Statistiques non locales dans les images : modélisation, estimation et échantillonnage

Valentin de Bortoli

université
PARIS-SACLAY

Non-local statistics in images: modelisation, estimation and sampling.

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 574, Ecole Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat: mathématiques appliquées
Unité de recherche: Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, 91190, Gif-sur-Yvette, France.
Référent: : ENS Paris-Saclay

Thèse présentée et soutenue à Paris, le 6 juillet 2020, par

Valentin DE BORTOLI

Composition du jury:

Yann Gousseau Professeur, Telecom ParisTech	Président
Bernard Delyon Professeur, IRMAR, Université de Rennes 1	Rapporteur
Stéphane Mallat Professeur, Collège de France et ENS Paris	Rapporteur
Alain Durmus Maître de conférence, Centre Borelli, ENS Paris-Saclay	Examinateur
Fabrice Gamboa Professeur, IMT, Université de Toulouse	Examinateur
Agnès Desolneux Directrice de recherche CNRS, Centre Borelli, ENS Paris-Saclay	Directrice
Bruno Galerne Professeur, Institut Denis Poisson, Université d'Orléans	Codirecteur
Arthur Leclaire Maître de conférence, IMB, Université de Bordeaux	Codirecteur

Thèse de doctorat

NNT: NNT2020UPASN020

Remerciements

Je tiens tout d'abord à remercier mes encadrants de thèse, Agnès Desolneux, Bruno Galerne et Arthur Leclaire. Votre complémentarité a fait de vous une équipe de choc pour l'encadrement de ma thèse. Agnès, merci pour ta disponibilité face à mes (très) nombreuses questions. Ton recul scientifique m'a été précieux tout au long de ma thèse. Bruno, merci de m'avoir guidé dans le monde de l'image, mes codes t'en remercient. Arthur, merci pour ta bienveillance et ta rigueur scientifique. Je garde en mémoire tes relectures détaillées et nos longues discussions (scientifiques ou non) autour d'un café. Merci aussi pour ton engagement auprès des étudiants, dont j'ai eu la chance de faire partie.

Je remercie sincèrement Bernard Delyon et Stéphane Mallat d'avoir accepté de rapporter ma thèse. Votre intérêt pour mon travail m'honore. J'adresse également mes remerciements à mes examinateurs Fabrice Gamboa et Yann Gousseau.

Je tiens également à remercier Alain Durmus qui m'a fait découvrir le monde des chaînes de Markov. Tes idées et ta force de travail m'ont réellement inspiré. J'espère que nous continuerons longtemps à travailler ensemble.

J'en profite également pour remercier Eric Moulines pour l'intérêt porté à mes travaux et son invitation à Moscou. Je remercie mes coauteurs Marcelo Pereyra, Ana Fernandez Vidal, Umut Şimşekli et Kimia Nadjahi. Merci Marcelo pour m'avoir invité à Édimbourg et pour ton enthousiasme. Merci Umut pour ta bonne humeur (et pour m'avoir fait découvrir certaines spécialités turques !).

Dans la "team Markov chains" je remercie également mes camarades de thèse Pablo, Achille et Aurélien. Les compétences en russe d'Achille et en espagnol de Pablo nous aurons épargné bien des tracas. Je vous souhaite bon courage pour la suite !

Je tiens à remercier l'ensemble du personnel du CMLA (ou centre Borelli) qui m'a accompagné avec beaucoup de patience lors de mes démarches administratives (ou lors de la recherche de papiers introuvables) : Delphine Laverne, Alina Müller, Virginie Pauchont, Sandra Doucet et Véronique Almadovar.

Je remercie également toute l'équipe enseignante du département de mathématiques de l'ENS Cachan (ou Paris Saclay) : Frédéric Pascal, Alain Trouvé, Sandrine Dallaporta et Tuong-Huy Nguyen. Merci pour votre dévouement. J'en profite également pour remercier une nouvelle fois Arthur pour son engagement vis à vis des élèves. Je remercie aussi Miguel et Argyris pour leur bonne humeur et leur humour lors des réunions de préparation de TP.

Merci à tous les doctorant-e-s du CMLA (ou centre Borelli). En particulier merci à Xavier (qui a relu avec attention nombre de mes preuves fausses), Pierre et ses histoires rocambolesques, Tina et Pashmina pour leur bonne humeur et leur énergie à toute épreuve. Merci à Jérémy qui a bien voulu réparer mes bourdes informatiques (promis je fais la pub de vpv). Merci aussi à la team hispanophone dont Mariano et Roger (partenaires de trop peu de parties de foot).

Merci aussi à tous les doctorant-e-s du bureau 725-C1. Malgré ma présence fantomatique je ne vous

ai pas oubliés ! Merci à Anne-Sophie, Alkéos, Cambyse, Alexandre et Antoine. Merci à Claire, ma "grande soeur de thèse", toujours motivée et enthousiaste. Merci à Rémi qui incarne la relève. Merci aussi aux "moutons" Anaïs, Pierre (le doyen de la bande), Vincent (bon courage pour la rédaction) et Newton.

Je me remercie également les Suricates cachanais et affiliés : Clément, Corentin, Margo, Quentin, Manon, Ron, Camille, Sarah et Antoine. Merci Clément pour ton énergie débordante. Merci Corentin pour ton humour et ton amour du Lot. Merci Margo pour ta passion pour le soleil et les soirées. Je me souviendrai longtemps de nos apéritifs cachanais. Merci Quentin, colocataire incroyable qui a su partager mon amour du riz, des listes et des productions cinématographiques chinoises à bas budget.

Merci à toute ma famille. Merci à mes grands parents pour leur gentillesse. Merci à Pascal et Antonia pour leur bienveillance, promis nous dégusterons bientôt de la tomme au marc. Merci à mon frère Jules. Ton implication dans nos parties de Monopoly m'a accompagné durant la fin de ma thèse. Merci à mes parents pour leur soutien. Merci de m'avoir toujours accompagné et encouragé dans mes choix. C'est vraiment grâce à vous que j'en suis là aujourd'hui.

Enfin, merci à Jeanne. Ton soutien sans faille, ta gentillesse et ta bonne humeur me portent depuis cinq ans et encore plus ces derniers mois. J'espère être à la hauteur lorsque viendra ton tour. Je nous souhaite de profiter ensemble des tous les moments qui nous attendent.

Contents

Chapter 1

Introduction (Français)

Dans cette thèse on s'intéresse au traitement statistique (modélisation, inférence et échantillonnage) de mesures non locales dans les images. On se concentre principalement sur deux de ces mesures : une notion de redondance spatiale définie a contrario et des statistiques issues de réseaux de neurones profonds. Plus précisément, on va tenter de répondre aux questions suivantes :

- Qu'est-ce que la redondance spatiale ? Comment l'identifier dans les images naturelles ?
- Quels modèles stochastiques d'images permettent d'incorporer des contraintes sur les réponses de réseaux de neurones ? Comment obtenir des échantillons de ces modèles ?

En Section 1.1, on rappelle l'importance de la notion de redondance spatiale dans le cadre d'une théorie de la perception visuelle : la théorie de la gestalt. On introduit également la méthode a contrario qui constitue le cadre statistique de cette approche et on rappelle ses liens avec la théorie statistique des tests (voir [DMM08, Chapter 15.3.2] pour une discussion sur les liens entre la méthode a contrario et la théorie des tests multiples). Pour utiliser une telle méthode, il est nécessaire de définir un modèle d'image de fond (ou de bruit). Ces modèles sont donnés par des champs aléatoires gaussiens qui définissent une classe particulière d'images de texture.

Le problème général de la synthèse de textures par l'exemple est rappelé en Section 1.2. Il s'agit de l'interprétation via le traitement d'images d'un problème plus général : la synthèse de champs aléatoires. On présente un exemple d'algorithme de synthèse de textures basé sur des statistiques d'ondelettes et plus tard étendu aux statistiques de réseaux de neurones. Cet algorithme peut s'interpréter comme une procédure de maximisation d'entropie sous contraintes presque-sûres. En utilisant le principe de maximum d'entropie qui correspond à une relaxation du problème précédent on est en mesure de définir explicitement des distributions sur l'espace des images satisfaisant des contraintes sur les réponses d'un réseau de neurones convolutionnel. Afin de résoudre le problème de maximisation associé, on s'appuie sur des algorithmes d'optimisation stochastique dont on rappelle les principes.

En supposant que la distribution des images qui vérifient les statistiques non locales est donnée (par le principe de maximum d'entropie par exemple), on s'intéresse au problème d'échantillonnage de telles lois en Section 1.3. On décrit l'algorithme de Langevin non-ajusté qui permet un échantillonnage efficace de lois en grande dimension. L'analyse de la convergence de discrétisations de diffusions permet d'établir les propriétés favorables de cet algorithme vis-à-vis de la dimension du problème initial.

Les contributions de cette thèse sont détaillées en Section 1.4. Un algorithme basé sur la méthode a contrario pour détecter la redondance spatiale est décrit en Section 1.4.1. Les applications de

cet algorithme au traitement d'images, et notamment à des problèmes de débruitage et de détection de périodicité sont énoncées en Section 1.4.2. En Section 1.4.3, on présente un théorème portant sur l'ergodicité pour diverses distances de Wasserstein de certaines chaînes de Markov issues de modèles fonctionels auto-régressifs. Ces résultats quantitatifs de convergence sur les chaînes de Markov sont utilisés pour assurer la convergence de l'algorithme d'approximation stochastique décrit en Section 1.4.4 : l'algorithme *Stochastic Optimization with Unadjusted Langevin* (SOUL). En Section 1.4.5, on montre que certains modèles de synthèse de texture par maximum d'entropie peuvent être obtenus via l'algorithme SOUL. Enfin, on présente quelques résultats de synthèse de texture par l'exemple dans la Section 1.4.6. Il est également à noter qu'une revue de littérature est proposée au début de chaque chapitre.

1.1 Redondance spatiale, méthodes a contrario et champs aléatoires

1.1.1 Une première définition et théorie de la gestalt

La notion de "redondance spatiale", comme celle de "texture" ou de "similarité", est un concept dont la définition peut varier d'un auteur à l'autre au gré des applications et des différents points de vue adoptés. Avant de présenter un cadre statistique rigoureux, on va tenter d'apporter quelques éléments de réponse concernant les attentes visuelles et perceptuelles vis à vis du concept de redondance spatiale.

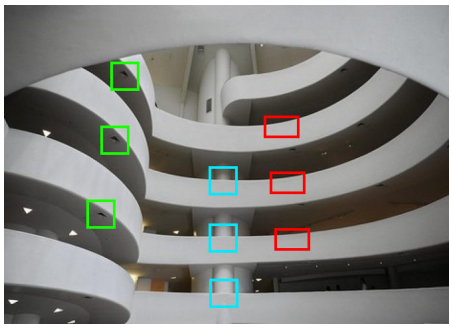
Intuitivement, cette redondance correspond à la répétition d'un motif dans une image et doit donc s'envisager dans un cadre local. Pour répondre à cette nécessité, on va considérer des *patchs* (ou imagettes) dans une image. Supposons qu'une image x soit assimilée à une fonction $x : \mathsf{E} \to \mathsf{V}$, où E est un domaine de $\mathbb{Z}^2$ ou $\mathbb{R}^2$ correspondant à l'ensemble des positions des pixels de l'image, et V est un sous-ensemble de $\mathbb{R}$ (pour une image en niveaux de gris) ou de $\mathbb{R}^3$ (pour une image couleur) correspondant à l'ensemble des valeurs des pixels de l'image. Un patch correspond alors à la restriction de x à un sous-domaine de E noté w. Plus précisement, pour un sous-domaine $\mathsf{w} \subset \mathsf{E}$ on considère l'opérateur P_{w} tel que pour toute image $x : \mathsf{E} \to \mathsf{V}$ on a $P_{\mathsf{w}}(x) : \mathsf{w} \to \mathsf{V}$ et pour toute position $p \in \mathsf{w}$, $P_{\mathsf{w}}(x)(p) = x(p)$. Cet opérateur extrait de l'image x le patch correspondant au domaine w.

Soit un domaine $\mathsf{w}_1 \subset \mathsf{E}$ et sa version translatée de vecteur $t \in \mathbb{Z}^2$, $\{p + t : p \in \mathsf{w}_1\}$. On dit que les patchs associés sont similaires si $s(P_{\mathsf{w}_1}(x), P_{\mathsf{w}_1}(\tau_{-t}(x)))$ a une faible valeur, où $s : \mathsf{V}^{\mathsf{w}_1} \times \mathsf{V}^{\mathsf{w}_1} \to [0, +\infty)$ est une fonction de similarité que l'on définira plus tard et où pour tout $p \in \mathsf{E}$, $\tau_t(x)(p) = x(p - t)$ (on a supposé que E était stable par translation). On définit alors la redondance spatiale de la manière suivante : "Une image est redondante spatialement si un grand nombre de ses patchs sont similaires."

Un premier exemple de redondance peut s'observer dans la Figure 1.1-(a). La redondance spatiale permet l'émergence de structures dans l'image (les différents étages dans la photographie du musée du Guggenheim à New York). Cependant, ce n'est pas le seul principe organisateur pour comprendre une image. Parmi ces principes on peut citer la couleur, la forme, le matériau, la continuité, etc. Certaines de ces propriétés peuvent être identifiées dans le tableau de Kandinsky "Black Lines" (1913) Figure 1.1-(b).

La redondance spatiale permet l'identification de nouvelles propriétés de l'image en ne considérant que les positions des patchs similaires, interprétées comme un processus de points. On réduit alors l'analyse de l'image à l'analyse géométrique d'un processus de points. Les propriétés perceptuelles associées à de tels ensembles ont été l'objet d'étude de la théorie de la *gestalt* (forme en allemand).

Cette théorie de la perception a pour origine les travaux sur les illusions d'optique de Wertheimer, Koffka et Köhler [Wer23; Kof13; Köh92]. D'un point de vue visuel elle repose sur le principe de groupe-



(a)

(b)

Figure 1.1: **Lois de groupements.** Une photographie de l'intérieur du musée Guggenheim à New York (a) présente de nombreuses redondances spatiales identifiées par les patchs aux bords rouge, vert et cyan. Au contraire, aucune redondance spatiale n'apparaît, au premier abord, dans le tableau de Kandinsky "Black Lines" (1913) mais l'œuvre obéit à d'autres règles d'organisation (couleur, courbure, convexité).

ment : les composants d'une image peuvent certes être compris indépendamment les uns des autres mais prennent un sens nouveau lorsqu'ils sont associés. Wertheimer [WW59] déclare ainsi :

> The basic thesis of gestalt theory might be formulated thus: there are contexts in which what is happening in the whole cannot be deduced from the characteristics of the separate pieces, but conversely; what happens to a part of the whole is, in clearcut cases, determined by the laws of the inner structure of its whole.
>
> *Max Wertheimer (1959)*

En Section 3.2 on étudiera comment la redondance spatiale et les principes de la théorie de la gestalt peuvent être utilisés dans le cadre du traitement d'images. On se concentrera sur deux applications : le débruitage par patchs, voir Section 3.2.4 et l'analyse de périodicité, voir Section 3.2.5.

1.1.2 Méthodes a contrario pour l'image

De nombreux efforts ont été menés pour aboutir à une présentation mathématique des principes de groupement [Lin97], mais c'est véritablement avec les travaux de David Lowe [Low12] que la théorie statistique de la gestalt a pris sa forme moderne. Ce cadre mathématique repose sur la méthode a contrario. Un évènement (l'alignement de points, la redondance spatiale de patchs) est dit significatif s'il a une très faible probabilité de se produire dans un modèle de fond (ou modèle a contrario ou modèle de bruit). Cette règle a par la suite été nommée principe de Helmholtz dans les travaux de Desolneux, Moisan et Morel [DMM08], en référence aux études de Helmholtz sur les illusions d'optique [Hel25]. Dans le cadre a contrario, le modèle de fond doit être choisi en fonction des propriétés de l'image que l'on veut discriminer. Afin de rendre compte de ce modèle statistique, on considère un espace probabilisé sous-jacent $(\Omega, \mathcal{F}, \mathbb{P})$.

Ce cadre est similaire à celui de la théorie statistique des tests dont on rappelle les principes pour l'étude des méthodes a contrario. Le but de la théorie des tests est de déterminer des critères afin de différencier une hypothèse nulle d'une hypothèse alternative. Soit U une variable aléatoire à valeurs dans $(\mathsf{V}^{\mathsf{E}}, \mathcal{B}(\mathsf{V}^{\mathsf{E}}))$, où $\mathcal{B}(\mathsf{V}^{\mathsf{E}})$ est l'ensemble des boréliens de V^{E}, et V^{E} est muni de la topologie produit.

On dit que U est un champ aléatoire, voir [Adl81]. On note $\mathbb{P}_U$ la distribution de U, *i.e.* la mesure image de $\mathbb{P}$ par U. Soit $\rho_0 \in \mathscr{P}(\mathsf{V}^{\mathsf{E}})$, où $\mathscr{P}(\mathsf{V}^{\mathsf{E}})$ est l'ensemble des mesures de probabilités sur $\mathcal{B}(\mathsf{V}^{\mathsf{E}})$. L'hypothèse nulle est alors $\mathbb{P}_U = \rho_0$ et l'hypothèse alternative est $\mathbb{P}_U \neq \rho_0$. Dans le cadre a contrario, ρ_0 est le modèle de fond (ou modèle de bruit, ou encore modèle a contrario). En Section 1.1.3 on propose un choix de modèle de fond pour l'analyse de la redondance spatiale. On définit la statistique de test suivante,

$$\mathrm{NFA} = \sum_{i=1}^{N} \mathbb{1}_{f_i(U) \in \mathsf{A}} \ ,$$

où $N \in \mathbb{N}^\star$, $(f_i)_{i \in \{1,\dots,N\}}$ est une famille de fonctions mesurables telle que pour tout $i \in \{1,\dots,N\}$, $f_i : (\mathsf{V}^{\mathsf{E}}, \mathcal{B}(\mathsf{V}^{\mathsf{E}})) \to (\mathsf{V}^{\mathsf{E}}, \mathcal{B}(\mathsf{V}^{\mathsf{E}}))$. La statistique de test NFA est le nombre de fausses alarmes. Dans le cadre du traitement d'images, l'évènement A représente le plus souvent une propriété d'intérêt de l'image, par exemple $\mathsf{A} = \{x \in \mathsf{V}^{\mathsf{E}} : \|P_{\mathsf{w}}(x) - P_{\mathsf{w}}(x_0)\| \leq \varepsilon\}$ où $x_0 \in \mathsf{V}^{\mathsf{E}}$ et $\varepsilon > 0$. En choisissant un tel ensemble A et $(f_i)_{i \in \{1,\dots,N\}} = (\tau_t)_{t \in \mathsf{E}}$ (on a supposé que E était stable par translation), le nombre de fausses alarmes est élevé dans un modèle de fond U si un grand nombre de patchs de U sont similaires à ceux de x_0.

On rejette l'hypothèse nulle si le nombre de fausses alarmes est plus grand qu'un nombre maximal de fausses alarmes, $\mathrm{NFA}_{\max}$. Le risque de première espèce α est alors $\alpha = \mathbb{P}(\mathrm{NFA} \geq \mathrm{NFA}_{\max})$. En utilisant l'inégalité de Markov on a

$$\alpha = \mathbb{P}\left(\mathrm{NFA} \geq \mathrm{NFA}_{\max}\right) \leq \mathrm{NFA}_{\max}^{-1} \sum_{i=1}^{N} \mathbb{P}\left(f_i(U) \in \mathsf{A}\right) \ .$$

Dans le cadre a contrario on dit que l'évènement A est ε-significatif si $\sum_{i=1}^{N} \mathbb{P}\left(f_i(U) \in \mathsf{A}\right) \leq \varepsilon$. On peut alors donner une interprétation a contrario de la statistique de test NFA de la manière suivante : pour tout $i \in \{1,\dots,N\}$ on considère d_i tel que pour tout $x \in \mathsf{V}^{\mathsf{E}}$, $d_i = \mathbb{1}_{\mathsf{A}}(f_i(x))$. On dit que l'indice $i \in \{1,\dots,N\}$ est détecté dans l'image x si $d_i(x) = 1$. On a alors $\mathrm{NFA} = \sum_{i=1}^{N} d_i(U)$. Si A est un évènement ε-significatif on obtient

$$\mathbb{E}\left[\mathrm{NFA}\right] = \mathbb{E}\left[\sum_{i=1}^{N} d_i(U)\right] \leq \varepsilon \ .$$

Autrement dit, ε donne une borne sur le nombre moyen de détections dans le modèle de fond. Dans le cadre de la détection de redondance la famille $(f_i)_{i \in \{1,\dots,N\}}$ sera donnée par $(\tau_t)_{t \in \mathsf{E}}$ (en supposant que E est stable par translation), voir Section 3.2. On a donc $N = |\mathsf{E}|$ (où $|\mathsf{E}|$ est le cardinal de E). Étant donné un patch $u_0 : \mathsf{w} \to \mathsf{V}$ on définit alors

$$\mathsf{A} = \left\{x \in \mathsf{V}^{\mathsf{E}} : s(P_{\mathsf{w}}(x), u_0) \leq \mathrm{v}\right\} \ ,$$

où $\mathrm{v} \in \mathbb{R}$ est une valeur à fixer et s une fonction de similarité (par exemple la norme $\|\cdot\|_2$). Dans ce cadre, un décalage $t \in \mathsf{E}$ est détecté dans une image x si et seulement si $s(P_{\mathsf{w}}(\tau_t(x)), u_0) \leq \mathrm{v}$. Si le modèle de fond ρ_0 est stationaire, *i.e.* si la loi de U est invariante par translation, alors on a pour tout $t \in \mathsf{E}$, $\mathbb{P}_{\tau_t(U)} = \mathbb{P}_U = \rho_0$. Afin de fixer v, on considère une borne sur le nombre moyen de détections ε dans le modèle de fond. Il convient donc de choisir $\mathrm{v} \in \mathbb{R}$ tel que

$$\rho_0(\mathsf{A}) = \mathbb{P}\left(s(P_{\mathsf{w}}(U), u_0) \leq \mathrm{v}\right) = \varepsilon / |\mathsf{E}| \ ,$$

où on rappelle que s est une fonction de similarité. En notant CDF la fonction de répartition de $s(P_{\mathsf{w}}(U), u_0)$ et ICDF sa fonction de répartition inverse, on peut alors déterminer si un décalage $t \in \mathsf{E}$

(a) (b)

Figure 1.2: **Un algorithme de détection de segments.** En (a) on présente l'image originale et en (b) une image binaire. Les segments noirs correspondent aux segments détectés par l'algorithme LSD, voir http://demo.ipol.im/demo/gjmr_line_segment_detector pour une démonstration en ligne.

est détecté de la manière suivante : $t \in \mathsf{E}$ est détecté si et seulement si $s(P_{\mathsf{w}}(\tau_t(x)), u_0) \leq \mathrm{ICDF}(\varepsilon/|\mathsf{E}|)$ ou $\mathrm{CDF}(s(P_{\mathsf{w}}(\tau_t(x)), u_0)) \leq \varepsilon/|\mathsf{E}|$. En appliquant cette procédure pour chaque décalage possible, on obtient une image binaire qui correspond aux décalages détectés dans l'image x, *i.e.* $(d_t(x))_{t\in\mathsf{E}} \in \{0,1\}^{\mathsf{E}}$. Cette image binaire correspond à un processus de points qui peut être analysé selon les principes de la théorie de la gestalt.

On a ici présenté le cadre a contrario et donné un exemple pour l'étude de la redondance spatiale dans les images naturelles. Néanmoins il existe de nombreuses autres applications de l'approche a contrario pour l'étude des lois de groupements de la gestalt [DMM00; DMM01; ADV03; Cao04; VG+08; Dav+18]. Par exemple, l'algorithme *Line Segment Detection* (LSD) [VG+08] a pour but de détecter les alignements dans les images naturelles. Dans ce cadre on définit $\mathsf{A} = \{x \in \mathsf{V}^{\mathsf{E}},\ k \text{ points sont alignés dans } r_0(x)\}$ où $r_0(x)$ est une sous-image de x de forme rectangulaire et $k \in \mathbb{N}^\star$. La notion d'alignement est définie sur les angles des gradients de l'image, voir [VG+08]. Soit $(f_i)_{i\in\{1,\dots,N\}}$ l'ensemble des translations-dilatations-rotations de E, voir [VG+08] pour une définition et un dénombrement de ces transformations. Les auteurs de [VG+08] proposent un modèle de fond ρ_0 tel que $\rho_0(\mathsf{A})$ peut être calculé explicitement. Étant donné un niveau de significativité ε, l'ensemble des détections correspond à l'ensemble des rectangles détectés pour ce niveau. On peut alors superposer ces rectangles pour produire une image binaire. C'est le résultat de l'algorithme LSD illustré en Figure 1.2.

Dans cette thèse on présente un algorithme de détection pour l'analyse de similarité (à la fois entre deux images et au sein d'une même image) basé sur les principes décrits précédemment. Étant donné une image x on peut alors obtenir une image des décalages détectés. On utilise cette information pour améliorer un algorithme de débruitage par patch, voir Section 3.2.4, et pour conduire une analyse de périodicité, voir Section 3.2.5. Afin de conclure notre présentation du cadre statistique utilisé il convient de décrire la classe de champs aléatoires utilisée pour définir le modèle de fond ρ_0 : les champs aléatoires gaussiens.

1.1.3 Texture et champs aléatoires

On rappelle qu'un champ aléatoire U est une variable aléatoire à valeurs dans $(\mathsf{V}^{\mathsf{E}}, \mathcal{B}(\mathsf{V}^{\mathsf{E}}))$ [Adl81]. Ainsi, pour tout élément $\omega \in \Omega$, on a $U(\omega) \in \mathsf{V}^{\mathsf{E}}$. Par abus de notation, on omet la dépendance du champs en ω lorsqu'il n'existe pas d'ambiguïté. Un champ aléatoire U est dit gaussien si pour tout

(a) (b) (c)

Figure 1.3: **Quelques exemples de textures réelles.** La texture (a) ne présente pas de structure particulière tandis que la texture (c) est extrêmement régulière. La texture (b) présente une régularité intermédiaire puisque les épines de pins sont disposées en cercle autour de chaque branche.

ensemble de points $(p_1, \dots, p_n) \in \mathsf{E}^n$, $(U(p_1), \dots, U(p_n))$ est un vecteur gaussien. Dans cette thèse, le modèle de fond utilisé dans l'analyse statistique de la redondance spatiale introduite en Section 1.2.2 sera un modèle de champ aléatoire gaussien. On rappelle que ce modèle de fond correspond à l'hypothèse nulle dans le cadre de la détection de redondance spatiale. De manière informelle, le modèle de fond est un modèle d'images dont chaque réalisation n'exhibe pas (ou peu) de redondance spatiale.

On peut trouver des exemples de champs aléatoires pour la modélisation d'images, et plus spécifiquement pour la synthèse de texture, dans les travaux pionniers de Cross et Jain, [CJ83]. On rappelle brièvement l'objet de la synthèse de texture par l'exemple : étant donné une texture de départ x_0 (la texture exemple), est-il possible de trouver un champ aléatoire U tel que les réalisations de U ressemblent à x_0 sans que ces réalisations soient des copies de x_0 ? On répertorie quelques exemples de texture en Figure 1.3. Un facteur déterminant pour la classification de textures est leur régularité. Une texture très structurée, telle que celle présentée en Figure 1.3-(c) sera appelée *macrotexture*. Au contraire, une texture n'exhibant aucune propriété de structure particulière, telle que celle présentée en Figure 1.3-(a), sera appelée *microtexture*.

Van Wijk [Wij91] introduit l'utilisation de *spots* pour la génération de texture. Le procédé de synthèse est le suivant : on sélectionne un motif (le *spot*) et on génère un processus de points (processus de Poisson ou processus de Bernoulli) sur la grille sous-jacente à l'image, l'ensemble E. Il s'agit ensuite de centrer chacun des spots sur les points du domaine. Ce point de vue est plus tard étendu [GGM11] dans le cas où E est une grille périodique, *i.e.*, $\mathsf{E} = \mathbb{Z}/(M\mathbb{Z}) \times \mathbb{Z}/(N\mathbb{Z})$ avec $M, N \in \mathbb{N}$. Les auteurs considèrent la limite lorsque le nombre de points tend vers l'infini. En notant x_0 le spot et après renormalisation et recentrage, on identifie un champ limite gaussien U tel que pour toutes positions du domaine $p_1, p_2 \in \mathsf{E}$,

$$\mathbb{E}\left[U(p_1)\right] = |\mathsf{E}|^{-1} \sum_{p \in \mathsf{E}} x_0(p) \ , \qquad \mathrm{Cov}\left[U(p_1), U(p_2)\right] = |\mathsf{E}|^{-1} \sum_{p \in \mathsf{E}} x_0(p_1 - p_2 + p) x_0(p) \ . \tag{1.1}$$

Notons que le champ U est un champ stationnaire, ce qui permet d'assurer une certaine homogénéité en espace (mais qui ne suffit pas à caractériser l'ensemble des textures). Ce champ aléatoire est un candidat de choix pour un modèle stochastique de texture via l'exemple x_0. De nombreuses expériences [GL17; GLM14; GLR18] montrent que ce champ aléatoire gaussien permet effectivement de reproduire fidèlement les microtextures dans la plupart des cas, voir Figure 1.4. Par contre, cette méthode échoue

(a) (b)

Figure 1.4: **Champs gaussien et microtexture.** Les images (exemples) de microtexture sont exposées en (a). En (b) on présente des réalisations du champ gaussien donné par (1.1). Image extraite de [GGM11].

à reproduire des textures possédant des structures complexes, ce qui est le cas des macrotextures, voir Figure 1.5.

On tire de ces deux figures la conclusion suivante : "le modèle gaussien associé à x_0 ne conserve que l'information de microtexture de x_0". Ce désavantage pour la synthèse de texture structurée peut être utilisé en notre faveur dans le cadre de la détection a contrario. En effet, on peut exploiter cette information en choisissant ce champ aléatoire comme modèle de fond, *i.e.* comme hypothèse nulle dans le cadre d'un test pour identifier la redondance spatiale dans une image. Dans le Chapitre 3, en combinant les méthodes a contrario et les modèles de fond donnés par des champs gaussiens on introduit un algorithme de détection de la redondance spatiale et on propose de nouveaux algorithmes de débruitage et détection de périodicité dont on peut démontrer certaines propriétés théoriques.

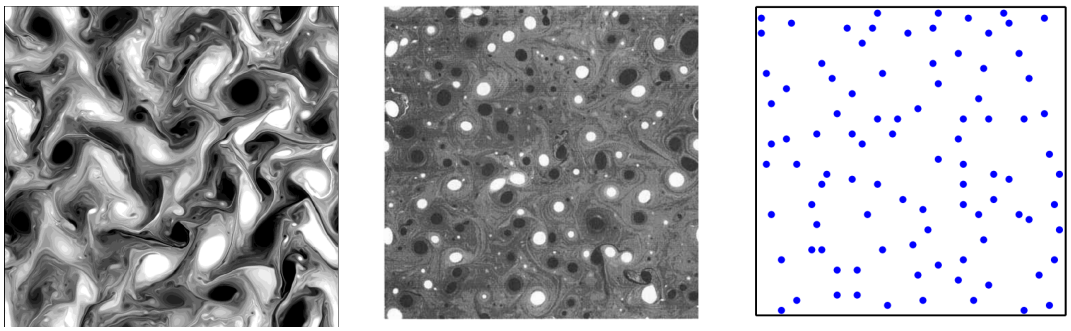
Jusqu'à maintenant on s'est intéressé à l'estimation de la redondance spatiale dans les images naturelles. Cette redondance peut être identifiée en réfutant l'hypothèse de la simple coïncidence dans un modèle a contrario. Cette réfutation a l'avantage de ne nécessiter que des modèles d'image non-structurés (par exemple des champs gaussiens). Malheureusement, la simplicité de ce processus de réfutation cache une limitation du modèle : il ne fournit aucune procédure pour enrichir le modèle naïf avec les structures détectées, *i.e.* il ne permet pas l'échantillonnage de champs aléatoires plus complexes.

1.2 Synthèse de champs aléatoires et principe de maximum d'entropie

On a vu dans la section précédente que les microtextures, *i.e.* les champs aléatoires sans dépendance à longue portée, peuvent être échantillonnées de manière satisfaisante en utilisant des champs aléatoires gaussiens. Néanmoins de nombreux champs aléatoires étudiés en mécanique des fluides, astro-

(a) (b)

Figure 1.5: **Champ gaussien et macrotexture.** L'image de macrotexture exemple est exposée en (a). En (b) on présente une réalisation du champ gaussien donné par (1.1). Image extraite de [GGM11].



(a) (b) (c)

Figure 1.6: **Champs aléatoires.** En (a) et (b) on présente quelques images de champs turbulents issus de modèles de mécanique des fluides. Images extraites de [Hel+95]. En (c) on présente une réalisation d'un processus déterminantal qui est un processus de points avec des contraintes de répulsion. Image extraite de [KT11].

physique ou encore en traitement d'images ne sont pas des microtextures et présentent des interactions aux longues échelles. On donne quelques exemples issus de la physique en Figure 1.6.

Dans cette thèse on s'intéresse au problème de synthèse de champs aléatoires du point de vue du traitement d'images, via la synthèse de textures par l'exemple. Ce problème constitue une première étape vers l'échantillonnage de champs aléatoires structurés.

1.2.1 Synthèse de texture paramétrique

On rappelle que le problème de synthèse de texture par l'exemple correspond à la problématique suivante. Étant donné une texture initiale x_0, comment synthétiser des nouvelles images telles que :

- leur contenu structurel soit similaire à celui de x_0,
- les images ne soient pas de simples reproductions de x_0 mais contiennent une part d'innovation.

On recense deux approches principales [Raa+17] pour traiter ce problème: l'approche *paramétrique* et l'approche *non-paramétrique* (ou approche par patchs). Dans cette thèse, on se concentre sur les

méthodes paramétriques. On reporte à la Section 5.2.1 une revue de littérature sur le sujet. Dans la suite on note $\mathsf{X} = \mathsf{V}^{\mathsf{E}}$ et on suppose que $\mathsf{X} \in \mathcal{B}(\mathbb{R}^d)$ pour un certain $d \in \mathbb{N}$.

Un des premiers algorithmes paramétriques pour la synthèse de texture par l'exemple a été proposé par Portilla et Simoncelli [PS00]. On commence par identifier un certain nombre de statistiques (locales ou non locales) qui vont constituer un dictionnaire de contraintes pour décrire la texture cible. Étant donné $p \in \mathbb{N}$ et une famille de fonctions mesurables $(f_i)_{i\in\{1,\dots,p\}}$ telle que pour tout $i \in \{1,\dots,p\}$, $f_i : (\mathsf{X}, \mathcal{B}(\mathsf{X})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et $x_0 \in \mathsf{X}$, où x_0 est la texture cible et $(f_i)_{i\in\{1,\dots,p\}}$ sont des fonctions de contrainte, on définit $\mathsf{Y} = \{x \in \mathsf{X},\ f_i(x) = f_i(x_0)\}$. Cet ensemble est appelé *ensemble de Julesz* dans [PS00]. Supposons qu'il existe une fonction mesurable $\Pi : (\mathsf{X}, \mathcal{B}(\mathsf{X})) \to (\mathsf{Y}, \mathcal{B}(\mathsf{Y}))$, alors tout champ aléatoire U à valeurs dans X donne un champ aléatoire à valeurs dans Y par composition avec Π. Malheureusement, il est très rare de disposer d'une fonction Π de manière explicite.

Si pour tout $i \in \{1,\dots,p\}$ on a accès à $\Pi_i : (\mathsf{X}, \mathcal{B}(\mathsf{X})) \to (\mathsf{X}, \mathcal{B}(\mathsf{X}))$ telle que pour tout $x \in \mathsf{X}$, $f_i(\Pi_i(x)) = f_i(x_0)$, on peut alors définir la suite suivante

$$X_{n+1} = \Pi_{n-\lfloor n/p \rfloor p+1}(X_n)\ , \tag{1.2}$$

où X_0 est une variable aléatoire à valeur dans X. Cette approche a été utilisée par Heeger et Bergen [HB95] dans le cas où pour tout $i \in \{1,\dots,p\}$, Π_i correspond à une égalisation d'histogrammes. Il n'existe pas de résultat général de convergence d'un tel algorithme sauf dans le cas où $(\Pi_i)_{i\in\{1,\dots,p\}}$ est une collection d'opérateurs strictement quasi non-expansifs [BC11, Corollaire 4.50, Théorème 5.23]. En particulier, on peut obtenir des résultats de convergence si $(\Pi_i)_{i\in\{1,\dots,p\}}$ est une collection d'opérateurs de projection sur des ensembles convexes et fermés. Portilla et Simoncelli remplacent (1.2) par une projection dans la direction du gradient au point courant, qui ne requiert que l'information de gradient des fonctions de contraintes $(f_i)_{i\in\{1,\dots,p\}}$, voir [PS00, Section 1.5].

Si Heeger et Bergen utilisent des données d'histogrammes pour construire leur ensemble de contraintes, Portilla et Simoncelli identifient 710 contraintes pour définir leur modèle d'image. Parmi celles-ci, on trouve : les statistiques marginales des coefficients d'une décomposition en ondelettes, les corrélations entre certains coefficients de la décomposition en ondelettes et des statistiques de phase. Gatys [GEB15] a étendu l'algorithme [PS00] en considérant des statistiques de matrices de Gram pour différentes couches d'un réseau de neurones pré-entrainé. Ce dernier algorithme qui donne des résultats proches de l'état de l'art, consiste en une descente de gradient pour une fonction de perte $\ell : \mathsf{X} \to [0, +\infty)$ donnée pour tout $x \in \mathsf{X}$ par

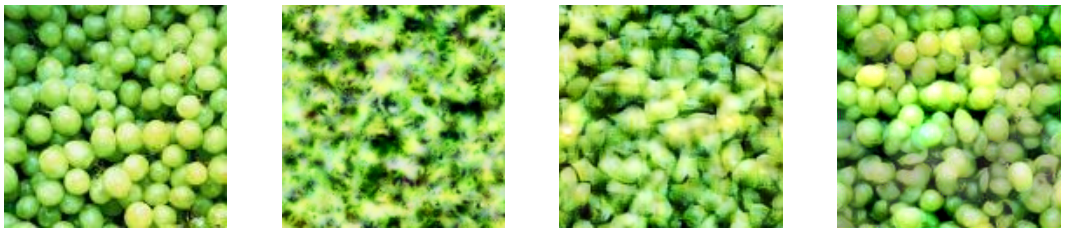
$$\ell(x) = \sum_{i=1}^{L} \lambda_i \|G_i(x) - G_i(x_0)\|_{\mathrm{Fr}}^2\ ,$$

où $(\lambda_i)_{i\in\{1,\dots,L\}} \in [0, +\infty)^L$ est une suite de poids, $G_i(x)$ est la matrice de Gram de la sortie du réseau de neurones VGG-19 [SZ14] appliqué à x à la couche i et $\|\cdot\|_{\mathrm{Fr}}$ est la norme de Frobenius. En Figure 1.7, on illustre les trois algorithmes décrits pour la synthèse d'une macrotexture.

1.2.2 Le principe de maximum d'entropie

Puisque les méthodes de synthèse considérées dans cette thèse relèvent du domaine paramétrique, le problème de synthèse de texture par l'exemple peut être reformulé comme un problème inverse. Dans cette section, on suppose que X est un espace topologique muni de la tribu associée $\mathcal{B}(\mathsf{X})$. Étant donné $p \in \mathbb{N}$ et une famille de fonctions mesurables $(f_i)_{i\in\{1,\dots,p\}}$ telle que pour tout $i \in \{1,\dots,p\}$ $f_i : (\mathsf{X}, \mathcal{B}(\mathsf{X})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et $x_0 \in \mathsf{X}$, on cherche π, probabilité sur $\mathcal{B}(\mathsf{X})$, telle que pour tout $i \in \{1,\dots,p\}$,

$$\int_{\mathsf{X}} \mathbb{1}_{f_i(x)\neq f_i(x_0)} \mathrm{d}\pi(x) = 0\ , \tag{1.3}$$



(a) (b) (c) (d)

Figure 1.7: **Quelques algorithmes de synthèse de texture.** En (a) on présente l'image de texture exemple. En (b) on donne le résultat obtenu avec la synthèse pyramidale proposée dans [HB95] En (c) on donne la synthèse obtenue avec [PS00] où les contraintes spatiales sont données par des ondelettes. Enfin en (d), on applique la méthode décrite dans [GEB15] pour générer la texture. Cette méthode, où les contraintes sont données par des sorties de réseaux de neurones, permet de synthésiser les structures complexes de l'image.

c'est-à-dire que $f_i = f_i(x_0)$, π presque-sûrement. On va ici considérer une version relâchée de (1.3), où les contraintes sont imposées en espérance plutôt que presque-sûrement,

$$\int_{\mathsf{X}} |f_i(x)| \mathrm{d}\pi(x) < +\infty \;, \qquad \int_{\mathsf{X}} f_i(x) \mathrm{d}\pi(x) = f_i(x_0) \;. \tag{1.4}$$

Dans la suite, on note $F : \mathsf{X} \to \mathbb{R}^p$, telle que pour tout $x \in \mathsf{X}$, $F(x) = (f_1(x) - f_1(x_0), \ldots, f_p(x) - f_p(x_0))$ et on suppose que la famille $\{f_i \,:\, i \in \{1, \ldots, p\}\} \cup 1$ est libre.

Évidemment, le problème inverse tel qu'énoncé est fortement mal posé et possède une solution triviale, $\pi = \delta_{x_0}$. Une manière de s'affranchir de ce problème est de chercher la distribution la plus uniforme possible étant donné ces contraintes. Dans le cadre des probabilités discrètes, Jaynes [Jay57] donne une solution utilisant l'entropie de Shannon [Sha48]. Supposons que $\mathsf{X} = \{1, \ldots, M\}$ avec $M \in \mathbb{N}$. Soit π une mesure de probabilité sur X. On définit l'entropie au sens de Shannon de la manière suivante

$$H(\pi) = -\sum_{x \in \mathsf{X}} \pi(x) \log(\pi(x)) \;,$$

où $0 \log(0) = 0$. On cherche alors à maximiser H sous les contraintes (1.4). Le modèle que l'on obtient est appelé modèle *macrocanonique*. Si on considère les contraintes (1.3) le modèle obtenu est le modèle *microcanonique*, étudié dans [BM18]. Les liens entre ces deux modèles (lorsque la dimension de l'espace d'image tend vers l'infini) sont discutés dans [BM18]. Dans cette thèse on s'intéresse au modèle macrocanonique mais on s'attachera à illustrer les liens que ce modèle entretient avec le microcanonique dans le Chapitre 5.

Maximiser H sous les contraintes (1.4) revient à la minimisation d'une fonction convexe de $\mathbb{R}^p$ sous contraintes linéaires et Jaynes obtient la solution $\pi^\star$ suivante : pour tout $x \in \mathsf{X}$ on a,

$$\pi^\star(x) = \exp[-\langle \theta^\star, F(x) \rangle] \Big/ \sum_{y \in \mathsf{X}} \exp[-\langle \theta^\star, F(y) \rangle] \;,$$

où $\theta^\star \in \mathbb{R}^p$ correspond aux multiplicateurs de Lagrange du problème contraint (1.4) et satisfait

$$\theta^\star \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^p} L(\theta) \;, \qquad L(\theta) = \log \left(\sum_{x \in \mathsf{X}} \exp[-\langle \theta, F(x) \rangle] \right) \;,$$

et L est appelé la *fonction de log-partition*. Jaynes [Jay79] identifie ce principe comme une généralisation du "principe d'indifférence" : en l'absence de contraintes, $\pi^\star$ doit être choisie uniforme. On se reportera à [Jay79] pour une discussion sur les origines du principe du maximum d'entropie. En traitement d'image, le principe de maximum d'entropie a d'abord été utilisé pour des problématiques de restauration [Wer+77; GD78; SB84; Bes86] avant d'être employé pour la synthèse d'image [ZWM98].

L'extension du principe de maximum d'entropie au cas où X n'est plus fini est délicate. On suppose désormais que $\mathsf{X} = \mathbb{R}^d$ avec $d \in \mathbb{N}$. Commençons par remarquer qu'il n'existe pas d'équivalent au principe d'indifférence dans ce cadre, *i.e.*, il n'existe pas de distribution de probabilité uniforme sur $\mathbb{R}^d$. De plus, en considérant pour toute mesure de probabilité π telle la quantité suivante lorsqu'elle est bien définie

$$H(\pi) = -\int_{\mathsf{X}} (\mathrm{d}\pi/\mathrm{d}\lambda)(x) \log\left[(\mathrm{d}\pi/\mathrm{d}\lambda)(x)\right] \mathrm{d}x \ ,$$

on a $H(\pi) \in [-\infty, +\infty]$ et non plus $[0, +\infty]$ comme dans le cas discret. On distingue alors deux approches. D'un côté [BLN96; TV93; Lóf8] établissent des équivalents du principe de maximum d'entropie pour une modification de l'entropie de Shannon en utilisant des techniques issues de l'analyse fonctionnelle (espaces de Birnbaum-Orlicz) et d'optimisation convexe dans des espaces de Banach. De l'autre, Csiszár et ses co-auteurs [Csi75; Csi84; Csi96; CGG99] munissent l'espace X d'une *mesure de probabilité de référence*, notée μ par la suite, et remplace l'entropie de Shannon de la probabilité π par l'opposé de la divergence de Kullback-Leibler [Kul97] entre π et μ. Cette divergence est définie entre deux mesures de probabilités ν_1, ν_2 par

$$\mathrm{KL}\left(\nu_1|\nu_2\right) = \begin{cases} \displaystyle\int_{\mathsf{X}} \log[(\mathrm{d}\nu_1/\mathrm{d}\nu_2)(x)]\mathrm{d}\nu_1(x) & \text{si } \nu_1 \ll \nu_2 \ , \\ +\infty & \text{sinon} \ . \end{cases}$$

Notons que $\mathrm{KL}\left(\nu_1|\nu_2\right) \geq 0$ avec égalité si et seulement si $\nu_1 = \nu_2$. Dans [Csi75], Csiszár parvient à généraliser le résultat de Jaynes en remplaçant l'entropie de Shannon par l'opposé de la divergence de Kullback-Leibler vis-à-vis de μ. Plus précisement on considère le problème suivant

$$\pi^\star \in \operatorname*{arg\,min}_{\pi \in \mathscr{P}_F} \mathrm{KL}\left(\pi|\mu\right) \ . \tag{1.5}$$

où $\mathscr{P}_F$ est l'ensemble des mesures de probabilités tel que pour tout $\pi \in \mathscr{P}_F$, $\int_{\mathsf{X}} F(x)\mathrm{d}\pi(x) = 0$. S'il existe une solution $\pi^\star$ à (1.5) qui admet une densité par rapport à μ alors, en utilisant [Csi75, Théorème 3.1], on a pour tout $x \in \mathsf{X}$,

$$(\mathrm{d}\pi^\star/\mathrm{d}\mu)(x) = \exp[-\langle\theta^\star, F(x)\rangle] \Big/ \int_{\mathsf{X}} \exp[-\langle\theta^\star, F(y)\rangle]\mathrm{d}\mu(y) \ , \tag{1.6}$$

pour un certain $\theta^\star \in \mathbb{R}^p$. Dans le même article, des conditions pour assurer l'existence d'un tel minimum sont établies. En effet, si pour tout $(\varepsilon_1, \ldots, \varepsilon_p) \in \mathrm{B}(0, r)$ avec $r > 0$, il existe π tel que $\mathrm{KL}\left(\pi|\mu\right) < +\infty$ et π satisfait (1.6) en remplaçant pour tout $i \in \{1, \ldots, p\}$, $f_i(x_0)$ par $f_i(x_0) + \varepsilon_i$ et si pour tout $\theta \in \mathbb{R}^p$, $\int_{\mathsf{X}} \exp[-\langle\theta, F(x)\rangle]\mathrm{d}\mu(x) < +\infty$ alors [Csi75, Théorème 3.3] assure l'existence d'une distribution exponentielle solution du problème de maximum d'entropie. Dans le cadre de la synthèse de textures considéré dans cette thèse, ce théorème est difficilement applicable et d'autres conditions suffisantes pour l'existence du modèle doivent être explorées.

Dans le Chapitre 5 on s'intéresse à ces modèles de maximum d'entropie dans le cadre de la synthèse de texture. On établit notamment des conditions faciles à vérifier sur le modèle F et sur la mesure de probabilité de référence μ afin que la solution du problème de minimisation existe. En particulier, dans le cas où F est donné par un réseau de neurones convolutionnel, on donne un certificat permettant de s'assurer de l'existence du modèle de maximum d'entropie.

1.2.3 Optimisation stochastique

Afin de mettre en place des algorithmes de synthèse de texture basés sur l'approche macrocanonique, nous devons répondre à deux problèmes : (a) Comment approcher $\theta^\star$ dans (1.6) ? (b) Étant donné $\theta^\star$, comment échantillonner selon la densité $\pi^\star$ donnée en (1.6) ? On présente les outils utilisés dans cette thèse pour traiter le problème de l'échantillonnage dans la Section 1.3. On s'intéresse ici au problème (a) en supposant qu'il est possible d'obtenir des échantillons de π_θ pour tout $\theta \in \mathbb{R}^p$, où pour tout $x \in \mathsf{X}$,

$$(\mathrm{d}\pi_\theta/\mathrm{d}\mu)(x) = \exp[-\langle \theta, F(x)\rangle] \Big/ \int_{\mathsf{X}} \exp[-\langle \theta, F(y)\rangle]\mathrm{d}\mu(y) \ .$$

Sous conditions sur μ et F, on montre dans le Chapitre 5 que $\theta^\star$ in (1.6) est le minimiseur de la fonction de log-partition $L : \mathbb{R}^p \to \mathbb{R}$ donnée pour tout $\theta \in \mathbb{R}^p$ par

$$L(\theta) = \log \left(\int_{\mathsf{X}} \exp[-\langle \theta, F(x)\rangle]\mathrm{d}\mu(x) \right) \ .$$

On va considérer une méthode de minimisation du premier ordre, c'est-à-dire que l'on va considérer une suite $(\theta_n)_{n\in\mathbb{N}}$ telle que pour tout $n \in \mathbb{N}$, θ_{n+1} est une fonction de θ_n et $\nabla L(\theta_n)$. Cependant, le gradient de L ne peut pas être calculé explicitement en pratique car il s'exprime sous la forme d'une intégrale vis-à-vis de π_θ : pour tout $\theta \in \mathbb{R}^p$ on a

$$\nabla L(\theta) = -\int_{\mathsf{X}} F(x)\mathrm{d}\pi_\theta(x) = -\mathbb{E}_{\pi_\theta}\left[F\right] \ .$$

Par contre, cette quantité peut être approchée. On considère un estimateur sans biais de $\nabla L(\theta) = -\mathbb{E}_{\pi_\theta}\left[F\right]$ de la manière suivante. Étant donné $(X_k)_{k\in\{1,\dots,M\}}$, $M \in \mathbb{N}$ échantillons indépendants de loi π_θ, $-(1/M)\sum_{k=1}^M F(X_k)$ est un estimateur de $\nabla L(\theta)$. On considère alors l'algorithme de gradient stochastique défini par la récurrence suivante : $\theta_0 \in \mathbb{R}^p$ et pour tout $n \in \mathbb{N}$

$$\theta_{n+1} = \theta_n + (\delta_{n+1}/M_{n+1}) \sum_{k=1}^{M_{n+1}} F(X_k^n) \ , \tag{1.7}$$

où $(\delta_n)_{n\in\mathbb{N}}$ est une suite de pas, $(M_n)_{n\in\mathbb{N}}$ une suite à valeurs dans $\mathbb{N}$ et pour tout $n \in \mathbb{N}$, $(X_k^n)_{k\in\{1,\dots,M_n\}}$ une suite d'échantillons indépendants de π_{θ_n}.

Dans l'algorithme (1.7), on suppose que l'on a accès à des échantillons indépendants de π_θ pour tout $\theta \in \Theta$. Il est possible de considérer des généralisations de (1.7) où la suite d'estimateurs n'est plus donnée par $(-(1/M_{n+1})\sum_{k=1}^{M_{n+1}} F(X_k^n))_{n\in\mathbb{N}}$ mais par $(\nabla L(\theta_n) + e_{\theta_n}(Y_{n+1}))_{n\in\mathbb{N}}$ où $(Y_n)_{n\in\mathbb{N}}$ est un processus stochastique sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $(\mathsf{Y}, \mathcal{Y})$, où $e : \mathbb{R}^p \times \mathsf{Y} \to \mathbb{R}^p$. On peut alors considérer l'algorithme de gradient stochastique général associé à la récurrence suivante : $\theta_0 \in \mathbb{R}^p$ et pour tout $n \in \mathbb{N}$

$$\theta_{n+1} = \theta_n - \delta_{n+1}\left\{\nabla L(\theta_n) + e_{\theta_n}(Y_{n+1})\right\} \ , \tag{1.8}$$

Par exemple, dans (1.7) on peut définir pour tout $n \in \mathbb{N}$,

$$e_{\theta_n}((X_k^n)_{k\in\{1,\dots,M_{n+1}\}}) = -(1/M_{n+1}) \sum_{k=1}^{M_{n+1}} F(X_k^n) - \nabla L(\theta_n) \ . \tag{1.9}$$

L'étude des méthodes d'approximation stochastique remonte aux travaux pionniers de Robbins et Monro [RM51] et de Kiefer et Wolfowitz [KW52]. Ces schémas sont abondamment utilisés dans le domaine de

l'apprentissage automatique et en particulier pour l'entraînement des réseaux de neurones profonds [BLC05]. Il existe de nombreux résultats de convergence pour ces méthodes, voir [MP84; BMP90; DJ93; Ben96; Del96; KY03]. En utilisant un résultat très général de convergence presque-sûre basé sur l'existence de fonctions de Lyapunov [DLM99, Theorem 2] et le théorème de Sard [Sar42] on obtient que, presque-sûrement, $\theta^\star = \lim_{n\to+\infty} \theta_n$ existe et $\theta^\star \in \{\theta \,:\, \nabla L(\theta) = 0\}$. Dans notre cas, puisque L est convexe, on peut donner des résultats de convergence quantitatifs non-asymptotiques, voir [SZ13; BM11; AFM17] par exemple. En particulier sous certaines conditions, [AFM17] montre que pour tout $n \in \mathbb{N}$

$$\mathbb{E}[L(\hat{\theta}_n)] - \min_{\theta\in\mathbb{R}^p} L(\theta) \leq \mathbb{E}\left[(1/2)\left\|\theta_0 - \theta^\star\right\|^2 - \sum_{k=0}^{n-1} \delta_{k+1}\langle \theta_k - \delta_{k+1}\nabla L(\theta_k) - \theta^\star, e_{\theta_k}(Y_{k+1})\rangle + \sum_{k=1}^{n} \delta_k^2 \left\|e_{\theta_k}(Y_{k+1})\right\|^2\right] \Big/ \sum_{k=1}^{n} \delta_k \ ,$$

où $\hat{\theta}_n = \sum_{k=1}^n \delta_k\theta_k / \sum_{k=1}^n \delta_k$. Dans notre cas, il est en général impossible de pouvoir échantillonner de manière i.i.d. selon π_θ. Néanmoins, il existe de nombreux algorithmes permettant de construire des chaînes de Markov ciblant π_θ. Plus précisément, nous considérons des schémas d'approximation de la forme (1.8) et (1.9), où pour tout $n \in \mathbb{N}$, $\{X_k^n, k \in \{1,\dots,M_{n+1}\}\}$ est une chaîne de Markov qui converge vers une version, potentiellement biaisée, de π_{θ_n}. En utilisant les propriétés d'ergodicité de la chaîne de Markov sous-jacente, on peut obtenir des résultats de convergence quantitatifs concernant la convergence de l'algorithme de gradient stochastique, voir [AFM17; DB+19].

1.3 Échantillonnage par MCMC, dynamique de Langevin et convergence de chaîne de Markov

On suppose désormais que $\mathsf{X} = \mathbb{R}^d$ et qu'on dispose d'une distribution de probabilités π sur l'espace des images $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ (obtenue par exemple, via le principe de maximum d'entropie) et que celle-ci admet une densité par rapport à la mesure de Lebesgue donnée pour tout $x \in \mathbb{R}^d$ par

$$(\mathrm{d}\pi/\mathrm{d}\lambda)(x) = \exp[-U(x)] \Big/ \int_{\mathbb{R}^d} \exp[-U(y)]\mathrm{d}y \ , \tag{1.10}$$

où $U : \mathbb{R}^d \to \mathbb{R}$ est une fonction mesurable telle que $\int_{\mathbb{R}^d} \exp[-U(y)]\mathrm{d}y < +\infty$. Par analogie avec la physique statistique, on appelle U le *potentiel*. On cherche à obtenir des échantillons de π. Dans le cadre de la synthèse de texture, ces nouveaux échantillons constitueront de nouvelles images présentant les caractéristiques perceptuelles imposées par π.

1.3.1 Échantillonnage

On commence par décrire quelques méthodes classiques pour l'échantillonnage d'une loi de probabilité ainsi que leurs limitations. Si $d = 1$ alors il est possible d'utiliser la méthode d'inversion à partir de la connaissance de la fonction de répartition. Malheureusement cette méthode est spécifique au cas $d = 1$ et nécessite de connaître la fonction de répartition qui est souvent inconnue. La méthode de rejet permet d'échantillonner des variables aléatoires en dimension $d \in \mathbb{N}$ avec $d > 1$ sans la connaissance de la fonction de répartition ni de la constante de normalisation. Par contre, cette méthode s'avère souvent inefficace puisque, à moins que la densité de proposition soit bien choisie, un grand nombre d'échantillons est rejeté.

Les méthodes envisagées jusqu'ici produisent des échantillons selon la loi π mais ne sont absolument pas adaptées à l'échantillonnage en grande dimension. Plutôt que d'imposer que tous les échantillons suivent la même loi que π, on va construire une suite d'échantillons $(X_n)_{n\in\mathbb{N}}$ dont la loi tend vers π dans un sens que l'on précisera rigoureusement plus tard. Le premier algorithme introduit pour répondre à ce problème est celui de Metropolis-Hastings [Met+53; Has70; Pes73] dont on rappelle ici les principales étapes. Soit R le noyau markovien donné pour tout $x \in \mathbb{R}^d$ et $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ par

$$\mathrm{R}(x, \mathsf{A}) = \delta_x(\mathsf{A}) \int_{\mathbb{R}^d} (1 - \alpha(x, y)) q(x, y) \mathrm{d}y + \int_{\mathsf{A}} \alpha(x, y) q(x, y) \mathrm{d}y \ ,$$

où q est une densité de transition appelée *densité de proposition*, $\alpha : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ est le *taux d'acceptation* et δ_x est la masse de Dirac en x. Soit $X_0 \in \mathbb{R}^d$, on définit pour tout $n \in \mathbb{N}$

$$X_{n+1} = (1 - W_{n+1}) X_n + W_{n+1} Y_{n+1} \ ,$$

avec Y_{n+1} distribuée selon $q(X_n, \cdot)$ conditionnellement à X_n, et W_{n+1} une variable de Bernoulli de paramètre $\alpha(X_n, Y_{n+1})$ conditionnellement à (X_n, Y_{n+1}). On obtient alors que pour tout $n \in \mathbb{N}$, X_n a pour distribution $\mathrm{R}^n(X_0, \cdot)$. Si les deux mesures μ_1, μ_2 données pour tout $\mathsf{A} \in \mathcal{B}(\mathbb{R}^{2d})$ par

$$\mu_1(\mathsf{A}) = \int_{\mathbb{R}^{2d}} \mathbb{1}_{\mathsf{A}}(x, y) \mathrm{R}(x, \mathrm{d}y) \mathrm{d}\pi(x) \ , \qquad \mu_2(\mathsf{A}) = \int_{\mathbb{R}^{2d}} \mathbb{1}_{\mathsf{A}}(y, x) \mathrm{R}(x, \mathrm{d}y) \mathrm{d}\pi(x) \ , \tag{1.11}$$

sont égales alors R est réversible par rapport à π, *i.e.* R est auto-adjoint dans $\mathrm{L}^2(\pi)$, et, en particulier, admet π pour loi invariante, $\pi\mathrm{R} = \pi$. Dans le cas où π admet une densité par rapport à la mesure de Lebesgue et si on pose pour tout $x, y \in \mathbb{R}^d$

$$\alpha(x, y) = \min\left(1, \frac{(\mathrm{d}\pi/\mathrm{d}\lambda)(y) q(y, x)}{(\mathrm{d}\pi/\mathrm{d}\lambda)(x) q(x, y)}\right) \ ,$$

alors (1.11) est vérifiée. Notons que cette méthode ne nécessite pas la connaissance de la constante de normalisation. Il reste à choisir la densité de proposition q. Si on fait le choix d'une densité symétrique $q(x, y) = (2\pi\sigma^2)^{-1/2} \exp[-\|x - y\|^2/(2\sigma^2)]$ avec $\sigma > 0$ par exemple, alors le taux d'acceptation se simplifie et on a $\alpha(x, y) = (\mathrm{d}\pi/\mathrm{d}\lambda)(y)/(\mathrm{d}\pi/\mathrm{d}\lambda)(x)$. D'autres choix sont possibles et on peut considérer la densité suivante pour tout $x, y \in \mathbb{R}^d$

$$q(x, y) = (4\pi\gamma)^{-1/2} \exp\left[\|y - x - \gamma \nabla U(x)\|^2 / (2\gamma)\right] \ , \tag{1.12}$$

où $\gamma > 0$ et on rappelle que $-U$ est la log-densité de π (1.10), supposée différentiable sur $\mathbb{R}^d$. Cette loi de proposition correspond à la mise à jour intermédiaire suivante pour tout $n \in \mathbb{N}$

$$Y_{n+1} = X_n - \gamma \nabla U(X_n) + \sqrt{2\gamma} Z_{n+1} \ , \tag{1.13}$$

où $(Z_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires gaussiennes indépendantes centrées de matrice de covariance identité. Le nouvel état de la chaîne est alors donné par Y_{n+1} avec probabilité $\alpha(X_n, Y_{n+1})$ et par X_n avec probabilité $1 - \alpha(X_n, Y_{n+1})$. La mise à jour (1.13) consiste en une étape de descente de gradient avec un pas γ, *i.e.* $\mathcal{T}_1(x) = x - \gamma \nabla U(x)$, et un ajout de bruit gaussien de variance 2γ, *i.e.* $\mathcal{T}_{2,n}(x) = x + \sqrt{2\gamma} Z_{n+1}$. On a alors $Y_{n+1} = \mathcal{T}_{2,n}(\mathcal{T}_1(X_n))$. Ainsi, (1.13) correspond à une étape de descente de gradient perturbé où la perturbation est donnée par l'opérateur $\mathcal{T}_{2,n}$. On verra en Section 1.3.2 que (1.13) correspond à la discrétisation d'une dynamique continue de loi invariante π.

L'algorithme associé à la loi de proposition (1.12) a été introduit par Besag dans un commentaire de [GM94] et est désormais connu sous le nom d'algorithme de Metropolis Langevin ajusté (MALA).

Le taux d'acceptation dépend alors des positions $x, y \in \mathbb{R}^d$ mais aussi du paramètre $\gamma > 0$, le pas de discrétisation.

Si le taux d'acceptation est trop bas (le pas est trop grand) alors la chaîne de Markov n'explore pas la distribution puisque de nombreuses itérations n'entrainent aucun mouvement car elles sont rejetées. Si le taux d'acceptation est trop élevé (le pas est trop petit) alors la chaîne de Markov explore inefficacement la distribution. Dans la suite on va considérer des chaînes de Markov non-ajustées, c'est-à-dire sans étape d'acceptation-rejet. Ces chaînes de Markov non ajustées ne convergent pas vers π en général, mais vers une autre mesure de probabilité dont la distance à π peut être contrôlée.

1.3.2 Dynamique de Langevin discrète et continue

On considère l'algorithme de Langevin non ajusté ULA [RT96], consistant à accepter la proposition (1.13) quel que soit le taux d'acceptation. On obtient ainsi la chaîne de Markov suivante : $X_0 \in \mathbb{R}^d$ et pour tout $n \in \mathbb{N}$

$$X_{n+1} = X_n - \gamma \nabla U(X_n) + \sqrt{2\gamma} Z_{n+1} \ , \tag{1.14}$$

où $(Z_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires gaussiennes d-dimensionelles, indépendantes et identiquement distribuées de moyenne 0 et de matrice de covariance identité. Pour tout $\gamma > 0$, on note R_γ le noyau markovien défini pour tout $x \in \mathbb{R}^d$ et $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ par

$$\mathrm{R}_\gamma(x, \mathsf{A}) = (4\pi\gamma)^{-1/2} \int_{\mathsf{A}} \exp[-\|x - \gamma \nabla U(x) - y\|^2/(2\gamma)] \mathrm{d}y \ .$$

Ces dernières années, l'algorithme ULA a suscité un grand intérêt de la part des communautés de statistiques computationnelles et d'apprentissage. En effet, puisque la relation de récurrence (1.14) ne requiert que la connaissance de ∇U, ULA peut être implémenté de manière efficace dans le cas où le potentiel est donné par un réseau de neurones profond en utilisant l'auto-différenciation. Les liens que cet algorithme entretient avec des techniques classiques d'optimisation comme la descente de gradient stochastique, qui a prouvé sa grande efficacité dans un contexte d'apprentissage [BLC05; Nem+09], en font un algorithme de choix pour l'échantillonnage en grande dimension. En vue de ces applications, il est nécessaire d'obtenir des résultats de convergence quantitatifs et non-asymptotiques afin d'évaluer qualitativement la performance de ULA, et de ses variantes, sur ces problèmes d'apprentissage en haute dimension.

Les premières analyses de convergence de cet algorithme vers une loi stationaire s'appuie sur une comparaison avec la version continue de cet algorithme : la dynamique de Langevin. Celle-ci est associée à l'équation différentielle stochastique (EDS) suivante

$$\mathrm{d}\mathbf{X}_t = -\nabla U(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \ , \tag{1.15}$$

où $(\mathbf{B}_t)_{t\geq 0}$ est un mouvement Brownien d-dimensionel. Les origines de cette dynamique remontent à Langevin [Lan08] et à ses travaux sur le mouvement des particules.

Si $U \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R})$ et ∇U est Lipschitz alors pour toute condition initiale $\mathbf{X}_0$ il existe une unique solution forte globale $(\mathbf{X}_t)_{t\geq 0}$ [IW89, Chapitre 4, Théorème 2.3, Théorème 2.4]. On suppose désormais qu'il existe une unique solution globale de (1.15). On introduit $(\mathrm{P}_t)_{t\geq 0}$ la famille de noyaux markoviens définie de la manière suivante : pour tout $x \in \mathbb{R}^d$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ et $t \geq 0$, $\mathrm{P}_t(x, \mathsf{A}) = \mathbb{P}(\mathbf{X}_t \in \mathsf{A})$, où $\mathbf{X}_0 = x$. Il est alors possible de montrer que π donnée par (1.10) est invariante par $(\mathrm{P}_t)_{t\geq 0}$, *i.e.* pour tout $t \geq 0$, $\pi\mathrm{P}_t = \pi$, voir [Dur16]. On s'intéresse désormais à la convergence de $(\mathrm{P}_t)_{t\geq 0}$ pour différentes distances sur les espaces de mesures, parmi elles la variation totale et les distances de Wasserstein.

La variation totale d'une mesure signée finie μ, notée $\|\mu\|_{\mathrm{TV}}$, est donnée par

$$\|\mu\|_{\mathrm{TV}} = (1/2) \sup_{f \in \mathrm{L}^\infty(\mathbb{R}^d, \mathbb{R}), \|f\|_\infty \leq 1} \left| \int_{\mathbb{R}^d} f(x) \mathrm{d}\mu(x) \right| \ .$$

Soit deux mesures de probabilités μ et ν sur $\mathcal{B}(\mathbb{R}^d)$. Une mesure de probabilité ζ sur $\mathcal{B}(\mathbb{R}^{2d})$ est un plan de transférence entre μ et ν si pour tout $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\zeta(\mathsf{A} \times \mathbb{R}^d) = \mu(\mathsf{A})$ et $\zeta(\mathbb{R}^d \times \mathsf{A}) = \nu(\mathsf{A})$. On note alors $\mathbf{T}(\mu, \nu)$ l'ensemble des plans de transférence entre μ et ν. Pour toute fonction de coût, $\mathbf{c} : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$ mesurable on définit alors

$$\mathbf{W}_{\mathbf{c}}(\mu, \nu) = \inf_{\zeta \in \mathbf{T}(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathbf{c}(x, y) \mathrm{d}\zeta(x, y) \ .$$

Si $\mathbf{c}(x, y) = \|x - y\|^p$ pour $p \geq 1$ alors on note $\mathbf{W}_p = \mathbf{W}_{\mathbf{c}}^{1/p}$. $\mathbf{W}_p$ est une métrique sur l'espace des mesures de probabilité qui admettent un moment d'ordre p, voir [Vil09, Definition 6.1], appelée la distance de Wasserstein d'ordre p.

L'ergodicité de $(\mathrm{P}_t)_{t \geq 0}$ pour la variation totale est acquise sous une condition faible de "courbure à l'infini", voir [RT96] par exemple. Plus précisément si $U \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R})$ satisfait

$$\inf_{x \in \mathbb{R}^d} \langle \nabla U(x), x \rangle > -\infty \ ,$$

alors pour tout $x \in \mathbb{R}^d$, $\lim_{t \to +\infty} \|\delta_x \mathrm{P}_t - \pi\|_{\mathrm{TV}} = 0$. Depuis, il a été aussi montré que la convergence de $(\mathrm{P}_t)_{t \geq 0}$ est géométrique sous certaines hypothèses sur U pour de nombreuses distances sur les probabilités (Wasserstein d'ordre p, $p \in \mathbb{N}^\star$, variation totale, V-norme). En Section 1.3.3, on rappelle quelques-uns de ces résultats. Une fois la convergence (géométrique ou sous-géométrique) de $(\mathrm{P}_t)_{t \geq 0}$ établie il est naturel de s'intéresser au problème utile à l'expérimentateur : la convergence de la chaîne discrète (1.14). On illustre la convergence de ULA de manière empirique pour un modèle de mélange de gaussiennes en Figure 1.8. En effet, si [TT90; Mil95] obtiennent des approximations fortes et faibles d'ordre un entre les schémas (1.15) et (1.14), ces résultats ne permettent pas de déduire de la convergence de $(\mathrm{P}_t)_{t \geq 0}$ celle de $(\mathrm{R}_\gamma^n)_{n \in \mathbb{N}}$.

Puisque R_γ satisfait des conditions de régularité, l'existence de la probabilité invariante π_γ est assurée sous des conditions de Foster-Lyapunov [Dou+18, Théorème 12.3.3]. Néanmoins, de nombreuses applications (calcul de la variance d'un estimateur MCMC [Bro+18] ou inégalités de concentration [Ebe16] par exemple) requièrent un contrôle quantitatif de la vitesse de convergence de la chaîne de Markov vers sa mesure invariante. Ce n'est que récemment que la convergence (géométrique et quantitative) de la chaîne associée à (1.14) a été établie [DT12; DM17; DM19; Dal17b; Dal17a] pour différentes distances sur l'espace des probabilités (Wasserstein d'ordre p, $p \in \mathbb{N}^\star$, variation totale, V-norme).

1.3.3 Convergence de discrétisations de diffusions

On considère alors le cadre suivant qui est une généralisation de (1.15) et constitue une famille particulière de modèles fonctionels auto-régressifs.

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t \ , \tag{1.16}$$

où $b \in \mathrm{C}(\mathbb{R}^d, \mathbb{R}^d)$ et $\mathbf{X}_0 \in \mathbb{R}^d$. De la même manière que dans la section précédente, on considère la discrétisation d'Euler-Maruyama de ce processus de diffusion. Soit $\gamma > 0$, pour tout $n \in \mathbb{N}$ on définit

$$X_{n+1} = X_n + \gamma b(X_n) + \sqrt{\gamma} Z_{n+1} \ , \tag{1.17}$$

où $\gamma > 0$ et $X_0 = \mathbf{X}_0$ et $(Z_n)_{n \in \mathbb{N}}$ est une suite de variables aléatoires gaussiennes d-dimensionelles, indépendantes et identiquement distribuées de moyenne 0 et de matrice de covariance identité. De la même manière que précédemment on peut considérer les noyaux markoviens $(\mathrm{P}_t)_{t \geq 0}$ et $(\mathrm{R}_\gamma^n)_{n \in \mathbb{N}}$ associés à (1.16) et (1.17). Dans le cas continu, on dit que $(\mathrm{P}_t)_{t \geq 0}$ est géométriquement ergodique pour

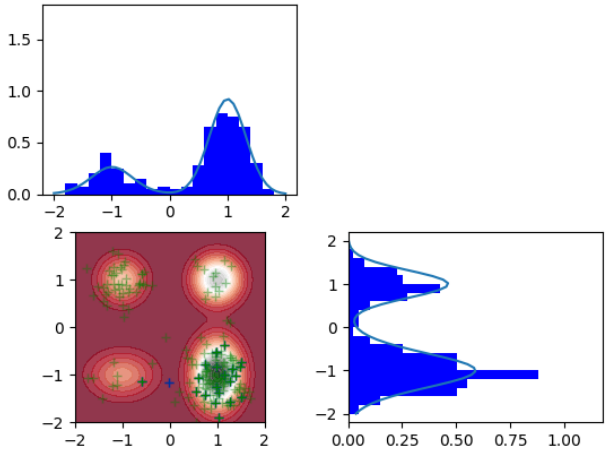


Figure 1.8: **L'algorithme de Langevin non ajusté.** Dans cette figure on présente 10^4 itérations de ULA pour l'échantillonnage d'un mélange de gaussiennes en dimension 2. Les croix vertes correspondent à l'ensemble des itérations de l'algorithme et la croix bleu correspond à la dernière itération. Après 10^4 itérations, on remarque que les marginales correspondant aux projections sur l'axe des abscisses et l'axe des ordonnées sont bien identifiées, même si la convergence n'est pas encore atteinte.

la distance $\mathbf{d}$ (distance sur les probabilités) s'il existe $C \geq 0$ et $\rho \in [0,1)$ tels que pour tout $x, y \in \mathbb{R}^d$ et $t \geq 0$

$$\mathbf{d}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq C\rho^t \mathbf{d}(\delta_x, \delta_y) \ .$$

Dans le cas discret, on dit que $(\mathrm{R}_\gamma^n)_{n\in\mathbb{N}}$ est géométriquement ergodique pour la distance $\mathbf{d}$ (distance sur les probabilités) s'il existe $C \geq 0$ et $\rho \in [0,1)$ tel que pour tout $x, y \in \mathbb{R}^d$ et $n \in \mathbb{N}$

$$\mathbf{d}(\delta_x \mathrm{R}_\gamma^n, \delta_y \mathrm{R}_\gamma^n) \leq C\rho^{\gamma n} \mathbf{d}(\delta_x, \delta_y) \ ,$$

avec ρ indépendant de $\gamma \in (0, \bar{\gamma}]$ pour $\bar{\gamma} > 0$. Notons qu'ici, on ne s'intéresse qu'à la convergence géométrique des chaînes de Markov mais il est également possible d'obtenir des taux de convergence sous-géométriques en considérant des hypothèses plus faibles que celles considérées dans ce manuscrit [But14; DFM16; FM03; VK04]. On reporte à la Section 4.1.1 un historique des résultats d'ergodicité pour les chaînes de Markov. On présente maintenant quelques travaux récents qui conduisent à des convergences géométriques avec des taux précis.

On considère la situation où le champ de vecteurs b est contractant à l'infini, *i.e.* il existe $R \geq 0$ tel que pour tout $x, y \in \mathbb{R}^d$, $\|x - y\| \geq R$, $\langle b(x) - b(y), x - y\rangle \leq -\mathtt{m}^+ \|x - y\|^2$ où $\mathtt{m}^+ > 0$. On suppose également que le champ est Lipschitz et satisfait une condition de Lipschitz unilatérale, *i.e.* il existe $\mathtt{m} \in \mathbb{R}$ tel que pour tout $x, y \in \mathbb{R}^d$, $\langle b(x) - b(y), x - y\rangle \leq -\mathtt{m}\|x - y\|^2$. Dans ce cadre, Eberle et Majka [EM19] montrent qu'on obtient une convergence géométrique pour la distance de Wasserstein associée à la fonction de coût suivante $\mathbf{c}_a$ définie pour tout $x, y \in \mathbb{R}^d$ par

$$\mathbf{c}_a(x, y) = a\mathbb{1}_{\Delta^c}(x, y) + f_a(\|x - y\|) \ ,$$

où $a \geq 0$ et f_a est explicite et donnée dans [EM19, Équation (2.53)] et $\Delta = \{(x, x) \ : \ x \in \mathbb{R}^d\}$. Plus précisément, il existe $\bar{\gamma} > 0$ et $C \geq 0$ tel que pour tout $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ et $n \in \mathbb{N}$ on a

$$\mathbf{W}_{\mathbf{c}_a}(\delta_x \mathrm{R}_\gamma^n, \delta_y \mathrm{R}_\gamma^n) \leq C\rho_a^{\gamma n} \mathbf{W}_{\mathbf{c}_a}(\delta_x, \delta_y) \ ,$$

avec

$$\log(\log^{-1}(\rho_a^{-1})) \simeq -\mathtt{m}R^2/c_1 \ , \text{ où } c_1 = 16^{-1}\int_{1/4}^{3/8}(1-\mathrm{e}^{u-1/2})\boldsymbol{\varphi}(u)\mathrm{d}u \leq 0.00051 \ ,$$

et pour tout $t \in \mathbb{R}$, $\boldsymbol{\varphi}(t) = (2\pi)^{-1/2}\exp[-t^2/2]$. Ces résultats impliquent la convergence pour la distance de Wasserstein $\mathbf{W}_1$ (aussi appelée distance de Kantorovitch-Rubinstein) et pour la variation totale TV.

Majka, Mijatović, et Szpruch ont étendu l'approche précédente à la distance de Wasserstein $\mathbf{W}_2$, voir [MMS18]. Plus précisément, sous les mêmes conditions que précédemment, les auteurs obtiennent le résultat suivant : il existe $\bar{\gamma} > 0$ et $C \geq 0$ tel que pour tout $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ et $n \in \mathbb{N}$

$$\mathbf{W}_2(\delta_x \mathrm{R}_\gamma^n, \delta_y \mathrm{R}_\gamma^n) \leq C\rho_b^{n\gamma/2}(\|x-y\| + \|x-y\|^{1/2}) \ ,$$

avec

$$\log(\log^{-1}(\rho_b^{-1})) \simeq \mathtt{L}R^2/(6c_2) \ ,$$
$$\text{où } c_2 = 4\min\left(\int_0^{1/2} u^2(1-\mathrm{e}^{u-1/2})\boldsymbol{\varphi}(u)\mathrm{d}u, (1-\mathrm{e}^{-1})\int_0^{1/2} u^3\boldsymbol{\varphi}(u)\mathrm{d}u\right) \leq 0.0072 \ ,$$

Les résultats [EM19; MMS18] se basent sur la discrétisation des arguments utilisés par Eberle [Ebe16] dans le but d'obtenir des résultats de convergence similaires pour les processus discrets. En effet, si on pose $\mathbf{c}_c(x,y) = f_c(\|x-y\|)$ pour tout $x, y \in \mathbb{R}^d$, où f_b est défini par [Ebe16, Equation (2.6)], Eberle [Ebe16] obtient que pour tout $x, y \in \mathbb{R}^d$ et $t \geq 0$

$$\mathbf{W}_{\mathbf{c}_c}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq \rho_c^t \mathbf{c}_c(x,y) \ ,$$

avec

$$\log(\log^{-1}(\rho_b^{-1})) \simeq -\mathtt{m}R^2/4 \ .$$

Ce résultat implique une convergence géométrique de $(\delta_x \mathrm{P}_t)_{t\geq 0}$ en distance de Wasserstein $\mathbf{W}_1$ avec un taux ρ_b. Luo et Wang [LW16b] ont par la suite étendu ces convergences pour les processus continus au cas de la distance de Wasserstein $\mathbf{W}_p$ avec $p \geq 1$. On récapitule ces résultats dans le Tableau 1.1. Celui-ci sera complété dans la Section 4.1 où l'on s'intéresse à ce problème et on obtient de nouveaux taux de convergence (valables aussi bien pour le processus discret que pour le processus continu) en utilisant des conditions de minoration et de dérive.

1.4 Organisation et contributions

Cette thèse est divisée en trois parties principales : une approche a contrario de la redondance spatiale, une étude de l'échantillonnage et de l'inférence en haute dimension et une application à la synthèse de texture. Cette division est également thématique : théorie des tests statistiques et des champs aléatoires, théorie des chaînes de Markov et de l'optimisation stochastique et théorie de l'information et applications au traitement d'images.

Dans une première partie, on développe un algorithme basé sur des méthodes a contrario pour la détection de redondance spatiale à partir d'une étude sur la loi de certaines fonctions de similarités dans des champs gaussiens.

Référence	Distance de Wasserstein	Distance de contrôle	(D)	(C)	(TN)
[EM19]	$\|\cdot\|_{\mathrm{TV}}$	$\mathbb{1}_{\Delta^c}(x,y) + \|x-y\|$	✓		7840
	$\mathbf{W}_1$	$\|x-y\|$	✓		4536
[MMS18]	$\mathbf{W}_2$	$\|x-y\| + \|x-y\|^{1/2}$	✓		332
[Ebe16]	$\mathbf{W}_1$	$\|x-y\|$		✓	1
[LW16b]	$\mathbf{W}_p$	$\|x-y\| + \|x-y\|^{1/p}$		✓	$1-\mathtt{m}^+/\mathtt{m}$

Table 1.1: Chaque ligne du tableau se comprend comme suit : pour la "Distance de Wasserstein" $\mathbf{W}_{\mathbf{c}_1}$ et pour la "Distance de contrôle" $\mathbf{c}_2$ si (D) est coché alors il existe $\bar{\gamma} > 0$ et $C \geq 0$ tel que pour tout $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ et $n \in \mathbb{N}$, $\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{R}_\gamma^n, \delta_y \mathrm{R}_\gamma^n) \leq C\rho^{n\gamma}\mathbf{c}_2(x,y)$ avec ρ qui vérifie $-4\log(\log^{-1}(\rho))/(\mathtt{m}R^2) \simeq \beta$ où β est l'entrée de (TN), le taux normalisé. De même, si (C) est coché alors il existe $C \geq 0$ tel que pour tout $x, y \in \mathbb{R}^d$ et $t \geq 0$, $\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{P}_t^n, \delta_y \mathrm{P}_t^n) \leq C\rho^t \mathbf{c}_2(x,y)$. Les distances utilisées sont données par [EM19, Equation (2.53)], [MMS18, Equation (2.11)], [Ebe16, Equation (2.6)] and [LW16b, Equation (2.4)].

Ensuite, on présente des résultats d'ergodicité sur la convergence de discrétisations de diffusions ainsi qu'un algorithme d'optimisation stochastique : *Stochastic Optimization with Unadjusted Langevin* (SOUL) basé sur la discrétisation d'Euler-Maruyama de la dynamique de Langevin.

Finalement, on présente une application de ces résultats dans le domaine de la synthèse de texture par l'exemple. Après avoir obtenu de nouveaux résultats sur le modèle de maximum d'entropie on applique l'algorithme SOUL pour le problème correspondant. Cette étude est appuyée par un grand nombre d'expériences.

1.4.1 Chapitre 3, Section 3.1

Dans la Section 3.1, on introduit une notion de redondance spatiale via des fonctions similarités. Ces fonctions sont définies sur les patchs et permettent donc une évaluation locale de cette similarité. On distingue alors deux cas : l'autosimilarité $\mathscr{A}$ (un patch d'une image est similaire à d'autres patchs de cette même image) et la similarité par modèle (ou *template*) $\mathscr{T}$ (un patch d'une image modèle est similaire à d'autres patchs d'une autre image).

Définition 1.4.1. *Soit x et y deux fonctions définies sur $\mathsf{E} \subset \mathbb{R}^2$ ou $\mathbb{Z}^2$. Soit $\mathsf{w} \subset \mathsf{E}$ un domaine de patch. Quand cette quantité est bien définie on introduit l'autosimilarité avec domaine de patch w et décalage $t \in \mathbb{R}^2$ ou $\mathbb{Z}^2$ tels que $t + \mathsf{w} \subset \mathsf{E}$*

$$\mathscr{A}(x, t, \mathsf{w}) = s\left(P_{\mathsf{w}}(x), P_{\mathsf{w}}(\tau_{-t}(x))\right) ,$$

où s est une fonction de similarité $s : \mathsf{V}^{\mathsf{w}} \times \mathsf{V}^{\mathsf{w}} \to \mathbb{R}$ (par exemple la norme $\|\cdot\|_p$ avec $p \geq 1$). De la même manière on introduit la similarité par modèle

$$\mathscr{T}(x, y, t, \mathsf{w}) = s\left(P_{\mathsf{w}}(y), P_{\mathsf{w}}(\tau_{-t}(x))\right) .$$

Ces deux notions se distinguent si on suppose que l'image modèle est fixée et que les autres images sont des réalisations de champs aléatoires définis sur $\mathbb{Z}^2$ ou $\mathbb{R}^2$. L'autosimilarité et la similarité par modèle sont alors des variables aléatoires à valeurs dans $\mathbb{R}$. Comme annoncé en Section 1.2.2, il est nécessaire, afin d'appliquer la méthode a contrario, d'approcher la distribution de $\mathscr{A}(U, t, \mathsf{w})$ et

$\mathscr{T}(U, x, t, \mathsf{w})$ dans le cas où x est une image et U est donné par un modèle de fond, *i.e.* dans le cas où U est un champ aléatoire gaussien. On commence par donner des résultats asymptotiques lorsque la taille des patchs tend vers l'infini dans le cas où le champ aléatoire est gaussien, stationaire et indépendant à longue distance.

Théorème 1. *Soit $(m_k)_{k\in\mathbb{N}}$ et $(n_k)_{k\in\mathbb{N}}$ deux suites d'entiers croissantes et $(\mathsf{w}_k)_{k\in\mathbb{N}}$ tel que pour tout $k \in \mathbb{N}$, $\mathsf{w}_k = [\![0, m_k]\!] \times [\![0, n_k]\!]$. Soit U un champ aléatoire gaussien tel que $\mathrm{Cov}[U(p_1)U(p_2)] = 0$ pour $\|p_1 - p_2\|_\infty$ assez grand. Pour les fonctions de similarité suivantes :*

- $s(x, y) = \|x - y\|_p$ *avec* $p \in [0, +\infty)$,
- $s(x, y) = \|x - y\|_p^p$ *avec* $p \in [0, +\infty)$,
- $s(x, y) = -\langle x, y\rangle$,
- $s(x, y) = -\langle x, y\rangle / (\|x\| \, \|y\|)$,

il existe μ, σ^2 des fonctions à valeurs réelles définies sur $\mathbb{Z}^2$, et $(\alpha_k)_{k\in\mathbb{N}}$ une suite positive telle que pour tout $t \in \mathbb{Z}^2 \setminus \{0\}$ on a

(a) $\lim_{k\to+\infty} \alpha_k^{-1} \mathscr{A}(U, t, \mathsf{w}_k) \underset{a.s}{=} \mu(t)$;

(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(\alpha_k^{-1} \mathscr{A}(U, t, \mathsf{w}_k) - \mu(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \sigma^2(t)\right)$.

Ce théorème peut être étendu au cas de la similarité par modèle. Notons que l'hypothèse de gaussianité pourrait être relâchée mais est conservée ici afin d'obtenir des espérances et des variances explicites dans notre extension du théorème central limite.

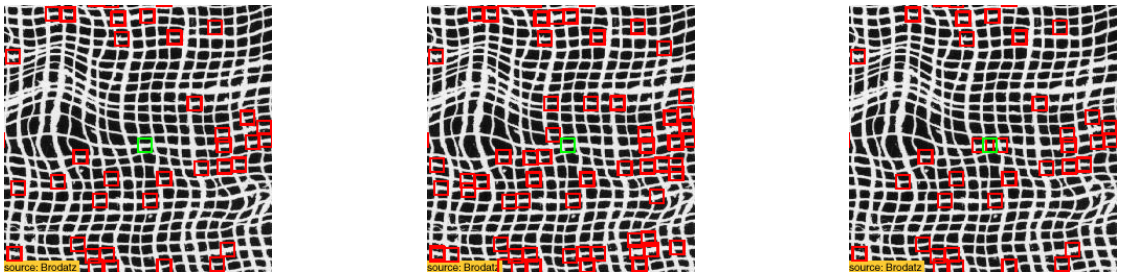
Dans le cas où la fonction de similarité est donnée par la norme ℓ_2 au carré, on peut trouver une expression non-asymptotique de l'autosimilarité comme combinaison linéaire de variables aléatoires indépendantes de loi $\chi_2(1)$ où les coefficients de la combinaison linéaire sont donnés par les valeurs propres d'une matrice symétrique et Toeplitz. On propose ensuite un algorithme qui approche efficacement cette distribution en utilisant une méthode des moments d'ordre 3 ainsi qu'une méthode d'approximation des valeurs propres :

- on approche les valeurs propres de la matrice de Toeplitz symétrique par les valeurs propres de la projection de cette matrice sur l'ensemble des matrices circulantes (les valeurs propres sont alors données par la transformée de Fourier),
- une fois les coefficients approchés, on utilise la méthode WoodF [Woo89] qui est une méthode des moments d'ordre 3. La distribution approximante est donnée par une loi de Fischer-Snedecor.

La précision des approximations proposées est illustrée en Figure 1.9. On propose alors un algorithme de détection de la redondance spatiale des patchs en utilisant la méthode a contrario, *i.e.* en approchant les quantités du type $\mathbb{P}(\mathscr{A}(U, t, \mathsf{w}) \leq \alpha)$. Cette première section est issue de l'article "Redundancy in Gaussian random fields" accepté à ESAIM: Probability and Statistics.

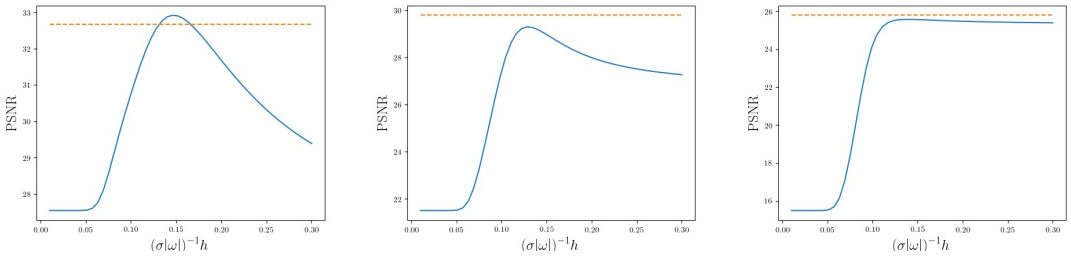
1.4.2 Chapitre 3, Section 3.2

Dans la Section 3.2, après avoir introduit un cadre a contrario pour l'analyse de l'autosimilarité on s'intéresse aux applications de l'algorithme décrit dans la Section 3.1. Pour employer des méthodes



(a) $1513s, n_d = 52$ (b) $200s, n_d = 116$ (c) $4.77s, n_d = 50$

Figure 1.9: **Détection de redondance spatiale.** Dans chaque image, le patch vert correspond au patch dont on veut détecter la redondance. Dans la figure (a), on utilise l'approche a contrario avec modèle de fond gaussien. Aucune approximation (aussi bien sur le calcul des valeurs propres que sur le calcul des probabilités) n'est effectuée. Dans la figure (b) on réduit le temps de calcul en utilisant l'approximation des valeurs propres proposées. Dans la figure (c), on utilise les deux approximations proposées (projection matricielle et méthode des moments).



(a) $\sigma = 10$ (b) $\sigma = 20$ (c) $\sigma = 40$

Figure 1.10: **Influence des paramètres sur le** PSNR**.** Dans cette figure on présente l'évolution du PSNR (ordonnées) pour différentes valeurs du paramètre de filtrage dans l'algorithme original NL-means [DAG10] sur une image pour différents niveaux de bruit σ. La ligne orange correspond au PSNR obtenu sur la même image en considérant notre modification de l'algorithme original NL-means.

a contrario il est nécessaire de calculer les quantités du type $\mathbb{P}(\mathscr{A}(U, t, \mathsf{w}) \leq \alpha)$ où $\mathscr{A}$ est la fonction d'autosimilarité, U est un champ aléatoire (le modèle de fond de la méthode a contrario), t est un vecteur de décalage qui permet de comparer les patchs de l'image U ayant pour indices w et $\mathsf{w} + t$ (où la somme doit être comprise au sens de Minkoswki). Cette quantité peut être facilement calculée en utilisant l'algorithme proposé dans la Section 3.1. Une fois les notions de similarités définies et le cadre statistique établi on peut exploiter la notion de redondance spatiale dans le cadre du traitement d'images. En modifiant la procédure de choix des patchs dans l'algorithme de débruitage par moyennes non locales (*Non-Local Means*, NL-means) on remplace le choix d'un seuil (ou d'une variance) dans le cadre de cet algorithme par le choix d'un nombre de fausses alarmes dans un modèle de fond. Ce choix est robuste vis-à-vis des différentes images utilisées contrairement au choix d'un seuil, voir Figure 1.10. On montre que cette simple modification permet d'obtenir de meilleurs résultats que l'algorithme NL-means classique avec un contrôle sur la reconstruction des patchs. La deuxième application concerne la détection de périodicité. Étant donné une image, on applique notre algorithme de détection de patchs similaires. On obtient alors une image binaire où un pixel à la valeur 1 si le patch centré autour de la

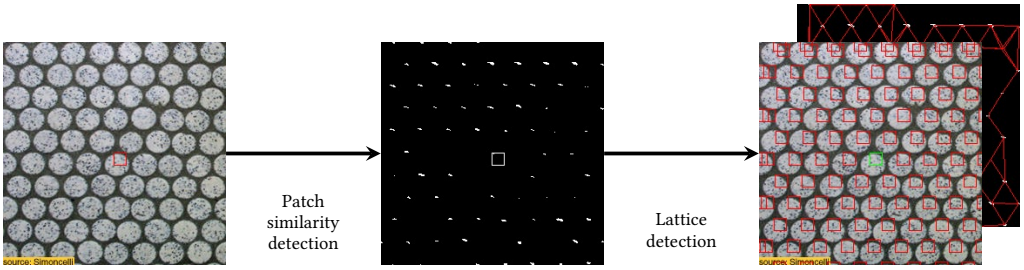


Figure 1.11: **Identification de réseaux.** La première étape correspond à la détection des patchs redondants comme illustré en Figure 1.9. La seconde partie de l'algorithme correspond à l'identification de la base du réseau en utilisant un algorithme de type Maximum A Posteriori, voir Théorème 2. Afin d'illustrer le résultat obtenu on superpose un patch (en rouge) sur chaque point du réseau identifié.

position de ce pixel est similaire au patch modèle et 0 sinon. On considère alors le graphe associé. On note $V \in \mathbb{R}^{p\times 2}$ l'ensemble des arêtes du graphe, avec $p \in \mathbb{N}$, et on considère le modèle aléatoire suivant.

Définition 1.4.2. *Soit V une variable aléatoire à valeurs dans $\mathbb{R}^{p\times 2}$ avec $p \in \mathbb{N}$. On dit que V satisfait l'hypothèse du réseau approché s'il existe une base $B = (b_1, b_2)$ de $\mathbb{R}^2$ et $\sigma > 0$ telle pour tout $\ell \in \{1, \dots, p\}$, il existe $(m_\ell, n_\ell) \in \mathbb{Z}^2$ tels que*

$$\mathcal{L}(V_\ell) = \mathcal{L}(m_\ell b_1 + n_\ell b_2 + \sigma Z_\ell) \ ,$$

où $(Z_\ell)_{\ell\in\{1,\dots,p\}}$ est une famille de variables aléatoires gaussiennes centrées réduites indépendantes. On note $M = (m_\ell, n_\ell)_{\ell\in\{1,\dots,p\}}$.

On peut alors proposer un algorithme de type Maximum a Posteriori pour estimer B et M. On obtient alors deux suites $(B_n)_{n\in\mathbb{N}}$ et $(M_n)_{n\in\mathbb{N}}$ qui vérifient le théorème suivant.

Théorème 2. *Pour tout $\sigma > 0$, $(B_n)_{n\in\mathbb{N}}$ et $(M_n)_{n\in\mathbb{N}}$ convergent en un nombre fini d'itérations.*

En Figure 1.11, on décrit l'algorithme complet d'identification de réseau sur un exemple. Cet algorithme est appliqué avec succès sur des images de cristallographie et permet également de classer des textures selon leur degré de périodicité. Cette présentation est issue de l'article "Patch redundancy in images: a statistical testing framework and some applications" publié à SIAM Imaging Sciences.

1.4.3 Chapitre 4, Section 4.1

Après avoir proposé une définition de la redondance spatiale et un cadre statistique pour l'identifier on s'intéresse à la possibilité d'obtenir des échantillons d'une loi de probabilité qui impose la redondance spatiale. Pour ce faire, il est nécessaire de développer des algorithmes d'échantillonnage dont la complexité n'explose pas avec la dimension même dans un cadre non-convexe. On mène cette étude pour des algorithmes de type Langevin dans la Section 4.1. Le but de cette section est d'étudier des modèles autorégressifs donnés par

$$X_{n+1} = X_n + \gamma b(X_n) + \sqrt{\gamma} Z_{n+1} \ . \tag{1.18}$$

Le principal résultat obtenu est le suivant.

Théorème 3. *Supposons qu'il existe $\mathtt{m} \in \mathbb{R}$, $\mathtt{m}^+ > 0$ et $\mathtt{L}, R \geq 0$ tels que pour tout $x, y \in \mathbb{R}^d$*

$$\|b(x) - b(y)\| \leq \mathtt{L} \|x - y\| \ , \qquad \langle b(x) - b(y), x - y\rangle \leq -\mathtt{m} \|x - y\|^2 \ ,$$

Référence	Distance de Wasserstein	Distance de contrôle	(D)	(C)	(TN)
[EM19]	$\|\cdot\|_{\mathrm{TV}}$	$\mathbb{1}_{\Delta^c}(x,y) + \|x-y\|$	✓		7840
	$\mathbf{W}_1$	$\|x-y\|$	✓		4536
[MMS18]	$\mathbf{W}_2$	$\|x-y\| + \|x-y\|^{1/2}$	✓		332
[Ebe16]	$\mathbf{W}_1$	$\|x-y\|$		✓	1
[LW16b]	$\mathbf{W}_p$	$\|x-y\| + \|x-y\|^{1/p}$		✓	$1-\mathtt{m}^+/\mathtt{m}$
Cette thèse	$\|\cdot\|_{\mathrm{TV}}$	$\mathbb{1}_{\Delta^c}(x,y) + \|x-y\|$	✓	✓	$(1-\mathrm{e}^{2\mathtt{m}R^2})^{-1}$
	$\mathbf{W}_1$	$\|x-y\|$	✓	✓	idem
	$\mathbf{W}_p$	$\|x-y\| + \|x-y\|^{1/\alpha}$	✓	✓	idem

et si $\|x-y\| \geq R$,

$$\langle b(x) - b(y), x-y\rangle \leq -\mathtt{m}^+ \|x-y\|^2 \ .$$

Alors il existe $\bar\gamma > 0$, $D_{\bar\gamma,1}, D_{\bar\gamma,2}, E_{\bar\gamma} \geq 0$ *et* $\lambda_{\bar\gamma}, \rho_{\bar\gamma} \in [0,1)$ *avec* $\lambda_{\bar\gamma} \leq \rho_{\bar\gamma}$, *tels que* $\gamma \in (0,\bar\gamma]$, $x, y \in \mathbb{R}^d$ *et* $k \in \mathbb{N}$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \lambda_{\bar\gamma}^{k\gamma/4}[D_{\bar\gamma,1}\mathbf{c}(x,y) + D_{\bar\gamma,2}\mathbb{1}_{\Delta^c}(x,y)] + E_{\bar\gamma}\rho_{\bar\gamma}^{k\gamma/4}\mathbb{1}_{\Delta^c}(x,y) \ ,$$

où $\mathbf{c}(x,y) = \mathbb{1}_{\Delta^c}(x,y)(1 + \|x-y\|/R)$, $\Delta = \{(x,x) \ : \ x \in \mathbb{R}^d\}$ *et* R_γ *est le noyau de Markov associé à* (1.18).

On trouve également que $\log(\log(\rho^{-1}))$ est de l'ordre de $\mathtt{m}R^2/(4(1-\mathrm{e}^{\mathtt{m}R^2}))$ et ne dépend pas de la dimension d. Comme corollaire du Théorème 3, on obtient l'ergodicité géométrique de ULA mais aussi d'autres algorithmes comme l'algorithme projeté suivant :

$$X_{n+1} = \Pi_{\mathsf{K}}[X_n + \gamma b(X_n) + \sqrt{\gamma} Z_{n+1}] \ , \tag{1.19}$$

où Π_{K} est la projection sur le compact convexe K. En considérant l'algorithme (1.19) et en passant à la limite à la fois en espace (en considérant une suite de compacts $(\mathsf{K}_n)_{n\in\mathbb{N}}$ telle que $\bigcup_{n\in\mathbb{N}} \mathsf{K}_n = \mathbb{R}^d$) et en temps (en considérant une suite de pas $(\gamma_n)_{n\in\mathbb{N}}$ telle que $\lim_{n\to+\infty}\gamma_n = 0$), on étend les taux obtenus dans le cas discret au cas où la dynamique est continue, *i.e.* on obtient que sous les conditions précédentes pour tout $x, y \in \mathbb{R}^d$ et $t \geq 0$

$$\mathbf{d}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq C\rho^t \mathbf{d}(\delta_x, \delta_y) \ ,$$

où $C \geq 0$, $\rho \in [0,1)$ et $\mathbf{d}$ est soit la variation totale TV, soit la distance de Wasserstein d'ordre 1, $\mathbf{W}_1$. Ce résultat étend les taux de convergence obtenus pour la distance de Wasserstein d'ordre 1 à d'autres distances comme la variation totale, [Ebe16]. Il améliore également des résultats de convergence obtenus récemment [EM19]. On complète le Tableau 1.1 par les contributions issues du Théorème 3. Notons que contrairement à la majeure partie des approches actuelles [EM19; EGZ18; EGZ19; Ebe16; MMS18; LW16b; Che+18] qui s'appuient sur l'introduction d'une distance bien choisie, notre preuve utilise uniquement des outils classiques pour l'étude des chaînes de Markov à espace d'états général : les conditions de minoration et les conditions de dérive de Foster-Lyapunov. Ce travail est tiré de "Convergence of diffusion and their discretizations: from continuous to discrete processes and back" soumis aux Annales de l'Institut Henri Poincaré.

1.4.4 Chapitre 4, Section 4.2

L'obtention de contrôles fins concernant la convergence ergodique de l'algorithme ULA et de ses variantes nous permet d'envisager cet algorithme d'échantillonnage comme étape intermédiaire d'un algo-

rithme de minimisation stochastique. En effet, dans la Section 4.2 on considère le problème de minimisation d'une fonction objectif f différentiable et telle que pour tout $\theta \in \mathbb{R}^p$, $\nabla f(\theta) = \int_{\mathbb{R}^d} H_\theta(x) \mathrm{d}\pi_\theta(x)$. Cette étude est motivée par des applications dans le cadre de l'estimation de paramètres de régularisation pour des modèles d'images (débruitage, déflouage, démêlage spectral) et dans le cadre de l'estimation des paramètres des modèles de maximum d'entropie comme on le verra dans les sections suivantes. On propose alors un algorithme de descente de gradient stochastique où à chaque étape, le gradient est estimé via une méthode de Monte-Carlo par chaîne de Markov, voir l'Algorithme 2.

Algorithm 1 Stochastic Optimization via Unadjusted Langevin (SOUL)

Inputs:
$\theta_0 \in \mathsf{K}$, $X_0^0 \in \mathbb{R}^d$, $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$, $(m_n)_{n\in\mathbb{N}}$, N
for $n \in \{1, \dots, N-1\}$ **do**
  if $n \geq 1$ **then**
    $X_0^n \leftarrow X_{m_{n-1}}^{n-1}$
  end if
  for $k \in \{0, \dots, m_n - 1\}$ **do**
    $Z_{k+1}^n \leftarrow$ sample $\mathrm{N}(0, \mathrm{Id})$
    $X_{k+1}^n \leftarrow X_k^n - \gamma_n \nabla_x \log \pi_\theta(X_k^n) + \sqrt{2\gamma_n} Z_{k+1}^n$
  end for
  $\Delta_{\theta_n} \leftarrow \frac{1}{m_n} \sum_{k=1}^{m_n} H_{\theta_n}(X_k^n)$
  $\theta_{n+1} \leftarrow \Pi_{\mathsf{K}}[\theta_n + \delta_{n+1} \Delta_{\theta_n}]$
end for
Outputs:
$\hat{\theta}_N = \left\{\sum_{n=1}^N \delta_n \theta_n\right\} \Big/ \left\{\sum_{n=1}^N \delta_n\right\}$

Si la chaîne de Markov utilisée à chaque itération est un algorithme de type ULA alors on nomme cet algorithme SOUL pour *Stochastic Optimization with Unadjusted Langevin*. La procédure peut en fait être étendue pour n'importe quelle chaîne de Markov qui converge géométriquement vers sa mesure invariante et dans ce cas là, on nomme l'algorithme SOUK pour *Stochastic Optimization with Unadjusted Kernel*. En s'appuyant sur des résultats récents d'optimisation stochastique [AFM17] on obtient des contrôles explicites (à la fois presque-sûrement et en espérance) sur la quantité $(f(\hat{\theta}_N))_{N\in\mathbb{N}}$, dans le cas où f est convexe, où pour tout $N \in \mathbb{N}$

$$\hat{\theta}_N = \sum_{k=1}^N \delta_k \theta_k \Big/ \sum_{k=1}^N \delta_k \ ,$$

et $(\delta_k)_{k\in\mathbb{N}}$ est la suite de pas utilisée dans l'algorithme de descente de gradient stochastique.

Théorème 4. *Supposons que $f : \mathbb{R}^p \to \mathbb{R}$ est différentiable, convexe et s'il existe $L \geq 0$ tel que pour tout $\theta_1, \theta_2 \in \mathbb{R}^p$,*

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq L \|\theta_1 - \theta_2\| \ .$$

Soit $(\delta_n)_{n\in\mathbb{N}}$ et $(\gamma_n)_{n\in\mathbb{N}}$ deux suites de réels positifs décroissantes et $(m_n)_{n\in\mathbb{N}}$ une suite d'entiers telles que

$$\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty \ , \qquad \sum_{n=0}^{+\infty} \delta_{n+1} \gamma_n^{1/2} < +\infty \ , \qquad \sum_{n=0}^{+\infty} \delta_{n+1} / (m_n \gamma_n) < +\infty \ .$$

Soit $\{(X_k^n)_{k\in\{0,\dots,m_n\}} : n \in \mathbb{N}\}$ et $(\theta_n)_{n\in\mathbb{N}}$ donnés par l'Algorithme 2. Supposons de plus que les conditions de Théorème 3 sont satisfaites pour $b = \nabla \log \pi_\theta$ uniformément en $\theta \in \mathsf{K}$. Alors pour $\sup_{n\in\mathbb{N}}(\delta_n + \gamma_n)$ assez petit on a :

(a) $(\theta_n)_{n\in\mathbb{N}}$ *converge presque-sûrement vers* $\theta^\star \in \arg\min_{\mathsf{K}} f$ *;*

(b) de plus, presque-sûrement il existe $C \geq 0$ *tel que* $n \in \mathbb{N}^\star$

$$\left\{ \sum_{k=1}^{n} \delta_k f(\theta_k) \Big/ \sum_{k=1}^{n} \delta_k \right\} - \min_{\mathsf{K}} f \leq C \Big/ \left(\sum_{k=1}^{n} \delta_k \right) \ .$$

Ce théorème permet de donner une caractérisation asymptotique de $(\hat{\theta}_N)_{N\in\mathbb{N}}$ de l'Algorithme 2. On peut également obtenir des résultats non-asymptotiques pour la quantité $\mathbb{E}[\sum_{k=1}^{n} \delta_k f(\theta_k) / \sum_{k=1}^{n} \delta_k]$. Les résultats obtenus étendent le cadre d'optimisation stochastique introduit dans [AFM17]. En particulier, on ne suppose pas que le noyau de Markov est Lipschitz vis-à-vis de ses paramètres.

Cette section est issue de l'article "Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. Application to maximum marginal likelihood and empirical Bayesian estimation" soumis à Statistics and Computing. Notons également qu'une extension de cet algorithme au cas où f n'est plus nécessairement différentiable et des applications dans le cadre de l'estimation de paramètres de régularisation avec une approche empirique bayésienne sont présentées dans "Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach" soumis à SIAM Journal on Imaging Sciences.

1.4.5 Chapitre 5, Section 5.1

Dans la Section 5.1 on présente une application du principe de maximum d'entropie pour la synthèse de texture. On étend des résultats sur l'existence et l'unicité de modèles de maximum d'entropie dans le cas où les probabilités étudiées ne sont plus discrètes. On donne notamment des conditions explicites sur les fonctions de contrainte et la mesure de probabilité de référence qui déterminent le modèle pour que la distribution de maximum d'entropie existe et soit donnée par $(\mathrm{d}\pi^\star/\mathrm{d}\mu)(x) \propto \exp[-\langle \theta^\star, F(x)\rangle]$ où μ est une mesure de probabilité de référence, $\theta^\star \in \mathbb{R}^p$ et $F : \mathbb{R}^d \to \mathbb{R}^p$ est un ensemble de contraintes.

Proposition 1.4.3. *Supposons qu'il existe* $\alpha \geq 0$, $\alpha' > \alpha$ *tels que*

- F *est continue et il existe* $C_\alpha \geq 0$ *tel que* $\sup_{x\in\mathbb{R}^d} \{\|F(x)\|(1+\|x\|^\alpha)^{-1}\} \leq C_\alpha < +\infty$,
- *il existe* $\eta > 0$ *tel que* $\int_{\mathbb{R}^d} \exp[\eta\|x\|^{\alpha'}]\mathrm{d}\mu(x) < +\infty$.

Alors on a les résultats suivants :

(a) Si pour tout $\theta \in \mathbb{R}^p$ *avec* $\|\theta\| = 1$ *on a* $\mu(\{x \in \mathbb{R}^d : \langle F(x), \theta\rangle < 0\}) > 0$*, alors la solution du problème de maximum d'entropie (1.5) est donnée par* $\pi^\star$ *dont la densité par rapport à* μ *est donnée pour tout* $x \in \mathbb{R}^d$ *par*

$$(\mathrm{d}\pi^\star/\mathrm{d}\mu)(x) = \exp\left[-\langle \theta^\star, F(x)\rangle\right] \Big/ \int_{\mathbb{R}^d} \exp\left[-\langle \theta^\star, F(y)\rangle\right] \mathrm{d}\mu(y) \ ,$$

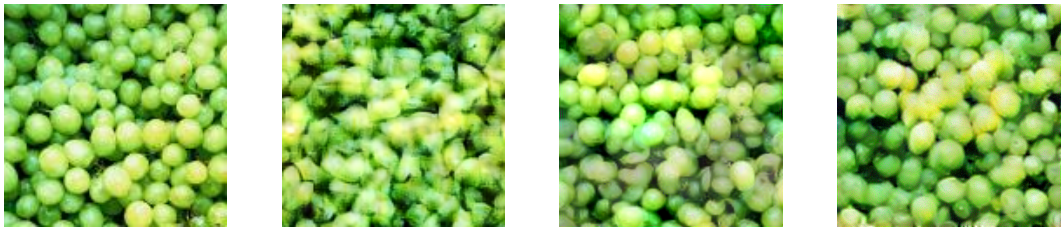
où $\theta^\star$ *est la solution du problème d'optimisation convexe suivant*

$$\min_{\theta\in\mathbb{R}^p} \left\{ \int_{\mathbb{R}^d} \exp[-\langle \theta, F(x)\rangle] \mathrm{d}\mu(x) \right\} \ .$$

(b) En particulier la Proposition 1.4.3-(a) est satisfaite si $\mu(\mathsf{A}) > 0$ *pour tout* $\mathsf{A} \subset \mathbb{R}^d$ *ouvert, et si* F *est continue et qu'il existe* $x \in F^{-1}(\{0\})$ *tel que* F *est différentiable en* x *et* $\det(\mathrm{D}F(x)\mathrm{D}F(x)^\top) > 0$.

(a) (b) (c) (d)

Figure 1.12: **Quelques exemples de synthèse.** Les images originales sont données en (b) et (d) et la synthèse en (a) et (c).



(a) (b) (c) (d)

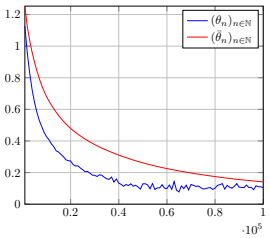
Figure 1.13: **Quelques algorithmes de synthèse.** En (a) on présente l'image de texture exemple. En (b) on donne la synthèse obtenue avec [PS00] où les contraintes spatiales sont données par des ondelettes. En (c), on applique la méthode décrite dans [GEB15] pour générer la texture. Enfin en (d), on présente les résultats obtenus avec notre algorithme

Dans ce cas, on montre que le problème d'optimisation convexe infini-dimensionnel sur les mesures se traduit en un problème d'optimisation convexe fini-dimensionel en utilisant une généralisation de la dualité lagrangienne. On trouve alors que $\theta^\star$ est minimum de la fonction de log-partition $L : \mathbb{R}^p \to \mathbb{R}$ donnée pour tout $\theta \in \mathbb{R}^p$ par $L(\theta) = \int_{\mathbb{R}^d} \exp[-\langle \theta, F(x)\rangle] \mathrm{d}\mu(x)$. La condition suffisante pour l'existence d'un modèle de maximum d'entropie est alors plus facile à vérifier que celle donnée par [Csi75]. Dans le cas où F est donnée par les sorties d'un réseau de neurones, on donne un certificat qui permet de vérifier l'existence du maximum d'entropie.

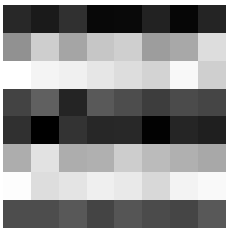
1.4.6 Chapitre 5, Section 5.2

La log-partition vérifie alors les hypothèses nécessaires pour appliquer l'algorithme SOUL. Notons que la formulation du problème de synthèse de texture via le principe de maximum d'entropie date des travaux de Zhu, Wu et Mumford [ZWM98]. L'utilisation de la dynamique de Langevin pour résoudre ce problème a été suggérée pour la première fois par [LZW16]. Notre étude théorique de l'existence, de l'unicité et de la caractérisation du maximum d'entropie nous permet d'affirmer que la mesure recherchée s'écrit sous forme exponentielle. On est alors dans les hypothèses de l'algorithme SOUL et on peut démontrer la convergence de notre algorithme dans la cadre de la synthèse de texture.

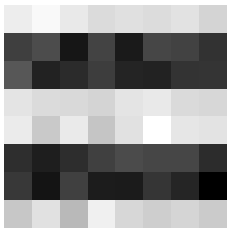
On établit des liens entre notre approche et l'approche microcanonique proposée récemment [BM18]. En particulier on montre que le modèle microcanonique peut s'envisager comme la limite de modèles



(a)



(b) $n = 100000$



(c)

Figure 1.14: **Convergence des paramètres.** En (a) on affiche l'erreur normalisée entre $(\theta_n)_{n\in\mathbb{N}}$ (bleu) et les poids optimaux ainsi qu'entre $(\hat{\theta}_n)_{n\in\mathbb{N}}$ (rouge) et les poids optimaux. En (b) on présente un échantillon du modèle pour $n = 100000$ et en (c) on présente l'image de texture exemple.

macrocanoniques successifs, voir Proposition 1.4.4.

Proposition 1.4.4. *Supposons qu'il existe $\eta > 0$ tel que $\int_{\mathbb{R}^d} \exp[\eta\|x\|^2]\mathrm{d}\mu(x) < +\infty$ et que F soit donné par un réseau de neurones qui vérifie une condition de coercivité, voir Proposition 5.2.3 pour une formulation précise. Alors, il existe $\varepsilon_0 > 0$ tel que pour tout $\varepsilon \in (0, \varepsilon_0)$, le maximum d'entropie π_ε existe. Si de plus $\mu(F^{-1}(\{0\})) > 0$, alors $\lim_{\varepsilon\to 0} \pi_\varepsilon = \pi_\infty$, avec pour tout $x \in \mathbb{R}^d$, $(\mathrm{d}\pi_\infty/\mathrm{d}\mu)(x) = \mathbb{1}_{F^{-1}(\{0\})}(x)/\mu(F^{-1}(\{0\}))$.*

Enfin, on vérifie la validité de l'algorithme SOUL dans deux cadres :

- synthèse de textures gaussiennes,
- synthèse de textures structurées.

Dans le cadre de la synthèse de textures gaussiennes, on cherche le modèle gaussien qui présente la même autocorrelation qu'une texture exemple. On peut alors caractériser facilement la validité du modèle puisque les poids optimaux peuvent être calculés explicitement, voir Figure 1.14.

Dans le cadre de la synthèse de textures structurées, on obtient des résultats comparables à l'état de l'art, voir Figure 1.12 et Figure 1.13 qui complète Figure 1.7. On étudie expérimentalement la capacité d'innovation de notre algorithme et on propose des extensions de celui-ci, notamment pour le transfert de style. Les travaux des deux sections précédentes sont issus de "Maximum entropy methods for texture synthesis: theory and practice" soumis à SIMODS et de "Macrocanonical models for texture synthesis" présenté à Scale Space and Variational Methods in Computer Vision.

1.4.7 Publications présentées dans cette thèse

Publications dans des revues à comités de lecture internationales

1) V. De Bortoli, A. Desolneux, B. Galerne, A. Leclaire. *Patch redundancy in images: a statistical testing framework and some applications* 2019 - Publié à SIAM Journal on Imaging Sciences (SIIMS). Voir Chapitre 3, Section 3.2.

2) V. De Bortoli, A. Desolneux, B. Galerne, A. Leclaire. *Redundancy in Gaussian random fields* 2018 - Accepté à ESAIM Probability and Statistics. Voir Chapitre 3, Section 3.1.

Pré-publications et travaux soumis

1) V. De Bortoli, A. Durmus. *Convergence of diffusion and their discretizations: from continuous to discrete processes and back* 2019 - Soumis. Voir Chapitre 4, Section 4.1.

2) V. De Bortoli, A. Durmus, M. Pereyra, A. F.Vidal. *Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. Application to maximum marginal likelihood and empirical Bayesian estimation* 2019 - Soumis. Voir Chapitre 4, Section 4.2.

3) V. De Bortoli, A. Desolneux, A. Durmus, B. Galerne, A. Leclaire. *Maximum entropy methods for texture synthesis: theory and practice* 2019 - Soumis. Voir Chapitre 5;

Publications dans des conférences à comité de lectures internationales

1) V. De Bortoli, A. Desolneux, B. Galerne, A. Leclaire. *Macrocanonical models for texture synthesis* 2019 - Présenté à Scale Space and Variational Methods in Computer Vision (SSVM). Version courte de “Maximum entropy methods for texture synthesis: theory and practice”

Chapter 2

Introduction (English)

In this thesis, we are interested in the analysis of non-local statistics in images (modelling, estimation, sampling). We mainly focus on two non-local statistics: spatial redundancy and deep neural network outputs. More precisely, we will tackle the following questions:

- What is spatial redundancy? How to detect such redundancy in natural images?
- How to design stochastic models of natural images with neural network-based constraints? How to sample from such models?

In Section 2.1, we recall the importance of the spatial redundancy in the setting of the *gestalt* theory. We also introduce the a contrario method which constitutes a statistical framework for the gestalt theory. We then draw links between the a contrario approach and the statistical hypothesis testing theory (see also [DMM08, Chapter 15.3.2] for a discussion on the links between the a contrario approach and multiple testing theory). In order to use such methods, we need to specify some a contrario (or noise) models. In our case, these models will be given by Gaussian random fields. These random fields define a particular class of textures.

The general problem of exemplar-based texture synthesis is recalled in Section 2.2. This is the image processing approach to a much more general problem: the synthesis of random fields. We present a texture synthesis algorithm based on wavelet features. This algorithm can be cast as an optimization scheme in order to maximize the entropy of the model under some geometrical constraints. The constraints of the model are enforced almost surely. If we consider the relaxed entropy maximization problem where the constraints are enforced in expectation and not almost surely, then we can explicitly compute the optimal measure under mild conditions on the constraints. These constraints will be satisfied in the case where the constraints are given by a differentiable neural network.

Assuming that the distribution of the images which satisfy some non-local statistics is known, via the maximum entropy principle for instance, we turn to the problem of sampling from such probability distributions in Section 2.3. We describe the unadjusted Langevin algorithm (ULA). This Markov chain allows for the efficient sampling of general target probability distributions even when the dimension of the state space is large. We present some recent quantitative results obtained for a class of functional autoregressive models. Using these results we give quantitative bounds on the convergence of ULA for various distances.

Our main contributions are detailed in Section 2.4. We describe an algorithm based on the a contrario methodology to detect spatial redundancy in Section 2.4.1. We apply this algorithm to image processing

tasks (denoising and periodicity detection) in Section 2.4.2. In Section 2.4.3 we present ergodicity results for a class of functional autoregressive Markov chains. These quantitative results are used to obtain the convergence of a stochastic approximation algorithm introduced in Section 2.4.4: the Stochastic Optimization with Unadjusted Langevin (SOUL) algorithm. In Section 2.4.5, we show that some maximum entropy based texture models can be obtained using the SOUL algorithm. Finally we present some visual applications of our previous results in Section 2.4.6. Each chapter contains a literature review of the subject under study.

2.1 Spatial redundancy, a contrario methods and random fields

2.1.1 A first definition and gestalt theory

The notion of "spatial redundancy", as the ones of "texture" or "similarity", varies from one author to another, depending on the theoretical setting. Before introducing a rigorous statistical framework, we provide some perceptual and visual insights with regards to the notion of spatial redundancy.

Intuitively, an image is spatially redundant if it contains a repeated small pattern. Therefore, the notion of spatial redundancy is highly dependent on the degree of locality we consider. In order to meet this requirement we consider *patches* in the image. Patches are defined as follow: assume that an image x is given by $x : \mathsf{E} \to \mathsf{V}$, where $\mathsf{E} \subset \mathbb{Z}^2$ and $\mathsf{V} \subset \mathbb{R}$ (for a grey-level image) or $\mathbb{R}^3$ (for a color image). A patch is simply the restriction of x to some subset of E. More precisely, let $\mathsf{w} \subset \mathsf{E}$ and consider P_{w} such that for any $x : \mathsf{E} \to \mathsf{V}$, $P_{\mathsf{w}}(x) : \mathsf{w} \to \mathsf{V}$ and for any $p \in \mathsf{w}$, $P_{\mathsf{w}}(p) = x(p)$. This operator extracts from x a patch corresponding to the domain w.

Let $s : \mathsf{V}^{\mathsf{w}_1} \times \mathsf{V}^{\mathsf{w}_1} \to [0, +\infty)$ a similarity function which will be precised later. Consider one domain w_1 and its translated version by a vector $t \in \mathbb{Z}^2$, $\{p+t \, : \, p \in \mathsf{w}_1\}$. We say that the two patches associated with these domains are similar if $s(P_{\mathsf{w}_1}(x), P_{\mathsf{w}_1}(\tau_{-t}(x)))$ is small (in a sense which will also be precised later) and where for any $p \in \mathsf{E}$, $\tau_t(x)(p) = x(p-t)$ and E is stable under translation. We say that an image is spatially redundant if this similarity occurs for a large number of subdomains. This leads us to consider the following principle: "An image is spatially redundant if a large number of its patches are similar".

A first instance of spatial redundancy can be observed in Figure 2.1-(a). In this case, it is clear that spatial redundancy gives rise to new structures in the image. In this picture of the New York Guggenheim museum one can clearly identify some periodicity. However, spatial redundancy is not the only tool at our disposal to understand what the features which play a role into the global perceptual description of the image are. For instance, color, form, material and continuity are key elements for describing images. Some of these features can be observed in the painting "Black Lines" by Kandinsky (1913) Figure 2.1-(b).

Using spatial redundancy one can infer new perceptual properties on the image by studying the point process given by the positions of similar patches. Hence, the study of some of the perceptual features of the image boils down to the study of point processes. The perceptual characteristics of such geometric patterns were extensively studied in the context of the gestalt (form in German) theory.

With their pioneering works on optical illusions [Wer23; Kof13; Köh92], Weirtheimer, Koffka and Kölher propose a theory of human perception. One of its most important principle is the grouping principle: even though the perceptual components can be understood independently, they can be perceived differently once we take into account their intricate relationships. Weirtheimer [WW59] states:

> "The basic thesis of gestalt theory might be formulated thus: there are contexts in which what is happening in the whole cannot be deduced from the characteristics of the separate

(a) (b)

Figure 2.1: A photograph of the interior of the New York Guggenheim museum (a) presents numerous spatial redundancies highlighted by red, green and cyan patches. On the contrary, at first glance, no spatial redundancy is identified in the painting "Black Lines" by Kandinsky (1913). However this does not mean that the work of Kandinsky makes no sense from a perceptual point of view, since other geometrical tools can be applied to analyze the painting (color, curvature, convexity, etc) .

> pieces, but conversely; what happens to a part of the whole is, in clearcut cases, determined by the laws of the inner structure of its whole. ”
>
> *Max Wertheimer (1959)*

In Section 3.2 we study how spatial redundancy along with the principles of the gestalt theory can be used for image processing purposes. We focus on two main applications: denoising, see Section 3.2.4, and periodicity detection, see Section 3.2.5.

2.1.2 A contrario methods for image processing

There have been many attempts at developing a mathematical presentation of the grouping laws of the gestalt theory [Lin97] but in this thesis we focus on the approach developed by David Lowe [Low12] which is based on the a contrario method. An event (points alignement, similarity between two patches) is said to be significant if it has a very low probability in a background model (or noise model or a contrario model). This rule was later named Helmholtz's principle in the work of Desolneux, Moisan and Morel [DMM08] in reference to Helmholtz's studies on optical illusions [Hel25]. In the a contrario framework, the background model must be chosen so that the properties we want to identify in the natural image are uncommon. We now consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

This setting is similar to the one of the statistical hypothesis testing theory whose basic properties are recalled below. The aim of the test theory is to provide criteria in order to choose between a null hypothesis and an alternative hypothesis. Let U be a random variable taking values in $(\mathsf{V}^{\mathsf{E}}, \mathcal{B}(\mathsf{V}^{\mathsf{E}}))$, where $\mathcal{B}(\mathsf{V}^{\mathsf{E}})$ is the set of Borel sets and V^{E} is equipped with its product topology. We say that U is a random field [Adl81]. We denote by $\mathbb{P}_U$ the distribution of U, *i.e.* the pushforward measure by U. Let $\rho_0 \in \mathscr{P}(\mathsf{V}^{\mathsf{E}})$, where $\mathscr{P}(\mathsf{V}^{\mathsf{E}})$ is the set of probability measure on $\mathcal{B}(\mathsf{V}^{\mathsf{E}})$. The null hypothesis is $\mathbb{P}_U = \rho_0$ and the alternative hypothesis is $\mathbb{P}_U \neq \rho_0$. In the a contrario setting, ρ_0 is the background model (or noise model or a contrario model). In Section 2.1.3, we investigate different choices for the background

model when studying spatial redundancy. We define the following statistics

$$\mathrm{NFA} = \sum_{i=1}^{N} \mathbb{1}_{f_i(U) \in \mathsf{A}} \ ,$$

where $N \in \mathbb{N}^\star$, $(f_i)_{i \in \{1,\dots,N\}}$ is a family of measurable functions such that for any $i \in \{1,\dots,N\}$, $f_i : (\mathsf{V}^\mathsf{E}, \mathcal{B}(\mathsf{V}^\mathsf{E})) \to (\mathsf{V}^\mathsf{E}, \mathcal{B}(\mathsf{V}^\mathsf{E}))$. The NFA is the number of false alarms, or number of false positive. In the image processing framework, the event A represents some interest property of the image, for instance $\mathsf{A} = \{x \in \mathsf{V}^\mathsf{E} \ : \ \|P_\mathsf{w}(x) - P_\mathsf{w}(x_0)\| \leq \varepsilon\}$ where $x_0 \in \mathsf{V}^\mathsf{E}$ and $\varepsilon > 0$. In this case, if we consider $(f_i)_{i \in \{1,\dots,N\}} = (\tau_t)_{t \in \mathsf{E}}$ and E invariant under translation, then the number of false alarms is large in a noise model U if many patches of U are similar to the ones of x_0.

We reject the null hypothesis if the number of false alarms is larger than a maximal number of false alarms $\mathrm{NFA}_{\max}$. The type I error α is given by $\alpha = \mathbb{P}(\mathrm{NFA} \geq \mathrm{NFA}_{\max})$. Using Markov inequality we have

$$\alpha = \mathbb{P}\left(\mathrm{NFA} \geq \mathrm{NFA}_{\max}\right) \leq \mathrm{NFA}_{\max}^{-1} \sum_{i=1}^{N} \mathbb{P}\left(f_i(U) \in \mathsf{A}\right) \ .$$

In the a contrario framework we say that the event A is ε-significant if $\sum_{i=1}^{N} \mathbb{P}\left(f_i(U) \in \mathsf{A}\right) \leq \varepsilon$. We give an a contrario interpretation of the number of false alarms NFA. For any $i \in \{1,\dots,N\}$ we consider d_i such that for any $x \in \mathsf{V}^\mathsf{E}$, $d_i = \mathbb{1}_\mathsf{A}(f_i(x))$. We say that the index $i \in \{1,\dots,N\}$ is detected in the image x if $d_i(x) = 1$. We obtain that $\mathrm{NFA} = \sum_{i=1}^{N} d_i(U)$. If A is a ε-significant event, we get

$$\mathbb{E}\left[\mathrm{NFA}\right] = \mathbb{E}\left[\sum_{i=1}^{N} d_i(U)\right] \leq \varepsilon \ .$$

In other words, ε is an upper-bound on the expected number of detections in the background model. In the study of spatial redundancy $(f_i)_{i \in \{1,\dots,N\}}$ is given by $(\tau_t)_{t \in \mathsf{E}}$ (assuming that E is invariant under translation), see Section 3.2. Therefore we obtain $N = |\mathsf{E}|$ (where $|\mathsf{E}|$ is the cardinality of E). Given a patch $u_0 : \mathsf{w} \to \mathsf{V}$ we define

$$\mathsf{A} = \left\{x \in \mathsf{V}^\mathsf{E} \ : \ s(P_\mathsf{w}(x), u_0) \leq \mathrm{v}\right\} \ ,$$

where $\mathrm{v} \in \mathbb{R}$ is some value to be fixed and s is some similarity function, $\|\cdot\|_2$ for instance. In this case, an offset $t \in \mathsf{E}$ is detected in an image x if and only if $s(P_\mathsf{w}(\tau_t(x)), u_0) \leq \mathrm{v}$. If the background model ρ_0 is stationary, *i.e.* if the distribution of U is invariant by translation, then we get for any $t \in \mathsf{E}$, $\mathbb{P}_{\tau_t(U)} = \mathbb{P}_U = \rho_0$. In order to fix v, we consider an upper-bound on the expected number of detections ε in the background model. Hence, we choose $\mathrm{v} \in \mathbb{R}$ such that

$$\rho_0(\mathsf{A}) = \mathbb{P}\left(s((P_\mathsf{w}(U), u_0) \leq \mathrm{v})\right) = \varepsilon / |\mathsf{E}| \ .$$

We denote CDF the cumulative distribution function of $s(P_\mathsf{w}(U), u_0)$ and ICDF its inverse cumulative distribution function. An offset $t \in \mathsf{E}$ is detected if and only if $s(P_\mathsf{w}(\tau_t(x)), u_0) \leq \mathrm{ICDF}(\varepsilon/|\mathsf{E}|)$ or $\mathrm{CDF}(s(P_\mathsf{w}(\tau_t(x)), u_0)) \leq \varepsilon/|\mathsf{E}|$. Considering this methodology for each possible offset, we obtain a binary image which corresponds to the offsets detected in the image x, *i.e.* $(d_t(x))_{t \in \mathsf{E}} \in \{0,1\}^\mathsf{E}$. This image corresponds to some point process which might be analyzed using the gestalt theory principles.

Until now we have presented the a contrario methodology and its application to the identification of spatial redundancy in natural images. However, there exist many other successful applications of the a contrario approach in the study of the grouping laws of the gestalt [DMM00; DMM01; ADV03; Cao04; VG+08; Dav+18]. For instance, the Line Segment Detection (LSD) algorithm [VG+08] aims at identifying

(a) (b)

Figure 2.2: **A line detection algorithm.** In (a) we present the original image and in (b) a binary image. The black lines correspond to the lines detected by the LSD algorithm, see http://demo.ipol.im/demo/gjmr_line_segment_detector/ for an online demonstration.

alignements in natural images. In this framework we define $\mathsf{A} = \{x \in \mathsf{V}^{\mathsf{E}},\ k \text{ points are aligned in } r_0(x)\}$ where $r_0(x)$ is a rectangular sub-image of x and $k \in \mathbb{N}$. The notion of alignement is defined by the angles between the gradients of the original image, see [VG+08]. Let $(f_i)_{i\in\{1,\dots,N\}}$ be the set of translations-dilatations-rotations of E, see [VG+08] for a precise definition and a tally of such transformations. The authors of [VG+08] propose a background model ρ_0 such that $\rho_0(\mathsf{A})$ can be explicitly computed. Given a significance level ε, the detections correspond to the set of detected rectangles for this level. If we overlay the detected rectangles we obtain some binary image. This image is the output of the LSD algorithm illustrated in Figure 2.2.

In this thesis we present a detection algorithm for the analysis of similarities based on the principles of the gestalt theory and on the a contrario method. Given an image x we can obtain a binary image of the detected offsets. We use this information to improve an existing patch denoising algorithm, see Section 3.2.4, and to conduct a periodicity analysis, see Section 3.2.5. We now turn to the class of random fields used as a background model ρ_0: the Gaussian random fields.

2.1.3 Texture and random fields

A random field U is a $(\mathsf{V}^{\mathsf{E}}, \mathcal{B}(\mathsf{V}^{\mathsf{E}}))$-valued random variable [Adl81]. For any $\omega \in \Omega$, $V(\omega) \in \mathsf{V}^{\mathsf{E}}$. By a slight abuse of notation we will omit the dependency with respect to ω when there is no ambiguity. A random field U is said to be Gaussian if for any set of points $(p_1, \dots, p_n) \in \mathsf{E}^n$, $(U(p_1), \dots, U(p_n))$ is a Gaussian vector. In this thesis, the background model used for the a contrario analysis conducted in Section 2.1.2 will correspond to the null-hypothesis for the detection of spatial redundancy. Informally, the background model is an image model such that each realization does not exhibit spatial redundancy.

In the context of image processing, random fields were used by Cross and Jain for texture synthesis in their pioneering work [CJ83]. We briefly recall the purpose of exemplar-based texture synthesis. Given an input texture x_0 (the exemplar texture), we aim at finding a random field U such that the realizations of U look like x_0 but are not verbatim copies of x_0. We give some examples of textures in Figure 2.3. A key feature for ordering textures is their regularity. A structured texture, as the one shown in Figure 2.3-(c) will be called a *macrotexture*. On the contrary, a texture which does not exhibit any salient structure, such as the one shown in Figure 2.3-(a), will be called *microtexture*.

(a) (b) (c)

Figure 2.3: **Examples of textures.** Some examples of natural textures. The first image (a) does not exhibit any particular structure whereas the texture shown in (c) is extremely structured. The image (b) presents some intermediate organization since the pine thorns respect some rotation invariance with respect to the branch.

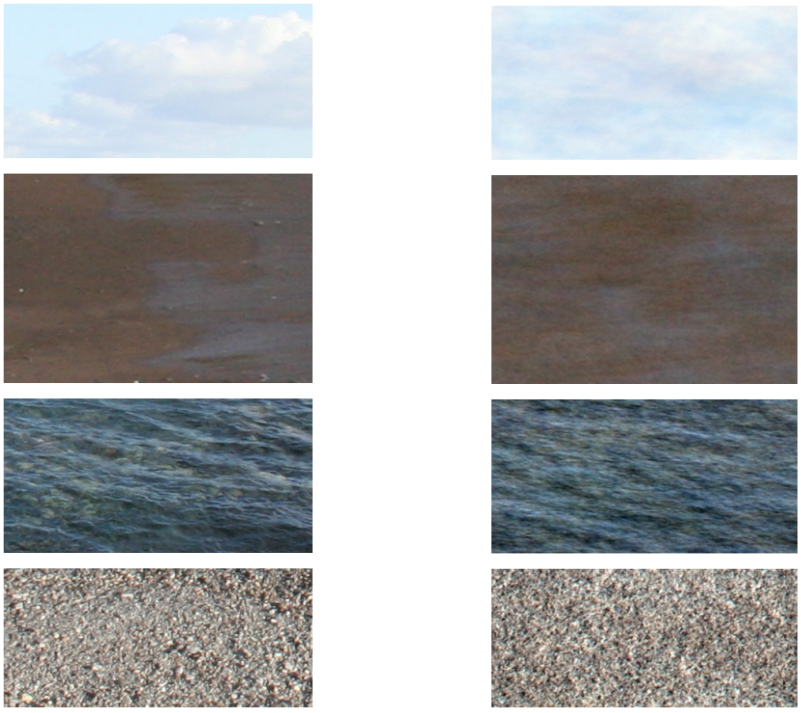
Van Wijk [Wij91] was among the first to use spots for texture synthesis. The synthesis process is the following. First, select a spot (some geometric pattern) and sample a point process (Bernoulli or Poisson process) on the underlying grid of the image, *i.e.* the set E. Then, we center the spot around each of the sampled points. This point of view was later extended [GGM11] in the case where $\mathsf{E} = \mathbb{Z}/(M\mathbb{Z}) \times \mathbb{Z}/(N\mathbb{Z})$ with $M, N \in \mathbb{N}$. The authors then consider the limit of this process when the number of points goes to infinity (in the case of a Bernoulli process) or when the intensity goes to infinity (in the case of a Poisson process). In both cases, recalling that x_0 is the exemplar texture, we obtain a Gaussian random field (after renormalization and centering) such that for any $p_1, p_2 \in \mathsf{E}$

$$\mathbb{E}\left[U(p_1)\right] = |\mathsf{E}|^{-1} \sum_{p \in \mathsf{E}} x_0(p) \ , \qquad \mathrm{Cov}\left[U(p_1), U(p_2)\right] = |\mathsf{E}|^{-1} \sum_{p \in \mathsf{E}} x_0(p_1 - p_2 + p) x_0(p) \ . \tag{2.1}$$

Note that U is a stationary random field and hence exhibits some homogeneity, which, however, is not sufficient to describe the set of textures. Many experiments [GL17; GLM14; GLR18] illustrate how the Gaussian random field U is a good model for microtexture exemplar-based texture synthesis, see Figure 2.4. However, this method fails for the more complicated problem of macrotexture exemplar-based texture synthesis, see Figure 2.5.

From these last two experiments we formulate the following principle: "the Gaussian random field associated with x_0 discards the spatial redundancy information while conserving the microtexture information". Although this principle is a drawback for the task of structured texture synthesis, it can be used to consider a contrario models based on Gaussian random fields. In Chapter 3, combining a contrario methods and Gaussian random fields background models we propose an algorithm to detect spatial redundancy and design new denoising and periodicity detection algorithms.

In this section we have addressed the problem of identifying spatial redundancy in natural images. Some of these redundant structures can be discovered by rejecting the null hypothesis in an a contrario model. Since we proceed by rejection, we only need to consider naive texture models. However, the simplicity of such a methodology is also its limitation as it does not provide us with a way to enrich our models in order to take into account these notions of spatial redundancy, *i.e.* we cannot sample more complex random fields.



(a) (b)

Figure 2.4: **Gaussian random field and microtexture.** The microtexture exemplar textures are given in (a) and in (b) we present some realizations of the Gaussian random field given by (2.1). Figure extracted from [GGM11].

2.2 Random field synthesis and maximum entropy principle

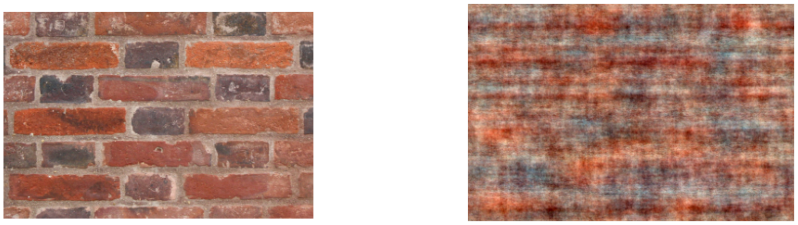
In the previous section we showed that microtextures, *i.e.* random fields with no long-range dependency, can be efficiently sampled using Gaussian random fields. However, numerous random fields studied in fluid mechanics, astrophysics or image processing are not microtextures and exhibit long-range interactions. We present some examples of such random fields in Figure 2.6.

In this thesis we are interested in the problem of sampling random fields from the point of view of image processing, via the exemplar-based texture synthesis. This problem constitutes the first step towards the sampling of structured random fields.

2.2.1 Parametric texture synthesis

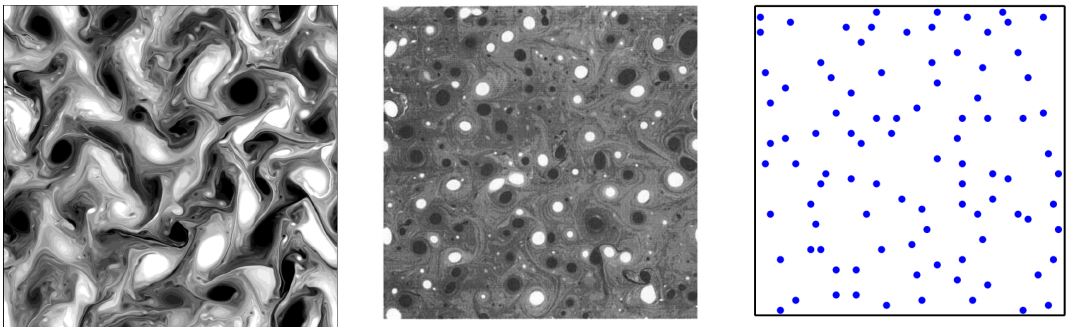
We recall that the exemplar-based texture synthesis problem can be described as follows. Given an input texture image x_0, how can we produce new images such that:

- the visual properties of the new images are close to the ones of x_0,
- the new images are not verbatim copies of x_0.



(a) (b)

Figure 2.5: **Gaussian random field and macrotexture.** The macrotexture exemplar textures are given in (a) and in (b) we present some realizations of the Gaussian random field given by (2.1). Figure extracted from [GGM11].



(a) (b) (c)

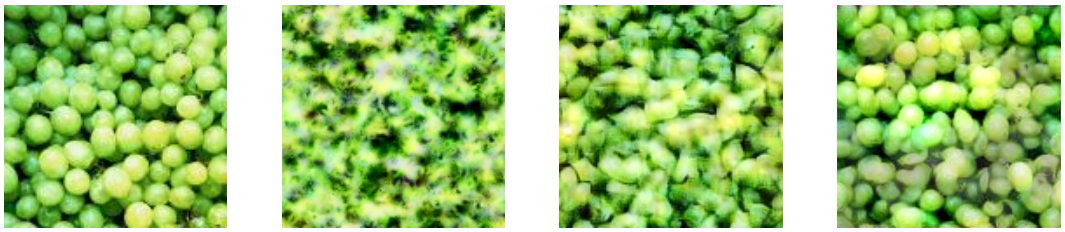
Figure 2.6: **Random fields.** In (a) and (b) we present some realizations of turbulent random fields. Images extracted from [Hel+95]. In (c), we present a realization of a determinantal point process which is a point process with repulsion constraints. Image extracted from [KT11].

There exist two main approaches to this problem [Raa+17]: the *parametric* methods and the *non-parametric* methods (or patch methods). In this thesis we focus on parametric methods. We refer to Section 5.2.1 for a literature review on the subject of parametric texture synthesis. In what follows we set $\mathsf{X} = \mathsf{V}^{\mathsf{E}}$ and we suppose that $\mathsf{X} \in \mathcal{B}(\mathbb{R}^d)$ for some $d \in \mathbb{N}$.

One of the first parametric method for texture synthesis is the algorithm proposed by Portilla and Simoncelli [PS00]. We start by identifying a certain number of statistics (local or non-local) which we gather in a constraints dictionary. Given $p \in \mathbb{N}$ and a family of measurable functions $(f_i)_{i\in\{1,\dots,p\}}$ such that for any $i \in \{1,\dots,p\}$, $f_i : (\mathsf{X}, \mathcal{B}(\mathsf{X})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $x_0 \in \mathsf{X}$, where x_0 is the input texture and $(f_i)_{i\in\{1,\dots,p\}}$ are some constraints, we define $\mathsf{Y} = \{x \in \mathsf{X},\ f_i(x) = f_i(x_0)\}$. This set is called the *Julesz ensemble* in [PS00]. Assume that there exists a measurable function $\Pi : (\mathsf{X}, \mathcal{B}(\mathsf{X})) \to (\mathsf{Y}, \mathcal{B}(\mathsf{Y}))$ such that for any random field U taking values X we get that $\Pi(U)$ takes values in Y, then we are provided with an easy way to generate random variables taking values in Y. Unfortunately, it is often difficult to obtain such a mapping.

If for any $i \in \{1,\dots,p\}$ we have access to $\Pi_i : (\mathsf{X}, \mathcal{B}(\mathsf{X})) \to (\mathsf{X}, \mathcal{B}(\mathsf{X}))$ such that for any $x \in \mathsf{X}$, $f_i(\Pi_i(x)) = f_i(x_0)$, then we can define

$$X_{n+1} = \Pi_{n-\lfloor n/p \rfloor p+1}(X_n)\ , \tag{2.2}$$



(a) (b) (c) (d)

Figure 2.7: **Some texture synthesis algorithms.** In (a) we present the exemplar texture. In (b) we show the results obtained with the pyramidal synthesis proposed in [HB95]. In (c) we observe the results obtained with [PS00] and in (d) we apply the method of [GEB15]. This method, where the constraints are given by neural network outputs, is the only one which succeeds in synthesizing the complex structures of the image.

where X_0 is some random variable taking values in X. This approach was considered by Heeger and Bergen in [HB95], in the case where for any $i \in \{1, \dots, p\}$, Π_i corresponds to some histogram egalization. There exists no general convergence results for such an algorithm, except in the case where $(\Pi_i)_{i\in\{1,\dots,p\}}$ is a collection of strictly quasinonexpansive operators [BC11, Corollary 4.50, Theorem 5.23]. In particular, the results hold if $(\Pi_i)_{i\in\{1,\dots,p\}}$ is a collection of projection operators onto convex and closed sets. Portilla and Simoncelli replace (2.2) by a projection in the direction of the gradient, see [PS00, Section 1.5]. This operation only requires the gradient of the constraint functions $(f_i)_{i\in\{1,\dots,p\}}$.

Heeger and Bergen use histogram data to build their constraints dictionary, whereas Portilla and Simoncelli identify 710 constraints to define their image model (marginal statistics of wavelet coefficients, correlation between wavelet coefficients and phase statistics). Gatys [GEB15] extended the algorithm presented in [PS00] by considering statistics of Gram matrices for different layers of a pretrained convolutional neural network, VGG-19 [SZ14]. The algorithm of [GEB15] consists in a gradient descent for the loss function $\ell : \mathsf{X} \to [0, +\infty)$ given for any $x \in \mathsf{X}$ by

$$\ell(x) = \sum_{i=1}^{L} \lambda_i \|G_i(x) - G_i(x_0)\|_{\mathrm{Fr}}^2 ,$$

where $(\lambda_i)_{i\in\{1,\dots,L\}} \in [0, +\infty)^L$ is a sequence of weights, $G_i(x)$ is the Gram matrix of the neural network output and $\|\cdot\|_{\mathrm{Fr}}$ is the Frobenius norm. This last algorithm yields state-of-the-art visual results. In Figure 5.1, we compare the three described algorithms for exemplar-based texture synthesis.

2.2.2 Principle of maximum entropy

All the models we consider in this thesis are parametric methods. In this case, the exemplar-based texture synthesis problem can be reframed as an inverse problem as follows. In this section, we assume that X is some topological space and consider $\mathcal{B}(\mathsf{X})$ the associated Borel sigma-field. Let $p \in \mathbb{N}$ and consider $(f_i)_{i\in\{1,\dots,p\}}$ such that for any $i \in \{1, \dots, p\}$, $f_i : (\mathsf{X}, \mathcal{B}(\mathsf{X})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $x_0 \in \mathsf{X}$. We aim at finding π such that for any $i \in \{1, \dots, p\}$

$$\int_{\mathsf{X}} \mathbb{1}_{f_i(x)\neq f_i(x_0)} \mathrm{d}\pi(x) = 0 , \tag{2.3}$$

i.e. $f_i = f_i(x_0)$, π almost surely. We consider a relaxed version of (2.3), where the constraint is imposed in expectation instead of almost surely,

$$\int_{\mathsf{X}} |f_i(x)| \mathrm{d}\pi(x) < +\infty \ , \qquad \int_{\mathsf{X}} f_i(x) \mathrm{d}\pi(x) = f_i(x_0) \ . \tag{2.4}$$

In what follows we denote $F : \mathsf{X} \to \mathbb{R}^p$ such that for any $x \in \mathsf{X}$, $F(x) = (f_1(x) - f_1(x_0), \dots, f_p(x) - f_p(x_0))$ and we assume that the family $\{f_i \ : \ i \in \{1, \dots, p\}\} \cup 1$ is linearly independent.

The inverse problem given by (2.4) is an ill-posed inverse problem since it admits the following trivial solution $\pi = \delta_{x_0}$. One way to tackle this problem is to search for the "most uniform" distribution given the constraints (2.4). In the context of discrete probabilities, Jaynes [Jay57] gives a solution using Shannon entropy [Sha48]. Assume that $\mathsf{X} = \{1, \dots, M\}$ with $M \in \mathbb{N}$. Let π be a probability measure on X. The Shannon entropy associated with π is given by

$$H(\pi) = -\sum_{x \in \mathsf{X}} \pi(x) \log(\pi(x)) \ ,$$

with the convention that $0 \log(0) = 0$. We aim at maximizing H under the constraints (2.4). The obtained model is the *macrocanonical* model. If we consider the model given by the constraints (2.3), we obtain the *microcanonical* model, studied in [BM18]. The links between the two models when the dimension of the image space grows towards infinity are discussed in [BM18]. In this thesis, we focus on macrocanonical models but we will illustrate the links between this model and the microcanonical one in 5.

Maximizing H under the constraints (2.4) can be cast as a convex minimization problem under linear constraints. In this case, Jaynes obtains the following solution $\pi^\star$ given for any $x \in \mathsf{X}$.

$$\pi^\star(x) = \exp[-\langle \theta^\star, F(x) \rangle] \Big/ \sum_{y \in \mathsf{X}} \exp[-\langle \theta^\star, F(y) \rangle] \ ,$$

where $\theta^\star \in \mathbb{R}^p$ corresponds to the Lagrange multipliers of the associated dual problem:

$$\theta^\star \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^p} L(\theta) \ , \qquad L(\theta) = \log \left(\sum_{x \in \mathsf{X}} \exp[-\langle \theta, F(x) \rangle] \right) \ ,$$

where L is the *log-partition function*. Jaynes [Jay79] identifies this principle as a generalization of the "principle of indifference": in the absence of additional information, $\pi^\star$ must be chosen to be uniform. We report to [Jay79] for a discussion on the origin of the principle of maximum entropy. In image this principle was first used for image restoration [Wer+77; GD78; SB84; Bes86] before it was used in texture synthesis [ZWM98];

The extension of the principle of maximum entropy to the case where X is no longer finite is more complicated. We will now assume that $\mathsf{X} = \mathbb{R}^d$ with $d \in \mathbb{N}^\star$. First we highlight that there is no equivalent of the "principle of indifference" in this context, *i.e.* there is no uniform probability distribution on $\mathbb{R}^d$. Moreover, we have that for any probability measure π such that the entropy $H(\pi)$ is well-defined,

$$H(\pi) = -\int_{\mathsf{X}} (\mathrm{d}\pi / \mathrm{d}\lambda)(x) \log \left[(\mathrm{d}\pi / \mathrm{d}\lambda)(x) \right] \mathrm{d}x \ ,$$

we have that $H(\pi) \in (-\infty, +\infty]$. In [BLN96; TV93; Lóf8] the authors establish equivalent of the maximum entropy principle by modifying the Shannon entropy and using techniques from functional analysis (Birnbaum-Orlicz spaces) and convex optimization in Banach spaces. On the other hand, Csiszár

and his co-authors [Csi75; Csi84; Csi96; CGG99] equip the space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ with a *reference probability measure* μ and replace the Shannon entropy by the opposite of the Kullback-Leibler divergence [Kul97] between π and μ. The Kullback-Leibler divergence between two probability measures ν_1 and ν_2 is given by

$$\mathrm{KL}\left(\nu_1|\nu_2\right) = \begin{cases} \displaystyle\int_{\mathsf{X}} \log[(\mathrm{d}\nu_1/\mathrm{d}\nu_2)(x)]\mathrm{d}\nu_1(x) & \text{if } \nu_1 \ll \nu_2 \ , \\ +\infty & \text{otherwise} \ . \end{cases}$$

Note that $\mathrm{KL}\left(\nu_1|\nu_2\right) \geq 0$ with equality if and only if $\nu_1 = \nu_2$. In [Csi75], the results of Jaynes are extended using this approach. We replace the Shannon entropy in the maximization by the opposite Kullback-Leibler divergence with respect to the reference probability measure μ, *i.e.* we consider

$$\pi^\star \in \operatorname*{arg\,min}_{\pi \in \mathscr{P}_F} \mathrm{KL}\left(\pi|\mu\right) \ . \tag{2.5}$$

where $\mathscr{P}_F$ is the set of probability measures such that for any $\pi \in \mathscr{P}_F$, $\int_{\mathsf{X}} F(x)\mathrm{d}\pi(x) = 0$. In this case, if the solution of the maximum entropy problem exists (2.5) and is dominated by μ, we have for any $x \in \mathsf{X}$

$$(\mathrm{d}\pi^\star/\mathrm{d}\mu)(x) = \exp[-\langle\theta^\star, F(x)\rangle]/\int_{\mathsf{X}} \exp[-\langle\theta^\star, F(y)\rangle]\mathrm{d}\mu(y) \ , \tag{2.6}$$

where $\theta^\star \in \mathbb{R}^p$ is the solution to the maximum entropy problem. In the same article, [Csi75], conditions are given to ensure the existence of a solution to the maximum entropy problem. If for any $(\varepsilon_1, \ldots, \varepsilon_p) \in \mathrm{B}(0, r)$ with $r > 0$, there exists π such that $\mathrm{KL}\left(\pi|\mu\right) < +\infty$ and π satisfies (2.4), replacing for any i $i \in \{1, \ldots, p\}$, $f_i(x_0)$ by $f_i(x_0) + \varepsilon_i$, and if for any $\theta \in \mathbb{R}^p$, $\int_{\mathsf{X}} \exp[-\langle\theta, F(x)\rangle]\mathrm{d}\mu(x) < +\infty$ then [Csi75, Theorem 3.3] ensures the existence of a solution to the maximum entropy problem. In the setting of exemplar-based texture synthesis, this theorem is difficult to apply and we derive other sufficient conditions to ensure the existence of such a model.

In Chapter 5 we use the principle of maximum entropy in the case of texture synthesis. We establish easy-to-check conditions on the model F and the probability measure μ for the existence of a solution to hold. In particular, when F is given by a neural network we provide a certificate which ensures the existence of maximum entropy model.

2.2.3 Stochastic optimization

Before deriving texture synthesis algorithms based on the macrocanonical approach described in the previous section, we must answer the two following questions:

(a) How to estimate $\theta^\star$ in (2.6)?

(b) Given $\theta^\star$, how to sample from the distribution $\pi^\star$ given in (2.6)?

In this section, we present the tools used in this thesis to tackle the problem of sampling in Section 2.3. In this section, we are interested in the problem (a) and assume that we have access to samples from the distribution π_θ for any $\theta \in \mathbb{R}^p$, where for any $x \in \mathsf{X}$

$$(\mathrm{d}\pi_\theta/\mathrm{d}\mu)(x) = \exp[-\langle\theta, F(x)\rangle] \Big/ \int_{\mathsf{X}} \exp[-\langle\theta, F(y)\rangle]\mathrm{d}\mu(y) \ .$$

Under conditions on μ and F, we show in Chapter 5 that $\theta^\star$ in (2.6) is the minimizer of the log-partition function $L : \mathbb{R}^p \to \mathbb{R}$ given for any $\theta \in \mathbb{R}^p$ by

$$L(\theta) = \log\left(\int_{\mathsf{X}} \exp[-\langle\theta, F(x)\rangle]\mathrm{d}\mu(x)\right) \ .$$

We consider a first-order minimization method, *i.e.* we consider a sequence $(\theta_n)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}$, θ_{n+1} is a function of θ_n and $\nabla L(\theta_n)$. However, the gradient of L cannot be explicitly computed in practice since it is expressed as an integral with respect to π_θ: for any $\theta \in \mathbb{R}^p$

$$\nabla L(\theta) = -\int_{\mathsf{X}} F(x) \mathrm{d}\pi_\theta(x) = -\mathbb{E}_{\pi_\theta}\left[F\right] \ .$$

Nevertheless, this quantity can be approached. We consider an unbiased estimator of $\nabla L(\theta) = -\mathbb{E}_{\pi_\theta}\left[F\right]$ as follows. Assuming that we have access to $M \in \mathbb{N}$ samples from π_θ, $-(1/M)\sum_{k=1}^{M} F(X_k)$ is an estimator of $\nabla L(\theta)$. We now consider the stochastic gradient algorithm associated with the following recursion: $\theta_0 \in \mathbb{R}^p$ and for any $n \in \mathbb{N}$

$$\theta_{n+1} = \theta_n + (\delta_{n+1}/M_{n+1}) \sum_{k=1}^{M_{n+1}} F(X_k^n) \ , \tag{2.7}$$

where $(\delta_n)_{n\in\mathbb{N}}$ is a sequence of stepsizes, $(M_n)_{n\in\mathbb{N}}$ is a sequence of batchsizes and for any $n \in \mathbb{N}$, $(X_k^n)_{k\in\{1,\ldots,M_n\}}$ is a sequence of independent samples from π_{θ_n}.

In (2.7) we assume that we have access to samples from π_θ. We now consider a generalization of (2.7) where the sequence of estimators is no longer given by $(-(1/M_{n+1})\sum_{k=1}^{M_{n+1}} F(X_k^n))_{n\in\mathbb{N}}$ but by $(\nabla L(\theta_n) + e_{\theta_n}(Y_{n+1}))_{n\in\mathbb{N}}$ where $(Y_n)_{n\in\mathbb{N}}$ is a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $(\mathsf{Y}, \mathcal{Y})$, and where $e : \mathbb{R}^p \times \mathsf{Y} \to \mathbb{R}^p$. We now consider the following general stochastic gradient algorithm associated with the following recursion: $\theta_0 \in \mathbb{R}^p$ and for any $n \in \mathbb{N}$

$$\theta_{n+1} = \theta_n - \delta_{n+1}\left\{\nabla L(\theta_n) + e_{\theta_n}(Y_{n+1})\right\} \ , \tag{2.8}$$

For instance, in (2.7) we can define for any $n \in \mathbb{N}$ $n \in \mathbb{N}$,

$$e_{\theta_n}((X_k^n)_{k\in\{1,\ldots,M_{n+1}\}}) = -(1/M_{n+1}) \sum_{k=1}^{M_{n+1}} F(X_k^n) - \nabla L(\theta_n) \ . \tag{2.9}$$

The study of such stochastic approximation methods can be traced back to the pioneering work of Robbins and Monro [RM51] and Kiefer and Wolfowitz [KW52]. Nowadays, these methodologies are widely used to train neural networks [BLC05]. There exist many general convergence results concerning these methods, see [MP84; BMP90; DJ93; Ben96; Del96; KY03].

Using a very general result on the almost-sure convergence of such schemes based on the existence of Lyapunov function [DLM99, Theorem 2] and the Sard theorem [Sar42] we obtain that, almost surely, $\theta^\star = \lim_{n\to+\infty} \theta_n$ exists and $\theta^\star \in \{\theta \ : \ \nabla L(\theta) = 0\}$. However, since L is convex we can be even more precise and provide quantitative non-asymptotic convergence results, see [SZ13; BM11; AFM17] for instance. In particular, under condition, [AFM17] shows that for any $n \in \mathbb{N}$

$$\mathbb{E}[L(\hat{\theta}_n)] - \min_{\theta\in\mathbb{R}^p} L(\theta) \leq \mathbb{E}\left[(1/2)\left\|\theta_0 - \theta^\star\right\|^2 \right.$$
$$\left. - \sum_{k=0}^{n-1} \delta_{k+1}\langle \theta_k - \delta_{k+1}\nabla L(\theta_k) - \theta^\star, e_{\theta_k}(Y_{k+1})\rangle + \sum_{k=1}^{n} \delta_k^2 \left\|e_{\theta_k}(Y_{k+1})\right\|^2\right] \Bigg/ \sum_{k=1}^{n} \delta_k \ ,$$

where $\hat{\theta}_n = \sum_{k=1}^{n} \delta_k\theta_k / \sum_{k=1}^{n} \delta_k$. In our case, we do not have access to independent samples from π_θ. However there exist numerous algorithms which consider Markov chains targetting π_θ. More precisely, we will consider approximation schemes of the form (2.8) et (2.9), where for any $n \in \mathbb{N}$, $\{X_k^n, k \in \{1, \ldots, M_{n+1}\}\}$ is a Markov chain which converges towards a (biased) version of π_{θ_n}. Using the ergodic properties of the underlying Markov chain, we can obtain quantitative convergence results for the general stochastic approximation algorihm, see [AFM17; DB+19].

2.3 MCMC sampling, Langevin dynamics and convergence of Markov chains

We now assume that $\mathsf{X} = \mathbb{R}^d$ and that we are provided with a probability distribution π onto the image space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. In most cases, this distribution π will be obtained using the maximum entropy principle. Assume that the distribution π admits a density with respect to the Lebesgue measure given for any $x \in \mathbb{R}^d$ by

$$(\mathrm{d}\pi/\mathrm{d}\lambda)(x) = \exp[-U(x)] \Big/ \int_{\mathbb{R}^d} \exp[-U(y)]\mathrm{d}y \ , \tag{2.10}$$

where $U : \mathbb{R}^d \to \mathbb{R}$ is a measurable function such that $\int_{\mathbb{R}^d} \exp[-U(y)]\mathrm{d}y < +\infty$. By analogy with statistical physics, we call U the *potential function*. We now aim at obtaining samples from π. In the context of texture synthesis, these new samples will be examples of images exhibiting the same perceptual characteristics as x_0, the exemplar texture.

2.3.1 Statistical sampling

First, we recall some basic sampling methods and their limitations. Assume that $d = 1$. Then one can use the inverse transform sampling given the knowledge of the cumulative distribution function. Unfortunately, this method is specific to the case where $d = 1$ and requires the knowledge of the cumulative distribution function. The rejection sampling method can be used to sample from random variables in dimension $d > 1$ without the knowledge of the cumulative distribution function or the normalization constant. However, it is often inefficient since the percentage of rejected samples is usually high for complex statistical problems.

The two previous methods produce samples which have exactly the same distribution as π but are inefficient for high dimensional problems. Instead of imposing that all the samples are distributed according to π, we are going to consider a sequence of samples $(X_n)_{n\in\mathbb{N}}$ such that the distribution of X_n will converge towards π in a sense which will be precised later. One of the first methodology introduced to deal with this problem is the Metropolis-Hastings algorithm [Met+53; Has70; Pes73]. We recall its main ingredients here. Let R be the Markov kernel given for any $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\mathrm{R}(x, \mathsf{A}) = \delta_x(\mathsf{A}) \int_{\mathbb{R}^d} (1 - \alpha(x, y))q(x, y)\mathrm{d}y + \int_{\mathsf{A}} \alpha(x, y)q(x, y)\mathrm{d}y \ ,$$

where for any $x \in \mathbb{R}^d$, $q(x, \cdot)$ is some density called the *proposal density*, $\alpha : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ is the *acceptance ratio* and δ_x is the Dirac mass at x. We consider the following recursion: $X_0 \in \mathbb{R}^d$ and for any $n \in \mathbb{N}$

$$X_{n+1} = (1 - W_{n+1})X_n + W_{n+1}Y_{n+1} \ ,$$

with Y_{n+1} distributed according to $q(X_n, \cdot)$ conditionally to X_n, W_{n+1} a random Bernoulli variable with parameter $\alpha(X_n, Y_{n+1})$ conditionally to X_n and Y_{n+1}. In this case, X_n is distributed according to $\mathrm{R}^n(X_0, \cdot)$ for all $n \in \mathbb{N}^\star$. If the two measures μ_1, μ_2 given for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\mu_1(\mathsf{A}) = \int_{\mathbb{R}^{2d}} \mathbb{1}_{\mathsf{A}}(x, y)\mathrm{R}(x, \mathrm{d}y)\mathrm{d}\pi(x) \ , \qquad \mu_2(\mathsf{A}) = \int_{\mathbb{R}^{2d}} \mathbb{1}_{\mathsf{A}}(y, x)\mathrm{R}(x, \mathrm{d}y)\mathrm{d}\pi(x) \ , \tag{2.11}$$

are equal then R is reversible with respect to π, *i.e.* R is self-adjoint in $\mathrm{L}^2(\pi)$. As a consequence, π is invariant with respect to R, *i.e.* $\pi\mathrm{R} = \pi$. In the case where π admits a density with respect to the Lebesgue measure if we have for any $x, y \in \mathbb{R}^d$

$$\alpha(x, y) = \min\left(1, \frac{(\mathrm{d}\pi/\mathrm{d}\lambda)(y)q(y, x)}{(\mathrm{d}\pi/\mathrm{d}\lambda)(x)q(x, y)}\right) \ ,$$

then (2.11) is satisfied. Note that this method does not require the knowledge of the normalization constant. We still have to choose the proposal density q. If one chooses a symmetric proposal density like $q(x,y) = (2\pi\sigma^2)^{-1/2}\exp[-\|x-y\|^2/(2\sigma^2)]$ with $\sigma > 0$, then we obtain the following simpler acceptation rate $\alpha(x,y) = (\mathrm{d}\pi/\mathrm{d}\lambda)(y)/(\mathrm{d}\pi/\mathrm{d}\lambda)(x)$. Other choices can be considered. For instance we can choose for any $x, y \in \mathbb{R}^d$

$$q(x,y) = (4\pi\gamma)^{-1/2}\exp\left[\|y - x - \gamma\nabla f(x)\|^2/(2\gamma)\right] , \tag{2.12}$$

where $\gamma > 0$, we recall that $-U$ is the log-density of π, see (2.10) and U is assumed to be differentiable. This proposal corresponds to the following update for all $n \in \mathbb{N}$

$$Y_{n+1} = X_n - \gamma\nabla U(X_n) + \sqrt{2\gamma}Z_{n+1} , \tag{2.13}$$

where $(Z_n)_{n\in\mathbb{N}}$ is a sequence of d-dimensional independent Gaussian random variables with zero mean and identity covariance matrix. The new state of the Markov chain is given by Y_{n+1} with probability $\alpha(X_n, Y_{n+1})$ and by X_n with probability $1 - \alpha(X_n, Y_{n+1})$. The update (2.13) consists in one step of gradient descent with stepsize γ, *i.e.* $\mathcal{T}_1(x) = x - \gamma\nabla U(x)$, followed by the addition of some Gaussian noise $\mathcal{T}_{2,n}(x) = x + \sqrt{2\gamma}Z_{n+1}$. We get that $Y_{n+1} = \mathcal{T}_{2,n}(\mathcal{T}_1(X_n))$. Therefore, (2.13) corresponds to one step of perturbed gradient descent where the perturbation is given by the operator $\mathcal{T}_{2,n}$. We will see in Section 2.3.2 that (2.13) corresponds to the discretization of the continuous dynamics with invariant measure π.

The algorithm associated with the proposal (2.12) was introduced by Besag in a commentary of [GM94] and is now known as the Metropolis Adjusted Langevin Algorithm (MALA). The acceptance rate depends on the position $x, y \in \mathbb{R}^d$ but also on the stepsize parameter $\gamma > 0$.

If the acceptance rate is too small (the stepsize is too big) then the Markov chain does not move because a lot of iterations consist in remaining at the same position since they are rejected. If the acceptance rate is too large (the stepsize is too small) then the Markov chain is not exploratory enough. In what follows we are going to consider non-adjusted Markov chain, *i.e.* Markov chains without acceptance-reject step. The price of this simplification is the loss of the knowledge of the invariant distribution which will no longer be π but some other probability distribution. However, we will be able to control its distance to π.

2.3.2 Discrete and continuous time Langevin dynamics

The unadjusted algorithm associated with (2.13) is called Unadjusted Langevin Algorithm (ULA) [RT96]. We obtain the following Markov chain: let $X_0 \in \mathbb{R}^d$ and set for any $n \in \mathbb{N}$

$$X_{n+1} = X_n - \gamma\nabla U(X_n) + \sqrt{2\gamma}Z_{n+1} , \tag{2.14}$$

where $(Z_n)_{n\in\mathbb{N}}$ is a sequence of d-dimensional independent Gaussian random variables with zero mean and identity covariance matrix. For any $\gamma > 0$ we denote R_γ the Markov kernel given for any $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\mathrm{R}_\gamma(x, \mathsf{A}) = (4\pi\gamma)^{-1/2}\int_{\mathsf{A}} \exp[-\|x - \gamma\nabla U(x) - y\|^2/(2\gamma)]\mathrm{d}y$$

In the recent years, ULA has attracted a lot of attention from the statistical and machine learning communities. Indeed, since the recursion (2.14) only requires the knowledge of ∇U, ULA can be efficiently implemented in the case where the potential function is given by neural network outputs using automatic differenciation. Since ULA can be adapted to handle stochastic gradient just like the stochastic

gradient descent which is extensively used in deep learning [BLC05; Nem+09] due to its efficiency, it is a prominent algorithm when it comes to sampling in high dimension. In order to theoretically investigate how ULA and its variant behave in high dimensional settings, it is necessary to obtain sharp quantitative non-asymptotic results with regards to the convergence of $(\mathrm{R}_\gamma^n)_{n\in\mathbb{N}}$ towards its invariant distribution.

The first convergence results for this algorithm are based on the comparison of the discrete-time process with its associated continuous-time process: the Langevin dynamics. This continuous-time process is given by the following Stochastic Differential Equation

$$\mathrm{d}\mathbf{X}_t = -\nabla U(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \ , \tag{2.15}$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion. This dynamics was first considered by Langevin [Lan08] who studied the dynamics of particles in fluids.

If $U \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R})$ and ∇U is Lipschitz continuous then for any initial condition $\mathbf{X}_0$, there exists a unique strong global solution $(\mathbf{X}_t)_{t\geq 0}$ [IW89, Chapter 4, Theorem 2.3, Theorem 2.4]. We now assume that there exists a unique strong global solution. We consider $(\mathrm{P}_t)_{t\geq 0}$ the family of Markov kernels defined as follows. For all $x \in \mathbb{R}^d$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ and $t \geq 0$ we have $\mathrm{P}_t(x, \mathsf{A}) = \mathbb{P}\left(\mathbf{X}_t \in \mathsf{A}\right)$, where $\mathbf{X}_0 = x$. Then one can show that π given by (2.10) is invariant with respect to $(\mathrm{P}_t)_{t\geq 0}$, see [Dur16]. We are now interested in the convergence of $(\mathrm{P}_t)_{t\geq 0}$ for various distances on the space of probability measures (total variation, Wasserstein distances of order p, V-norm).

The total variation of a finite signed measure μ, denoted $\|\mu\|_{\mathrm{TV}}$ is given by

$$\|\mu\|_{\mathrm{TV}} = (1/2) \sup_{f\in \mathrm{L}^\infty(\mathbb{R}^d,\mathbb{R}),\|f\|_\infty\leq 1} \left| \int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x) \right| \ .$$

Let μ and ν be two probability measures on $\mathcal{B}(\mathbb{R}^d)$. A probability measure ζ on $\mathcal{B}(\mathbb{R}^{2d})$ is a transference plan between μ and ν if for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\zeta(\mathsf{A}\times\mathbb{R}^d) = \mu(\mathsf{A})$ and $\zeta(\mathbb{R}^d \times \mathsf{A}) = \nu(\mathsf{A})$. We then note $\mathbf{T}(\mu,\nu)$ the set of all transference plans between μ and ν. For any measurable cost function $\mathbf{c} : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$, we define

$$\mathbf{W}_{\mathbf{c}}(\mu,\nu) = \inf_{\zeta\in\mathbf{T}(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathbf{c}(x,y)\mathrm{d}\zeta(x,y) \ .$$

If $\mathbf{c}(x,y) = \|x-y\|^p$ with $p \geq 1$ then we denote $\mathbf{W}_p = \mathbf{W}_{\mathbf{c}}^{1/p}$. $\mathbf{W}_p$ is a metric on the space of probability measures which admit a moment of order p, see [Vil09, Definition 6.1]. $\mathbf{W}_p$ is called the Wasserstein distance of order p.

The ergodicity of $(\mathrm{P}_t)_{t\geq 0}$ in total variation holds under mild "curvature at infinity" conditions [RT96]. More precisely, if $U \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R})$ satisfies

$$\inf_{x\in\mathbb{R}^d} \langle \nabla U(x), x\rangle > -\infty \ ,$$

then $\lim_{t\to+\infty} \|\mathrm{P}_t - \pi\|_{\mathrm{TV}} = 0$. Since [RT96], these ergodicity results have been improved and it has been proven that, under curvature conditions, $(\mathrm{P}_t)_{t\geq 0}$ converges towards π for many distances (Wasserstein of order p for $p \in \mathbb{N}^\star$, total variation, V-norm). Once the convergence (geometric or sub-geometric) of $(\mathrm{P}_t)_{t\geq 0}$ is established, it is possible to provide similar estimates on the convergence of the associated chain. Note that standard weak and strong approximations [TT90; Mil95] do not provide the expected convergence results as these results are finite-time estimates.

Since R_γ is strongly Feller, R_γ admits an invariant probability measure under weak Foster-Lyapunov conditions [Dou+18, Theorem 12.3.3]. The analysis of the geometric ergodicity of the discrete-time process was established recently [DT12; DM17; DM19; Dal17b; Dal17a] for various distances on the space of probability measures (Wasserstein of order p for $p \in \mathbb{N}^\star$, total variation, V-norm).

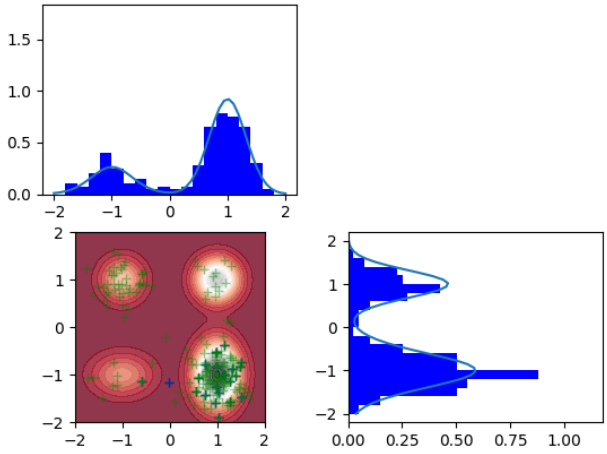


Figure 2.8: **The Unadjusted Langevin Algorithm.** In this figure we present 10^4 iterations of ULA for the sampling of a mixture of Gaussian in dimension 2. The green crosses correspond to all the iterations of the algorithm and the blue cross corresponds to the last iteration. After 10^4 iterations, we remark that the marginals corresponding to the projections on the two axes are well approximated.

2.3.3 Convergence of discretizations of diffusions

We now consider the following Stochastic Differential Equation which is a generalization of the one considered in (2.15) and forms a special class of functional auto-regressive models

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t \ , \tag{2.16}$$

where $b \in \mathrm{C}(\mathbb{R}^d, \mathbb{R}^d)$ and $\mathbf{X}_0 \in \mathbb{R}^d$. Similarly to the previous section we consider the Euler-Maruyama discretization of this diffusion process. Let $\gamma > 0$. For any $n \in \mathbb{N}$ let

$$X_{n+1} = X_n + \gamma b(X_n) + \sqrt{\gamma} Z_{n+1} \ , \tag{2.17}$$

where $X_0 = \mathbf{X}_0$ and $(Z_n)_{n\in\mathbb{N}}$ a sequence of d-dimensional independent Gaussian random variables with zero mean and identity covariance matrix. We consider the Markov kernels $(\mathrm{P}_t)_{t\geq 0}$ and $(\mathrm{R}^n_\gamma)_{n\in\mathbb{N}}$ associated with (2.16) and (2.17). In the continuous case we say that $(\mathrm{P}_t)_{t\geq 0}$ is geometrically ergodic with respect to the distance $\mathbf{d}$ if there exist $C \geq 0$ and $\rho \in [0, 1)$ such that for any $x, y \in \mathbb{R}^d$ and $t \geq 0$

$$\mathbf{d}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq C\rho^t \mathbf{d}(\delta_x, \delta_y) \ .$$

In the case where $(\mathrm{R}^n_\gamma)_{n\in\mathbb{N}}$ is geometrically ergodic with respect to the distance $\mathbf{d}$ if there exist $C \geq 0$ and $\rho \in [0, 1)$ such that for any $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$ such that

$$\mathbf{d}(\delta_x \mathrm{R}^n_\gamma, \delta_y \mathrm{R}^n_\gamma) \leq C\rho^{\gamma n} \mathbf{d}(\delta_x, \delta_y) \ .$$

Note that it is also possible to obtain sub-geometric convergence rates under weaker assumptions than the ones we consider in this thesis, see [But14; DFM16; FM03; VK04]. We refer to Section 4.1.1 for a literature review on ergodicity results for Markov chains. We now present some recent works on geometric convergence results with explicit dependency with respect to the parameters of the problem.

We consider the following situation where the vector field is contractive at infinity, *i.e.*there exists $R \geq 0$ such that for any $x, y \in \mathbb{R}^d$ with $\|x-y\| \geq R$, $\langle b(x)-b(y), x-y\rangle \leq -\mathtt{m}^+\|x-y\|^2$ where $\mathtt{m}^+ > 0$. We also assume that b is Lipschitz regular and satisfies some one-sided Lipschitz condition, *i.e.* $\langle b(x)-b(y), x-y\rangle \leq -\mathtt{m}\|x-y\|^2$. In this setting, Eberle and Majka [EM19] show geometric convergence results for the Wasserstein distance associated with cost function $\mathbf{c}_a$ defined for any $x, y \in \mathbb{R}^d$ by

$$\mathbf{c}_a(x,y) = a\mathbb{1}_{\Delta^c}(x,y) + f_a(\|x-y\|) \;,$$

where $a \geq 0$ and f_a explicit in [EM19, Equation (2.53)] and $\Delta = \{(x,x) \;:\; x \in \mathbb{R}^d\}$. More precisely, there exist $\bar{\gamma} > 0$ and $C \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$, we have

$$\mathbf{d}(\delta_x \mathrm{R}_\gamma^n, \delta_y \mathrm{R}_\gamma^n) \leq C\rho_a^{\gamma n}\mathbf{d}(\delta_x, \delta_y) \;,$$

with

$$\log(\log^{-1}(\rho_a^{-1})) \simeq -\mathtt{m}R^2/c_1 \;, \text{ where } c_1 = 16^{-1}\int_{1/4}^{3/8}(1-\mathrm{e}^{u-1/2})\boldsymbol{\varphi}(u)\mathrm{d}u \leq 0.00051 \;,$$

and for any $t \in \mathbb{R}$, $\boldsymbol{\varphi}(t) = (2\pi)^{-1/2}\exp[-t^2/2]$. These results imply the geometric convergence of the Markov chain for the Wasserstein distance $\mathbf{W}_1$ (also called Kantorovitch-Rubinstein distance) and for the total variation TV.

Majka, Mijatović, and Szpruch recently extended these results to $\mathbf{W}_2$. More precisely, under the same conditions as before, the authors obtain the following result: there exist $C \geq 0$ and $\bar{\gamma} > 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$

$$\mathbf{W}_2(\delta_x \mathrm{R}_\gamma^n, \delta_y \mathrm{R}_\gamma^n) \leq C\rho_b^{n\gamma/2}(\|x-y\| + \|x-y\|^{1/2}) \;,$$

with

$$\log(\log^{-1}(\rho_b^{-1})) \simeq \mathtt{L}R^2/(6c_2) \;,$$
$$\text{where } c_2 = 4\min\left(\int_0^{1/2} u^2(1-\mathrm{e}^{u-1/2})\boldsymbol{\varphi}(u)\mathrm{d}u, (1-\mathrm{e}^{-1})\int_0^{1/2} u^3\boldsymbol{\varphi}(u)\mathrm{d}u\right) \leq 0.0072 \;,$$

The results presented in [EM19; MMS18] are based on the discretization of the arguments used by Eberle in [Ebe16]. Doing so, the auhtors obtain similar convergence results for the discrete-time processes associated with the continuous-time processes under study in [Ebe16]. Indeed, if we set $\mathbf{c}_c(x,y) = f_c(\|x-y\|)$ for any $x, y \in \mathbb{R}^d$, where f_b is defined [Ebe16, Equation (2.6)], Eberle [Ebe16] obtain that for any $x, y \in \mathbb{R}^d$ and $t \geq 0$

$$\mathbf{W}_{\mathbf{c}_c}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq \rho_c^t \mathbf{c}_c(x,y) \;,$$

with

$$\log(\log^{-1}(\rho_b^{-1})) \simeq -\mathtt{m}R^2/4 \;.$$

This result implies geometric convergence rate of $(\delta_x \mathrm{P}_t)_{t\geq 0}$ in Wasserstein distance $\mathbf{W}_1$ with rate ρ_b. Luo and Wang [LW16b] extend these convergence results for Wasserstein distances of order p, $\mathbf{W}_p$, with $p \geq 1$. We summarize these results in Table 2.1. This table will be completed in Section 4.1, in which we derive new convergence rates (valid for both discrete-time and continuous-time processes) using minorization and Foster-Lyapunov drift conditions.

Reference	Wasserstein distance	Control distance	(D)	(C)	(TN)
[EM19]	$\|\cdot\|_{\mathrm{TV}}$	$\mathbb{1}_{\Delta^{\mathrm{c}}}(x,y) + \|x-y\|$	✓		7840
	$\mathbf{W}_1$	$\|x-y\|$	✓		4536
[MMS18]	$\mathbf{W}_2$	$\|x-y\| + \|x-y\|^{1/2}$	✓		332
[Ebe16]	$\mathbf{W}_1$	$\|x-y\|$		✓	1
[LW16b]	$\mathbf{W}_p$	$\|x-y\| + \|x-y\|^{1/p}$		✓	$1-\mathtt{m}^+/\mathtt{m}$

Table 2.1: Every line of the table reads as follows. Suppose "Wasserstein distance" reads $\mathbf{W}_{\mathbf{c}_1}$ and "distance bound" reads $\mathbf{c}_2(x,y)$ then: if (D) is checked, there exist $C \geq 0$ and $\rho \in [0,1)$ such that for any $x, y \in \mathbb{R}^d$ and $k \in \mathbb{N}$, $\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq C\rho^{k\gamma}\mathbf{c}_2(x,y)$ for γ small enough. If (C) is checked, there exist $C \geq 0$ and $\rho \in [0,1)$ such that for any $x, y \in \mathbb{R}^d$ and $t \geq 0$, $\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq C\rho^t \mathbf{c}_2(x,y)$. In addition, if the normalized rate "(NR)" reads β, we have $-4\log(\log^{-1}(\rho^{-1}))/(\mathtt{m}R^2) \simeq \beta$ (with $\mathtt{m}$ replaced by $-\mathtt{L}$ in the case of [MMS18]). Note that for the sake of simplicity we omit the dependency with respect to $\bar{\gamma}$ in the present analysis. The exact distances used in papers with which we compare our results are given in [EM19, Equation (2.53)], [MMS18, Equation (2.11)], [Ebe16, Equation (2.6)] and [LW16b, Equation (2.4)].

2.4 Organization and contributions

This thesis is divided into three main parts: an a contrario approach of spatial redundancy, a study of sampling and inference in high dimension with an application to texture synthesis. This separation is mainly thematic: statistical test theory and random fields, Markov chain theory and stochastic optimization, information theory and applications to image processing.

In a first part we present an algorithm based on a contrario methods for the detection of spatial redundancy from a theoretical study of the distribution of similarity function outputs for Gaussian random fields.

Second, we prove ergodicity results for a special class of functional autoregressive models. We then present a stochastic optimization algorithm: *Stochastic Optimization with Unadjusted Langevin* (SOUL) based on the Euler-Maruyama discretization of the Langevin dynamic.

Finally we present an application of these results for the problem of exemplar-based texture synthesis. We obtain a target distribution applying new results on the maximum entropy principle. We then proceed to sample from this distribution applying the SOUL algorithm.

2.4.1 Chapter 3, Section 3.1

In Section 3.1, we introduce a notion of spatial redundancy using similarity functions. These functions are defined on patches and allow for a local evaluation of this similarity. We then distinguish between two cases: the autosimilarity $\mathscr{A}$ (a patch in an image is similar to other patches in the same image) and the template similarity $\mathscr{T}$ (a patch in an image is similar to other patches in another image).

Definition 2.4.1. *Let x and y be two functions defined on $\mathsf{E} \subset \mathbb{R}^2$ or $\mathbb{Z}^2$. Let $\mathsf{w} \subset \mathsf{E}$ be a patch domain. When it is defined we consider the autosimilarity with patch domain w and offset $t \in \mathbb{R}^2$ or $\mathbb{Z}^2$ such that $t + \mathsf{w} \subset \mathsf{E}$ given by*

$$\mathscr{A}(x,t,\mathsf{w}) = s\left(P_{\mathsf{w}}(x), P_{\mathsf{w}}(\tau_{-t}(x))\right) ,$$

where s is a similarity function $s : \mathsf{V}^{\mathsf{w}} \times \mathsf{V}^{\mathsf{w}} \to \mathbb{R}$ (for instance $\|\cdot\|_p$ with $p \geq 1$). Similarly we consider

the template similarity

$$\mathscr{T}(x, y, t, \mathsf{w}) = s\left(P_{\mathsf{w}}(y), P_{\mathsf{w}}(\tau_{-t}(x))\right).$$

If these two notions coincide in a deterministic setting, they are different if we assume that the template image is fixed and that the other images are realizations of a certain random field defined over $\mathbb{Z}^2$ or $\mathbb{R}^2$. In this context, autosimilarity and template similarity are real-valued random variables. As announced in Section 2.1.2, in order to apply the a contrario method to detect spatial redundancy, we need to approximate $\mathscr{A}(U, t, \mathsf{w})$ and $\mathscr{T}(U, x, t, \mathsf{w})$ in the case where x is an image and U is the background model, *i.e.* in the case where U is a Gaussian random field. We begin with limit theorems in the case where the size of patches grows towards infinity and the underlying random field is Gaussian, stationary with long-range independence.

Theorem 2.4.2. *Let $(m_k)_{k\in\mathbb{N}}$ and $(n_k)_{k\in\mathbb{N}}$ be two non-decreasing sequences of $\mathbb{N}$ and $(\mathsf{w}_k)_{k\in\mathbb{N}}$ such that for any $k \in \mathbb{N}$, $\mathsf{w}_k = [\![0, m_k]\!] \times [\![0, n_k]\!]$. Let U be a Gaussian random field such that $\mathrm{Cov}[U(p_1)U(p_2)] = 0$ for $\|p_1 - p_2\|_\infty$ large enough. For the following similarity functions:*

- $s(x, y) = \|x - y\|_p$ *with* $p \in [0, +\infty)$,
- $s(x, y) = \|x - y\|_p^p$ *with* $p \in [0, +\infty)$,
- $s(x, y) = -\langle x, y\rangle$,
- $s(x, y) = -\langle x, y\rangle / (\|x\|\,\|y\|)$,

there exist μ, σ^2 non-negative and defined on $\mathbb{Z}^2$, and $(\alpha_k)_{k\in\mathbb{N}}$ a positive sequence such that for any $t \in \mathbb{Z}^2 \setminus \{0\}$ we have

(a) $\lim_{k\to+\infty} \alpha_k^{-1} \mathscr{A}(U, t, \mathsf{w}_k) \underset{a.s}{=} \mu(t)$;

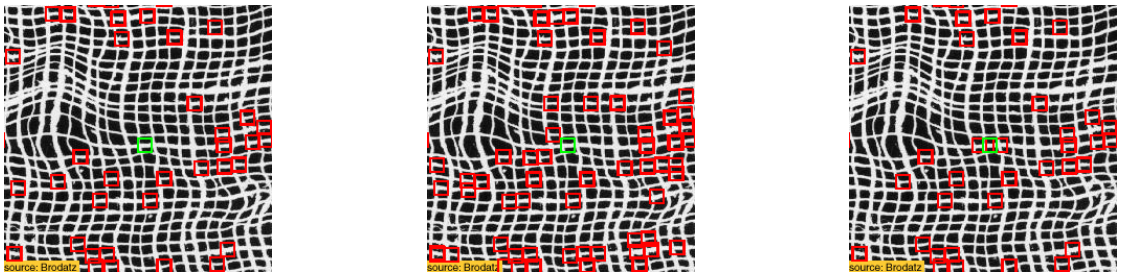
(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(\alpha_k^{-1} \mathscr{A}(U, t, \mathsf{w}_k) - \mu(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \sigma^2(t)\right)$.

This theorem can be extended to the case of template similarity. Note that the Gaussian assumption can be lifted but allows for explicit computations of expectations and variance in the central limit theorem.

In the case where the similarity function is given by the squared ℓ_2 norm, then we obtain a non - asymptotic expression for the autosimilarity which consists in a linear combination of independent $\chi_2(1)$ random variables where the coefficients of this linear combination are given by the eigenvalues of some Toeplitz and symmetric matrix. We then propose an algorithm which efficiently approximates this distribution using a moment method of order 3 as well as an approximation of the eigenvalues:

- we estimate the eigenvalues of the Toeplitz symmetric matrix using the eigenvalues of its projection onto the set of circulant matrices, whose eigenvalues are obtained using the Fourier transform,
- once the coefficients are approximated, we use the WoodF [Woo89] method which is a third order moment method. The approximating distribution is a Fischer-Snedecor distribution.

The precision of these approximation is illustrated in Figure 2.9. We then derive a spatial redundancy detection algorithm based on patches which rely on the a contrario method. This first section is adapted from the paper "Redundancy in Gaussian random fields" accepted at ESAIM: Probability and Statistics.



(a) $1513s, n_d = 52$ (b) $200s, n_d = 116$ (c) $4.77s, n_d = 50$

Figure 2.9: **Spatial redundancy detection.** In each image the green patch corresponds to the target patch in the similarity functions. In (a) we use the a contrario approach with a Gaussian random field background model in order to detect the redundancy. No approximation is made in this case and the computing time is extensive. In (b) we approximate the eigenvalues of the Toeplitz matrix using the projection method. In (c) we combine the matrix projection with moment method to obtain a fast method.

2.4.2 Chapter 3, Section 3.2

In Section 3.2, we first introduce an a contrario framework for the detection of autosimilarity in natural images. We then give some applications of the algorithm described in Section 3.1. In order to use a contrario methods it is necessary to obtain estimates of the type $\mathbb{P}(\mathscr{A}(U, t, \mathsf{w}) \leq \alpha)$ where $\mathscr{A}$ is the autosimilarity function, U the random field (the background model of the a contrario method) and t is some offset which allows for the comparison between the patch with support w and the patch with support $\mathsf{w} + t$ (where the sum should be understood in the Minkowski sense). This probability can be efficiently estimated using the algorithm of Section 3.1. Once the similarity functions are defined and the statistical testing framework is set we exploit the notion of spatial redundancy in the context of image processing. First we provide a modification of the celebrated *Non-Local Means*, NL-means, algorithm. We replace the threshold (or variance) parameters in the original algorithm by a number of false alarms parameter. This new parameter is robust with respect to the choice of image and noise level. We show how this simple modification provides better results than a classical NL-means algorithm and prove that, using the a contrario framework, we can obtain estimates on the reconstruction of the image in probability. The second application of our algorithm is periodicity detection.

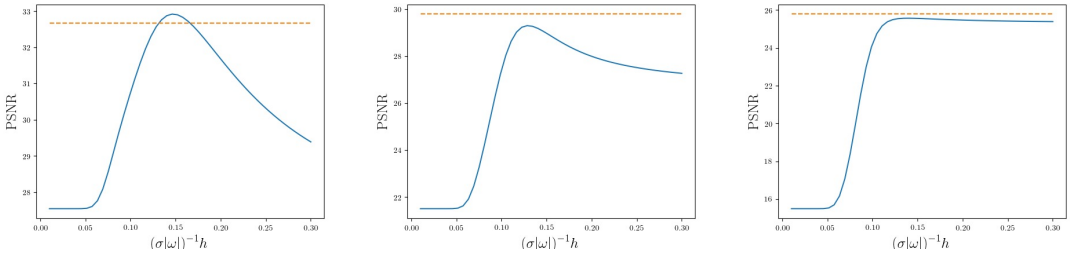
Given an image using the algorithm described in Section 3.1, we obtain a binary image such that a pixel has a value of 1 if the patch centered around this position is detected and 0 otherwise. We then consider the graph associated with this binary map as well as a lattice hypothesis. We note $V \in \mathbb{R}^{p\times 2}$ the edges of the graph, where $p \in \mathbb{N}$ is the number of edges, and consider the following model.

Definition 2.4.3. *Let V be a random variable taking values in $\mathbb{R}^{p\times 2}$ with $p \in \mathbb{N}$. We say that V satisfies the approximate lattice hypothesis if there exist a basis $B = (b_1, b_2)$ and $\sigma > 0$ such that for any $\ell \in \{1, \dots, p\}$, there exists $(m_\ell, n_\ell) \in \mathbb{Z}^2$ such that*

$$\mathcal{L}(V_\ell) = \mathcal{L}(m_\ell b_1 + n_\ell b_2 + \sigma Z_\ell) \ ,$$

where $(Z_\ell)_{\ell\in\{1,\dots,p\}}$ is a family of independent centered Gaussian variables with identity covariance matrix. We note $M = (m_\ell, n_\ell)_{\ell\in\{1,\dots,p\}}$.

Using classical inference tools such as the Maximum A Posteriori we are able to identify the basis



(a) $\sigma = 10$ (b) $\sigma = 20$ (c) $\sigma = 40$

Figure 2.10: **Influence of the parameters on the** PSNR**.** In this figure we present, for 3 different noise levels, the evolution of the PSNR for different values of the filtering parameter in the original NL-means algorithm [DAG10]. The orange line corresponds to the PSNR we obtain using our algorithm, which is a modification of the original NL-means algorithm.

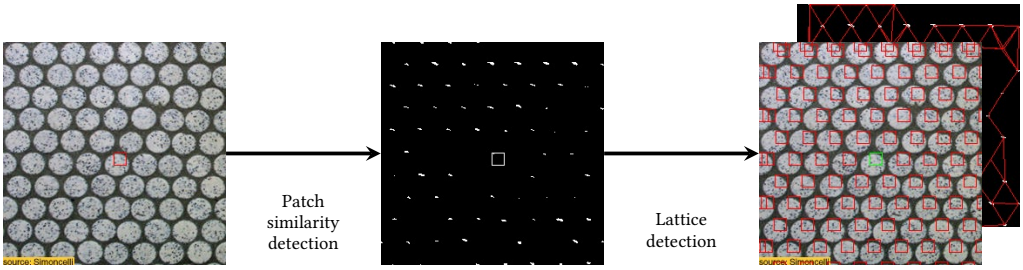


Figure 2.11: **Lattice estimation.** The first step corresponds to the detection of redundant patches as illustrated in Figure 2.9. The second part of the algorithm corresponds to the estimation of the basis of the lattice using a MAP algorithm, see Theorem 2.4.4. In order to illustrate the obtained result, we overlay a red square of the size of the original patch on each detected point of the lattice.

vectors of the underlying module. These vectors fully characterize the underlying lattice (if it exists). The full algorithm is described in Figure 2.11.

Theorem 2.4.4. *For any $\sigma > 0$, $(B_n)_{n\in\mathbb{N}}$ and $(M_n)_{n\in\mathbb{N}}$ converge in a finite number of iterations.*

This algorithm is successfully applied on cristallography images and also permits to classify texture images based on their degree of periodicity. This section is adapted from "Patch redundancy in images: a statistical testing framework and some applications" published in SIAM Imaging Sciences.

2.4.3 Chapter 4, Section 4.1

After having proposed a definition for the spatial redundancy and a statistical framework to identify it in natural images we turn to the problem of sampling images which exhibit such redundancy as well as complex structures. In order to do so we must develop algorithms whose complexity does not explode with the dimension even in a non-convex case. We conduct this study for a specific case of functional auto-regressive models in Section 4.1. The goal of this section is to study Markov chains of the form

$$X_{n+1} = X_n + \gamma b(X_n) + \sqrt{\gamma} Z_{n+1} \ . \tag{2.18}$$

Our main result is the following.

Theorem 2.4.5. *Assume that there exist* $\mathtt{m} \in \mathbb{R}$, $\mathtt{m}^+ > 0$ *and* $\mathtt{L}, R \geq 0$ *such that for any* $x, y \in \mathbb{R}^d$

$$\|b(x) - b(y)\| \leq \mathtt{L} \|x - y\| \ , \qquad \langle b(x) - b(y), x - y \rangle \leq -\mathtt{m} \|x - y\|^2 \ ,$$

and if $\|x - y\| \geq R$,

$$\langle b(x) - b(y), x - y \rangle \leq -\mathtt{m}^+ \|x - y\|^2 \ .$$

Then, there exist $\bar{\gamma} > 0$, $D_{\bar{\gamma},1}, D_{\bar{\gamma},2}, E_{\bar{\gamma}} \geq 0$ *and* $\lambda_{\bar{\gamma}}, \rho_{\bar{\gamma}} \in [0, 1)$ *with* $\lambda_{\bar{\gamma}} \leq \rho_{\bar{\gamma}}$, *such that* $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \lambda_{\bar{\gamma}}^{k\gamma/4} [D_{\bar{\gamma},1} \mathbf{c}(x, y) + D_{\bar{\gamma},2} \mathbb{1}_{\Delta^c}(x, y)] + E_{\bar{\gamma}} \rho_{\bar{\gamma}}^{k\gamma/4} \mathbb{1}_{\Delta^c}(x, y) \ ,$$

where $\mathbf{c}(x, y) = \mathbb{1}_{\Delta^c}(x, y)(1 + \|x - y\| / R)$, $\Delta = \{(x, x) \ : \ x \in \mathbb{R}^d\}$ *and* R_γ *is the Markov kernel associated with* (2.18).

In addition, we obtain that $\log(\log(\rho_{\bar{\gamma}}^{-1}))$ is of order $\mathtt{m} R^2 / (4(1 - \mathrm{e}^{\mathtt{m} R^2}))$ and does not depend on the dimension d. This result improves convergence results recently obtained [EM19]. As a corollary we obtain the geometric ergodicity of ULA but also the ergodicity of the projected algorithm associated with the following recusion: $X_0 \in \mathbb{R}^d$ an for any $n \in \mathbb{N}$,

$$X_{n+1} = \Pi_{\mathsf{K}}[X_n + \gamma b(X_n) + \sqrt{\gamma} Z_{n+1}] \ , \tag{2.19}$$

where Π_{K} is the orthogonal projection onto the compact and convex set K. Considering (2.19), a sequence of compact and convex sets $(\mathsf{K}_n)_{n \in \mathbb{N}}$ such that $\bigcup_{n \in \mathbb{N}} \mathsf{K}_n$ and a sequence of stepsizes $(\gamma_n)_{n \in \mathbb{N}}$ such that $\lim_n \gamma_n = 0$, we extend the results we obtain in the discrete-time setting to the case continuous dynamic. More precisely, we obtain that, under the same conditions as the ones derived in the discrete-time setting, for any $x, y \in \mathbb{R}^d$ and $t \geq 0$

$$\mathbf{d}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq C \rho^t \mathbf{d}(\delta_x, \delta_y) \ ,$$

with $C \geq 0$, $\rho \in [0, 1)$ and $\mathbf{d}$ is either the total variation distance or the Wasserstein distance of order 1, $\mathbf{W}_1$. This result extends the convergence rates obtained for the Wasserstein distance of order 1, $\mathbf{W}_1$, to the total variation norm [Ebe16]. We complete Table 2.1 with the consequences of Theorem 2.4.5. Note that contrary to most of the recent approaches [EM19; EGZ18; EGZ19; Ebe16; MMS18; LW16b; Che+18] which rely on the use of an appropriate metric, our proof only uses classical tools for the study of Markov chains (minorization conditions and Foster-Lyapunov drift conditions). This work is adapted from "Convergence of diffusion and their discretizations: from continuous to discrete processes and back" submitted to the Annals of the Henri Poincaré Institute.

2.4.4 Chapter 4, Section 4.2

Once we have derived sharp convergence estimates for ULA and its variants, we turn to the task of stochastic optimization and propose an algorithm with ULA as an intermediate sampling step. Indeed in Section 4.2 we consider the following problem: given a differentiable objective function such that for any $\theta \in \mathbb{R}^p$, $\nabla f(\theta) = \int_{\mathbb{R}^d} H_\theta(x) \mathrm{d}\pi_\theta(x)$, how to find a point $\theta^\star$ which minimizes this function? This study is motivated by applications in the context of regularization parameters estimation (denoising, deblurring and spectral unmixing) and in the context of parameter estimation in maximum entropy models as we will see in the next section. We propose a stochastic gradient descent algorithm in which the gradient is estimated by a Monte-Carlo Markov chain at each iteration. If the underlying Markov

Reference	Wasserstein distance	Control distance	(D)	(C)	(TN)
[EM19]	$\|\cdot\|_{\mathrm{TV}}$	$\mathbb{1}_{\Delta^c}(x,y) + \|x-y\|$	✓		7840
	$\mathbf{W}_1$	$\|x-y\|$	✓		4536
[MMS18]	$\mathbf{W}_2$	$\|x-y\| + \|x-y\|^{1/2}$	✓		332
[Ebe16]	$\mathbf{W}_1$	$\|x-y\|$		✓	1
[LW16b]	$\mathbf{W}_p$	$\|x-y\| + \|x-y\|^{1/p}$		✓	$1-\mathtt{m}^+/\mathtt{m}$
This thesis	$\|\cdot\|_{\mathrm{TV}}$	$\mathbb{1}_{\Delta^c}(x,y) + \|x-y\|$	✓	✓	$(1-\mathrm{e}^{2\mathtt{m}R^2})^{-1}$
	$\mathbf{W}_1$	$\|x-y\|$	✓	✓	idem
	$\mathbf{W}_p$	$\|x-y\| + \|x-y\|^{1/\alpha}$	✓	✓	idem

chain is ULA then the algorithm is called SOUL (Stochastic Optimization with Unadjusted Langevin). In fact, this optimization procedure can be extended to any ergodic Markov kernel. In this case, the algorithm is called SOUK (Stochastic Optimization with Unadjusted Kernel). Using recent results from the optimization literature [AFM17], we obtain explicit bounds (both almost surely and in expectation) with regards to the distance between $(f(\bar{\theta}_n))_{n\in\mathbb{N}}$ and $\min_{\mathbb{R}^p} f$, in the case where f is convex and for any $n \in \mathbb{N}$

$$\bar{\theta}_n = \sum_{k=1}^{n} \delta_k \theta_k \Big/ \sum_{k=1}^{n} \delta_k \ ,$$

and $(\delta_k)_{k\in\mathbb{N}}$ is the sequence of stepsizes used in the stochastic gradient algorithm.

Algorithm 2 Stochastic Optimization via Unadjusted Langevin (SOUL)

Inputs:
$\theta_0 \in \mathsf{K}$, $X_0^0 \in \mathbb{R}^d$, $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$, $(m_n)_{n\in\mathbb{N}}$, N
for $n \in \{1, \dots, N-1\}$ **do**
  if $n \geq 1$ **then**
    $X_0^n \leftarrow X_{m_{n-1}}^{n-1}$
  end if
  for $k \in \{0, \dots, m_n - 1\}$ **do**
    $Z_{k+1}^n \leftarrow$ sample $\mathrm{N}(0, \mathrm{Id})$
    $X_{k+1}^n \leftarrow X_k^n - \gamma_n \nabla_x \log \pi_\theta(X_k^n) + \sqrt{2\gamma_n} Z_{k+1}^n$
  end for
  $\Delta_{\theta_n} \leftarrow \frac{1}{m_n} \sum_{k=1}^{m_n} H_{\theta_n}(X_k^n)$
  $\theta_{n+1} \leftarrow \Pi_{\mathsf{K}}[\theta_n + \delta_{n+1} \Delta_{\theta_n}]$
end for
Outputs:
$\hat{\theta}_N = \left\{\sum_{n=1}^N \delta_n \theta_n\right\} \Big/ \left\{\sum_{n=1}^N \delta_n\right\}$

Theorem 2.4.6. *Assume that $f : \mathbb{R}^p \to \mathbb{R}$ is differentiable, convex and if there exists $L \geq 0$ such that for any $\theta_1, \theta_2 \in \mathbb{R}^p$,*

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq L \|\theta_1 - \theta_2\| \ .$$

Let $(\delta_n)_{n\in\mathbb{N}}$ and $(\gamma_n)_{n\in\mathbb{N}}$ be two decreasing sequences of non-negative real numbers and $(m_n)_{n\in\mathbb{N}}$ be a sequence of integers such that

$$\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty \ , \qquad \sum_{n=0}^{+\infty} \delta_{n+1} \gamma_n^{1/2} < +\infty \ , \qquad \sum_{n=0}^{+\infty} \delta_{n+1}/(m_n \gamma_n) < +\infty \ .$$

Let $\{(X_k^n)_{k\in\{0,\dots,m_n\}} : n \in \mathbb{N}\}$ *and* $(\theta_n)_{n\in\mathbb{N}}$ *be given by Algorithm 2. Assume that the conditions of Theorem 2.4.5 are satisfied for* $b = \nabla \log \pi_\theta$ *uniformly in* $\theta \in \mathsf{K}$*. Then for* $\sup_{n\in\mathbb{N}}(\delta_n + \gamma_n)$ *small enough we have*

(a) $(\theta_n)_{n\in\mathbb{N}}$ *converges a.s. towards* $\theta^\star \in \arg\min_{\mathsf{K}} f$ *;*

(b) In addition, there exists $C \geq 0$ *a.s. such that for any* $n \in \mathbb{N}^\star$

$$\left\{ \sum_{k=1}^n \delta_k f(\theta_k) \middle/ \sum_{k=1}^n \delta_k \right\} - \min_{\mathsf{K}} f \leq C \middle/ \left(\sum_{k=1}^n \delta_k \right) .$$

This theorem describes the asymptotic behavior of $(\hat{\theta}_N)_{N\in\mathbb{N}}$ in Algorithm 2. We can also provide non-asymptotic results for $\mathbb{E}[\sum_{k=1}^n \delta_k f(\theta_k)/\sum_{k=1}^n \delta_k]$. These results generalize and extend the ones obtained in [AFM17].

This section is adapted from "Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. Application to maximum marginal likelihood and empirical Bayesian estimation" submitted at Statistics and Computing. An extension of this algorithm to the case where f is no longer differentiable and applications for the estimation of the regularization parameters in Empirical Bayes models can be found in "Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach" submitted at SIAM Imaging Sciences.

2.4.5 Chapter 5, Section 5.1

In Section 5.1, we present an application of the maximum entropy principle for exemplar-based texture synthesis. We extend existing results on the existence and uniqueness results on maximum entropy models in the case where the space is no longer finite. In particular, we give explicit conditions on the constraint functions of the model and on the reference probability measure for the maximum entropy distribution to exist. If the maximum entropy model exists, its distribution $\pi^\star$ is given by $(\mathrm{d}\pi^\star/\mathrm{d}\mu)(x) \propto \exp[-\langle\theta^\star, F(x)\rangle]$ where μ is a reference probability measure, $\theta^\star \in \mathbb{R}^p$ and $F : \mathbb{R}^d \to \mathbb{R}^p$ is a constraint.

Proposition 2.4.7. *Assume that there exist* $\alpha \geq 0$, $\alpha' > \alpha$ *such that*

- F *is continuous and there exists* $C_\alpha \geq 0$ *such that* $\sup_{x\in\mathbb{R}^d}\{\|F(x)\|(1+\|x\|^\alpha)^{-1}\} \leq C_\alpha < +\infty$,
- *There exists* $\eta > 0$ *such that* $\int_{\mathbb{R}^d} \exp[\eta\|x\|^{\alpha'}]\mathrm{d}\mu(x) < +\infty$.

Then, the following results hold:

(a) If for any $\theta \in \mathbb{R}^p$ *with* $\|\theta\| = 1$ *we have* $\mu(\{x \in \mathbb{R}^d : \langle F(x), \theta\rangle < 0\}) > 0$*, then the solution of the maximum entropy problem is given by* $\pi^\star$ *where for any* $x \in \mathbb{R}^d$

$$(\mathrm{d}\pi^\star/\mathrm{d}\mu)(x) = \exp\left[-\langle\theta^\star, F(x)\rangle\right] / \int_{\mathbb{R}^d} \exp\left[-\langle\theta^\star, F(y)\rangle\right] \mathrm{d}\mu(y) ,$$

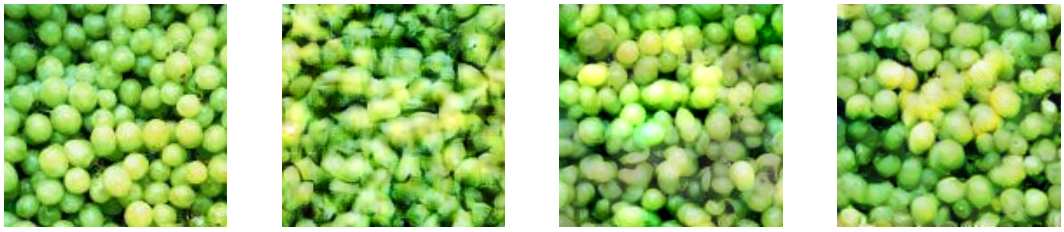
and where $\theta^\star$ *is the solution to the following convex optimization problem.*

$$\min_{\theta\in\mathbb{R}^p} \left\{ \int_{\mathbb{R}^d} \exp[-\langle\theta, F(x)\rangle]\mathrm{d}\mu(x) \right\} .$$

(b) In particular Proposition 2.4.7-(a) is satisfied if $\mu(\mathsf{A}) > 0$ *for all* $\mathsf{A} \subset \mathbb{R}^d$ *open, and if* F *is continuous and there exists* $x \in F^{-1}(\{0\})$ *such that* F *is differentiable with respect to* x *and* $\det(\mathrm{D}F(x)\mathrm{D}F(x)^\top) > 0$.

(a) (b) (c) (d)

Figure 2.12: **Some examples of texture synthesis.** The original images are presented in (b) and (d) and the synthesis results in (a) and (c).



(a) (b) (c) (d)

Figure 2.13: **Some texture synthesis algorithms.** In (a) we present the exemplar texture. In (b) we show the obtained synthesis using [PS00]. In (c) we use the method of [GEB15] in order to generate the texture image. In (d) we present our results.

In this case we show that the infinite-dimensional convex optimization problem can be reframed as a finite-dimensional convex optimization problem using a generalization of Lagrangian duality. We obtain that $\theta^\star$ is such that $L(\theta^\star)$ is the minimum of the log-partition function $L : \mathbb{R}^p \to \mathbb{R}$ given for any $\theta \in \mathbb{R}^p$ by $L(\theta) = \int_{\mathbb{R}^d} \exp[-\langle \theta, F(x)\rangle] \mathrm{d}\mu(x)$. The necessary condition for the existence of a maximum entropy model is easier to check than the one given in [Csi75]. If the constraints are given by a neural network then we provide a certificate to check that such a model exists.

2.4.6 Chapter 5, Section 5.2

We can then apply the SOUL algorithm to the log-partition in order to approximate $\theta^\star$. Note that the first formulation of the maximum entropy texture-synthesis heuristics can be found in the work of Zhu, Lu and Mumford [ZWM98] and that the use of the Langevin dynamics in this case was first suggested in [LZW16]. Finally, in Section 5.2 we present our results in the case of texture synthesis. More precisely, we establish links between our approach and the microcanonical models recently proposed by [BM18], see Proposition 2.4.8.

Proposition 2.4.8. *Assume that there exists $\eta > 0$ such that $\int_{\mathbb{R}^d} \exp[\eta \|x\|^2] \mathrm{d}\mu(x) < +\infty$ and F are given by a neural network which satisfies some coercivity condition, see Proposition 5.2.3 for a precise statement. Then, there exists $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0)$, the maximum entropy model π_ε exists. In addition, if $\mu(F^{-1}(\{0\})) > 0$, then $\lim_{\varepsilon \to 0} \pi_\varepsilon = \pi_\infty$, then for any $x \in \mathbb{R}^d$, $(\mathrm{d}\pi_\infty / \mathrm{d}\mu)(x) = \mathbb{1}_{F^{-1}(\{0\})}(x) / \mu(F^{-1}(\{0\}))$.*

(a) (b) $n = 100000$ (c)

Figure 2.14: **Convergence of the parameters.** In (a) we display the normalized errror between $(\theta_n)_{n\in\mathbb{N}}$ (blue) and the optimal weights as well as the error between $(\hat{\theta}_n)_{n\in\mathbb{N}}$ (red) and the optimal weights. In (b) we present a sample from the model for $n = 100000$ and in (c) we present the exemplar texture.

We check the validity of the SOUL algorithm in two settings:

- Gaussian texture synthesis,
- structured texture synthesis.

In the case of the Gaussian texture synthesis the constraints functions are given by the autocorrelation function. Therefore, we are looking for texture with the same autocorrelation as some target texture. We show that this model is Gaussian, compute its covariance matrix and show that we indeed estimate correctly this covariance matrix, see Figure 2.14.

We then turn to the more challenging case where the constraints are given by a pretrained convolutional neural network for exemplar-based texture synthesis. We obtain results which are comparable to the state-of-the-art methods, see Figure 2.13. In Figure 2.12 we complete Figure 5.1. We study the innovation capacity of our algorithm as well as some variants and we present an extension of this algorithm for style transfer. This work is adapted from "Maximum entropy methods for texture synthesis: theory and practice" submitted at SIMODS and "Macrocanonical models for texture synthesis" presented at Scale Space and Variational Methods in Computer Vision.

2.4.7 Publications of this thesis

Papers published in international journals

1) V. De Bortoli, A. Desolneux, B. Galerne, A. Leclaire. *Patch redundancy in images: a statistical testing framework and some applications.* 2019 - Published in SIAM Journal on Imaging Sciences (SIIMS). See Chapter 3, Section 3.2.

2) V. De Bortoli, A. Desolneux, B. Galerne, A. Leclaire. *Redundancy in Gaussian random fields.* 2018 - Accepted in ESAIM Probability and Statistics. See Chapter 3, Section 3.1.

Preprints and submitted papers

1) V. De Bortoli, A. Durmus. *Convergence of diffusion and their discretizations: from continuous to discrete processes and back.* 2019 - Submitted. See Chapter 4, Section 4.1.

2) V. De Bortoli, A. Durmus, M. Pereyra, A. F.Vidal. *Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. Application to maximum marginal likelihood and empirical Bayesian estimation.* 2019 - Submitted. See Chapter 4, Section 4.2.

3) V. De Bortoli, A. Desolneux, A. Durmus, B. Galerne, A. Leclaire. *Maximum entropy methods for texture synthesis: theory and practice.* 2019 - Submitted. See Chapter 5;

International conference publications

1) V. De Bortoli, A. Desolneux, B. Galerne, A. Leclaire. *Macrocanonical models for texture synthesis.* 2019 - Presented at Scale Space and Variational Methods in Computer Vision (SSVM). Short version of "Maximum entropy methods for texture synthesis: theory and practice"

Notation

Let A, B and C three sets with $\mathsf{C} \subset \mathsf{B}$ and $f : \mathsf{A} \to \mathsf{B}$, we set $f^{\leftarrow}(\mathsf{C}) = \{x \in \mathsf{A} : f(x) \in \mathsf{C}\}$. For any $\mathsf{A} \subset \mathsf{B}$ and $f : \mathsf{B} \to \mathsf{C}$ we denote $f|_{\mathsf{A}}$ the restriction of f to A. Let $d \in \mathbb{N}^\star$ and $\langle \cdot, \cdot \rangle$ be a scalar product over $\mathbb{R}^d$, and $\| \cdot \|$ be the corresponding norm. The complement of a set $\mathsf{A} \subset \mathbb{R}^d$, is denoted by A^{c}. For any $\mathsf{A} \in \mathbb{R}^d$, we denote $\mathrm{int}(\mathsf{A})$ its interior, $\bar{\mathsf{A}}$ its closure, $\partial \mathsf{A} = \bar{\mathsf{A}} \cap \mathrm{int}(\mathsf{A})^{\mathrm{c}}$ its boundary. Let $\mathsf{A} \subset \mathbb{R}^d$ and $R \geq 0$, we denote $\mathrm{diam}(\mathsf{A}) = \sup_{(x,y) \in \mathsf{A}^2} \|x - y\|$ and $\Delta_{\mathsf{A},R} = \{(x, y) \in \mathsf{A}^2 : \|x - y\| \leq R\} \subset \mathbb{R}^{2d}$ and $\Delta_{\mathsf{A}} = \Delta_{\mathsf{A},0} = \{(x, x) : x \in \mathsf{A}\}$.

$\mathcal{B}(\mathbb{R}^d)$ denotes the Borel σ-field of $\mathbb{R}^d$, $\mathbb{F}(\mathbb{R}^d, \mathbb{R}^p)$ the set of all $\mathbb{R}^p$-valued Borel measurable functions on $\mathbb{R}^d$. If $p = 1$, we write $\mathbb{F}(\mathbb{R}^d, \mathbb{R}^p) = \mathbb{F}(\mathbb{R}^d)$ and define for $f \in \mathbb{F}(\mathbb{R}^d)$,

$$\|f\|_\infty = \inf\{t \geq 0 : \lambda(\{x \in \mathbb{R}^d : |f(x)| > t\}) = 0\} ,$$

where λ is the Lebesgue measure over $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. An open ball of $\mathbb{R}^d$ for the Euclidean distance with center $x_0 \in \mathbb{R}^d$ and radius $r > 0$ is denoted $\mathrm{B}(x_0, \mathbb{R}^d)$.

For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $f \in \mathbb{F}(\mathbb{R}^d, \mathbb{R}^p)$, a μ-integrable function, denote by $\mu(f)$ the integral of f with respect to (w.r.t.) μ, *i.e.* $\mu(f) = \int_{\mathbb{R}^d} f(x) \mathrm{d}\mu(x)$. If $f = \mathbb{1}_{\mathsf{A}}$ for some measurable set A then we denote $\mu(\mathbb{1}_{\mathsf{A}}) = \mu(\mathsf{A})$. If μ is the Lebesgue measure then we denote $\mathrm{Vol}(\mathsf{A}) = \lambda(\mathsf{A})$. Let $f \in \mathbb{F}(\mathbb{R}^d)$ then for any probability measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ we denote by $f_\sharp \mu$ the pushforward measure of μ by f.

Let U be an open set of $\mathbb{R}^d$. We denote by $\mathrm{C}^k(\mathsf{U}, \mathbb{R}^p)$ the set of $\mathbb{R}^p$-valued k-continuously differentiable functions. The differential of $f \in \mathrm{C}^k(\mathsf{U}, \mathbb{R}^p)$ is denoted $\mathrm{d}f$ and its Jacobian matrix $\mathrm{D}f$. Let $\mathrm{C}^k(\mathsf{U})$ stand for $\mathrm{C}^k(\mathsf{U}, \mathbb{R})$. Let $f : \mathsf{U} \to \mathbb{R}$, we denote by ∇f, the gradient of f if it exists. f is said to be $\mathtt{m}$-strongly convex with $\mathtt{m} \geq 0$ if for all $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$,

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y) - (\mathtt{m}/2)t(1 - t) \|x - y\|^2 .$$

We recall that if $f : \mathsf{U} \to \mathbb{R}$ is twice differentiable at point $a \in \mathbb{R}^d$, its Laplacian is given by $\Delta f(a) = \sum_{i=1}^d \partial^2 f(a)/\partial x_i^2$.

For any $\alpha > 0$, let $\mathcal{P}_\alpha$ be the set of probability measures over $\mathcal{B}(\mathbb{R}^d)$ such that $\int_{\mathbb{R}^d} \|x\|^\alpha \mathrm{d}\pi(x) < +\infty$. Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space, and

$$\mathrm{L}^2(\Omega, \mathcal{G}) = \{X : X \text{ is a real-valued random variable on } \Omega \text{ such } \mathbb{E}[X^2] < +\infty\} .$$

Let μ, ν be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We write $\mu \ll \nu$ if μ is absolutely continuous w.r.t. ν and $\mathrm{d}\mu/\mathrm{d}\nu$ an associated density.

Let $\mathbb{M}(\mathcal{X})$ be the set of finite signed measures over $(\mathsf{X}, \mathcal{X})$ and $\mu \in \mathbb{M}(\mathcal{X})$. Let $V \in \mathbb{F}(\mathbb{R}^d, [1, +\infty))$. We define the V-norm for any $f \in \mathbb{F}(\mathsf{X}, \mathbb{R})$ and the V-total variation norm for any $\mu \in \mathbb{M}(\mathcal{X})$ as follows

$$\|f\|_V = \|f/V\|_\infty , \qquad \|\mu\|_V = (1/2) \sup_{f \in \mathbb{F}(\mathsf{X}, \mathbb{R}), \|f\|_V \leq 1} \left| \int_{\mathbb{R}^d} f(x) \mathrm{d}\mu(x) \right| .$$

In the case where $V = 1$ this norm is called the total variation norm of μ. Let μ, ν be two probability measures over $\mathcal{X}$, *i.e.* two elements of $\mathbb{M}(\mathcal{X})$ such that $\mu(\mathsf{X}) = \nu(\mathsf{X}) = 1$. A probability measure ζ over $\mathcal{X}^{\otimes 2}$ is said to be a transference plan between μ and ν if for any $\mathsf{A} \in \mathcal{X}$, $\zeta(\mathsf{A} \times \mathcal{X}) = \mu(\mathsf{A})$ and $\zeta(\mathcal{X} \times \mathsf{A}) = \nu(\mathsf{A})$. We denote by $\mathbf{T}(\mu, \nu)$ the set of all transference plans between μ and ν. Let $\mathbf{c} \in \mathbb{F}(\mathsf{X} \times \mathsf{X}, [0, +\infty))$. We define the Wasserstein metric/distance $\mathbf{W}_{\mathbf{c}}(\mu, \nu)$ between μ and ν by

$$\mathbf{W}_{\mathbf{c}}(\mu, \nu) = \inf_{\zeta \in \mathbf{T}(\mu,\nu)} \int_{\mathsf{X}^2} \mathbf{c}(x, y) \mathrm{d}\zeta(x, y) \ .$$

Note that the term Wasserstein metric/distance is an abuse of terminology since $\mathbf{W}_{\mathbf{c}}$ is only a real metric on a subspace of probability measures on X under additional conditions on $\mathbf{c}$, *e.g.* if $\mathbf{c}$ is a metric on $\mathbb{R}^d$, see [Vil09, Definition 6.1]. If $\mathbf{c}(x, y) = \|x - y\|^p$ for $p \geq 1$, the Wasserstein distance of order p is defined by $\mathbf{W}_p = \mathbf{W}_{\mathbf{c}}^{1/p}$. Assume that $\mathbf{c}(x, y) = \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x, y) W(x, y)$ with $W \in \mathbb{F}(\mathsf{X} \times \mathsf{X}, [0, +\infty))$ such that W is symmetric, satisfies the triangle inequality, *i.e.* for any $x, y, z \in \mathsf{X}$, $W(x, z) \leq W(x, y) + W(y, z)$, and for any $x, y \in \mathsf{X}$, $W(x, y) = 0$ implies $x = y$. Then $\mathbf{c}$ is a metric over X^2 and the associated Wasserstein cost, denoted by $\mathbf{W}_{\mathbf{c}}$, is an extended metric. Note that if $W(x, y) = \{V(x) + V(y)\}/2$ then $\mathbf{W}_{\mathbf{c}}(\mu, \nu) = \|\mu - \nu\|_V$, see [Dou+18, Theorem 19.1.7].

The Kullback-Leibler divergence, or relative entropy, of μ from ν is defined by

$$\mathrm{KL}\left(\mu|\nu\right) = \begin{cases} \int_{\mathbb{R}^d} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\nu(x) & \text{if } \mu \ll \nu \ , \\ +\infty & \text{otherwise} \ . \end{cases}$$

If μ and ν are probability measures, the relative entropy takes values in $[0, +\infty]$.

Let $\mathcal{Z}$ be a σ-field. We say that $\mathrm{P} : \mathsf{X} \times \mathcal{Z} \to [0, +\infty)$ is a Markov kernel if for any $x \in \mathsf{X}$, $\mathrm{P}(x, \cdot)$ is a probability measure over $\mathcal{Z}$ and for any $\mathsf{A} \in \mathcal{Z}$, $\mathrm{P}(\cdot, \mathsf{A}) \in \mathbb{F}(\mathsf{X}, [0, +\infty))$. Let $\mathsf{Y} \in \mathcal{B}(\mathbb{R}^d)$ be equipped with $\mathcal{Y}$ the trace of $\mathcal{B}(\mathbb{R}^d)$ over Y, $\mathrm{P} : \mathsf{X} \times \mathcal{Z}$ and $\mathrm{Q} : \mathsf{Y} \times \mathcal{Z}$ be two Markov kernels. We say that $\mathrm{K} : \mathsf{X} \times \mathsf{Y} \to \mathcal{Z}^{\otimes 2}$ is a Markov coupling kernel if for any $(x, y) \in \mathsf{X} \times \mathsf{Y}$, $\mathrm{K}((x, y), \cdot)$ is a transference plan between $\mathrm{P}(x, \cdot)$ and $\mathrm{Q}(y, \cdot)$.

We take the convention that $\prod_{k=p}^{n} = 1$ and $\sum_{k=p}^{n} = 0$ for $n, p \in \mathbb{N}$, $n < p$. If $x, y \in \mathbb{C}^d$ with $d \in \mathbb{N}$ we define the periodic convolution between x and y and denote $z = x * y \in \mathbb{C}^d$, the element z such that for any $i \in \{0, \ldots, d-1\}$, $z(i) = \sum_{k=0}^{d-1} x(k) y(i-k)$, where x and y are extended over $\mathbb{Z}$ by periodicity. We also denote $\check{x} \in \mathbb{C}^d$ such that for any $i \in \{0, \ldots, d-1\}$, $\check{x}(i) = x(-i)$ and x is extended over $\mathbb{Z}$ by periodicity. For any $x \in \mathbb{C}^d$, $\mathcal{F}(x)$ (respectively $\mathcal{F}^{-1}(x) \in \mathbb{C}^d$) stands for the Fourier transform (respectively the inverse Fourier transform), defined for any $j \in \{0, \ldots, d-1\}$ by

$$\mathcal{F}(x)(j) = \sum_{k=0}^{d-1} x(j) \mathrm{e}^{-2\mathrm{i}\pi jk/d} \ , \qquad \mathcal{F}^{-1}(x)(j) = d^{-1} \sum_{k=0}^{d-1} x(j) \mathrm{e}^{2\mathrm{i}\pi jk/d} \ .$$

Note that we have $\mathcal{F}^{-1}(\mathcal{F}(x)) = x$. For any $z \in \mathbb{C}$, we denote by $\Re(z)$ the real part of z and by $\Im(z)$ its imaginary part. We denote by $\mathbb{A}_{n_2, n_1}(\mathbb{R})$ the vector space of affine operators from $\mathbb{R}^{n_1}$ to $\mathbb{R}^{n_2}$ and for any $A \in \mathbb{A}_{n_2, n_1}$, $\tilde{A}$ is the linear part of A. Finally, $\mathrm{S}_d(\mathbb{R})$ is the space of $d \times d$ real symmetric matrices.

Chapter 3

Spatial redundancy and a contrario methods

3.1 Redundancy in Gaussian random fields

3.1.1 Abstract

Stochastic geometry [Chi+13; Bad13; SW08] aims at describing the arrangement of random structures based on the knowledge of the distribution of geometrical elementary patterns (point processes, random closed sets, etc.). When the considered patterns are functions over some topological space, we can study the geometry of the associated random field. For example, centering a kernel function at each point of a Poisson point process gives rise to the notion of shot-noise random field [Dal71; Ric77; Ric44]. We can then study the perimeter or the Euler-Poincaré characteristic of the excursion sets among other properties [BD16; AST10]. In the present work we will focus on the geometrical notion of redundancy of local windows in random fields. We say that a local window in a random field is redundant if it is "similar" to other local windows in the same random field. The similarity of two local windows is defined as the output of some similarity function computed over these local windows. The lower is the output, the more similar the local windows are.

Identifying such spatial redundancy is a fundamental task in the field of image processing. For instance, in the context of denoising, Buades et al. in [BCM05], propose the Non-Local means algorithm in which a noisy patch is replaced by a weighted mean over all similar patches. Other examples can be found in the domains of inpainting [CPT04] and video coding [JJ81]. Spatial redundancy is also of crucial importance in exemplar-based texture synthesis, where we aim at sampling images with the same perceptual properties as an input exemplar texture. If Gaussian random fields [Wij91; GGM11; Lec15; Xia+14] give good visual results for input textures with no, or few, spatial redundancy, they fail when it comes to sampling structured textures (brick walls, fabric with repeated patterns, etc.). In this case, more elaborated models are needed [GEB15; LZW16]. In this work, we derive explicit probability distribution functions for the random variables associated with the output of similarity functions computed on local windows of random fields. The knowledge of such functions allows us to conduct rigorous statistical testing on the spatial redundancy in natural images.

In order to compute these explicit distributions we will consider specific random fields over specific topological spaces. First, the random fields will be defined either over $\mathbb{R}^2$ (or $\mathbb{T}^2$, where $\mathbb{T}^2$ is the 2-

dimensional torus, when considering periodicity assumptions on the field), or over $\mathbb{Z}^2$ (or $(\mathbb{Z}/(M\mathbb{Z}))^2$, with $M \in \mathbb{N}$ when considering periodicity assumptions on the field). Each of these spaces is embedded with its classical topology. The first case is the *continuous setting*, whereas the second one is the *discrete setting*. In image processing, the most common framework is the finite discrete setting. The discrete setting ($\mathbb{Z}^2$) can be used to define asymptotic properties when the size of images grows or when their resolution increases [BM18], whereas continuous settings are needed in specific applications where, for instance, rotation invariant models are required [UTY95]. All the considered random fields will be Gaussian. This assumption will allow us to explicitly derive moments of some similarity functions computed on local windows of the random field. Once again, another reason for this restriction comes from image processing. Indeed, given an input image, we can compute its first and second-order statistics. Sampling from the associated Gaussian random field gives examples of images which preserve the covariance structure but lose the global arrangement of the input image. Investigating redundancy of such fields is a first step towards giving a mathematical description of this lack of structure.

Finding measurements which correspond to the ones of our visual system is a long-standing problem in image processing. It was considered in the early days of texture synthesis and analyzed by Julesz [Yel93; Jul81; Jul62] who formulated the conjecture that textures with similar first-order statistics (first conjecture) or that textures with similar first and second-order statistics (second conjecture) could not be discriminated by the human eye. Even if both conjectures were disproved [DF81], the work of Gatys et al. [GEB15] suggests that second-order statistics of image features are enough to characterize a broad range of textures. To compute features on images we embed them in a higher dimensional space. This operation can be conducted using linear filtering [PS00] or convolutional neural networks [GEB15] for instance. Some recent works examine the response of convolutional neural network to elementary geometrical pattern [New+18], giving insight about the perceptual properties of such a lifting. In the present work, we focus on another embedding given by considering a square neighborhood, called a patch, around each pixel. This embedding, is exploited in many image processing tasks such as inpainting [HS14], denoising [BCM05; LBM13], texture synthesis [EL99; EF01; LB06; RDM16], etc.

In the special case where the similarity functions are given by the ℓ^2 norm, explicit distributions can be inferred even in the non-asymptotic case. Calculating this distribution exactly is demanding since it requires the knowledge of some covariance matrix eigenvalues as well as an efficient method to compute cumulative distribution functions of quadratic forms of Gaussian random variables. We propose an efficient algorithm to approximate this distribution. In [Bor+19], this algorithm is applied to denoising and periodicity detection problems in an *a contrario* framework.

Section 3.1 is organized as follows. We recall basic notions of Gaussian random fields in general settings in Section 3.1.2. Similarity functions to be evaluated on these random fields, as well as their statistical properties, are described in Section 3.1.2. We give the asymptotic properties of these similarity functions in Gaussian random fields in the discrete setting in Section 3.1.3 and in the continuous setting in Section 3.1.3. It is shown in Section 3.1.3 that the Gaussian asymptotic approximation is valid only for large patches. In order to overcome this problem we consider an explicit formulation of the probability distribution function for a particular similarity function: the square ℓ^2 norm. The computations are conducted in the finite discrete case in Section 3.1.4. We also derive an efficient algorithm to compute these probability distribution functions. Similar non-asymptotic expressions are given in the continuous case in Section 3.1.4. Technical proofs and additional results on multidimensional central limit theorems are presented in the Appendices.

3.1.2 Similarity functions and random fields

Gaussian random fields

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Following [Adl81], a random field over a topological space E is defined as a measurable mapping $U : \Omega \to \mathbb{R}^{\mathsf{E}}$. Thus, for all ω in Ω, $U(\omega)$ is a function over E and, for any $\omega \in \Omega$ and $p \in \mathsf{E}$, $U(\omega)(p)$ is a real number. For the sake of clarity we will omit ω in what follows.

We say that a random field U is of order $r > 0$ if for any finite sequence $(p_1, \dots, p_n) \in \mathsf{E}^n$ with $n \in \mathbb{N}$, the vector $V = (U(p_1), \dots, U(p_n))$ satisfy $\mathbb{E}\left[\|V\|_r^r\right] < +\infty$. Assuming that U is a second-order random field, we define the mean function of U, $m : \mathsf{E} \to \mathbb{R}$ as well as its covariance function, $C : \mathsf{E}^2 \to \mathbb{R}$ for any $p_1, p_2 \in \mathsf{E}^2$ by

$$m(p_1) = \mathbb{E}\left[U(p_1)\right] \qquad \text{and} \quad C(p_1, p_2) = \mathbb{E}\left[(U(p_1) - m(p_1))(U(p_2) - m(p_2))\right] .$$

Assuming that E has a group structure, a random field U is said to be stationary if for any finite sequence $(p_1, \dots, p_n) \in \mathsf{E}^n$ with $n \in \mathbb{N}$ and $t \in \mathsf{E}$, the vector $(U(p_1), \dots, U(p_n))$ and $(U(p_1 + t), \dots, U(p_n + t))$ have same distribution. A second-order random field U is said to be stationary in the weak sense if its mean function is constant and if for all $p_1, p_2 \in \mathsf{E}$, $C(p_1, p_2) = C(p_1 - p_2, 0)$. In this case the covariance of U is fully characterized by its auto-covariance function $\Gamma : \mathsf{E} \to \mathbb{R}$ given for any $p \in \mathsf{E}$ by

$$\Gamma(p) = C(p, 0) .$$

A random field U is said to be a Gaussian random field if, for any sequence $(p_1, \dots, p_n) \in \mathsf{E}^n$ with $n \in \mathbb{N}$, the vector $(U(p_1), \dots, U(p_n))$ is a n-dimensional Gaussian random vector. The distribution of a Gaussian random field is entirely characterized by its mean and covariance functions. As a consequence, the notions of stationarity and weak stationarity are the same for Gaussian random fields.

Since the applications we are interested in are image processing tasks, we consider the case where $\mathsf{E} = \mathbb{R}^2$ (in the continuous setting) and $\mathsf{E} = \mathbb{Z}^2$ (in the discrete setting). In Section 3.1.2 we will consider Lebesgue integrals of random fields and thus need integrability condition for U over compact sets. Let $\mathsf{K} = [a, b] \times [c, d]$ be a compact rectangular domain in $\mathbb{R}^2$. Continuity requirements on the function C imply that $\int_K g(p)U(p)\mathrm{d}p$ is well-defined as a quadratic mean limit for real-valued functions g over E such that $\int_{K\times K} g(p_1)g(p_2)C(p_1, p_2)\mathrm{d}p_1\mathrm{d}p_2 < +\infty$, see [Lin13]. However, we are interested in almost sure quantities and thus we want the integral to be defined almost surely over rectangular windows. The existence of a continuous modification of a random field, ensures the almost sure existence of Riemann integrals over rectangular windows. The following assumptions will ensure continuity almost surely, see Lemma 3.1.1 whose proof can be found in [Adl81, Theorem 1.4.1] and [Pot09, Lemma 4.2, Lemma 4.3, Theorem 4.5]. We define $D : \mathsf{E} \times \mathsf{E} \to \mathbb{R}$ such that for any $p_1, p_2 \in \mathsf{E}$

$$D(p_1, p_2) = \mathbb{E}\left[(U(p_1) - U(p_2))^2\right] = C(p_1, p_1) + C(p_2, p_2) - 2C(p_1, p_2) + (m(p_1) - m(p_2))^2 .$$

A1. *U is a second-order random field and there exist $M, \eta, \alpha > 0$ such that for any $p_1 \in \mathsf{E}$ and $p_2 \in \mathrm{B}(p_1, \eta) \cap \mathsf{E}$ with $p_2 \neq p_1$ we have*

$$D(p_1, p_2) \leq M\|p_1 - p_2\|_2^2 |\log(\|p_1 - p_2\|_2)|^{-2-\alpha} .$$

This assumption can be considerably weakened in the case of a stationary Gaussian random field.

A2. *U is a stationary Gaussian random field and there exist $M, \eta, \alpha > 0$ such that for any $p_1 \in \mathsf{E}$ and $p_2 \in \mathrm{B}(p_1, \eta) \cap \mathsf{E}$ with $p_2 \neq p_1$ we have*

$$D(p_1, p_2) \leq M|\log(\|p_1 - p_2\|_2)|^{-1-\alpha} .$$

Lemma 3.1.1. *Assume* **A1** *or* **A2**. *In addition, assume that for any* $p_1 \in \mathsf{E}$, $m(p) = 0$. *Then there exists a modification of* U, *i.e. a random field* $\widetilde{U}$ *such that for any* $p_1 \in \mathsf{E}$, $\mathbb{P}(U(p) = \widetilde{U}(p)) = 1$, *and for any* $\omega \in \Omega$, $\widetilde{U}(\omega)$ *is continuous over* E.

In the rest of the section we always replace U by its continuous modification $\widetilde{U}$. Note that in the discrete case all random fields are continuous with respect to the discrete topology. In Section 3.1.3 and Section 3.1.4, we will assume that U is a stationary Gaussian random field with zero mean. Asymptotic theorems derived in the next section remain true in broader frameworks, however restricting ourselves to stationary Gaussian random fields allows for explicit computations of asymptotic quantities in order to numerically assess the rate of convergence.

Similarity functions

In order to evaluate redundancy in random fields, we first need to derive a criterion for comparing random fields. We introduce similarity functions which take rectangular restrictions of random fields as inputs.

When comparing local windows of random fields (patches), two cases can occur. We can compare a patch with a patch extracted from the same image. We call this situation *internal matching*. Applications can be found in denoising [BCM05] or inpainting [CPT04] where the information of the image itself is used to perform the image processing task. On the other hand, we can compare a patch with a patch extracted from another image. We call this situation *template matching*. An application can be found in the non-parametric exemplar-based texture synthesis algorithm proposed by Efros and Leung, [EL99].

The ℓ_2 norm is the usual way to measure the similarity between patches [LBM13] but many other measurements exist, corresponding to different structural properties, see Figure 3.1.

Definition 3.1.2. *Let* $P, Q \in \mathbb{R}^{\mathsf{w}}$ *with* $\mathsf{w} \subset \mathbb{R}^2$ *or* $\mathsf{w} \subset \mathbb{Z}^2$. *When it is defined we introduce*

(a) the ℓ_q*-similarity,* $s_q(P,Q) = \left(\int_{p\in\mathsf{w}} |P(p) - Q(p)|^q \mathrm{d}p\right)^{1/q}$, *with* $q \in (0, +\infty)$;

(b) the ℓ_∞*-similarity,* $s_\infty(P,Q) = \sup_{p\in\mathsf{w}}(|P(p) - Q(p)|)$;

(c) the q*-th power of the* ℓ_q*-similarity,* $s_{q,q}(P,Q) = s_q(P,Q)^q$, *with* $q \in (0, +\infty)$;

(d) the scalar product similarity, $s_{sc}(P,Q) = -\langle P, Q\rangle = \frac{1}{2}\left(s_{2,2}(P,Q) - \|P\|_2^2 - \|Q\|_2^2\right)$;

(e) the cosine similarity, $s_{\cos}(P,Q) = s_{sc}(P,Q)(\|P\|_2\|Q\|_2)^{-1}$, *if* $\|P\|_2\|Q\|_2 \neq 0$.

Depending on the case $\mathrm{d}p$ *is either the Lebesgue measure or the discrete measure over* w.

The locality of the measurements is ensured by the fact that these functions are defined on patches, *i.e.* local windows. Following conditions (1) and (3) in [DDT12] one can check that similarity functions (a), (c) and (e) satisfy the following properties

- (Symmetry) $s(P,Q) = s(Q,P)$;
- (Maximal self-similarity) $s(P,P) \leq s(P,Q)$;
- (Equal self-similarities) $s(P,P) = s(Q,Q)$.

Note that since s_{sc}, the scalar product similarity, is homogeneous in P, maximal self-similarity and equal self-similarity properties are not satisfied. All introduced similarities satisfy the symmetry condition and s_∞ satisfies the maximal self-similarity property. In [DDT12], the authors present many other similarity functions all relying on statistical properties such as likelihood ratios, joint likelihood criteria and mutual information kernels. In this section we focus only on similarity functions defined directly in the spatial domain.

Definition 3.1.3. *Let x and y be two functions defined over a domain $\mathsf{E} \subset \mathbb{R}^2$ or $\mathbb{Z}^2$. Let $\mathsf{w} \subset \mathsf{E}$ be a patch domain. Let $P_\mathsf{w}(x) = x|_\mathsf{w}$ be the restriction of x to the patch domain w. When it is defined we introduce the auto-similarity with patch domain w and offset $t \in \mathbb{R}^2$ or $\mathbb{Z}^2$ such that $t + \mathsf{w} \subset \mathsf{E}$,*

$$\mathscr{A}_i(x, t, \mathsf{w}) = s_i\left(P_\mathsf{w}(x), (\tau_{-t} \circ P_{t+\mathsf{w}})(x)\right),$$

where s_i corresponds to s_q with $q \in (0, +\infty]$, $s_{q,q}$ with $q \in (0, +\infty)$, s_{sc} or s_{cos} and for any $p \in \mathsf{w}$, $(\tau_{-t} \circ P_{t+\mathsf{w}})(p) = x(p + t)$. Similarly, when it is defined, we introduce the template similarity with patch w and offset t by

$$\mathscr{T}_i(x, y, t, \mathsf{w}) = s_i\left(P_\mathsf{w}(y), (\tau_{-t} \circ P_{t+\mathsf{w}})(x)\right).$$

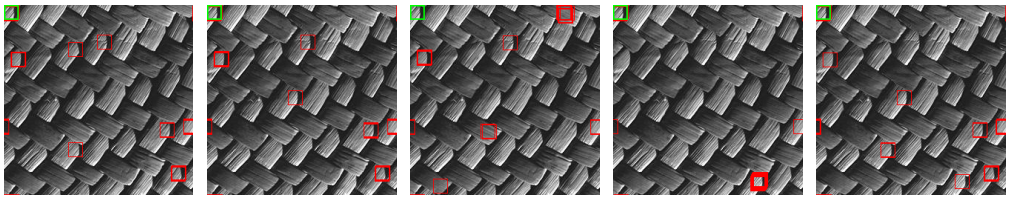
Note that in the finite discrete setting, *i.e.* $\mathsf{E} = (\mathbb{Z}/(M\mathbb{Z}))^2$ with $M \in \mathbb{N}$, the definition of $\mathscr{A}$ and $\mathscr{T}$ can be extended to any patch domain $\mathsf{w} \subset \mathbb{Z}^2$ by replacing x by $\dot{x}$, its periodic extension to $\mathbb{Z}^2$. A similar extension can be derived in the finite continuous setting, *i.e.* $\mathsf{E} = \mathbb{T}^2$.

Suppose we evaluate the scalar product auto-similarity $\mathscr{A}_{sc}(U, t, \mathsf{w})$ with U a random field. Then the auto-similarity function is a random variable and its expectation depends on the second-order statistics of U. In the template case, the expectation of $\mathscr{T}_{sc}(U, y, t, \mathsf{w})$ depends on the first-order statistics of U. This shows that auto-similarity and template similarity can exhibit very different behaviors even for the same similarity functions.

In the discrete case, it is well-known that the ℓ_2 norm does not behave well in large-dimensional spaces and is a poor measure of structure, due to the curse of dimensionality. Thus, considering x and y two images, $s_2(x, y)$, the ℓ_2 template similarity on full images, does not yield interesting information about the perceptual differences between x and y. The template similarity $\mathscr{T}_2(x, y, 0, \mathsf{w})$ avoids this effect by considering patches which reduces the dimension of the data (if the cardinality of w, denoted $|\mathsf{w}|$, is small) and also allows for fast computation of similarity mappings, see Figure 3.1 for a comparison of the different similarity functions on a natural image.

We extract patches from images as follow. For each position in the image we consider a square w centered around this position. This operation is called patch lifting. In Figure 3.2, we investigate the behavior of patch lifting on different Gaussian random fields. Roughly speaking, patches are said to be similar if they are clustered in the patch space. Using Principal Component Analysis we illustrate that patches are more scattered in Gaussian white noise than in the Gaussian random field $U = f * W$ (with periodic convolution, *i.e.* for any $p_1 \in \mathsf{E}$, $f * W(p_1) = \sum_{p_2 \in \mathsf{E}} W(p_2)\dot{f}(p_1 - p_2)$ where $\dot{f}$ is the periodic extension of f to $\mathbb{Z}^2$), where W is a Gaussian white noise over E (a finite discrete grid) and f is the indicator function of a rectangle non reduced to a single pixel.

We continue this investigation in Figure 3.3 in which we present the closest patches (of size 10×10), for the ℓ_2 norm, in two Gaussian random fields $U = f * W$ (where the convolution is periodic) for different functions f called spots, [Gal16]. The more regular f is, the more similar the patches are. Limit cases are $f = 0$ (all patches are constant and thus all the patches are similar) and $f = \delta_0$, *i.e.* $U = W$.



s_2 s_1 s_∞ s_{sc} $s_{\cos}$

Figure 3.1: ***Structural properties of similarity functions.*** In this experiment the image size is 512×512 and the patch size is 20×20. We show the 20 closest patches (red squares) to the upper-left patch (green square) among all patches for different similarity functions. All introduced similarity functions, see Definition 3.1.2, correctly identify the structure of the patch, *i.e.* a large clear part with diagonal textures and a dark ray on the right side of the patch, except for s_∞ which is too sensitive to outliers. Similarities s_2, s_1 and $s_{\cos}$ have analogous behaviors and identify correct regions. Similarity s_{sc} is too sensitive to contrast and, as it selects a correct patch, it gives too much importance to illumination.

We introduce the notion of autocorrelation. Let $f \in \ell_2(\mathbb{Z}^2)$. We denote by Γ_f the autocorrelation of f, *i.e.* $\Gamma_f = f * \check{f}$ where for any $p \in \mathbb{Z}^2$, $\check{f}(p) = f(-p)$ and define the associated random field to a square-integrable function f as the stationary Gaussian random field U such that for any $p \in \mathsf{E}$

$$\mathbb{E}\left[U(p)\right] = 0 \qquad \text{and} \quad \Gamma(p) = \Gamma_f(p)\ .$$

In Figure 3.4, we compare the patch spaces of natural images and the ones of their associated random fields. Since the associated Gaussian random fields lose all global structures, most of the spatial information is discarded. This situation can be observed in the patch space. In the natural images, patches containing the same highly spatial information (such as a white diagonal) are close for the ℓ_2 norm. In Gaussian random field since this highly spatial information is lost, close patches for the ℓ_2 norm are not necessarily perceptually close.

3.1.3 Asymptotic results

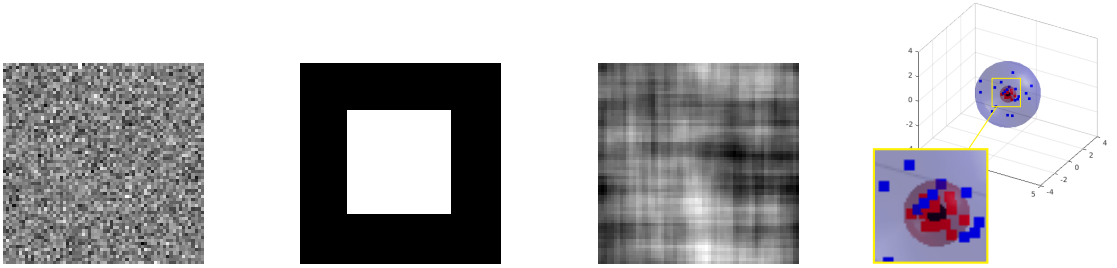
In this section we aim at giving explicit asymptotic expressions for the probability distribution functions of the auto-similarity and the template similarity in both discrete and continuous settings. Using general versions of the law of large numbers and central limit theorems we will derive Gaussian asymptotic approximations.

Additional assumptions are required in the case of template matching since we use an exemplar input image y to compute $\mathscr{T}_i(U, y, t, \mathsf{w})$. Let $y \in \mathbb{R}^{\mathsf{E}}$, where E is $\mathbb{R}^2$ or $\mathbb{Z}^2$. We denote by $(y_k)_{k\in\mathbb{N}}$ the sequence of the restriction of y to w_k, extended to $\mathbb{Z}^2$ (or $\mathbb{R}^2$) by zero-padding, *i.e.* $y_k(p) = 0$ for $p \notin \mathsf{w}_k$. We suppose that $\lim_{k\to+\infty} |\mathsf{w}_k| = +\infty$, where $|\mathsf{w}_k|$ is the Lebesgue measure, respectively the cardinality, of w_k if $\mathsf{E} = \mathbb{R}^2$, respectively $\mathsf{E} = \mathbb{Z}^2$. Note that the following assumptions are well-defined for both continuous and discrete settings.

A3. *The function y is bounded on E.*

The following assumption ensures the existence of spatial moments of any order for the function v.

A4. *For any $m, n \in \mathbb{N}$, there exist $\beta_m \in \mathbb{R}\backslash\{0\}$ and $\gamma_{m,n} \in \mathbb{R}^{\mathsf{E}}$ such that*



(a) (b) (c) (d)

Figure 3.2: **Gaussian models and spatial redundancy** In this experiment we illustrate the notion of spatial redundancy in two models. In (A), we present a 64×64 Gaussian white noise. (B) shows an indicator function f . In (C), we present a realization of the Gaussian random field defined by $f * W$ (with periodic convolution) where W is a Gaussian white noise over E (domain of size 64×64). Note that f was chosen so that the two Gaussian random fields (A) and (C) have the same gray-level distribution for each pixel. To each pixel position in (A) and (C) we associate the surrounding patch, with patch domain w (of size 3×3). Hence, for each image (A) and (C) we obtain $64 \times 64 = 5096$ vectors each of size $3 \times 3 = 9$. These 9-dimensional vectors are projected in a 3-dimensional space using Principal Component Analysis. In the subfigure (D), we display the 20 vectors closest to 0 in each case: Gaussian white noise model (in blue) and the Gaussian random field (C) (in red). The radius of the blue, respectively red, sphere represents the maximal ℓ_2 norm of these 20 vectors in the Gaussian white noise model, respectively in model (C). Since the radius of the blue sphere is larger than the red one the points are more scattered in the patch space of (A) than in the patch space of (B). This implies that there is more spatial redundancy in (C) than in (A), which is expected.

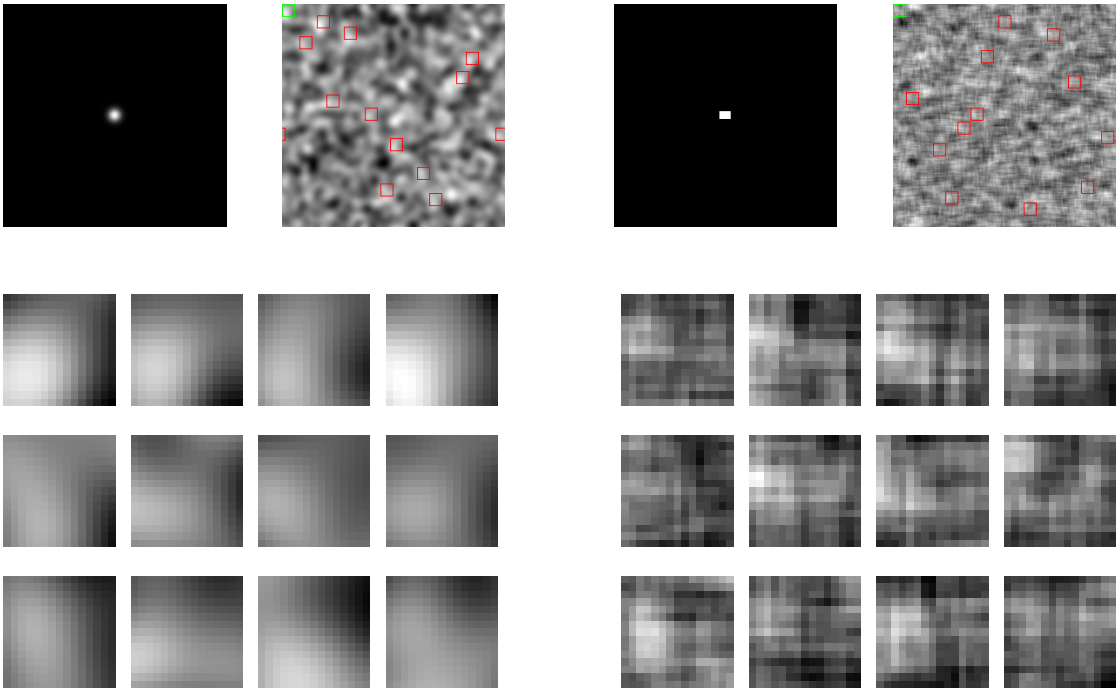


Figure 3.3: **Patch similarity in Gaussian random fields** In this figure we show two examples of Gaussian random fields in the discrete periodic case. On the left of the first row we show a Gaussian spot f and a realization of the Gaussian random field $U = f * W$, where the convolution is periodic and W is a Gaussian white noise. The associated random field is smooth and isotropic. The random field $U = f * W$ associated with a rectangular plateau f is no longer smooth nor isotropic. Images are displayed on the right of their respective spot. For each setting (Gaussian spot or rectangular spot) we present 12 patches of size 15×15. In each case the top-left patch is the top-left patch in the presented realization of the random field, shown in green. Following from the top to the bottom and from the left to the right are the closest patches in the patch space for the ℓ_2 norm. We discard patches which are spatially too close (if w_1 and w_2 are two patch domains we impose $\sup_{p_1,p_2} \|p_1 - p_2\|_\infty \geq 10$).

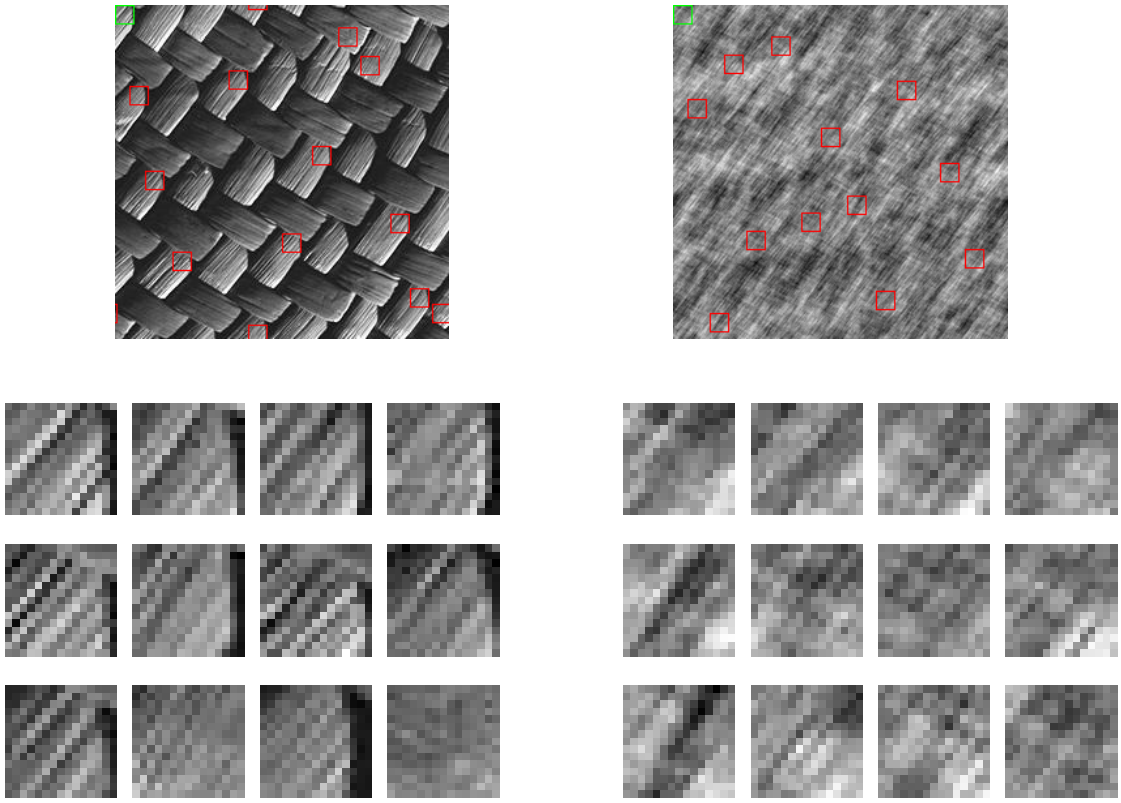


Figure 3.4: **Natural images and Gaussian random fields** In this experiment we present the same image, f, which was used in Figure 3.1 and the associated Gaussian random field $U = f * W$, where the convolution is periodic and W is a Gaussian white noise. As in Figure 3.3 we present under each image the top-left patch (of size 15×15 and shown in green in the original images) and its 11 closest matches for the ℓ_2 similarity. We discard patches which are spatially too close (if w_1 and w_2 are two patch domains we impose $\sup_{p_1,p_2} \|p_1 - p_2\|_\infty \geq 10$). Note that if a structure is clearly identified in the real image (black and white diagonals) and is retrieved in every patch, it is not as clear in the Gaussian random field.

(a) $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2}\left(|\mathsf{w}_k|^{-1}\int_{\mathsf{w}_k} y_k^{2m}(p)\mathrm{d}p - \beta_m\right) = 0$;

(b) *for any* $\mathsf{K}\subset\mathsf{E}$ *compact,* $\lim_{k\to+\infty}\sup_{p_1\in\mathsf{K}}\{|\mathsf{w}_k|^{-1}\int_{p_2\in\mathsf{w}_k} y_k^{2m}(p_2)y_k^{2n}(p_1+p_2)\mathrm{d}p_2-\gamma_{m,n}(p_1)|\} = 0$.

Note that in the case where E is discrete, the uniform convergence on compact sets introduced in (b) is equivalent to the pointwise convergence.

A5. *There exists* $\gamma\in\mathbb{R}^{\mathsf{E}}$ *with for any* $\mathsf{K}\subset\mathsf{E}$ *compact,*

$$\lim_{k\to+\infty}\sup_{p_1\in\mathsf{K}}\{|\mathsf{w}_k|^{-1}\int_{p_2\in\mathsf{w}_k} y_k(p_2)y_k(p_1+p_2)\mathrm{d}p_2-\gamma(p_1)|\} = 0 \ .$$

Discrete case

In the discrete case, we consider a random field U over $\mathbb{Z}^2$ and compute local similarity measurements. The asymptotic approximation is obtained when the patch size grows to infinity. In Theorem 3.1.4 and Theorem 3.1.6 we obtain Gaussian asymptotic probability distribution in the auto-similarity case and in the template similarity case. In Proposition 3.1.5 and Proposition 3.1.7 we give explicit mean and variance for the Gaussian approximations.

Theorem 3.1.4. *Let* $(m_k)_{k\in\mathbb{N}},(n_k)_{k\in\mathbb{N}}$ *be two positive increasing integer sequences and* $(\mathsf{w}_k)_{k\in\mathbb{N}}$ *be the sequence of subsets defined for any* $k\in\mathbb{N}$ *by,* $\mathsf{w}_k = [\![0,m_k]\!]\times[\![0,n_k]\!]$. *Let* $f:\mathbb{Z}^2\to\mathbb{R}$, $f\neq 0$ *with finite support,* W *a Gaussian white noise over* $\mathbb{Z}^2$ *and* $U = f*W$. *For* $i\in\{q,(q,q),sc,\cos\}$ *with* $q\in(0,+\infty)$ *there exist* μ_i,σ_i^2, *real valued functions on* $\mathbb{Z}^2$, *and* $(\alpha_{i,k})_{k\in\mathbb{N}}$ *a positive sequence such that for any* $t\in\mathbb{Z}^2\backslash\{0\}$ *we have*

(a) $\lim_{k\to+\infty}\alpha_{i,k}^{-1}\mathscr{A}_i(U,t,\mathsf{w}_k)\underset{a.s}{=}\mu_i(t)$;

(b) $\lim_{k\to+\infty}|\mathsf{w}_k|^{1/2}\left(\alpha_{i,k}^{-1}\mathscr{A}_i(U,t,\mathsf{w}_k)-\mu_i(t)\right)\underset{\mathcal{L}}{=}\mathsf{N}\left(0,\sigma_i^2(t)\right)$,

where $\mathsf{N}(m,\sigma^2)$ *is the one-dimensional Gausian distribution with mean* m *and variance* σ^2.

The asymptotics derived in Theorem 3.1.4 can be extended to vectors of autosimilarities, *i.e.* selecting $(t_j)_{j\in\{1\dots N\}}$ a finite number of shifts the results of Theorem 3.1.4 hold for the sequence $((\mathscr{A}_i(U,t_j,\mathsf{w}_k))_{j\in\{1\dots N\}})_{k\in\mathbb{N}}$. Note that in theorem 3.1.4 if t varies with k such that for any $k\in\mathbb{N}$, $(\mathsf{w}_k+t_k)\cap\mathsf{w}_k=\emptyset$ then similar results can be obtained with the usual law of large numbers and central limit theorem since true independence hold.

Proof. The proof is divided into three parts. First we show 1 and 2 for $i=(q,q)$ and extends the result to $i=q$. Then we show 1 and 2 for $i=sc$. Finally, we show 1 and 2 for $i=\cos$.

1. Let $q\in(0,+\infty)$, $t\in\mathbb{Z}^2\backslash\{0\}$ and define $V_{q,t}$ for any $p\in\mathbb{Z}^2$ by, $V_{q,t}(p)=|U(p)-U(p+t)|^q$. We remark that for any $k\in\mathbb{N}$ we have

$$\mathscr{A}_{q,q}(U,t,\mathsf{w}_k)=\sum_{p\in\mathsf{w}_k}V_{q,t}(p) \ .$$

First, remark that U is R-independent with $R>0$, see Lemma 3.1.20. Since for any $p\in\mathbb{Z}^2$, $V_{q,t}(p)$ depends only on $U(p)$ and $U(p+t)$ we have that $V_{q,t}$ is $R_t = R+\|t\|_\infty$-independent. Since U is

stationary, so is $V_{q,t}$. The random field $V_{q,t}$ admits moments of every order since it is the q-th power of the absolute value of a Gaussian random field. Thus $V_{q,t}$ is a R_t-independent second-order stationary random field. We can apply Lemma 3.1.21 and we get

(a) $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \mathscr{A}_{q,q}(U,t,\mathsf{w}_k) \underset{a.s.}{=} \mu_{q,q}(t)$;

(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(|\mathsf{w}_k|^{-1} \mathscr{A}_{q,q}(U,t,\mathsf{w}_k) - \mu_{q,q}(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \sigma^2_{q,q}(t)\right)$.

with $\mu_{q,q}(t) = \mathbb{E}\left[V_{q,t}(0)\right]$ and $\sigma_{q,q}(t)^2 = \sum_{p_1\in\mathbb{Z}^2} \mathrm{Cov}\left[V_{q,t}(p_1), V_{q,t}(0)\right]$. By continuity of the q-th root over $[0,+\infty)$ we get 1 for $i = q$ with

$$\alpha_{q,k} = |\mathsf{w}_k|^{1/q} , \qquad \mu_q(t) = \mu_{q,q}(t)^{1/q} .$$

By Lemma 3.1.22 we get that $\mathbb{E}\left[(U(0)-U(t))^2\right] = 2(\Gamma_f(0) - \Gamma_f(t)) > 0$ thus $\mu_{q,q}(t) = \mathbb{E}\left[V_{q,t}(0)\right] > 0$. Since the q-th root is continuously differentiable on $(0,+\infty)$ we can apply the Delta method, see [Cra99], and we get 2 for $i = q$ with

$$\alpha_{p,k} = |\mathsf{w}_k|^{1/q} , \qquad \mu_q(t) = \mu_{q,q}(t)^{1/q} , \qquad \sigma_q(t)^2 = q^{-2}\sigma_{q,q}(t)^2 \mu_{q,q}(t)^{2/q-2} . \tag{3.1}$$

2. We now prove the theorem for $i = sc$. Let $t \in \mathbb{Z}^2\backslash\{0\}$ and define $V_{sc,t}$ for any $p \in \mathbb{Z}^2$, $V_{sc,t}(p) = -U(p)U(p+t)$. We remark that for any $k \in \mathbb{N}$ we have

$$\mathscr{A}_{sc}(U,t,\mathsf{w}_k) = \sum_{p\in\mathsf{w}_k} V_{sc,t}(p) .$$

Since for any $p_1 \in \mathbb{Z}^2$, $V_{sc,t}(p)$ depends only on $U(p)$ and $U(p+t)$, we have that $V_{sc,t}$ is $R_t = R + \|t\|_\infty$-independent. Since U is stationary, so is $V_{sc,t}$. The random field $V_{sc,t}$ admits moments of every order since it is a product of Gaussian random fields. Thus $V_{sc,t}$ is a R_t-independent second-order stationary random field. We can again apply Lemma 3.1.21 and we get

(a) $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \mathscr{A}_{sc}(U,t,\mathsf{w}_k) \underset{a.s.}{=} \mu_{sc}(t)$;

(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(|\mathsf{w}_k|^{-1} \mathscr{A}_{sc}(U,t,\mathsf{w}_k) - \mu_{sc}(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \sigma^2_{sc}(t)\right)$,

with $\mu_{sc}(t) = \mathbb{E}\left[V_{sc,t}(0)\right]$ and $\sigma_{sc}(t)^2 = \sum_{p_1\in\mathbb{Z}^2} \mathrm{Cov}\left[V_{sc,t}(p_1), V_{sc,t}(0)\right]$, which concludes the proof.

3. Finally, we consider the case $i = \cos$. Let $t \in \mathbb{Z}^2\backslash\{0\}$ and define $V_{\cos,t}$ for any $p \in \mathbb{Z}^2$,

$$V_{\cos,t}(p) = \begin{pmatrix} -U(p)U(p+t) \\ U(p)^2 \\ U(p+t)^2 \end{pmatrix} .$$

We remark that for any $k \in \mathbb{N}$ we have

$$\mathscr{A}_{s_{\cos}}(U,t,\mathsf{w}_k) = h\left(|\mathsf{w}_k|^{-1} \sum_{p\in\mathsf{w}_k} V_{\cos,t}(p)\right) ,$$

with $h(x,y,z) = xy^{-1/2}z^{-1/2}$. Since U is stationary, so is $V_{\cos,t}$. The random field $V_{\cos,t}$ admits moments of every order since it is a vector of products of Gaussian random fields. Thus $V_{\cos,t}$ is a R_t-independent second-order stationary random field. We can apply Lemma 3.1.21 and there exist $\tilde{\mu}_{\cos}(t)$ and $\tilde{C}_{\cos}(t)$ such that

(a) $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} V_{\cos,t} \underset{a.s.}{=} \tilde{\mu}_{\cos}(t)$;

(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(|\mathsf{w}_k|^{-1} V_{\cos,t} - \tilde{\mu}_{\cos}(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \tilde{C}_{\cos}(t)\right)$.

We conclude the proof using the multivariate Delta method, [Cra99].

□

In the following proposition we give explicit values for the constants involved in the law of large numbers and the central limit theorem derived in Theorem 3.1.4. We introduce the following quantities for $k, \ell \in \mathbb{N}$ and $j \in [\![0, k \wedge \ell]\!]$, where $k \wedge \ell = \min(k, \ell)$,

$$q_\ell = \frac{(2\ell)!}{\ell!\, 2^\ell} , \qquad r_{j,k,\ell} = q_{k-j} q_{\ell-j} \binom{2k}{2j} \binom{2\ell}{2j} (2j)! . \tag{3.2}$$

We also denote $r_{j,\ell} = r_{j,\ell,\ell}$. Note that for all $\ell \in \mathbb{N}$, $r_{0,\ell} = q_\ell^2$ and $\sum_{j=0}^{\ell} r_{j,\ell} = q_{2\ell}$. We also introduce the following functions:

$$\Delta_f(t,p) = 2\Gamma_f(p) - \Gamma_f(p+t) - \Gamma_f(p-t) , \quad \widetilde{\Delta}_f(t,p) = \Gamma_f(p)^2 + \Gamma_f(p+t)\Gamma_f(p-t) . \tag{3.3}$$

Note that Δ_f is a second-order statistic on the Gaussian field $U = f * W$ with W a Gaussian white noise over $\mathbb{Z}^2$, whereas $\widetilde{\Delta}_f$ is a fourth-order statistic on the same random field.

Proposition 3.1.5. *In Theorem 3.1.4 we have the following constants for any $t \in \mathbb{Z}^2 \setminus \{0\}$.*

(i) If $i = q$ with $q = 2\ell$ and $\ell \in \mathbb{N}$, then for all $k \in \mathbb{N}$, we get that $\alpha_{p,k} = |\mathsf{w}_k|^{1/(2\ell)}$ and

$$\mu_q(t) = q_\ell^{1/(2\ell)} \Delta_f(t,0)^{1/2} \quad \text{and} \quad \sigma_q(t)^2 = \frac{q_\ell^{1/\ell-2}}{(2\ell)^2} \sum_{j=1}^{\ell} r_{j,\ell} \left(\frac{\|\Delta_f(t,\cdot)\|_{2j}}{\Delta_f(t,0)}\right)^{2j} \Delta_f(t,0) ,$$

where $(r_{i,jk})_{i,j,k\in\mathbb{N}}$ and $(q_k)_{k\in\mathbb{N}}$ are given in (3.2).

(ii) If $i = sc$, then for all $k \in \mathbb{N}$, we get that $\alpha_{sc,k} = |\mathsf{w}_k|$ and

$$\mu_{sc}(t) = \Gamma_f(t) \quad \text{and} \quad \sigma_{sc}(t)^2 = \sum_{p\in\mathbb{Z}^2} \widetilde{\Delta}_f(t,p) .$$

(iii) if $i = \cos$, then for all $k \in \mathbb{N}$, we get that $\alpha_{\cos,k} = 1$ and

$$\mu_{\cos}(t) = \Gamma_f(t)/\Gamma_f(0) \text{and}$$
$$\sigma_{\cos}(t)^2 = \Gamma_f(0)^{-2} \left\{ \|\Gamma_f\|_2^2 \left(1 + 2\frac{\Gamma_f(t)^2}{\Gamma_f(0)^2}\right) - 4\frac{\Gamma_f(t)}{\Gamma_f(0)} \Gamma_f * \breve{\Gamma}_f(t) + \Gamma_f * \breve{\Gamma}_f(2t) \right\} .$$

Proof. The proof is postponed to Appendix 3.1.5. □

For example we have

$$\mu_2(t) = \Delta_f(t,0)^{1/2} , \qquad \mu_4(t) = 3^{1/4} \Delta_f(t,0)^{1/2}$$
$$\sigma_2^2(t) = \frac{1}{2} \frac{\|\Delta_f(t,\cdot)\|_2^2}{\Delta_f(t,0)} , \qquad \sigma_4^2(t) = 2\sqrt{3} \frac{\|\Delta_f(t,\cdot)\|_2^2}{\Delta_f(t,0)} + \frac{\sqrt{3}}{6} \frac{\|\Delta_f(t,\cdot)\|_4^4}{\Delta_f(t,0)^3} .$$

We now derive similar asymptotic properties in the template similarity case.

Theorem 3.1.6. *Let $(m_k)_{k\in\mathbb{N}}, (n_k)_{k\in\mathbb{N}}$ be two positive increasing integer sequences and $(\mathsf{w}_k)_{k\in\mathbb{N}}$ be the sequence of subsets defined for any $k \in \mathbb{N}$, $\mathsf{w}_k = [\![0, m_k]\!] \times [\![0, n_k]\!]$. Let $f : \mathbb{Z}^2 \to \mathbb{R}$, $f \neq 0$ with finite support, W a Gaussian white noise over $\mathbb{Z}^2$, $U = f * W$ and let y, a real valued function on $\mathbb{Z}^2$. For $i = \{q, (q,q), sc, \cos\}$ with $q = 2\ell$ and $\ell \in \mathbb{N}$, if $i = q$ or (q,q) assume* **A**3 *and* **A**4*, if $i = sc$ assume* **A**3 *and* **A**5 *and if $i = \cos$ assume* **A**3*,* **A**4 *and* **A**5*. Then there exist $\mu_i, \sigma_i^2 \in \mathbb{R}$ and $(\alpha_{i,k})_{k\in\mathbb{N}}$ a positive sequence such that for any $t \in \mathbb{Z}^2$ we get*

1. $\lim_{k\to+\infty} \alpha_{i,k}^{-1} \mathscr{T}_i(U, y, t, \mathsf{w}_k) \underset{a.s}{=} \mu_i$;
2. $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(\alpha_{i,k}^{-1} \mathscr{T}_i(U, y, t, \mathsf{w}_k) - \mu_i(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \sigma_i^2\right)$.

Note that contrarily to Theorem 3.1.4 we could not obtain such a result for all $q \in (0, +\infty)$ but only for even integers. Indeed, the convergence of the sequence $\left(|\mathsf{w}_k|^{-1}\mathbb{E}\left[\mathscr{T}_{q,q}(U, y, t, \mathsf{w}_k)\right]\right)_{k\in\mathbb{N}}$, which is needed in order to apply Theorem 3.1.18, is not trivial. Assuming that y is bounded it is easy to show that $\left(|\mathsf{w}_k|^{-1}\mathbb{E}\left[\mathscr{T}_{q,q}(U, y, t, \mathsf{w}_k)\right]\right)_{k\in\mathbb{N}}$ is also bounded and we can deduce the existence of a convergent subsequence. In the general case, for Theorem 3.1.6 to hold with any $p \in (0, +\infty)$, we must check that for any $t \in \mathsf{E}$, there exist $\mu_{q,q}(t) > 0$ and $\sigma_{q,q}^2(t) \geq 0$ such that

(a) $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(|\mathsf{w}_k|^{-1}\mathbb{E}\left[\mathscr{T}_{q,q}(U, y, t, \mathsf{w}_k)\right] - \mu_{q,q}(t)\right) = 0$;

(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \operatorname{Var}\left[\mathscr{T}_{q,q}(U, y, t, \mathsf{w}_k)\right] = \sigma_{q,q}^2(t)$.

We now turn to the proof of Theorem 3.1.6.

Proof. As for the proof of theorem 3.1.4, the proof is divided into three parts. First we show 1 and 2 for $i = (q,q)$ and extends the result to $i = q$. Then we show 1 and 2 for $i = sc$. Finally, we show 1 and 2 for $i = \cos$.

1. Let $q = 2\ell$ with $\ell \in \mathbb{N}$, $t \in \mathbb{Z}^2$ and define $V_{q,t}$ the random field on $\mathbb{Z}^2$ for any $p \in \mathbb{Z}^2$, by $V_{p,t}(p) = |y(p) - U(p+t)|^q$. We remark that for any $k \in \mathbb{N}$ we have

$$\mathscr{T}_{q,q}(U, v, t, \mathsf{w}_k) = \sum_{p\in\mathsf{w}_k} V_{q,t}(p) .$$

By Lemma 3.1.20, U is R-independent with $R > 0$. Since for any $p \in \mathbb{Z}^2$ we have that $V_{q,t}(p)$ depends only on $U(p+t)$ we also have that $V_{q,t}$ is R-independent. We define the random field $V_{q,t}^{\infty}$ for any $p \in \mathbb{Z}^2$, $V_{q,t}^{\infty}(p) = (\|y\|_{\infty} + U(p+t))^q$. We have that $V_{q,t}^{\infty}(p) + \mathbb{E}[V_{q,t}^{\infty}(0)]$ uniformly almost surely dominates $V_{q,t}(p) - \mathbb{E}[V_{q,t}(p)]$. The random field $V_{q,t}^{\infty}$ admits moments of every order since it is the q-th power of the absolute value of a Gaussian random field and is stationary because U is. Thus $V_{q,t}$ is a R_t-independent random field and $V_{q,t}(p) - \mathbb{E}[V_{q,t}(p)]$ is uniformly stochastically dominated by $V_{q,t}^{\infty}(p) + \mathbb{E}[V_{q,t}^{\infty}(0)]$, a second-order stationary random field. Using **A**4 and Lemma 3.1.23, we can apply Theorem 3.1.18 and Theorem 3.1.19 and we get

(a) $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \mathscr{T}_{q,q}(U, y, t, \mathsf{w}_k) \underset{a.s.}{=} \mu_{q,q}(t)$;

(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(|\mathsf{w}_k|^{-1} \mathscr{T}_{q,q}(U, y, t, \mathsf{w}_k) - \mu_{q,q}(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \sigma_{q,q}^2(t)\right)$.

Note that since U is stationary we have for any $t \in \mathbb{Z}^2$, $\mu_{q,q} = \mu_{q,q}(0) = \mu_{q,q}(t)$ and $\sigma_{q,q}^2 = \sigma_{q,q}^2(0) = \sigma_{q,q}^2(t)$. By continuity of the q-th root over $[0, +\infty)$ we get 1 for $i = p$ with

$$\alpha_{q,k} = |\mathsf{w}_k|^{1/p} , \qquad \mu_q = \mu_{q,q}^{1/q} .$$

By Lemma 3.1.23, we have that $\mu_{q,q} > 0$. Since the q-th root is continuously differentiable on $(0, +\infty)$ we can apply the Delta method and we get 2 for $i = q$ with

$$\alpha_{p,k} = |\mathsf{w}_k|^{1/p} , \qquad \mu_q = \mu_{q,q}^{1/q} , \qquad \sigma_q^2 = \sigma_{q,q}^2 \mu_{q,q}^{2/q-2} q^{-2} . \tag{3.4}$$

2. We now prove the theorem for $i = sc$. Let $t \in \mathbb{Z}^2$ and define $V_{sc,t}$ the random field on $\mathbb{Z}^2$ such that for any $p \in \mathbb{Z}^2$, $V_{sc,t}(p) = -y(p)U(p+t)$. We remark that for any $k \in \mathbb{N}$ we have

$$\mathscr{T}_{sc}(U, y, t, \mathsf{w}_k) = \sum_{p \in \mathsf{w}_k} V_{sc,t}(p) .$$

It is clear that for any $k \in \mathbb{N}$, $\mathscr{T}_{sc}(U, y, t, \mathsf{w}_k)$ is a R-independent Gaussian random variable with $\mathbb{E}[\mathscr{T}_{sc}(U, y, t, \mathsf{w}_k)] = 0$ and

$$\begin{aligned} \mathrm{Var}\left[\mathscr{T}_{sc}(U, y, t, \mathsf{w}_k)\right] &= \sum_{p_1, p_2 \in \mathsf{w}_k} \mathbb{E}\left[V_{sc,t}(p_1) V_{sc,t}(p_2)\right] \\ &= \sum_{p_1, p_2 \in \mathsf{w}_k} y(p_1) y(p_2) \Gamma_f(p_1 - p_2) = \sum_{p_1 \in \mathbb{Z}^2} \Gamma_f(p_1) y_k * \check{y}_k(p_1) , \end{aligned}$$

where we recall that y_k is the restriction of y to w_k. The last sum is finite since $\mathrm{supp} f$ finite implies that $\mathrm{supp}\Gamma_f$ is finite. Using **A**5 we obtain that for any $k \in \mathbb{N}$,

$$\sum_{p_1 \in \mathsf{w}_k} \left(\mathbb{E}\left[V_{sc,t}\right](p_1) - \mu_{sc}\right) = 0 , \qquad \lim_{k \to +\infty} |\mathsf{w}_k|^{-1} \sum_{p_1, p_2 \in \mathsf{w}_k} \mathrm{Cov}\left[V_{sc,t}(p_1), V_{sc,t}(p_2)\right] = \sigma_{sc}^2 , \tag{3.5}$$

with $\mu_{sc} = 0$ and $\sigma_{sc}^2 = \sum_{p_1 \in \mathbb{Z}^2} \Gamma_f(p_1)\gamma(p_1)$, where γ is given in **A**5. Since $V_{sc,t}$ is a R-independent second-order random field using (3.5) we can apply Theorem 3.1.18 and Theorem 3.1.19 to conclude.

3. We now consider the case $i = \cos$. First, notice that

$$\begin{aligned} &\mathscr{T}_{\cos}(U, y, t, \mathsf{w}_k) \\ &\qquad = |\mathsf{w}_k|^{-1} \mathscr{T}_{sc}(U, y, t, \mathsf{w}_k) \left[\left(|\mathsf{w}_k|^{-1} \sum_{p \in \mathsf{w}_k} y(p)^2 \right)^{1/2} \left(|\mathsf{w}_k|^{-1} \sum_{p \in \mathsf{w}_k} U(p)^2 \right)^{1/2} \right]^{-1} . \end{aligned}$$

Using that $\lim_{k \to +\infty} |\mathsf{w}_k|^{-1} \mathscr{T}_{sc}(U, y, t, \mathsf{w}_k) = 0$, $\lim_{k \to +\infty} |\mathsf{w}_k|^{-1} \sum_{p \in \mathsf{w}_k} U(p)^2 = \Gamma_f(0)$ by Lemma 3.1.21 almost surely and $\lim_{k \to +\infty} |\mathsf{w}_k|^{-1} \sum_{p_1 \in \mathsf{w}_k} y(p_1)^2 = \beta_1 \neq 0$ by **A**4, we get that

$$\lim_{k \to +\infty} \mathscr{T}_{\cos}(U, v, t, \mathsf{w}_k) = 0 .$$

In addition, using Slutsky's theorem and the fact that $\lim_{k \to +\infty} |\mathsf{w}_k|^{-1/2} \mathscr{T}_{sc}(U, y, t, \mathsf{w}_k) = \mathsf{N}(0, \sigma_{sc}^2)$ we obtain that $\lim_{k \to +\infty} |\mathsf{w}_k|^{-1/2} \mathscr{T}_{cos}(U, y, t, \mathsf{w}_k) = \mathsf{N}(0, \sigma_{cos}^2)$ with

$$\sigma_{cos}^2 = \frac{\langle \gamma, \Gamma_f \rangle}{\beta_1 \Gamma_f(0)} .$$

□

Proposition 3.1.7. *In Theorem 3.1.6 we have the following constants for any $t \in \mathbb{Z}^2$.*

(i) If $i = q$ with $q = 2\ell$ and $\ell \in \mathbb{N}$, then we get that $\alpha_{q,k} = |\mathsf{w}_k|^{1/q}$, and

$$\mu_q = \left(\sum_{j=0}^{\ell} \binom{2\ell}{2j} q_{\ell-j} \Gamma_f(0)^{-j} \beta_j \right)^{1/2\ell} \Gamma_f(0)^{1/2} ,$$

$$\sigma_q^2 = \left(\sum_{i,j=0}^{\ell} \binom{2\ell}{2i} \binom{2\ell}{2j} \sum_{m=1}^{\ell-i\wedge\ell-j} r_{m,\ell-i,\ell-j} \Gamma_f(0)^{-(i+j+2m)} \left\langle \Gamma_f^{2m}, \gamma_{i,j} \right\rangle \right)$$
$$\times \left(\sum_{j=0}^{\ell} \binom{2\ell}{2j} q_{\ell-j} \Gamma_f(0)^{-j} \beta_j \right)^{1/\ell-2} \frac{\Gamma_f(0)}{(2\ell)^2} ,$$

where $(\beta_j)_{j\in\mathbb{N}}$, $(\gamma_{i,j})_{i,j\in\mathbb{N}}$ are given in **A**4 *and $(r_{i,jk})_{i,j,k\in\mathbb{N}}$ and $(q_k)_{k\in\mathbb{N}}$ are given in* (3.2).

(ii) If $i = sc$ then for all $k \in \mathbb{N}$, we get that $\alpha_{sc,k} = |\mathsf{w}_k|$ and

$$\mu_{sc} = 0 , \quad \sigma_{sc}^2 = \langle \gamma, \Gamma_f \rangle .$$

(iii) If $i = s_{\cos}$ then for all $k \in \mathbb{N}$, we get that $\alpha_{s_{\cos},k} = 1$ and

$$\mu_{s_{\cos}} = 0 , \quad \sigma_{s_{\cos}}^2 = \frac{\langle \gamma, \Gamma_f \rangle}{\beta_1 \Gamma_f(0)} .$$

Proof. The proof is postponed to Appendix 3.1.5. □

For example we have

$$\mu_2 = (2\Gamma_f(0) + \beta_1)^{1/2} , \quad \mu_4 = (3\Gamma_f(0)^2 + 12\Gamma_f(0)^3\beta_1 + \beta_2)^{1/4} ,$$
$$\sigma_2^2 = \frac{1}{4} \frac{\|\Gamma_f\|_2^2}{\Gamma_f(0)} (2 + \Gamma_f(0)^{-1}\beta_1)^{-1} ,$$
$$\sigma_4^2 = \frac{1}{16} \left(288\Gamma_f(0)^{-1}\|\Gamma_f\|_2^2 + 144\Gamma_f(0)^{-2}\langle \Gamma_f^2, \gamma_{0,1} \rangle + 24\Gamma_f(0)^{-3}\|\Gamma_f\|_4^4 + \Gamma_f(0)^{-3}\langle \Gamma_f^2, \gamma \rangle \right)$$
$$\times \left(3 + 12\Gamma_f(0)^{-1}\beta_1 + \Gamma_f(0)^{-2}\beta_2 \right)^{-3/2} .$$

Note that the limit mean and standard deviation do not depend on the offset anymore. Indeed, template similarity functions are stationary in t. If v has finite support then **A**4 holds with $\beta_i = 0$ and $\gamma_{i,j} = 0$ as soon as $i \neq 0$ or $j \neq 0$. Remarking that $\beta_0 = 1$ and $\gamma_{0,0} = 1$ we obtain that

$$\mu_q = q_\ell^{1/(2\ell)} \Gamma_f(0)^{1/2} , \quad \sigma_q^2 = \frac{q_\ell^{1/\ell-2}}{(2\ell)^2} \sum_{j=1}^{\ell} r_{j,\ell} \left(\frac{\|\Gamma_f\|_{2j}}{\Gamma_f(0)} \right)^{2j} \Gamma_f(0) .$$

Continuous case

We now turn to the continuous setting. Theorem 3.1.8, respectively Theorem 3.1.10, is the continuous counterpart of Theorem 3.1.4, respectively Theorem 3.1.6.

Theorem 3.1.8. *Let $(m_k)_{k\in\mathbb{N}}, (n_k)_{k\in\mathbb{N}}$ be two positive increasing integer sequences and $(\mathsf{w}_k)_{k\in\mathbb{N}}$ be the sequence of subsets defined for any $k \in \mathbb{N}$ by, $\mathsf{w}_k = [0, m_k] \times [0, n_k]$. Let U be a zero-mean Gaussian random field over $\mathbb{R}^2$ with covariance function Γ. Assume **A**2 and that Γ has finite support. For $i \in \{q, (q, q), sc, \cos\}$ with $q \in (0, +\infty)$ there exist μ_i, σ_i^2, real valued functions on $\mathbb{R}^2$, and $(\alpha_{i,k})_{k\in\mathbb{N}}$ a positive sequence such that for any $t \in \mathbb{R}^2 \setminus \{0\}$ we get*

1. $\lim_{k\to+\infty} \alpha_{i,k}^{-1} \mathscr{A}_i(U, t, \mathsf{w}_k) \underset{a.s}{=} \mu_i(t)$;
2. $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(\alpha_{i,k}^{-1} \mathscr{A}_i(U, t, \mathsf{w}_k) - \mu_i(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \sigma_i^2(t)\right)$.

Proof. The proof is the same as the one of Theorem 3.1.4 replacing Lemma 3.1.21 and Lemma 3.1.22 by Lemma 3.1.26 and Lemma 3.1.27. □

Proposition 3.1.9. *Constants given in Proposition 3.1.5 apply to Theorem 3.1.8 provided that Γ_f is replaced by Γ in (3.3).*

Proof. The proof is the same as the one of Proposition 3.1.5. □

Theorem 3.1.10. *Let $(m_k)_{k\in\mathbb{N}}, (n_k)_{k\in\mathbb{N}}$ be two positive increasing integer sequences and $(\mathsf{w}_k)_{k\in\mathbb{N}}$ be the sequence of subsets defined for any $k \in \mathbb{N}$ by, $\mathsf{w}_k = [0, m_k] \times [0, n_k]$. Let U be a zero-mean Gaussian random field over $\mathbb{R}^2$ with covariance function Γ. Assume **A**2 and that Γ has finite support. For $i \in \{q, (q, q), sc, \cos\}$ with $p \in (0, +\infty)$, if $i = q$ or (q, q) assume **A**3 and **A**4, if $i = sc$ assume **A**3 and **A**5 and if $i = \cos$ assume **A**3, **A**4 and **A**5. Then there exist $\mu_i, \sigma_i^2 \in \mathbb{R}$ and $(\alpha_{i,k})_{k\in\mathbb{N}}$ a positive sequence such that for any $t \in \mathbb{R}^2$ we get*

1. $\lim_{k\to+\infty} \alpha_{i,k}^{-1} \mathscr{T}_i(U, y, t, \mathsf{w}_k) \underset{a.s.}{=} \mu_i$;
2. $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(\alpha_{i,k}^{-1} \mathscr{T}_i(U, y, t, \mathsf{w}_k) - \mu_i(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \sigma_i^2\right)$.

Proof. The proof is the same as the one of Theorem 3.1.6. □

Proposition 3.1.11. *Constants given in Proposition 3.1.7 apply to Theorem 3.1.10 provided that Γ_f is replaced by Γ in (3.3).*

Proof. The proof is similar to the one of Proposition 3.1.7. □

Speed of convergence

In the discrete setting, Theorem 3.1.4 justifies the use of a Gaussian approximation to compute $\mathscr{A}_i(U, t, \mathsf{w})$. However this asymptotic behavior strongly relies on the increasing size of the patch domains. We define the patch size to be $|\mathsf{w}|$, the cardinality of w, and the spot size $|\mathrm{supp} f|$ to be the cardinality of the support of the spot f. The quantity of interest is the ratio $r = \frac{\text{patch size}}{\text{spot size}}$. If $r \gg 1$ then the Gaussian random field associated to f can be well approximated by a Gaussian white noise from the patch perspective. If $r \approx 1$ this approximation is not valid and the Gaussian approximation is no longer accurate, see Figure 3.5. We say that an offset t is *detected* in a Gaussian random field if $\mathscr{A}_i(U, t, \mathsf{w}) \leq a(t)$ for some threshold $a(t)$. In the experiments presented in Figure 3.6 and Table 3.1 the threshold is given by the asymptotic Gaussian inverse cumulative distribution function evaluated at some quantile. The parameters of the Gaussian random variable are given by Proposition 3.1.5. We find that except for small spot sizes and

	5	10	15	20	40	70
1	0.3	1.4	3.2	4.6	7.4	9.0
2	0.3	0.4	1.2	2.2	5.8	8.5
5	0.3	0.4	0.4	0.5	1.3	4.1
10		0.4	0.5	0.5	0.4	1.4
15			0.5	0.5	0.5	0.5
20				0.5	0.5	0.5
25					0.5	0.5

	5	10	15	20	40	70
1	18.1	11.6	10.9	10.4	10.1	10.0
2	34.2	16.5	12.8	11.5	10.4	9.9
5	93.9	49.3	30.8	20.9	13.2	11.5
10		86.7	57.6	46.0	19.7	14.5
15			83.9	63.8	30.0	18.2
20				79.5	36.7	24.7
25					51.5	26.6

Table 3.1: **Asymptotic properties** Number of detections with different patch domains from 5×5 to 70×70 and spot domains from 1×1 to 25×25 for the $s_{2,2}$ (left table) or s_{sc} (right table) auto-similarity function. We only consider patch domains larger than spot domains. We generate 5000 Gaussian random field images of size 256×256 for each setting (with spot the indicator of the spot domain). We set $\alpha = 10/256^2$. For each setting we compute $a(t)$ the inverse cumulative distribution function of $\mathsf{N}(\mu_i(t), \sigma_i^2(t))$ evaluated at quantile α, with μ_i and σ_i^2 given by Proposition 3.1.5. For each pair of patch size and spot size we compute $\sum_{t \in \mathsf{E}} \mathbb{1}_{\mathscr{A}_i(u,t,\mathsf{w}) \leq a(t)}$, namely the number of detections, for all the 5000 random fields samples. The empirical averages are displayed in the table. If $\mathscr{A}_i(u, t, \mathsf{w})$ had Gaussian distribution with parameters given by Proposition 3.1.5 then the number in each cell would be $\sum_{t \in \mathsf{E}} \mathbb{P}\left(\mathscr{A}_i(U, t, \mathsf{w}) \leq a(t)\right) \approx 10$.

large patches, *i.e.* $r \gg 1$, the approximation is not valid. More precisely, let $U = f * W$ with f a finitely supported function over $\mathbb{Z}^2$ and W a Gaussian white noise over $\mathbb{Z}^2$. Let $\mathsf{w} \subset \mathbb{Z}^2$ and let E_0 be a finite subset of $\mathbb{Z}^2$. We compute $\sum_{t \in \mathsf{E}_0} \mathbb{1}_{\mathscr{A}_i(U,t,\mathsf{w}) \leq a(t)}$, with $a(t)$ defined by the inverse cumulative distribution function of quantile $10/|\mathsf{E}_0|$ for the Gaussian $\mathsf{N}(\mu, \sigma^2)$ where μ, σ^2 are given by Theorem 3.1.4 and Proposition 3.1.5. Note that $a(t)$ would satisfy $\mathbb{P}\left(\mathscr{A}_i(U, t, \mathsf{w}) \leq a(t)\right) \approx 10/|\mathsf{E}_0|$ if the approximation for the cumulative distribution function was correct. In other words, if the Gaussian asymptotic was always valid, we would have a number of detections equal to 10 independently of r. This is clearly not the case in Table 3.1. One way to interpret this is by looking at the left tail of the approximated distribution for $s_{2,2}$ and s_{sc} on Figure 3.5. For s_{sc} the histogram is above the estimated curve, see (a) in Figure 3.6 for example. Whereas for $s_{2,2}$ the histogram is under the estimated curve. Thus for s_{sc} we expect to obtain more detections than what is predicted whereas we will observe the opposite behavior for $s_{2,2}$. This situation is also illustrated for similarities s_2 and s_{sc} in Figure 3.6 in which we compare the asymptotic cumulative distribution function with the empirical one.

In the next section we address this problem by studying non-asymptotic cases for the $s_{2,2}$ auto-similarity function in both continuous and discrete settings.

3.1.4 A non-asymptotic case: internal Euclidean matching

Discrete periodic case

In this section E is a finite rectangular domain in $\mathbb{Z}^2$. We fix $\mathsf{w} \subset \mathsf{E}$. We also define f a function over E. We consider the Gaussian random field $U = f * W$ (we consider the periodic convolution) with W a Gaussian white noise over E.

In the previous section, we derived asymptotic properties for similarity functions. However, a necessary condition for the asymptotic Gaussian approximation to be valid is for the spot size to be very small when compared to the patch size. This condition is not often met and non-asymptotic techniques must be developed. For instance it should be noted that the distribution of the s_{sc} template similarity, $\mathscr{T}_{sc}(U, y, t, \mathsf{w})$, is Gaussian for every w. We might also derive a non-asymptotic expression for the template similarity in the cosine case if the Gaussian model is a white noise model. In what fol-

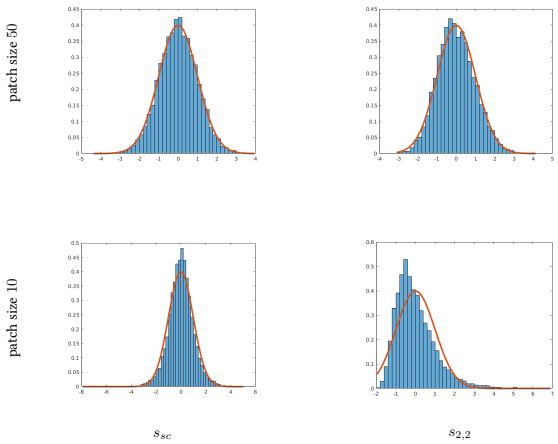
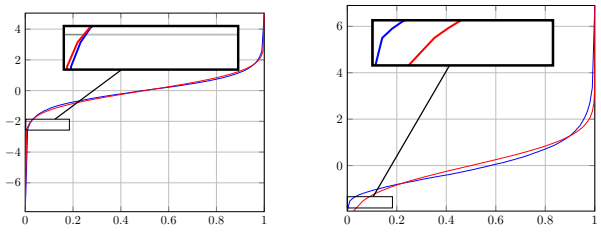


Figure 3.5: **Gaussian moment matching** In this experiment, 10^4 samples of 128×128 Gaussian images are computed with a spot of size 5×5 (the spot is the indicator of this square). Scalar product auto-similarities and squared ℓ_2 auto-similarities are computed for a fixed offset $(70, 100)$. We then plot the normalized histogram of these values. The red curve corresponds to the standard Gaussian $\mathsf{N}(0, 1)$. On the top row $r = 100 \gg 1$ and the Gaussian approximation is valid. On the bottom row $r \approx 1$ and the Gaussian approximation is not valid.



(a) s_{sc} (b) s_2

Figure 3.6: **Theoretical and empirical cumulative distribution function** This experiment illustrates the non-Gaussianity in Figure 3.5. In both cases, the red curve is the inverse cumulative distribution function of the standard Gaussian and the blue curve is the empirical inverse cumulative distribution function of normalized auto-similarity functions computed with 10^4 samples of Gaussian models. We present auto-similarity results obtained for $t = (70, 100)$ and similarity function s_{sc} (on the left) and s_2 (on the right). We note that for rare events, see the magnified region, the theoretical inverse cumulative distribution function is above the empirical inverse cumulative distribution function. The opposite behavior is observed for similarity s_2. These observations are in accordance with the findings of Table 3.1.

lows we restrict ourselves to the auto-similarity framework and consider the square of the ℓ_2 norm auto-similarity function, *i.e.* $\mathscr{A}_{2,2}(U,t,\mathsf{w})$. In this case we present an efficient method to compute the cumulative distribution function of the auto-similarity function even in the non-asymptotic case.

Proposition 3.1.12. *Let $\mathsf{E} = (\mathbb{Z}/M\mathbb{Z})^2$ with $M \in \mathbb{N}$, $\mathsf{w} \subset \mathsf{E}$, $f \in \mathbb{R}^{\mathsf{E}}$ and $U = f * W$ where W is a Gaussian white noise over E. The following equality holds for any $t \in \mathsf{E}$ up to a change of the underlying probability space*

$$\mathscr{A}_{2,2}(U,t,\mathsf{w}) \underset{a.s}{=} \sum_{k=0}^{|\mathsf{w}|-1} \lambda_k(t,\mathsf{w}) Z_k \ , \tag{3.6}$$

with Z_k independent chi-square random variables with parameter 1 and $\lambda_k(t,\mathsf{w})$ the eigenvalues of the covariance matrix C_t associated with function $\Delta_f(t,\cdot)$, see (3.3), restricted to w, i.e for any $p_1, p_2 \in \mathsf{w}$, $C_t(p_1,p_2) = \Delta_f(t, p_1 - p_2)$.

Proof. Let $t \in \mathsf{E}$ and V_t be given for any $p \in \mathsf{E}$ by $V_t(p) = U(p) - U(p+t)$. It is a Gaussian vector with mean 0 and covariance matrix C_V given for any $p_1, p_2 \in \mathsf{E}$ by

$$C_V(p_1,p_2) = 2\Gamma_f(p_1-p_2) - \Gamma_f(p_1-p_2-t) - \Gamma_f(p_1-p_2+t) = \Delta_f(t,p_1-p_2) \ .$$

The covariance of the random field $P_{\mathsf{w}}(V_t)$, the restriction of V_t to w, is given by the restriction of C_V to w. This new covariance matrix, C_t, is symmetric and the spectral theorem ensures that there exists an orthonormal basis $\mathscr{B}$ such that C_t is diagonal when expressed in $\mathscr{B}$. Thus we obtain that $P_{\mathsf{w}}(V_t) = \sum_{e_k \in \mathscr{B}} \langle P_{\mathsf{w}}(V_t), e_k \rangle e_k$. It is clear that, for any $k \in [\![0, |\mathsf{w}| - 1]\!]$, $\langle P_{\mathsf{w}}(V_t), e_k \rangle$ is a Gaussian random variable with mean 0 and variance $e_k^T C_t e_k = \lambda_k(t,\mathsf{w}) \geq 0$. We set $\mathsf{K} = \{k \in [\![0, |\mathsf{w}|-1]\!],\ \lambda_k(t,\mathsf{w}) \neq 0\}$ and define X a random vector in $\mathbb{R}^{|\mathsf{w}|}$ such that

$$X_k = \lambda_k(t,\mathsf{w})^{-1/2} \langle P_{\mathsf{w}}(V_t), e_k \rangle, \text{ if } k \in \mathsf{K} \ , \qquad \text{and } X_{\mathsf{K}_-} = Y \ ,$$

where X_{K_-} is the restriction of X to the indices of $\mathsf{K}_- = [\![0, |\mathsf{w}| - 1]\!] \backslash \mathsf{K}$ and Y is a standard Gaussian random vector on $\mathbb{R}^{|\mathsf{K}_-|}$ independent from the sigma field generated by $\{(X_k), k \in \mathsf{K}\}$. By construction we have $\mathbb{E}[X_k X_\ell] = 0$ if $\ell \in \mathsf{K}$ and $k \in \mathsf{K}_-$, or $\ell \in \mathsf{K}_-$ and $k \in \mathsf{K}_-$. Suppose now that $k, \ell \in \mathsf{K}$. We obtain that

$$\mathbb{E}[X_k X_\ell] = \lambda_k(t,\mathsf{w})^{-1/2} \lambda_\ell^{-1/2}(t,\mathsf{w}) \mathbb{E}\left[e_k^T C_t e_\ell\right] = 0 \ .$$

Thus X is a standard Gaussian random vector and we have $P_{\mathsf{w}}(V_t) = \sum_{k=0}^{|\mathsf{w}|-1} \lambda_k^{1/2}(t,\mathsf{w}) X_k e_k$, where the equality holds almost surely. We get that

$$\mathscr{A}_{2,2}(U,t,\mathsf{w}) = \|P_{\mathsf{w}}(V_t)\|_2^2 = \sum_{e_k \in \mathscr{B}} \langle P_{\mathsf{w}}(V_t), e_k \rangle^2 = \sum_{k=0}^{|\mathsf{w}|-1} \lambda_k(t,\mathsf{w}) X_k^2 \ .$$

Setting $Z_k = X_k^2$ concludes the proof. □

Note that if $\mathsf{w} = \mathsf{E}$ then we obtain that the covariance matrix C_t is block-circulant with circulant blocks and the eigenvalues are given by the discrete Fourier transform.

In order to compute the true cumulative distribution function of the auto-similarity square ℓ_2 norm we need to: 1) compute the eigenvalues of a covariance matrix in $\mathcal{M}_{|\mathsf{w}|}(\mathbb{R})$; 2) compute the cumulative distribution function of a positive-weighted sum of independent chi-square random variable with weights given by the computed eigenvalues. Storing all covariance matrices for each offset t is not feasible. For instance considering a patch of size 10×10 and an image of size 512×512 we have approximately 2.6×10^9 coefficients to store, *i.e.* 10.5GB in float precision. In the rest of the section we

suppose that t and w are fixed and we denote by C_t the covariance matrix associated to the restriction of $\Delta_f(t,\cdot)$ to $\mathsf{w}+(-\mathsf{w})$. In Proposition 3.1.13 we propose a method to approximate the eigenvalues of C_t by using its specific structure. Indeed, as a covariance matrix, C_t is symmetric and positive and, since its associated Gaussian random field is stationary, it is block-Toeplitz with Toeplitz blocks, *i.e.* is block-diagonally constant and each block has constant diagonals. In the one-dimensional case these properties translate into symmetry, positivity and Toeplitz properties of the covariance matrix. Proposition 3.1.13 is stated in the one-dimensional case for the sake of simplicity but two-dimensional analogous can be derived. Note that this approximation is not always sharp as shown in Figure 3.7.

We recall that the Frobenius norm of a matrix of size $n \times n$ is the ℓ_2 norm of the associated vector of size n^2.

Proposition 3.1.13. *Let b be a function defined over $[\![-(n-1), n-1]\!]$ with $n \in \mathbb{N}\setminus\{0\}$. We define $T_b(j,\ell) = b(j-\ell)$ for $j,\ell \in [\![0, n-1]\!]$. The matrix T_b is a circulant matrix if and only if b is n-periodic. T_b is symmetric if and only if b is symmetric. Let b be symmetric, defining $\Pi(T_b)$ the projection of T_b onto the set of symmetric circulant matrix for the Frobenius product, we obtain that*

1. *the projection satisfies $\Pi(T_b) = T_c$ with $c(j) = (1-j/n)\, b(j) + (j/n) b(n-j)$ for all $j \in [\![0, n-1]\!]$ and c is extended by n-periodicity to $\mathbb{Z}$;*

2. *the eigenvalues of $\Pi(T_b)$ are given by $(2\,\mathrm{Re}(\hat{d}(j)) - b(0))_{j\in[\![0,n-1]\!]}$ with $d(j) = (1-j/n)\, b(j)$, and $\hat{d}$ is the discrete Fourier transform over $[\![0, n-1]\!]$;*

3. *let $(\lambda_j)_{j\in[\![1,n]\!]}$ be the sorted eigenvalues of T_b and $(\tilde{\lambda}_j)_{j\in[\![1,n]\!]}$ the sorted eigenvalues of $\Pi(T_b)$ (in the same order). For any $j \in [\![1, n]\!]$, we have $|\lambda_j - \tilde{\lambda}_j| \leq \|T_b - \Pi(T_b)\|_{\mathrm{Fr}}$;*

4. *if T_b is positive-definitive then $\Pi(T_b)$ is positive-definite.*

Proof. 1. Let T_c be an element of the symmetric circulant matrices set. Minimizing $\|T_b - T_c\|_{\mathrm{Fr}}^2$ in $c(j)_{j\in[\![0,n-1]\!]}$ we get that $c(j)$ satisfies for any $j \in [\![0, n-1]\!]$

$$c(j) = \operatorname*{argmin}_{s\in\mathbb{R}} \left(2(n-j)(s-b(j))^2 + 2j(s-b(n-j))^2\right) ,$$

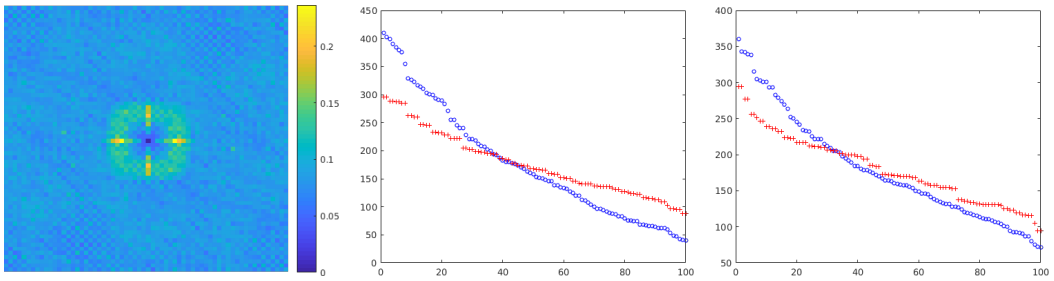
which gives the result.

2. Since $T_c = \Pi(T_b)$ is circulant, its eigenvalues are given by the discrete Fourier transform of c. We have that if $i \neq 0$ then $c(i) = \dot{d}(j) + \dot{d}(-j)$ with $d(j) = (1-j/n)\, b(j)$ and $\dot{d}$ its extension to $\mathbb{Z}$ by n-periodicity. We also have $c(0) = b(0)$. We conclude the proof by taking the discrete Fourier transform of c.

3. The proof of the Lipschitz property on the sorted eigenvalues of symmetric matrices with respect to the ℓ_2 matricial norm can be found in [Cia82]. We conclude using the fact that the ℓ_2 matricial norm is upper-bounded by the Frobenius norm.

4. This result is a special case of the spectrum contraction property of the projection [CJY91, Theorem 2].

□



(a) (b) (c)

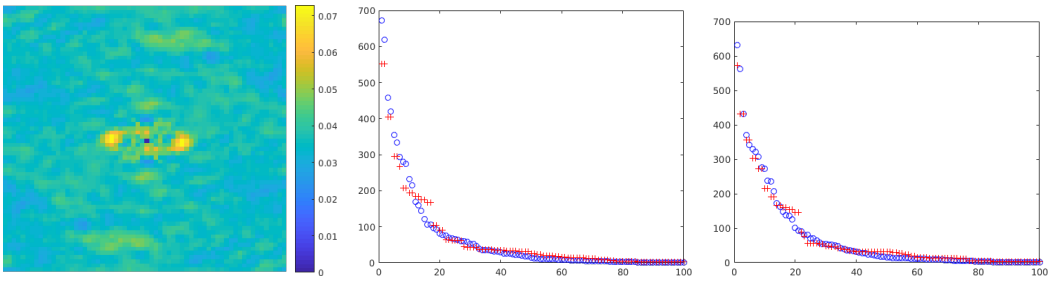
Figure 3.7: **Eigenvalues approximation** We consider a Gaussian random field generated with $f * W$ with W a Gaussian white noise and f is a fixed sample of an independent Gaussian white noise over E. We consider patches of size 10×10 and study the approximation of the eigenvalues for the covariance matrix of the random field restricted to a domain of size 10×10, similarly to Proposition 3.1.12. (A) shows the Normalized Root-Mean Square Deviation between the eigenvalues computed with standard routines and the ones given by the approximation for each offset, see (3.7). Offset zero is at the center of the image. (B) and (C) illustrate the properties of Proposition 3.1.13. Blue circles correspond to the 100 eigenvalues computed with MATLAB routine for offset $(5, 5)$ in (B), respectively $(10, 20)$ in (C), and red crosses correspond to the 100 approximated eigenvalues for the same offsets. Note that a standard routine takes $273s$ for 10×10 patches on 256×256 images whereas it only takes $1.11s$ when approximating the eigenvalues using the discrete Fourier transform.

In Figure 3.7 we display the behavior of the projection for the eigenvalues in the two-dimensional case. The measure we consider is the Normalized Root Mean Square Deviation

$$\mathrm{NRMSD} = \frac{\left(|\mathsf{w}|^{-1} \sum_{k=0}^{|\mathsf{w}|-1} |\tilde{\lambda}_k(t, \mathsf{w}) - \lambda_k(t, \mathsf{w})|^2 \right)^{1/2}}{\max \left(\lambda_k(t, \mathsf{w})\right)_{k \in [\![0, |\mathsf{w}|-1]\!]} - \min \left(\lambda_k(t, \mathsf{w})\right)_{k \in [\![0, |\mathsf{w}|-1]\!]}}, \tag{3.7}$$

with $\tilde{\lambda}_k(t, \mathsf{w})$ the approximation of the eigenvalues, for every possible offset in the image and $\lambda_k(t, \mathsf{w})$ the true eigenvalues, for every possible offset. Computing the eigenvalues of the projection is done via Fast Fourier Transform (FFT) which is faster than standard routines for computing eigenvalues of Toeplitz matrices. The major cons of using such an approximation is that it may not be valid for small offsets $t \in \mathsf{E}$ as shown in Figure 3.7. However, in most cases the random field is smooth and in this case, see Figure 3.8, the approximation is satisfactory. We also highlight that for similarity detection purposes, see Figure 3.9, the level of precision achieved by our approximation is satisfactory, see [Bor+19].

Suppose the approximation of the eigenvalues is valid, we need an efficient algorithm to compute the distribution of the associated positive-weighted sum of chi-square random variables in (3.6). Exact computation has been derived by Imhof in [Imh61] but requires to compute heavy integrals. This exact method, named Imhof method in the following, will be used as a baseline for other algorithms. Numerous methods such as differential equations [Dav77], series truncation [KJB67], negative binomial mixtures [OZ81] approaches were later introduced but all require stopping criteria such as truncation criteria which can be hard to set. We focus on cumulant methods which generalize and refine the Gaussian approximations used in Section 3.1.3. These methods rely on computing moments of the original distribution and then fitting a known probability distribution function to the objective distribution using these moments. Bodenham et al. in [BA16] show that the following methods can be efficiently computed:



(a) (b) (c)

Figure 3.8: **Eigenvalues approximation** Same study as the one conducted in Figure 3.7 with $f = \mathbb{1}_{\{1,2,3\}^2}$. Note that in this case the approximation is better than the one presented in Figure 3.7.

- Gaussian approximation (discarded due to its poor results for small patches as illustrated in Section 3.1.3),
- Hall-Buckley-Eagleson [Hal83; BE88] (HBE), (three moments fitted Gamma distribution),
- Wood F [Woo89] (three moments fitted Fischer-Snedecor distribution).

Other methods such as the Lindsay-Pilla-Basak-4 method, which relies on the computation of eight moments, are slower than HBE by a factor 350 at least, see [BA16], and are thus discarded. In Figure 3.9 we investigate the trade-off between computational speed and accuracy of these methods for the task of detection.

The experiments conducted in Figure 3.9 show that the HBE approximation does not give good results when evaluating the probability of rare events. This was already noticed by Bodenham et al. in [BA16] who stated that "Hall–Buckley–Eagleson method is recommended for most practitioners [...]. However, [...], for very small probability values, either the Wood F or the Lindsay–Pilla–Basak method should be used".

Continuous periodic case

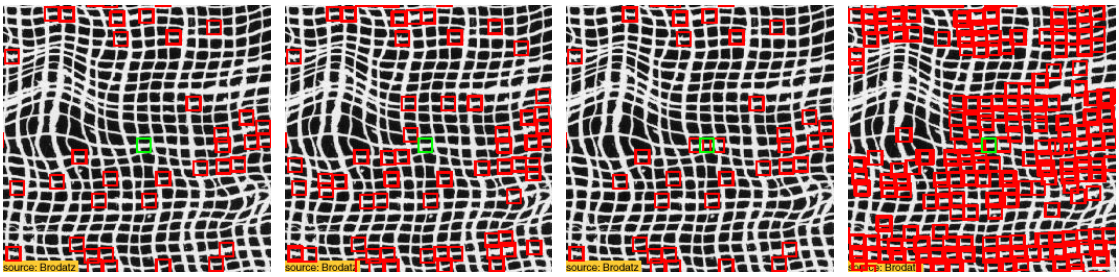
To conclude we show that a similar non-asymptotic study can be conducted in continuous settings.

Proposition 3.1.14. *Let* $\mathsf{E} = \mathbb{T}^2$, $\mathsf{w} \subset \mathsf{E}$ *and let* U *be a zero-mean Gaussian random field on* E *with covariance function* Γ. *Assume* **A2**, *then the following equality holds for any* $t \in \mathsf{E}$ *up to a change of the underlying probability space*

$$\mathscr{A}_{2,2}(U, t, \mathsf{w}) \underset{a.s}{=} \sum_{k \in \mathbb{N}} \lambda_k(t, \mathsf{w}) Z_k,$$

with Z_k *independent chi-square random variables with parameter 1 and* $\lambda_k(t, \mathsf{w})$ *the eigenvalues of the kernel* C_t *associated with function* $\Delta(t, \cdot) = 2\Gamma(t) - \Gamma(\cdot + \mathrm{t}) - \Gamma(\cdot - \mathrm{t})$ *restricted to* w, *i.e. for any* $p_1, p_2 \in \mathsf{w}$, $C_t(p_1, p_2) = \Delta(t, p_1 - p_2)$.

Proof. We consider the stationary Gaussian random field $P_{\mathsf{w}}(V_t)$ over w defined by the restriction to w where for any $p \in \mathsf{E}$ by $V_t(p) = U(p) - U(p + t)$. The Karhunen-Loeve theorem [GS91] ensures the



(a) $1513s, n_d = 52$ (b) $200s, n_d = 116$ (c) $4.77s, n_d = 50$ (d) $4.64s, n_d = 671$

Figure 3.9: **Similarity detection** In this figure we illustrate the accuracy of the different proposed approximations of the cumulative distribution function of $\mathscr{A}_{2,2}(U, t, \mathsf{w})$. We say that an offset t is *detected* in an image if $\mathscr{A}_{2,2}(u, t, \mathsf{w}) \leq a(t)$ for some threshold $a(t) \in \mathbb{R}$. In every image, in green we display the patch domain w (in the center of the image) and in red we display the shifted patch domain for detected offsets with function $a(t)$ such that for any $t \in \mathsf{E}$, $\mathbb{P}\left(\mathscr{A}_{2,2}(U, t, \mathsf{w}) \leq a(t)\right) = 1/256^2$, where U is given by the Gaussian random field $f * W$ where f is the original image of fabric and W is a Gaussian white noise over $\mathsf{E} = 256 \times 256$. Approximations of the cumulative distribution function of $\mathscr{A}_{2,2}(U, t, \mathsf{w})$ lead to approximations of $a(t)$. The most precise approximation is given in (A) where the eigenvalues are computed using a MATLAB routine and the cumulative distribution function is given by the Imhof method. In (B) we approximate the eigenvalues using the projection described in Proposition 3.1.13 and still use the Imhof method. It yields twice as many detections. In (C) Wood F method is used instead of Imhof's yielding less detections but performing seven times faster. Interestingly errors seem to compensate and the obtained result with Wood F method is very close to the results obtained with the baseline algorithm in (A). In (D) HBE method is used instead of Imhof's, in this case we obtain too many detections, *i.e.* the approximation of the cumulative distribution function is not valid.

existence of $(\lambda_k(t,\mathsf{w}))_{k\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$, $(X_k)_{k\in\mathbb{N}}$ a sequence of independent unit Gaussian random variables and $(e_k)_{k\in\mathbb{N}}$ a sequence of orthonormal function over $\mathrm{L}^2(\mathsf{w})$ such that

$$\lim_{n\to+\infty} \sup_{p\in\mathsf{w}} \mathbb{E}\left[\left|P_{\mathsf{w}}(V_t)(p) - \sum_{k=0}^{n} \sqrt{\lambda_k(t,\mathsf{w})} e_k(p) X_k\right|^2\right] = 0 \ , \tag{3.8}$$

We define the sequence $(I_n)_{n\in\mathbb{N}} = (\int_{\mathsf{w}} (\sum_{k=0}^{n} \sqrt{\lambda_k(t,\mathsf{w})} e_k(p) X_k)^2 \mathrm{d}p)_{n\in\mathbb{N}}$. We have, using the Cauchy-Schwarz inequality on $\mathrm{L}^2(\Omega\times\mathsf{w})$ and (3.8)

$$\begin{aligned}
\mathbb{E}\left[|\mathscr{A}_{2,2}(U,t,\mathsf{w}) - I_n|\right] &\leq \mathbb{E}\left[\int_{\mathsf{w}} \left|P_{\mathsf{w}}(V_t)^2(p) - \left(\sum_{k=0}^{n} \sqrt{\lambda_k(t,\mathsf{w})} e_k(p) X_k\right)^2\right| \mathrm{d}p\right] \\
&\leq \mathbb{E}\left[\int_{\mathsf{w}} (P_{\mathsf{w}}(V_t)(p) - \sum_{k=0}^{n} \sqrt{\lambda_k(t,\mathsf{w})} e_k(p) X_k)^2 \mathrm{d}p\right]^{1/2} \\
&\quad \times \mathbb{E}\left[\int_{\mathsf{w}} (P_{\mathsf{w}}(V_t)(p) + \sum_{k=0}^{n} \sqrt{\lambda_k(t,\mathsf{w})} e_k(p) X_k)^2 \mathrm{d}p\right]^{1/2} \\
&\leq 2\mathbb{E}\left[\mathscr{A}_{2,2}(U,t,\mathsf{w})\right]^{1/2} \int_{\mathsf{w}} \mathbb{E}\left[(P_{\mathsf{w}}(V_t)(p) - \sum_{k=0}^{n} \sqrt{\lambda_k(t,\mathsf{w})} e_k(p) X_k)^2\right] \mathrm{d}p \ ,
\end{aligned} \tag{3.9}$$

where we used the Fubini theorem in the last inequality. Using the dominated convergence theorem in (3.9) with domination given by

$$\sup_{n\in\mathbb{N}} \sup_{p\in\mathsf{w}} \mathbb{E}[(P_{\mathsf{w}}(V_t)(p) - \sum_{k=0}^{n} \sqrt{\lambda_k(t,\mathsf{w})} e_k(p) X_k)^2]$$

we conclude that $(I_n)_{n\in\mathbb{N}}$ converges to $\mathscr{A}_{2,2}(U,t,\mathsf{w})$ in $\mathrm{L}^1(\Omega)$. Thus there exists a subsequence of $(I_n)_{n\in\mathbb{N}}$ which converges almost surely to $\mathscr{A}_{2,2}(U,t,\mathsf{w})$. We also have that

$$I_n = \int_{\mathsf{w}} (\sum_{k=0}^{n} \sqrt{\lambda_k(t,\mathsf{w})} e_k(p) X_k)^2 \mathrm{d}p = \sum_{k=0}^{n} \lambda_k(\mathsf{w},k) X_k^2 \ ,$$

by orthonormality and thus the sequence $(I_n)_{n\in\mathbb{N}}$ is almost surely non-decreasing. We get that $(I_n)_{n\in\mathbb{N}}$ converges almost surely to $\mathscr{A}_{2,2}(U,t,\mathsf{w})$ which can be rewritten as

$$\mathscr{A}_{2,2}(U,t,\mathsf{w}) = \sum_{k\in\mathbb{Z}} \lambda_k(t,\mathsf{w}) X_k^2 \quad \text{a.s..}$$

The characterization of $(\lambda_k(t,\mathsf{w}), e_k(p))$ is given by the Karhunen-Loeve theorem and $e_k(p)$ is solution of the following Fredholm equation for all $p \in \mathsf{w}$

$$\int_{\mathsf{w}} \Delta(t, p-p_2) e_k(p_2) \, \mathrm{d}p_2 = \lambda_k(t,\mathsf{w}) e_k(p) \ .$$

Setting $Z_k = X_k^2$ concludes the proof. □

Note that if $\mathsf{w} = \mathbb{T}^2$ then the solution of the Fredholm equation is given by the Fourier series of Γ.

3.1.5 Technical results

Multidimensional Central Limit Theorems

In this section we provide an extension of [Jan88, Theorem 2] to the multidimensional case.

We recall the notion of dependency graph as introduced in [Jan88]. Let $(X_i)_{i\in\mathbb{N}}$ be $\mathbb{R}^d$-valued random variables. A graph is a dependency graph for $(X_i)_{i\in\mathbb{N}}$ if the two following conditions are satisfied.

1. There is a one-to-one correspondence between $(X_i)_{i\in\mathbb{N}}$ and the vertices of the graph.
2. If two sets of vertices are not connected then the corresponding random variables are independent.

Theorem 3.1.15. *Let $(X_{i,j})_{(i,j)\in\mathbb{N}^2}$ be a sequence of $\mathbb{R}^d$-valued random variables and $(N_n)_{n\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$. For any $n \in \mathbb{N}$, assume that there exists $A_n, M_n \geq 0$ such that for any $j \in \mathbb{N}$, $\|X_{n,j}\| \leq A_n$ and that the dependency graph of $(X_{n,j})_{j\in\mathbb{N}}$ is of degree M_n at most. For any $n \in \mathbb{N}$ let $S_n = \sum_{j=1}^{N_n} X_{n,j}$ and $C_n = \operatorname{Cov}[S_n]$. Assume that there exists $m_0 \in \mathbb{N}$ and $C \in \mathcal{M}_d(\mathbb{R})$ such that for any $n \in \mathbb{N}$, $\lim_{n\to+\infty}(N_n/M_n)^{1/m_0} M_n A_n = 0$ and $\lim_{n\to+\infty} C_n = C$. Then, $S_n - \mathbb{E}[S_n]$ converges (in the weak sense) towards $\mathsf{N}(0, C)$.*

Proof. Let $a \in \mathbb{R}^d$ and consider $(X^a_{i,j})_{(i,j)\in\mathbb{N}^2}$ such that for any $i, j \in \mathbb{N}$, $X^a_{i,j} = \langle X_{i,j}, a\rangle$. We also introduce for any $n \in \mathbb{N}$, $S^a_n = \sum_{j=1}^{N_n} x^a_{n,j}$. Assume that $a^\top C a = 0$. Then, using the Bienaymé-Tchebychev inequality, we have for any $\varepsilon > 0$

$$\lim_{n\to+\infty} \mathbb{P}\left(|S^a_n - \mathbb{E}[S^a_n]| > \varepsilon\right) \leq \lim_{n\to+\infty} \varepsilon^{-2} a^\top C_n a = 0 \,.$$

Hence, $\langle a, S_n - \mathbb{E}[S_n]\rangle$ converges (in the weak sense) towards $\langle a, Z\rangle$ with Z a d-dimensional Gaussian random variable with zero mean and covariance matrix C. If $a^\top C_n a \neq 0$ then using [Jan88] we have that $\langle a, S_n - \mathbb{E}[S_n]\rangle$ converges (in the weak sense) towards $\langle a, Z\rangle$. We conclude using the Cramér-Wold theorem, [CW36, Theorem 1]. □

Similarly to [Jan88, Theorem 2], we can replace the condition $\|X_{n,i}\| \leq A_n$ by the following condition: for any $a \in \mathbb{R}^d$ with $a \neq 0$

$$\lim_{n\to+\infty} M_n \sum_{j=1}^{N_n} \mathbb{E}\left[\|X_{n,j}\|^2 \mathbb{1}_{\|X_{n,j}\| > A_n \|a\|}\right] = 0 \,.$$

Indeed, this implies that for any $a \in \mathbb{R}^d$, $\lim_{n\to+\infty} M_n \sum_{j=1}^{N_n} \mathbb{E}[\langle a, X_{n,j}\rangle^2 \mathbb{1}_{|\langle a, X_{n,j}\rangle| > A_n}] = 0$, which is the Lindeberg type condition identified in [Jan88, Theorem 2].

Asymptotic theorems – discrete case

We start by introducing two notions which will be crucial in order to derive a law of large numbers and a central limit theorem in broad settings. The R-independence, see Definition 3.1.16, ensures long-range independence whereas stochastic domination will replace integrability conditions in the standard law of large numbers or central limit theorem.

The notion of R-independence generalizes to $\mathbb{R}^2$ and $\mathbb{Z}^2$ the associated one-dimensional concept, see [Bil95] and its extension to $\mathbb{N}^2$ [ST11], [MST08].

Definition 3.1.16. *Let $d \in \mathbb{N}$, $\mathsf{E} = \mathbb{R}^2$ or $\mathsf{E} = \mathbb{Z}^2$ and V be a d-dimensional random field over E. Let $\mathsf{K}_1, \mathsf{K}_2 \subset \mathsf{E}$ be two compact sets, and $V|_{\mathsf{K}_i}$ be the restriction of V to K_i, $i \in \{1, 2\}$. We say that V is R-independent, with $R \geq 0$, if $V|_{\mathsf{K}_1}$ is independent from $V|_{\mathsf{K}_2}$ as soon as $d_\infty(\mathsf{K}_1, \mathsf{K}_2) = \min_{p_1 \in \mathsf{K}_1, p_2 \in \mathsf{K}_2} \|p_1 - p_2\|_\infty > R$.*

Note that in the case of $\mathsf{E} = \mathbb{Z}^2$, compacts sets K_1 and K_2 are finite sets of indices. This notion of R-independence will replace the traditional assumption of independence in asymptotic theorems.

Definition 3.1.17. *Let $\mathsf{E} = \mathbb{R}^2$ or $\mathsf{E} = \mathbb{Z}^2$ and let $V, \widetilde{V}$ be real random fields over E. We say that:*

(a) $\widetilde{V}$ uniformly stochastically dominates V if for any $\alpha \geq 0$ and $p \in \mathsf{E}$, $\mathbb{P}(V(p) \geq \alpha) \leq \mathbb{P}(\widetilde{V}(p) \geq \alpha)$;

(b) $\widetilde{V}$ uniformly a.s. dominates V if for any $p \in \mathsf{E}$, $V(p) \leq \widetilde{V}(p)$ a.s..

Note that if $\widetilde{V}$ uniformly a.s. dominates V then $\widetilde{V}$ uniformly stochastically dominates $\widetilde{V}$.

The following theorem is a two-dimensional law of large numbers with weak dependence assumptions. It is a slight modification of Corollary 4.1 (ii) in [ST11].

Theorem 3.1.18. *Let $d \in \mathbb{N}$. Let $(m_k)_{k\in\mathbb{N}}, (n_k)_{k\in\mathbb{N}}$ be two positive increasing integer sequences and $(\mathsf{w}_k)_{k\in\mathbb{N}}$ be the sequence of subsets such that for any $k \in \mathbb{N}$, $\mathsf{w}_k = [\![0, m_k]\!] \times [\![0, n_k]\!]$. Let V be a d-dimensional R-independent random field over $\mathbb{Z}^2$, with $R \geq 0$, such that $\|V(p) - \mathbb{E}[V(p)]\|$ is uniformly stochastically dominated by $\widetilde{V}$, a real second-order stationary random field over $\mathbb{Z}^2$. Then V is a second-order random field. In addition, assume that there exists $\mu \in \mathbb{R}^d$ such that $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \sum_{p\in\mathsf{w}_k} \mathbb{E}[V(p)] = \mu$. Then it holds that*

$$\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \sum_{p\in\mathsf{w}_k} V(p) \underset{a.s}{=} \mu \,. \tag{3.10}$$

Proof. Without loss of generality we can assume that $d = 1$ and that for any $p \in \mathbb{Z}^2$, $\mathbb{E}[V(p)] = 0$. In order to apply Corollary 4.1 (ii) in [ST11] we must check that:

(a) V is $R-$independent ;

(b) $|V|$ is uniformly stochastically dominated by a random field $\widetilde{V}$ and there exists $r \in [1, 2[$ such that for any $p \in \mathbb{Z}^2$, $\mathbb{E}[\widetilde{V}^r(p) \log^+(\widetilde{V}(p))]$ is finite.

First, (a) is given in the statement of Theorem 3.1.18 and $|V|$ is uniformly stochastically dominated by the random field $\widetilde{V}_0$ defined for any $p \in \mathbb{Z}^2$ by $\widetilde{V}_0(p) = \widetilde{V}(0)$. Since $\mathbb{E}[\widetilde{V}(0)^2]$ is finite so is $\mathbb{E}[\widetilde{V}(0) \log^+(\widetilde{V}(0))]$ which implies (b). Then it holds that

$$\lim_{k\to+\infty} \sum_{p\in\mathsf{w}_k} (V(p) - \mathbb{E}[V(p)]) \underset{a.s}{=} 0 \,.$$

Using that $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \sum_{p\in\mathsf{w}_k} \mathbb{E}[U(p)] = \mu$, we conclude the proof. □

We now turn to an extension of the central limit theorem to two-dimensional random fields with weak dependence assumptions. This result is a consequence of Theorem 3.1.15.

Theorem 3.1.19. *Under the hypotheses of Theorem 3.1.18 and assuming that there exist $\mu \in \mathbb{R}^d$ and $C \in \mathcal{M}_d(\mathbb{R})$ such that*

(a) $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1/2} \sum_{p\in\mathsf{w}_k} (\mathbb{E}[V(p)] - \mu) = 0$;

(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \sum_{p_1,p_2\in\mathsf{w}_k} \mathrm{Cov}[V(p_1), V(p_2)] = C$.

Then it holds that

$$\lim_{k\to+\infty} |\mathsf{w}_k|^{-1/2} \sum_{p\in\mathsf{w}_k} (V(p) - \mu) \underset{\mathcal{L}}{=} \mathsf{N}(0, C) \ . \tag{3.11}$$

Proof. For any $i, j \in \mathbb{N}$, let $X_{i,j} = (V(p_j) - \mathbb{E}[V(p_j)])|\mathsf{w}_i|^{-1/2}$ with $(p_j)_{j\in\mathbb{N}}$ such that for any $k \in \mathbb{N}$, $\{V(p_j),\ j \in [\![1, |\mathsf{w}_k|]\!]\} = \{V(p), p \in \mathsf{w}_k\}$. For any $n \in \mathbb{N}$, let $N_n = |\mathsf{w}_n|$. Then, we have that for any $n \in \mathbb{N}$, $\sum_{j=1}^{N_n} X_{n,j} = |\mathsf{w}_k|^{-1/2} \sum_{p\in\mathsf{w}_k} (V(p) - \mathbb{E}[V(p)])$. Since V is R-independent each vertex of the dependency graph of $(X_{i,j})_{i,j\in\mathbb{N}^2}$ has its degree bounded by $(2R+1)^2$ and therefore for any $n \in \mathbb{N}$, $M_n = (2R+1)^2$. For any $n \in \mathbb{N}$, let $A_n = |\mathsf{w}_n|^\alpha$ with $\alpha \in (1/3, 1/2)$. Using that $\tilde{V}$ uniformly stochastically dominates $(\|V(p) - \mathbb{E}[V(p)]\|)_{p\in\mathbb{Z}^2}$ we obtain that for any $a \in \mathbb{R}^d$

$$\begin{aligned}
&\sum_{j=1}^{N_n} \mathbb{E}\left[\|X_{n,j}\|^2 \mathbb{1}_{\|X_{n,j}\|^2 > A_n^2\|a\|^{-2}}\right] \\
&\quad = |\mathsf{w}_n|^{-1} \sum_{p\in\mathsf{w}_n} \mathbb{E}\left[\|V(p) - \mathbb{E}[V(p)]\|^2 \mathbb{1}_{\|V(p)-\mathbb{E}[V(p)]\|^2 > A_n^2\|a\|^{-2}|\mathsf{w}_n|}\right] \\
&\quad = |\mathsf{w}_n|^{-1} \sum_{p\in\mathsf{w}_n} \int_0^{+\infty} \mathbb{P}\left(\|V(p) - \mathbb{E}[V(p)]\|^2 \geq \max(A_n^2\|a\|^{-2}|\mathsf{w}_n|, t)\right) \mathrm{d}t \\
&\quad \leq |\mathsf{w}_n|^{-1} \sum_{p\in\mathsf{w}_n} \int_0^{+\infty} \mathbb{P}\left(\tilde{V}(p) \geq \max(A_n^2\|a\|^{-2}|\mathsf{w}_n|, t)\right) \mathrm{d}t \\
&\quad \leq \mathbb{E}\left[\tilde{V}(0) \mathbb{1}_{\tilde{V}(0) > A_n^2\|a\|^{-2}|\mathsf{w}_n|}\right] \ .
\end{aligned}$$

Hence, since $\lim_{n\to+\infty} A_n^2|\mathsf{w}_n| = \lim_{n\to+\infty} |\mathsf{w}_n|^{1-2\alpha} = +\infty$ we get that

$$\lim_{n\to+\infty} \sum_{j=1}^{N_n} \mathbb{E}\left[\|X_{n,j}\|^2 \mathbb{1}_{\|X_{n,j}\|^2 > A_n^2\|a\|^{-2}}\right] = 0 \ . \tag{3.12}$$

Letting $m_0 = 3$ we get that

$$\lim_{n\to+\infty} (N_n/M_n)^{1/m_0} M_n A_n = (2R+1)^{2(1+1/3)} |\mathsf{w}_n|^{1/3-\alpha} = 0 \ . \tag{3.13}$$

In addition, we have that for any $n \in \mathbb{N}$

$$C_n = \mathrm{Cov}\left[\sum_{j=1}^{N_n} X_{n,j}\right] = |\mathsf{w}_n|^{-1} \mathrm{Cov}\left[\sum_{p\in\mathsf{w}_n} V(p)\right] = |\mathsf{w}_n|^{-1} \sum_{p_1,p_2\in\mathsf{w}_n} \mathrm{Cov}[V(p_1), V(p_2)] \ . \tag{3.14}$$

Hence, combining (3.12), (3.13), (3.14), (b) and Theorem 3.1.15, we get that

$$\lim_{k\to+\infty} |\mathsf{w}_k|^{-1/2} \sum_{p\in\mathsf{w}_k} (V(p) - \mathbb{E}[V(p)]) \underset{\mathcal{L}}{=} \mathsf{N}(0, C) \ . \tag{3.15}$$

Combining (3.15) and (a) concludes the proof. □

The following lemma explicits a class of Gaussian random fields over $\mathbb{Z}^2$ such that the R-independence property holds for some $R \geq 0$.

Lemma 3.1.20. *Let $f \in \mathbb{R}^{\mathbb{Z}^2}$ with finite support $\mathrm{supp} f \subset [\![-r, r]\!]^2$, where $r \in \mathbb{N}$. Let W be a Gaussian white noise over $\mathbb{Z}^2$ and $V = f * W$ then V is a R-independent second-order random field with $R = 2r$.*

Proof. V is a Gaussian random field such that for any $p_1, p_2 \in \mathbb{Z}^2$

$$\begin{aligned} &\mathbb{E}\left[V(p_1)\right] = 0 \ , \\ &\mathrm{Cov}\left[V(p_1), V(p_2)\right] = \sum_{p_1', p_2' \in \mathbb{Z}^2} f(p_1 - p_1') f(p_2 - p_2') \, \mathrm{Cov}\left[W(p_1'), W(p_2')\right] = \Gamma_f(p_1 - p_2) \ . \end{aligned} \tag{3.16}$$

Note that since $\mathrm{supp} f \subset [\![-r, r]\!]$ we have $\mathrm{supp} \Gamma_f \subset [\![-R, R]\!]$ with $R = 2r$. For any $p_1, p_2 \in \mathbb{Z}^2$ such that $\|p_1 - p_2\|_\infty > R$, using (3.16), we obtain

$$\mathrm{Cov}\left[V(p_1), V(p_2)\right] = \Gamma_f(p_1 - p_2) = 0 \ . \tag{3.17}$$

Let $\mathsf{K}_1, \mathsf{K}_2 \subset \mathbb{Z}^2$ two finite sets with $\sup_{p \in \mathsf{K}_1, p_2 \in \mathsf{K}_2} \|p - p_2\|_\infty > R$ and consider $V|_{\mathsf{K}_i}$ the restriction of V to K_i for $i = \{1, 2\}$. Using (3.17), we get that for any $p \in \mathsf{K}_1$, $p_2 \in \mathsf{K}_2$ we have

$$\mathrm{Cov}\left[V|_{\mathsf{K}_1}(p), V|_{\mathsf{K}_2}(p_2)\right] = 0 \ .$$

As a consequence, $\mathrm{Cov}\left[V|_{\mathsf{K}_1}, V|_{\mathsf{K}_2}\right] = 0$ and $V|_{\mathsf{K}_1}$ and $V|_{\mathsf{K}_2}$ are uncorrelated. Since $V|_{\mathsf{K}_1}, V|_{\mathsf{K}_2}$ are Gaussian random fields we get that $V|_{\mathsf{K}_1}, V|_{\mathsf{K}_2}$ are R-independent. □

The following lemma gives specific conditions on random fields in order for Theorems 3.1.18 and 3.1.19 to hold.

Lemma 3.1.21. *Let $d \in \mathbb{N}$. Let $(m_k)_{k \in \mathbb{N}}, (n_k)_{k \in \mathbb{N}}$ be two positive increasing integer sequences and $(\mathsf{w}_k)_{k \in \mathbb{N}}$ be the sequence of subsets given for any $k \in \mathbb{N}$ by, $\mathsf{w}_k = [\![0, m_k]\!] \times [\![0, n_k]\!]$. Let V be a d-dimensional R-independent second-order stationary random field over $\mathbb{Z}^2$, with $R \geq 0$. Then for all $k \in \mathbb{N}$*

(a) $|\mathsf{w}_k|^{-1} \sum_{p \in \mathsf{w}_k} \mathbb{E}\left[V(p)\right] = \mathbb{E}\left[V(0)\right]$;

(b) $\lim_{k \to +\infty} |\mathsf{w}_k|^{-1} \sum_{p_1, p_2 \in \mathsf{w}_k} \mathrm{Cov}\left[V(p_1), V(p_2)\right] = \sum_{p \in \mathbb{Z}^2} \mathrm{Cov}\left[V(p), V(0)\right]$.

In addition, (3.10) and (3.11) hold with $\mu = \mathbb{E}\left[V(0)\right]$ and $C = \sum_{p \in \mathbb{Z}^2} \mathrm{Cov}\left[V(p), V(0)\right]$ which is finite.

Proof. First, (a) is immediate by stationarity. Concerning (b), for any $k \in \mathbb{N}$ we have by stationarity

$$\begin{aligned} |\mathsf{w}_k|^{-1} \sum_{p_1, p_2 \in \mathsf{w}_k} \mathrm{Cov}\left[V(p_1), V(p_2)\right] &= |\mathsf{w}_k|^{-1} \sum_{p_1, p_2 \in \mathsf{w}_k} \mathrm{Cov}\left[V(p_1 - p_2), V(0)\right] \\ &= \sum_{p \in \mathbb{Z}^2} \mathrm{Cov}\left[V(p), V(0)\right] g_k(p) \ , \end{aligned}$$

where $g_k \in \mathbb{R}^{\mathbb{Z}^2}$ satisfies for any $p \in \mathbb{Z}^2$, $g_k(p) = |\mathsf{w}_k|^{-1} \mathbb{1}_{\mathsf{w}_k} * \check{\mathbb{1}}_{\mathsf{w}_k}(p)$. For any $k \in \mathbb{N}$, $p \in \mathbb{Z}^2$ we have $0 \leq g_k(p) \leq 1$ and $\lim_{k \to +\infty} g_k(p) = 1$. For any $p \in \mathbb{Z}^2$ such that $\|p\|_\infty > R$, $\mathrm{Cov}\left[V(p), V(0)\right] = 0$ and then $\sum_{p \in \mathbb{Z}^2} |\mathrm{Cov}\left[V(p), V(0)\right]| < +\infty$. Using the dominated convergence theorem we get that

$$\lim_{k \to +\infty} |\mathsf{w}_k|^{-1} \sum_{p_1, p_2 \in \mathsf{w}_k} \mathrm{Cov}\left[V(p_1), V(p_2)\right] = \sum_{p \in \mathbb{Z}^2} \mathrm{Cov}\left[V(p), V(0)\right] \ .$$

We obtain (3.10) and (3.11) by applying Theorem 3.1.18 and Theorem 3.1.19. □

Lemma 3.1.22. *Let $f : \mathbb{Z}^2 \to \mathbb{R}$, $f \neq 0$, a function with finite support. Then it holds for any $t \in \mathbb{Z}^2$, $\Gamma_f(t) \leq \Gamma_f(0)$, with equality if and only if $t = 0$.*

Proof. For any $t \in \mathbb{Z}^2$, let $\tau_t f = f(\cdot + t)$. By the definition of the autocorrelation Γ_f and using the Cauchy-Schwarz inequality we get that for any $t \in \mathbb{Z}^2$

$$\Gamma_f(t) = \langle \tau_t f, f \rangle \leq \|f\|_2^2 \leq \Gamma_f(0) \ ,$$

with equality if and only if $f = \alpha \tau_t f$, with $\alpha \neq 0$ since $f \neq 0$. This implies that $\mathrm{supp}\tau_t(f) = \mathrm{supp} f$. As a consequence $t = 0$, which concludes the proof. □

The following lemma ensures that items (a) and (b) in Theorem 3.1.19 are satisfied in the template similarity case when assuming summability conditions over y.

Lemma 3.1.23. *Under the hypotheses of Theorem 3.1.4, assuming **A**4 with $\ell \in \mathbb{N}$ and $q = 2\ell$. There exist $\mu_{q,q} > 0$ and $\sigma_{q,q} \geq 0$ such that for any $t \in \mathsf{E}$*

(a) $\lim_{k \to +\infty} |\mathsf{w}_k|^{1/2} \left(|\mathsf{w}_k|^{-1} \mathbb{E}\left[\mathscr{T}_{q,q}(U, y, t, \mathsf{w}_k) \right] - \mu_{q,q}(t) \right) = 0$;

(b) $\lim_{k \to +\infty} |\mathsf{w}_k|^{-1} \operatorname{Var}\left[\mathscr{T}_{q,q}(U, y, t, \mathsf{w}_k) \right] = \sigma_{q,q}^2(t)$.

Proof. (a) For any $k \in \mathbb{N}$ we have that

$$\begin{aligned}
\mathbb{E}\left[\mathscr{T}_{q,q}(U, y, t, \mathsf{w}_k) \right] &= \sum_{p \in \mathsf{w}_k} \mathbb{E}\left[(y(p) - U(p+t))^{2\ell} \right] \\
&= \sum_{j=0}^{2\ell} \binom{2\ell}{j} \sum_{p \in \mathsf{w}_k} (-1)^j y(p)^j \mathbb{E}\left[U(p)^{2\ell - j} \right] \\
&= \sum_{j=0}^{\ell} \binom{2\ell}{2j} \sum_{p \in \mathsf{w}_k} y(p)^{2j} \mathbb{E}\left[U(p)^{2(\ell - j)} \right] \\
&= \sum_{j=0}^{\ell} \binom{2\ell}{2j} \mathbb{E}\left[U(0) \right]^{2(\ell - j)} \sum_{p \in \mathsf{w}_k} y(p)^{2j} \ .
\end{aligned}$$

Let $\mu_{q,q} = \sum_{j=0}^{\ell} \binom{2\ell}{2j} \mathbb{E}\left[U(0) \right]^{2(\ell - j)} \beta_j$ and using **A**4-(a) we get that

$$\lim_{k \to +\infty} |\mathsf{w}_k|^{1/2} \left(|\mathsf{w}_k|^{-1} \mathbb{E}\left[\mathscr{T}_{q,q}(U, v, t, \mathsf{w}_k) \right] - \mu_{q,q}(t) \right) = 0 \ .$$

Now since $\mu_{q,q} \geq \mathbb{E}\left[U(0)^{2\ell} \right] \geq \mathbb{E}\left[U(0)^2 \right]^{\ell} \geq \Gamma_f(0) > 0$ we have that $\mu_{q,q} > 0$.

(b) For any $k \in \mathbb{N}$ we have that

$$\begin{aligned}
\operatorname{Var}\left[\mathscr{T}_{q,q}(U, v, t, \mathsf{w}_k) \right] &= \sum_{p_1, p_2 \in \mathsf{w}_k} \operatorname{Cov}\left[(U(p_1) - y(p_1))^{2\ell}, (U(p_2) - v(p_2))^{2\ell} \right] \\
&= \sum_{p_1, p_2 \in \mathsf{w}_k} \sum_{i,j=0}^{\ell} \binom{2\ell}{2i} \binom{2\ell}{2j} y(p_1)^{2i} y(p_2)^{2j} \operatorname{Cov}\left[U(p_1)^{2(\ell - i)}, U(p_2)^{2(\ell - j)} \right] \\
&= \sum_{p_1, p_2 \in \mathbb{Z}^2} \sum_{i,j=0}^{\ell} \binom{2\ell}{2i} \binom{2\ell}{2j} y_k(p_1)^{2i} y_k(p_1 + p_2)^{2j} \operatorname{Cov}\left[U(p_2)^{2(\ell - i)}, U(0)^{2(\ell - j)} \right]
\end{aligned}$$

$$= \sum_{i,j=0}^{\ell} \binom{2\ell}{2i}\binom{2\ell}{2j} \left\langle y_k^{2i} * \check{y}_k^{2j}, \mathrm{Cov}\left[U(\cdot)^{2(\ell-i)}, U(0)^{2(\ell-j)}\right]\right\rangle .$$

Let $\sigma_{q,q} = \sum_{i,j=0}^{\ell} \binom{2\ell}{2i}\binom{2\ell}{2j} \langle \gamma_{i,j}, \mathrm{Cov}\left[U(\cdot)^{2(\ell-i)}, U(0)^{2(\ell-j)}\right]\rangle$. Using **A**4-(b) we can conclude.

□

Note that this lemma is also valid in the continuous case.

Asymptotic theorems – continuous case

We now turn to the continuous setting. We start by stating the continuous counterparts of Theorem 3.1.18 and Theorem 3.1.19. The following theorem, given here for completeness, can be found with different assumptions (in the one-dimensional case) in [Lin13].

Theorem 3.1.24. *Let $d \in \mathbb{N}$. Let $(m_k)_{k\in\mathbb{N}}, (n_k)_{k\in\mathbb{N}}$ be two positive increasing integer sequences and $(\mathsf{w}_k)_{k\in\mathbb{N}}$ be the sequence of subsets given for any $k \in \mathbb{N}$ by, $\mathsf{w}_k = [0, m_k] \times [0, n_k]$. Let V be a d-dimensional R-independent random field over $\mathbb{R}^2$, with $R \geq 0$, such that $\|V(p) - \mathbb{E}[V(p)]\|$ is uniformly stochastically dominated by $\widetilde{V}$, a stationary random field of order $r > 2$ over $\mathbb{R}^2$. Then V is a second-order random field. In addition, assume V is sample path continuous and that there exists $\mu \in \mathbb{R}^d$ given by $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \int_{p\in\mathsf{w}_k} \mathbb{E}\left[V(p)\right] \mathrm{d}p = \mu$. Then it holds that*

$$\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \int_{p\in\mathsf{w}_k} V(p)\mathrm{d}p \underset{a.s.}{=} \mu . \tag{3.18}$$

Proof. Without loss of generality we can assume that $d = 1$ and that for any $p \in \mathsf{E}$, $\mathbb{E}\left[V(p)\right] = 0$. Let $(\sigma_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ given for any $k \in \mathbb{N}$ by

$$\sigma_k^2 = \mathbb{E}\left[\left(k^{-2}\int_{\mathsf{E}_k} V(p)\mathrm{d}p\right)^2\right] , \tag{3.19}$$

with $\mathsf{E}_k = [0, k]^2$. Since V is R-independent, for any $p_1, p_2 \in \mathsf{E}$ such that $\|p_1 - p_2\|_\infty > R$, we have $C(p_1, p_2) = 0$. Hence for k large enough we obtain

$$\begin{aligned}\int_{\mathsf{E}_k}\int_{\mathsf{E}_k} C(p_1,p_2)\mathrm{d}p_1\mathrm{d}p_2 &\leq \int_{p_1\in\mathsf{E}_k}\int_{\|p_2\|_\infty\leq R} |C(p_1, p_1+p_2)|\mathrm{d}p_2\mathrm{d}p \\ &\leq k^2|\bar{\mathrm{B}}_\infty(0,R)| \sup_{(p_1,p_2)\in\mathsf{E}_k\times\bar{\mathrm{B}}_\infty(0,R)} |C(p_1,p_1+p_2)| .\end{aligned} \tag{3.20}$$

Using that $\widetilde{V}$ uniformly stochastically dominates $|V|$, the stationarity of $\widetilde{V}$, and the Cauchy-Schwarz inequality, we obtain for any $p_1, p_2 \in \mathsf{E}$,

$$|C(p_1, p_1+p_2)| = |\mathbb{E}\left[V(p_1)V(p_1+p_2)\right]| \leq \mathbb{E}\left[\widetilde{V}^2(p_1)\right]^{1/2} \mathbb{E}\left[\widetilde{V}^2(p_1+p_2)\right]^{1/2} \leq \mathbb{E}\left[\widetilde{V}^2(0)\right] . \tag{3.21}$$

Combining (3.19), (3.20) and (3.21) we get that for any $k \in \mathbb{N}$

$$\sigma_k^2 \leq Mk^{-2} ,$$

with $M = |\bar{\mathrm{B}}_\infty(0,R)|\mathbb{E}[\widetilde{V}^2(0)]$. The series $\sum_{k\in\mathbb{N}}\sigma_k^2$ converges and $\sum_{k\in\mathbb{N}}(k^{-2}\int_{\mathsf{E}_k}V(p)\mathrm{d}p)^2$ is finite almost surely. This proves that $\lim_{k\to+\infty}k^{-2}\int_{\mathsf{E}_k}V(p)\mathrm{d}p = 0$ almost surely. Using [CL04, p. 95] we get that

$$\lim_{k\to+\infty}\sup_{\mathsf{E}_k\subset\mathsf{w}\subset\mathsf{E}_{k+1}}\left||\mathsf{w}|^{-1}\int_{\mathsf{w}}V(p)\mathrm{d}p - k^{-2}\int_{\mathsf{E}_k}V(p)\mathrm{d}p\right| \underset{a.s.}{=} 0\ .$$

Combining this result with $\lim_{k\to+\infty}k^{-2}\int_{\mathsf{E}_k}V(p)\mathrm{d}p = 0$ we obtain that

$$lim_{k\to+\infty}|\mathsf{w}_k|^{-1}\int_{p\in\mathsf{w}_k}V(p)\mathrm{d}p \underset{a.s.}{=} 0\ .$$

□

The following theorem is an application of [IL89, Theorem 1.7.1].

Theorem 3.1.25. *Under the hypotheses of Theorem 3.1.24 and assuming that there exist $\mu\in\mathbb{R}^d$ and $C\in\mathcal{M}_d(\mathbb{R})$ such that*

(a) $\lim_{k\to+\infty}|\mathsf{w}_k|^{-1/2}\int_{p\in\mathsf{w}_k}(\mathbb{E}[V(p)]-\mu)\,\mathrm{d}p = 0$;

(b) $\lim_{k\to+\infty}|\mathsf{w}_k|^{-1}\int_{p_1,p_2\in\mathsf{w}_k}\mathrm{Cov}[V(p_1),V(p_2)]\,\mathrm{d}p_1\mathrm{d}p_2 = C$.

Then it holds that

$$\lim_{k\to+\infty}|\mathsf{w}_k|^{-1/2}\int_{p\in\mathsf{w}_k}(V(p)-\mu)\,\mathrm{d}p \underset{\mathcal{L}}{=} \mathsf{N}(0,C)\ . \tag{3.22}$$

Proof. Let $a\in\mathbb{R}^d$. We consider the d-dimensional random field ξ over $\mathbb{R}^2$ defined for any $p\in\mathbb{R}^2$ by $\xi(p) = V(p)-\mathbb{E}[V(p)]$. We define also the weight functions $(g_n)_{n\in\mathbb{N}}$ given for any $n\in\mathbb{N}$ by $g_n(p) = |\mathsf{w}_n|^{-1/2}\mathbb{1}_{p\in\mathsf{w}_n}$. For any $n\in\mathbb{N}$, let $S_n = \int_{\mathbb{R}^2}g_n(p)\xi(p)\mathrm{d}p$. We have for any $n\in\mathbb{N}$,

$$S_n = |\mathsf{w}_n|^{-1/2}\int_{p\in\mathsf{w}_n}(V(p)-\mathbb{E}[V(p)])\mathrm{d}p\ .$$

Let ξ^a be the one-dimensional random field over $\mathbb{R}^2$ such that for any $p\in\mathbb{R}^2$, $\xi^a(p) = \langle a,\xi(p)\rangle$ and $(S_n^a)_{n\in\mathbb{N}}$ be the sequence of real-valued random variables such that for any $n\in\mathbb{N}$, $S_n^a = \langle a,S_n\rangle$. Then for any $n\in\mathbb{N}$, $S_n^a = \int_{\mathbb{R}^2}g_n(p)\xi^a(p)\mathrm{d}p$. Using (b) we have that

$$\lim_{n\to+\infty}\mathbb{E}\left[(S_n^a)^2\right] = \lim_{n\to+\infty}|\mathsf{w}_n|^{-1}\int_{p_1,p_2\in\mathsf{w}_n}a^\top\mathrm{Cov}[V(p_1),V(p_2)]\,a = a^\top Ca\ . \tag{3.23}$$

By assumption, ξ^a is stochastically dominated by $\|a\|\widetilde{V}$ and therefore for any $p\in\mathbb{R}^2$ we have

$$\mathbb{E}[|\xi^a|^r] < +\infty\ . \tag{3.24}$$

Combining (3.23), (3.24), the fact that V is R-independent, [IL89, Theorem 1.7.1] and (a) we obtain that

$$\lim_{k\to+\infty}\left\langle a,|\mathsf{w}_k|^{-1/2}\int_{p\in\mathsf{w}_k}(V(p)-\mu)\,\mathrm{d}p\right\rangle \underset{\mathcal{L}}{=} \mathsf{N}(0,a^\top Ca)\ .$$

We conclude the proof upon using the Cramér-Wold theorem, [CW36, Theorem 1].

□

The following lemmas are the continuous versions of Lemma 3.1.21 and 3.1.22.

Lemma 3.1.26. *Let $d \in \mathbb{N}$. Let $(m_k)_{k\in\mathbb{N}}, (n_k)_{k\in\mathbb{N}}$ be two positive increasing integer sequences and $(\mathsf{w}_k)_{k\in\mathbb{N}}$ be the sequence of subsets such that for any $k \in \mathbb{N}$, $\mathsf{w}_k = [0, m_k] \times [0, n_k]$. Let V be a d-dimensional, R-independent random field of order $r > 2$ over $\mathbb{R}^2$, with $R \geq 0$. Assume that V is sample path continuous, then for all $k \in \mathbb{N}$*

(a) $|\mathsf{w}_k|^{-1} \int_{p\in\mathsf{w}_k} \mathbb{E}\left[V(p)\right] \mathrm{d}p = \mathbb{E}\left[V(0)\right]$;

(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} \int_{p_1,p_2\in\mathsf{w}_k} \mathrm{Cov}\left[V(p_1), V(p_2)\right] \mathrm{d}p\mathrm{d}p_2 = \int_{p\in\mathbb{R}^2} \mathrm{Cov}\left[V(p), V(0)\right] \mathrm{d}p$.

In addition, (3.18) and (3.22) hold with $\mu = \mathbb{E}\left[V(0)\right]$ and $C = \int_{p\in\mathbb{R}^2} \mathrm{Cov}\left[V(p), V(0)\right] \mathrm{d}p$.

Proof. (a) The proof is immediate since for any $p \in \mathbb{R}^2$, $\mathbb{E}\left[V(p)\right] = \mathbb{E}\left[V(0)\right]$.

(b) For any $k \in \mathbb{N}$ we have by stationarity

$$|\mathsf{w}_k|^{-1} \int_{p_1,p_2\in\mathsf{w}_k} \mathrm{Cov}\left[V(p_1), V(p_2)\right] \mathrm{d}p_1\mathrm{d}p_2 = |\mathsf{w}_k|^{-1} \int_{p_1,p_2\in\mathsf{w}_k} \mathrm{Cov}\left[V(p_1 - p_2), V(0)\right] \mathrm{d}p_1\mathrm{d}p_2 \ .$$

By the Fubini-Lebesgue theorem we obtain that for any $k \in \mathbb{N}$,

$$|\mathsf{w}_k|^{-1} \int_{p_1,p_2\in\mathsf{w}_k} \mathrm{Cov}\left[V(p_1), V(p_2)\right] \mathrm{d}p_1\mathrm{d}p_2 = \int_{p\in\mathbb{R}^2} \mathrm{Cov}\left[V(p), V(0)\right] g_k(p)\mathrm{d}p \ ,$$

where $g_k \in \ell_\infty(\mathbb{R}^2)$ satisfies for any $p \in \mathbb{R}^2$, $g_k(p) = |\mathsf{w}_k|^{-1} \mathbb{1}_{\mathsf{w}_k} * \check{\mathbb{1}}_{\mathsf{w}_k}(p)$. For any $k \in \mathbb{N}$, $p \in \mathbb{R}^2$ we have $0 \leq g_k(p) \leq 1$ and $\lim_{k\to+\infty} g_k(p) = 1$. For any $p \in \mathbb{R}^2$ such that $\|p\|_\infty > R_t$, $\mathrm{Cov}\left[V(p), V(0)\right] = 0$ and then

$$\int_{p\in\mathbb{R}^2} |\mathrm{Cov}\left[V(p), V(0)\right]| \mathrm{d}p < +\infty \ .$$

Using the dominated convergence theorem we get that

$$|\mathsf{w}_k|^{-1} \int_{p_1,p_2\in\mathsf{w}_k} \mathrm{Cov}\left[V(p_1), V(p_2)\right] \mathrm{d}p_1\mathrm{d}p_2 = \int_{p\in\mathbb{R}^2} \mathrm{Cov}\left[V(p), V(0)\right] \mathrm{d}p \ ,$$

Since V is R-independent we conclude the proof by applying Theorem 3.1.24 and 3.1.25.

□

Lemma 3.1.27. *Let Γ be a function over $\mathbb{R}^2$, $\Gamma \neq 0$, such that for any $p_1, p_2 \in \mathbb{R}^2$, $C(p_1, p_2) = \Gamma(p_1 - p_2)$ with C the covariance function of V a second-order random field over $\mathbb{R}^2$. Assume that Γ has finite support. Then it holds for any $t \in \mathbb{R}^2$, $\Gamma(t) \leq \Gamma(0)$, with equality if and only if $t = 0$.*

Proof. Upon replacing for any $p \in \mathbb{R}^2$, $V(p)$ by $V(p) - \mathbb{E}\left[V(p)\right]$ we suppose that $\mathbb{E}\left[V(p)\right] = 0$. Using the Cauchy-Schwarz inequality and the stationarity of V we get for any $t \in \mathbb{R}^2$ and $p \in \mathbb{R}^2$

$$\Gamma(t) = \mathbb{E}\left[V(p + t)V(p)\right] \leq \mathbb{E}\left[V(p + t)^2\right]^{1/2} \mathbb{E}\left[V(p)^2\right]^{1/2} \leq \mathbb{E}\left[V(p)^2\right] \leq \Gamma(0) \ .$$

with equality if and only if $V(p + t) = \alpha(p)V(p)$ with $\alpha(p) \in \mathbb{R}$. Since V is stationary and $V \neq 0$ we have that for any $p, t \in \mathbb{R}^2$, $\mathbb{E}[V(p + t)^2] = \mathbb{E}[V(p)^2] > 0$. Thus $\alpha(p)^2 = 1$ and for all $n \in \mathbb{N}$, $V(nt) = \pm V(0)$. If $t \neq 0$ then there exists $n \in \mathbb{N}$ such that $nt \notin \mathrm{supp}\Gamma$ and then we have

$$0 = \Gamma(nt) = \mathbb{E}\left[V(nt)V(0)\right] = \pm\mathbb{E}\left[V(0^2)\right] \neq 0 \ ,$$

which is absurd. Thus the equality in the inequality holds if and only if $t = 0$. □

Explicit constants

In order to derive precise constants in Theorem 3.1.4 and Theorem 3.1.6 we use the following lemma which is a consequence of the Isserlis formula [Iss18].

Lemma 3.1.28. *Let U and V be two zero-mean, real-valued Gaussian random variable and $k, \ell \in \mathbb{N}$. We have*

$$\mathbb{E}\left[U^{2k}V^{2\ell}\right] = \sum_{j=0}^{k\wedge\ell} r_{j,k,\ell}\mathbb{E}\left[U^2\right]^{k-j}\mathbb{E}\left[V^2\right]^{\ell-j}\mathbb{E}\left[UV\right]^{2j} \quad \text{and}$$

$$\mathrm{Cov}\left[U^{2k}, V^{2\ell}\right] = \sum_{j=1}^{k\wedge\ell} r_{j,k,\ell}\mathbb{E}\left[U^2\right]^{k-j}\mathbb{E}\left[V^2\right]^{\ell-j}\mathbb{E}\left[UV\right]^{2j} \ ,$$

with $r_{j,k,\ell}$ defined by (3.2).

Proof. Let $k, \ell \in \mathbb{N}$. Using Isserlis formula [Iss18] we obtain that $\mathbb{E}\left[U^{2k}V^{2\ell}\right]$ is the sum over all the partitions in pairs of $\{\underbrace{U,\dots,U}_{2k \text{ times}},\underbrace{V,\dots,V}_{2\ell \text{ times}}\}$ of the product of the expectations given by a pair partition. Given a pair partition we identify three different cases, $\{U,U\}$, $\{V,V\}$ and $\{U,V\}$. We only need to count the number of times each case appears in the sum. We denote the number of $\{U,U\}$ couples in a given pair partition p by $n_{U,U}(p)$. In the same fashion we define $n_{U,V}(p)$ and $n_{V,V}(p)$. We have $2k = 2n_{U,U}(p) + n_{U,V}(p)$ which proves that $n_{U,V}(p)$ is even. We denote by $\mathscr{P}_j$ the number of pair partitions p such that $n_{U,V}(p) = 2j$, with $j \in [\![0, k\wedge\ell]\!]$.

The cardinality of $\mathscr{P}_j$ is given by $r_{j,k,\ell}$. Indeed, in order to select $2j$ pair $\{U,V\}$ we select $2j$ elements among $2k$ (selection of replicates of U), same for V which gives $\binom{2k}{2j}\binom{2\ell}{2j}$ possibilities. Considering all the bijections between these elements we construct all the possible $2j$ pairs $\{U,V\}$. Given $2j$ pairs $\{U,V\}$ we must construct $k-j$ pairs $\{U,U\}$ and $\ell-j$ pairs $\{V,V\}$ in order to obtain a pair partition of $\mathscr{P}_j$. The number of pairs partition of a set with $\ell-j$ elements is given $q_{\ell-j}$. As a consequence we obtain for all $j \in [\![0, k\wedge\ell]\!]$

$$|\mathscr{P}_j| = q_{k-j}q_{\ell-j}\binom{2k}{2j}\binom{2\ell}{2j}(2j)! = r_{j,k,\ell} \ .$$

Summing over $j \in [\![0, k\wedge\ell]\!]$ we obtain all the possible pair partition and we get

$$\mathbb{E}\left[U^{2k}V^{2\ell}\right] = \sum_{j=0}^{k\wedge\ell} r_{j,k,\ell}\mathbb{E}\left[U^2\right]^{k-j}\mathbb{E}\left[V^2\right]^{\ell-j}\mathbb{E}\left[UV\right]^{2j} \ . \tag{3.25}$$

Using that $r_{0,k} = q_k^2$, respectively $r_{0,\ell} = q_\ell^2$ and $\mathbb{E}\left[U^{2k}\right] = q_k\mathbb{E}\left[U^2\right]^k$, respectively $\mathbb{E}\left[V^{2\ell}\right] = q_\ell\mathbb{E}\left[V^2\right]^\ell$, we obtain that the first term in the sum of (3.25) is equal to $\mathbb{E}\left[U^{2k}\right]\mathbb{E}\left[V^{2\ell}\right]$. Hence by removing this term we obtain the covariance and conclude the proof. □

Proof of Proposition 3.1.5. The proof is divided into three parts. First we consider the case $i = q$ then the case $i = sc$ and finally the case $i = \cos$.

1. Let $i = q$ with $q = 2\ell$ and $\ell \in \mathbb{N}$, $t \in \mathbb{Z}^2\setminus\{0\}$ and V_t the Gaussian random field given for any $p \in \mathbb{Z}^2$ by $V_t(p) = U(p) - U(p+t)$. Note that for all $p \in \mathbb{Z}^2$ we have $V_t(p)^{2\ell} = V_{q,t}(p)$. For any $p_1, p_2 \in \mathbb{Z}^2$

we have

$$\begin{aligned}&\mathbb{E}\left[V_t(p_1)\right] = 0\ ,\\ &\mathrm{Cov}\left[V_t(p_1), V_t(p_2)\right] = 2\Gamma_f(p_1-p_2) - \Gamma_f(p_1-p_2-t) - \Gamma_f(p_1-p_2+t) = \Delta_f(t, p_1-p_2)\ ,\end{aligned} \tag{3.26}$$

with Δ_f given by (3.3). We show in proof of Theorem 3.1.4, see (3.1), that for any $t \in \mathbb{Z}^2 \setminus \{0\}$

$$\mu_q(t) = \mathbb{E}\left[V_t^{2\ell}(0)\right]^{1/2\ell}\ , \qquad \sigma_q(t)^2 = \sum_{p\in\mathbb{Z}^2} \mathrm{Cov}\left[V_t^{2\ell}(p), V_t^{2\ell}(0)\right] \mathbb{E}\left[V_t^{2\ell}(0)\right]^{1/\ell-2} / (2\ell)^2\ . \tag{3.27}$$

Combining (3.26), (3.27) and Lemma 3.1.28 we get that

(a) $\mu_q(t) = q_{2\ell}^{1/(2\ell)} \Delta_f(t,0)^{1/2}$;

(b) $\sigma_q(t)^2 = \sum_{p\in\mathbb{Z}^2} \left(\sum_{j=1}^{\ell} r_{j,\ell} \Delta_f(t,0)^{2(\ell-j)} \Delta_f(t,p)^{2j} \right) q_\ell^{1/\ell-2} \Delta_f(t,0)^{1-2\ell} / (2\ell)^2$.

Exchanging the sums in (b) we get $\sigma_q(t)^2 = \frac{q_\ell^{1/\ell-2}}{(2\ell)^2} \sum_{j=1}^{\ell} r_{j,\ell} \left(\frac{\|\Delta_f(t,\cdot)\|_{2j}}{\Delta_f(t,0)} \right)^{2j} \Delta_f(t,0)$.

2. Let $i = sc$, $t \in \mathbb{Z}^2 \setminus \{0\}$ and $V_{sc,t}$ be a Gaussian random field given for any $p \in \mathbb{Z}^2$, by $V_t(p) = U(p)U(p+t)$. For any $p_1, p_2 \in \mathbb{Z}^2$ we have

$$\begin{aligned}&\mathbb{E}\left[V_{sc,t}(p_1)\right] = \Gamma_f(t)\ ,\\ &\mathrm{Cov}\left[V_{sc,t}(p_1), V_{sc,t}(p_2)\right] = \Gamma_f(p_1-p_2) - \Gamma_f(p_1-p_2-t)\Gamma_f(p_1-p_2+t) = \widetilde{\Delta}_f(t, p_1-p_2)\ ,\end{aligned} \tag{3.28}$$

with $\widetilde{\Delta}_f$ given by (3.3). We show in the proof of Theorem 3.1.4, see (3.1), that for any $t \in \mathbb{Z}^2 \setminus \{0\}$

$$\mu_{sc}(t) = \mathbb{E}\left[V_{sc,t}(0)\right]\ , \qquad \sigma_{sc}(t)^2 = \sum_{p\in\mathbb{Z}^2} \mathrm{Cov}\left[V_{sc,t}(p), V_{sc,t}(0)\right]\ . \tag{3.29}$$

Combining (3.28) and (3.29) we get that

(a) $\mu_{sc}(t) = \Gamma_f(t)$;

(b) $\sigma_{sc}(t)^2 = \sum_{p\in\mathbb{Z}^2} \widetilde{\Delta}_f(t,p)$,

which concludes the proof in the case $i = sc$.

3. We now consider the case $i = \cos$. Recall that in the proof of Theorem 3.1.4 we show that

$$\mathscr{A}_{s_{\cos}}(U, t, \mathsf{w}_k) = h\left(|\mathsf{w}_k|^{-1} \sum_{p\in\mathsf{w}_k} V_{\cos,t}(p) \right)\ ,$$

where for any $x \in \mathbb{R}$, $y, z > 0$

$$h(x,y,z) = xy^{-1/2}z^{-1/2}\ , \qquad V_{\cos,t}(p) = \begin{pmatrix} -U(p)U(p+t) \\ U(p)^2 \\ U(p+t)^2 \end{pmatrix}\ .$$

Applying Lemma 3.1.21 there exist $\tilde{\mu}_{\cos}(t)$ and $\tilde{C}_{\cos}(t)$ such that

(a) $\lim_{k\to+\infty} |\mathsf{w}_k|^{-1} V_{\cos,t} \underset{a.s.}{=} \tilde{\mu}_{\cos}(t)$;

(b) $\lim_{k\to+\infty} |\mathsf{w}_k|^{1/2} \left(|\mathsf{w}_k|^{-1} V_{\cos,t} - \tilde{\mu}_{\cos}(t)\right) \underset{\mathcal{L}}{=} \mathsf{N}\left(0, \tilde{C}_{\cos}(t)\right)$,

with

$$\begin{aligned}
\tilde{\mu}_{\cos}(t) &= \begin{pmatrix} \Gamma_f(t) \\ \Gamma_f(0) \\ \Gamma_f(0) \end{pmatrix} , \\
\tilde{C}_{\cos}(t) &= \begin{pmatrix} \|\Gamma_f\|^2 + \Gamma_f * \breve{\Gamma}_f(2t) & 2\Gamma_f * \breve{\Gamma}_f(t) & 2\Gamma_f * \breve{\Gamma}_f(t) \\ 2\Gamma_f * \breve{\Gamma}_f(t) & 2\|\Gamma_f\|^2 & 2\|\Gamma_f\|^2 \\ 2\Gamma_f * \breve{\Gamma}_f(t) & 2\|\Gamma_f\|^2 & 2\|\Gamma_f\|^2 \end{pmatrix} .
\end{aligned} \tag{3.30}$$

In addition, for any $x \in \mathbb{R}$, $y, z > 0$

$$\nabla h(x, y, z) = \begin{pmatrix} y^{-1/2} z^{-1/2} \\ -(1/2) x y^{-3/2} z^{-1/2} \\ -(1/2) x y^{-1/2} z^{-3/2} \end{pmatrix} .$$

Combining this result, (3.30) and the multivariate Delta method we get that

$$\begin{aligned}
\mu_{\cos}(t) &= h(\Gamma_f(t), \Gamma_f(0), \Gamma_f(0)) = \Gamma_f(t)/\Gamma_f(0) , \\
\sigma_{\cos}(t)^2 &= \nabla h(\Gamma_f(t), \Gamma_f(0), \Gamma_f(0))^\top M \nabla h(\Gamma_f(t), \Gamma_f(0), \Gamma_f(0)) \\
&= \Gamma_f(0)^{-2} \left\{ \|\Gamma_f\|_2^2 \left(1 + 2\frac{\Gamma_f(t)^2}{\Gamma_f(0)^2}\right) - 4\frac{\Gamma_f(t)}{\Gamma_f(0)} \Gamma_f * \breve{\Gamma}_f + \Gamma_f * \breve{\Gamma}_f(2t) \right\} ,
\end{aligned}$$

where

$$M = \begin{pmatrix} \|\Gamma_f\|^2 + \Gamma_f * \breve{\Gamma}_f(2t) & 2\Gamma_f * \breve{\Gamma}_f(t) & 2\Gamma_f * \breve{\Gamma}_f(t) \\ 2\Gamma_f * \breve{\Gamma}_f(t) & 2\|\Gamma_f\|^2 & 2\|\Gamma_f\|^2 \\ 2\Gamma_f * \breve{\Gamma}_f(t) & 2\|\Gamma_f\|^2 & 2\|\Gamma_f\|^2 \end{pmatrix} ,$$

which concludes the proof. □

Proof of Proposition 3.1.7. The proof is divided in two parts. First we treat the case $i = q$ then the case $i = sc$ and $i = \cos$. Let $q = 2\ell$ with $\ell \in \mathbb{N}$. Lemma 3.1.23 gives us that

$$\begin{aligned}
\mu_{q,q} &= \sum_{j=0}^{\ell} \binom{2\ell}{2j} \mathbb{E}\left[U(0)\right]^{2(\ell-j)} \beta_j , \\
\sigma_{q,q} &= \sum_{i,j=0}^{\ell} \binom{2\ell}{2i} \binom{2\ell}{2j} \left\langle \gamma_{i,j}, \mathrm{Cov}\left[U(\cdot)^{2(\ell-i)}, U(0)^{2(\ell-j)}\right] \right\rangle .
\end{aligned}$$

Using Lemma 3.1.28 we obtain that

$$\begin{aligned}
\mu_{q,q} &= \Gamma_f(0)^\ell \sum_{j=0}^{\ell} \binom{2\ell}{2j} q_{\ell-j} \Gamma_f(0)^{-j} \beta_j , \\
\sigma_{q,q} &= \sum_{i,j=0}^{\ell} \binom{2\ell}{2i} \binom{2\ell}{2j} \sum_{m=1}^{\ell-i\wedge\ell-j} r_{m,k,\ell} \left\langle \gamma_{i,j}, \Gamma_f^{2m} \right\rangle \Gamma_f(0)^{2\ell-i-j-2m} .
\end{aligned}$$

We conclude using (3.4). For $i = sc$ and $i = \cos$, the result is given in the proof of Theorem 3.1.6. □

3.2 Patch redundancy in images

3.2.1 Abstract

In many image processing applications, using local information combined with the knowledge of long-range spatial arrangement is crucial. The spatial redundancy on sub-images called patches, encodes the small scale structure of the image as well as its large scale organization. More precisely, local information is encoded in the patch content and the large scale organization is contained in the redundancy of this information across the patches of the image. For example, patch-based inpainting techniques, such as [CPT04; HS14], assign patches of a known region to patches of an unknown region. Namely, each patch position on the border of the unknown region is associated to an offset corresponding to the best patch according to the partial available information. In [HS14] the authors replace the search on the whole image by a search among the most redundant offsets in the known region. This allows the authors of [HS14] to retrieve long-range spatial structure in the unknown part of the image. Another famous application of spatial redundancy can be found in denoising, with the seminal work (Non-Local means) of Buades and coauthors [BCM05], in which the authors propose to replace a noisy patch by the mean over all spatially redundant patches.

Last but not least, spatial redundancy is of crucial importance in exemplar-based texture synthesis. In this section we define textures as images containing repeated patterns but also reflecting randomness in the arrangement of these patterns. Among textures, one important class is given by the microtextures in which no individual object can be clearly delimited. In the periodic case, a more precise definition will be given in Definition 3.2.4. These microtexture models can be described by Gaussian random fields [Wij91; GGM11; Lec15; Xia+14]. Parametric models using features such as wavelet transform coefficients [PS00], scattering transform coefficients [SM13] or convolutional neural network outputs [GEB15] have been proposed in order to derive image models with more structure. On the other hand, non-parametric patch-based algorithms such as [EL99; EF01; Kwa+03; RDM16; GLR18] propose to use most similar patches in order to fill the new texture images, similarly to inpainting techniques.

All these techniques lift images in spaces with dimensions higher than the original image space, and make use of the redundancy of the lifting to extract important structural information. There exist two main types of lifting: feature extraction or patch extraction. Feature extraction relies on the use of filters, linear or non-linear, which aim at selecting substantial local information. Among popular kernels are oriented and multiscale filters, which happened to be identified as early processing in mammal vision systems [Dau85; HW59]. These last years have seen the rise of neural networks in which the filter dictionary is no longer given as an input but learned through a data-driven optimization procedure [SZ14]. On the other hand, patch-based methods rely on the assumption that image processing tasks are simplified when conducted in the higher dimensional patch space.

Every analysis performed in a lifted space, built via feature extraction or patch extraction, relies on the comparison of points in this space. In patch-based lifted spaces, we aim at finding dissimilarity functions such that two patches are visually close if the dissimilarity measurement between them is small. In this section we focus on the square Euclidean distance but other choices could be considered [WSB03; Wan+04; DDT12].

This leads us to consider a statistical hypothesis testing framework to assess similarity (or dissimilarity) between patches. The null hypothesis is defined as the absence of local structural similarities in the image. Reciprocally the alternative hypothesis is defined as the presence of such similarities. There exists a wide variety of tractable models exhibiting no similarity at long-range, like Gaussian random fields [Wij91; GGM11; Lec15; Xia+14] or spatial Markov random fields [CJ83], whereas sampling and inference in very structured models rely on optimization procedures and may be computationally ex-

pensive, their distribution being the limit of some Markov chain [ZWM98; LZW16] or some stochastic optimization procedure [BM18]. This encourages us to consider an *a contrario* approach, *i.e.* we do not consider the alternative hypothesis and focus on rejecting the null hypothesis. This framework was successfully applied in many areas of image processing [Dav+18; DMM00; DMM01; ADV03; Cao04] and aims at identifying structure events in images. This statistical model takes its roots in the fundamental work of the Gestalt theory [DMM08]. One of its principle, the non-accidentalness principle [Low12] or Helmholtz principle [Zhu99; DMM01], states that no structure is perceived in a noise model. To be precise, in our case of interest, we want to assess that no spatial redundancy is perceived in microtexture models. This methodology allows us to only design a locally structured background model to define a null hypothesis. Combining *a contrario* principles and patch-based measures, we propose an algorithm to identify auto-similarities in images.

We then turn to the implementation of such an algorithm and illustrate the diversity of its possible applications with three examples: denoising, lattice extraction, and periodicity ranking of textures. In our denoising application we propose a modification of the celebrated Non-Local means algorithm [BCM05] (NL-means) by inserting a threshold in the selection of similar patches. Using an *a contrario* model we are able to give probabilistic control on the patch reconstruction.

We then focus on periodicity detection and, more precisely, lattice extraction. Periodicity in images was described as an important feature in early mathematical vision [HSD73]. Most of the proposed methods to analyze periodicity rely on global measurements such as the modulus of the Fourier transform [MMN83] or the autocorrelation [LWY97]. These global techniques are widely used in crystallography where lattice properties, such as the angle between basis vectors, are fundamental [MB15; SL14]. Since all of our measurements are local, we are able to identify periodic similarities even in images which are not periodic but present periodic parts, for instance if two crystal structures are present in a single crystallography image. We draw a link between the introduced notion of auto-similarity and the inertia measurement in co-occurence matrices [HSD73]. We then introduce our lattice proposal algorithm which combines a detection map, *i.e.* the output of our redundancy detection algorithm, and graphical model techniques, as in [Par+09], in order to extract lattice basis vectors.

Our last application concerns texture ranking. Since the definition of texture is broad and covers a wide range of images, it is a natural question to identify criteria in order to distinguish textures. In [LCT03], the authors use a classical measure for distinguishing textures: regularity. In this work, we narrow this criterion and restrict ourselves to the study of periodicity in texture images. The proposed graphical model inference naturally gives a quantitative measurement for texture periodicity ranking. We give an example of ranking on 25 images of the Brodatz set.

The rest of this chapter is organized as follows. An *a contrario* framework for local similarity detection is proposed in Section 3.2.2. In the *a contrario* framework, a background model, corresponding to the null hypothesis, is required. The consequence of choosing Gaussian models as background models is investigated and a redundancy detection algorithm is proposed in Section 3.2.3. The rest of the section is dedicated to some examples of application of the proposed framework. After reviewing one of the most popular method in image denoising we introduce a denoising algorithm in Section 3.2.4 and present our experimental results in Section 3.2.4. Local dissimilarity measurements can be used as periodicity detectors. The link between the locality of the introduced functions and the literature on periodicity detection problems is investigated in Section 3.2.5. An algorithm for detecting lattices in images is given in Section 3.2.5 and numerical results are presented in Section 3.2.5. In our last experiment in Section 3.2.5, we introduce a criterion for measuring texture periodicity. We conclude our study and discuss future work in Section 3.2.6.

3.2.2 An a contrario framework for auto-similarity

In this section, we only consider the autosimilarity $\mathscr{A}_{2,2}(x,t,\mathsf{w})$, see Definition 3.1.3 and denote it $\mathscr{A}(x,t,\mathsf{w})$. This auto-similarity computes the squared distance between a patch of u defined on a domain w and the patch of u defined by the domain w shifted by the offset vector t. In what follows, we introduce an *a contrario* framework on the auto-similarity. This framework will allow us to derive an algorithm for detecting spatial redundancy in natural images. In this section we fix an image domain $\mathsf{E} \subset \mathbb{Z}^2$ and a patch domain $\mathsf{w} \subset \mathsf{E}$. We recall that our final aim is to design a criterion that will answer the following question: are two given patches similar? This criterion will be given by the comparison between the value of a dissimilarity function and a threshold a. We will define the threshold a so that few similarities are identified in the null hypothesis model, *i.e.* similarity does not occur "just by chance". Thus we can reformulate the initial question: is the similarity output of a dissimilarity function between two patches small enough? Or, to be more precise, how can we set the threshold a in order to obtain a criterion for assessing similarity between patches?

This formulation agrees with the *a contrario* framework [DMM08] which states that geometrical and/or perceptual structure in an image is meaningful if it is a rare event in a background model. This general principle is sometimes called the Helmholtz principle [Zhu99] or the non-accidentalness principle [Low12]. Therefore, in order to control the number of similarities identified in the background model, we study the probability density function of the auto-similarity function with input random image U over E. We will denote by $\mathbb{P}_0$ the probability distribution of U over $\mathbb{R}^{\mathsf{E}}$, the images over E. We will assume that $\mathbb{P}_0$ is a microtexture model, see Definition 3.2.4 below for a precise definition of such a model. We define the following significant event which encodes spatial redundancy: $\mathscr{A}(u,t,\mathsf{w}) \leq a(t)$, where a, the threshold function, is defined over the offsets ($t \in \mathbb{Z}^2$) but also depends on other parameters such as w or $\mathbb{P}_0$. The dependency of a with respect to t cannot be omitted. For instance, even in a Gaussian white noise W, the probability distribution function of $\mathscr{A}(W,t,\mathsf{w})$ depends on t.

The Number of False Alarms (NFA) is a crucial quantity in the *a contrario* methodology. A false alarm is defined as an occurrence of the significant event in the background model $\mathbb{P}_0$. We recall that in our model the significant event is patch redundancy. This test must be conducted for every possible configurations of the significant event, *i.e.* in our case we test every possible offset t. The NFA is then defined as the expectation of the number of false alarms over all possible configurations. Bounding the NFA ensures that the probability of identifying k offsets with spatial redundancy is also bounded, see Proposition 3.2.3. In what follows we give the definition of the NFA in the spatial redundancy context.

Definition 3.2.1. *Let $U \sim \mathbb{P}_0$, where $\mathbb{P}_0$ is a background microtexture model. We define the auto-similarity probability map* AP *for any $t \in \mathsf{E}$, $\mathsf{w} \subset \mathsf{E}$ and $a : \mathsf{E} \to \mathbb{R}$ by*

$$\mathsf{AP}(t,\mathsf{w},a) = \mathbb{P}_0\left(\mathscr{A}(U,t,\mathsf{w}) \leq a(t)\right) . \tag{3.31}$$

We define the auto-similarity expected number of false alarms ANFA *by*

$$\mathsf{ANFA}(\mathsf{w},a) = \sum_{t\in\mathsf{E}} \mathsf{AP}(t,\mathsf{w},a) . \tag{3.32}$$

Note that $\mathsf{AP}(t,\mathsf{w},a)$ corresponds to the probability that $\mathsf{w}+t$ is similar to w in the background model U. For any $t \in \mathsf{E}$, the cumulative distribution function of the auto-similarity random variable $\mathscr{A}(U,t,\mathsf{w})$ under $\mathbb{P}_0$ evaluated at value $\alpha(t)$ is given by $\mathsf{AP}(t,\mathsf{w},\alpha(t))$. We denote by $q \mapsto \mathsf{AP}^{-1}(t,\mathsf{w},q)$ the inverse cumulative distribution function, potentially defined by a generalized inverse ($\mathsf{AP}^{-1}(t,\mathsf{w},q) = \inf\{\alpha(t) \in \mathbb{R},\ \mathsf{AP}(t,\mathsf{w},\alpha(t)) \geq q\}$), of the auto-similarity random variable for a fixed offset t, with $q \in (0,1)$ a quantile. We now have all the tools to control the number of detected offsets in the background model.

Definition 3.2.2. *Let $x : \mathsf{E} \to \mathbb{R}$ be an image, $\mathsf{w} \subset \mathsf{E}$ a patch domain, and $a : \mathsf{E} \to \mathbb{R}$. An offset t is said to be detected with respect to a, if $\mathscr{A}(x, t, \mathsf{w}) \leq a(t)$.*

Note that a detected offset in $U \sim \mathbb{P}_0$ corresponds to a false alarm in the *a contrario* model. In what follows we suppose that the cumulative distribution function of $\mathscr{A}(U, t, \mathsf{w})$ is invertible for every $t \in \mathsf{E}$. This ensures that for any $t \in \mathsf{E}$ and $q \in (0, 1)$ we have

$$\mathsf{AP}\left(t, \mathsf{w}, \mathsf{AP}^{-1}\left(t, \mathsf{w}, q\right)\right) = q \, . \tag{3.33}$$

Proposition 3.2.3. *Let $\mathrm{NFA}_{max} \geq 0$ and for all $t \in \mathsf{E}$ define $a(t) = \mathsf{AP}^{-1}\left(t, \mathsf{w}, \mathrm{NFA}_{max} / |\mathsf{E}|\right)$. We have that for any $n \in \mathbb{N} \backslash \{0\}$,*

$$\mathsf{ANFA}(\mathsf{w}, a) = \mathrm{NFA}_{max} \quad \textit{and} \quad \mathbb{P}_0\left(\textit{“at least } n \textit{ offsets are detected in } U\textit{”}\right) \leq n^{-1} \, \mathrm{NFA}_{max} \, .$$

Proof. Using (3.32), and $a(t) = \mathsf{AP}^{-1}\left(t, \mathsf{w}, \mathrm{NFA}_{\max} / |\mathsf{E}|\right)$, we get

$$\mathsf{ANFA}(\mathsf{w}, a) = \sum_{t \in \mathsf{E}} \mathsf{AP}(t, \mathsf{w}, a) = \sum_{t \in \mathsf{E}} \mathsf{AP}\left(t, \mathsf{w}, \mathsf{AP}^{-1}\left(t, \mathsf{w}, \mathrm{NFA}_{\max} / |\mathsf{E}|\right)\right) = \mathrm{NFA}_{\max} \, ,$$

where the last equality is obtained using (3.33). Concerning the upper-bound, we have, using the Markov inequality and (3.31), for any $n \in \mathbb{N} \backslash \{0\}$

$$\begin{aligned} \mathbb{P}_0\left(\text{“at least } n \text{ offsets are detected in } U\text{”}\right) &= \mathbb{P}_0\left(\sum_{t \in \mathsf{E}} \mathbb{1}_{\mathscr{A}(U,t,\mathsf{w}) \leq a(t)} \geq n\right) \\ &\leq n^{-1} \sum_{t \in \mathsf{E}} \mathbb{E}\left[\mathbb{1}_{\mathscr{A}(U,t,\mathsf{w}) \leq a(t)}\right] \leq n^{-1} \, \mathrm{NFA}_{\max} \, , \end{aligned}$$

where $\mathbb{1}_{\mathscr{A}(U,t,\mathsf{w}) \leq a(t)} = 1$ if $\mathscr{A}(U, t, \mathsf{w}) \leq a(t)$ and 0 otherwise. □

Thus, setting a as in Proposition 3.2.3, an offset $t \in \mathsf{E}$ is detected for an image $x : \mathsf{E} \to \mathbb{R}$ if

$$\mathscr{A}(x, t, \mathsf{w}) \leq \mathsf{AP}^{-1}\left(t, \mathsf{w}, \mathrm{NFA}_{\max} / |\mathsf{E}|\right) \, . \tag{3.34}$$

This *a contrario* detection framework can then be simply rewritten as 1) computing the auto-similarity function with input image u, 2) thresholding the obtained dissimilarity map with the inverse cumulative distribution function of the computed dissimilarity function under $\mathbb{P}_0$. The computed threshold depends on the offset and Proposition 3.2.3 ensures probabilistic guarantees on the expected number of detections under $\mathbb{P}_0$. Using the inverse property of the inverse cumulative distribution function and (3.34), we obtain that an offset is detected if and only if

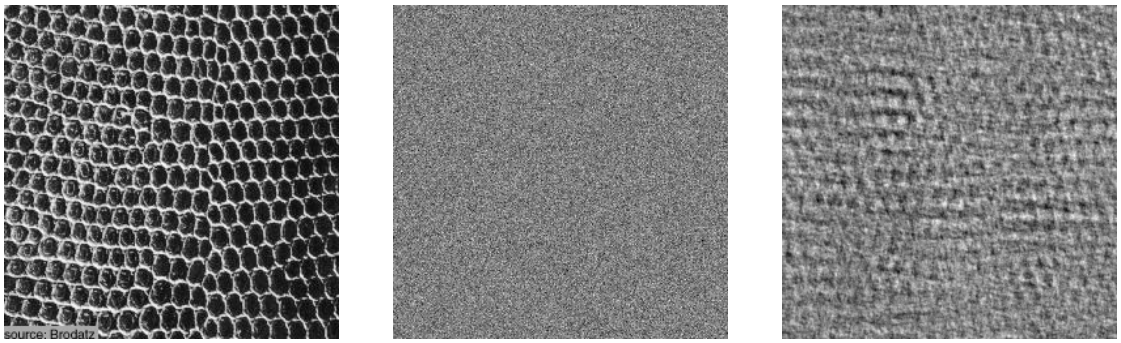
$$\mathbb{P}_0\left(\mathscr{A}(U, t, \mathsf{w}) \leq \mathscr{A}(x, t, \mathsf{w})\right) = \mathsf{AP}\left(t, \mathsf{w}, \mathscr{A}(x, t, \mathsf{w})\right) \leq \mathrm{NFA}_{\max} / |\mathsf{E}| \, . \tag{3.35}$$

Therefore, the thresholding operation can be conducted either on $\mathscr{A}(x, t, \mathsf{w})$, see (3.34), or on $\mathsf{AP}\left(t, \mathsf{w}, \mathscr{A}(x, t, \mathsf{w})\right)$, see (3.35). This property will be used in Section 3.2.3 to define a similarity detection algorithm based on the evaluation of $\mathscr{A}(x, t, \mathsf{w})$.

3.2.3 Gaussian model and detection algorithm

Choice of background model

In this section we compute $\mathsf{AP}\left(t, \mathsf{w}, \alpha\right)$, *i.e.* the cumulative distribution function of the similarity function under the null hypothesis model, with a Gaussian background model. Indeed, if the background



(a) (b) (c)

Figure 3.10: **Examples of microtexture models** In (a) we present an original 256×256 image. In (b) and (c) we derive two microtexture models. In (b) we present a Gaussian white noise and in (c) the microtexture model given by (3.36). Note that (c) shows more local structure than (b).

model is simply a Gaussian white noise the similarities identified by the *a contrario* algorithm are the ones that are not likely to be present in the Gaussian white noise image model. More generally, we consider stationary Gaussian random fields defined in the following way: we introduce an image $f : \mathsf{E} \to \mathbb{R}$ which contains the microtexture information we want to discard in our *a contrario* model. In what follows we give the definition of the microtexture model associated to f.

Definition 3.2.4. *Let $f : \mathsf{E} \to \mathbb{R}$, we define the associated microtexture model U by setting, $U = f * W$, where $*$ is the periodic convolution operator over E given for any $p_1 \in \mathsf{E}$ by $f * W(p_1) = \sum_{p_2 \in \mathsf{E}} W(p_2) \dot{f}(p_1 - p_2)$ and W is a white noise over E, i.e. $(W(p))_{p \in \mathsf{E}}$ are i.i.d. $\mathsf{N}(0,1)$ random variables.*

Note that with this definition a microtexture is an instance of the Gaussian random fields studied in Section 3.1 in the discrete case. Given an image $x : \mathsf{E} \to \mathbb{R}$, a microtexture model can be derived considering

$$m_x = \sum_{p \in \mathsf{E}} x(p)/|\mathsf{E}| \ , \quad \text{and} \quad U = |\mathsf{E}|^{-1/2}(x - m_x) * W \ . \tag{3.36}$$

Note that if U is given by (3.36) we have for any $p_1, p_2 \in \mathsf{E}$

$$\mathbb{E}\left[U(p_1)\right] = 0 \quad \text{and} \quad \mathrm{Cov}\left[U(p_1), U(p_2)\right] = |\mathsf{E}|^{-1} \sum_{p_3 \in \mathsf{E}} (\dot{x}(p_3) - m_x)(\dot{x}(p_3 - (p_2 - p_1))) - m_x) \ .$$

We refer to [GGM11] for a mathematical study of this model. In Section 3.1 we conducted a study of the behavior of numerous similarity functions in this model.

Detection algorithm

In this section, E is a finite square domain in $\mathbb{Z}^2$. We fix $\mathsf{w} \subset \mathsf{E}$. We also define f, a function over E. We consider the Gaussian random field $U = f * W$, where W is a Gaussian white noise over E. We denote by Γ_f the autocorrelation of f, *i.e.* $\Gamma_f = f * \check{f}$ where for any $p \in \mathsf{E}$, $\check{f}(p) = f(-p)$. We recall that the function Δ_f, see (3.3) is defined for any $t, p \in \mathsf{E}$ by

$$\Delta_f(t,p) = 2\Gamma_f(p) - \Gamma_f(p+t) - \Gamma_f(p-t) \ .$$

We also recall the following non-asymptotic result, see Proposition 3.1.12.

Proposition 3.2.5. *Let* $\mathsf{E} = [\![0, M-1]\!]^2$ *with* $M \in \mathbb{N}\backslash\{0\}$, $\mathsf{w} \subset \mathsf{E}$, $f : \mathsf{E} \to \mathbb{R}$ *and* $U = f * W$ *where* W *is a Gaussian white noise over* E. *Then, for any* $t \in \mathsf{E}$, $\mathscr{A}(U, t, \mathsf{w})$ *has the same distribution as* $\sum_{k=0}^{|\mathsf{w}|-1} \lambda_k(t, \mathsf{w}) Z_k$, *with* Z_k *independent chi-square random variables with parameter 1 and* $\lambda_k(t, \mathsf{w})$ *the eigenvalues of the covariance matrix* C_t *associated with function* $\Delta_f(t, \cdot)$ *restricted to* w, *defined in* (3.3), *i.e for any* $p_1, p_2 \in \mathsf{w}$, $C_t(p_1, p_2) = \Delta_f(t, p_1 - p_2)$.

As a consequence if $f = \delta_0$, *i.e.* U is a Gaussian white noise, and $\{p + t, p \in \mathsf{w}\} \cap \mathsf{w} = \emptyset$, *i.e.* there is no overlapping between the patch domain w and its shifted version, then $\mathscr{A}(U, t, \mathsf{w})$ is a chi-square random variable with parameter $|\mathsf{w}|$.

In order to compute the cumulative distribution function of a quadratic form of Gaussian random variables we must deal with two issues: 1) the computation of the eigenvalues $\lambda_k(t, \mathsf{w})$ might be time-consuming and efficient methods must be developed ; 2) the exact computation of the cumulative distribution function of a quadratic form of Gaussian random variables requires the use of heavy integrals, see [Imh61]. In Section 3.1.4 we introduced a projection method in order to easily compute approximated eigenvalues, with equality when $\mathsf{w} = \mathsf{E}$, see Proposition 3.1.13. The so-called Wood F method (see [Woo89; BA16]) shows the best trade-off between accuracy and computational cost to approximate the cumulative distribution function of quadratic forms in Gaussian random variables with given weights. It is a moment method of order 3, fitting a Fisher-Snedecor distribution to the empirical one. Note that in [LTZ09] another moment method of order 3 is proposed. In what follows, we assume that we can compute the cumulative distribution function of $\mathscr{A}(U, t, \mathsf{w})$ and we refer to Section 3.1.4 for further details.

In Algorithm 3 we propose an *a contrario* framework for spatial redundancy detection. We suppose that u and w are provided by the user. Using Proposition 3.2.3 and (3.35) , we say that an offset is detected if $\mathsf{AP}(t, \mathsf{w}, \mathscr{A}(x, t, \mathsf{w})) \leq \mathrm{NFA}_{\max} / |\mathsf{E}|$. The value $\mathrm{NFA}_{\max}$ is supposed to be set by the user. The background model used in the auto-similarity detection is the one given in (3.36). Therefore, Proposition 3.1.12 and the discussion that follows can be used to compute an approximation of $\mathsf{AP}(t, \mathsf{w}, \mathscr{A}(x, t, \mathsf{w}))$. In Figure 3.11 we apply Algorithm 3 to a texture image.

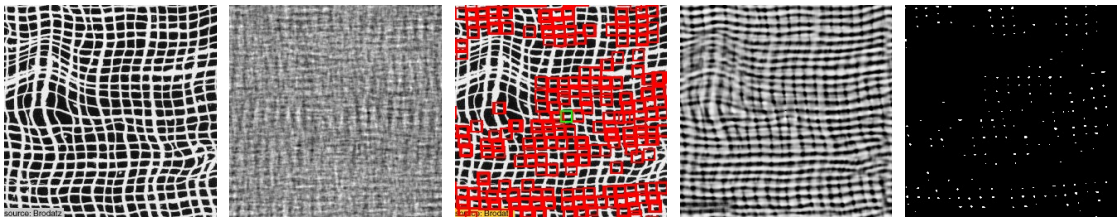
Algorithm 3 Auto-similarity detection

function AUTOSIM-DETECTION(u, w, $\mathrm{NFA}_{\max}$)
  for $t \in \mathsf{E}$ **do**
    val $\leftarrow \mathscr{A}(x, t, \mathsf{w})$
    $P_{map}(t) \leftarrow \mathsf{AP}(t, \mathsf{w}, \text{val})$ ▷ $\mathsf{AP}(t, \mathsf{w}, \text{val})$ approximation detailed in Section 3.2.3
    $D_{map}(t) \leftarrow \mathbb{1}_{P_{map}(t) \leq \mathrm{NFA}_{\max}/|\mathsf{E}|}$
  end for
  return the images P_{map}, D_{map}
end function

3.2.4 Denoising

NL-means and a contrario framework

In this section we apply the *a contrario* framework to the context of image denoising and propose a simple modification of the celebrated image denoising algorithm Non-Local Means (NL-means). This algorithm was introduced in the seminal paper of Buades et al. [BCM05] and was inspired by the work of Efros and Leung in texture synthesis [EL99]. It was also independently introduced in [AW06]. This



(a) (b) (c) (d) (e)

Figure 3.11: **Outputs of Algorithm 3** In (a) we present an original 256×256 image. In (b) we present the associated microtexture model given by (3.36). In (c) the green patch is the input patch, *i.e.* $P_{\mathsf{w}}(u)$. In this experiment $\mathrm{NFA}_{\max}$ is set to 1. In (d), respectively (e), we present the output P_{map}, respectively D_{map}, of Algorithm 3. In (c) we show in red the patches corresponding to the identified offsets in P_{map}.

algorithm relies on the simple idea that denoising operations can be conducted in the lifted patch space. In this space the usual Euclidean distance acts as a good similarity detector and we can obtain a denoised patch by averaging all the patches with weights that depend on this Euclidean distance. Usually the weight function is set to have exponential decay, but it was suggested in [Goo+08; Sal10; DAG10] to use compactly supported weight functions in order to avoid the loss of isolated details. Since its introduction, many algorithms derived from NL-means have been proposed in order to embed the algorithm in general statistical frameworks [DAG11; LBM13] or to take into account the underlying geometry of the patch space [HBD17]. Among the state-of-the-art denoising algorithms, see [Leb+12] for a review, we consider Block-Matching and 3D Filtering (BM3D) [Dab+07] to compare our algorithm with.

There exist several works combining *a contrario* models and denoising tasks. Coupier et al. in [CDY05] propose to combine morphological filters and a testing hypothesis framework to remove impulse noise. In [DD13] Delon and Desolneux compare different statistical frameworks to perform denoising with Gaussian noise or impulse noise. The *a contrario* model was also successfully used to deal with speckle noise [FAI05] and quasi-periodic noise [Sur15], and rely on the thresholding of wavelet or Fourier coefficients. In [KB08], Kervrann and Boulanger derive approximated probabilistic thresholds using χ_2 probability distribution functions. In [Wu+13] the authors propose a testing framework in order to estimate thresholds. The expressions they derive also relies on an approximation of the probability distribution of the squared Euclidean norm between two patches in Gaussian white noise.

Following a standard extension procedure of the NL-means algorithm we consider a threshold version of it, see Algorithm 4. In what follows we fix a "clean", or original, image x_0 defined over E, a finite rectangular domain of $\mathbb{Z}^2$, a noisy image $x = x_0 + \sigma w$, with w a realization of a standard Gaussian random field W and $\sigma > 0$ the standard deviation of the noise. In all of our experiments we assume that σ is known. Note that there exist several algorithms to estimate σ from real images, see [Pon+07] for instance. Our goal is to retrieve x_0 based on the information in x. We consider the lifted version of x in a patch space. Let w_0 be a centered 8×8 patch domain. For a patch window $\mathsf{w} = p + \mathsf{w}_0$ the patch search window T will be defined by

$$\mathsf{T} = \left\{ t \in \mathbb{Z}^2,\ t + \mathsf{w} \subset \mathsf{E},\ \|t\|_\infty \leq c \right\}\ , \tag{3.37}$$

with $c \in \mathbb{N}$. $|\mathsf{T}|$ denotes the cardinality of T. There exists a large literature concerning the setting of c and w_0, see [DAG10]. Note that the locality of the patch window was assessed to be a crucial feature of NL-means [GZW11]. For any $p \in \mathsf{E}$ we define

$$\mathsf{T}_p = \{ t \in \mathsf{E} \ :\ p \in t + \mathsf{w} \subset \mathsf{E} \}$$

Assume we have a collection of denoised patches $\hat{P}(p, \mathsf{w})$ for all patch domains w, we obtain a pixel at position p in the denoised image $\hat{x}$ using the following average, see [BCM11],

$$\hat{x}(p) = |\mathsf{T}_p|^{-1} \sum_{t \in \mathsf{T}_p} \hat{P}(x, t + \mathsf{w})(p) \ . \tag{3.38}$$

We now introduce our modification of NL-means. We assume that we are provided a threshold function a. The choice of such a function is discussed in Proposition 3.2.6.

Algorithm 4 NL-means threshold

```
function NL-means-threshold(u, σ, w_0, c, a)
    for p ∈ ℤ², p + w_0 ⊂ E do
        w ← p + w_0
        T ← defined by (3.37)
        N_w(x) ← 0
        P̂(x, w) ← 0
        for t ∈ T do
            if 𝒜(x, t, w) ≤ σ² a(t) then          ▷ always true for t = 0
                P̂(x, w) ← N_w(x)/(N_w(x)+1) P̂(x, w) + 1/(N_w(x)+1) P_{t+w}(x)   ▷ see Definition 3.1.3
                N_w(x) ← N_w(x) + 1
            end if
        end for
    end for
    x̂ ← defined by (3.38)
    return P̂(x, ·), x̂
end function
```

Note here that the output denoised version of the patch $\hat{P}(x, \mathsf{w})$ satisfies

$$\hat{P}(x, \mathsf{w}) = \sum_{t \in \mathsf{T}} \lambda_t P_{t+\mathsf{w}}(x) \ , \qquad \lambda_t = \mathbb{1}_{\mathscr{A}(x,t,\mathsf{w}) \leq a(t)} \left(\sum_{\mathsf{s} \in \mathsf{T}} \mathbb{1}_{\mathscr{A}(x,\mathsf{s},\mathsf{w}) \leq a(\mathsf{s})} \right)^{-1} .$$

In the original NL-means method, we have

$$\lambda_t = \exp\left[-\mathscr{A}(x, t, \mathsf{w})/h^2\right] \left(\sum_{t \in \mathsf{T}} \exp\left[-\mathscr{A}(x, t, \mathsf{w})/h^2\right] \right)^{-1} . \tag{3.39}$$

Setting h is not trivial and depends on many parameters (patch size, search window size, content of the original image). As in Algorithm 4, we denote $N_{\mathsf{w}}(x) = \sum_{t \in \mathsf{T}} \mathbb{1}_{\mathscr{A}(x,t,\mathsf{w}) \leq a(t)}$. The following proposition, similar to Proposition 3.2.3, gives a method for setting a. We say that an offset t is a false alarm in a Gaussian white noise if the associated patch is used in the denoising algorithm, *i.e.* if $\mathscr{A}(W, t, \omega) \leq a(t)$, similarly to Definition 3.2.2.

Some discussion is in order here. Although the definition of false alarm given here is consistent with the one given in Definition 3.2.2 it is not the classical choice for denoising applications. Indeed, if the image to denoise is a white noise, the underlying image is its mean, *i.e.* the constant image equal to zero. Therefore, in a white noise every offset should be selected and therefore an event in this background model is an offset whose associated patch is not used in the denoising algorithm, *i.e.* $\mathscr{A}(W, t, \omega) > a(t)$. This is the point of view we adopt in the paper associated with this section, see [Bor+19].

In Proposition 3.2.6 we choose a in order to control the number of false alarms with high probability.

Proposition 3.2.6. *Let* $\mathrm{NFA}_{max} \in [0, |\mathsf{T}|]$, T *given in* (3.37) *and let* $a : \mathsf{E} \to \mathbb{R}$ *be defined for any* $t \in \mathsf{E}$ *by*

$$a(t) = \mathsf{AP}^{-1}\left(t, \mathsf{w}, \mathrm{NFA}_{max}/|\mathsf{T}|\right) ,$$

with background model being a Gaussian white noise W*, i.e.* $f = \delta_0$ *in Definition 3.2.4. Let* T *be defined in* (3.37) *and* $N_{\mathsf{w}}(W) \in \{0, \dots, \mathsf{T}\}$ *the random number of selected patches used to denoise the patch* $P_{\mathsf{w}}(W)$*, see Algorithm 4. Then for any* $n \in \mathbb{N}\setminus\{0\}$ *it holds that*

$$\mathbb{P}_0\left(N_{\mathsf{w}}(W) \geq n\right) \leq \frac{\mathrm{NFA}_{max}}{n} .$$

Proof. Using the Markov inequality, we have

$$\mathbb{P}_0\left(N_{\mathsf{w}}(W) \geq n\right) \leq n^{-1} \sum_{t \in \mathsf{T}} \mathbb{E}\left[\mathbb{1}_{\mathscr{A}(W,t,\mathsf{w}) \leq a(t)}\right] \leq n^{-1}\, \mathrm{NFA}_{\max} .$$

□

In this case the null hypothesis $\mathbb{P}_0$ is given by a standard Gaussian random field, which is a special case of the Gaussian random field models introduced in Section 3.2.3. In the next proposition, using the *a contrario* framework, we obtain probabilistic guarantees on the distance between the reconstructed patch $\hat{P}(x, \mathsf{w})$ and the true patch $P_{\mathsf{w}}(x_0)$.

Proposition 3.2.7. *Let* $U = x_0 + \sigma W$*, where* W *is a standard Gaussian white noise over* E*,* $x_0 : \mathsf{E} \to \mathbb{R}$ *and* $\sigma > 0$*. Let* $p \in \mathsf{E}$ *and* $\mathsf{w} = p + \mathsf{w}_0$ *be a fixed patch and let* $\mathrm{NFA}_{max} \in [0, |\mathsf{T}|]$*. We introduce the random set* $\hat{\mathsf{T}} = \{t \in \mathsf{T} : \mathscr{A}(x, t, \mathsf{w}) \leq \sigma^2 a(t)\}$ *(the selected offsets) with* $a(t) = \mathsf{AP}^{-1}(t, \mathsf{w}, \mathrm{NFA}_{max}/|\mathsf{T}|)$ *as in Proposition 3.2.6 and* T *defined in* (3.37)*. Let* $a_{\mathsf{T}} = \max_{t \in \mathsf{T}} a(t)$*. Then for any* $a_W > 0$*, setting* $\varepsilon_W = 1 - \mathbb{P}(\|P_{\mathsf{w}}(W)\|_2^2 \leq a_W \mid \hat{\mathsf{T}})$*, we have*

$$\mathbb{P}\left(\|\hat{P}(x, \mathsf{w}) - P_{\mathsf{w}}(x_0)\|_2 \leq \sigma(a_{\mathsf{T}}^{1/2} + a_W^{1/2}) \mid \hat{\mathsf{T}}\right) \geq 1 - \varepsilon_W .$$

Proof. We have for any $t \in \hat{\mathsf{T}}$

$$\begin{aligned}
\|P_{t+\mathsf{w}}(x) - P_{\mathsf{w}}(x_0)\|_2 &\leq \|P_{t+\mathsf{w}}(x) - P_{\mathsf{w}}(x) + P_{\mathsf{w}}(x) - P_{\mathsf{w}}(x_0)\|_2 \\
&\leq \|P_{t+\mathsf{w}}(x) - P_{\mathsf{w}}(x)\|_2 + \|P_{\mathsf{w}}(x) - P_{\mathsf{w}}(x_0)\|_2 \\
&\leq \sigma a_{\mathsf{T}}^{1/2} + \sigma \|P_{\mathsf{w}}(W)\|_2 .
\end{aligned}$$

This gives the following event inclusion for any $t \in \hat{\mathsf{T}}$,

$$\left\{\|P_{\mathsf{w}}(W)\|_2 \leq a_W^{1/2}\right\} \subset \left\{\|P_{t+\mathsf{w}}(x) - P_{\mathsf{w}}(x_0)\|_2 \leq \sigma(a_{\mathsf{T}}^{1/2} + a_W^{1/2})\right\} ,$$

We also have that by definition of ε_W

$$\begin{aligned}
&\mathbb{P}\left(\|\hat{P}(x, \mathsf{w}) - P_{\mathsf{w}}(x_0)\|_2 \leq \sigma(a_{\mathsf{T}}^{1/2} + a_W^{1/2}) \mid \hat{\mathsf{T}}\right) \\
&\qquad \geq \mathbb{P}\left(\bigcap_{t \in \hat{\mathsf{T}}} \{\|P_{t+\mathsf{w}}(x) - P_{\mathsf{w}}(x_0)\|_2^2 \leq \sigma^2 (a_{\mathsf{T}}^{1/2} + a_W^{1/2})^2\} \mid \hat{\mathsf{T}}\right) \\
&\qquad \geq \mathbb{P}\left(\|P_{\mathsf{w}}(W)\|_2^2 \leq a_W \mid \hat{\mathsf{T}}\right) \geq 1 - \varepsilon_W .
\end{aligned}$$

□

In our applications we use Algorithm 4 with $a(t) = \mathsf{AP}^{-1}(t, \mathsf{w}, \mathrm{NFA}_{\max}/|\mathsf{T}|)$. Therefore we need to compute $a(t) = \mathsf{AP}^{-1}(t, \mathsf{w}, \mathrm{NFA}_{\max}/|\mathsf{T}|)$ with a Gaussian white noise background model. We recall that in Section 3.2.3, using Proposition 3.1.12, we give a method to compute this quantity in general Gaussian settings. In the case of a Gaussian white noise, the next proposition shows that the eigenvalues can be computed without approximation.

Proposition 3.2.8. *Let $t = (t_x, t_y) \in \mathbb{Z}^2 \setminus \{0\}$, C_t as in Proposition 3.1.12 with $f = \delta_0$ and $\mathsf{w} = [\![0, p-1]\!]^2$, with $p \in \mathbb{N}$. We have, expressing C_t in the basis corresponding to the raster scan order on the x-axis*

$$C_t = \begin{pmatrix} B_0 & B_1 & \dots & B_{p-1} \\ B_1^\top & B_0 & \ddots & \vdots \\ \vdots & \ddots & B_0 & B_1 \\ B_{p-1}^\top & \dots & B_1^\top & B_0 \end{pmatrix} + 2\mathrm{Id}\,, \quad \begin{cases} B_\ell = D_{|t_y|} \in \mathcal{M}_p(\mathbb{R}) & \text{if } \ell = |t_x| \\ B_\ell = 0 & \text{otherwise} \end{cases}$$

where D_j is a zero matrix with ones on the j-th diagonal. The eigenvalues of C_t are given by $\lambda_{m,k} = 4\sin^2\left(\frac{k\pi}{2m}\right)$ with multiplicity $r_{m,k}$ where $m \in [\![2, q+1]\!]$, $k \in [\![1, m-1]\!]$ and $q = \lceil \frac{p}{|t_x| \vee |t_y|} \rceil$. For any $m \in [\![2, q+1]\!]$, $k \in [\![1, m-1]\!]$ it holds

(a) for any $k' \in [\![1, m-1]\!]$, $r_{m,k} = r_{m,k'}$;

(b) $r_{m,k} = 2|t_x||t_y|$ if $2 \le m < q$;

(c) $r_{m,k} = r_x r_y$ if $m = q+1$;

(d) $\sum_{m=2}^{q+1} \sum_{k=1}^{m-1} r_{m,k} = p^2$,

with $r_x = \left(\lceil \frac{p}{|t_x|} \rceil - q\right)|t_x| + |t_x| - p_x$, where $p_x = |t_x| \lceil \frac{p}{|t_x|} \rceil - p$. We define r_y in the same manner. A similar proposition holds if $t_y \neq 0$.

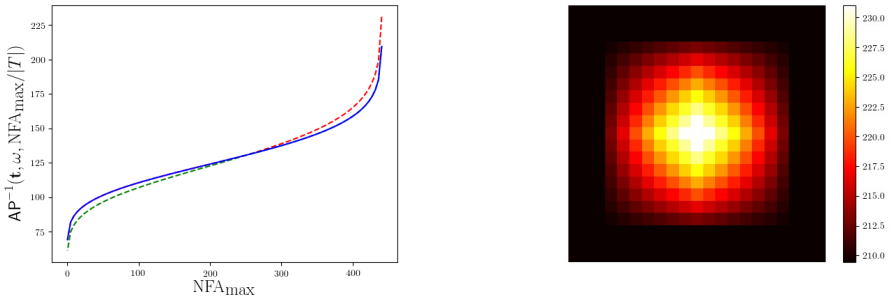
Proof. The proof is postponed to Appendix A. □

This property allows us to compute exactly the eigenvalues appearing in Proposition 3.1.12. In Figure 3.12 we illustrate that $a(t)$ for fixed patch size (8×8) and patch search window (21×21). Thus in our implementation we assume that $a(t) = \mathsf{AP}^{-1}(t, \mathsf{w}, \mathrm{NFA}_{\max}/|\mathsf{T}|)$ is constant and set its value to the mean of $a(t)$ over $t \in \mathsf{T}$.

The result in Figure 3.12-(b) might seem contradictroy at first. Indeed, in a white noise model, if two patches are spatially close one could expect that they are perceptually close (since they are correlated). Therefore one could expect that $a(t) \le a(t')$ if $\|t\|_\infty \le \|t'\|_\infty$. This is true if $\mathrm{NFA}_{\max}$ is small. However, if $\mathrm{NFA}_{\max}$ is close to $|\mathsf{T}|$ (as it is the case in Figure 3.12-(b)) then we have to take into account the fact that even if spatially close patches are correlated they are not always positively correlated. Therefore for high values of $\mathrm{NFA}_{\max}$ we have $a(t) \ge a(t')$ if $\|t\|_\infty \le \|t'\|_\infty$. From a statistical point of view, this means that the distribution of the autosimilarity is more spread for small offsets. We illustrate this property on a toy example in Section 3.2.4

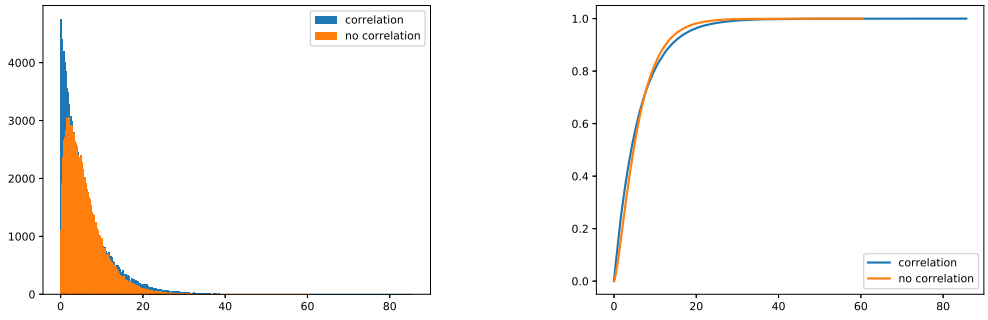
Some experimental results

In the following paragraph we present and comment some results of our threshold NL-means algorithm, see Algorithm 4. We recall that we use $a(t) = \sum_{t \in T} \mathsf{AP}^{-1}(t, \mathsf{w}, \mathrm{NFA}_{\max}/|T|)/|T|$. In Figure 3.14 we



(a) (b)

Figure 3.12: **Thresholds dependency in $\mathrm{NFA}_{\max}$** In (a) we display $a(t) = \mathsf{AP}^{-1}(t, \mathsf{w}, \mathrm{NFA}_{\max}/|\mathsf{T}|)$ as a function of $\mathrm{NFA}_{\max}$. The patch size is fixed to 8×8 and the offsets t satisfy $\|t\|_\infty \leq 10$, hence $|\mathsf{T}| = 21^2 = 441$. The red dashed line is given by $\max_{t \in \mathsf{T}} a(t)$ and the green dashed line by $\min_{t \in \mathsf{T}} a(t)$. The blue line represents the value obtained considering the simplifying assumption that patch domains do not overlap, see Proposition 3.1.12 and the remark that follows. The maximal increase between the maximum of $a(t)$ and the minimum of $a(t)$ is of 13.0%. In (b) we display the mapping $t \mapsto a(t)$ for $\mathrm{NFA}_{\max} = 440$, the central pixel corresponds to $\mathrm{t} = 0$. Note that $a(t)$ decreases as $\|t\|$ increases and is constant when, $\|t\|_\infty \geq 8$.



(a) (b)

Figure 3.13: **The distribution of the autosimilarity** We consider the following toy example. In the "no correlation" case we compute $\sum_{i=1}^{3}(X_i - Y_i)^2$ where X and Y are independent and identically distributed 3-dimensional Gaussian random variables with zero mean and identity covariance matrix. In the "correlation" case we compute $\sum_{i=1}^{3}(X_i - X_{i+1})^2$ with $X_4 = X_1$. In (a) we represent the histograms we obtain with 10^5 experiments and in (b) the corresponding empirical cumulative distribution functions. In both cases, it is clear that the distribution of the "correlation" random variable is more spread than the "no correlation" one.

(a) (b) (c) PSNR $= 29.81, \delta_t = 0.46s$ (d) PSNR $= 29.29, \delta_t = 0.37s$

Figure 3.14: **Visual results** In (a) we present an original image (Barbara) scaled between 0 and 255. In (b) we add Gaussian white noise with $\sigma = 10$. We recall that the patch domain is fixed to w_0 being a 8×8 square. In (c) we present the denoised results with NL-means threshold, Algorithm 4, where $\mathrm{NFA}_{\max} = 437$, which corresponds to 1% of rejected patches in the search window of a Gaussian white noise. In (d) we present the results obtained with the traditional NL-means algorithm with $h = 0.13\sigma|\mathsf{w}|$ (optimal h for this noise level and this image with regard to the PSNR measure). The results are the same on the texture area for (c) and (d). The perceptual results on the zoomed region are satisfying, even though some regions are too smooth compared to the original image (a). In (c) and (d), δ_t is the running time of the algorithm. We can observe that our algorithm is slightly slower than NL-means.

first present comparison with the NL-means algorithm. Perceptual results as well as Peak Signal to Noise Ratio (PSNR) measurements [1] are commented. We also present the running time of the original NL-means algorithm and ours. The experiments were conducted with the following computer specifications: 16G RAM, 4 Intel Core i7-7500U CPU (2.70GHz). Results on other images than Barbara are displayed in Figure 3.15.

If the threshold $a(t)$ is high, *i.e.* $\mathrm{NFA}_{\max} \gg |T|$ then almost no patch is rejected, which means that almost all patches are used in the denoising process. In consequence, the output denoised image is very smooth. This smoothness is a correct guess for constant patches. However, this proposition does not hold when the region contains details. Indeed, in this case details are lost due to the averaging process. By setting a conservative threshold, *e.g.* $\mathrm{NFA}_{\max}/|T| \approx 0.9$, for example, we reject all the patches for which the structure does not strongly match the one of the input patch, see Figure 3.16. This conservative property of the algorithm ensures that we can control the loss of information in the denoised image, see Proposition 3.2.7. However, if no patch, other than the input patch itself, is detected as similar we highly overfit the original noise. Many algorithms such as BM3D, see [Dab+07], solve this problem by treating this case as an exception, applying a specific denoising method in this situation. We show the differences between our version of NL-means and BM3D in Figure 3.17 .

In Figure 3.18, we show that Algorithm 4 performs better than the original NL-means algorithm. By setting $\mathrm{NFA}_{\max}/|T| = 0.99$ we obtain that the PSNR of the denoised image is better than the one of NL-means for nearly every value of h.

Let us emphasize that our goal is not to provide a new state-of-the-art denoising algorithm. Indeed we never obtain better denoising results than the BM3D algorithm. However, our algorithm slightly improves the original NL-means algorithm. It shows that statistical testing can be efficiently used to measure the similarity between patches and therefore provides a robust way to perform the weighted average in this algorithm.

[1] $\mathrm{PNSR}(x, y) = 10 \log_{10}\left(\frac{\max_{\mathsf{E}} x^2}{\|x-y\|_2^2}\right)$.

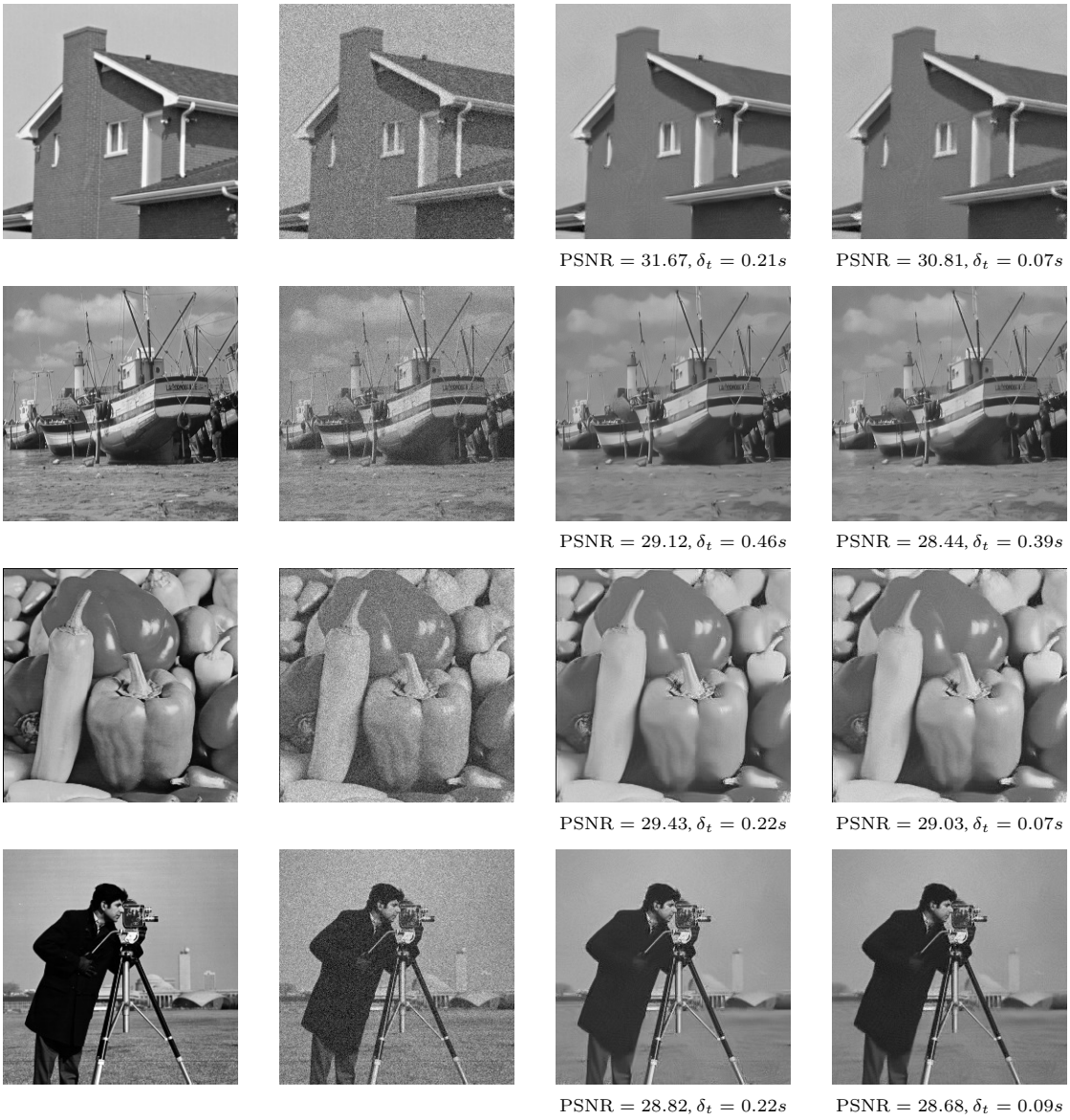
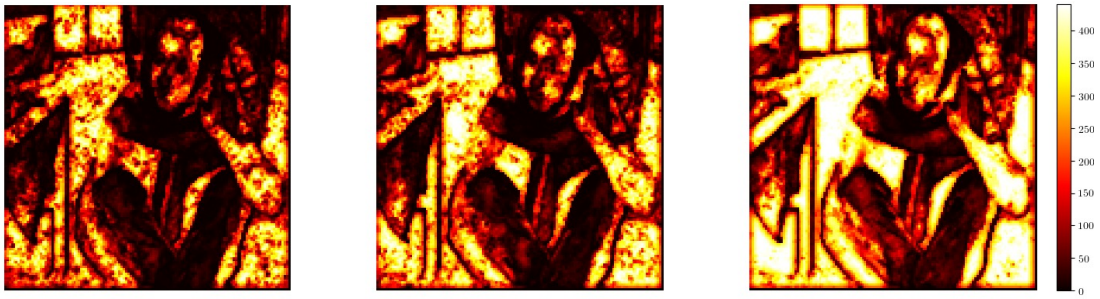


Figure 3.15: **NL-means comparison** In this figure we compare Algorithm 4 with the traditional NL-means algorithm. Here w_0 is fixed to be a 8×8 square. The first column contains clean images, the second column represents the same images corrupted by a Gaussian noise with $\sigma = 20$. The third column is the output of Algorithm 4 with $\mathrm{NFA}_{\max}$ fixed to 437 and the last column is the output of the NL-means algorithm for the optimal value of h (with regards to the PSNR), see (3.39). Perceptual results and PSNR are comparable, even though our algorithm yields slightly better PSNR values. We also present the running times δ_t of both algorithm. Our algorithm is slower than NL-means as it computes the threshold before running the NL-means algorithm.

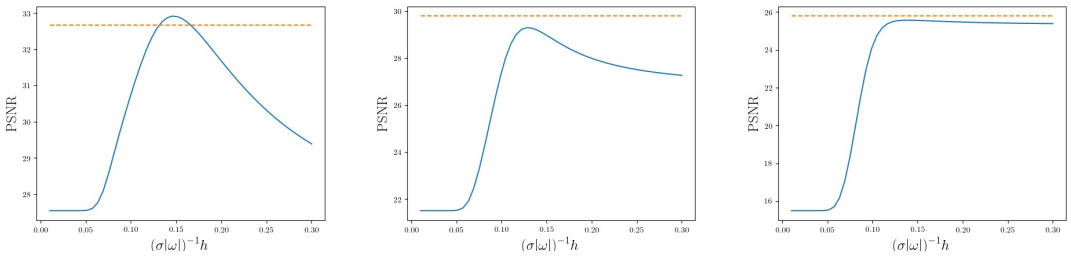


(a) $\mathrm{NFA}_{\max}/|T| = 0.8$ (b) $\mathrm{NFA}_{\max}/|T| = 0.9$ (c) $\mathrm{NFA}_{\max}/|T| = 0.99$

Figure 3.16: **Number of detections** In this figure we present, for each denoised pixel, the number of detected offsets used to compute the denoised patch, *i.e.* the cardinality of $\hat{\mathsf{T}}$, see Proposition 3.2.7. A white pixel means that the number of detected offsets is maximal and a black pixel means that the number of detected offsets is 1, *i.e.* the patch is not denoised. As $\mathrm{NFA}_{\max}$ decreases the number of detected offsets increases. Note that $|\hat{\mathsf{T}}|$ is maximal, *i.e.* equals to $21^2 = 441$, for constant regions. For $\mathrm{NFA}_{\max}/|T| = 0.9$, pixels located in textured neighborhoods use in average 20 to 40 patches to perform denoising.

(a) original (b) BM3D (c) $\mathrm{NFA}_{\max}/|T| = 0.01$ (d) $\mathrm{NFA}_{\max}/|T| = 0.1$

Figure 3.17: **Comparison with BM3D** We compare Algorithm 4 to BM3D [Dab+07]. The original image (Barbara) is presented in (a). We consider a noisy version of the input image with $\sigma = 20$. In (b) we present the ouput of BM3D, with default parameters, see [Leb12]. The result in (c) corresponds to the output of Algorithm 4 with $\mathrm{NFA}_{\max}/|T| = 0.99$. The output (c) is too blurry compared to (b). In order to correct this behavior we set $\mathrm{NFA}_{\max}/|T| = 0.9$ in (d), *i.e.* increase the global threshold and some improvements are noticeable. However the image remains blurry and artifacts due to the overfitting of the noise appear, this is known as the *rare patch effect* in [DAG11]. For instance, some patches in the scarf are not denoised anymore.



(a) $\sigma = 10$ (b) $\sigma = 20$ (c) $\sigma = 40$

Figure 3.18: PSNR **study** In this figure we present the evolution of the PSNR for different values of the parameter h of the original NL-means method, see (3.39), in blue, computed on the Barbara image. The x-axis represents $h/(\sigma|\mathsf{w}|)$. The orange dashed line is the PSNR obtained for the threshold NL-means algorithm (Algorithm 4) with $\mathrm{NFA}_{\max}/|T| = 0.99$. Except for low levels of noise the proposed method gives better PSNR values than the original implementation of NL-means algorithm for every choice of h.

3.2.5 Periodicity analysis

Existing algorithms

In the following sections we use our patch similarity detection algorithm, see Algorithm 3, to analyze images exhibiting periodicity features. Let $\mathsf{E} \subset \mathbb{Z}^2$ be a finite domain and $\mathsf{w} \subset \mathsf{E}$ a finite patch domain.

Periodicity detection is a long-standing problem in texture analysis [ZT80]. First algorithms used the quantization of images, relying on co-occurrence matrices and statistical tools like χ_2 tests or κ tests. Global methods extract peaks in the frequency domain (Fourier spectrum) [MMN83] or in the spatial domain (autocorrelation). In [HSD73] the notion of inertia is introduced. It is defined for any $t \in \mathsf{E}$ by $\mathcal{I}(t) = \sum_{(i,j)\in[\![0,N_g]\!]^2}(i-j)^2\left(\sum_{p\in\mathsf{E}}\mathbb{1}_{\dot{x}(p)=i}\mathbb{1}_{\dot{x}(p+t)=j}\right)$, where x is a quantized image on N_g+1 gray levels. In [CH80], the authors show that the local minima of the inertia measurement can be used to assess periodicity. Similarly, we introduce the w-inertia for any $t \in \mathsf{E}$ by $\mathcal{I}_{\mathsf{w}}(t) = \sum_{(i,j)\in[\![0,N_g]\!]^2}(i-j)^2\left(\sum_{p\in\mathsf{w}}\mathbb{1}_{\dot{x}(p)=i}\mathbb{1}_{\dot{x}(p+t)=j}\right)$. The following proposition extends to a local framework results from [OLS99].

Proposition 3.2.9. *Let $u \in \mathbb{R}^{\mathsf{E}}$. Suppose that u is quantized, i.e. there exists $N_g \in \mathbb{N}$ such that for any $p \in \mathsf{E}$, $u(p) \in [\![0, N_g]\!]$. We have $\mathcal{I}_{\mathsf{w}}(t) = \mathscr{A}(x,t,\mathsf{w})$.*

Proof. For any $t \in \mathsf{E}$ we have

$$\begin{aligned}\mathcal{I}_{\mathsf{w}}(t) &= \sum_{(i,j)\in[\![0,N_g]\!]^2}(i-j)^2\sum_{p\in\mathsf{w}}\mathbb{1}_{\dot{x}(p)=i}\mathbb{1}_{\dot{x}(p+t)=j}\\ &= \sum_{p\in\mathsf{w},(i,j)\in[\![0,N_g]\!]^2}(i-j)^2\mathbb{1}_{\dot{x}(p)=i}\mathbb{1}_{\dot{x}(p+t)=j} = \sum_{p\in\mathsf{w}}(\dot{x}(p)-\dot{x}(p+t))^2 = \mathscr{A}(x,t,\mathsf{w}).\end{aligned}$$

□

If $\mathsf{w} = \mathsf{E}$ then the w-inertia statistics is exactly the inertia introduced in [HSD73] and the result is due to [OLS99].

Algorithm and properties

Lattice detection is closely related to periodicity analysis, since identifying a lattice is similar to extracting periodic or pseudo-periodic structures up to deformations and approximations. A state-of-the-art algorithm proposed in [Par+09] uses a recursive framework which consists in 1) a lattice model proposal based on detectors such as Kanade-Lucas-Tomasi (KLT) feature trackers [LK81], 2) spatial tracking using inference in a probabilistic graphical model, 3) spatial warping correcting the lattice deformations in the original image. In this section we propose a new algorithm for lattice detection. The lattice proposal step 1) is replaced by an Euclidean auto-similarity matching detection (see Section 3.2.3 and Algorithm 3) where the patch domain w is fixed. Using these detections we build a graph with a few nodes (usually approximately 20 nodes for a 256×256 image). We use the same notation for the detection mapping $t \mapsto \mathbb{1}_{\mathscr{A}_i(x,t,\mathsf{w}) \leq a(t)}$ *i.e.* the D_{map} output of Algorithm 3, which is a binary function over the offsets, and the set of detected offsets. We recall that two pixel coordinates p_1 and p_2 are said to be 8-connected if $p_1 = p_2 + (\delta_1, \delta_2)$ with $\delta_1, \delta_2 \in \{-1, 0, 1\}$. The graph $\mathscr{G} = (V, E)$ is then built as follows:

- **Vertices**: for each 8-connected component, $\mathscr{C}_k$ in D_{map} we note p_k one position for which the minimum of $\mathscr{A}(x, t, \mathsf{w})$ over $\mathscr{C}_k$ is achieved. The set of vertices V is defined as $V = (p_k)_{k \in [\![1, N_{\mathscr{C}}]\!]}$ where $N_{\mathscr{C}}$ is the number of 8-connected components in D_{map} ;
- **Edges**: each vertex is linked with its four nearest neighbors in the sense of the Euclidean distance, defining four unoriented edges.

Referring to the three steps of [Par+09] we present our model to replace step 2) (*i.e.* the inference in a probabilistic graphical model) and introduce the approximated lattice hypothesis defined on a graph.

Definition 3.2.10. *Let $\mathscr{G} = (V, E)$ be a random graph with $V \subset \mathbb{R}^2$. We say that $\mathscr{G}$ follows the approximated lattice hypothesis if there exists a basis $B = (b_1, b_2)$ of $\mathbb{R}^2$ and, for each edge $\mathrm{e} \in E$, a couple of integers $(m_\mathrm{e}, n_\mathrm{e}) \in \mathbb{Z}^2$ such that e has the same distribution as $m_\mathrm{e} b_1 + n_\mathrm{e} b_2 + \sigma Z_\mathrm{e}$, with $(Z_\mathrm{e})_{\mathrm{e} \in E}$ independent standard Gaussian random variables in $\mathbb{R}^2$ and $\sigma > 0$. We denote by M the vector of all coefficients $(m_\mathrm{e}, n_\mathrm{e})_{\mathrm{e} \in E} \in \mathbb{Z}^{2|E|}$.*

Our goal is to perform inference in the statistical model defined by the following log-posterior

$$\mathscr{L}(B, M, \sigma^2 | E) = -2(|E| + 1)\log(\sigma^2) - \frac{1}{2\sigma^2} \underbrace{\left(\sum_{\mathrm{e} \in E} \| m_\mathrm{e} b_1 + n_\mathrm{e} b_2 - \mathrm{e} \|^2 + r(B, M) \right)}_{q(B, M | E)}, \tag{3.40}$$

where $r(B, M) = \delta_B \|B\|_2^2 + \delta_M \|M\|_2^2$ with $\delta_B, \delta_M \geq 0$. A discussion on the dependence of the model on the hyperparameters (δ_B, δ_M) is conducted in Figure 3.19. Finding the MLE of this full log-posterior is a non-convex, integer problem. However performing the minimization alternatively on B and M is easier since at each step we only have a quadratic function to minimize. Minimizing a positive-definite quadratic function over $\mathbb{Z}^2$ is equivalent to finding the vector of minimum norm in a lattice. This last formulation is known as the Shortest Vector Problem (SVP) which is a challenging problem [Mic01] (though it is not known if it is a NP-hard problem). We replace this minimization procedure over a lattice by a minimization over $\mathbb{R}^2$ followed by a rounding of this relaxed solution.

For any $\sigma > 0$ we denote by $\mathscr{L}_n(\sigma) = \mathscr{L}(B_n, M_n, \sigma^2 | E)$, with $n \in \mathbb{N}$, the log-posterior sequence.

Algorithm 5 Lattice detection – Alternate minimization

function ALTERNATE-MINIMIZATION(E, δ_B, δ_M, N_{it})
  $M_0 \leftarrow 0$
  $B_0 \leftarrow$ initialization procedure ▷ initialization discussed in the end of Section 3.2.5
  for $n \leftarrow 0$ to $N_{it} - 1$ **do**
    $\tilde{M} \leftarrow \underset{M \in \mathbb{R}^{2|E|}}{\operatorname{argmin}}\ q(B_n, M|E)$ ▷ expression given in Proposition 3.2.11
    if $q\left(E|B_n, [\tilde{M}]\right) < q\left(E|B_n, M_n\right)$ **then** ▷ $[\cdot]$ is the nearest integer operator
      $M_{n+1} \leftarrow [\tilde{M}]$
    else
      $M_{n+1} \leftarrow M_n$
    end if
    $B_{n+1} \leftarrow \underset{B \in \mathbb{R}^4}{\operatorname{argmin}}\ q(B, M_{n+1}|E)$ ▷ expression given in Proposition 3.2.11
  end for
  $\sigma^2_{N_{it}} \leftarrow \underset{\sigma^2 \in \mathbb{R}_+}{\operatorname{argmin}}\ -\mathscr{L}(B_{N_{it}}, M_{N_{it}}, \sigma^2|E)$
  return $B_{N_{it}}, M_{N_{it}}, \sigma^2_{N_{it}}$
end function

Proposition 3.2.11 (Alternate minimization update rule). *In Algorithm 5, we get for any* $n \in \mathbb{N}$

$$\tilde{M} = \left(\Lambda_{B_n} \otimes \mathrm{Id}_{|E|}\right)^{-1} E_{B_n} \in \mathbb{R}^{2|E|}\,, \qquad B_{n+1} = \left(\Lambda_{M_{n+1}} \otimes \mathrm{Id}_2\right)^{-1} E_{M_{n+1}} \in \mathbb{R}^4\,,$$

with $\otimes$ *the tensor product between matrices and*

(a) $\Lambda_B = \begin{pmatrix} \|b_1\|^2 + \delta_B & \langle b_1, b_2\rangle \\ \langle b_1, b_2\rangle & \|b_2\|^2 + \delta_B \end{pmatrix}$, $\Lambda_M = \begin{pmatrix} \|M_1\|^2 + \delta_M & \langle M_1, M_2\rangle \\ \langle M_1, M_2\rangle & \|M_2\|^2 + \delta_M \end{pmatrix}$;

(b) $E_B = \begin{pmatrix} (\langle \mathrm{e}, b_1\rangle)_{\mathrm{e}\in E} \\ (\langle \mathrm{e}, b_2\rangle)_{\mathrm{e}\in E} \end{pmatrix}$, $E_M = \begin{pmatrix} \sum_{\mathrm{e}\in E} m_{\mathrm{e}}\mathrm{e} \\ \sum_{\mathrm{e}\in E} n_{\mathrm{e}}\mathrm{e} \end{pmatrix}$.

Proof. The proof is postponed to Appendix B. □

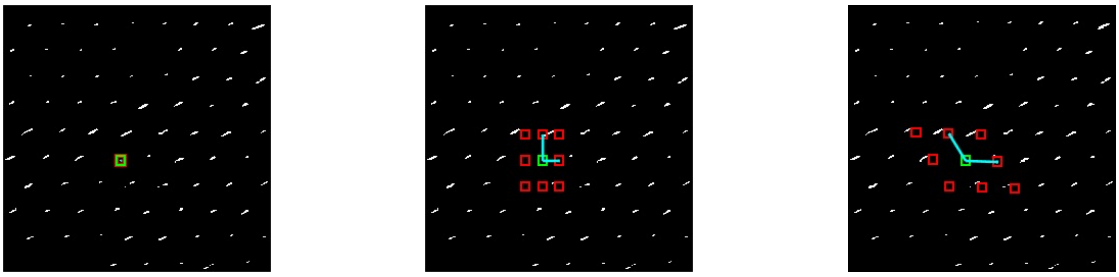
Note that if B is orthogonal, *i.e.* $\langle b_1, b_2\rangle = 0$ then Λ_B is diagonal and the proposed method is the exact solution to the minimization problem over $\mathbb{Z}^2$.

Theorem 3.2.12 (Convergence in finite time). *For any* $\sigma > 0$, $(\mathscr{L}_n(\sigma))_{n\in\mathbb{N}}$ *is a non-decreasing sequence. In addition,* $(B_n)_{n\in\mathbb{N}}$ *and* $(M_n)_{n\in\mathbb{N}}$ *converge in a finite number of iterations.*

Proof. $(\mathscr{L}_n(\sigma))_{n\in\mathbb{N}}$ is non-decreasing since for any $n \in \mathbb{N}$, $\mathscr{L}_n(\sigma) \leq \mathscr{L}(B_n, M_{n+1}, \sigma^2|E) \leq \mathscr{L}_{n+1}(\sigma)$. Let us show that the sequences $(M_n)_{n\in\mathbb{N}}$ and $(B_n)_{n\in\mathbb{N}}$ are bounded. Because $(\mathscr{L}_n(\sigma))_{n\in\mathbb{N}}$ is non-decreasing, the sequence $(q(B_n, M_n|E))_{n\in\mathbb{N}}$ is non-increasing. We obtain that

$$\delta_M \|M_n\|^2 \leq q(B_0, M_0|E)\,, \qquad \delta_B \|B_n\|^2 \leq q(B_0, M_0|E)\,.$$

The sequence $(M_n)_{n\in\mathbb{N}}$ is bounded thus we can extract a converging subsequence. Since $(M_n)_{n\in\mathbb{N}}$ takes value in $\mathbb{Z}^{2|E|}$, this subsequence is stationary with value M. Let $n_0 \in \mathbb{N}$ be the first time we hit



(a) $\delta_M = 0$ $\delta_B = 0$ (b) $\delta_M = 5$ $\delta_B = 10^{-1}$ (c) $\delta_M = 9$ $\delta_B = 10^{-1}$

Figure 3.19: **Influence of hyperparameters** In this experiment we assess the importance of the hyperparameters. We consider Algorithm 5 with input graph a detection map, output of Algorithm 3. The initialization in the three cases is the canonical basis $((0,1),(1,0))$. In (a), since the initial basis vectors are a local minimum to the optimization problem, the algorithm converges after one iteration. However, this is not perceptually satisfying. Setting $\delta_M = 5$ and $\delta_B = 10^{-1}$ in (b) the true observed lattice is a sub-lattice of the output lattice of Algorithm 5. Increasing δ_M up to 9, in (c) we obtain a perceptually correct lattice. For δ_M larger than 10, the basis vectors go to 0. Only the regularizing term is minimized by the optimization procedure and the data-attachment term is not taken into account. Experimentally we found that the choice of δ_M is more flexible and that $\delta_M \in (1, 20)$ gives satisfying perceptual results if the initialization heuristics proposed in Section 3.2.5 is chosen.

value M. Let $n \in \mathbb{N}$, with $n \geq n_0 + 1$, there exists $n_1 \in \mathbb{N}$, with $n_1 \geq n$ such that $M_{n_1} = M_{n_0}$ thus

$$\mathscr{L}_{n_0}(\sigma) \leq \mathscr{L}_{n_0+1}(\sigma) \leq \mathscr{L}_n(\sigma) \leq \mathscr{L}(B_{n_1-1}, M_{n_1}, \sigma^2|E) \leq \mathscr{L}(B_{n_1-1}, M_{n_0}, \sigma^2|E) \leq \mathscr{L}_{n_0}(\sigma) \ .$$

Hence for every $n \geq n_0 + 1$, $\mathscr{L}_n(\sigma) = \mathscr{L}(B_n, M_n, \sigma^2|E) = \tilde{\mathscr{L}}(\sigma)$. Suppose there exists $n \geq n_0 + 1$ such that $M_n \neq M_{n+1}$ this means that $\mathscr{L}(B_n, M_{n+1}, \sigma^2|E) > \mathscr{L}_n(\sigma)$ (because of lines 6 and 7 of Algorithm 5) which is absurd. Thus $(M_n)_{n\in\mathbb{N}}$ is stationary and so is $(B_n)_{n\in\mathbb{N}}$. □

In Algorithm 5 M_0 is initialized with zero and B_0 is defined as an orthonormal (up to a dilatation factor) direct basis where the first vector is given by an edge with median norm in E.

Experimental results

Combining the results of Section 3.2.5 and Section 3.2.3 we obtain an algorithm to extract lattices in images, see Figure 3.20. In what follows we perform lattice detection using Algorithm 3 in order to extract auto-similarity given a patch in an original image u, which implies that the patch domain w is set by the user. Recall that in Algorithm 3, the eigenvalues of the covariance matrix in Proposition 3.1.12 are approximated, and that the cumulative distribution function of the quadratic form in Gaussian random variables is computed via the Wood F method [Woo89]. Lattice detection is performed using Algorithm 5 with parameters $\delta_M = 10$ and $\delta_B = 10^{-2}$.

Escher paving In this section we study art images, Escher pavings, with strongly periodic structure. We investigate the following parameters of our lattice detection algorithm:redundancy2:

(a) background microtexture model $\mathbb{P}_0$,

(b) NFA_{max} parameter in Algorithm 3,

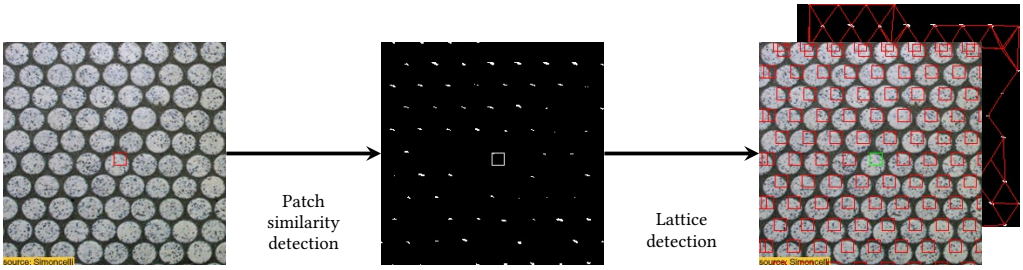


Figure 3.20: **Lattice proposal algorithm** Lattice detection and extraction in images require a patch from the user and compute a binary image containing all the offsets with correct similarity as well as a lattice matching the underlying graph. The patch auto-similarity detection step was presented in Section 3.2.3. The lattice detection step was presented in Section 3.2.5. The first image is the input, the second one is the output of the detection algorithm. In the last step we show the original image with red squares placed on the computed lattice. Behind this image, the unoriented edges of the graph are shown in red.

(c) patch domain w.

Microtexture model We confirm that the choice of the microtexture model will influence the detected geometrical structures. The more structured is the background noise model the less we obtain detections. This situation is considered in Figure 3.21.

$\mathrm{NFA}_{\max}$ parameter Using a more adapted microtexture model as background model we gain robustness compared to other less structured models such as a Gaussian white noise. However, $\mathrm{NFA}_{\max}$ must be set carefully otherwise two situations can occur:

(a) if $\mathrm{NFA}_{\max}$ is too high, too many detections can be obtained (true perceptual detections are not differentiated from false positives) ;

(b) if $\mathrm{NFA}_{\max}$ is too low, we fail to identify important perceptual structures in the image.

We observe that a general good practice is to set $\mathrm{NFA}_{\max}$ equal to 10, see Figure 3.22. However, if the input patch is corrupted one may increase this parameter up to 10^2 or 10^3, see Figure 3.27 and Figure 3.28.

Patch position Patch position and size are crucial in our detection model, since we rely on local properties of the image. As shown in Figure 3.23 these parameters should be carefully selected by the user. However, for particular applications such as lattice extraction for crystallographic purposes, there exist procedures to extract primitive cells [MB15].

Crystallography images Defect localization, noise reduction, correction of crystalline structures in images are central tasks in crystallography. Usually, they require the knowledge of the geometry of a perfect underlying crystal. In our experiments we manually identify the geometry of the periodic crystal, which allows for multiple structures in one image, provided a user input of the primitive cell in a lattice. This primitive cell extraction could be automated [MB15]. In Figure 3.24, we present an example of multiple geometry extraction. Statistics like angle and period can be retrieved using the

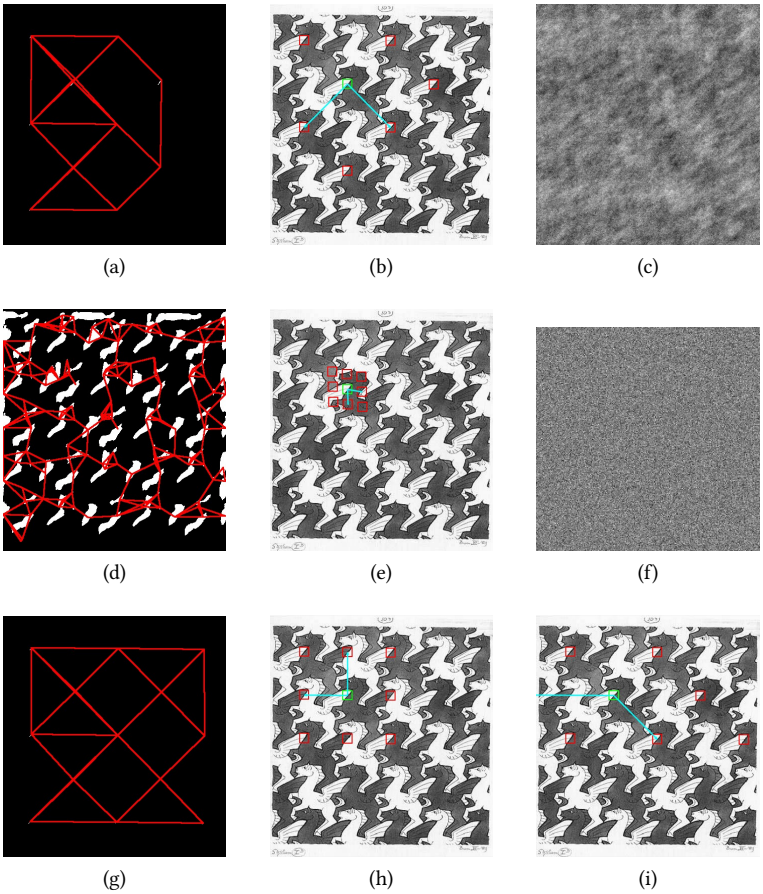
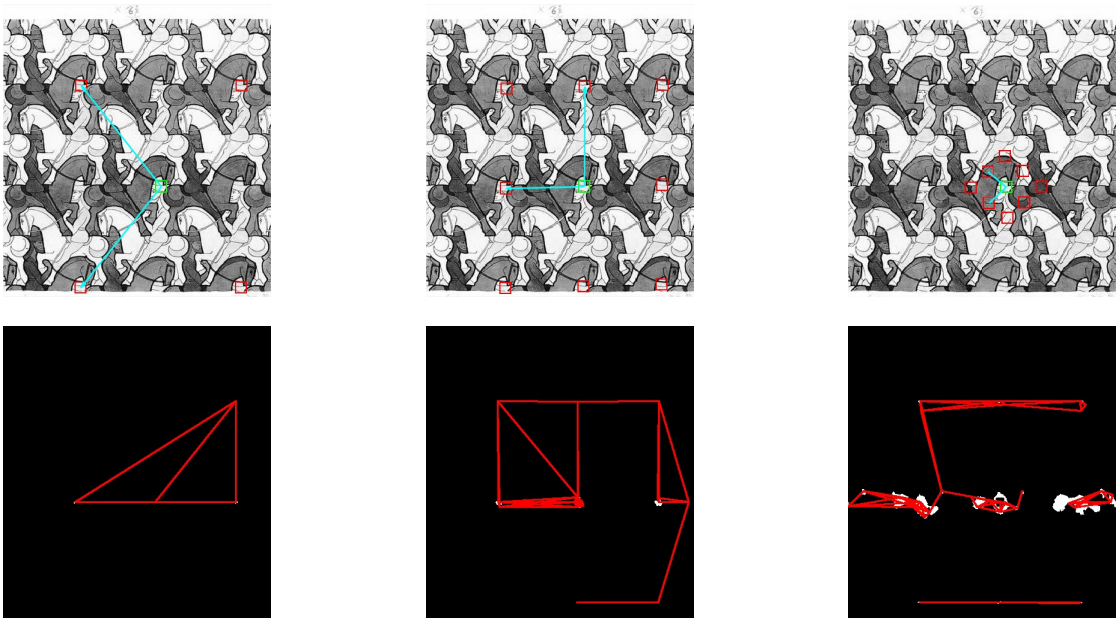
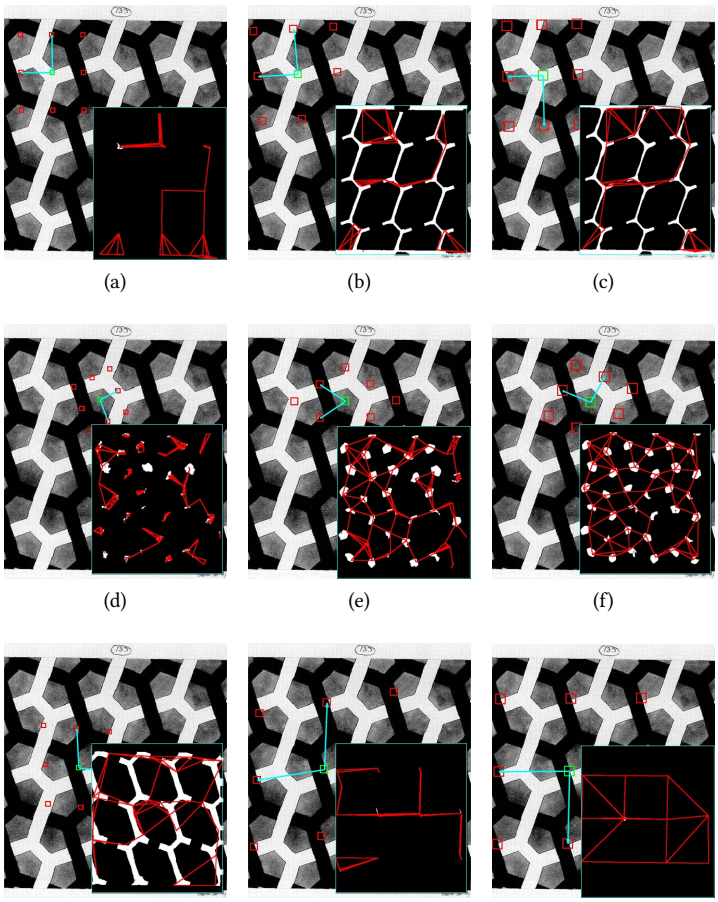


Figure 3.21: **Choice of the microtexture model** In this experiment we discuss the choice of the *a contrario* background microtexture model. In the left column we display the graph obtained after the detection step. In the middle column we superpose the proposed lattice on the original image. The original patch is drawn in green, obtained basis lattice vectors are in cyan, and red squares are placed onto the proposed lattice. In (a) and (b) the microtexture model is given by (3.36) and $\mathrm{NFA}_{\max}$ is set to 10. A sample of this model is presented in (c). Obtained results match the perceptual lattice. In (d), (e), (g) and (h) the microtexture model is a Gaussian white noise model with variance equal to the empirical variance of the original image. Sample from this Gaussian white noise is presented in (f). In (d) and (e), $\mathrm{NFA}_{\max}$ is set to 10. This leads to an excessive number of detections in the input image. In order to obtain the perceptual lattice found in (b) with a Gaussian white noise model we must set the $\mathrm{NFA}_{\max}$ parameter to 10^{-111}. Results are presented in experiments (g), (h) and (i). Image (h) is also an example for which the median initialization for B_0 in Algorithm 5 identifies a non satisfying local minimum. This situation is corrected in (i) with random initialization for B_0. In (h) final log-posterior value is -565.5 which is inferior to the final log-posterior value in (i): -542.1. Thus (i) gives a better local maximum of the full log-posterior than (h).



(a) (b) (c)

Figure 3.22: **Choice of Number of False Alarms** In this experiment we discuss the choice of the $\mathrm{NFA}_{\max}$ parameter in the *a contrario* framework in the case where the underlying microtexture model is given by (3.36). Each column corresponds to a pair of images: the returned lattice and its associated underlying graph. In (a), $\mathrm{NFA}_{\max}$ is set to 1. Detections are correct, there are not enough points to precisely retrieve the perceptual lattice. In (b), $\mathrm{NFA}_{\max}$ is set to 10. The estimated lattice is correct. In (c), $\mathrm{NFA}_{\max}$ is set to 10^3. In this case we obtain false detections which lead to an incorrect final lattice. Note that large detection zones in the binary image (c) are due to the non-validity of the Wood F approximation for some offsets. This behavior is also present in (a) and (b) but less noticeable.

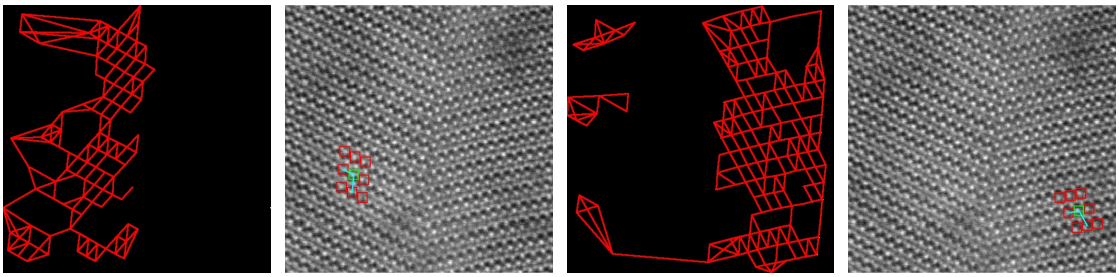


(a) (b) (c)

(d) (e) (f)

(g) 10×10 (h) 15×15 (i) 20×20

Figure 3.23: **Influence of patch size and patch position** For each experiment $\mathrm{NFA}_{\max}$ is set to 10^4, *i.e.* 4 % of the pixels. In most cases lower $\mathrm{NFA}_{\max}$ could be used but setting a high $\mathrm{NFA}_{\max}$ ensures that we always get detections even if the patch only contains microtexture information. Each row corresponds to a lattice proposal with same patch position but different patch sizes: 10×10 for the left column, 15×15 for the middle one and 20×20 for the right one. Each image represents the superposed proposed lattice on the original image. On the bottom-right of each image we display the underlying graph as well as the binary detection. On the first row the patch contains only a white region with a few gray pixels. The influence of these pixels is visible for small patch sizes (a) but is no longer taken into account for larger patch sizes, (b) and (c). On the second row the patch contains gray microtexture which has some local structure. We identify large similarity regions and no perceptual lattice is retrieved in (d), (e) and (f). The situation is different on the third row. The 10×10 patch contains only uniform black information in (g), but the situation changes as the patch sizes grows. In (h), the patch intersects black, gray and white zones. The graph is much sparser and the lattice is close to the perceptual one. In (i), the patch size is large enough to cover large areas of the three gray levels and the perceptual lattice is identified.



(a) (b) (c) (d)

Figure 3.24: **Lattice extraction** In this experiment we consider a crystallographic image (an orthorombic NiZr alloy) and set $\mathrm{NFA}_{\max}$ to 10^2. Two lattices are present in this image and they are correctly identified in (b) and (d). Note that in (a), respectively in (c), mostly points in the left, respectively right, part of the image are identified, thus yielding correct lattice identification. Points which should be identified and are discarded nonetheless correspond to regions in which we observe contrast variation. Image courtesy of Denis Gratias.

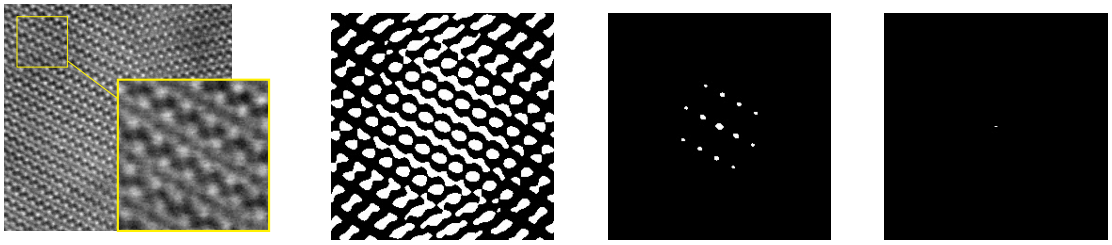
estimated basis vectors. This image contains two lattices and the locality of our measurements allows for the detection of both structures. Using windowed Fourier transform could be efficient to obtain local measurements on the periodicity of these images since the information is highly frequential. However in order to obtain the same detection map as Algorithm 3 one must carefully set the threshold parameter, $\mathrm{NFA}_{\max}$. This situation is illustrated in Figure 3.25.

Finally we assess the precision of our measurements by comparing our results with a model used in crystallography, see Figure 3.26. We indeed retrieve one of the possible bases used to describe these lattices. However, the symmetry constraints are not present in the identified basis. To obtain another basis, one must relax the regularization parameters. A more natural way to obtain the desired primitive cell would be to introduce symmetry constraints in the graphical model formulation in (3.40).

Natural images Identifying lattices in natural images is a more challenging task since we have to deal with image artifacts. In this section we investigate the effect on the detection of the background clutter in natural images, see Figure 3.27, and the effect of the camera position, see Figure 3.28.

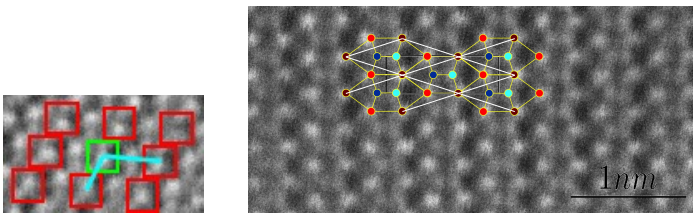
Preprocessing Due to the occlusions occurring in natural images, if a lattice is superposed over a real photograph, carefully selecting structural elements might not be enough in order to retrieve the periodicity. Indeed, if we observe a repetition of the lattice pattern, the background does not necessarily contain any repetition and thus makes the detection more complicated. In order to avoid such a problem we propose to introduce a preprocessing step in our algorithm. This preprocessing step will be encoded in a linear filter h. Suppose U is a sample from a Gaussian model with function f then $h * U$ is a sample from a Gaussian model with function $h * f$. Thus all the properties derived earlier remain valid with this linear operation. In Figure 3.27, we set h to be a Laplacian operator [2]. This operation allows us to avoid contrast issues.

[2]We use a discrete Laplacian operator Δ such that for any $p = (x_1, x_2)$, we get that $\Delta(x)(p_1, p_2) = (x(p_1 + 1, p_2) + x(p_1 - 1, p_2) + x(p_1, p_2 + 1) + x(p_1, p_2 - 1) - 4x(p))/4$, where boundaries are handled periodically.



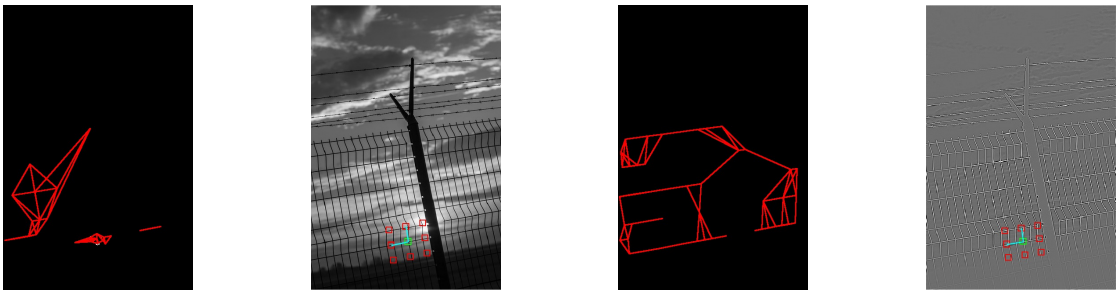
(a) (b) 90% (c) 95% (d) 99%

Figure 3.25: **Comparison with Fourier based methods** Since the original image can be segmented in two highly periodic components, Fourier methods might be well-adapted to the lattice extraction task. In (a) we present a sub-image of the original alloy. We compute the autocorrelation of this sub-image and threshold it. This operation gives us a detection map, like Algorithm 3. In (b) the threshold is set to 90% percent of the maximum value of the autocorrelation. Too many points are identified. In (d) the threshold is set to 99% and only one point is identified. Correct lattice is identified in (c).



(a) (b)

Figure 3.26: **Agreement with crystallography models** In (a) we perform a zoom on of the lattice identified in Figure 3.24 and compare it to the one identified by crystallographists in (b). (a) is a zoomed rotated version of a crystalline structure similar to (b). The output lattice in (a) is the same as the one in (b). Indeed in (b) the red points, for instance, form a lattice. A possible basis for this lattice is given by the vectors of a parallelogramm. Up to rotation these basis vectors match the one identified in (a). However, the parallelogramm basis is a symmetric and thus is not chosen by chemists since it does not reflect the geometry of the alloy. The preferred basis is given by the symmetric rhombus (white edges in (b)). Image courtesy of Denis Gratias.



(a) (b) (c) (d)

Figure 3.27: **Preprocessing and filtering** In (a) and (c) we display the graphs obtained with Algorithm 3 applied on images (b) and (d). In (b) and (d) the original image is superposed with the estimated lattice (vectors in cyan and proposed patches in red). In (a) and (b), $\mathrm{NFA}_{\max}$ was set to 10^5 which corresponds to 35 % of detection in the associated *a contrario* model. Lower $\mathrm{NFA}_{\max}$ did not give enough points to conduct the lattice proposal step. We obtain a visually satisfying lattice. In (c) and (d) we apply a simple preprocessing, a Laplacian filter, to the image and set $\mathrm{NFA}_{\max}$ to 10. The detection figure is much cleaner and the estimation makes much more sense from a perceptual point of view. Note that, as in (b), the proposed lattice does not exactly match the fence periodicity. This is due to: 1) the initialization of the algorithm and the structure of the graph in the alternate minimization algorithm 2) the fact that the horizontal periodicity is broken by the black post.

Homography In the previous experiments we suppose that the lattice structure was in front of the camera. In many cases this assumption is not true and there exists an homography that matches the deformed lattice in the image to a true lattice. Our algorithm makes the assumption that the lattice is viewed in a frontal way and fails otherwise. However, locally, this assumption is true and we can observe partial match of the lattices in Figure 3.28.

Texture ranking

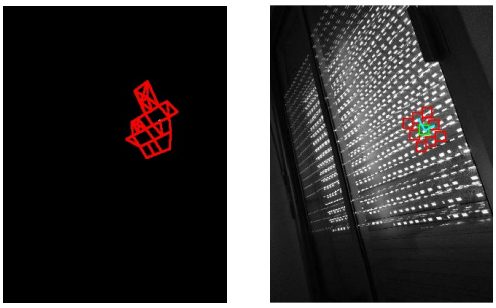
We conclude these experiments by showing that this simple graphical model can be used to perform ranking among texture images, sorting them according to their degree of periodicity. We say that an image has high periodicity degree if a lattice structure can be well fitted to the image. We introduce a criterion for evaluating the relevance of the lattice hypothesis. Let u be an image over E, let $\mathsf{w} \subset \mathsf{E}$ be a patch domain and a be as in Proposition 3.2.3 with $\mathrm{NFA}_{\max}$ set by the user.

Definition 3.2.13. *Let $\{t \in \mathsf{E},\ \mathscr{A}(x,t,\mathsf{w}) \leq a(t)\}$ be the set of detected offsets and $N_{\mathscr{C}}$ its number of connected components as defined in Section 3.2.5. Let also $(\widehat{B}, \widehat{M}, \widehat{\sigma})$ be the estimated parameters using Algorithm 5. We define the periodicity criterion c_{per} as*

$$c_{per}(x) = \frac{\pi \widehat{\sigma}^2}{N_{\mathscr{C}} |\det(\hat{b}_1, \hat{b}_2)|}\ , \tag{3.41}$$

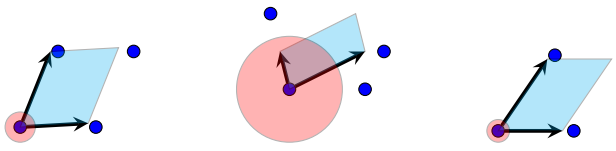
where $\widehat{B} = (\hat{b}_1, \hat{b}_2)$.

The criterion c_{per} simply computes the ratio between the error area of Algorithm 5, *i.e.* the error made when considering the approximated lattice hypothesis, see Definition 3.2.10, and the area of the



(a) (b)

Figure 3.28: **Homography and locality** In this experiment $\mathrm{NFA}_{\max}$ was set to 10^3. Note that the detected graph is localized around the original patch in (a). In (b) we superpose the proposed lattice onto the original image. The lattice proposal is valid in a small neighborhood around the original patch. However it is not valid for the whole image.



(a) (b) (c)

Figure 3.29: **Parameters configurations** In each figure, dark blue points correspond to the vertex of the graph $\mathscr{G}$. Arrows represent the estimated basis $\widehat{B}$ and its associated parallelogram is displayed in blue. Red disks have radius $\widehat{\sigma}$. In (a) there exists a lattice in the graph which corresponds to some perceptual lattice on the vertices. In this case $\pi\widehat{\sigma}^2/(|\det(\hat{b}_1, \hat{b}_2)|)$, the ratio of the uncertainty area and the primitive area, is small. In (b) no perceptual lattice is identified onto the vertices and the estimated lattice vectors are not valid. This is expressed with a high ratio $\pi\widehat{\sigma}^2/(|\det(\hat{b}_1, \hat{b}_2)|)$. In (c), the graph $\mathscr{G}$ contains only three vertices. Thus the lattice approximation is nearly optimal (up to regularization factors) and the ratio $\pi\widehat{\sigma}^2/(|\det(\hat{b}_1, \hat{b}_2)|)$ is very small. However, three points are not perceptually identified as a lattice. Hence considering c_{per} we take into account the number of points and we may retrieve that (a) is considered more periodic than (b) and (c).

parallelogram defined by the output basis vectors. If we have enough detections this quantity is supposed to be small when the approximated lattice hypothesis holds and large when it does not. Nonetheless, we introduce a dependency in the number of detections. Indeed, even if no lattice is perceived, the hypothesis in Definition 3.2.10 may still hold if the number of detected offsets is small, see Figure 3.29.

In the experiment presented in Figure 3.30 we sort 25 texture images based on the c_{per} criterion. Images are of size 256×256. Since the identified graph highly depends on the patch position and the patch size, for each image we uniformly sample 150 patch positions and set the patch size to 20×20. For each set of parameters we find a lattice using Algorithm 3 and Algorithm 5 with parameters $\mathrm{NFA}_{\max} = 1$, $\delta_M = 10$, $\delta_B = 10^{-2}$ and $N_{it} = 10$. A statistical study of our ranking is presented in Figure 3.31. Note that, from a perceptual point of view, from (a) to (n) all textures are periodic except for (f), (j) and (k) which are examples for which our algorithm fails. However, from (o) to (y), no texture is periodic.

3.2.6 Summary

In this section we introduce a statistical model, the *a contrario* framework, to analyze spatial redundancy in images. We propose a general algorithm for detecting redundancy in natural images. It relies on Gaussian random fields as background models and takes advantage of the links between the ℓ^2 norm and Gaussian densities. The *a contrario* formulation provides us with a statistically sound way of thresholding distances in order to assess similarity between patches. In this rationale we replace the task of manually setting thresholds by the selection of a Number of False Alarms.

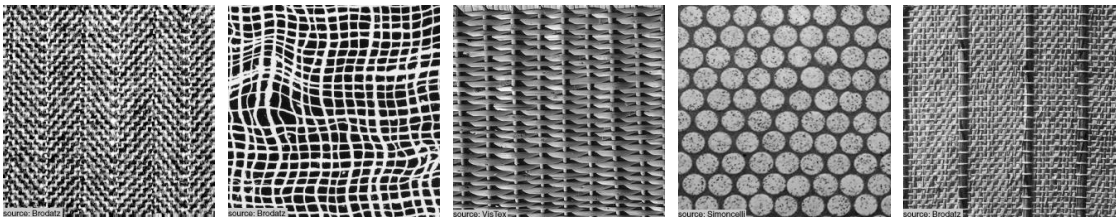
We illustrate our contribution with three examples in various domains of image processing. Introducing a simple modification of the NL-means algorithm we show that similarity detection (in this case, dissimilarity detection) in a theoretical *a contrario* framework can easily be embedded in any image denoising pipeline. For instance, the threshold we introduced could be integrated into the Non-Local Bayes algorithm [LBM13] in order to estimate mean and covariance matrices with probabilistic guarantees. The generality of our model allows for several extensions for non-Gaussian noises [DDT09] or to take into account the geometry of the patch space [HBD17; WM13].

Turning to periodicity detection we propose a novel graphical model using the output of Algorithm 3 in order to extract lattices from images. In this model, lattice extraction is formulated as the maximization of some log-likelihood defined on a graph. We prove the finite-time convergence of Algorithm 5. We provide image experiments illustrating the role of the hyperparameters in our study and we assess the importance of selecting adaptive Gaussian random fields as background models. We remark that the expected number of false alarms parameter is linked to the choice of the input patch and give a range of possible values for $\mathrm{NFA}_{\max}$ settings. We also illustrate its possible application in crystallography as it correctly identifies underlying lattices in alloys. This rationale could be used to identify symmetry groups (wallpaper groups) in alloys, following the work of [LCT03]. Finally our method is tested on natural images where some of its limits such as perspective defect or sensitivity to occlusion phenoma are identified. It must be noted that our method could easily be extended to color images by considering $\mathbb{R}^3$-valued instead of real-valued images. Our last application consists in giving a quantitative criterion for periodicity texture ranking. This criterion is based on the parameters estimated in Algorithm 5.

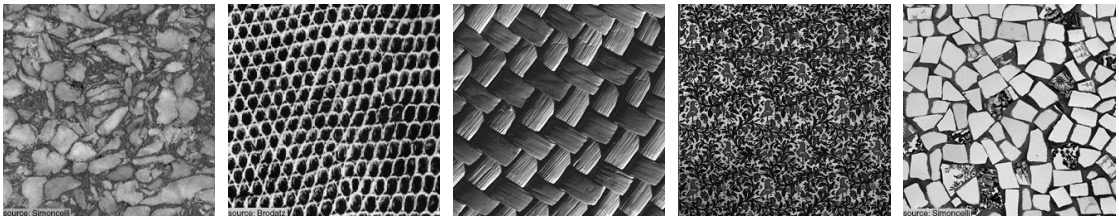
3.2.7 Proofs and additional results

Eigenvalues

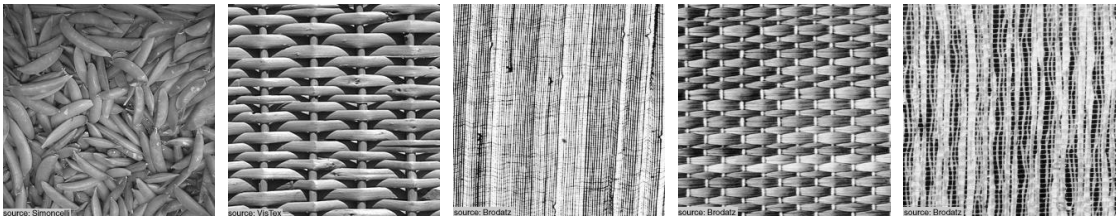
Proof of Proposition 3.2.8. We fix $\mathrm{t} \neq 0$ with $\|\mathrm{t}\|_\infty < p$ and denote $C = C_{\mathrm{t}}$. Without loss of generality we consider that $t_x > 0$ and $t_y > 0$. We consider X an eigenvector of C. Let $\Omega_0 = (\Omega - \mathrm{t}) \cap \Omega^c$ and



(a) -9.75 (b) -9.42 (c) -9.12 (d) -9.00 (e) -8.80



(f) -8.24 (g) -8.24 (h) -7.99 (i) -7.80 (j) -7.77



(k) -7.74 (l) -7.72 (m) -7.47 (n) -7.26 (o) -7.21

(p) -7.20 (q) -7.19 (r) -7.17 (s) -6.92 (t) -6.86

(u) -6.78 (v) -6.65 (w) -6.56 (x) -6.30 (y) -6.16

Figure 3.30: **Texture ranking** The c_{per} criterion, defined in (3.41), is computed for each setting. We associate to each image the median of the 150 criterion values and sort the images accordingly. (a) corresponds to the lowest criterion, *i.e.* the most periodic image according to c_{per} criterion. (y) corresponds to the largest criterion, *i.e.* the least periodic image according to c_{per}. Under each image we give the logarithm of the median c_{per} values.

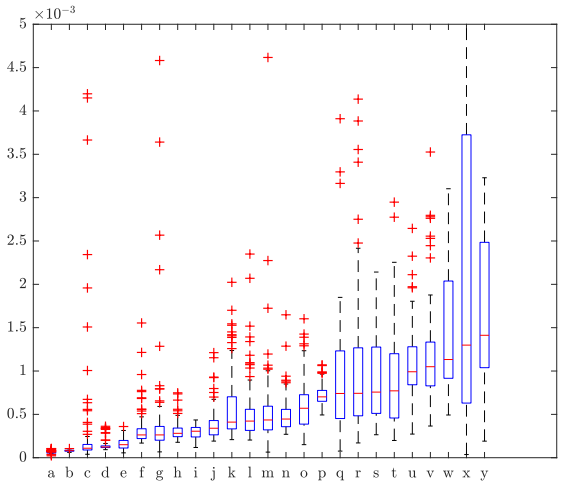


Figure 3.31: **Boxplot for c_{per} values** In this figure we present a boxplot of the c_{per} values, defined in (3.41), used to rank textures images in Figure 3.30. We recall that we use 150 random patch positions in order to compute the c_{per} values. Letters on the x-axis correspond to the textures in Figure 3.30. For each texture we present its median c_{per} value. The lower, respectively upper, limit of the blue box corresponds to 25%, respectively 75% of the computed c_{per} values. The dashed line corresponds to the confidence interval with level 0.07 under normality assumption. Points outside this interval are plotted in red and the graphics was clipped between 0 and 5×10^{-3}. The size of the confidence interval grows with the median value. It must be noted that the overlapping of the blue boxes might explain some inconsistencies of our ranking. Another source of errors lie in the model which assumes that if a texture is periodic its pattern is described by a 20×20 patch. In order to perform a more robust ranking a multiscale approach should be preferred.

the function $J:\ \Omega_0 \to [\![2,+\infty[\![$ such that for any $\mathrm{x}_0 \in \Omega_0$

$$J(\mathrm{x}_0) = \arg\min\{k \in \mathbb{N}\setminus\{0\}\,,\ \mathrm{x}_0 + k\mathrm{t} \notin \Omega\}\ .$$

It is clear that $I = \{(k,m), k \in [\![1, m-1]\!],\ m \in J(\Omega_0)\}$ is in bijection with Ω. Let $\mathrm{x}_0 \in \Omega_0$, $m = J(\mathrm{x}_0)$ and $k \in [\![1, m-1]\!]$. We define $X_{\mathrm{x}_0,k}$ over $\mathbb{Z}^2$ such that

$$X_{\mathrm{x}_0,k}(\mathrm{x}_0 + \ell\mathrm{t}) = \sin\left(\frac{\ell k\pi}{m}\right) \text{ for } \ell \in [\![1, m-1]\!]\ , \qquad 0 \text{ otherwise}\ .$$

Using that $\sin(a+b)+\sin(a-b) = 2\sin(a)\cos(b)$, we have for any $\mathrm{x} \in \mathbb{Z}^2$

$$X_{\mathrm{x}_0,k}(\mathrm{x}+\mathrm{t}) - 2\cos\left(\frac{k\pi}{m}\right) X_{\mathrm{x}_0,k}(\mathrm{x}) + X_{\mathrm{x}_0,k}(\mathrm{x}-\mathrm{t}) = 0\ .$$

This implies that for any $\mathrm{x} \in \mathbb{Z}^2$

$$2X_{\mathrm{x}_0,k}(\mathrm{x}) - X_{\mathrm{x}_0,k}(\mathrm{x}+\mathrm{t}) - X_{\mathrm{x}_0,k}(\mathrm{x}-\mathrm{t}) = \left[2 - 2\cos\left(\frac{k\pi}{m}\right)\right] X_{\mathrm{x}_0,k}(\mathrm{x}) = 4\sin^2\left(\frac{k\pi}{m}\right) X_{\mathrm{x}_0,k}(\mathrm{x})\ .$$

Thus the one-dimensional vector (given by the raster-scan order on the x-axis) of the restriction of $X_{\mathrm{x}_0,k}$ is an eigenvector of C associated with eigenvalue $4\sin^2\left(\frac{k\pi}{m}\right)$.

Using that I is in bijection with Ω we get that the number of vectors $(X_{\mathrm{x}_0,k})$ is $|\Omega|$. We show that this family of vectors is linearly-independent. Let $\Lambda_{\mathrm{x}_0,k} \in \mathbb{R}$ such that

$$\sum_{\mathrm{x}_0 \in \Omega_0} \sum_{k=1}^{J(\mathrm{x}_0)-1} \Lambda_{\mathrm{x}_0,k} X_{\mathrm{x}_0,k} = 0\ .$$

Since $X_{\mathrm{x}_0,k}$ and $X_{\mathrm{y}_0,k'}$ have different support if and only if $\mathrm{x}_0 \neq \mathrm{y}_0$ we get that for any $\mathrm{x}_0 \in \Omega_0$, $\sum_{k=1}^{J(\mathrm{x}_0)-1} \Lambda_{\mathrm{x}_0,k} X_{\mathrm{x}_0,k} = 0$. This gives that $(\Lambda_{\mathrm{x}_0,k})_{k\in[\![1,J(\mathrm{x}_0)-1]\!]}$ is in the kernel of the matrix $(\sin(\ell k\pi/(J(\mathrm{x}_0)-1)))_{1\leq j,\ell\leq J(\mathrm{x}_0)-1}$. Since the sinus discrete transform is invertible we obtain that for any $\mathrm{x}_0 \in \Omega_0$ and $k \in [\![1, J(\mathrm{x}_0)-1]\!]$, $\Lambda_{\mathrm{x}_0,k} = 0$. Thus the family $X_{\mathrm{x}_0,k}$ is a basis of eigenvectors.

We aim at computing the cardinality of $K_{k,m} = \{X_{\mathrm{x}_0,k}, J(\mathrm{x}_0) = m\}$. By definition, in Proposition 3.2.8, $r_{k,m} = |K_{k,m}|$. First note that $|K_{k',m}| = |K_{k,m}|$. We give the following decomposition $\Omega_0 = \Omega_x \cup \Omega_y \cup \Omega_{x,y}$ with

$$\Omega_x = [\![-t_x,-1]\!] \times [\![0, p-1-t_y]\!],\ \Omega_x = [\![0, p-1-t_x]\!] \times [\![-t_y,-1]\!],\ \Omega_{x,y} = [\![-t_x,-1]\!] \times [\![-t_y,-1]\!]\ .$$

Note that for all $\mathrm{x}_0 \in \Omega_0$ we have that $\mathrm{x}_0 + (q+1)\mathrm{t} \notin \Omega$, with $q = \lceil \frac{p}{|t_x|\vee|t_y|} \rceil$. Thus $J(\Omega_0) \subset [\![2, q+1]\!]$. Let $m \in [\![2, q-1]\!]$. The cardinality of $K_{k,m}$ is the cardinality of $J^{-1}(m)$. Let $\mathrm{x}_0 \in \Omega_x$ we have

$$\mathrm{x}_0 = (i_0, j_0) \in K_{k,m} \ \Leftrightarrow\ \begin{cases} i_0 + mt_x \geq p \\ \text{or} \\ j_0 + mt_y \geq p \end{cases} \text{and} \begin{cases} i_0 + (m-1)t_x \leq p-1 \\ \text{and} \\ j_0 + (m-1)t_y \leq p-1 \end{cases}\ .$$

Since $\mathrm{x}_0 \in \Omega_x$ we have $i_0 + mtx \leq p-1$, hence

$$\mathrm{x}_0 = (i_0, j_0) \in K_{k,m} \ \Leftrightarrow\ j_0 + mt_y \geq p \text{ and } j_0 + (m-1)t_y \leq p-1\ .$$

Thus $|\Omega_x \cap J^{-1}(m)| = t_x t_y$. Similarly we get that $|\Omega_y \cap J^{-1}(m)| = t_x t_y$ and $\Omega_{x,y} \cap J^{-1}(m) = \emptyset$. Thus, $|K_{k,m}| = 2t_x t_y$.

We have computed $|K_{k,m}|$ for every $m \in [\![2, q-1]\!]$. In order to complete our study it only remains to compute $|K_{k,q+1}|$, since $|K_{k,q}|$ can be deduced from the summability condition and from $|K_{k,m}| = |K_{k',m}|$. We only compute $|K_{k,q+1}|$. We remark that $\Omega_x \cap J^{-1}(q+1) = \Omega_y \cap J^{-1}(q+1) = \emptyset$. Let $\mathrm{x}_0 \in \Omega_{x,y}$ then $\mathrm{x}_0 = -\mathrm{t} + (x, y)$ with $x \in [\![0, t_x - 1]\!]$ and $y \in [\![0, t_y - 1]\!]$. We obtain the following equivalence

$$\mathrm{x}_0 \in J^{-1}(q+1) \;\Leftrightarrow\; \begin{cases} -t_x + x + (q+1)t_x \geq p \\ \text{or} \\ -t_y + y + (q+1)t_y \geq p \end{cases} \text{and} \begin{cases} -t_x + x + qt_x \leq p-1 \\ \text{and} \\ -t_y + y + qt_y \leq p-1 \end{cases} .$$

Since $qt_x \geq p$ or $qt_y \geq p$ we obtain that the first condition is always satisfied. Thus we get

$$\mathrm{x}_0 \in J^{-1}(q+1) \;\Leftrightarrow\; x \leq p - 1 - (q-1)t_x \text{ and } y \leq p - 1 - (q-1)t_y \; .$$

Using that $p - 1 - (q-1)t_x = \left(\lceil \frac{p}{t_x} \rceil - q\right) t_x + t_x - 1 - p_x$, we conclude the proof. □

Update rules

We derive the proof of Proposition 3.2.11.

Proof. Computing the minimum of $q(B, M|E)$ for fixed $B \in \mathbb{R}^4$, respectively fixed $M \in \mathbb{R}^{2|E|}$, gives the update rule for M, respectively for B. We obtain that

$$\begin{aligned} q(B, M|E) &= \sum_{\mathrm{e} \in E} m_{\mathrm{e}}^2 \|b_1\|^2 + \sum_{\mathrm{e} \in Eb} n_{\mathrm{e}}^2 \|b_2\|^2 + 2 \sum_{e \in E} m_{\mathrm{e}} n_{\mathrm{e}} \langle b_1, b_2 \rangle \\ &\quad - 2 \sum_{e \in E} m_{\mathrm{e}} \langle b_1, \mathrm{e} \rangle - 2 \sum_{e \in E} n_{\mathrm{e}} \langle b_2, \mathrm{e} \rangle + r(B, M) \\ &= B^T \left(\Lambda_M \otimes \mathrm{Id}_2\right) B - 2\langle B, E_M \rangle + \alpha(M) \\ &= \| \left(\Lambda_M \otimes \mathrm{Id}_2\right)^{\frac{1}{2}} B - \left(\Lambda_M \otimes \mathrm{Id}_2\right)^{\frac{-1}{2}} E_M \|^2 + \alpha(M) \\ &= \| \left(\Lambda_M \otimes \mathrm{Id}_2\right)^{\frac{1}{2}} \left(B - \left(\Lambda_M \otimes \mathrm{Id}_2\right)^{-1} E_M \right) \|^2 + \alpha(M) \; , \end{aligned}$$

where $\alpha(M)$ depends only on M. Similar derivation goes for B and we obtain the proposed update rules. □

Chapter 4

Stochastic Optimization with Unadjusted Langevin Algorithm

4.1 Convergence of diffusions and their discretizations

4.1.1 Abstract

The study of the convergence of Markov processes in general state space is a very attractive and active field of research motivated by applications in mathematics, physics and statistics [JM17]. Among the many works on the subject, we can mention the pioneering results from [NT78; NT82; NT83] using the renewal approach. Then, the work of [Pop77; MT92] paved the way for the use of *Foster-Lyapunov drift conditions* [Fos53; Bre99] which, in combination of an appropriate *minorization condition*, implies (f, r)-ergodicity on general state space, drawing links with control theory, see [TT94; Dou+04; JR02]. This approach was successively applied to the study of Markov chains in numerous papers [Cha93; CT91; RP94] and was later extended and used in the case of continuous-time Markov processes in [Kha11; MT93b; MT93c; DMT95; GM06; FR05; DFG09; Ver97; DKM17]. However, most of these results establish convergence in total variation or in V-norm and are non-quantitative. Explicit convergence bounds in the same metrics for Markov chains have been established in [Ros95; DMR04; Ros02; RT99; JH04; LT96; MT94; For02], driven by the need for stopping rules for Markov Chain Monte Carlo (MCMC) simulations. To the authors' knowledge, the techniques developed in these papers have not been adapted to continuous-time Markov processes, except in [RR96]. One of the main reason is that deriving quantitative minorization conditions for continuous-time processes seems to be even more difficult than for their discrete counterparts [EGZ18]. Indeed, in most cases, the constants which appear in minorization conditions are either really pessimistic or hard to quantify accurately [JH01; QH18].

Since the last decade, in order to avoid the use of minorization conditions, other metrics than the total variation distance, or V-norm, have been considered. In particular, Wasserstein metrics have shown to be very interesting in the study of Markov processes and to derive quantitative bounds of convergence as well as in the study of perturbation bounds for Markov chains [RS18; PS14]. For example, [Oll09; JO10; Pau16] introduced the notion of Ricci curvature of Markov chains and its use to derive precise bounds on variance and concentration inequalities for additive functionals. Following [HM11], [HMS11] generalizes the Harris' theorem for V-norms to handle more general Wasserstein type metrics. In the same spirit, [But14] establishes conditions which imply subgeometric convergence in Wasserstein distance of Markov processes. In addition, the use of Wasserstein distance has been successively applied to

the study of diffusion processes and MCMC algorithms. In particular, [Ebe16; EGZ18] establish explicit convergence rates for diffusions and McKean Vlasov processes. Regarding analysis of MCMC methods, [HSV14] establishes geometric convergence of the pre-conditioned Crank-Nicolson algorithm. Besides, [DM19; DK19; Cha+18; Bak+19] study the computational complexity in Wasserstein distance to sample from a log-concave density on $\mathbb{R}^d$ using appropriate discretizations of the overdamped Langevin diffusion. One key idea introduced in [HMS11] and [Ebe16] is the construction of an appropriate metric designed specifically for the Markov process under consideration. The approach of [Ebe16] has then been generalized in [Che+18; MMS18]. While this approach leads to quantitative results in the case of diffusions or their discretization, we can still wonder if appropriate minorization conditions can be found to derive similar bounds using classical results cited above.

In this section, we show that for a class of functional auto-regressive models, sharp minorization conditions hold using an iterated Markov coupling kernel. As a result new quantitative convergence bounds can be obtained combining this conclusion and drift inequalities for well-suited Lyapunov functionals. We apply them to the study of the Euler-Maruyama discretization of diffusions with identity covariance matrix under various curvature assumptions on the drift. The rates of convergence we derive in weighted total variation metric in this case improve the one recently established in [EM19]. Indeed, while recent papers have established precise bounds between the n-th iterate of the Euler-Maruyama scheme and π in different metrics (*e.g.* total variation or Wasserstein distances), the convergence of the associated Markov kernel is in general needed to obtain quantitative bounds on the mean square error or concentration inequalities for additive functionals, see [DM19; JO10].

In the second part of this section, we show how the results we derive for functional auto-regressive models can be used to establish explicit convergence rates for diffusion processes. First, we show that, under proper conditions on a sequence of discretizations, the distance in some metric between the distributions of the diffusion at time t with different starting points can be upper bounded by the limit of the distance between the corresponding discretizations, when the discretization stepsize decreases towards zero. Similarly, in [KM17] general Markov processes are approximated by hidden Markov models under a continuous Foster-Lyapunov assumption. Second, we design appropriate discretizations satisfying the necessary conditions we obtain and which belong to the class of functional autoregressive models we study. Therefore, under the same curvature conditions as in the discrete case, we get quantitative convergence rates for diffusions by taking the limit in the bounds we derived for the Euler-Maruyama discretizations. Finally, the rates we obtain scale similarly with respect to the parameters of the problem under consideration to the ones given in [Ebe16; EGZ18] for the Kantorovitch-Rubinstein distance, and improve them in the case of the total variation norm. Note that in the diffusion case, earlier results were derived in [CW97; CW95; Wan94].

The rest of Section 4.1 is organized as follows. For reader's convenience and to motivate our results, we begin in Section 4.1.2, with one of their applications to the specific case of a diffusion over $\mathbb{R}^d$ with identity covariance matrix and its Euler-Maruyama discretization, in the case where the drift function is strongly convex at infinity. In Section 4.1.3, we present our main convergence results regarding a class of functional autoregressive models. We then specialize them to the Euler-Maruyama discretization of diffusions under various assumptions on the drift function in Section 4.1.4. Section 4.1.5 deals with the convergence of diffusion processes with identity covariance matrix. More precisely, in Section 4.1.5, we derive sufficient conditions for the convergence of such processes based on a sequence of well-suited discretizations. In Section 4.1.5, we apply our results to the continuous counterparts of the situations considered in Section 4.1.4. Finally, new quantitative convergence bounds for discrete Markov chains on general state spaces are given in Section 4.1.6. For ease of presentation we gather the proofs and generalizations of our results in Section 4.1.7.

4.1.2 Motivation and illustrative example

Non-contractive setting

In this section, we motivate our work with applications of our main results to one specific example. Let $b : \mathbb{R}^d \to \mathbb{R}^d$ be a drift function, $(\mathbf{B}_t)_{t\geq 0}$ be a d-dimensional Brownian motion and assume that the SDE

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t \ , \tag{4.1}$$

admits a unique strong solution $(\mathbf{X}_t)_{t\geq 0}$ on $\mathbb{R}_+$ for any starting point $\mathbf{X}_0 = x \in \mathbb{R}^d$. We denote by $(\mathrm{P}_t)_{t\geq 0}$ its associated Markov semigroup. We consider the Euler-Maruyama discretization of this SDE, *i.e.* the homogeneous Markov chain $(X_k)_{k\in\mathbb{N}}$, starting from $X_0 = x \in \mathbb{R}^d$ and defined by the following recursion: for any $k \in \mathbb{N}$

$$X_{k+1} = X_k + \gamma b(X_k) + \sqrt{\gamma} Z_{k+1} \ , \tag{4.2}$$

where $\gamma > 0$ is a stepsize and $(Z_k)_{k\in\mathbb{N}^\star}$ is a sequence of i.i.d. d-dimensional Gaussian random variables with zero mean and identity covariance matrix. We denote by R_γ its associated Markov kernel.

The first consequence of the results established in this section is the explicit convergence of the Markov chain defined by (4.2) in a distance which is a mix of the total variation distance and the Wasserstein distance of order 1, under the assumption that b is Lipschitz and strongly convex at infinity.

Theorem 4.1.1. *Assume that there exist* $\mathtt{m} \in \mathbb{R}$, $\mathtt{m}^+ > 0$ *and* $\mathtt{L}, R \geq 0$ *such that for any* $x, y \in \mathbb{R}^d$

$$\|b(x) - b(y)\| \leq \mathtt{L}\|x - y\| \ , \qquad \langle b(x) - b(y), x - y\rangle \leq -\mathtt{m}\|x - y\|^2 \ , \tag{4.3}$$

and if $\|x - y\| \geq R$,

$$\langle b(x) - b(y), x - y\rangle \leq -\mathtt{m}^+\|x - y\|^2 \ . \tag{4.4}$$

Then there exist $\bar\gamma > 0$, $D_{\bar\gamma,1}, D_{\bar\gamma,2}, E_{\bar\gamma} \geq 0$ *and* $\lambda_{\bar\gamma}, \rho_{\bar\gamma} \in [0, 1)$ *with* $\lambda_{\bar\gamma} \leq \rho_{\bar\gamma}$, *which can be explicitly computed, such that for any* $\gamma \in (0, \bar\gamma]$, $x, y \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \lambda_{\bar\gamma}^{k\gamma/4}[D_{\bar\gamma,1}\mathbf{c}(x, y) + D_{\bar\gamma,2}\mathbb{1}_{\Delta^c}(x, y)] + E_{\bar\gamma}\rho_{\bar\gamma}^{k\gamma/4}\mathbb{1}_{\Delta^c}(x, y) \ , \tag{4.5}$$

where $\mathbf{c}(x, y) = \mathbb{1}_{\Delta^c}(x, y)(1 + \|x - y\|/R)$, $\Delta = \{(x, x) \ : \ x \in \mathbb{R}^d\}$ *and* R_γ *is the Markov kernel associated with* (4.2).

Proof. The result is a direct consequence of Theorem 4.1.13 and the corresponding discussion in Section 4.1.4. □

This result is derived as a specific case of a more general theorem for a class of functional autoregressive models, see Theorem 4.1.8 and Section 4.1.3. Its proof relies on the use of an extended Foster-Lyapunov drift assumption as well as a minorization condition on the Markov chain (4.2). As an important consequence, curvature assumptions on the drift (such as strong convexity at infinity) can be omitted if we instead assume some Foster-Lyapunov condition, similarly to [Ebe16, Theorem 6.1] and [EGZ18, Theorem 2.1].

The result derived in Theorem 4.1.1 has several important applications which we gather in the following corollary.

Corollary 4.1.2. *Assume that there exist* $\mathtt{m} \in \mathbb{R}$, $\mathtt{m}^+ > 0$ *and* $\mathtt{L}, R \geq 0$ *such that* (4.3) *and* (4.4) *are satisfied. Then, there exist* $\bar\gamma > 0$, $E_{\bar\gamma,1}, E_{\bar\gamma,2} \geq 0$ *such that for any* $\gamma \in (0, \bar\gamma]$, $x, y \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$ *we have*

$$\|\delta_x \mathrm{R}_\gamma^k - \delta_y \mathrm{R}_\gamma^k\|_{\mathrm{TV}} \leq \mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq E_{\bar\gamma,1}\rho_{\bar\gamma}^{k\gamma/4}\mathbf{c}(x, y) \ , \tag{4.6}$$

$$\mathbf{W}_1(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq E_{\bar{\gamma},2} \rho_{\bar{\gamma}}^{k\gamma/4} \|x - y\| \ , \tag{4.7}$$

where $\mathbf{c}(x,y) = \mathbb{1}_{\Delta^c}(x,y)(1 + \|x-y\|/R)$, $\Delta = \{(x,x) \ : \ x \in \mathbb{R}^d\}$ *and* $\rho_{\bar{\gamma}}$ *is given in* (4.5). *In addition, for any* $p \in \mathbb{N}$ *and* $\alpha \in (p, +\infty)$ *there exists* $E_{\bar{\gamma},\alpha} \geq 0$ *such that for any* $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$ *we have*

$$\mathbf{W}_p(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq E_{\bar{\gamma},\alpha} \rho_{\bar{\gamma}}^{k\gamma/(4\alpha)} (\|x-y\| + \|x-y\|^{1/\alpha}) \ . \tag{4.8}$$

The constants $\bar{\gamma}$, $\{E_{\bar{\gamma},i} \ : \ i = 1,2,3\}$ *and* $E_{\bar{\gamma},\alpha}$ *can be explicitly computed.*

Proof. The estimate (4.6) is a direct consequences of Theorem 4.1.1. The two inequalities (4.7) and (4.8) follow from Corollary 4.1.14. □

Note that the same rate $\rho_{\bar{\gamma}}$ appears in the inequalities (4.5), (4.6), (4.7) and (4.8). Section 4.1.5 is devoted to the extension of our discrete-time results to their continuous-time counterparts. Note also that Theorem 4.1.3 and its consequences still hold if we only assume a local Lipschitz assumption, see the condition **B**5.

Theorem 4.1.3. *Assume that there exist* $\mathtt{m} \in \mathbb{R}$, $\mathtt{m}^+ > 0$ *and* $\mathtt{L}, R \geq 0$ *such that* (4.3) *and* (4.4) *are satisfied. Then there exist* $D_1, D_2, E \geq 0$ *and* $\lambda, \rho \in [0,1)$ *with* $\lambda \leq \rho$ *such that for any* $x, y \in \mathbb{R}^d$ *and* $t \geq 0$

$$\|\delta_x \mathrm{P}_t - \delta_y \mathrm{P}_t\|_{\mathrm{TV}} \leq \mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq \lambda^{t/4}[D_1 \mathbf{c}(x,y) + D_2 \mathbb{1}_{\Delta^c}(x,y)] + E\rho^{t/4} \mathbb{1}_{\Delta^c}(x,y) \ , \tag{4.9}$$

where $\mathbf{c}(x,y) = \mathbb{1}_{\Delta^c}(x,y)(1 + \|x-y\|/R)$, $\Delta = \{(x,x) \ : \ x \in \mathbb{R}^d\}$, $(\mathrm{P}_t)_{t \geq 0}$ *is the Markov semigroup associated with* (4.1) *and*

$$D_1 = \lim_{\bar{\gamma} \to 0} D_{\bar{\gamma},1} \ , \quad D_2 = \lim_{\bar{\gamma} \to 0} D_{\bar{\gamma},2} \ , \quad E = \lim_{\bar{\gamma} \to 0} E_{\bar{\gamma}} \ , \quad \lambda = \lim_{\bar{\gamma} \to 0} \lambda_{\bar{\gamma}} \ , \quad \rho = \lim_{\bar{\gamma} \to 0} \rho_{\bar{\gamma}} \ ,$$

and $D_{\bar{\gamma},1}, D_{\bar{\gamma},2}, E_{\bar{\gamma}}, \lambda_{\bar{\gamma}}, \rho_{\bar{\gamma}}$ *are given in Theorem 4.1.1.*

Proof. This result follows from Theorem 4.1.21. □

Note that the constants D_1, D_2, E, λ and ρ have explicit expressions, see the corresponding discussion in Section 4.1.5 after Theorem 4.1.21. In addition, the rate ρ and λ in (4.9) are independent of the dimension d. This is a significant improvement compared to the convergence results in total variation derived in [EGZ18, Theorem 2.1] which imply a convergence rate which scales exponentially in the dimension, under the setting we consider. Similarly, we derive a continuous counterpart of Corollary 4.1.2 in the continuous time setting, see Corollary 4.1.22.

As stated before, the convergence rates $\rho_{\bar{\gamma}}$, ρ, $\lambda_{\bar{\gamma}}$, λ, given in Theorem 4.1.1 and Theorem 4.1.3 can be explicitly computed. More precisely, we obtain the following expressions (up to logarithmic terms) with respect to the parameters $\mathtt{m}$, $\mathtt{L}$ and R in the case $-\mathtt{m}R^2 \gg 1$, see Section 4.1.4, Theorem 4.1.13, Equations (4.46) and (4.47):

$$\log(\log^{-1}(\rho_{\bar{\gamma}}^{-1})) \simeq -(\mathtt{m}R^2/4) \sup_{\gamma \in (0,\bar{\gamma}]} \left\{ \left(1 - \frac{\gamma \mathtt{L}^2}{2\mathtt{m}}\right) \left(1 - \exp\left[\frac{R^2(2\mathtt{m} - \gamma\mathtt{L}^2)}{1 - 2\mathtt{m}\gamma + \gamma^2 \mathtt{L}^2}\right]\right)^{-1} \right\} \tag{4.10}$$

$$\log(\log^{-1}(\rho^{-1})) \simeq -(\mathtt{m}R^2/4) \times (1 - \mathrm{e}^{2\mathtt{m}R^2})^{-1} \ , \tag{4.11}$$

$$\log(\lambda_{\bar{\gamma}}) = -\mathtt{m}^+/2 + \bar{\gamma}\mathtt{L}^2/4 \ , \qquad \log(\lambda) = -\mathtt{m}^+/2 \ .$$

where $\simeq$ denotes equality up to logarithmic factors.

It is sensible to obtain two different convergence rates $\lambda_{\bar\gamma}, \rho_{\bar\gamma}$ (resp. λ, ρ) in Theorem 4.1.1 (resp. in Theorem 4.1.3), one characterizing the forgetting of the initial distance between the two starting points $x, y \in \mathbb{R}^d$, corresponding to a burn-in period, and the other one characterizing the effective convergence. In addition, note that $\lambda_{\bar\gamma} \ll \rho_{\bar\gamma}$ and $\lambda \ll \rho$ if $-\mathtt{m}R^2 \gg 1$.

We now compare these results and the rates obtained in (4.10)-(4.11) with recent works studying the convergence of the Markov chain defined by (4.2) and/or the corresponding diffusion process (4.1) in the same framework, *i.e.* under the conditions (4.3) and (4.4). Note that the same conclusions hold under more general Foster-Lyapunov drift conditions but it would make the comparison more involved. Note also we are still able to derive convergence results under weaker curvature assumptions on the drift. The discussion is postponed to Section 4.1.4 for the discrete setting and Section 4.1.5 for the continuous setting.

First, a major difference between our work and the ones mentioned below is that we use a completely different technique to establish our results. Indeed, all of them follow the approach initiated in [Ebe11], designing a suitable coupling and distance function of the form $\mathbf{c}(x, y) = f(\|x - y\|)$, for any $x, y \in \mathbb{R}^d$, with $f : \mathbb{R}_+ \to \mathbb{R}_+$, to obtain a geometric contraction in $\mathbf{W}_\mathbf{c}$ for either the Markov chain (4.2) or the diffusion (4.1) under the conditions (4.3)-(4.4). In this section, we follow a different path and derive convergence estimates using minorization and Foster-Lyapunov drift conditions, adapting the technique used in [Dou+18] and the references therein. It has been thought for a long time that such an approach only gives very pessimistic convergence bounds [EGZ18]. We now compare more specifically our results with the ones obtained following the work of [Ebe11] and show that in fact our technique inspired by classical methods to establish geometric convergence of Markov chains gives very sharp estimates, improving and simplifying the results obtained in the existing literature. This discussion and its conclusion are summarized in Table 4.1. In the rest of this section, $C \geq 0$ stands for a positive constant which may be different at each occurrence.

First we compare our work with the results of [EM19] which extend to the discrete setting the estimates of [Ebe16]. The authors use the following cost function defined for any $x, y \in \mathbb{R}^d$ by

$$\mathbf{c}_a(x, y) = a\mathbb{1}_{\Delta^c}(x, y) + f_a(\|x - y\|) \ ,$$

where $a \geq 0$ and f_a is given in [EM19, Equation (2.53)]. Note that the cost $\mathbf{c}_a$ is close to the one introduced in Theorem 4.1.1. Then, [EM19, Theorem 2.10] states that if $a \in [2\gamma^{1/2}, \mathbf{\Phi}_E(R)]$ where $\mathbf{\Phi}_E$ is given in [EM19, Theorem 2.10], then there exist $\bar\gamma_a > 0$ and $\rho_a \in [0, 1)$ such that for any $\gamma \in (0, \bar\gamma_a]$, $x, y \in \mathbb{R}^d$ and $k \in \mathbb{N}$,

$$\mathbf{W}_{\mathbf{c}_a}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \rho_a^{k\gamma} \mathbf{c}_a(x, y) \ . \tag{4.12}$$

Compared to our results Theorem 4.1.1, (4.12) only gives one convergence rate ρ_a and does not dissociate the forgetting of the initial distance between the starting points $x, y \in \mathbb{R}^d$ from the long-term behavior. In addition, a may depend on γ, since it is required that $a \in [2\gamma^{1/2}, \mathbf{\Phi}(R)]$ and $\bar\gamma_a < \bar\gamma$ where $\bar\gamma$ is given by Theorem 4.1.1. Omitting the dependency of a and ρ_a with respect to γ for the sake of simplicity, and applying [EM19, Theorem 2.10] yield

$$\log(\log^{-1}(\rho_a^{-1})) \simeq -\mathtt{m}R^2/c_1 \ , \text{ with } c_1 = 16^{-1} \int_{1/4}^{3/8} (1 - \mathrm{e}^{u-1/2}) \boldsymbol{\varphi}(u) \mathrm{d}u \leq 0.00051 \ , \tag{4.13}$$

where for any $t \in \mathbb{R}$, $\boldsymbol{\varphi}(t) = (2\pi)^{-1/2} \exp(-t^2/2)$. It is worth noticing that in the case we are interested in, $-\mathtt{m}R^2 \gg 1$, we obtain that our rate given by (4.10) satisfies $\rho_{\bar\gamma} \ll \rho_a$ (also omitting dependency of $\rho_{\bar\gamma}$ with respect to γ).

Let $\mathbf{c}_b$ be defined for any $x, y \in \mathbb{R}^d$ by $\mathbf{c}_b(x, y) = f_b(\|x - y\|)$ with f_b given in [EM19, Equation (2.68)]. Then, [EM19, Theorem 2.12] implies that there exist $\bar\gamma_b > 0$ and $\rho_b \in [0, 1)$ such that for any

$\gamma \in (0, \bar{\gamma}_b]$, $x, y \in \mathbb{R}^d$ and $k \in \mathbb{N}$,

$$\mathbf{W}_{\mathbf{c}_b}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \rho_b^{k\gamma} \mathbf{c}_b(x, y) \ . \tag{4.14}$$

Note that (4.12) implies convergence bounds both with respect to $\mathbf{W}_1$ and the total variation distance whereas (4.14) implies convergence bounds with respect to $\mathbf{W}_1$ only. Once again, omitting the dependency with respect to γ, we obtain that the rate satisfies

$$\log(\log^{-1}(\rho_b^{-1})) \simeq -49\mathtt{m}R^2/(6c_2) \ ,$$

with

$$c_2 = 4 \min \left(\int_0^{1/2} u^2 (1 - \mathrm{e}^{u-1/2}) \boldsymbol{\varphi}(u) \mathrm{d}u, (1 - \mathrm{e}^{-1}) \int_0^{1/2} u^3 \boldsymbol{\varphi}(u) \mathrm{d}u \right) \leq 0.0072 \ , \tag{4.15}$$

and we obtain that our rate given by (4.10) satisfies $\rho_{\bar{\gamma}} \ll \rho_a$ when $-\mathtt{m}R^2 \gg 1$.

We now compare our results with the ones derived in [MMS18]. For fair comparison, since [MMS18] does not assume a one-sided Lipschitz condition but only a global Lipschitz condition we set $\mathtt{m} = -\mathtt{L}$ in the next paragraph. In this section, we extend the techniques of [EM19; Ebe16] to deal with $\mathbf{W}_2$. It is shown in [MMS18, Theorem 2.1] that there exist $\bar{\gamma}_c > 0$ and $\rho_c \in [0, 1)$ such that for any $\gamma \in (0, \bar{\gamma}_c]$, $x, y \in \mathbb{R}^d$ and $k \in \mathbb{N}$,

$$\mathbf{W}_{\mathbf{c}_c}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \rho_c^{k\gamma} \mathbf{c}_c(x, y) \ ,$$

with $\mathbf{c}_c$ given for any $x, y \in \mathbb{R}^d$ by $\mathbf{c}_c(x, y) = f_c(\|x - y\|)$ and f_c given in [EM19, Equation (2.11)]. Note that this result implies convergence bounds with respect to $\mathbf{W}_1$ and $\mathbf{W}_2$. In particular, we have for any $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ and $k \in \mathbb{N}$,

$$\mathbf{W}_2(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq C \rho_c^{k\gamma/2} \mathbf{c}_c^{1/2}(x, y) \leq C \rho_c^{k\gamma/2} (\|x - y\| + \|x - y\|^{1/2}) \ .$$

In addition, it holds that

$$\log(\log^{-1}(\rho_c^{-1})) \simeq \mathtt{L}R^2/(6c_2) \ \text{ where } c_2 \text{ is defined by (4.15)} \ , \tag{4.16}$$

and therefore our rate also satisfies $\rho_{\bar{\gamma}} \ll \rho_c$ when $\mathtt{L}R^2 \gg 1$.

The results of [EM19; MMS18] both extend, and generalize, in the discrete-time setting the techniques used in [Ebe16]. In the latter, contraction results for the semigroup $(\mathrm{P}_t)_{t \geq 0}$ are obtained with respect to $\mathbf{W}_{\mathbf{c}_e}$, where for any $x, y \in \mathbb{R}^d$, $\mathbf{c}_e(x, y) = f_e(\|x - y\|)$ and f_e is defined by [Ebe16, Equation (2.6)]. In particular, in [Ebe16, Corollary 2.3], it is shown that there exists $\rho_e \in [0, 1)$ such that for any $x, y \in \mathbb{R}^d$ and $t \geq 0$

$$\mathbf{W}_{\mathbf{c}_e}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq \rho_e^t \mathbf{c}_e(x, y) \ .$$

Note that this result implies convergence bounds in $\mathbf{W}_1$, see [Ebe16, Corollary 2.3]. The rate is given [Ebe16, Lemma 2.9] and, in the case $-\mathtt{m}R^2 \gg 1$, we have

$$\log(\log^{-1}(\rho_e^{-1})) \simeq -\mathtt{m}R^2/4 \ , \tag{4.17}$$

which is better than our rate in the continuous-time case[1]. However, note that we derive our results in $\mathbf{W}_1$ from our estimates with respect to $\mathbf{W}_{\mathbf{c}}$ with $\mathbf{c}$ given in Theorem 4.1.1, which controls both $\mathbf{W}_1$

[1]Note that in [Ebe16, Lemma 2.9, Equation (2.18)] the stated result implies that $\log(\log^{-1}(\rho^{-1})) \simeq \mathtt{L}R^2/8$ if $\kappa(r) \geq -\mathtt{L}r$ for any $r \geq 0$, where κ is defined in [Ebe16, p.5]. However, note that if b is L-Lipschitz then $\kappa(r) \geq -2\mathtt{L}$ and (4.17) follows.

and the total variation norm. Also, the discrepancy between our rate and the one of (4.17) is controlled by $(\mathrm{e}^{-2\mathtt{m}R^2}-1)^{-1}$ which is small when $-\mathtt{m}R^2$ is large.

Finally we compare our continuous-time results with the ones of [LW16b]. It is shown in [LW16b, Theorem 1.3] that for any $p>1$ there exist $\rho_f \in [0,1)$ and $C \geq 0$ such that for any $x,y \in \mathbb{R}^d$ and $t \geq 0$

$$\mathbf{W}_p(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq C\rho_f^t \left\{ \|x-y\| + \|x-y\|^{1/p} \right\} , \tag{4.18}$$

and the rate is given in [LW16b, Theorem 1.3] by

$$\log(\log^{-1}(\rho_f^{-1})) = (-\mathtt{m}+\mathtt{m}^+)R^2/4 .$$

The additional term $\mathtt{m}^+R^2/4$ does not appear in our rates[2]. As a consequence our rate is better as soon as

$$\mathtt{m}^+ \geq -\mathtt{m}/(\mathrm{e}^{-2\mathtt{m}R^2}-1) .$$

Table 4.1 gives a summary of the comparisons we made above.

Reference	Wasserstein distance	distance bound	(D)	(C)	(NR)
[EM19]	$\|\cdot\|_{\mathrm{TV}}$	$\mathbb{1}_{\Delta^c}(x,y) + \|x-y\|$	✓		7840
	$\mathbf{W}_1$	$\|x-y\|$	✓		4536
[MMS18]	$\mathbf{W}_2$	$\|x-y\| + \|x-y\|^{1/2}$	✓		332
[Ebe16]	$\mathbf{W}_1$	$\|x-y\|$		✓	1
[LW16b]	$\mathbf{W}_p$	$\|x-y\| + \|x-y\|^{1/p}$		✓	$1-\mathtt{m}^+/\mathtt{m}$
Ours	$\|\cdot\|_{\mathrm{TV}}$	$\mathbb{1}_{\Delta^c}(x,y) + \|x-y\|$	✓	✓	$(1-\mathrm{e}^{2\mathtt{m}R^2})^{-1}$
	$\mathbf{W}_1$	$\|x-y\|$	✓	✓	idem
	$\mathbf{W}_p$	$\|x-y\| + \|x-y\|^{1/\alpha}$	✓	✓	idem

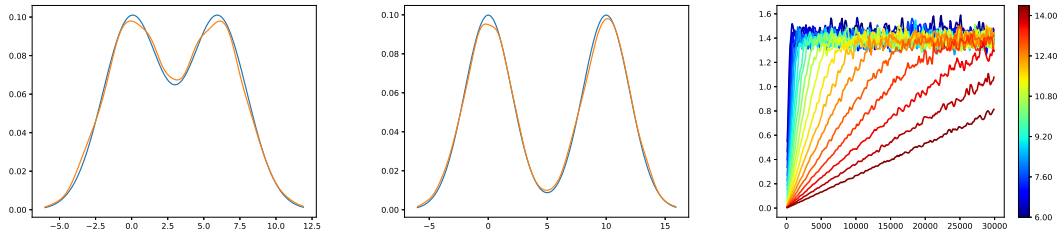
Table 4.1: Every line of the table reads as follows. Suppose "Wasserstein distance" reads $\mathbf{W}_{\mathbf{c}_1}$ and "distance bound" reads $\mathbf{c}_2(x,y)$ then: if (D) is checked, there exist $C \geq 0$ and $\rho \in [0,1)$ such that for any $x,y \in \mathbb{R}^d$ and $k \in \mathbb{N}$, $\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq C\rho^{k\gamma}\mathbf{c}_2(x,y)$ for γ small enough. If (C) is checked, there exist $C \geq 0$ and $\rho \in [0,1)$ such that for any $x,y \in \mathbb{R}^d$ and $t \geq 0$, $\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq C\rho^t \mathbf{c}_2(x,y)$. In addition, if the normalized rate "(NR)" reads β we have $-4\log(\log^{-1}(\rho^{-1}))/(\mathtt{m}R^2) \simeq \beta$ (with $\mathtt{m}$ replaced by $-\mathtt{L}$ in the case of [MMS18]). Note that for the sake of simplicity we omit the dependency with respect to $\bar{\gamma}$ in the present analysis. The exact distances used in papers with which we compare our results, are given in [EM19, Equation (2.53)], [MMS18, Equation (2.11)], [Ebe16, Equation (2.6)] and [LW16b, Equation (2.4)]. Note that $p \in \mathbb{N}$ and $\alpha \in (p, +\infty)$.

An illustrative example

We now consider a toy example to justify the setting under study in the previous section. Consider the following Gaussian mixture distribution π whose Radon-Nikodym density with respect to the Lebesgue measure λ is given for any $x \in \mathbb{R}$ by

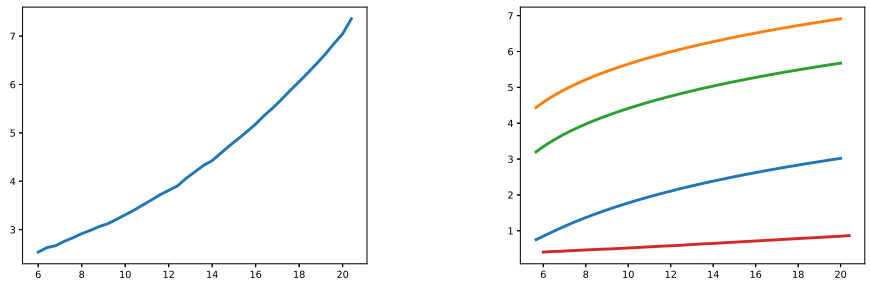
$$(\mathrm{d}\pi/\mathrm{d}\lambda)(x) = (2\sqrt{2\pi\sigma^2})^{-1} \exp[-x^2/(2\sigma^2)] + (2\sqrt{2\pi\sigma^2})^{-1} \exp[-(x-m)^2/(2\sigma^2)] ,$$

[2] Similarly to [Ebe16], in [LW16b, Theorem 1.3] the stated result implies that $\log(\log^{-1}(\rho^{-1})) \simeq \mathtt{L}R^2/2$ if $\kappa(r) \leq \mathtt{L}r$ for any $r \geq 0$ and $\kappa(r) \leq -\mathtt{m}^+ r$ for $r \geq R$, where κ is defined in [LW16b, Equation (1.4)]. However, note that if b is L-Lipschitz and $\mathtt{m}^+$ strongly convex outside of $\bar{\mathrm{B}}(0,R)$ we have $\kappa(r) \leq \mathtt{L}r/2$ for any $r \geq 0$ and $\kappa(r) \leq -\mathtt{m}^+ r/2$ for any $r \geq R$ and (4.18) follows.



(a) (b) (c)

Figure 4.1: In (a) and (b), the blue curve is the theoretical log-partition and in orange the estimated log-partition of $\delta_{x_0} \mathrm{R}_\gamma^n$ at iteration $n = 10000$ with $\gamma = 0.1$. The estimation of the log-partition is performed using Gaussian kernels and 1000 points sampled from 1000000 points using a bootstrap procedure. In (a), $m = 6$ and $\sigma = 2$ and in (b) $m = 10$ and $\sigma = 2$. In (c) we illustrate the behavior of $-\log_{10}(\|\delta_{x_0} \mathrm{R}_\gamma^n - \pi\|_{\mathrm{TV}})$ for $\sigma = 2$ and m between 6 and 14 (color blue to red). Note that the precision saturates since $\pi \neq \pi_\gamma$.



(a) (b)

Figure 4.2: In (a) we present $\log(\log^{-1}(\rho_{\mathrm{exp}}^{-1}))$ as m varies. In (b) we present $\log_{10}(\log(\log^{-1}(\rho^{-1})))$ with $\rho \leftarrow \rho_{\mathrm{exp}}$ (red), ρ given by (4.11) (blue), ρ given by (4.16) (green) and ρ given by (4.13) (orange).

where $\sigma > 0$ and $m \geq 0$. For any $x \in \mathbb{R}$, we have $(\mathrm{d}\pi/\mathrm{d}\lambda)(x) \propto \mathrm{e}^{-U(x-m/2)}$ and for any $\bar{x} \in \mathbb{R}$

$$U(\bar{x}) = \bar{x}^2/(2\sigma^2) - \log\left[\cosh(m\bar{x}/(2\sigma^2))\right] \ ,$$

Note that U' is L-Lipschitz with $\mathtt{L} = \sigma^{-2} \max\{1, (m/(2\sigma))^2 - 1\}$ and that U is convex if and only if $m \leq 2\sigma$. Also, we obtain that $b = -U'$ satisfies (4.3) with $\mathtt{L} = \sigma^{-2} \max\{1, (m/(2\sigma))^2 - 1\}$, $R = 2m$, $\mathtt{m}^+ = 1/(2\sigma^2)$.

We now consider the Markov chain (4.2) with $b = -U'$ and its associated Markov kernel R_γ for $\gamma > 0$. Let $x_0 \in \mathbb{R}$ and we define $\log(\rho_{\mathrm{exp}})$ to be the slope of the function $n \mapsto \log(\|\delta_{x_0} \mathrm{R}_\gamma^n - \pi\|_{\mathrm{TV}})$. Note that this slope is computed only until $\log(\|\delta_{x_0} \mathrm{R}_\gamma^n - \pi\|_{\mathrm{TV}})$ reaches a given precision, since for $\gamma > 0$ small enough there exists a probability measure π_γ such that $\|\delta_{x_0} \mathrm{R}_\gamma^n - \pi_\gamma\|_{\mathrm{TV}} \to 0$ and $\pi_\gamma \neq \pi$. In what follows we compare $\log(\rho_{\mathrm{exp}})$ with our estimates.

Let $\theta = m/(2\sigma)$ and assume that $\theta \geq \sqrt{2}$. Note that in this case $\mathtt{L}R^2 = 16\theta^2(\theta^2 - 1)$. Let ρ be the rate we identify in (4.11). Up to logarithmic terms we have $\log(\log^{-1}(\rho^{-1})) \simeq 4\theta^2(\theta^2 - 1)/(1 -$

$\mathrm{e}^{-32\theta^2(\theta^2-1)}$). In Figure 4.1 and Figure 4.2, we fix $\sigma = 2$ and study the behavior of $\log(\rho_{\exp})$ and $\log(\rho)$ w.r.t. m. In particular, Figure 4.2-(b) illustrates that the rates we obtain are much closer to the ones estimated by our numerical simulations.

4.1.3 Quantitative convergence bounds for a class of functional autoregressive models

Let $\mathsf{X} \in \mathcal{B}(\mathbb{R}^d)$ endowed with the trace of $\mathcal{B}(\mathbb{R}^d)$ on X denoted by $\mathcal{X} = \{\mathsf{A} \cap \mathsf{X} : \mathsf{A} \in \mathcal{B}(\mathbb{R}^d)\}$. In this section we consider the Markov chain $(X_k)_{k\in\mathbb{N}}$ defined by $X_0 \in \mathsf{X}$ and the following recursion: for any $k \in \mathbb{N}$

$$X_{k+1} = \Pi\left(\mathcal{T}_\gamma(X_k) + \sqrt{\gamma}\, Z_{k+1}\right) , \tag{4.19}$$

where $\{\mathcal{T}_\gamma : \gamma \in (0, \bar{\gamma}]\}$ is a family of measurable functions from X to $\mathbb{R}^d$ with $\bar{\gamma} > 0$, $\gamma \in (0, \bar{\gamma}]$ is a stepsize, $(Z_k)_{k\in\mathbb{N}^\star}$ is a sequence of i.i.d. d-dimensional zero mean Gaussian random variables with covariance identity and $\Pi : \mathbb{R}^d \to \mathsf{X}$ is a measurable function. The Markov chain $(X_k)_{k\in\mathbb{N}}$ defined by (4.19) is associated with the Markov kernel R_γ defined on $\mathsf{X} \times \mathcal{B}(\mathbb{R}^d)$ for any $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\mathrm{R}_\gamma(x, \mathsf{A}) = (2\pi\gamma)^{-d/2} \int_{\Pi^{\leftarrow}(\mathsf{A})} \exp\left[-(2\gamma)^{-1}\|y - \mathcal{T}_\gamma(x)\|^2\right] \mathrm{d}y , \tag{4.20}$$

where $\Pi^{inv}(\mathsf{A}) = \{y \in \mathbb{R}^d : \Pi(x) \in \mathsf{A}\}$, Note that for any $x \in \mathsf{X}$, $\mathrm{R}_\gamma(x, \mathsf{X}) = 1$ and therefore, R_γ given in (4.20) is also a Markov kernel over $\mathsf{X} \times \mathcal{X}$.

In this section we state explicit convergence results for R_γ for some Wasserstein distances and discuss the rates we obtain. These results rely on appropriate minorization and Foster-Lyapunov drift conditions. We first derive the minorization condition for the n-th iterate of R_γ. To do so, we consider a Markov coupling kernel K_γ for R_γ for any $\gamma \in (0, \bar{\gamma}]$, *i.e.* for any $x, y \in \mathbb{R}^d$, $\mathrm{K}_\gamma((x, y), \cdot)$ is a transference plan between $\mathrm{R}_\gamma(x, \cdot)$ and $\mathrm{R}_\gamma(y, \cdot)$. Indeed, in that case, by [Dou+18, Theorem 19.1.6], we have for any $x, y \in \mathsf{X}$, $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}^\star$,

$$\|\delta_x \mathrm{R}_\gamma^n - \delta_y \mathrm{R}_\gamma^n\|_{\mathrm{TV}} \le \mathrm{K}_\gamma^n((x, y), \Delta_{\mathsf{X}}^{\mathrm{c}}) , \tag{4.21}$$

where $\Delta_{\mathsf{X}} = \{(x, x) : x \in \mathsf{X}\}$. We consider a projected version of the discrete reflection coupling [BDJ98] which is the discrete counterpart of the coupling introduced in [LR86]. For any $x, y, z \in \mathbb{R}^d$, $\gamma \in (0, \bar{\gamma}]$, let

$$\mathrm{e}(x, y) = \begin{cases} \mathrm{E}(x, y)/\|\mathrm{E}(x, y)\| & \text{if } \mathcal{T}_\gamma(x) \neq \mathcal{T}_\gamma(y) \\ 0 & \text{otherwise} \end{cases} , \qquad \mathrm{E}(x, y) = \mathcal{T}_\gamma(y) - \mathcal{T}_\gamma(x) ,$$

and

$$\mathcal{S}_\gamma(x, y, z) = \mathcal{T}_\gamma(y) + (\mathrm{Id} - 2\mathrm{e}(x, y)\mathrm{e}(x, y)^\top) z ,$$
$$p_\gamma(x, y, z) = 1 \wedge \frac{\boldsymbol{\varphi}_\gamma(\|\mathrm{E}(x, y)\| - \langle \mathrm{e}(x, y), z\rangle)}{\boldsymbol{\varphi}_\gamma(\langle \mathrm{e}(x, y), z\rangle)} ,$$

where $\boldsymbol{\varphi}_\gamma$ is the one dimensional zero mean Gaussian distribution function with variance γ. Let $(U_k)_{k\in\mathbb{N}^\star}$ be a sequence of i.i.d. uniform random variables on $[0, 1]$ independent of $(Z_k)_{k\in\mathbb{N}^\star}$. Define the Markov chain $(X_k, Y_k)_{k\in\mathbb{N}}$ starting from $(X_0, Y_0) \in \mathsf{X}^2$ by the recursion: for any $k \in \mathbb{N}$,

$$\tilde{X}_{k+1} = \mathcal{T}_\gamma(X_k) + \sqrt{\gamma} Z_{k+1} ,$$
$$\tilde{Y}_{k+1} = \begin{cases} \tilde{X}_{k+1} & \text{if } \mathcal{T}_\gamma(X_k) = \mathcal{T}_\gamma(Y_k) , \\ W_{k+1}\tilde{X}_{k+1} + (1 - W_{k+1})\mathcal{S}_\gamma(X_k, Y_k, \sqrt{\gamma} Z_{k+1}) & \text{otherwise} , \end{cases}$$

where $W_{k+1} = \mathbb{1}_{(-\infty,0]}(U_{k+1} - p(X_k, Y_k, \sqrt{\gamma} Z_{k+1}))$ and finally set

$$(X_{k+1}, Y_{k+1}) = (\Pi(\tilde{X}_{k+1}), \Pi(\tilde{Y}_{k+1})) \ . \tag{4.22}$$

The Markov chain $(X_k, Y_k)_{k\in\mathbb{N}}$ is associated with the Markov kernel K_γ on $\mathsf{X}^2 \times \mathcal{X}^{\otimes 2}$ given for all $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{X}^{\otimes 2}$ by

$$\begin{aligned}
\mathrm{K}_\gamma((x,y),\mathsf{A}) &= \frac{\mathbb{1}_{\Delta_{\mathbb{R}^d}}(\mathcal{T}_\gamma(x), \mathcal{T}_\gamma(y))}{(2\pi\gamma)^{d/2}} \int_{\mathbb{R}^d} \mathbb{1}_{\Pi_\mathsf{A}}(\tilde{x}, \tilde{x}) \mathrm{e}^{-\frac{\|\tilde{x} - \mathcal{T}_\gamma(x)\|^2}{2\gamma}} \mathrm{d}\tilde{x} \\
&+ \frac{\mathbb{1}_{\Delta^{\mathrm{c}}_{\mathbb{R}^d}}(\mathcal{T}_\gamma(x), \mathcal{T}_\gamma(y))}{(2\pi\gamma)^{d/2}} \left[\int_{\mathbb{R}^d} \mathbb{1}_{\Pi_\mathsf{A}}(\tilde{x}, \tilde{x}) p_\gamma\left(x, y, \tilde{x} - \mathcal{T}_\gamma(x)\right) \mathrm{e}^{-\frac{\|\tilde{x} - \mathcal{T}_\gamma(x)\|^2}{2\gamma}} \mathrm{d}\tilde{x} \right. \\
&\left. + \int_{\mathbb{R}^d} \mathbb{1}_{\Pi_\mathsf{A}}\left(\tilde{x}, \mathcal{S}_\gamma\left(x, y, \tilde{x} - \mathcal{T}_\gamma(x)\right)\right) \left\{1 - p_\gamma\left(x, y, \tilde{x} - \mathcal{T}_\gamma(x)\right)\right\} \mathrm{e}^{-\frac{\|\tilde{x} - \mathcal{T}_\gamma(x)\|^2}{2\gamma}} \mathrm{d}\tilde{x} \right] ,
\end{aligned} \tag{4.23}$$

where $\Pi_\mathsf{A} = (\Pi, \Pi)^{\leftarrow}(\mathsf{A})$ and $\Delta_{\mathbb{R}^d} = \{(x, x) \ : \ x \in \mathbb{R}^d\}$. Note that marginally, by definition, the distribution of X_{k+1} given X_k is $\mathrm{R}_\gamma(X_k, \cdot)$. It is well-know (see *e.g.* [BDJ98, Section 3.3]) that $\tilde{Y}_{k+1}$ and $\mathcal{T}_\gamma(Y_k) + \sqrt{\gamma} Z_{k+1}$ have the same distribution given Y_k, and therefore the distribution of Y_{k+1} given Y_k is $\mathrm{R}_\gamma(Y_k, \cdot)$. As a result, for any $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathsf{X}$, $\mathrm{K}_\gamma((x,y), \cdot)$ is a transference plan between $\mathrm{R}_\gamma(x, \cdot)$ and $\mathrm{R}_\gamma(y, \cdot)$.

find ./ -type f -iname "" -print0 | xargs -0 sed -i '/searchstring/s/durmus2016high/durmus2016high/g' As emphasized previously, based on (4.21), to study convergence of R_γ for $\gamma \in (0, \bar{\gamma}]$<, we first give upper bounds for $\mathrm{K}^n_\gamma((x,y), \Delta^{\mathrm{c}}_\mathsf{X})$ for any $x, y \in \mathsf{X}$ and $n \in \mathbb{N}^\star$ under appropriate conditions on $\mathcal{T}_\gamma$ and Π.

A1. *The function $\Pi : \mathbb{R}^d \to \mathsf{X}$ is non expansive: i.e. for any $x, y \in \mathbb{R}^d$, $\|\Pi(x) - \Pi(y)\| \le \|x - y\|$.*

Note that **A**1 is satisfied if Π is the proximal operator [BC11, Proposition 12.27] associated with a convex lower semi-continuous function $\mathrm{f} : \mathbb{R}^d \to (-\infty, +\infty]$. For example, if $\mathrm{f}(x) = \sum_{i=1}^d |x_i|$, the associated proximal operator is the soft thresholding operator [PB14, Section 6.5.2]. If f is the convex indicator of a closed convex set $\mathsf{C} \subset \mathbb{R}^d$, defined by $\mathrm{f}(x) = 0$ for $x \in \mathsf{C}$, $\mathrm{f}(x) = +\infty$ otherwise, the proximal operator is simply the orthogonal projection onto C by [BC11, Example 12.21] and we define for any $x \in \mathbb{R}^d$

$$\Pi_\mathsf{C}(x) = \operatorname*{arg\,min}_{y \in \mathsf{C}} \|y - x\| \ . \tag{4.24}$$

First, the class of Markov chains defined by (4.19) contains Euler-Maruyama discretizations of diffusion processes with identity diffusion matrix and for which $\Pi = \mathrm{Id}$. Our results will be specified for this particular case in Section 4.1.4. Second, for the applications that we have in mind, the use of Markov chains defined by (4.19) with $\Pi \neq \mathrm{Id}$ satisfying **A**1, has been proposed based on optimization literature to sample non-smooth log-concave densities [DMP18; BEL15; DMM19; Ber18]. Finally, we will also make use of (4.19) with $\Pi = \Pi_{\mathsf{K}_n}$, where Π_{K_n} is defined by (4.24) with $\mathsf{C} \leftarrow \mathsf{K}_n$, and $(\mathsf{K}_n)_{n\in\mathbb{N}^\star}$ is a sequence of increasing compact sets of $\mathbb{R}^d$, to derive our results on diffusion processes in Section 4.1.5.

We now consider the following assumption on $\{\mathcal{T}_\gamma \ : \ \gamma \in (0, \bar{\gamma}]\}$. Let $\mathsf{A} \in \mathcal{B}(\mathbb{R}^{2d})$.

A2 (A). *There exists $\kappa : (0, \bar{\gamma}] \to \mathbb{R}$ such that for any $\gamma \in (0, \bar{\gamma}]$ and $(x, y) \in \mathsf{A} \cap \mathsf{X}^2$*

$$\|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\|^2 \le (1 + \gamma\kappa(\gamma)) \|x - y\|^2 \ . \tag{4.25}$$

Further, one of the following conditions holds for any $\gamma \in (0, \bar{\gamma}]$: (i) $\kappa(\gamma) < 0$; (ii) $\kappa(\gamma) \le 0$; (iii) $\kappa(\gamma) > 0$.

If $\mathcal{T}_\gamma(x) = x + \gamma b(x)$ and b is L-Lipschitz we have that $\mathbf{A2}(\mathbb{R}^d)$ holds for any with $\kappa(\gamma) = \mathrm{L}(2+\gamma\mathrm{L})$. Note that $\mathbf{A2}(\mathsf{X}^2)$-(i) or $\mathbf{A2}(\mathsf{X}^2)$-(ii) imply that for any $\gamma \in (0, \bar{\gamma}]$, $\mathcal{T}_\gamma$ is non-expansive itself (see $\mathbf{A1}$). For $\kappa : (0, \bar{\gamma}] \to \mathbb{R}$ and $\ell \in \mathbb{N}^\star$, $\gamma \in (0, \bar{\gamma}]$ such that $\gamma\kappa(\gamma) \in (-1, +\infty)$, define

$$\Xi_n(\kappa) = \gamma \sum_{k=1}^{n} (1 + \gamma\kappa(\gamma))^{-k} \, . \tag{4.26}$$

The following theorem gives a generalization of a minorization condition on autoregressive models [DM19, Section 6].

Theorem 4.1.4. *Let $\mathsf{A} \in \mathcal{B}(\mathbb{R}^{2d})$ and assume $\mathbf{A1}$ and $\mathbf{A2}(\mathsf{A})$. Let $(X_k, Y_k)_{k\in\mathbb{N}}$ be defined by (4.22) with $(X_0, Y_0) = (x, y) \in \mathsf{A} \cap \mathsf{X}^2$ and $\gamma \in (0, \bar{\gamma}]$. Then for any $n \in \mathbb{N}^\star$*

$$\mathbb{P}\left(X_n \neq Y_n \text{ and for any } k \in \{1, \dots, n-1\}, (X_k, Y_k) \in \mathsf{A}\right) \leq \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x, y) \left\{ 1 - 2\mathbf{\Phi}\left(-\frac{\|x - y\|}{2\Xi_n^{1/2}(\kappa)} \right) \right\} \, ,$$

where $\mathbf{\Phi}$ is the cumulative distribution function of the Gaussian distribution with zero mean and unit variance on $\mathbb{R}$.

Proof. The proof is a simple application of Theorem 4.1.43 in Section 4.1.7. □

Based on Theorem 4.1.4, since $\mathbb{P}(X_n \neq Y_n) = \mathrm{K}^n((x, y), \Delta_{\mathsf{X}}^{\mathrm{c}})$ where $(X_k, Y_k)_{k\in\mathbb{N}}$ is defined by (4.22) with $(X_0, Y_0) = (x, y) \in \mathsf{X}^2$, we can derive minorization conditions for the Markov kernel R_γ^n with $n \in \mathbb{N}^\star$ for any $\gamma \in (0, \bar{\gamma}]$ depending on the assumption we make on κ in $\mathbf{A2}(\mathsf{X}^2)$. More precisely, these minorization conditions are derived using $\mathrm{K}_\gamma^{\ell\lceil 1/\gamma \rceil}$ with $\ell \in \mathbb{N}^\star$. This is a requirement to obtain sharp bounds in the limit $\gamma \to 0$. Indeed, for any $x, y \in \mathsf{X}$, based only on the results of Theorem 4.1.4, we get that for any $\ell \in \mathbb{N}^\star$, $\lim_{\gamma\to 0} \|\delta_x \mathrm{R}_\gamma^\ell - \delta_y \mathrm{R}_\gamma^\ell\|_{\mathrm{TV}} \leq 1$, whereas the following proposition implies that for any $\ell \in \mathbb{N}^\star$, $\lim_{\gamma\to 0} \|\delta_x \mathrm{R}_\gamma^{\ell\lceil 1/\gamma \rceil} - \delta_y \mathrm{R}_\gamma^{\ell\lceil 1/\gamma \rceil}\|_{\mathrm{TV}} < 1$.

Proposition 4.1.5. *Let $\mathsf{A} \in \mathcal{B}(\mathbb{R}^{2d})$ and assume $\mathbf{A1}$ and $\mathbf{A2}(\mathsf{A})$ hold. Let $(X_k, Y_k)_{k\in\mathbb{N}}$ be defined by (4.22) with $(X_0, Y_0) = (x, y) \in \mathsf{A} \cap \mathsf{X}^2$ and $\gamma \in (0, \bar{\gamma}]$. Then for any $\ell \in \mathbb{N}^*$ and $\gamma \in (0, \bar{\gamma}]$,*

$$\mathbb{P}\left(X_{\ell\lceil\gamma\rceil} \neq Y_{\ell\lceil\gamma\rceil} \text{ and for any } k \in \{1, \dots, n-1\}, (X_k, Y_k) \in \mathsf{A}\right) \leq 1 - 2\mathbf{\Phi}\left(-\alpha^{-1/2}(\kappa, \gamma, \ell)\|x - y\|/2\right) \, , \tag{4.27}$$

where

(a) $\alpha(\kappa, \gamma, \ell) = -\kappa^{-1}(\gamma)\left[\exp(-\ell\kappa(\gamma)) - 1\right]$ *if* $\mathbf{A2}(\mathsf{A})$-(i) *holds ;*

(b) $\alpha(\kappa, \gamma, \ell) = \ell$ *if* $\mathbf{A2}(\mathsf{A})$-(ii) *holds ;*

(c) $\alpha(\kappa, \gamma, \ell) = \kappa^{-1}(\gamma)\left[1 - \exp\left\{-\ell\kappa(\gamma)/(1 + \gamma\kappa(\gamma))\right\}\right]$ *if* $\mathbf{A2}(\mathsf{A})$-(iii) *holds.*

Proof. The proof is postponed to Section 4.1.7. □

Depending on the conditions imposed on κ defined in $\mathbf{A2}(\mathsf{X}^2)$, we obtain the following consequences of Proposition 4.1.5 which establish, either an explicit convergence bound in total variation for R_γ, or a quantitative minorization condition satisfied by this kernel.

Corollary 4.1.6. *Assume* **A**1 *and* **A**2(X^2).

(a) If **A**2(X^2)-(i) *holds and* $\kappa_- = \sup_{\gamma\in(0,\bar\gamma]} \kappa(\gamma) < 0$. *Then, for any* $\gamma \in (0,\bar\gamma]$, R_γ *admits a unique invariant probability measure* π_γ *and we have for any* $\gamma \in (0,\bar\gamma]$, $x \in \mathbb{R}^d$ *and* $\ell \in \mathbb{N}^\star$,

$$\|\delta_x \mathrm{R}_\gamma^{\ell\lceil 1/\gamma\rceil} - \pi_\gamma\|_{\mathrm{TV}} \leq 1 - 2\int_{\mathbb{R}^d} \mathbf{\Phi}\left\{-(-\kappa_-)^{1/2}\|x-y\|/\{2(\exp(-\ell\kappa_-)-1)^{1/2}\}\right\} \mathrm{d}\pi_\gamma(y) \; .$$

(b) If **A**2(X^2)-(ii) *holds and, in addition, assume that for any* $\gamma \in (0,\bar\gamma]$, R_γ *admits an invariant probability measure* π_γ, *then we have for any* $\gamma \in (0,\bar\gamma]$, $x \in \mathbb{R}^d$ *and* $\ell \in \mathbb{N}^\star$,

$$\|\delta_x \mathrm{R}_\gamma^{\ell\lceil 1/\gamma\rceil} - \pi_\gamma\|_{\mathrm{TV}} \leq 1 - 2\int_{\mathbb{R}^d} \mathbf{\Phi}\left\{-\|x-y\|/(2\ell^{1/2})\right\} \mathrm{d}\pi_\gamma(y) \; .$$

Proof. The proof is postponed to Section 4.1.7. □

In other words, if $\mathcal{T}_\gamma$ is a contractive mapping, see **A**2(X^2)-(i), then for $x \in \mathbb{R}^d$ the convergence of $(\delta_x \mathrm{R}_\gamma^{\ell\lceil 1/\gamma\rceil})_{\ell\in\mathbb{N}^\star}$ to π_γ in total variation is exponential in ℓ. If $\mathcal{T}_\gamma$ is non expansive, see **A**2(X^2)-(ii), and R_γ admits an invariant probability measure π_γ, for any $x \in \mathbb{R}^d$, the convergence of $(\delta_x \mathrm{R}_\gamma^{\ell\lceil 1/\gamma\rceil})_{\ell\in\mathbb{N}^\star}$ to π_γ in total variation is linear in $\ell^{1/2}$. In the case where $\mathcal{T}_\gamma$ is non expansive, see **A**2(X^2)-(ii), or simply Lipschitz, see **A**2(X^2)-(iii) and no additional assumption is made, we do not directly obtain contraction in total variation but only minorization conditions.

Corollary 4.1.7. *Assume* **A**1 *and* **A**2(X^2). *Then, for any* $\gamma \in (0,\bar\gamma]$,

(a) if **A**2(X^2)-(ii) *holds, for any* $x, y \in \mathsf{X}$ *with* $\|x-y\| \leq M$ *with* $M \geq 0$ *and* $\ell \in \mathbb{N}^\star$ *with* $\ell \geq \lceil M^2 \rceil$,

$$\mathrm{K}_\gamma^{\ell\lceil 1/\gamma\rceil}((x,y),\Delta_{\mathsf{X}}^{\mathrm{c}}) \leq 1 - 2\mathbf{\Phi}\left(-1/2\right) \; ; \tag{4.28}$$

(b) if **A**2(X^2)-(iii) *holds, for any* $x, y \in \mathsf{X}$ *and* $\ell \in \mathbb{N}^\star$,

$$\mathrm{K}_\gamma^{\ell\lceil 1/\gamma\rceil}((x,y),\Delta_{\mathsf{X}}^{\mathrm{c}}) \leq 1 - 2\mathbf{\Phi}\left\{-(1+\bar\gamma)^{1/2}(1+\kappa_+)^{1/2}\,\|x-y\|\,/2\right\} \; , \tag{4.29}$$

where $\kappa_+ = \sup_{\gamma\in(0,\bar\gamma]} \kappa(\gamma)$.

Proof. The proof is postponed to Section 4.1.7. □

In our application below, we are mainly interested in the case where R_γ satisfies a geometric drift condition. Let $(\mathsf{Y},\mathcal{Y})$ be a measurable space, $\lambda \in (0,1)$, $A \geq 0$, $V : \mathsf{Y} \to [1,+\infty)$ be a measurable function and $\mathsf{C} \in \mathcal{Y}$.

$\mathbf{D}_{\mathrm{d}}(V,\lambda,A,\mathsf{C})$. *A Markov kernel* R *on* $\mathsf{Y}\times\mathcal{Y}$ *satisfies the discrete Foster-Lyapunov drift condition if for all* $y \in \mathsf{Y}$

$$\mathrm{R}V(y) \leq \lambda V(y) + A\mathbb{1}_{\mathsf{C}}(y) \; .$$

The index d in $\mathbf{D}_{\mathrm{d}}$ stands for "discrete" as we will introduce the continuous-time counterpart of this drift condition, denoted by $\mathbf{D}_{\mathrm{c}}$, in Section 4.1.5. Note that this drift condition implies the existence of an invariant probability measure if R is a Feller kernel and the level sets of V are compact, see [Dou+18, Theorem 12.3.3]. In the sequel, we are interested in establishing convergence results in the Wasserstein metric $\mathbf{W}_{\mathbf{c}}$ associated with the cost

$$\mathbf{c} : (x, y) \mapsto \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x, y) W(x, y) \tag{4.30}$$

where $W : \mathsf{X} \times \mathsf{X} \to [0, +\infty)$ satisfies for any $x, y, z \in \mathsf{X}$, $W(x, y) = W(y, x)$, $W(x, z) \leq W(x, y) + W(y, z)$ and $W(x, y) = 0$ implies that $x = y$. Note that under these conditions on W, $\mathbf{c}$ defines a metric on $\mathbb{R}^d$. Let μ, ν be two probability measures over $\mathcal{X}$, we highlight three cases.

- total variation: if $W = 1$ then $\mathbf{W}_{\mathbf{c}}(\mu, \nu) = \|\mu - \nu\|_{\mathrm{TV}}$;
- V-norm: if $W(x, y) = \{V(x) + V(y)\}/2$ where $V : \mathbb{R}^d \to [1, +\infty)$ is measurable then $\mathbf{W}_{\mathbf{c}}(\mu, \nu) = \|\mu - \nu\|_V$;
- total variation + Kantorovitch-Rubinstein metric: if $W(x, y) = 1 + \vartheta \|x - y\|$ with $\vartheta > 0$, then by definition of Wasserstein metrics, $\mathbf{W}_{\mathbf{c}}(\mu, \nu) \geq \|\mu - \nu\|_{\mathrm{TV}} + \vartheta \mathbf{W}_1(\mu, \nu)$.

We now state convergence bounds for Markov kernels which satisfy one of the conclusions of Corollary 4.1.7. Indeed, in order to deal with the two assumptions **A**2(X^2)-(ii) and **A**2(X^2)-(iii) together, we provide a general result regarding the contraction of R_γ in the metric $\mathbf{W}_{\mathbf{c}}$ for some cost function $\mathbf{c}$ on X^2. This result is based on an abstract condition on $\tilde{\mathrm{K}}_\gamma^{\lceil 1/\gamma \rceil} \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}$, which is satisfied under **A**2(X^2)-(ii) or **A**2(X^2)-(iii) by Corollary 4.1.7 with $\tilde{\mathrm{K}}_\gamma \leftarrow \mathrm{K}_\gamma$, and a drift condition for $\tilde{\mathrm{K}}_\gamma$, where $\tilde{\mathrm{K}}_\gamma$ is a Markov coupling kernel for R_γ. We recall that for any $M \geq 0$,

$$\Delta_{\mathsf{X},M} = \{(x, y) \in \mathsf{X} : \|x - y\| \leq M\} . \tag{4.31}$$

Theorem 4.1.8. *Assume that there exist $\lambda \in (0, 1)$, $A \geq 0$, $\tilde{M}_{\mathrm{d}} > 0$, a measurable function $W : \mathsf{X} \times \mathsf{X} \to [1, +\infty)$, $\mathsf{C} \in \mathcal{X}^{\otimes 2}$ with $\mathsf{C} \subset \Delta_{\mathsf{X},\tilde{M}_{\mathrm{d}}}$ and for any $\gamma \in (0, \bar{\gamma}]$, $\tilde{\mathrm{K}}_\gamma$ a Markov coupling kernel for R_γ satisfying $\mathbf{D}_{\mathrm{d}}(W, \lambda^\gamma, A\gamma, \mathsf{C})$. Further, assume that for any $\gamma \in (0, \bar{\gamma}]$, Δ_{X} is absorbing for $\tilde{\mathrm{K}}_\gamma$, i.e. for any $x \in \mathsf{X}$, $\tilde{\mathrm{K}}_\gamma \mathbb{1}_{\Delta_{\mathsf{X}}}(x, x) = 1$, and that there exists $\boldsymbol{\Psi} : (0, \bar{\gamma}] \times \mathbb{N}^\star \times \mathbb{R}_+ \to [0, 1]$ such that for any $\gamma \in (0, \bar{\gamma}]$, $\ell \in \mathbb{N}^\star$ and $x, y \in \mathsf{X}$*

$$\tilde{\mathrm{K}}_\gamma^{\ell \lceil 1/\gamma \rceil}((x, y), \Delta_{\mathsf{X}}^{\mathrm{c}}) \leq 1 - \boldsymbol{\Psi}(\gamma, \ell, \|x - y\|) , \tag{4.32}$$

and for any $M \geq 0$, $\inf_{(x,y) \in \Delta_{\mathsf{X},M}} \boldsymbol{\Psi}(\gamma, \ell, \|x - y\|) > 0$. Then, for any $\gamma \in (0, \bar{\gamma}]$, $\ell \in \mathbb{N}^\star$ and $x, y \in \mathsf{X}$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \tilde{\mathrm{K}}_\gamma^k \mathbf{c}(x, y) \leq \lambda^{k\gamma/4} [\bar{D}_1 \mathbf{c}(x, y) + \bar{D}_2 \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x, y)] + \bar{C}_1 \bar{\rho}_1^{k\gamma/4} \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x, y) , \tag{4.33}$$

where

$$\begin{aligned}
\bar{D}_1 &= 1 + 4A \log^{-1}(1/\lambda)/\lambda^{\bar{\gamma}} , \quad \bar{D}_2 = \bar{D}_1 A \lambda^{-(1+\bar{\gamma})\ell}(1 + \bar{\gamma})\ell , \\
\bar{C}_1 &= 8A \log^{-1}(1/\bar{\rho}_1)/\bar{\rho}_1^{\bar{\gamma}} , \quad \log(\bar{\rho}_1) = \{\log(\lambda) \log(1 - \bar{\varepsilon}_{\mathrm{d},1})\} / \{-\log(\bar{c}_1) + \log(1 - \bar{\varepsilon}_{\mathrm{d},1})\} , \\
\bar{c}_1 &= \tilde{B}_{\mathrm{d}} + A\lambda^{-(1+\bar{\gamma})\ell}(1 + \bar{\gamma})\ell , \\
\bar{\varepsilon}_{\mathrm{d},1} &= \inf_{\gamma \in (0,\bar{\gamma}],\ (x,y) \in \Delta_{\mathsf{X},\tilde{M}_{\mathrm{d}}}} \boldsymbol{\Psi}(\gamma, \ell, \|x - y\|) \qquad \tilde{B}_{\mathrm{d}} = \sup_{(x,y) \in \mathsf{C}} W(x, y) .
\end{aligned}$$

In addition, if $\bar{\gamma} \leq 1$ and $\bar{\varepsilon}_{\mathrm{d},1} \leq 1 - \mathrm{e}^{-1}$, then

$$\log^{-1}(\bar{\rho}_1^{-1}) \leq \left[1 + \log(\tilde{B}_{\mathrm{d}}) + \log(1 + 2A\ell) + 2\ell \log(\lambda^{-1})\right] \Big/ \left[\log(\lambda^{-1}) \bar{\varepsilon}_{\mathrm{d},1}\right] .$$

Proof. The proof is postponed to Section 4.1.7. □

We emphasize that (4.32) is satisfied under **A**2(X^2)-(ii) or **A**2(X^2)-(iii) by Corollary 4.1.7 with $\tilde{\mathrm{K}}_\gamma \leftarrow \mathrm{K}_\gamma$.

Further, note that in (4.74), the leading term, $\bar{C}_1 \bar{\rho}_1^{k\gamma/4}$, does not depend on $x, y \in \mathsf{X}$. Indeed, the rate in front of the initial conditions $W(x,y)$ is given by $\lambda^{\gamma/4}$ which is always smaller than $\bar{\rho}_1^{\gamma/4}$. Therefore, Theorem 4.1.8 implies in particular that for any $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathsf{K}$ and $k \in \mathbb{N}$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \bar{\rho}_1^{k\gamma/4} [\bar{D}_1 + \bar{D}_2 + \bar{C}_1] \mathbf{c}(x,y) \; . \tag{4.34}$$

We conclude this section with two propositions which highlight the usefulness of the conclusions of Theorem 4.1.8 to establish convergence estimates with respect to different metrics. First, in Proposition 4.1.9, under additional conditions on $\mathbf{\Psi}$ and on W (which will be satisfied in our applications, see Corollary 4.1.14) we get a similar result to (4.34) replacing $\mathbf{c}$ by $(x,y) \mapsto \|x - y\|$, *i.e.* replacing $\mathbf{W}_{\mathbf{c}}$ by $\mathbf{W}_1$.

Proposition 4.1.9. *Assume that the conditions of Theorem 4.1.8 are satisfied with for any $x, y \in \mathsf{X}$, $W(x,y) = 1 + \vartheta \|x - y\|$, where $\vartheta > 0$. In addition, assume that the following conditions hold.*

(i) For any $\gamma \in (0, \bar{\gamma}]$, $t \mapsto \mathbf{\Psi}(\gamma, 1, t)$ is convex on $\mathbb{R}_+$, admits a right-derivative at 0, denoted by $\mathbf{\Psi}'(\gamma, 1, 0)$, and $\mathbf{a} = \inf_{\gamma \in (0,\bar{\gamma}]} \mathbf{\Psi}'(\gamma, 1, 0) > -\infty$.

(ii) There exists $\varkappa \geq 0$ such that for any $x, y \in \mathsf{X}$, $\tilde{\mathrm{K}}_\gamma \|x - y\| \leq (1 + \gamma\varkappa) \|x - y\|$.

Then there exist $\bar{D}_3 \geq 0$ and $\bar{\rho}_1 \in [0,1)$ such that for any $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathsf{X}$ and $k \in \mathbb{N}$

$$\mathbf{W}_1(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \tilde{\mathrm{K}}_\gamma^k \|x - y\| \leq \bar{D}_3 \bar{\rho}_1^{k\gamma/4} \|x - y\| \; , \tag{4.35}$$

with $\bar{\rho}_1$ given in Theorem 4.1.8 and $\bar{D}_3$ explicit in the proof.

Proof. The proof is postponed to Section 4.1.7. □

As a consequence, if X is closed, the Markov kernel R_γ admits a unique invariant probability measure π_γ, *i.e.* $\pi_\gamma = \pi_\gamma \mathrm{R}_\gamma$, using [GD03, Chapter 1, 6, A.1], since

$$\mathscr{P}_1(\mathsf{X}) = \{\mu \text{ probability measure on } (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \; : \; \int_{\mathbb{R}^d} \|x\| \, \mathrm{d}\mu(x) < +\infty\} \; ,$$

endowed with $\mathbf{W}_1$ is complete, see [Vil09, Theorem 6.18]. Further, for any $\gamma \in (0, \bar{\gamma}]$, using [Mey67, Theorem 1], there exists a distance $\mathbf{d}_\gamma$ on $\mathscr{P}_1(\mathsf{X})$, topologically equivalent to $\mathbf{W}_1$, such that $\mathscr{P}_1$ is complete and for any $x, y \in \mathbb{R}^d$ and $k \in \mathbb{N}$

$$\mathbf{d}_\gamma(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \rho^{k\gamma/4} \mathbf{d}_\gamma(\delta_x, \delta_y) \; .$$

A similar result to (4.35) in Proposition 4.1.9 can be derived when replacing $\mathbf{W}_1$ by $\mathbf{W}_p$ with $p \in \mathbb{N}$, if we assume some Foster-Lyapunov condition with respect to $(x,y) \mapsto \|x - y\|^p$.

Proposition 4.1.10. *Assume that there exist $\bar{\rho} \in (0,1]$, $\bar{D} \geq 0$ and for any $\gamma \in (0, \bar{\gamma}]$, $\tilde{\mathrm{K}}_\gamma$ a Markov coupling kernel for R_γ satisfying for any $x, y \in \mathsf{X}$ and $k \in \mathbb{N}$*

$$\tilde{\mathrm{K}}_\gamma^k \|x - y\| \leq \bar{D} \bar{\rho}^{k\gamma} \|x - y\| \; .$$

In addition, assume that for any $\gamma \in (0, \bar{\gamma}]$ and $q \in \mathbb{N}$, $\tilde{\mathrm{K}}_\gamma$ satisfies $\mathbf{D}_{\mathrm{d}}((x,y) \mapsto \|x-y\|^q, \tilde{\lambda}_q^\gamma, \tilde{A}_q \gamma)$ with $\tilde{\lambda}_q \in (0,1]$ and $\tilde{A}_q \geq 0$. Then, for any $p \geq 1$ and $\alpha \in (p, +\infty)$ there exists $\bar{D}_{4,\alpha} \geq 0$ such that

$$\mathbf{W}_p(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \left(\tilde{\mathrm{K}}_\gamma^k \|x-y\|^p\right)^{1/p} \leq \bar{D}_{4,\alpha} \bar{\rho}^{k\gamma/\alpha} \left\{\|x-y\| + \|x-y\|^{1/\alpha}\right\} ,$$

with $\bar{D}_{4,\alpha}$ explicit in the proof.

Proof. The proof is postponed to Section 4.1.7. □

4.1.4 Application to the projected Euler-Maruyama discretization

Here we consider the case in which the operator $\mathcal{T}_\gamma$ in (4.19) is given by the discretization of a diffusion (4.1). More precisely, for $b : \mathbb{R}^d \to \mathbb{R}^d$, we study the projected Euler-Maruyama discretization associated to the diffusion with drift function b and diffusion coefficient Id, *i.e.* we consider the following assumption for $\mathsf{X} \subset \mathbb{R}^d$.

B1 (X). *X is assumed to be a closed convex (non-empty) subset of $\mathbb{R}^d$, $\Pi = \Pi_{\mathsf{X}}$ is the orthogonal projection onto X defined in (4.24) and*

$$\mathcal{T}_\gamma(x) = x + \gamma b(x) \text{ for any } \gamma > 0 \text{ and } x \in \mathsf{X} , \tag{4.36}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ is continuous.

Note that if $\mathsf{X} = \mathbb{R}^d$ and $\Pi = \mathrm{Id}$, then this scheme is the classical Euler-Maruyama discretization of a diffusion with drift b and diffusion coefficient Id. The application to the tamed Euler-Maruyama discretization of the results of Section 4.1.3 is given in Section 4.1.7. In what follows, we show the convergence in weighted total variation for the projected Euler-Maruyama discretization and discuss the dependency of the constants appearing in the bounds we obtain with respect to the properties we assume on the drift b. We first derive minorization conditions or convergence in total variation depending on the regularity/curvature assumption on the drift b in Section 4.1.4. Drift conditions and the ensuing convergence when combined with the minorization assumption are studied in Section 4.1.4.

Minorization condition

First, we show that some regularity/curvature conditions on the drift b imply condition $\mathbf{A}2(\mathsf{X}^2)$ for $\mathcal{T}_\gamma$ given by (4.36). Let $\mathtt{m} \in \mathbb{R}$.

B2. *There exists $\mathtt{L} \geq 0$ such that b is $\mathtt{L}$-Lipschitz, i.e. for any $x, y \in \mathsf{X}$, $\|b(x) - b(y)\| \leq \mathtt{L}\|x-y\|$ and $b(0) = 0$.*

B3 ($\mathtt{m}$). *For any $x, y \in \mathsf{X}$,*

$$\langle b(x) - b(y), x - y \rangle \leq -\mathtt{m} \|x-y\|^2 .$$

Note that **B**2 implies **B**3($-\mathtt{L}$). However, we are interested in the case where $|\mathtt{m}|$ is possibly strictly smaller than $\mathtt{L}$. If there exists $U \in \mathrm{C}^1(\mathsf{X})$ such that for any $x \in \mathsf{X}$, $b(x) = -\nabla U(x)$ and **B**3($\mathtt{m}$) holds with $\mathtt{m} = 0$, respectively $\mathtt{m} > 0$ then U is convex, respectively strongly convex. Note that **B**3(0) does not imply that $\mathcal{T}_\gamma$ given by (4.36) is non-expansive, therefore we consider the following assumption.

B4. *There exists $\mathtt{m}_b > 0$ such that for any $x, y \in \mathsf{X}$,*

$$\langle b(x) - b(y), x - y \rangle \leq -\mathtt{m}_b \|b(x) - b(y)\|^2 .$$

Note that **B**4 implies that **B**2 with $\mathtt{L} = \mathtt{m}_b^{-1}$ and **B**3(0) hold. Conversely, in the case where $\mathsf{X} = \mathbb{R}^d$ and there exists $U \in \mathrm{C}^1(\mathbb{R}^d)$ such that for any $x \in \mathbb{R}^d$, $b(x) = -\nabla U(x)$, [Nes04, Theorem 2.1.5] implies that under **B**2 and **B**3(0), **B**4 holds with $\mathtt{m}_b = \mathtt{L}^{-1}$. Based on Proposition 4.1.11 and assuming **B** 1, we obtain the following results on the Markov kernel R_γ defined by (4.20) with $\gamma > 0$.

Proposition 4.1.11. *Assume* **B**1(X) *holds for* $\mathsf{X} \subset \mathbb{R}^d$.

(a) If **B**2 *and* **B**3($\mathtt{m}$) *hold with* $\mathtt{m} \in \mathbb{R}$. *Then* (4.25) *in* **A**2(X^2) *holds for any* $\gamma > 0$ *with* $\kappa(\gamma) = -2\mathtt{m} + \mathtt{L}^2\gamma$. *In particular, if* $\mathtt{m} > 0$ *then* **A**2(X^2)-(i) *holds for any* $\bar{\gamma} < 2\mathtt{m}/\mathtt{L}^2$ *and if* $\mathtt{m} \leq 0$ *then* **A**2(X^2)-(iii) *holds for any* $\bar{\gamma} > 0$ *;*

(b) If **B**4 *holds, then* **A**2(X^2)-(ii) *holds with* $\kappa(\gamma) = 0$ *for any* $\bar{\gamma} \leq 2\mathtt{m}_b$.

Proof. The proof is postponed to Section 4.1.7 □

Combining Proposition 4.1.11 and Proposition 4.1.5 and/or Corollary 4.1.6, we can draw the following conclusions.

If **B**2 and **B**3($\mathtt{m}$) hold with $\mathtt{m} > 0$, then we obtain, by Proposition 4.1.11-(a) and Proposition 4.1.5-(a), that for any $\gamma \in \left(0, 2\mathtt{m}/\mathtt{L}^2\right)$ and $\ell \in \mathbb{N}^\star$, (4.27) holds with $\alpha = \alpha_-$ given by

$$\alpha_-(\kappa, \gamma, \ell) = -\frac{\exp(-\ell(-2\mathtt{m} + \mathtt{L}^2\gamma)) - 1}{-2\mathtt{m} + \mathtt{L}^2\gamma} \, .$$

In addition, Corollary 4.1.6-(a) implies that for any $\gamma \in \left(0, 2\mathtt{m}/\mathtt{L}^2\right)$ and $x \in \mathsf{X}$, $(\delta_x \mathrm{R}_\gamma^{\lceil 1/\gamma \rceil \ell})_{\ell \in \mathbb{N}}$ converges exponentially fast to its invariant probability measure π_γ in total variation, with a rate which does not depend on γ, but only on $\mathtt{m}$.

Under **B**4, combining Proposition 4.1.11-(b) and Corollary 4.1.7-(a) we obtain that on any compact set $\mathsf{K} \subset \mathsf{X}$, $\mathrm{R}_\gamma^{\lceil 1/\gamma \rceil \ell}$ satisfies the minorization condition (4.28) with $\ell \geq \mathrm{diam}(\mathsf{K})^2$. In addition, if R_γ admits an invariant probability measure π_γ, then Corollary 4.1.6-(b) implies that for any $\gamma \in (0, 2\mathtt{m}_b]$ and $x \in \mathsf{X}$, $(\delta_x \mathrm{R}_\gamma^{\lceil 1/\gamma \rceil \ell})_{\ell \in \mathbb{N}}$ converges linearly in $\ell^{1/2}$ to π_γ in total variation.

In the case where **B**2 and **B**3($\mathtt{m}$) are satisfied with $\mathtt{m} \in \mathbb{R}_-$, we obtain that for any $\gamma > 0$ and $\ell \in \mathbb{N}^\star$, (4.27) holds with $\alpha = \alpha_+$ given by

$$\begin{aligned} &\alpha_+(\kappa, \gamma, \ell) \\ &\quad = (-2\mathtt{m} + \mathtt{L}^2\gamma)^{-1} \left\{1 - \exp\left[-\ell(-2\mathtt{m} + \mathtt{L}^2\gamma)/(1 + \gamma(-2\mathtt{m} + \mathtt{L}^2\gamma))\right]\right\} \leq (-2\mathtt{m} + \mathtt{L}^2\gamma)^{-1} \, , \end{aligned} \tag{4.37}$$

which implies that the bound given by Proposition 4.1.5-(c) does not go to 0 when ℓ goes to infinity. Therefore we cannot directly conclude that the Markov chain converges in total variation. However, by Proposition 4.1.11-(a), Corollary 4.1.7-(b) shows that for any $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ and $\ell \in \mathbb{N}^\star$, $\mathrm{R}_\gamma^{\lceil 1/\gamma \rceil \ell}$ satisfies the minorization condition (4.29), with constants which only depend on $\mathtt{m}$ and $\mathtt{L}$. Note however that in (4.37) the influence of $\mathtt{m}$ is different than the one of $\mathtt{L}$ and this result justifies the two assumptions **B**2 and **B**3($\mathtt{m}$).

Drift conditions and convergence

In the sequel of this section, we consider several assumptions on the drift function b which imply Foster-Lyapunov drift conditions on the Markov coupling kernel K_γ defined in (4.23). These results in combination with Proposition 4.1.11 will allow us to use Theorem 4.1.8, see also Theorem 4.1.46 in Section 4.1.7.

Strongly convex at infinity First, we consider conditions on b which imply that R_γ for $\gamma \in (0, \bar{\gamma}]$, is geometrically convergent in a metric which dominates the total variation distance and the Wasserstein distance of order 1. This result will be an application of Theorem 4.1.8 and the constants we end up with are independent of the dimension d. To do so, we establish that there exists a Lyapunov function W for which K_γ satisfies for $\gamma \in (0, \bar{\gamma}]$, $\mathbf{D}_{\mathrm{d}}(W, \lambda^\gamma, A\gamma, \Delta_{\mathsf{X}, M_{\mathrm{d}}})$ where $\Delta_{\mathsf{X}, M_{\mathrm{d}}}$ is given by (4.31) and $M_{\mathrm{d}} \geq 0$ which do not depend on the dimension.

C1. *There exist $R_1 > 0$ and $\mathtt{m}_1^+ > 0$ such that for any $x, y \in \mathsf{X}$ with $\|x - y\| \geq R_1$,*

$$\langle b(x) - b(y), x - y \rangle \leq -\mathtt{m}_1^+ \|x - y\|^2 \ .$$

This assumption has been considered in [EM19; Ebe16; LW16b; MMS18] and is sometimes referred to as strong convexity of the drift b outside of the ball $\mathrm{B}(0, R_1)$, see Section 4.1.2 for an example of such a setting. In the next proposition, we derive the announced drift for $W_1 : \mathsf{X}^2 \to [1, +\infty)$ defined for any $x, y \in \mathsf{X}$ by

$$W_1(x, y) = 1 + \|x - y\| / R_1 \ . \tag{4.38}$$

Proposition 4.1.12. *Assume* **B**1(X) *for $\mathsf{X} \subset \mathbb{R}^d$,* **B**2, **B**3($\mathtt{m}$) *for $\mathtt{m} \in \mathbb{R}_-$ and* **C**1. *Let K_γ be defined by* (4.23) *and $\bar{\gamma} \in (0, 2\mathtt{m}_1^+ / \mathrm{L}^2)$. Then the following hold:*

(a) for any $\gamma \in (0, \bar{\gamma}]$, we have

$$\mathrm{K}_\gamma \|x - y\| \leq \|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\| \leq (1 + \gamma(-\mathtt{m} + \bar{\gamma}\mathrm{L}^2/2)) \|x - y\| \ .$$

(b) for any $\gamma \in (0, \bar{\gamma}]$, K_γ satisfies $\mathbf{D}_{\mathrm{d}}(W_1, \lambda^\gamma, A\gamma, \Delta_{\mathsf{X}, R_1})$ where Δ_{X, R_1} is given by (4.31) *and*

$$\lambda = \exp\left[-(\mathtt{m}_1^+ - \bar{\gamma}\mathrm{L}^2/2)/2\right] \ , \quad A = \mathtt{m}_1^+ - \mathtt{m} \ . \tag{4.39}$$

(c) for any $p \in \mathbb{N}$ with $p \geq 2$, there exist $\lambda_p \in (0, 1]$, $A_p \geq 0$ and $\gamma \in (0, \bar{\gamma}]$, such that K_γ satisfies $\mathbf{D}_{\mathrm{d}}((x, y) \mapsto \|x - y\|^p, \lambda_p^\gamma, A_p \gamma)$, with explicit constants given in the proof.

Proof. The proof is postponed to Section 4.1.7. □

Theorem 4.1.13. *Assume* **B**1(X) *for $\mathsf{X} \subset \mathbb{R}^d$,* **B**2 *and* **C**1. *Assume in addition either* **B**3($\mathtt{m}$) *for $\mathtt{m} \in \mathbb{R}_-$ or* **B**4. *Then the conditions and the conclusions of Theorem 4.1.8 hold with $\bar{\gamma}$, λ and A given by Proposition 4.1.12-(b), $\tilde{M}_{\mathrm{d}} = R_1$, K_γ given by* (4.23) *for any $\gamma \in (0, \bar{\gamma}]$, $W = W_1$ defined in* (4.38), *and for any $\gamma \in (0, \bar{\gamma}]$, $\ell \in \mathbb{N}^\star$ and $t > 0$,*

$$\textit{under } \mathbf{B}3(\mathtt{m}) \ , \boldsymbol{\Psi}(\gamma, \ell, t) = 2\boldsymbol{\Phi}\{-t/(2\Xi^{1/2}_{\ell\lceil 1/\gamma \rceil}(\kappa))\} \ , \tag{4.40}$$

$$\textit{under } \mathbf{B}4 \ , \boldsymbol{\Psi}(\gamma, \ell, t) = \begin{cases} 2\boldsymbol{\Phi}\{-1/2\} & \textit{if } \ell \geq \lceil R_1 \rceil^2 \textit{ and } t \leq R_1 \ , \\ 2\boldsymbol{\Phi}\{-t/(2\Xi^{1/2}_{\ell\lceil 1/\gamma \rceil}(\kappa))\} & \textit{otherwise} \ , \end{cases} \tag{4.41}$$

where κ is given in Proposition 4.1.11-(a) and $\Xi_{\ell\lceil 1/\gamma \rceil}$ in (4.26).

Proof. First, note that for any $\gamma > 0$, Δ_{X} is absorbing for K_γ by definition of the reflection coupling, see (4.23). We assume that **B**3($\mathtt{m}$) holds. Let $\bar{\gamma} \in (0, 2\mathtt{m}_1^+ / \mathrm{L}^2)$. Using Proposition 4.1.12-(b) we obtain that

W_1 given by (4.38) satisfies $\mathbf{D}_{\mathrm{d}}(W_1, \lambda^\gamma, A\gamma, \Delta_{\mathsf{X},R_1})$ for any $\gamma \in (0, \bar\gamma]$ with λ and A given in (4.39). Using Theorem 4.1.4, Proposition 4.1.11-(a), we have for any $\gamma \in (0, \bar\gamma]$, $\ell \in \mathbb{N}^\star$ and $x, y \in \mathsf{X}$

$$\mathrm{K}_\gamma^{\ell\lceil 1/\gamma\rceil}((x,y), \Delta_{\mathsf{X}}^{\mathrm{c}}) \le 1 - 2\mathbf{\Phi}(-\Xi^{-1/2}_{\ell\lceil 1/\gamma\rceil}(\kappa)\, \|x-y\|\, /2) \;,$$

where $\kappa(\gamma) = -2\mathtt{m} + \gamma \mathtt{L}^2$, which concludes the proof.

The proof under **B**4 follows the same lines upon noting that **B**4 implies that **B**3(0) holds and using Proposition 4.1.11-(b) instead of Proposition 4.1.11-(a). □

Let $\bar\gamma \in (0, \max(2\mathtt{m}_1^+/\mathtt{L}^2, 1))$, $\ell \in \mathbb{N}^\star$ specified below, $\lambda_{\bar\gamma,a}, \rho_{\bar\gamma,a} \in (0,1)$ and $D_{\bar\gamma,1,a}, D_{\bar\gamma,2,a}, C_{\bar\gamma,a} \ge 0$ the constants given by Theorem 4.1.13, such that for any $k \in \mathbb{N}$, $\gamma \in (0, \bar\gamma]$ and $x, y \in \mathsf{X}$

$$\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \le \mathrm{K}_\gamma^k \mathbf{c}_1(x,y) \le \lambda_{\bar\gamma,a}^{k\gamma/4}[D_{\bar\gamma,1,a}\mathbf{c}_1(x,y) + D_{\bar\gamma,2,a}\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}] + C_{\bar\gamma,a}\rho_{\bar\gamma,a}^{k\gamma/4} \;, \tag{4.42}$$

with $\mathbf{c}_1(x,y) = \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y)(1 + \|x-y\|/R_1)$ for any $x, y \in \mathsf{X}$. Note that by (4.34), this result implies that for any $k \in \mathbb{N}$, $\gamma \in (0, \bar\gamma]$ and $x, y \in \mathsf{X}$

$$\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \le \{D_{\bar\gamma,1,a} + D_{\bar\gamma,2,a} + C_{\bar\gamma,a}\}\, \rho_{\bar\gamma,a}^{k\gamma} \mathbf{c}_1(x,y) \;.$$

We now give upper-bounds on $\rho_{\bar\gamma,a}$. Note that using Theorem 4.1.8, we obtain that the following limits exist and do not depend on $\mathtt{L}$

$$\begin{aligned} &D_{1,a} = \lim_{\bar\gamma\to 0} D_{\bar\gamma,1,a} \;, \quad D_{2,a} = \lim_{\bar\gamma\to 0} D_{\bar\gamma,2,a} \;, \quad C_a = \lim_{\bar\gamma\to 0} C_{\bar\gamma,a} \;, \\ &\lambda_a = \lim_{\bar\gamma\to 0} \lambda_{\bar\gamma,a} \;, \quad \rho_a = \lim_{\bar\gamma\to 0} \rho_{\bar\gamma,a} \;. \end{aligned} \tag{4.43}$$

Once again, we point out that $\lambda_{\bar\gamma,a} \le \rho_{\bar\gamma,a}$ in Theorem 4.1.8. In the following discussion we assume that **B**1(X) for $\mathsf{X} \subset \mathbb{R}^d$, **B**2 and **C**1 hold. We now give upper bounds on the rate $\rho_{\bar\gamma,a}$ and ρ_a using Theorem 4.1.8 depending on the assumptions in Theorem 4.1.13.

(a) If **B**4 holds, set $\ell = \lceil R_1^2 \rceil$. Using that $2\mathbf{\Phi}(-1/2) \le 1 - \mathrm{e}^{-1}$ and choosing $\mathtt{m}_1^+$ sufficiently small such that the conditions of Theorem 4.1.8 hold, we have

$$\begin{aligned} \log^{-1}(\rho_{\bar\gamma,a}^{-1}) \le \big[1 + \log(2) + \log\big(1 + 2(1+R_1^2)\mathtt{m}_1^+\big) \\ +2(1+R_1^2)(\mathtt{m}_1^+ - \bar\gamma\mathtt{L}^2/2)\big] \big/ \big[(\mathtt{m}_1^+ - \bar\gamma\mathtt{L}^2/2)\mathbf{\Phi}\{-1/2\}\big] \;. \end{aligned} \tag{4.44}$$

Taking the limit $\bar\gamma \to 0$ in (4.44) and using that for any $t \ge 0$, $\log(1+t) \le t$, we get that

$$\log^{-1}(\rho_a^{-1}) \le (1 + \log(2))/(\mathtt{m}_1^+ \mathbf{\Phi}\{-1/2\}) + 4(1+R_1^2)/\mathbf{\Phi}\{-1/2\} \;. \tag{4.45}$$

The leading term in (4.45) is of order $\max(R_1^2, 1/\mathtt{m}_1^+)$, which corresponds to the one identified in [EM19, Theorem 2.8] and is optimal, see [Ebe16, Remark 2.10].

(b) If **B**3($\mathtt{m}$) holds with $\mathtt{m} \in \mathbb{R}_-$, set $\ell = \lceil R_1^2 \rceil$. Choosing $\mathtt{m}_1^+ > 0$ sufficiently small and $R_1, |\mathtt{m}|$ sufficiently large such that the conditions of Theorem 4.1.8 hold, we have

$$\begin{aligned} \log^{-1}(\rho_{\bar\gamma,a}^{-1}) \le \big[1 + \log(2) + \log\big(1 + 2(1+R_1^2)\{\mathtt{m}_1^+ - \mathtt{m}\}\big) \\ +2(1+R_1^2)(\mathtt{m}_1^+ - \bar\gamma\mathtt{L}^2/2)\big] \Big/ \Big[(\mathtt{m}_1^+ - \bar\gamma\mathtt{L}^2/2)\mathbf{\Phi}\{-\Xi^{-1/2}_{\lceil 1/\bar\gamma\rceil\lceil R_1^2\rceil}(\kappa)R_1/2\}\Big] \;. \end{aligned} \tag{4.46}$$

Taking the limit $\bar{\gamma} \to 0$ in this result and using (4.26), we get that

$$\log^{-1}(\rho_a^{-1}) \leq \left[1 + \log(2) + \log(1 + 2\{\mathtt{m}_1^+ - \mathtt{m}\}) + 2\mathtt{m}_1^+\right] \Big/ \left[\mathtt{m}_1^+ \mathbf{\Phi}\{-(-\mathtt{m})^{1/2} R_1/(2 - 2\mathrm{e}^{2\mathtt{m}R_1^2})^{1/2}\}\right] . \tag{4.47}$$

The comparison between this rate and the ones derived in recent works is conducted in Section 4.1.2. We extend our result to other Wasserstein metrics in the following proposition.

Corollary 4.1.14. *Assume* **B**1(X) *for* $\mathsf{X} \subset \mathbb{R}^d$, **B**2 *and* **C**1. *Assume in addition either* **B**3(m) *for* $\mathtt{m} \in \mathbb{R}_-$ *or* **B**4. *Then for any* $p \in \mathbb{N}$, $\alpha \in (p, +\infty)$, $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathsf{X}$ *and* $k \in \mathbb{N}$ *we have*

$$\mathbf{W}_1(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq D_{3,\bar{\gamma},a} \rho_{\bar{\gamma},a}^{k\gamma/4} \|x - y\| ,$$
$$\mathbf{W}_p(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq D_{\alpha,\bar{\gamma},a} \rho_{\bar{\gamma},a}^{k\gamma/(4\alpha)} \left\{\|x - y\| + \|x - y\|^{1/\alpha}\right\} ,$$

where $\rho_{\bar{\gamma},a}$, $D_{3,\bar{\gamma},a}$ *and* $D_{\alpha,\bar{\gamma},a}$ *are given in* (4.49), *Proposition 4.1.9 and Proposition 4.1.10 respectively.*

Proof. The proof is postponed to Section 4.1.7. □

Other curvature conditions We now derive uniform ergodic convergence in V-norm under weaker conditions than **C**1. The following assumption ensures that the radial part of b decreases faster than a linear function with slope $-\mathtt{m}_2^+ < 0$.

C2. *There exist* $R_2 \geq 0$ *and* $\mathtt{m}_2^+ > 0$ *such that for any* $x \in \bar{\mathrm{B}}(0, R_2)^{\mathrm{c}} \cap \mathsf{X}$,

$$\langle b(x), x \rangle \leq -\mathtt{m}_2^+ \|x\|^2 .$$

In the next proposition we derive a Foster-Lyapunov drift condition for $W_2 : \mathsf{X}^2 \to [1, +\infty)$ defined for any $x, y \in \mathsf{X}$ by

$$W_2(x, y) = 1 + \|x\|^2/2 + \|y\|^2/2 , \qquad \mathbf{c}_2(x, y) = \mathbb{1}_{\Delta_\mathsf{X}}(x, y) W_2(x, y) . \tag{4.48}$$

Note that for any $x, y \in \mathsf{X}$, $W_2(x, y) = \{V(x) + V(y)\}/2$ with $V(x) = 1 + \|x\|^2$.

Proposition 4.1.15. *Assume* **B**1(X) *for* $\mathsf{X} \subset \mathbb{R}^d$, **B**2, **B**3(m) *for* $\mathtt{m} \in \mathbb{R}_-$ *and* **C**2. *Then* K_γ *defined by* (4.23) *satisfies* $\mathbf{D}_{\mathrm{d}}(W_2, \lambda^\gamma, A\gamma, \bar{\mathrm{B}}(0, R) \times \bar{\mathrm{B}}(0, R))$ *for any* $\gamma \in (0, \bar{\gamma}]$ *where* $\bar{\gamma} \in (0, 2\mathtt{m}_2^+/\mathtt{L}^2)$ *and*

$$\lambda = \exp[-(\mathtt{m}_2^+ - \bar{\gamma}\mathtt{L}^2/2)] , \quad A = d + 2R_2^2(\mathtt{m}_2^+ - \mathtt{m}) + 2\mathtt{m}_2^+ , \quad R = \sqrt{2}\lambda^{-\bar{\gamma}} A^{1/2} \log^{-1/2}(1/\lambda) .$$

Proof. The proof is postponed to Section 4.1.7. □

Theorem 4.1.16. *Assume* **B**1(X) *for* $\mathsf{X} \subset \mathbb{R}^d$, **B**2 *and* **C**2. *Assume in addition either* **B**3(m) *for* $\mathtt{m} \in \mathbb{R}_-$ *or* **B**4. *Then the conditions and conclusions of Theorem 4.1.8 hold with* $W = W_2$ *defined in* (4.48), $\bar{\gamma}$, λ, A *and* $\tilde{M}_{\mathrm{d}} = 2R$ *given by Proposition 4.1.15, and* $\mathbf{\Psi}$ *given by* (4.40) *or* (4.41).

Proof. The proof is similar to the one of Theorem 4.1.13. □

Let $\bar{\gamma} \in (0, \max(2\mathtt{m}_2^+/\mathtt{L}^2, 1))$, $\ell \in \mathbb{N}^\star$ specified below, $\lambda_{\bar{\gamma},b}, \rho_{\bar{\gamma},b} \in (0,1)$ and $D_{\bar{\gamma},1,b}, D_{\bar{\gamma},2,b}, C_{\bar{\gamma},b} \geq 0$ the constants given by Theorem 4.1.16, such that for any $k \in \mathbb{N}$, $\gamma \in (0, \bar{\gamma}]$ and $x, y \in \mathsf{X}$

$$\mathbf{W}_{\mathbf{c}_2}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \mathrm{K}_\gamma^k \mathbf{c}_2(x,y) \leq \lambda_{\bar{\gamma},b}^{k\gamma/4} [D_{\bar{\gamma},1,b} \mathbf{c}_2(x,y) + D_{\bar{\gamma},2,b} \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}] + C_{\bar{\gamma},b} \rho_{\bar{\gamma},b}^{k\gamma/4} , \tag{4.49}$$

with $\mathbf{c}_2(x,y) = \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y)\{V(x) + V(y)\}/2$ for any $x, y \in \mathsf{X}$. Note that by (4.48), this result implies that for any $k \in \mathbb{N}$, $\gamma \in (0, \bar{\gamma}]$ and $x, y \in \mathsf{X}$

$$\|\delta_x \mathrm{R}_\gamma^k - \delta_y \mathrm{R}_\gamma^k\|_V \leq \{D_{\bar{\gamma},1,b} + D_{\bar{\gamma},2,b} + C_{\bar{\gamma},b}\} \, \rho_{\bar{\gamma},b}^{k\gamma} \mathbf{c}_2(x,y) .$$

Note that using Theorem 4.1.8, we obtain that the following limits exist and do not depend on $\mathtt{L}$

$$\begin{aligned} &D_{1,b} = \lim_{\bar{\gamma} \to 0} D_{\bar{\gamma},1,b} , \quad D_{2,b} = \lim_{\bar{\gamma} \to 0} D_{\bar{\gamma},2,b} , \quad C_b = \lim_{\bar{\gamma} \to 0} C_{\bar{\gamma},b} , \\ &\lambda_b = \lim_{\bar{\gamma} \to 0} \lambda_{\bar{\gamma},b} , \quad \rho_b = \lim_{\bar{\gamma} \to 0} \rho_{\bar{\gamma},b} . \end{aligned} \tag{4.50}$$

We now discuss the dependency of ρ_b with respect to the introduced parameters, depending on the sign of $\mathtt{m}$ and based on Theorem 4.1.8.

(a) If **B**4 holds, set $\ell = \left\lceil \tilde{M}_{\mathrm{d}}^2 \right\rceil$. Then, if we consider $\mathtt{m}_2^+$ sufficiently small and $|\mathtt{m}|$ and R_2 sufficiently large such that the conditions of Theorem 4.1.8 hold, we have

$$\log^{-1}(\rho_b^{-1}) \leq \left[1 + 2\log(1+R^2) + \log(1+2A) + 2(1+4R^2)\mathtt{m}_2^+\right] / \left[\mathtt{m}_2^+ \mathbf{\Phi}(-1/2)\right] . \tag{4.51}$$

Note that the leading term on the right hand side of this equation is of order R^2, *i.e.* of order $\max(R_2^2, d/\mathtt{m}_2^+)$.

(b) If **B**3($\mathtt{m}$) with $\mathtt{m} \in \mathbb{R}_-$, set $\ell = \left\lceil \tilde{M}_{\mathrm{d}}^2 \right\rceil$. Then, if we consider $\mathtt{m}_2^+$ sufficiently small and $|\mathtt{m}|$ and R_2 sufficiently large such that the conditions of Theorem 4.1.8 hold, we have

$$\begin{aligned} \log^{-1}(\rho_b^{-1}) \leq \left[1 + 2\log(1+R^2) + \log(1+2A) + 2(1+4R^2)\mathtt{m}_2^+\right] \\ \Big/ \left[\mathtt{m}_2^+ \mathbf{\Phi}\{-2(-\mathtt{m})^{1/2} R/(2 - 2\mathrm{e}^{2\mathtt{m}R^2})^{1/2}\}\right] , \end{aligned} \tag{4.52}$$

Note that the right hand side of (4.52) is exponential in $-\mathtt{m}R^2$, *i.e.* exponential in $-\mathtt{m}d/\mathtt{m}_2^+$ and $-R_2^2(\mathtt{m}_2^+ - \mathtt{m})\mathtt{m}/\mathtt{m}_2^+$.

We now consider a condition which enforces weak curvature outside of a compact set.

C3. *There exist* $R_3, \mathtt{a} \geq 0, \mathtt{k}_1, \mathtt{k}_2 > 0$, *such that for any* $x \in \mathbb{R}^d$

$$\langle b(x), x \rangle \leq -\mathtt{k}_1 \|x\| \mathbb{1}_{\bar{\mathrm{B}}(0,R_3)^{\mathrm{c}}}(x) - \mathtt{k}_2 \|b(x)\|^2 + \mathtt{a}/2 .$$

In the case where $\mathsf{X} = \mathbb{R}^d$, $\Pi_{\mathsf{X}} = \mathrm{Id}$ and there exists $U \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R})$ such that **B**2 and **B**3(0) hold with $b = -\nabla U$ and $\int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \mathrm{d}x < +\infty$, then there exist $R_3 \geq 0$ and $\mathtt{k}_1 > 0$ such that **C**3 holds with $\mathtt{k}_2 = \mathtt{a} = 0$, see [Bak+08, Lemma 2.2]. Define $V : \mathsf{X} \to [1, +\infty)$ for any $x \in \mathsf{X}$ by

$$V(x) = \exp(\mathtt{m}_3^+ \phi(x)) , \qquad \phi(x) = \sqrt{1 + \|x\|^2} , \qquad \mathtt{m}_3^+ \in (0, \mathtt{k}_1/2] . \tag{4.53}$$

We also define for any $x, y \in \mathsf{X}$,

$$W_3(x,y) = \{V(x) + V(y)\}/2 , \qquad \mathbf{c}_3(x,y) = \mathbb{1}_{\Delta_{\mathsf{X}}}(x,y) W_3(x,y) . \tag{4.54}$$

Proposition 4.1.17. *Assume* **B**1(X) *for* $\mathsf{X} \subset \mathbb{R}^d$ *and* **C**3. *Then for any* $\gamma \in (0, \bar{\gamma}]$, K_γ *defined by* (4.23) *satisfies* $\mathbf{D}_{\mathrm{d}}(W_3, \lambda^\gamma, A\gamma, \bar{\mathrm{B}}(0,R) \times \bar{\mathrm{B}}(0,R))$ *where* $\bar{\gamma} \in (0, 2\mathtt{k}_2)$, $R_4 = \max(1, R_3, (d+\mathtt{a})/\mathtt{k}_1)$ *and*

$$\lambda = \mathrm{e}^{-(\mathtt{m}_3^+)^2/2} ,$$
$$A = \exp\left[\bar{\gamma}(\mathtt{m}_3^+(d+\mathtt{a}) + (\mathtt{m}_3^+)^2)/2 + \mathtt{m}_3^+(1+R_4^2)^{1/2}\right](\mathtt{m}_3^+(d+\mathtt{a})/2 + (\mathtt{m}_3^+)^2) ,$$
$$R = \log(2\lambda^{-2\bar{\gamma}} A \log^{-1}(1/\lambda)) .$$

Proof. The proof is postponed to Section 4.1.7. □

Theorem 4.1.18. *Assume* **B**1(X) *for* $\mathsf{X} \subset \mathbb{R}^d$, **B**2 *and* **C**3. *Assume in addition either* **B**3(m) *for* $\mathtt{m} \in \mathbb{R}_-$ *or* **B**4. *Then the conditions and conclusions of Theorem 4.1.8 hold with* $W = W_2$ *defined in* (4.48), $\bar{\gamma}$, λ, A *and* $\tilde{M}_{\mathrm{d}} = 2R$ *given by Proposition 4.1.17, and* $\mathbf{\Psi}$ *given by* (4.40) *or* (4.41).

Proof. The proof is similar to the one of Theorem 4.1.13. □

The dependency of the rate given by Theorem 4.1.18 with respect to the constants is discussed in Section 4.1.7.

4.1.5 Quantitative convergence bounds for diffusions

Main results

In this section, we aim at deriving quantitative convergence bounds with respect to some Wasserstein metrics for diffusion processes under regularity and curvature assumptions on the drift b. Consider the following SDE

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t , \tag{4.55}$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion and $b : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous drift.

When there exists a unique strong solution $(\mathbf{X}_t)_{t\geq 0}$ of (4.55) for any starting point $\mathbf{X}_0 = x$, with $x \in \mathbb{R}^d$, we define the semi-group $(\mathrm{P}_t)_{t\geq 0}$ for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $x \in \mathbb{R}^d$ and $t \geq 0$ by $\mathrm{P}_t(x, \mathsf{A}) = \mathbb{P}(\mathbf{X}_t \in \mathsf{A})$. We now turn to establishing that $(\mathrm{P}_t)_{t\geq 0}$ converges for some Wasserstein metrics. In order to prove this result we will rely on discretizations of the SDE (4.55). If the conditions of Theorem 4.1.8 are satisfied, these discretized processes are uniformly geometrically ergodic and taking the limit when the discretization stepsize goes to zero, we obtain the convergence of the associated diffusion processes.

First, assume that b is Lipschitz regular. We establish in Theorem 4.1.19 that for any $T \geq 0$ and $x, y \in \mathbb{R}^d$, the Wasserstein distance $\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T)$ is upper-bounded by the upper limit when $m \to +\infty$ of $\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_{T/m}^m, \delta_y \mathrm{R}_{T/m}^m)$, where R_γ is given for any $\gamma > 0$ in (4.20) and $\mathbf{c}$ is given in (4.30).

Second, this result is extended in Theorem 4.1.20 to cover the case where b is no longer Lipschitz regular but only locally Lipschitz regular, see **B**5. Theorem 4.1.19 and Theorem 4.1.20 are applications of a more general theory developed in Section 4.1.5. Let $M \geq 0$, we consider for any $x \in \mathbb{R}^d$

$$V_M(x,y) = \exp[M\phi(x)] , \qquad \phi(x) = (1 + \|x\|)^{1/2} . \tag{4.56}$$

Theorem 4.1.19. *Assume* **B**2 *and* $\sup_{x\in\mathbb{R}^d} \langle x, b(x)\rangle < +\infty$. *Then, for any starting point* $\mathbf{X}_0 = x$, *with* $x \in \mathbb{R}^d$, *there exists a unique strong solution to* (4.55). *In addition, for any* $W : \mathbb{R}^d \times \mathbb{R}^d \to [1, +\infty)$ *satisfying* $\sup_{(x,y)\in\mathbb{R}^d\times\mathbb{R}^d} \{W(x,y)(V_M(x) + V_M(y))^{-1}\} < +\infty$ *with* $M \geq 0$ *and* V_M *given in* (4.56), *we get that for any* $x, y \in \mathbb{R}^d$ *and* $T \geq 0$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T) \leq \limsup_{m\to+\infty} \mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_{T/m}^m, \delta_y \mathrm{R}_{T/m}^m) ,$$

where $\mathbf{c}$ *is given by* (4.30), $(\mathrm{P}_t)_{t\geq 0}$ *is the semigroup associated with* (4.55) *and for any* $\gamma \in (0, \bar{\gamma}]$, R_γ *is the Markov kernel associated with* (4.19) *where* $\mathcal{T}_\gamma(x) = x + \gamma b(x)$, $\mathsf{X} = \mathbb{R}^d$ *and* $\Pi = \mathrm{Id}$.

Proof. The proof is postponed to Section 4.1.7. □

We now weaken the Lipschitz regularity assumption and consider the following condition on the drift b.

B5. b *is locally Lipschitz, i.e. for any* $M \geq 0$, *there exists* $\mathtt{L}_M \geq 0$ *such that for any* $x, y \in \bar{\mathrm{B}}(0, M)$, $\|b(x) - b(y)\| \leq \mathtt{L}_M \|x - y\|$ *and* $b(0) = 0$.

As a consequence, under a mild integrability assumption, which will be satisfied in all of our applications, we obtain the following generalization of Theorem 4.1.19.

Theorem 4.1.20. *Assume* **B**3(𝔪), **B**5 *and that* $\sup_{x\in\mathbb{R}^d} \langle x, b(x)\rangle < +\infty$. *Then, for any starting point* $\mathbf{X}_0 = x$, *with* $x \in \mathbb{R}^d$, *there exists a unique strong solution of* (4.55). *In addition assume that for any* $x \in \mathbb{R}^d$ *and* $T \geq 0$ *there exists* $\varepsilon_b > 0$ *such that*

$$\sup_{s\in[0,T]} \left\{ \delta_x \mathrm{P}_s \|b(x)\|^{2(1+\varepsilon_b)} \right\} < +\infty \,. \tag{4.57}$$

Then, for any $W : \mathbb{R}^d \times \mathbb{R}^d \to [1, +\infty)$ *satisfying* $\sup_{(x,y)\in\mathbb{R}^d\times\mathbb{R}^d} \{W(x, y)(V_M(x)+V_M(y))^{-1}\} < +\infty$ *with* $M \geq 0$ *and* V_M *given in* (4.56), *we get that for any* $x, y \in \mathbb{R}^d$ *and* $T \geq 0$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T) \leq \limsup_{n\to+\infty} \limsup_{m\to+\infty} \mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}^m_{T/m,n}, \delta_y \mathrm{R}^m_{T/m,n}) \,,$$

where for any $x, y \in \mathbb{R}^d$, $\mathbf{c}(x, y) = \mathbb{1}_{\Delta^c_{\mathsf{X}}}(x, y) W(x, y)$, $(\mathrm{P}_t)_{t\geq 0}$ *is the semigroup associated with* (4.55) *and for any* $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$, $\mathrm{R}_{\gamma,n}$ *is the Markov kernel associated with* (4.19) *where* $\mathcal{T}_\gamma(x) = x + \gamma b(x)$, $\mathsf{X} = \bar{\mathrm{B}}(0, n)$ *and* $\Pi = \Pi_{\bar{\mathrm{B}}(0,n)}$.

Proof. The proof is postponed to Section 4.1.7. □

Note that (4.57) holds under mild conditions on the drift function, see Proposition 4.1.30. In the next section we apply these results to diffusion processes and derive sharp convergence bounds in the case where b satisfies some curvature assumption, similarly to Section 4.1.4.

Applications

In this section, we combine the results of Theorem 4.1.20 with the convergence bounds for discrete processes derived in Section 4.1.4, in order to obtain convergence bounds for continuous processes that are solutions of (4.55).

Strongly convex at infinity

Theorem 4.1.21. *Assume either* **B**3(𝔪) *for* $\mathtt{m} \in \mathbb{R}_-$ *or* **B**4. *Assume* **C**1, **B**5 *and in addition* $\sup_{x\in\mathbb{R}^d} \{\|b(x)\|^{2(1+\varepsilon_b)} \mathrm{e}^{-\mathtt{m}_1^+ \|x\|^2}\} <$ $+\infty$ *for some* $\varepsilon_b > 0$. *Then, for any* $T \geq 0$, *and* $x, y \in \mathbb{R}^d$

$$\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T) \leq \lambda_a^{T/4} (D_{1,a} \mathbf{c}_1(x, y) + D_{2,a} \mathbb{1}_{\Delta^c}(x, y)) + C_a \rho_a^{T/4} \mathbb{1}_{\Delta^c}(x, y) \,,$$

with $D_{1,a}, D_{2,a}, C_a \geq 0$, $\lambda_a, \rho_a \in (0, 1)$ *given by* (4.43) *and for any* $x, y \in \mathbb{R}^d$, $\mathbf{c}_1(x, y) = \mathbb{1}_{\Delta^c_{\mathsf{X}}}(x, y) W_1(x, y)$ *with* $W_1(x, y) = 1 + \|x - y\| / R_1$.

Proof. Le $T \geq 0$ and $x, y \in \mathbb{R}^d$. Using Theorem 4.1.19 or Proposition 4.1.30 and Theorem 4.1.20 we have

$$\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T) \leq \limsup_{n \to +\infty} \limsup_{m \to +\infty} \mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{R}^m_{T/m,n}, \delta_y \mathrm{R}^m_{T/m,n}) \ .$$

Let $n \in \mathbb{N}$ and $m \in \mathbb{N}^\star$ such that $x, y \in \bar{\mathrm{B}}(0,n)$ and $T/m \leq 2\mathtt{m}_1^+/\mathrm{L}_n^2$. Since $\mathbf{B}1(\bar{\mathrm{B}}(0,n))$ holds and $\mathbf{B}5$ implies $\mathbf{B}2$ on $\bar{\mathrm{B}}(0,n)$, we can apply Theorem 4.1.13 and we get

$$\begin{aligned}
&\mathbf{W}_{\mathbf{c}_1}(\delta_x \mathrm{R}^m_{T/m,n}, \delta_y \mathrm{R}^m_{T/m,n}) \\
&\qquad \leq \lambda^{T/4}_{T/m,a}(D_{T/m,1,a}\mathbf{c}_1(x,y) + D_{T/m,2,a}\mathbb{1}_{\Delta^c}(x,y)) + C_{T/m,a}\rho^{T/4}_{T/m,a}\mathbb{1}_{\Delta^c}(x,y) \ ,
\end{aligned}$$

where $D_{T/m,1,a}, D_{T/m,2,a}, C_{T/m,a}, \lambda_{T/m,a}$ and $\rho_{T/m,a}$ are given in (4.49). In addition, these quantities admit limits $D_{1,a}, D_{2,a}, C_a \geq 0$ and $\lambda_a, \rho_a \in (0,1)$ when $m \to +\infty$ which do not depend on L_n, hence on n, see (4.43). □

Note that $\mathbf{B}2$ implies $\mathbf{B}5$ and $\sup_{x \in \mathbb{R}^d}\{\|b(x)\|^{2(1+\varepsilon_b)}\,\mathrm{e}^{-\mathtt{m}_1^+\|x\|^2}\} < +\infty$ for some $\varepsilon_b > 0$.

Corollary 4.1.22. *Assume either* $\mathbf{B}3(\mathtt{m})$ *for* $\mathtt{m} \in \mathbb{R}_-$ *or* $\mathbf{B}4$. *Assume* $\mathbf{C}1$, $\mathbf{B}5$ *and in addition* $\sup_{x \in \mathbb{R}^d}\{\|b(x)\|^{2(1+\varepsilon_b)}\,\mathrm{e}^{-\mathtt{m}_1^+\|x\|^2}\} < +\infty$ *for some* $\varepsilon_b > 0$. *Then, for any* $p \in \mathbb{N}$, $\alpha \in (p, +\infty)$, $T \geq 0$, *and* $x, y \in \mathbb{R}^d$ *we have*

$$\begin{aligned}
&\mathbf{W}_1(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T) \leq D_{3,a}\rho_a^{T/4}\,\|x-y\| \ , \\
&\mathbf{W}_p(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T) \leq D_{\alpha,a}\rho_a^{T/(4\alpha)}\left\{\|x-y\| + \|x-y\|^{1/\alpha}\right\} \ ,
\end{aligned}$$

where ρ_a *is given in* (4.43), $D_{3,a} = \lim_{\bar\gamma \to 0} D_{3,\bar\gamma,a}$ *and* $D_{\alpha,a} = \lim_{\bar\gamma \to 0} D_{\alpha,\bar\gamma,a}$ *with* $D_{3,\bar\gamma,a}$ *and* $D_{\alpha,\bar\gamma,a}$ *given in Corollary 4.1.14.*

Proof. The proof is similar to the one of Theorem 4.1.21. □

The discussion on the dependency of ρ_a with respect to the parameters of the problem conducted in Section 4.1.4 still holds. We distinguish the following cases, assuming that the conditions of Theorem 4.1.8 are satisfied.

(a) If $\mathbf{B}4$ holds, we have

$$\log^{-1}(\rho_a^{-1}) \leq (1+\log(2))/(\mathbf{\Phi}\{-1/2\}\mathtt{m}_1^+) + 4R_1^2/\mathbf{\Phi}\{-1/2\} \ . \tag{4.58}$$

The leading term in (4.58) is of order $\max(R_1^2, 1/\mathtt{m}_1^+)$, which corresponds to the one identified in [Ebe16, Lemma 2.9] and is optimal, see [Ebe16, Remark 2.10].

(b) If $\mathbf{B}3(\mathtt{m})$ holds with $\mathtt{m} \in \mathbb{R}_-$, we have

$$\begin{aligned}
\log^{-1}(\rho_a^{-1}) \leq \left[1 + \log(2) + \log(1 + 2\{\mathtt{m}_1^+ - \mathtt{m}\}\{1+R_1^2\}) + 2\mathtt{m}_1^+(1+R_1^2)\right] \\
\Big/\left[\mathtt{m}_1^+\mathbf{\Phi}\{-(-\mathtt{m})^{1/2}R_1/(2-2\mathrm{e}^{2\mathtt{m}R_1^2})^{1/2}\}\right] \ .
\end{aligned} \tag{4.59}$$

We now give an upper-bound for (4.59) when both R and $\mathtt{m}$ are large. For any $t \geq C$ with $C \geq 0$ we have

$$\mathbf{\Phi}(-t)^{-1} \leq \sqrt{2\pi}(1+C^{-2})t\mathrm{e}^{t^2/2} \ . \tag{4.60}$$

As a consequence if we also have $R_1 \geq 2$, $1 \leq -\mathtt{m}R_1^2$ and using that for any $t \in (0,1)$, $-\log(1-t) \leq t$ as well as (4.60) we get that $\log^{-1}(\rho_a^{-1}) \leq \log^{-1}(\rho_{\max}^{-1})$

$$\log^{-1}(\rho_{\max}^{-1}) = C\left[1 + \log(1 + 2\{\mathtt{m}_1^+ - \mathtt{m}\}\{1+R_1^2\}) + 2\mathtt{m}_1^+(1+R_1^2)\right]R_1(-\mathtt{m})^{1/2}$$

$$\times \exp\left[-\mathtt{m}R_1^2/(4-4\mathrm{e}^{2\mathtt{m}R_1^2})\right] \Big/ \left[\mathtt{m}_1^+(1-\mathrm{e}^{2\mathtt{m}R_1^2})^{1/2}\right] \ ,$$

with $C = 2(1+\log(2))\sqrt{\pi} \approx 6.00$.

For a comparison of our results with recent works, see Section 4.1.2.

Other curvature conditions

Theorem 4.1.23. *Assume either* **B**3($\mathtt{m}$) *for* $\mathtt{m} \in \mathbb{R}_-$ *or* **B**4. *Assume* **C**2, **B**5 *and in addition* $\sup_{x\in\mathbb{R}^d}\{\|b(x)\|^{2(1+\varepsilon_b)}\,\mathrm{e}^{-\mathtt{m}_2^+\|x\|^2}\} <$ $+\infty$ *for some* $\varepsilon_b > 0$. *Then for any* $T \geq 0$ *and* $x, y \in \mathbb{R}^d$

$$\|\delta_x \mathrm{P}_T - \delta_y \mathrm{P}_T\|_V \leq (D_{1,b} + D_{2,b} + C_b)\rho_b^T \mathbf{c}_2(x,y) \ ,$$

with $D_{1,b}, D_{2,b}, C_b \geq 0$ *and* $\rho_b \in (0,1)$ *given by* (4.50) *and* $\mathbf{c}_2$ *defined in* (4.48).

Proof. The proof is identical to the one of Theorem 4.1.21 upon replacing Theorem 4.1.13 by Theorem 4.1.16. □

Theorem 4.1.24. *Assume either* **B**3($\mathtt{m}$) *for* $\mathtt{m} \in \mathbb{R}_-$ *or* **B**4. *Assume* **C**3, **B**5 *and in addition* $\sup_{x\in\mathbb{R}^d}\|b(x)\|^{2(1+\varepsilon_b)}\,\mathrm{e}^{-\mathtt{k}_1(1+\|x\|)^1}$ $+\infty$ *for some* $\varepsilon_b > 0$. *Then for any* $T \geq 0$ *and* $x, y \in \mathbb{R}^d$

$$\|\delta_x \mathrm{P}_T - \delta_y \mathrm{P}_T\|_V \leq (D_{1,c} + D_{2,c} + C_c)\rho_c^T \mathbf{c}_3(x,y) \ ,$$

with $D_{1,c}, D_{2,c}, C_c \geq 0$ *and* $\rho_c \in (0,1)$ *given by Section 4.1.7 and* $\mathbf{c}_3$ *defined in* (4.54).

Proof. The proof is postponed to Section 4.1.7. □

The rates we obtain in Theorem 4.1.23, respectively Theorem 4.1.24, are identical to the ones derived taking the limit $\bar{\gamma} \to 0$ in Theorem 4.1.16, respectively Theorem 4.1.18. An upper bound on ρ_b, respectively ρ_c, is provided in (4.51) and (4.52), respectively Section 4.1.7.

From discrete to continuous processes

In this section we present the general theory which leads to Theorem 4.1.19 and Theorem 4.1.20. First, we derive bounds between the discrete and continuous process given a family of approximating drift functions in Section 4.1.5. Second, we show in Section 4.1.5 that under mild regularity assumptions on b such families can be explicitly constructed.

Quantitative convergence bounds for diffusion processes We recall that the SDE under study is given by

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t \ ,$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion and $b : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous drift. In the sequel we will always consider the following assumption.

L1. *There exists a unique strong solution of* (4.55) *for any starting point* $\mathbf{X}_0 = x$, *with* $x \in \mathbb{R}^d$.

Under **L**1, the Markov semigroup P_t, whose definition is given in Section 4.1.5, exists for any time $t \geq 0$. Consider the extended infinitesimal generator $\mathcal{A}$ associated with $(\mathrm{P}_t)_{t\geq 0}$ and defined for any $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$ by

$$\mathcal{A}f = (1/2)\Delta f + \langle \nabla f, b\rangle \ .$$

Let $V \in \mathrm{C}^2(\mathbb{R}^d, [1,+\infty))$, $\zeta \in \mathbb{R}$ and $B \geq 0$

$\mathbf{D}_{\mathrm{c}}(V, \zeta, B)$. *The extended infinitesimal generator $\mathcal{A}$ satisfies the continuous Foster-Lyapunov drift condition if for all $x \in \mathbb{R}^d$*

$$\mathcal{A}V(x) \leq -\zeta V(x) + B \, .$$

This assumption is the continuous counterpart of $\mathbf{D}_{\mathrm{d}}(V, \lambda, A, \mathbb{R}^d)$. We start by drawing a link between the continuous drift condition $\mathbf{D}_{\mathrm{c}}(V, \zeta, B)$ and the discrete drift condition $\mathbf{D}_{\mathrm{d}}(V, \lambda, A, \mathbb{R}^d)$. The result and its proof are standard [MT93c, Theorem 2.1] but are given here for completeness. Denote by $(\mathcal{F}_t)_{t\geq 0}$ the filtration associated with $(\mathbf{B}_t)_{t\geq 0}$ satisfying the usual conditions [IW89, Chapter I, Section 5].

Lemma 4.1.25. *Let $\zeta \in \mathbb{R}$, $B \geq 0$ and $V \in \mathrm{C}^2(\mathbb{R}^d, [1, +\infty))$ such that $\lim_{\|x\|\to+\infty} V(x) = +\infty$. Assume* **L**1 *and $\mathbf{D}_{\mathrm{c}}(V, \zeta, B)$.*

(a) If $B = 0$, then for any $x \in \mathbb{R}^d$, $(V(\mathbf{X}_t)\mathrm{e}^{\zeta t})_{t\geq 0}$ is a $(\mathcal{F}_t)_{t\geq 0}$-supermartingale where $(\mathbf{X}_t)_{t\geq 0}$ is the solution of (4.55) starting from $\mathbf{X}_0 = x$.

(b) For any $t_0 > 0$, P_{t_0} satisfies $\mathbf{D}_{\mathrm{d}}(V, \exp(-\zeta t_0), B(1 - \exp(-\zeta t_0))/\zeta, \mathbb{R}^d)$.

Proof. The proof is postponed to Section 4.1.7. □

Consider a family of drifts $\{b_{\gamma,n} : \mathbb{R}^d \to \mathbb{R}^d \, : \, \gamma \in (0, \bar{\gamma}] \, , n \in \mathbb{N}\}$ for some $\bar{\gamma} > 0$. For all $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}$, we denote by $\tilde{\mathrm{R}}_{\gamma,n}$ the Markov kernel associated with (4.19) where $\mathcal{T}_\gamma(x) = x + \gamma b_{\gamma,n}(x)$, $\mathsf{X} = \mathbb{R}^d$ and $\Pi = \mathrm{Id}$. We will show that under the following assumptions the family $\{\tilde{\mathrm{R}}_{\gamma,n}^{\lceil T/\gamma \rceil} : \mathbb{R}^d \to \mathbb{R}^d \, : \, \gamma \in (0, \bar{\gamma}] \, , n \in \mathbb{N}\}$ approximates P_T for $T \geq 0$ as $\gamma \to 0$ and $n \to +\infty$.

L2. *There exist $\beta > 0$ and $C_1 \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$, $b_{\gamma,n} \in \mathrm{C}(\mathbb{R}^d, \mathbb{R}^d)$ and for any $x \in \mathbb{R}^d$,*

$$\|b(x) - b_{\gamma,n}(x)\|^2 \leq C_1 \gamma^\beta \|b(x)\|^2 \, .$$

The following assumption is mainly technical and is satisfied in our applications.

L3. *There exists $\varepsilon_b > 0$ such that $\sup_{s\in[0,T]}\{\delta_x \mathrm{P}_s \|b(x)\|^{2(1+\varepsilon_b)}\} < +\infty$, for any $x \in \mathbb{R}^d$ and $T \geq 0$.*

By Lemma 4.1.25-(a), if $\mathbf{D}_{\mathrm{c}}(V, \zeta, 0)$ is satisfied with $\zeta \in \mathbb{R}$, it holds that for any starting point $x \in \mathbb{R}^d$, $\sup_{t\in[0,T]} \mathbb{E}[V(\mathbf{X}_t)] \leq \mathrm{e}^{-\zeta T} V(x)$, where $(\mathbf{X}_t)_{t\geq 0}$ is solution of (4.55) starting from x. Therefore, if $\|b(x)\|^{2(1+\varepsilon_b)} \leq V(x)$ for any $x \in \mathbb{R}^d$, **L**3 is satisfied.

The proof of the next result relies on the combination of the Girsanov theorem with estimates on the drift functions, adapting [DM17, Theorem 10]. Similar strategies have also been used in [Dal17b; FG19; RRT17].

Proposition 4.1.26. *Assume* **L**1, **L**2 *and* **L**3. *Let $V : \mathbb{R}^d \to [1, +\infty)$. In addition, assume that for any $n \in \mathbb{N}$, $T \geq 0$ and $x \in \mathbb{R}^d$*

$$\mathrm{P}_T V^2(x) < +\infty \, , \qquad \limsup_{m\to+\infty} \tilde{\mathrm{R}}_{T/m,n}^m V^2(x) < +\infty \, .$$

Then for any $n \in \mathbb{N}$, $T \geq 0$ and $x \in \mathbb{R}^d$

$$\lim_{m\to+\infty} \|\delta_x \mathrm{P}_T - \delta_x \tilde{\mathrm{R}}_{T/m,n}^m\|_V = 0 \, ,$$

where $(\mathrm{P}_t)_{t\geq 0}$ is the semigroup associated with (4.55) and for any $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}$, $\tilde{\mathrm{R}}_{\gamma,n}$ is the Markov kernel associated with (4.19) where $\mathcal{T}_\gamma(x) = x + \gamma b_{\gamma,n}(x)$ and $\Pi = \mathrm{Id}$.

Proof. The proof is postponed to Section 4.1.7. □

If $V = 1$, Proposition 4.1.26 implies that $\lim_{m\to+\infty} \|\delta_x \mathrm{P}_T - \delta_x \tilde{\mathrm{R}}^m_{T/m,n}\|_{\mathrm{TV}} = 0$. Let $V : \mathbb{R}^d \to [1, +\infty)$ and $\mathbf{c} : \mathbb{R}^d \times \mathbb{R}^d \to [1, +\infty)$ a distance such that for any $x, y \in \mathbb{R}^d$, $\mathbf{c}(x, y) \leq \{V(x) + V(y)\}/2$. Then, under the conditions of Proposition 4.1.26, we obtain that for any $T \geq 0$, $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^d$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T) \leq \limsup_{m\to+\infty} \mathbf{W}_{\mathbf{c}}(\delta_x \tilde{\mathrm{R}}^m_{T/m,n}, \delta_y \tilde{\mathrm{R}}^m_{T/m,n}) \;, \tag{4.61}$$

Therefore, if for any $T \geq 0$, $\mathbf{W}_{\mathbf{c}}(\delta_x \tilde{\mathrm{R}}^m_{T/m,n}, \delta_y \tilde{\mathrm{R}}^m_{T/m,n})$ can be bounded uniformly in m using Theorem 4.1.8, we obtain an explicit bound for $\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T)$ for any $T \geq 0$. As a consequence, this result easily implies non-asymptotic convergence bounds of $(\mathrm{P}_t)_{t\geq 0}$ to its invariant measure if it exists. However, in our applications, global Lipschitz regularity on $b_{T/m,n} : \mathbb{R}^d \to \mathbb{R}^d$ is needed in order to apply Theorem 4.1.8 to $\tilde{\mathrm{R}}_{T/m,n}$ for $T \geq 0$, $m \in \mathbb{N}^\star$ and $n \in \mathbb{N}$. To be able to deal with the fact that $b_{T/m,n}$ is non necessarily globally Lipschitz, we consider an appropriate sequence of projected Euler-Maruyama schemes associated to a sequence of subsets of $\mathbb{R}^d$, $(\mathsf{K}_n)_{n\in\mathbb{N}}$ satisfying the following assumption.

L4. *For any $n \in \mathbb{N}$, K_n is convex and closed, and $\bar{\mathrm{B}}(0, n) \subset \mathsf{K}_n$.*

Consider for any $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}$ the Markov chain associated (4.19), where for any $x \in \mathbb{R}^d$, $\mathcal{T}_\gamma(x) = x + \gamma b_{\gamma,n}(x)$, $\mathsf{X} = \mathsf{K}_n$ and $\Pi = \Pi_{\mathsf{K}_n}$, the projection on K_n. The Markov kernel associated with this chain is denoted $\mathrm{R}_{\gamma,n}$ for any $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}$. Assuming only local Lipschitz regularity we can apply Theorem 4.1.8 to the projected version of the Markov chain associated with $\mathrm{R}_{T/m,n}$. Therefore we want to replace $\tilde{\mathrm{R}}_{T/m,n}$ by $\mathrm{R}_{T/m,n}$ in (4.61). In order to do so we consider the following assumption on the family of drifts $\{b_{\gamma,n} \,;\, \gamma \in (0, \bar{\gamma}] \,, n \in \mathbb{N}\}$.

L5. *There exist $\tilde{A} > 0$ and $\tilde{V} : \mathbb{R}^d \to [1, +\infty)$ such that for any $n \in \mathbb{N}$ there exist $\tilde{E}_n \geq 0$, $\tilde{\varepsilon}_n > 0$ and $\bar{\gamma}_n \in (0, \bar{\gamma}]$ satisfying for any $\gamma \in (0, \bar{\gamma}_n]$ and $x \in \mathbb{R}^d$,*

$$\tilde{\mathrm{R}}_{\gamma,n}\tilde{V}(x) \leq \exp\left[\log(\tilde{A})\gamma(1 + \tilde{E}_n\gamma^{\tilde{\varepsilon}_n})\right]\tilde{V}(x) \;, \qquad \sup_{x\in\mathbb{R}^d}\left\{\|x\|/\tilde{V}(x)\right\} \leq 1 \;,$$

where for any $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}$, $\tilde{\mathrm{R}}_{\gamma,n}$ is the Markov kernel associated with (4.19) where $\mathcal{T}_\gamma(x) = x + \gamma b_{\gamma,n}(x)$ and $\Pi = \mathrm{Id}$.

Proposition 4.1.27. *Let $V : \mathbb{R}^d \to [1, +\infty)$. Assume* **L1**, **L4**, **L5** *and that for any $T \geq 0$, $x \in \mathbb{R}^d$*

$$\limsup_{n\to+\infty} \limsup_{m\to+\infty} \left(\mathrm{R}^m_{T/m,n} + \tilde{\mathrm{R}}^m_{T/m,n}\right) V^2(x) < +\infty \;.$$

Then for any $T \geq 0$ and $x \in \mathbb{R}^d$

$$\lim_{n\to+\infty} \limsup_{m\to+\infty} \|\delta_x \mathrm{R}^{mk}_{T/m,n} - \delta_x \tilde{\mathrm{R}}^{mk}_{T/m,n}\|_V = 0 \;,$$

Proof. the proof is postponed to Section 4.1.7. □

Based on Proposition 4.1.26 and Proposition 4.1.27, we have the following result which establishes a clear link between the convergence of the family of the projected Euler-Maruyama scheme $\{\mathrm{R}_{\gamma,n} : \gamma \in (0, \bar{\gamma}] \,, n \in \mathbb{N}\}$ and the semigroup $(\mathrm{P}_t)_{t\geq 0}$ associated with (4.55).

Theorem 4.1.28. *Let $W : \mathbb{R}^d \times \mathbb{R}^d \to [1, +\infty)$ and $V : \mathbb{R}^d \to [1, +\infty)$ satisfying for any $x, y \in \mathbb{R}^d$, $\sup_{(x,y)\in\mathbb{R}^d\times\mathbb{R}^d} W(x,y)\{V(x)+V(y)\}^{-1} < +\infty$. Assume* **L**1, **L**2, **L**3, **L**4 *and* **L**5. *In addition, assume that for any $T \geq 0$ and $x \in \mathbb{R}^d$*

$$\mathrm{P}_T V^2(x) < +\infty \;, \quad \limsup_{n\to+\infty} \limsup_{m\to+\infty} \left(\mathrm{R}^m_{T/m,n} + \tilde{\mathrm{R}}^m_{T/m,n}\right) V^2(x) < +\infty \;. \tag{4.62}$$

Then,

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T) \leq \limsup_{n\to+\infty} \limsup_{m\to+\infty} \mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}^m_{T/m,n}, \delta_y \mathrm{R}^m_{T/m,n}) \;,$$

where for any $x, y \in \mathbb{R}^d$, $\mathbf{c}(x,y) = \mathbb{1}_{\Delta^c_{\mathsf{X}}}(x,y) W(x,y)$, $(\mathrm{P}_t)_{t\geq 0}$ is the semigroup associated with (4.55) *and for any $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$, $\mathrm{R}_{\gamma,n}$ is the Markov kernel associated with* (4.19) *where $\mathcal{T}_\gamma(x) = x + \gamma b_{\gamma,n}(x)$, $\mathsf{X} = \mathsf{K}_n$ and $\Pi = \Pi_{\mathsf{K}_n}$, $\tilde{\mathrm{R}}_{\gamma,n}$ is the Markov kernel associated with* (4.19) *where $\mathcal{T}_\gamma(x) = x + \gamma b_{\gamma,n}(x)$, $\mathsf{X} = \mathbb{R}^d$ and $\Pi = \mathrm{Id}$.*

Proof. Let $T \geq 0$, $x, y \in \mathbb{R}^d$ and

$$C_V = 2 \sup_{(x,y)\in\mathbb{R}^d\times\mathbb{R}^d} W(x,y)\{V(x)+V(y)\}^{-1} < +\infty \;.$$

We have for any $n \in \mathbb{N}$ and $m \in \mathbb{N}^\star$ such that $T/m \leq \bar{\gamma}$

$$\begin{aligned}\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{P}_T, \delta_y \mathrm{P}_T) \leq\; & C_V \|\delta_x \mathrm{P}_T - \delta_x \tilde{\mathrm{R}}^m_{T/m,n}\|_V + C_V \|\delta_x \mathrm{R}^m_{T/m,n} - \delta_x \tilde{\mathrm{R}}^m_{T/m,n}\|_V \\ & + \mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}^m_{T/m,n}, \delta_y \mathrm{R}^m_{T/m,n}) \\ & + C_V \|\delta_y \mathrm{P}_T - \delta_y \tilde{\mathrm{R}}^m_{T/m,n}\|_V + C_V \|\delta_y \mathrm{R}^m_{T/m,n} - \delta_y \tilde{\mathrm{R}}^m_{T/m,n}\|_V \;,\end{aligned}$$

which concludes the proof upon combining Proposition 4.1.26 and Proposition 4.1.27. □

Explicit approximating family of drifts In this section we show that under regularity and curvature assumptions on the drift function b we can construct explicit families of approximating drift functions satisfying the assumptions of Theorem 4.1.28. The section is divided into two parts. First, we show under regularity conditions **L**1, **L**2, **L**3, **L**4 and **L**5 are satisfied. Second, we show, under similar conditions, that the summability assumptions (4.62) in Theorem 4.1.28 hold for $V \leftarrow V_M$ with $V_M : \mathbb{R}^d \to [1, +\infty)$ given by (4.56) for $M \geq 0$. We start with the case where b satisfies **B**2.

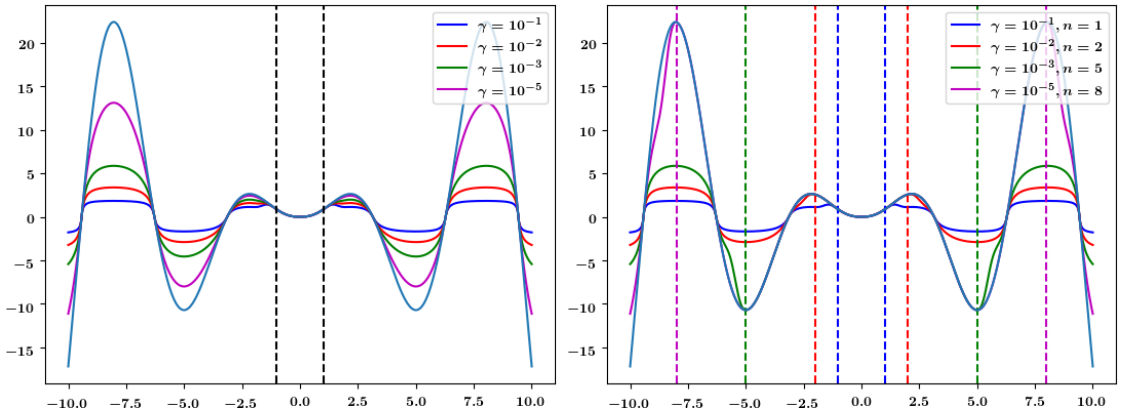
Proposition 4.1.29. *Assume* **B**2. *Let $\{\mathrm{b}_{\gamma,n} : \gamma \in (0, \bar{\gamma}], n \in \mathbb{N}\}$ be given for any $\gamma > 0$, $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$ by $\mathrm{b}_{\gamma,n}(x) = b(x)$. Let $\mathsf{K}_n = \mathbb{R}^d$ for any $n \in \mathbb{N}$. Then,* **L**1, **L**2, **L**3, **L**4 *and* **L**5 *are satisfied.*

Proof. The proof is postponed to Section 4.1.7. □

We now consider the more challenging case where **B**2 does not hold and is replaced by the weaker condition **B**5. In this setting, by [IW89, Chapter 4, Theorem 2.3], (4.55) admits a unique solution $(\mathbf{X}_t)_{t\in[0,+\infty)}$ with $\mathbf{X}_0 = x \in \mathbb{R}^d$ and let $e = \inf\{s \geq 0 : \|\mathbf{X}_s\| = +\infty\}$. In particular, the condition $e = +\infty$ is met a.s. if we assume that b is sub-linear [IW89, Chapter 4, Theorem 2.3] or that the condition $\mathbf{D}_{\mathrm{c}}(V, \zeta, 0)$ holds with $\zeta \in \mathbb{R}$ and $\lim_{\|x\|\to+\infty} V(x) = +\infty$ [Kha11, Theorem 3.5]. This last condition is satisfied for all the applications we consider in Section 4.1.5.

Proposition 4.1.30. *Assume* **B**3(m) *with $\mathtt{m} \in \mathbb{R}$ and* **B**5, *then* **L**1 *holds. In addition:*

(a) if there exists $\varepsilon_b > 0$ and $p \in \mathbb{N}^\star$ such that $\sup_{x\in\mathbb{R}^d}\{\|b(x)\|^{2(1+\varepsilon_b)}(1+\|x\|^{2p})^{-1}\} < +\infty$ then **L**3 *holds ;*



(a) (b)

Figure 4.3: In this figure we illustrate the approximation properties of the family of drift functions defined by (4.63). Let $b(x) = |x|^{1.5}\sin(x)$ and for any $n \in \mathbb{N}$, $\varphi_n(x) = d(x, \bar{\mathrm{B}}(0,+1)^{\mathrm{c}})^2/(d(x, \bar{\mathrm{B}}(0,n))^2 + d(x, \bar{\mathrm{B}}(0,n+1)^{\mathrm{c}})^2)$. In both figures the original drift is displayed in cyan and we fix $\alpha = 0.3$. In (a), we fix $n = 1$, represented by the black dashed lines, and observe the behavior of the drift functions for different values of $\gamma > 0$. In (b), we plot the drift for different $\gamma > 0$ and $n \in \mathbb{N}$.

(b) assume that **C**2 *holds and* $\sup_{x\in\mathbb{R}^d}\{\|b(x)\|^{2(1+\varepsilon_b)}\,\mathrm{e}^{-\mathtt{m}_2^+\|x\|^2}\} < +\infty$ *for some* $\varepsilon_b > 0$ *satisfying then* **L**3 *holds.*

Proof. The proof is postponed to Section 4.1.7. □

Proposition 4.1.30 gives conditions under which **L**1 and **L**3 hold. In addition, **L**4 is satisfied if we take for any $n \in \mathbb{N}$, $\mathsf{K}_n = \bar{\mathrm{B}}(0,n)$. Therefore, it only remains to find a family of drift functions which satisfies **L**2 and **L**5. To this end, consider the following family of drift functions $\{\mathrm{b}_{\gamma,n} : \gamma \in (0,\bar{\gamma}], n \in \mathbb{N}\}$ defined for any $\gamma > 0$, $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$ by

$$\mathrm{b}_{\gamma,n}(x) = \varphi_n(x)b(x) + (1-\varphi_n(x))\frac{b(x)}{1+\gamma^\alpha\|b(x)\|}\,, \tag{4.63}$$

with $\alpha < 1/2$ and $\varphi_n \in \mathrm{C}(\mathbb{R}^d,\mathbb{R})$ such that for any $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$,

$$\varphi_n(x) \in [0,1] \quad \text{and} \quad \varphi_n(x) = \begin{cases} 1 & \text{if } x \in \bar{\mathrm{B}}(0,n), \\ 0 & \text{if } x \in \bar{\mathrm{B}}(0,n+1)^{\mathrm{c}} \end{cases}. \tag{4.64}$$

An example of such a family is displayed in Figure 4.3.

Proposition 4.1.31. *Assume* **B**3(𝚖) *for* $\mathtt{m} \in \mathbb{R}$ *and* **B**5*, then* **L**2 *and* **L**5 *hold for the family* $\{\mathrm{b}_{\gamma,n} : \gamma \in (0,\bar{\gamma}], n \in \mathbb{N}\}$ *defined by* (4.63).

Proof. The proof is postponed to Section 4.1.7. □

The following proposition is a generalization of Proposition 4.1.29.

Proposition 4.1.32. *Assume* **B**3(m) *with* $\mathtt{m} \in \mathbb{R}$ *and* **B**5. *Let* $\{\mathrm{b}_{\gamma,n} : \gamma \in (0, \bar{\gamma}], n \in \mathbb{N}\}$ *be given for any* $\gamma > 0$, $n \in \mathbb{N}$ *and* $x \in \mathbb{R}^d$ *by* (4.63). *Let* $\mathsf{K}_n = \bar{\mathrm{B}}(0, n)$ *for any* $n \in \mathbb{N}$. *Then,* **L**1, **L**2, **L**4 *and* **L**5 *are satisfied.*

Proof. The proof is a straightforward combination of Proposition 4.1.30 and Proposition 4.1.31. □

In Proposition 4.1.33 and Proposition 4.1.34 we show that the second part of (4.62) holds under regularity assumptions on the drift function b.

Proposition 4.1.33. *Assume* **B**3(m) *for* $\mathtt{m} \in \mathbb{R}$ *and* **B**2, *then for any* $T, M \geq 0$ *and* $x \in \mathbb{R}^d$

$$\limsup_{m \to +\infty} \mathrm{R}^m_{T/m} V_M(x) < +\infty \ ,$$

with V_M *given in* (4.56) *and where for any* $\gamma \in (0, \bar{\gamma}]$, R_γ *is the Markov kernel associated with* (4.19) *where* $\mathcal{T}_\gamma(x) = x + \gamma b(x)$, $\mathsf{X} = \mathbb{R}^d$ *and* $\Pi = \mathrm{Id}$.

Proof. The proof is postponed to Section 4.1.7. □

Proposition 4.1.34. *Assume* **B**3(m) *for* $\mathtt{m} \in \mathbb{R}$ *and* **B**5, *then for any* $T, M \geq 0$ *and* $x \in \mathbb{R}^d$

$$\limsup_{n \to +\infty} \limsup_{m \to +\infty} \left(\mathrm{R}^m_{T/m,n} + \tilde{\mathrm{R}}^m_{T/m,n} \right) V_M(x) < +\infty \ ,$$

with V_M *given in* (4.56) *and where for any* $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$, $\mathrm{R}_{\gamma,n}$ *is the Markov kernel associated with* (4.19) *where* $\mathcal{T}_\gamma(x) = x + \gamma \mathrm{b}_{\gamma,n}(x)$, $\mathsf{X} = \bar{\mathrm{B}}(0, n)$ *and* $\Pi = \Pi_{\bar{\mathrm{B}}(0,n)}$, $\tilde{\mathrm{R}}_{\gamma,n}$ *is the Markov kernel associated with* (4.19) *where* $\mathcal{T}_\gamma(x) = x + \gamma \mathrm{b}_{\gamma,n}(x)$, $\mathsf{X} = \mathbb{R}^d$ *and* $\Pi = \mathrm{Id}$.

Proof. The proof is postponed to Section 4.1.7. □

Finally, we show that under mild curvature assumption on the drift function b, the first part of (4.62) holds.

Proposition 4.1.35. *Assume* **L**1 *and that* $\sup_{x \in \mathbb{R}^d} \langle b(x), x \rangle < +\infty$. *Then for any* $M \geq 0$, *there exists* $\zeta \in \mathbb{R}$ *such that* $\mathbf{D}_{\mathrm{c}}(V_M, \zeta, 0)$ *holds with* V_M *given in* (4.56). *In particular, for any* $T, M \geq 0$, $\mathrm{P}_T V_M(x) < +\infty$.

Proof. The proof is postponed to Section 4.1.7. □

4.1.6 Quantitative bounds for geometric convergence of Markov chains in Wasserstein distance

In this section, we establish new quantitative bounds for Markov chains in Wasserstein distance. We consider a Markov kernel P on the measurable space $(\mathsf{Y}, \mathcal{Y})$ equipped with the bounded semi-metric $\mathbf{d} : \mathsf{Y} \times \mathsf{Y} \to \mathbb{R}_+$, *i.e.* which satisfies the following condition.

H1. *For any* $x, y \in \mathsf{Y}$, $\mathbf{d}(x, y) \leq 1$, $\mathbf{d}(x, y) = \mathbf{d}(y, x)$ *and* $\mathbf{d}(x, y) = 0$ *if and only if* $x = y$.

Let K be a Markov coupling kernel for P. In this section, we assume the following condition on K.

H2 (K). *There exists* $\mathsf{C} \in \mathcal{Y}^{\otimes 2}$ *such that*

(i) there exist $\mathtt{n}_0 \in \mathbb{N}^\star$ *and* $\varepsilon > 0$ *such that for any* $x, y \in \mathsf{C}$, $\mathrm{K}^{\mathtt{n}_0}\mathbf{d}(x,y) \leq (1-\varepsilon)\mathbf{d}(x,y)$;

(ii) for any $x, y \in \mathsf{Y}$, $\mathrm{K}\mathbf{d}(x,y) \leq \mathbf{d}(x,y)$;

(iii) there exist $W_1 : \mathsf{Y}^2 \to [1, +\infty)$ *measurable,* $\lambda_1 \in (0,1)$ *and* $A_1 \geq 0$ *such that* K *satisfies* $\mathbf{D}_{\mathrm{d}}(W_1, \lambda_1, A_1, \mathsf{C})$.

We consider the Markov chain $(X_n, Y_n)_{n\in\mathbb{N}}$ associated with the Markov kernel K defined on the canonical space $((\mathsf{Y}\times\mathsf{Y})^{\mathbb{N}}, (\mathcal{Y}^{\otimes 2})^{\mathbb{N}})$ and denote by $\mathbb{P}_{(x,y)}$ and $\mathbb{E}_{(x,y)}$ the corresponding probability distribution and expectation respectively when $(X_0, Y_0) = (x, y)$. Denote by $(\mathcal{G}_n)_{n\in\mathbb{N}}$ the canonical filtration associated with $(X_n, Y_n)_{n\in\mathbb{N}}$. Note that for any $n \in \mathbb{N}$ and $x, y \in \mathsf{Y}$, under $\mathbb{P}_{(x,y)}$, (X_n, Y_n) is by definition a coupling of $\delta_x \mathrm{P}^n$ and $\delta_y \mathrm{P}^n$. The main result of this section is the following which by the previous observation implies quantitative bounds on $\mathbf{W}_{\mathbf{d}}(\delta_x\mathrm{P}^n, \delta_y\mathrm{P}^n)$.

Theorem 4.1.36. *Let* K *be a Markov coupling kernel for* P *and assume* **H**1 *and* **H**2(K). *Then for any* $n \in \mathbb{N}$ *and* $x, y \in \mathsf{Y}$,

$$\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\right] \leq \min\left[\rho^n(M_{\mathsf{C},\mathtt{n}_0}\boldsymbol{\Xi}(x,y,\mathtt{n}_0) + \mathbf{d}(x,y)), \rho^{n/2}(1+\mathbf{d}(x,y)) + \lambda_1^{n/2}\boldsymbol{\Xi}(x,y,\mathtt{n}_0)\right] ,$$

where

$$\begin{aligned} \boldsymbol{\Xi}(x,y,\mathtt{n}_0) &= W_1(x,y) + A_1\lambda_1^{-\mathtt{n}_0}\mathtt{n}_0 \\ \log(\rho) &= \log(1-\varepsilon)\log(\lambda_1)/[-\log(M_{\mathsf{C},\mathtt{n}_0}) + \log(1-\varepsilon)] \ , \\ M_{\mathsf{C},\mathtt{n}_0} &= \sup_{(x,y)\in\mathsf{C}} \boldsymbol{\Xi}(x,y,\mathtt{n}_0) = \sup_{(x,y)\in\mathsf{C}} [W_1(x,y)] + A_1\lambda_1^{-\mathtt{n}_0}\mathtt{n}_0 \ . \end{aligned} \tag{4.65}$$

In Theorem 4.1.36, we obtain geometric contraction for P in bounded Wasserstein metric $\mathbf{W}_{\mathbf{d}}$ since $\mathbf{d}$ is assumed to be bounded. To obtain convergence associated with unbounded Wasserstein metric associated with $W_2 : \mathsf{Y}^2 \to [0, +\infty)$, we consider the next assumption which is a generalized drift condition linking W_2 and the bounded semi-metric $\mathbf{d}$.

H3 (K). *There exist* $W_2 : \mathsf{Y}^2 \to [0, +\infty)$ *measurable,* $\lambda_2 \in (0,1)$ *and* $A_2 \geq 0$ *such that for any* $x, y \in \mathsf{Y}$,

$$\mathrm{K}W_2(x,y) \leq \lambda_2 W_2(x,y) + A_2\mathbf{d}(x,y) \ .$$

In the special case where $\mathbf{d}(x,y) = \mathbb{1}_{\Delta_{\mathsf{Y}}^{\mathrm{c}}}(x,y)$, $W_2(x,y) = \mathbb{1}_{\Delta_{\mathsf{Y}}^{\mathrm{c}}}(x,y)W_1(x,y)$ and for any $x \in \mathsf{Y}$, $\mathrm{K}((x,x), \Delta_{\mathsf{Y}}) = 1$, we obtain that $\mathbf{D}_{\mathrm{d}}(W_1, \lambda_1, A_1, \mathsf{Y})$ implies **H**3(K). The following result implies quantitative bounds on the Wasserstein distance $\mathbf{d}_{W_2}(\delta_x\mathrm{P}^n, \delta_y\mathrm{P}^n)$ for any $x, y \in \mathsf{Y}$ and $n \in \mathbb{N}^\star$.

Theorem 4.1.37. *Let* K *be a Markov coupling kernel for* P *and assume* **H**1, **H**2(K) *and* **H**3(K). *Then for any* $n \in \mathbb{N}$ *and* $x, y \in \mathsf{Y}$,

$$\begin{aligned} \mathbb{E}_{(x,y)}\left[W_2(X_n, Y_n)\right] &\leq \lambda_2^n W_2(x,y) \\ &+ A_2 \min\left[\tilde{\rho}^{n/4} r_\rho(\mathbf{d}(x,y) + \boldsymbol{\Xi}(x,y,\mathtt{n}_0)), \tilde{\rho}^{n/4} r_\rho(1+\mathbf{d}(x,y)) + \tilde{\lambda}^{n/4} r_\lambda \boldsymbol{\Xi}(x,y,\mathtt{n}_0)\right] , \end{aligned}$$

where

$$\begin{aligned} \tilde{\rho} &= \max(\lambda_2, \rho) \in (0,1) \ , & \tilde{\lambda} &= \max(\lambda_1, \lambda_2) \in (0,1) \ , \\ r_\rho &= 4\log^{-1}(1/\tilde{\rho})/\tilde{\rho} \ , & r_\lambda &= 4\log^{-1}(1/\tilde{\lambda})/\tilde{\lambda} \ , \end{aligned}$$

and $\boldsymbol{\Xi}(x,y,\mathtt{n}_0)$, $M_{\mathsf{C},\mathtt{n}_0}$ *and* ρ *are given in* (4.65).

Theorem 4.1.36 and Theorem 4.1.37 share some connections with [Ros95, Theorem 5], [HMS11] and [DM15] but hold under milder assumptions than the ones considered in these works. Compared to [HMS11] and [DM15], the main difference is that we allow here only a contraction for the $\mathtt{n}_0$-th iterate of the Markov chain (condition **H**2-(i)) which is necessary if we want to use Theorem 4.1.4 to obtain sharp quantitative convergence bounds for (4.19). Finally, [Ros95, Theorem 5] also considers minorization condition for the the $\mathtt{n}_0$-th iterate, however our results compared favourably for large $\mathtt{n}_0$. Indeed, Theorem 4.1.36 implies that the rate of convergence $\min(\rho, \lambda_1)$ is of the form $C\mathtt{n}_0^{-1}$ for $C \geq 0$ independent of $\mathtt{n}_0$. Applying [Ros95, Theorem 5], we found a rate of convergence of the form $C\mathtt{n}_0^{-2}$. Finally, a recent work [QH19] has established new results based on the technique used in [HMS11]. However, we were not able to apply them since they assume as in [HMS11], a contraction for $\mathtt{n}_0 = 1$ which does not imply sharp bounds on the situations we consider.

The rest of this section is devoted to the proof of Theorem 4.1.36 and Theorem 4.1.37. Denote by $\theta : (\mathsf{Y} \times \mathsf{Y})^{\mathbb{N}} \to (\mathsf{Y} \times \mathsf{Y})^{\mathbb{N}}$ the shift operator defined for any $(x_n, y_n)_{n\in\mathbb{N}} \in (\mathsf{Y} \times \mathsf{Y})^{\mathbb{N}}$ by $\theta((x_n, y_n)_{n\in\mathbb{N}}) = (x_{n+1}, y_{n+1})_{n\in\mathbb{N}}$. Define by induction, for any $m \in \mathbb{N}$, the sequence of $(\mathcal{G}_n)_{n\in\mathbb{N}}$-stopping times $(T^{(m)}_{\mathsf{C},\mathtt{n}_0})_{m\in\mathbb{N}}$, with $T^{(0)}_{\mathsf{C},\mathtt{n}_0} = 0$ and for any $m \in \mathbb{N}^\star$

$$\begin{aligned} T^{(m)}_{\mathsf{C},\mathtt{n}_0} &= \inf\left\{k \geq T^{(m-1)}_{\mathsf{C},\mathtt{n}_0} + \mathtt{n}_0 \,:\, (X_k, Y_k) \in \mathsf{C}\right\} \\ &= T^{(m-1)}_{\mathsf{C},\mathtt{n}_0} + \mathtt{n}_0 + \tilde{T}_{\mathsf{C}} \circ \theta^{T^{(m-1)}_{\mathsf{C}}+\mathtt{n}_0} = T^{(1)}_{\mathsf{C},\mathtt{n}_0} + \sum_{i=1}^{(m-1)} T^{(1)}_{\mathsf{C},\mathtt{n}_0} \circ \theta^{T^{(i)}_{\mathsf{C},\mathtt{n}_0}} , \qquad (4.66) \\ \tilde{T}_{\mathsf{C}} &= \inf\{k \geq 0 \,:\, (X_k, Y_k) \in \mathsf{C}\} . \end{aligned}$$

Note that $(T^{(m)}_{\mathsf{C},\mathtt{n}_0})_{m\in\mathbb{N}^\star}$ are the successive return times to C delayed by $\mathtt{n}_0 - 1$ and $\tilde{T}_{\mathsf{C}}$ is the first hitting time to C. We will use the following lemma which borrows from [DFM16] and [JT01, Lemma 3.1].

Lemma 4.1.38 ([DFM16, Proposition 14]). *Let K be a Markov coupling kernel for P and assume* **H**2(K)-(i)-(ii). *Then for any $n, m \in \mathbb{N}$, $x, y \in \mathsf{Y}$,*

$$\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\right] \leq (1-\varepsilon)^m \mathbf{d}(x, y) + \mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\mathbb{1}_{[n,+\infty]}(T^{(m)}_{\mathsf{C},\mathtt{n}_0})\right] .$$

Proof. Using **H**2(K)-(ii), we have that $(\mathbf{d}(X_n, Y_n))_{n\in\mathbb{N}}$ is a $(\mathcal{G}_n)_{n\in\mathbb{N}}$-supermartingale and therefore using the strong Markov property and **H**2(K)-(i) we obtain for any $m \in \mathbb{N}$ and $x, y \in \mathsf{Y}$ that

$$\begin{aligned} \mathbb{E}_{(x,y)}\left[\mathbf{d}(X_{T^{(m+1)}_{\mathsf{C},\mathtt{n}_0}}, Y_{T^{(m+1)}_{\mathsf{C},\mathtt{n}_0}})\right] &\leq \mathbb{E}_{(x,y)}\left[\mathbb{E}\left[\mathbf{d}(X_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}+\mathtt{n}_0}, Y_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}+\mathtt{n}_0})\middle|\mathcal{G}_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}}\right]\right] \\ &\leq (1-\varepsilon)\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}}, Y_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}})\right] . \qquad (4.67) \end{aligned}$$

Therefore by recursion and using (4.67) we obtain that for any $m \in \mathbb{N}$ and $x, y \in \mathsf{Y}$

$$\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}}, Y_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}})\right] \leq (1-\varepsilon)^m \mathbf{d}(x, y) . \qquad (4.68)$$

For any $n, m \in \mathbb{N}$ we have using (4.68) and that $(\mathbf{d}(X_n, Y_n))_{n\in\mathbb{N}}$ is a supermartingale,

$$\begin{aligned} &\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\right] \leq \mathbb{E}_{(x,y)}\left[\mathbf{d}(X_{n\wedge T^{(m)}_{\mathsf{C},\mathtt{n}_0}}, Y_{n\wedge T^{(m)}_{\mathsf{C},\mathtt{n}_0}})\right] \\ &\leq \mathbb{E}_{(x,y)}\left[\mathbf{d}(X_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}}, Y_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}})\mathbb{1}_{[0,n]}(T^{(m)}_{\mathsf{C},\mathtt{n}_0})\right] + \mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\mathbb{1}_{[n,+\infty]}(T^{(m)}_{\mathsf{C},\mathtt{n}_0})\right] \end{aligned}$$

$$\leq (1-\varepsilon)^m \mathbf{d}(x,y) + \mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\mathbb{1}_{[n,+\infty]}(T^{(m)}_{\mathsf{C},\mathtt{n}_0})\right] .$$

□

By Lemma 4.1.38 and since $\mathbf{d}$ is bounded by 1, we need to obtain a bound on $\mathbb{P}_{(x,y)}(T^{(m)}_{\mathsf{C},\mathtt{n}_0} \geq n)$ for $x, y \in \mathsf{Y}$ and $m, n \in \mathbb{N}^\star$. To this end, we will use the following proposition which gives an upper bound on exponential moment of the hitting times $(T^{(m)}_{\mathsf{C},\mathtt{n}_0})_{m\in\mathbb{N}^\star}$.

Lemma 4.1.39. *Let* K *be a Markov coupling kernel for* P *and assume* **H**2(K)-(iii). *Then for any* $x, y \in \mathsf{Y}$ *and* $m \in \mathbb{N}^\star$,

$$\mathbb{E}_{(x,y)}\left[\lambda_1^{-T^{(1)}_{\mathsf{C},\mathtt{n}_0}}\right] \leq \boldsymbol{\Xi}(x,y,\mathtt{n}_0) , \quad \mathbb{E}_{(x,y)}\left[\lambda_1^{-T^{(m)}_{\mathsf{C},\mathtt{n}_0}+T^{(1)}_{\mathsf{C},\mathtt{n}_0}}\right] \leq M^{m-1}_{\mathsf{C},\mathtt{n}_0} ,$$
$$\mathbb{E}_{(x,y)}\left[\lambda_1^{-T^{(m)}_{\mathsf{C},\mathtt{n}_0}}\right] \leq \boldsymbol{\Xi}(x,y,\mathtt{n}_0) M^{m-1}_{\mathsf{C},\mathtt{n}_0} ,$$

where $\boldsymbol{\Xi}(x,y,\mathtt{n}_0)$ *and* $M_{\mathsf{C},\mathtt{n}_0}$ *are defined in* (4.65).

Proof. We first show that for any $x, y \in \mathsf{Y}$ we have that $\mathbb{P}_{(x,y)}(\tilde{T}_{\mathsf{C}}) < +\infty$. Let $x, y \in \mathsf{Y}$. For any $n \in \mathbb{N}$ we have using **H**2(K)-(iii) and the Markov property

$$\mathbb{E}_{(x,y)}\left[W_1(X_{n+1}, Y_{n+1})|\mathcal{G}_n\right] \leq \lambda_1 W_1(X_n, Y_n) + A_1 \mathbb{1}_{\mathsf{C}}(X_n, Y_n) .$$

Therefore applying the comparison theorem [Dou+18, Theorem 4.3.1] we get that

$$(1-\lambda_1)\mathbb{E}_{(x,y)}\left[\sum_{k=0}^{\tilde{T}_{\mathsf{C}}-1} W_1(X_k, Y_k)\right] + \mathbb{E}_{(x,y)}\left[\mathbb{1}_{[0,+\infty)}(\tilde{T}_{\mathsf{C}}) W(X_{\tilde{T}_{\mathsf{C}}}, Y_{\tilde{T}_{\mathsf{C}}})\right] \leq W(x,y) .$$

Since for any $\tilde{x}, \tilde{y} \in \mathsf{Y}$, $1 \leq W_1(\tilde{x}, \tilde{y}) < +\infty$ we obtain that $(1-\lambda_1)\mathbb{E}_{(x,y)}[\tilde{T}_{\mathsf{C}}] \leq W(x,y)$ which implies $\mathbb{P}_{(x,y)}(\tilde{T}_{\mathsf{C}}) < +\infty$ since $\lambda_1 \in (0,1)$. We now show the stated result. Let $x, y \in \mathsf{Y}$ and $(\mathrm{S}_n)_{n\in\mathbb{N}}$ be defined for any $n \in \mathbb{N}$ by $\mathrm{S}_n = \lambda_1^{-n} W_1(X_n, Y_n)$. For any $n \in \mathbb{N}$ we have using **H**2(K)-(iii) and the Markov property

$$\begin{aligned}\mathbb{E}\left[\mathrm{S}_{n+1}|\mathcal{G}_n\right] &\leq \lambda_1^{-n} W_1(X_n, Y_n) + A_1 \lambda_1^{-(n+1)} \mathbb{1}_{\mathsf{C}}(X_n, Y_n) \\ &\leq \mathrm{S}_n + A_1 \lambda_1^{-(n+1)} \mathbb{1}_{\mathsf{C}}(X_n, Y_n) . \end{aligned} \tag{4.69}$$

Using the Markov property, the definition of $T^{(1)}_{\mathsf{C},\mathtt{n}_0}$ given in (4.66), the comparison theorem [Dou+18, Theorem 4.3.1], (4.69) and **H**2(K)-(iii) we obtain that

$$\begin{aligned}
&\mathbb{E}_{(x,y)}\left[\mathrm{S}_{T^{(1)}_{\mathsf{C},\mathtt{n}_0}}\right] = \mathbb{E}_{(x,y)}\left[\mathbb{E}_{(x,y)}\left[\mathrm{S}_{T^{(1)}_{\mathsf{C},\mathtt{n}_0}}\middle|\mathcal{G}_{\mathtt{n}_0}\right]\right] = \mathbb{E}_{(x,y)}\left[\mathbb{E}_{(x,y)}\left[\mathrm{S}_{\mathtt{n}_0+\tilde{T}_{\mathsf{C}}\circ\theta^{\mathtt{n}_0}}\middle|\mathcal{G}_{\mathtt{n}_0}\right]\right] \\
&= \mathbb{E}_{(x,y)}\left[\lambda_1^{-\mathtt{n}_0}\ \mathbb{E}_{(x,y)}\left[W_1(X_{\mathtt{n}_0+T^{(1)}_{\mathsf{C},\mathtt{n}_0}}, Y_{\mathtt{n}_0+T^{(1)}_{\mathsf{C},\mathtt{n}_0}})\lambda_1^{-T^{(1)}_{\mathsf{C},\mathtt{n}_0}}\middle|\mathcal{G}_{\mathtt{n}_0}\right]\right] \\
&\leq \mathbb{E}_{(x,y)}\left[\lambda_1^{-\mathtt{n}_0}\ \mathbb{E}_{(X_{\mathtt{n}_0},Y_{\mathtt{n}_0})}\left[\mathrm{S}_{\tilde{T}_{\mathsf{C}}}\right]\right] \\
&\leq \mathbb{E}_{(x,y)}\left[\lambda_1^{-\mathtt{n}_0}\ \mathbb{E}_{(X_{\mathtt{n}_0},Y_{\mathtt{n}_0})}\left[\mathrm{S}_{\tilde{T}_{\mathsf{C}}}\mathbb{1}_{[0,+\infty)}(\tilde{T}_{\mathsf{C}})\right]\right] \\
&\leq \mathbb{E}_{(x,y)}\left[\lambda_1^{-\mathtt{n}_0}\ \mathbb{E}_{(X_{\mathtt{n}_0},Y_{\mathtt{n}_0})}\left[\mathrm{S}_0 + A_1 \sum_{k=0}^{\tilde{T}_{\mathsf{C}}-1} \lambda_1^{-(k+1)} \mathbb{1}_{\mathsf{C}}(X_k, Y_k)\right]\right]
\end{aligned}$$

$$\leq \mathbb{E}_{(x,y)}\left[\lambda_1^{-\mathtt{n}_0} W_1(X_{\mathtt{n}_0}, Y_{\mathtt{n}_0})\right] \leq W_1(x,y) + A_1 \lambda_1^{-\mathtt{n}_0} \mathtt{n}_0 \ . \tag{4.70}$$

Combining (4.70) and the fact that for any $x, y \in \mathsf{Y}$, $W_1(x,y) \geq 1$, we obtain that

$$\mathbb{E}_{(x,y)}\left[\lambda_1^{-T^{(1)}_{\mathsf{C},\mathtt{n}_0}}\right] \leq W_1(x,y) + A_1 \lambda_1^{-\mathtt{n}_0} \mathtt{n}_0 \ . \tag{4.71}$$

We conclude by a straightforward recursion and using (4.71), the definition of $T^{(m)}_{\mathsf{C},\mathtt{n}_0}$ (4.66) for $m \geq 1$, the strong Markov property and the fact that for any $m \in \mathbb{N}^\star$, $(X_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}}, Y_{T^{(m)}_{\mathsf{C},\mathtt{n}_0}}) \in \mathsf{C}$. □

Proof of Theorem 4.1.36. Let $x, y \in \mathsf{Y}$ and $n \in \mathbb{N}$. By Lemma 4.1.38, Lemma 4.1.39, **H**1, the fact that $M_{\mathsf{C},\mathtt{n}_0} \geq 1$ and the Markov inequality, we have for any $m \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\right] &\leq (1-\varepsilon)^m \mathbf{d}(x,y) + \mathbb{P}_{(x,y)}\left[T^{(m)}_{\mathsf{C},\mathtt{n}_0} \geq n\right] \\
&\leq (1-\varepsilon)^m \mathbf{d}(x,y) + \lambda_1^n \, \mathbb{E}_{(x,y)}\left[\lambda_1^{-T^{(m)}_{\mathsf{C},\mathtt{n}_0}}\right] \\
&\leq (1-\varepsilon)^m \mathbf{d}(x,y) + \lambda_1^n M^m_{\mathsf{C},\mathtt{n}_0} \boldsymbol{\Xi}(x,y,\mathtt{n}_0) \ ,
\end{aligned}$$

where $\boldsymbol{\Xi}(x,y,\mathtt{n}_0)$ is given in Theorem 4.1.36. Combining this result and Lemma 4.1.39, we can conclude that $\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\right] \leq \rho^n (M_{\mathsf{C},\mathtt{n}_0} \boldsymbol{\Xi}(x,y,\mathtt{n}_0) + \mathbf{d}(x,y))$ setting

$$m = \lceil n \log(\lambda_1) / \{\log(1-\varepsilon) - \log(M_{\mathsf{C},\mathtt{n}_0})\} \rceil \ .$$

To show that $\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\right] \leq \rho^{n/2}(1 + \mathbf{d}(x,y)) + \lambda_1^{n/2} \boldsymbol{\Xi}(x,y,\mathtt{n}_0)$, first note that Lemma 4.1.38 and **H**1 imply that for any $m \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\right] \\
&\leq (1-\varepsilon)^m \mathbf{d}(x,y) + \mathbb{P}_{(x,y)}\left[T^{(m)}_{\mathsf{C},\mathtt{n}_0} - T^{(1)}_{\mathsf{C},\mathtt{n}_0} \geq n/2\right] + \mathbb{P}_{(x,y)}\left[T^{(1)}_{\mathsf{C},\mathtt{n}_0} \geq n/2\right] \\
&\leq (1-\varepsilon)^m \mathbf{d}(x,y) + \lambda_1^{n/2} \mathbb{E}_{(x,y)}\left[\lambda_1^{-T^{(m)}_{\mathsf{C},\mathtt{n}_0} + T^{(1)}_{\mathsf{C},\mathtt{n}_0}}\right] + \lambda_1^{n/2} \mathbb{E}_{(x,y)}\left[\lambda_1^{-T^{(1)}_{\mathsf{C},\mathtt{n}_0}}\right] \ ,
\end{aligned}$$

where we have used the Markov inequality in the last line. Combining this result and Lemma 4.1.39, we can conclude that $\mathbb{E}_{(x,y)}\left[\mathbf{d}(X_n, Y_n)\right] \leq \rho^{n/2}(1 + \mathbf{d}(x,y)) + \lambda_1^{n/2} \boldsymbol{\Xi}(x,y,\mathtt{n}_0)$ setting

$$m = \lceil n \log(\lambda_1) / \{2\log(1-\varepsilon) - 2\log(M_{\mathsf{C},\mathtt{n}_0})\} \rceil \ .$$

□

Proof of Theorem 4.1.37. Let $x, y \in \mathsf{Y}$ and $n \in \mathbb{N}$. Using **H**3(K), we obtain by recursion

$$\mathbb{E}_{(x,y)}\left[W_2(X_n, Y_n)\right] \leq \lambda_2^n W_2(x,y) + A_2 \sum_{k=0}^{n-1} \lambda_2^{n-1-k} \ \mathbb{E}_{(x,y)}\left[\mathbf{d}(X_k, Y_k)\right] \ . \tag{4.72}$$

Applying Theorem 4.1.36 we obtain

$$\sum_{k=0}^{n-1} \lambda_2^{n-1-k} \ \mathbb{E}_{(x,y)}\left[\mathbf{d}(X_k, Y_k)\right]$$

$$\leq \sum_{k=0}^{n-1} \lambda_2^{n-1-k} \min\left[\rho^k (M_{\mathsf{C},\mathfrak{n}_0}\Xi(x,y,\mathfrak{n}_0) + \mathbf{d}(x,y)), \rho^{k/2}(1+\mathbf{d}(x,y)) + \lambda^{k/2}\Xi(x,y,\mathfrak{n}_0)\right]$$
$$\leq \min\left[n\tilde{\rho}^{n-1}(\mathbf{d}(x,y) + \Xi(x,y,\mathfrak{n}_0)), n\tilde{\rho}^{n/2-1}(1+\mathbf{d}(x,y)) + n\tilde{\lambda}^{n/2-1}\Xi(x,y,\mathfrak{n}_0)\right] .$$

We conclude plugging this result in (4.72) and using that for any $n \in \mathbb{N}$ and $t \in (0,1)$, $nt^{n/2} \leq 4\log^{-1}(1/t)t^{n/4}$. □

4.1.7 Proofs and additional results

Proofs of Section 4.1.3

Proof of Proposition 4.1.5 First, we prove the following technical lemma.

Lemma 4.1.40. *Let $\bar{\gamma} > 0$ and $\kappa : (0,\bar{\gamma}] \to \mathbb{R}$, with $\kappa(\gamma)\gamma \in (-1,+\infty)$ for any $\gamma \in (0,\bar{\gamma}]$. We have for any $\gamma \in (0,\bar{\gamma}]$ such that $\kappa(\gamma) \neq 0$ and $\ell \in \mathbb{N}^\star$*

$$\Xi_{\ell\lceil 1/\gamma\rceil}(\kappa) = -\kappa^{-1}(\gamma)\left\{\exp\left[-\ell\left\lceil 1/\gamma\right\rceil \log\left\{1+\gamma\kappa(\gamma)\right\}\right] - 1\right\} ,$$

where $\Xi_{\ell\lceil 1/\gamma\rceil}(\kappa)$ is defined by (4.26) . In addition, for any $\ell \in \mathbb{N}^\star$ and $\gamma \in (0,\bar{\gamma}]$

(a) $\Xi_{\ell\lceil 1/\gamma\rceil}(\kappa) \geq \alpha_-(\kappa,\gamma,\ell) = -\kappa^{-1}(\gamma)\left[\exp(-\ell\kappa(\gamma)) - 1\right]$ if for any $\gamma \in (0,\bar{\gamma}]$, $\kappa(\gamma) < 0$;

(b) $\Xi_{\ell\lceil 1/\gamma\rceil}(\kappa) \geq \alpha_0(\kappa,\gamma,\ell) = \ell$ if for any $\gamma \in (0,\bar{\gamma}]$, $\kappa(\gamma) \leq 0$;

(c) $\Xi_{\ell\lceil 1/\gamma\rceil}(\kappa) \geq \alpha_+(\kappa,\gamma,\ell) = \kappa^{-1}(\gamma)\left[1 - \exp\left\{-\dfrac{\ell\kappa(\gamma)}{1+\gamma\kappa(\gamma)}\right\}\right]$ if for any $\gamma \in (0,\bar{\gamma}]$, $\kappa(\gamma) > 0$.

Proof. Let $\ell \in \mathbb{N}^\star$ and $\gamma \in (0,\bar{\gamma}]$. First note that the following equalities hold if $\kappa(\gamma) \neq 0$

$$\begin{aligned}
\Xi_{\ell\lceil 1/\gamma\rceil}(\kappa) &= \gamma \sum_{i=1}^{\ell\lceil 1/\gamma\rceil} (1+\gamma\kappa(\gamma))^{-i} \\
&= \gamma(1+\gamma\kappa(\gamma))^{-1}\frac{1-(1+\gamma\kappa(\gamma))^{-\ell\lceil 1/\gamma\rceil}}{1-(1+\gamma\kappa(\gamma))^{-1}} \\
&= -\kappa^{-1}(\gamma)\left\{\left[1+\gamma\kappa(\gamma)\right]^{-\ell\lceil 1/\gamma\rceil} - 1\right\} \\
&= -\kappa^{-1}(\gamma)\left\{\exp\left[-\ell\left\lceil 1/\gamma\right\rceil \log\left\{1+\gamma\kappa(\gamma)\right\}\right] - 1\right\} . \qquad (4.73)
\end{aligned}$$

We now give a lower-bound on $\Xi_{\ell\lceil 1/\gamma\rceil}(\kappa)$ depending on the condition satisfied by $\gamma \mapsto \kappa(\gamma)$.

(a) Assume that for any $\tilde{\gamma} \in (0,\bar{\gamma}]$, $\kappa(\tilde{\gamma}) < 0$. Using that $\log(1-t) \leq -t$ for $t \in (0,1)$, we obtain that

$$\exp\left[-\ell\left\lceil 1/\gamma\right\rceil \log\left\{1+\gamma\kappa(\gamma)\right\}\right] \geq \exp(-\ell\left\lceil 1/\gamma\right\rceil \gamma\kappa(\gamma)) \geq \exp(-\ell\kappa(\gamma)) ,$$

which together with (4.73) concludes the proof for Proposition 4.1.5-(a) .

(b) Assume that for any $\tilde{\gamma} \in (0,\bar{\gamma}]$, $\kappa(\tilde{\gamma}) \leq 0$. Then,

$$\Xi_{\ell\lceil 1/\gamma\rceil}(\kappa) = \gamma \sum_{i=1}^{\ell\lceil 1/\gamma\rceil} (1+\gamma\kappa(\gamma))^{-i} \geq \gamma\left\lceil 1/\gamma\right\rceil \ell \geq \ell .$$

(c) Assume that for any $\tilde{\gamma} \in (0, \bar{\gamma}]$, $\kappa(\tilde{\gamma}) > 0$. Using that $\log(1+t) \geq t/(1+t)$ for $t > 0$, we obtain that

$$\begin{aligned} \exp\left[-\ell \left\lceil 1/\gamma \right\rceil \log\left\{1 + \gamma\kappa(\gamma)\right\}\right] &\leq \exp\left[-(\ell/\gamma)\log\left\{1+\gamma\kappa(\gamma)\right\}\right] \\ &\leq \exp\left[-\ell\kappa(\gamma)/(1+\gamma\kappa(\gamma))\right] \ , \end{aligned}$$

which concludes the proof for Proposition 4.1.5-(a).

□

Proof of Proposition 4.1.5. The proof is a direct application of Theorem 4.1.4 and Lemma 4.1.40 with $\kappa(\gamma) = \kappa(\gamma)$. □

Proof of Corollary 4.1.6

(a) Consider $V : \mathsf{X} \to [1, +\infty]$ given for any $x \in \mathsf{X}$ by $V(x) = 1 + \|x\|$. Then since $\mathbf{A}2(\mathsf{X}^2)$ with $\sup_{\gamma \in (0,\bar{\gamma}]} \kappa(\gamma) \leq \kappa_- < 0$ holds, using the triangle inequality and the Cauchy-Schwarz inequality, we have for any $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathsf{X}$

$$\mathrm{R}_\gamma V(x) \leq \|\mathcal{T}_\gamma(x)\| + \sqrt{\gamma d} \leq (1 + \kappa_-\gamma)\|x\| + \|\mathcal{T}_\gamma(0)\| + \sqrt{\gamma d} + 1 \leq \lambda V(x) + A \ ,$$

with $\lambda \in (0,1)$ and $A \geq 0$. As a result, since for any $\gamma \in (0, \bar{\gamma}]$, R_γ is a Feller kernel and the level sets of V are compact, R_γ admits a unique invariant probability measure π_γ for any $\gamma \in (0, \bar{\gamma}]$ by [Dou+18, Theorem 12.3.3]. Then the last result is a straightforward consequence of Proposition 4.1.5-(a), (4.21) and the fact that for any $\ell \in \mathbb{N}^\star$ and $\gamma \in (0, \bar{\gamma}]$, $\alpha_-(\kappa, \gamma, \ell) \geq -(\exp(-\ell\kappa_-) - 1)/\kappa_-$ since $t \mapsto (\exp(\ell t) - 1)/t$ is increasing on $\mathbb{R}$.

(b) This result is a direct consequence of Proposition 4.1.5-(b), (4.21) and the fact that R_γ admits an invariant probability measure π_γ.

Proof of Corollary 4.1.7

(a) The proof is a direct application of Proposition 4.1.5-(b), the fact that $(X_k, Y_k) \in \mathsf{X}^2$ for any $k \in \mathbb{N}$ and that K_γ is the Markov kernel associated with $(X_k, Y_k)_{k \in \mathbb{N}}$.

(b) Consider the case where $\mathbf{A}2(\mathsf{X}^2)$-(iii) holds. Using that for any $t \geq 0$, $1 - \mathrm{e}^{-t} \geq t/(t+1)$ we obtain that for any $\gamma \in (0, \bar{\gamma}]$ and $\ell \in \mathbb{N}^\star$

$$\alpha_+(\kappa, \gamma, \ell) \geq \ell/(1 + (\ell + \bar{\gamma})\kappa(\gamma)) \geq (1 + (1+\bar{\gamma})\kappa_+)^{-1} \geq (1+\bar{\gamma})^{-1}(1+\kappa_+)^{-1} \ ,$$

where α_+ is given in Lemma 4.1.40-(c). Then, combining this result and Proposition 4.1.5-(c) complete the proof.

Proof of Theorem 4.1.8 We start with the following theorem.

Theorem 4.1.41. *Under the assumptions of Theorem 4.1.8, we have for any $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathsf{X}$ and $k \in \mathbb{N}$*

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \tilde{\mathrm{K}}_\gamma^k \mathbf{c}(x,y) \leq \lambda^{k\gamma/4}[D_{\gamma,1}\mathbf{c}(x,y) + D_{\gamma,2}\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y)] + \tilde{C}_\gamma \tilde{\rho}_\gamma^{k\gamma/4}\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y) \ , \quad (4.74)$$

where $\mathbf{W}_{\mathbf{c}}$ *is the Wasserstein metric associated with* $\mathbf{c}$ *defined by* (4.30),

$$
\begin{aligned}
D_{\gamma,1} &= 1 + 4A[\log(1/\lambda)\lambda^{\gamma}]^{-1} \;, \qquad D_{\gamma,2} = D_{\gamma,1}\left[A\lambda^{-\gamma\lceil 1/\gamma\rceil\ell}\gamma\,\lceil 1/\gamma\rceil\,\ell\right] \;, \\
\tilde{C}_{\gamma} &= 8A\log^{-1}(1/\tilde{\rho}_{\gamma})/\tilde{\rho}_{\gamma}^{\gamma} \;, \\
\log(\tilde{\rho}_{\gamma}) &= \{\log(\lambda)\log(1-\tilde{\varepsilon}_{\mathrm{d},\gamma})\}\,/\,\{-\log(\tilde{c}_{\gamma}) + \log(1-\tilde{\varepsilon}_{\mathrm{d},\gamma})\} \;, \\
\tilde{B}_{\mathrm{d}} &= \sup_{(x,y)\in\mathsf{C}} W(x,y) \;, \quad \tilde{c}_{\gamma} = \tilde{B}_{\mathrm{d}} + A\lambda^{-\gamma\lceil 1/\gamma\rceil\ell}\gamma\,\lceil 1/\gamma\rceil\,\ell \;, \\
\tilde{\varepsilon}_{\mathrm{d},\gamma} &= \inf_{(x,y)\in\Delta_{\mathsf{X},\tilde{M}_{\mathrm{d}}}} \mathbf{\Psi}(\gamma,\ell,\|x-y\|) \;.
\end{aligned}
$$

Proof. The proof of this proposition is an application of Theorem 4.1.37 in Section 4.1.6 with $\mathbf{d} \leftarrow \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}$ which satisfies **H**1. Let $\gamma \in (0,\bar{\gamma}]$. Then, since $\tilde{\mathrm{K}}_{\gamma}$ and $\mathbf{\Psi}$ satisfy $\mathbf{D}_{\mathrm{d}}(W,\lambda^{\gamma},A\gamma,\mathsf{C})$ and (4.32) respectively, and Δ_{X} is absorbing for $\tilde{\mathrm{K}}_{\gamma}$, **H**2($\mathrm{K}_{\gamma}$) and **H**3($\mathrm{K}_{\gamma}$) are satisfied. More precisely, for any $\gamma \in (0,\bar{\gamma}]$ setting $\tilde{\varepsilon}_{\mathrm{d},\gamma} = \inf_{(x,y)\in\Delta_{\mathsf{X},\tilde{M}_{\mathrm{d}}}} \mathbf{\Psi}(\gamma,\ell,\|x-y\|)$, then **H**2($\mathrm{K}_{\gamma}$)-(i) is satisfied since for any $x,y \in \mathsf{C} \subset \Delta_{\mathsf{X},\tilde{M}_{\mathrm{d}}}$,

$$
\begin{aligned}
\tilde{\mathrm{K}}_{\gamma}^{\lceil 1/\gamma\rceil\ell}\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y) &\le \left\{1 - \inf_{(x,y)\in\Delta_{\mathsf{X},\tilde{M}_{\mathrm{d}}}} \mathbf{\Psi}(\gamma,\ell,\|x-y\|)\right\}\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y) \\
&\le (1-\tilde{\varepsilon}_{\mathrm{d},\gamma})\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y) \;.
\end{aligned}
$$

H2(K_{γ})-(ii) is satisfied since for any $\gamma \in (0,\bar{\gamma}]$ and $x,y \in \mathsf{X}$, $\mathrm{K}_{\gamma}\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y) \le \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y)$. Finally, the conditions **H**2(K_{γ})-(iii) and **H**3(K_{γ}) hold using $\mathbf{D}_{\mathrm{d}}(W,\lambda^{\gamma},A\gamma,\mathsf{C})$ with $W_1 \leftarrow W$, $W_2 \leftarrow W\mathbf{d}$, $\lambda_1 = \lambda_2 \leftarrow \lambda^{\gamma}$, $A_1 = A_2 \leftarrow A\gamma$, $\mathfrak{n}_0 \leftarrow \ell\,\lceil 1/\gamma\rceil$. Applying Theorem 4.1.37, we obtain that for any $k \in \mathbb{N}$, $\gamma \in (0,\bar{\gamma}]$ and $x,y \in \mathsf{X}$

$$
\begin{aligned}
&\mathbf{W}_{\mathbf{c}}(\delta_x\mathrm{R}_{\gamma}^{k},\delta_y\mathrm{R}_{\gamma}^{k}) \\
&\le \lambda^{k\gamma}W(x,y) + A\gamma\left[\tilde{\rho}_{\gamma}^{k\gamma/4}r_1(1+\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y)) + \lambda^{k\gamma/4}r_2\mathbf{\Xi}(x,y,\ell\,\lceil 1/\gamma\rceil)\right] \\
&\le \lambda^{k\gamma/4}W(x,y) + 2r_1A\gamma\tilde{\rho}_{\gamma}^{k\gamma/4} + A\gamma r_2\lambda^{k\gamma/4}\mathbf{\Xi}(x,y,\ell\,\lceil 1/\gamma\rceil) \\
&\le \lambda^{k\gamma/4}(1+A\gamma r_2)\left[W(x,y) + A\gamma\lambda^{-\ell\lceil 1/\gamma\rceil\gamma}\ell\,\lceil 1/\gamma\rceil\,\gamma\right] + 2r_1A\gamma\tilde{\rho}_{\gamma}^{k\gamma/4} \;,
\end{aligned}
$$

where

$$
r_1 = 4\log^{-1}(1/\tilde{\rho}_{\gamma})/(\gamma\tilde{\rho}_{\gamma}^{\gamma}) \;, \quad r_2 = 4\log^{-1}(1/\lambda)/(\gamma\lambda^{\gamma}) \;.
$$

This concludes the proof of (4.74) upon using that Δ_{X} is absorbing for $\tilde{\mathrm{K}}_{\gamma}$. □

Proof of Theorem 4.1.8. The first part of the proof is straightforward using Theorem 4.1.41 and that $\lambda^{\gamma} \ge \lambda^{\bar{\gamma}}$.

By assumption on $\bar{\gamma}$ and λ, we have $\lambda^{-\gamma\lceil 1/\gamma\rceil\ell}\gamma\,\lceil 1/\gamma\rceil\,\ell \le \lambda^{-(1+\bar{\gamma})\ell}(1+\bar{\gamma})\ell$. As a result and using the fact that $\log(1-t) \le -t$ for any $t \in (0,1)$, $\log((1-\bar{\varepsilon}_{\mathrm{d},1})^{-1}) \le 1$ and $W(x,y) \ge 1$ for any $x,y \in \mathsf{X}$, we obtain that

$$
\begin{aligned}
\log^{-1}(\bar{\rho}_1^{-1}) &\le [\log(\lambda^{-1})\log((1-\bar{\varepsilon}_{\mathrm{d},\bar{\gamma}})^{-1})]^{-1}\,[1+\log(\bar{c}_2)] \\
&\le \left[\log(\lambda^{-1})\bar{\varepsilon}_{\mathrm{d},1}\right]^{-1}\left[1+\log(\tilde{B}_{\mathrm{d}}) + \log(1+2A\ell\lambda^{-2\ell})\right] \;, \\
&\le \left[\log(\lambda^{-1})\bar{\varepsilon}_{\mathrm{d},1}\right]^{-1}\left[1+\log(\tilde{B}_{\mathrm{d}}) + \log(1+2A\ell) + 2\ell\log(\lambda^{-1})\right] \;,
\end{aligned}
$$

which completes the proof. □

Proof of Proposition 4.1.9 Let $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathsf{X}$ and $k \in \mathbb{N}$. We divide the proof into two parts.

(a) If $k \leq \lceil 1/\gamma \rceil$. Then using we get that

$$\tilde{\mathrm{K}}_\gamma^k \left\| x - y \right\| \leq (1 + \gamma\varkappa)^k \left\| x - y \right\| \leq (1 + \gamma\varkappa)^{\lceil 1/\gamma \rceil} \left\| x - y \right\| \leq \exp[\varkappa(1 + \bar{\gamma})] \left\| x - y \right\| \ .$$

(b) If $k > \lceil 1/\gamma \rceil$ then using Theorem 4.1.8, (4.32) and $\rho_1 \geq \lambda$ we get that

$$\begin{aligned}
&\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \tilde{\mathrm{K}}_\gamma^{\lceil 1/\gamma \rceil} \mathrm{K}_\gamma^{k - \lceil 1/\gamma \rceil} \mathbf{c}(x, y) \\
&\leq \tilde{\mathrm{K}}_\gamma^{\lceil 1/\gamma \rceil} \left\{ \lambda^{(k - \lceil 1/\gamma \rceil)\gamma/4} [\bar{D}_1 W(x, y) + \bar{D}_2 \mathbb{1}_{\Delta^{\mathrm{c}}}(x, y)] + \bar{C}_2 \bar{\rho}_1^{(k - \lceil 1/\gamma \rceil)\gamma/4} \mathbb{1}_{\Delta^{\mathrm{c}}}(x, y) \right\} \\
&\leq \bar{\rho}_1^{(k - \lceil 1/\gamma \rceil)\gamma/4} \tilde{\mathrm{K}}_\gamma^{\lceil 1/\gamma \rceil} \left\{ (\bar{D}_1 + \bar{D}_2 + \bar{C}_1) \mathbb{1}_{\Delta^{\mathrm{c}}}(x, y) + \vartheta \bar{D}_1 \left\| x - y \right\| \right\} \\
&\leq \bar{\rho}_1^{k\gamma/4} \left\{ (\bar{D}_1 + \bar{D}_2 + \bar{C}_1)(1 - \mathbf{\Psi}(\gamma, 1, \left\| x - y \right\|)) + \vartheta \bar{D}_1 \exp[\varkappa(1 + \bar{\gamma})] \left\| x - y \right\| \right\} / \bar{\rho}_1^{(1+\bar{\gamma})/4} \\
&\leq -\mathbf{a}(\bar{D}_1 + \bar{D}_2 + \bar{C}_1) \bar{\rho}_1^{k\gamma/4} \left\| x - y \right\| / \bar{\rho}_1^{1/4} + \vartheta \bar{D}_1 \exp[\varkappa(1 + \bar{\gamma})] \bar{\rho}_1^{k\gamma/4} \left\| x - y \right\| / \bar{\rho}_1^{(1+\bar{\gamma})/4} \ ,
\end{aligned}$$

which concludes the proof upon noting that $\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \geq \vartheta \mathbf{W}_1(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k)$.

Proof of Proposition 4.1.10 Let $q \in \mathbb{N}$ and $\gamma \in (0, \bar{\gamma}]$. Using that $\tilde{\mathrm{K}}_\gamma$ satisfies $\mathbf{D}_{\mathrm{d}}((x, y) \mapsto \left\| x - y \right\|^q, \tilde{\lambda}_q^\gamma, \tilde{A}_q \gamma)$, we get that for any $x, y \in \mathsf{X}$ and $k \in \mathbb{N}$ we have

$$\tilde{\mathrm{K}}_\gamma^k \left\| x - y \right\|^q \leq \left\| x - y \right\|^q + \tilde{A}_q \gamma \sum_{\ell=0}^{k-1} \tilde{\lambda}_q^{\ell\gamma} \leq \left\| x - y \right\|^q + \tilde{A}_q \log^{-1}(1/\tilde{\lambda}_q) \tilde{\lambda}_q^{-\bar{\gamma}} \ . \tag{4.75}$$

Let $p \geq 1$, $\alpha \in (p, +\infty)$, $x, y \in \mathsf{X}$ and $k \in \mathbb{N}$ and consider $q = p(\alpha - 1)/(\alpha - p)$. Note that we have

$$(1 - 1/\alpha)p / \lceil q \rceil \leq (1 - 1/\alpha)p/q \leq 1 - p/\alpha \leq 1 \ .$$

Using this result, (4.75), Hölder's inequality, Jensen's inequality and that for any $a, b \geq 0$ and $r \geq 1$, $(a + b)^{1/r} \leq a^{1/r} + b^{1/r}$, we have

$$\begin{aligned}
\tilde{\mathrm{K}}_\gamma^k \left\| x - y \right\|^p &\leq \tilde{\mathrm{K}}_\gamma^k \left\{ \left\| x - y \right\|^{p(1-1/\alpha)} \left\| x - y \right\|^{p/\alpha} \right\} \\
&\leq \left(\tilde{\mathrm{K}}_\gamma^k \left\| x - y \right\|^{p(1-1/\alpha)/(1-p/\alpha)} \right)^{1-p/\alpha} \left(\tilde{\mathrm{K}}_\gamma^k \left\| x - y \right\| \right)^{p/\alpha} \\
&\leq \left(\tilde{\mathrm{K}}_\gamma^k \left\| x - y \right\|^q \right)^{1-p/\alpha} \bar{D}^{p/\alpha} \bar{\rho}^{k\gamma p/\alpha} \left\| x - y \right\|^{p/\alpha} \\
&\leq \left(\tilde{\mathrm{K}}_\gamma^k \left\| x - y \right\|^{\lceil q \rceil} \right)^{(1-p/\alpha)q/\lceil q \rceil} \bar{D}^{p/\alpha} \bar{\rho}^{k\gamma p/\alpha} \left\| x - y \right\|^{p/\alpha} \\
&\leq \left(\left\| x - y \right\|^{\lceil q \rceil} + \tilde{A}_{\lceil q \rceil} \log^{-1}(1/\tilde{\lambda}_{\lceil q \rceil}) \tilde{\lambda}_{\lceil q \rceil}^{-\bar{\gamma}} \right)^{(1-p/\alpha)q/\lceil q \rceil} \bar{D}^{p/\alpha} \bar{\rho}^{k\gamma p/\alpha} \left\| x - y \right\|^{p/\alpha} \\
&\leq \left(\left\| x - y \right\|^{\lceil q \rceil} + \tilde{A}_{\lceil q \rceil} \log^{-1}(1/\tilde{\lambda}_{\lceil q \rceil}) \tilde{\lambda}_{\lceil q \rceil}^{-\bar{\gamma}} \right)^{(1-1/\alpha)p/\lceil q \rceil} \bar{D}^{p/\alpha} \bar{\rho}^{k\gamma p/\alpha} \left\| x - y \right\|^{p/\alpha} \\
&\leq \left(\left\| x - y \right\|^{(1-1/\alpha)p} + \left\{ \tilde{A}_{\lceil q \rceil} \log^{-1}(1/\tilde{\lambda}_{\lceil q \rceil}) \tilde{\lambda}_{\lceil q \rceil}^{-\bar{\gamma}} \right\}^{(1-1/\alpha)p/\lceil q \rceil} \right) \bar{D}^{p/\alpha} \bar{\rho}^{k\gamma p/\alpha} \left\| x - y \right\|^{p/\alpha} \\
&\leq \left(\left\| x - y \right\|^p + \left\{ \tilde{A}_{\lceil q \rceil} \log^{-1}(1/\tilde{\lambda}_{\lceil q \rceil}) \tilde{\lambda}_{\lceil q \rceil}^{-\bar{\gamma}} \right\}^{(1-1/\alpha)p/\lceil q \rceil} \left\| x - y \right\|^{p/\alpha} \right) \bar{D}^{p/\alpha} \bar{\rho}^{k\gamma p/\alpha} \\
&\leq \bar{D}_{4,\alpha}^p \bar{\rho}^{k\gamma p/\alpha} (\left\| x - y \right\|^p + \left\| x - y \right\|^{p/\alpha}) \ ,
\end{aligned}$$

which completes the proof upon using that for any $a, b \geq 0$ and $p \geq 1$, $(a + b)^{1/p} \leq a^{1/p} + b^{1/p}$.

Proofs of Section 4.1.4

Proof of Proposition 4.1.11

(a) By **B**2 and **B**3(m) we have for any $\gamma > 0$ and $x, y \in \mathsf{X}$, $\|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\|^2 \leq (1 - 2\gamma\mathfrak{m} + \gamma^2\mathrm{L}^2) \|x - y\|^2 \leq (1 + \gamma\kappa(\gamma)) \|x - y\|^2$, which concludes the proof.

(b) We have for any $\gamma > 0$ and $x, y \in \mathsf{X}$, $\|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\|^2 \leq \|x - y\|^2 + \gamma(-2\mathfrak{m}_b + \gamma) \|b(x) - b(y)\|^2$. Then if $\gamma \leq 2\mathfrak{m}_b$, $\|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\|^2 \leq \|x - y\|^2$, which concludes the proof.

Proof of Proposition 4.1.12 Let $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathsf{X}$ and set $\mathrm{E} = \mathcal{T}_\gamma(y) - \mathcal{T}_\gamma(x)$. We divide the proof into three parts.

(a) First, we show that Proposition 4.1.12-(a) holds. If $\mathrm{E} = 0$ then the proposition is trivial, therefore we suppose that $\mathrm{E} \neq 0$ and let $\mathrm{e} = \mathrm{E}/\|\mathrm{E}\|$. Consider Z_1, a d-dimensional Gaussian random variable with zero mean and covariance identity. By (4.23) and the fact that Π_{X} is non expansive, we have for any $\gamma \in (0, \bar{\gamma}]$

$$\begin{aligned}
\mathrm{K}_\gamma \|x - y\| &\leq \mathbb{E}\left[(1 - p_\gamma(x, y, \sqrt{\gamma}Z_1)) \left\|(\mathcal{T}_\gamma(x) + \sqrt{\gamma}Z_1) - (\mathcal{T}_\gamma(y) + \sqrt{\gamma}(\mathrm{Id} - 2\mathrm{e}\mathrm{e}^\top)Z_1)\right\|\right] \\
&= \mathbb{E}\left[\left\|\mathrm{E} - 2\sqrt{\gamma}\mathrm{e}\mathrm{e}^\top Z_1\right\| (1 - p_\gamma(x, y, \sqrt{\gamma}Z_1))\right] \\
&= \int_{\mathbb{R}} \|\mathrm{E} - 2z\mathrm{e}\| \left\{\boldsymbol{\varphi}_\gamma(z) - (\boldsymbol{\varphi}_\gamma(z) \wedge \boldsymbol{\varphi}_\gamma(\|\mathrm{E}\| - z))\right\} \mathrm{d}z \\
&= \int_{-\infty}^{\|\mathrm{E}\|/2} (\|\mathrm{E}\| - 2z) \left\{\boldsymbol{\varphi}_\gamma(z) - \boldsymbol{\varphi}_\gamma(\|\mathrm{E}\| - z)\right\} \mathrm{d}z \leq \|\mathrm{E}\| \ , \qquad (4.76)
\end{aligned}$$

where we have used the change of variable $z \mapsto \|\mathrm{E}\| - z$ for the last line. We conclude this part of the proof upon using **B**2 and **B**3(m).

(b) Second, we show that Proposition 4.1.12-(b) holds. Consider the case $(x, y) \in \Delta^{\mathrm{c}}_{\mathsf{X}, R_1}$. By **B**2, **C**1, and since for any $t \in [-1, +\infty)$, $\sqrt{1 + t} \leq 1 + t/2$, we have that

$$\|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\| \leq (1 - 2\gamma\mathfrak{m}_1^+ + \gamma^2\mathrm{L}^2)^{1/2} \|x - y\| \leq (1 - \gamma\mathfrak{m}_1^+ + \gamma^2\mathrm{L}^2/2) \|x - y\| \ . \qquad (4.77)$$

Combining (4.76) and (4.77) and since $\gamma < 2\mathfrak{m}_1^+/\mathrm{L}^2$, we obtain that for any $(x, y) \in \Delta^{\mathrm{c}}_{\mathsf{X}, R_1}$,

$$\begin{aligned}
\mathrm{K}_\gamma W_1(x, y) &\leq (1 - \gamma\mathfrak{m}_1^+ + \gamma^2\mathrm{L}^2/2) \|x - y\| / R_1 + 1 \\
&\leq (1 - \gamma\mathfrak{m}_1^+/2 + \gamma^2\mathrm{L}^2/4)(1 + \|x - y\| / R_1) \leq \lambda^\gamma W_1(x, y) \ . \qquad (4.78)
\end{aligned}$$

Similarly, we obtain using Proposition 4.1.11-(a) that for any $(x, y) \in \Delta_{\mathsf{X}, R_1}$

$$\begin{aligned}
\mathrm{K}_\gamma W_1 &\leq (1 - \gamma\mathfrak{m} + \gamma^2\mathrm{L}^2/2) \|x - y\| / R_1 + 1 \\
&\leq (1 - \gamma\mathfrak{m}_1^+/2 + \gamma^2\mathrm{L}^2/4) \|x - y\| / R_1 + 1 + \gamma \left\{\mathfrak{m}_1^+/2 - \mathfrak{m} + \gamma\mathrm{L}^2/4\right\} \\
&\leq (1 - \gamma\mathfrak{m}_1^+/2 + \gamma^2\mathrm{L}^2/4) W_1(x, y) + \gamma \left[\mathfrak{m}_1^+ - \mathfrak{m}\right] \leq \lambda^\gamma W_1(x, y) + A\gamma \ . \qquad (4.79)
\end{aligned}$$

We conclude the proof upon combining (4.78) and (4.79).

(c) Finally we show that Proposition 4.1.12-(c) holds. Let $p \in \mathbb{N}$ with $p \geq 2$. Similarly to Proposition 4.1.12-(a) , we have

$$\mathrm{K}_\gamma \|x - y\|^p = \int_{\mathbb{R}} (\|\mathrm{E}\| - 2z)^p \boldsymbol{\varphi}_\gamma(z) \mathrm{d}z \ . \qquad (4.80)$$

For any $k \in \mathbb{N}$, let $c_k = \int_{\mathbb{R}} z^k \boldsymbol{\varphi}_1(z) \mathrm{d}z$ and

$$\kappa_{1,\gamma} = 1 - \gamma \mathtt{m}_1^+ + \gamma^2 \mathtt{L}^2/2 \;, \qquad \kappa_{2,\gamma} = \max(1, 1 - \gamma \mathtt{m} + \gamma^2 \mathtt{L}^2/2) \;, \qquad \mathbb{R} = \max(1, R_1) \;. \tag{4.81}$$

Note that for any $k \in \mathbb{N}$, $c_{2k+1} = 0$. Consider the case $\|x - y\| \geq \mathbb{R}$. Using (4.80), (4.81), **B**2, **C**1 we have

$$\begin{aligned}
\mathrm{K}_\gamma \|x - y\|^p &\leq \|\mathrm{E}\|^p + \sum_{k=2}^{p} \binom{p}{k} \|\mathrm{E}\|^{p-k} (2\gamma)^k c_k \\
&\leq \kappa_{1,\gamma} \|x - y\|^p + \sum_{k=2}^{p} \binom{p}{k} \|x - y\|^{p-k} (2\gamma)^k c_k \\
&\leq \kappa_{1,\gamma} \|x - y\|^p + \gamma c_{2p} 2^{2p} \max(1, \bar{\gamma})^p \|x - y\|^{p-2} \\
&\leq \kappa_{2,\gamma/2} \|x - y\|^p + \gamma \left\{ c_{2p} 2^{2p} \max(1, \bar{\gamma})^p \|x - y\|^{p-2} - \mathtt{m}_1^+ \|x - y\|^p / 2 \right\} \\
&\leq \kappa_{2,\gamma/2} \|x - y\|^p + \gamma \sup_{t \in [0, +\infty)} \left\{ c_{2p} 2^{2p} \max(1, \bar{\gamma})^p t^{p-2} - \mathtt{m}_1^+ t^p / 2 \right\} \;.
\end{aligned}$$

Note that we have for any $a \geq b \geq 0$ and $t \geq 0$

$$(1 + ta)^p - (1 + tb) \leq t \left\{ -b + \max(1, t)^p \sum_{k=1}^{p} \binom{p}{k} a^k \right\} \leq t \left\{ \max(1, t)^p (1 + a)^p - b \right\} \;. \tag{4.82}$$

Now, consider the case $\|x - y\| \leq \mathbb{R}$. Using (4.80), (4.81), (4.82), **B**2, **B**3($\mathtt{m}$) we have

$$\begin{aligned}
&\mathrm{K}_\gamma \|x - y\|^p - \kappa_{1,\gamma/2} \|x - y\|^p \leq (\kappa_{2,\gamma}^p - \kappa_{1,\gamma/2}) \|x - y\|^p + \gamma c_{2p} 2^{2p} \max(1, \bar{\gamma})^p \kappa_{2,\gamma}^p \mathbb{R}^{p-2} \\
&\leq \gamma c_{2p} 2^{2p} \max(1, \bar{\gamma})^p \kappa_{2,\gamma}^p \mathbb{R}^{p-2} + (\kappa_{2,\gamma}^p - \kappa_{1,\gamma/2}) \mathbb{R}^p \\
&\leq \gamma c_{2p} 2^{2p} \max(1, \bar{\gamma})^p \kappa_{2,\gamma}^p \mathbb{R}^{p-2} + \gamma \mathbb{R}^p \left\{ \max(1, \bar{\gamma})^p (1 - \mathtt{m}/2 + \mathtt{L}^2 \bar{\gamma}/4)^p + \mathtt{m}_1^+ \right\} \;,
\end{aligned}$$

which concludes the proof upon setting

$$\begin{aligned}
\lambda_p &= \exp[-\mathtt{m}_1^+/2 + \bar{\gamma} \mathtt{L}^2/4] \;, \\
A_p &= \max \left\{ A_{p,1}, A_{p,2} \right\} \;, \\
A_{p,1} &= \sup_{t \in [0, +\infty)} \left\{ c_{2p} 2^{2p} \max(1, \bar{\gamma})^p t^{p-2} - \mathtt{m}_1^+ t^p / 2 \right\} \;, \\
A_{p,2} &= c_{2p} 2^{2p} \max(1, \bar{\gamma})^p \kappa_{2,\gamma}^p \mathbb{R}^{p-2} + \mathbb{R}^p \left\{ \max(1, \bar{\gamma})^p (1 - \mathtt{m}/2 + \mathtt{L}^2 \bar{\gamma}/4)^p + \mathtt{m}_1^+ \right\} \;.
\end{aligned}$$

Proof of Corollary 4.1.14 Let $\bar{\gamma} > 0$. Then for any $\gamma \in (0, \bar{\gamma}]$, $\boldsymbol{\Psi} : t \mapsto 2\boldsymbol{\Phi}\{-t/(2\Xi_{\lceil 1/\gamma \rceil}^{1/2}(\kappa))\}$ is convex on $[0, +\infty)$, differentiable on $\mathbb{R}$, and for any $\gamma \in (0, \bar{\gamma}]$

$$\boldsymbol{\Psi}'(0) \geq -(\pi \inf_{\gamma \in (0, \bar{\gamma}]} \Xi_{\lceil 1/\gamma \rceil}(\kappa))^{-1/2} \;. \tag{4.83}$$

We divide the rest of the proof into two parts.

(a) First combining (4.83), Proposition 4.1.12-(a), (4.49), Theorem 4.1.13 and Proposition 4.1.9 shows that

$$\mathbf{W}_1(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq D_{3,\bar{\gamma},a} \rho_{\bar{\gamma},a}^{k\gamma/4} \|x - y\| \;.$$

(b) Second, combining Proposition 4.1.12-(c) and Proposition 4.1.10 shows that

$$\mathbf{W}_p(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq D_{\alpha,\bar{\gamma},a} \rho_{\bar{\gamma},a}^{k\gamma/(4\alpha)} \left\{ \|x - y\| + \|x - y\|^{1/\alpha} \right\} \;.$$

Proof of Proposition 4.1.15 We preface the proof by a technical result.

Lemma 4.1.42. *Let $\bar{\gamma} > 0$, such that for any $\gamma \in (0, \bar{\gamma}]$, P_γ is a Markov kernel and Q_γ is a Markov coupling kernel for P_γ. Assume that there exist $V : \mathsf{X} \to [1, +\infty)$ measurable, $\lambda \in (0, 1)$ and $A \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, P_γ satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, A\gamma, \mathsf{X})$. Let $W : \mathsf{X}^2 \to [1, +\infty)$ given for any $x, y \in \mathsf{X}$ by $W(x, y) = \{V(x) + V(y)\}/2$. The following properties hold.*

(a) Q_γ *satisfies* $\mathbf{D}_{\mathrm{d}}(W, \lambda^\gamma, A\gamma, \mathsf{X}^2)$

(b) if $\lim_{\|x\| \to +\infty} V(x) = +\infty$, Q_γ *satisfies* $\mathbf{D}_{\mathrm{d}}(W, \lambda^{\gamma/2}, A\gamma, \bar{\mathrm{B}}(0, R) \times \bar{\mathrm{B}}(0, R))$ *where* $R = \inf\{r \geq 0 \, : \,$ *for any* $x \in \bar{\mathrm{B}}(0, r)^{\mathrm{c}}$, $V(x) \geq 2(\lambda^{1/2})^{-2\bar{\gamma}} A \log^{-1}(1/\lambda^{1/2})\}$ *and* $\bar{\mathrm{B}}(0, R) \times \bar{\mathrm{B}}(0, R) \subset \Delta_{\mathsf{X}, 2R}$.

Proof. Let $\gamma \in (0, \bar{\gamma}]$ and $x, y \in \mathsf{X}$.

(a) Since $\delta_{(x,y)} \mathrm{Q}_\gamma$ is a transference plan between $\delta_x \mathrm{P}_\gamma$ and $\delta_y \mathrm{P}_\gamma$ we have

$$\mathrm{Q}_\gamma W(x, y) = \mathrm{Q}_\gamma \{V(x) + V(y)\}/2 = \mathrm{P}_\gamma V(x)/2 + \mathrm{P}_\gamma V(y)/2 \leq \lambda^\gamma W(x, y) + A\gamma \, .$$

(b) Let $x, y \in \mathsf{X}$. If $(x, y) \in \bar{\mathrm{B}}(0, R) \times \bar{\mathrm{B}}(0, R)$ then the result is immediate using Lemma 4.1.42-(a). Now, assume that $(x, y) \notin \bar{\mathrm{B}}(0, R) \times \bar{\mathrm{B}}(0, R)$. By definition of R, $\max(V(x), V(y)) \geq 4\lambda^{-\bar{\gamma}} A \log^{-1}(1/\lambda)$. Without loss of generality assume that $V(x) \geq V(y)$. Using this result, Lemma 4.1.42-(a) and that for any $b \geq a$, $(\mathrm{e}^b - \mathrm{e}^a) \geq \mathrm{e}^a (b - a)$, we have

$$\begin{aligned}
\mathrm{Q}_\gamma W(x, y) &\leq \lambda^\gamma W(x, y) + A\gamma \\
&\leq \lambda^{\gamma/2} W(x, y) + \gamma \left[A + \lambda^\gamma \{\log(\lambda) - \log(\lambda)/2\} W(x, y)\right] \\
&\leq \lambda^{\gamma/2} W(x, y) + \gamma \left[A - \lambda^{\bar{\gamma}} \log(\lambda^{-1}) V(x)/4\right] \leq \lambda^{\gamma/2} W(x, y) \, .
\end{aligned}$$

□

Proof of Proposition 4.1.15. Let $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathsf{X}$. We divide the proof into two parts. Using (4.19), **B** 2, **B**3(m), **C**2, that the projection Π_{X} is non expansive and $\gamma < 2\mathtt{m}_2^+ / \mathtt{L}^2$, we obtain for any $x \in \mathsf{X}$

$$\begin{aligned}
\mathrm{R}_\gamma V(x) &\leq 1 + \|x + \gamma b(x)\|^2 + \gamma d \\
&\leq 1 + \|x\|^2 + 2\gamma \langle x, b(x) \rangle + \gamma^2 \|b(x)\|^2 + \gamma d \\
&\leq (1 + \|x\|^2) \left[1 - \gamma(2\mathtt{m}_2^+ - \bar{\gamma}\mathtt{L}^2)\right] + \gamma \left(d + 2R_2^2 (\mathtt{m}_2^+ - \mathtt{m})_+ + 2\mathtt{m}_2^+\right) \, .
\end{aligned}$$

In addition, for any $x \in \mathsf{X}$, such that $\|x\| \geq 2A^{1/2} \log^{-1/2}(1/\lambda)$, we have $V(x) \geq 4A \log^{-1}(1/\lambda)$. We conclude the proof using Lemma 4.1.42-(b).

□

Proof of Proposition 4.1.17 Let $\gamma \in (0, \bar{\gamma}]$. Using the fact that Π_{X} is non expansive, the Log-Sobolev inequality, the fact that π is 1-Lipschitz, [BLM13, Theorem 5.5] and the Jensen inequality we obtain for any $x \in \mathbb{R}^d$

$$\begin{aligned}
\mathrm{R}_\gamma V(x) &\leq \exp\left[\mathtt{m}_3^+ \mathrm{R}_\gamma \phi(x) + \gamma (\mathtt{m}_3^+)^2/2\right] \leq \exp\left[\mathtt{m}_3^+ \sqrt{1 + \mathrm{R}_\gamma \|x\|^2} + \gamma (\mathtt{m}_3^+)^2/2\right] \\
&\leq \exp\left[\mathtt{m}_3^+ \sqrt{1 + \|\mathcal{T}_\gamma(x)\|^2 + \gamma d} + \gamma (\mathtt{m}_3^+)^2/2\right] \, . \qquad (4.84)
\end{aligned}$$

Let $x \in \mathbb{R}^d$. The rest of the proof is divided in two parts.

(a) In the first case, $\|x\| \geq R_4$. Since $\|x\| \geq R_3$ and $\gamma \leq 2\mathtt{k}_2$, we have using **C**3

$$\|\mathcal{T}_\gamma(x)\|^2 \leq \|x\|^2 - 2\gamma\mathtt{k}_1 \|x\| + \gamma(\gamma - 2\mathtt{k}_2) \|b(x)\|^2 + \gamma\mathtt{a} \leq \|x\|^2 - 2\gamma\mathtt{k}_1 \|x\| + \gamma\mathtt{a} \,. \tag{4.85}$$

Since $\|x\| \geq 1$ we have $2\|x\| \geq \phi(x)$ and therefore, using that $\|x\| \geq (d+\mathtt{a})/\mathtt{k}_1$, $2\mathtt{k}_1 \|x\| \geq 2\mathtt{m}_3^+ \phi(x) + d + \mathtt{a}$. This inequality, combined with the fact that for any $t \in (-1, +\infty)$, $\sqrt{1+t} \leq 1 + t/2$, yields

$$\begin{aligned} &\sqrt{1 + \|x\|^2 + \gamma(-2\mathtt{k}_1 \|x\| + d + \mathtt{a})} - \phi(x) \\ &\qquad \leq \gamma(-2\mathtt{k}_1 \|x\| + d + \mathtt{a})/(2\phi(x)) \leq -\gamma\mathtt{m}_3^+ \,. \end{aligned} \tag{4.86}$$

Combining (4.84), (4.85) and (4.86) we get

$$\mathrm{R}_\gamma V(x) \leq \lambda^\gamma V(x) \,.$$

(b) In the second case $\|x\| \leq R_4$. We have the following inequality using **C**3 and that $\gamma \leq 2\mathtt{k}_2$

$$\|\mathcal{T}_\gamma(x)\|^2 \leq \|x\|^2 + \gamma(\gamma - 2\mathtt{k}_2) \|b(x)\|^2 + \gamma c \leq \|x\|^2 + \gamma\mathtt{a} \,. \tag{4.87}$$

Combining (4.84), (4.87) and the fact that for any $t \in (-1, +\infty)$, $\sqrt{1+t} \leq 1 + t/2$ we get

$$\begin{aligned} \mathrm{R}_\gamma V(x) &\leq \exp\left[\gamma\mathtt{m}_3^+(d+\mathtt{a})/(2\phi(x)) + \gamma(\mathtt{m}_3^+)^2/2\right] V(x) \\ &\leq \exp\left[\gamma(\mathtt{m}_3^+(d+\mathtt{a}) + (\mathtt{m}_3^+)^2)/2\right] V(x) \,. \end{aligned} \tag{4.88}$$

Note that for any $c_1 \geq c_2$ and $t \in [0, \bar{t}]$ we have the following inequality

$$\mathrm{e}^{c_1 t} \leq \mathrm{e}^{c_2 t} + \mathrm{e}^{c_1 \bar{t}}(c_1 - c_2)t \,. \tag{4.89}$$

Combining (4.88) and (4.89) we get

$$\mathrm{R}_\gamma V(x) \leq \lambda^\gamma V(x) + \exp\left[\bar{\gamma}(\mathtt{m}_3^+(d+\mathtt{a}) + (\mathtt{m}_3^+)^2)/2\right] C_\mathtt{a}\gamma \,,$$

with $C_\mathtt{a} = (\mathtt{m}_3^+(d+\mathtt{a})/2 + (\mathtt{m}_3^+)^2) \exp(\mathtt{m}_3^+(1+R_4^2)^{1/2})$, which concludes the proof using Lemma 4.1.42.

Proofs of Section 4.1.5

Proof of Theorem 4.1.19 Combining Proposition 4.1.29, Proposition 4.1.33 and Proposition 4.1.35 in Theorem 4.1.28 concludes the proof.

Proof of Theorem 4.1.20 Combining Proposition 4.1.29, Proposition 4.1.33 and Proposition 4.1.35 in Theorem 4.1.28 concludes the proof.

Proof of Theorem 4.1.24 Let $T \geq 0$ and $x \in \mathbb{R}^d$. First, using Proposition 4.1.32 we have that **L**1, **L**2, **L**4 and **L**5 are satisfied. In addition, using Proposition 4.1.35 we get

$$\mathrm{P}_T V_{\mathtt{k}_1}(x) < +\infty \,,$$

where $V_{\mathtt{k}_1} = V_M$ with $M \leftarrow \mathtt{k}_1$ and V_M given in (4.56). Hence, since we have $\sup_{x \in \mathbb{R}^d} \|b(x)\|^{2(1+\varepsilon_b)} \mathrm{e}^{-\mathtt{k}_1(1+\|x\|)^{1/2}} < +\infty$, we get that **L**3 is satisfied.

In addition, using that $2\mathtt{m}_3^+ \leq \mathtt{k}_1$ we have

$$\mathrm{P}_T V^2(x) < +\infty \, .$$

Thus, the first part of (4.62) is satisfied. Second using Proposition 4.1.17 and replacing $\mathtt{m}_3^+ \leftarrow 2\mathtt{m}_3^+$ (which is valid since $\mathtt{m}_3^+ \leq \mathtt{k}_1/4$), we obtain that for any $n, m \in \mathbb{N}$, with $T/m < 2\mathtt{k}_2$, $\mathrm{R}_{T/m,n}$ and $\tilde{\mathrm{R}}_{T/m}$ satisfy $\mathbf{D}_{\mathrm{d}}(V^2, \lambda^{T/m}, AT/m, \mathsf{X}^2)$. Hence, for any $m \in \mathbb{N}$ with $T/m < 2\mathtt{k}_2$ we have

$$\mathrm{R}_{T/m,n}^m + \tilde{\mathrm{R}}_{T/m,n}^m V^2(x) \leq V^2(x) + ATm^{-1} \sum_{k \in \mathbb{N}} \lambda^{T/m} \leq V^2(x) + A \log^{-1}(1/\lambda) \lambda^{-\bar{\gamma}} \, .$$

Therefore, the second part of (4.62) is satisfied and we can apply Theorem 4.1.28. Using Theorem 4.1.18 and we get that for any $m, n \in \mathbb{N}$ with $x, y \in \mathbb{R}^d$ and $T/m \in (0, 2\mathtt{k}_2)$

$$\|\delta_x \mathrm{R}_{T/m,n}^m - \delta_y \mathrm{R}_{T/m,n}^m\|_V \leq C_{1/m,c} \rho_{1/m,c}^T \left\{V(x) + V(y)\right\} \, ,$$

where $C_{1/m,c} \geq 0$ and $\rho_{1/m,c} \in (0,1)$, see Section 4.1.7. We conclude upon noting that $C_{1/m,c}$ and $\rho_{1/m,c}$ admit limits C_c and ρ_c when $m \to +\infty$ which do not depend on n.

Proof of Lemma 4.1.25

(a) Let $x \in \mathbb{R}^d$ and let $(\mathbf{X}_t)_{t \geq 0}$ a solution of (4.55) starting from x. Define for any $k \in \mathbb{N}^\star$, $\tau_k = \inf\{t \geq 0 \, : \, \|\mathbf{X}_t\| \geq k\}$ and for any $t \geq 0$, $\mathbf{M}_t = \int_0^t \langle \nabla V(\mathbf{X}_s), \mathrm{d}B_s \rangle$. Using the Itô formula we obtain that for every $t \geq 0$ and $k \in \mathbb{N}^\star$

$$\begin{aligned} V(\mathbf{X}_{t \wedge \tau_k}) \mathrm{e}^{\zeta(t \wedge \tau_k)} &= \int_0^{t \wedge \tau_k} \left[\mathrm{e}^{\zeta(t \wedge \tau_k)} \mathcal{A} V(\mathbf{X}_u) + \zeta \mathrm{e}^{\zeta u} V(\mathbf{X}_u)\right] \mathrm{d}u + \mathbf{M}_{t \wedge \tau_k} + V(x) \\ &= V(\mathbf{X}_{s \wedge \tau_k}) \mathrm{e}^{\zeta(s \wedge \tau_k)} + \mathbf{M}_{t \wedge \tau_k} - \mathbf{M}_{s \wedge \tau_k} + \int_{s \wedge \tau_k}^{t \wedge \tau_k} \left[\mathrm{e}^{\zeta(t \wedge \tau_k)} \mathcal{A} V(\mathbf{X}_u) + \zeta \mathrm{e}^{\zeta u} V(\mathbf{X}_u)\right] \mathrm{d}u \\ &\leq V(\mathbf{X}_{s \wedge \tau_k}) \mathrm{e}^{\zeta(s \wedge \tau_k)} + \mathbf{M}_{t \wedge \tau_k} - \mathbf{M}_{s \wedge \tau_k} \, . \end{aligned}$$

Therefore since for any $k \in \mathbb{N}^\star$, $(\mathbf{M}_{t \wedge \tau_k})_{t \geq 0}$ is a $(\mathcal{F}_t)_{t \geq 0}$-martingale, we get for every $t \geq s \geq 0$ and $k \in \mathbb{N}^\star$

$$\mathbb{E}\left[V(\mathbf{X}_{t \wedge \tau_k}) \mathrm{e}^{\zeta(t \wedge \tau_k)} \middle| \mathcal{F}_s\right] \leq V(\mathbf{X}_{s \wedge \tau_k}) \mathrm{e}^{\zeta(s \wedge \tau_k)} \, ,$$

which concludes the first part of the proof taking $k \to +\infty$ and using Fatou's lemma.

(b) Similarly we have that $(V(\mathbf{X}_t)\mathrm{e}^{\zeta t} - B(1 - \exp(-\zeta t))/\zeta)_{t \geq 0}$ is a $(\mathcal{F}_t)_{t \geq 0}$-supermartingale which concludes the proof upon taking the expectation of $V(\mathbf{X}_t)\mathrm{e}^{\zeta t} - B(1 - \exp(-\zeta t))/\zeta$.

Proof of Proposition 4.1.26 Let $T \geq 0$, $x \in \mathbb{R}^d$, $n \in \mathbb{N}$ and $m \in \mathbb{N}^\star$ with $T/m \leq \bar{\gamma}$. Using [DM17, Lemma 24], we obtain

$$\begin{aligned} &\|\delta_x \mathrm{P}_T - \delta_x \tilde{\mathrm{R}}_{T/m,n}^m\|_V \\ &\quad \leq (1/\sqrt{2}) \left(\delta_x \mathrm{P}_T V^2(x) + \delta_x \tilde{\mathrm{R}}_{T/m,n}^m V^2(x)\right)^{1/2} \mathrm{KL}\left(\delta_x \mathrm{P}_T | \delta_x \tilde{\mathrm{R}}_{T/m,n}^m\right)^{1/2} \, . \end{aligned}$$

Let $M \geq 0$, $n \in \mathbb{N}^\star$ with $n^{-1} < \bar{\gamma}$, $x \in \mathbb{R}^d$ and $k \in \mathbb{N}$. Therefore, we only need to show that $\lim_{m \to +\infty} \mathrm{KL}(\delta_x \mathrm{P}_T | \delta_x \tilde{\mathrm{R}}_{T/m,n}^m) = 0$. Consider the two processes $(\mathbf{X}_t)_{t \in [0,T]}$ and $(\tilde{\mathbf{X}}_t)_{t \in [0,T]}$ defined by (4.55) with $\mathbf{X}_0 = \tilde{\mathbf{X}}_0 = x$ and

$$\mathrm{d}\tilde{\mathbf{X}}_t = \tilde{b}_{T/m,n}(t, (\tilde{\mathbf{X}}_s)_{s \in [0,T]}) \mathrm{d}t + \mathrm{d}\mathbf{B}_t \, , \qquad \tilde{\mathbf{X}}_0 = x \, ,$$

where for any $(\mathrm{w}_s)_{s\in[0,T]} \in \mathrm{C}([0,T],\mathbb{R}^d)$, $t\in[0,T]$,

$$\tilde{b}_{T/m,n}(t,(\mathrm{w}_s)_{s\in[0,T]}) = \sum_{i=0}^{m-1} b_{T/m,n}(\mathrm{w}_{iT/n})\mathbb{1}_{[iT/m,(i+1)T/m)}(t) \; . \tag{4.90}$$

Note for any $i\in\{0,\dots,m\}$, the distribution of $\tilde{\mathbf{X}}_{iT/m}$ is $\delta_x\tilde{\mathrm{R}}^i_{T/m,n}$. Using that b and $b_{T/m,n}$ are continuous and that $(\mathbf{X}_t)_{t\in[0,T]}$ and $(\tilde{\mathbf{X}}_t)_{t\in[0,T]}$ take their values in $\mathrm{C}([0,T],\mathbb{R}^d)$, we obtain that

$$\mathbb{P}\left(\int_0^T \|b(\mathbf{X}_t)\|^2\mathrm{d}t < +\infty\right) = 1 \; ,$$
$$\mathbb{P}\left(\int_0^T \|\tilde{b}_{T/m,n}(t,(\tilde{\mathbf{X}}_s)_{s\in[0,T]})\|^2\mathrm{d}t < +\infty\right) = 1 \; ,$$

and

$$\mathbb{P}\left(\int_0^T \|b(\mathbf{B}_t)\|^2\mathrm{d}t < +\infty\right) = 1 \; ,$$
$$\mathbb{P}\left(\int_0^T \|\tilde{b}_{T/m,n}(t,(\mathbf{B}_s)_{s\in[0,T]})\|^2\mathrm{d}t < +\infty\right) = 1 \; ,$$

where $(\mathbf{B}_t)_{t\in[0,T]}$ is the d-dimensional Brownian motion associated with (4.55). Therefore by [LS01, Theorem 7.7] the distributions of $(\mathbf{X}_t)_{t\in[0,T]}$ and $(\tilde{\mathbf{X}}_t)_{t\in[0,T]}$, denoted by μ^x and $\tilde{\mu}^x$ respectively, are equivalent to the distribution of the Brownian motion μ^x_B starting at x. In addition, μ^x admits a Radon-Nikodym density w.r.t. to μ^x_B and μ^x_B admits a Radon-Nikodym density w.r.t. to $\tilde{\mu}^x$, given μ^x_B-almost surely for any $(\mathrm{w}_t)_{t\in[0,T]} \in \mathrm{C}([0,T],\mathbb{R}^d)$ by

$$\frac{\mathrm{d}\mu^x}{\mathrm{d}\mu^x_B}((\mathrm{w}_t)_{t\in[0,T]}) = \exp\left((1/2)\int_0^T \langle b(\mathrm{w}_s),\mathrm{dw}_s\rangle - (1/4)\int_0^T \|b(\mathrm{w}_s)\|^2\mathrm{d}s\right) \; ,$$
$$\frac{\mathrm{d}\mu^x_B}{\mathrm{d}\tilde{\mu}^x}((\mathrm{w}_t)_{t\in[0,T]}) = \exp\left(-(1/2)\int_0^T \langle \tilde{b}_{T/m,n}(s,(\mathrm{w}_u)_{u\in[0,T]}),\mathrm{dw}_s\rangle \right.$$
$$\left. +(1/4)\int_0^T \|\tilde{b}_{T/m,n}(s,(\mathrm{w}_u)_{u\in[0,T]})\|^2\mathrm{d}s\right) \; .$$

Finally we obtain that μ^x_B-almost surely for any $(\mathrm{w}_s)_{s\in[0,T]} \in \mathrm{C}([0,T],\mathbb{R}^d)$

$$\frac{\mathrm{d}\mu^x}{\mathrm{d}\tilde{\mu}^x}((\mathrm{w}_t)_{t\in[0,T]}) = \exp\left((1/2)\int_0^T \langle b(\mathrm{w}_s) - \tilde{b}_{T/m,n}(s,(\mathrm{w}_u)_{u\in[0,T]}),\mathrm{dw}_s\rangle \right.$$
$$\left. +(1/4)\int_0^T \|\tilde{b}_{T/m,n}(s,(\mathrm{w}_u)_{u\in[0,T]})\|^2 - \|b(\mathrm{w}_s)\|^2\mathrm{d}s\right) \; . \tag{4.91}$$

Now define for any $(\mathrm{w}_s)_{s\in[0,T]} \in \mathrm{C}([0,T],\mathbb{R}^d)$ and $t\in[0,T]$

$$b_{T/m}(t,(\mathrm{w}_s)_{s\in[0,T]}) = \sum_{i=0}^{m-1} b(\mathrm{w}_{iT/m})\mathbb{1}_{[iT/m,(i+1)T/m)}(t) \; . \tag{4.92}$$

Using (4.55), (4.90), (4.91), **L**2, and for any $a_1, a_1 \in \mathbb{R}^d$, $\|a_1 - a_2\|^2 \leq 2(\|a_1\|^2 + \|a_2\|^2)$, we obtain that

$$
\begin{aligned}
2\mathrm{KL}\left(\delta_x \mathrm{P}_T | \delta_x \tilde{\mathrm{R}}^m_{T/m,n}\right) &\leq 2^{-1}\mathbb{E}\left[\int_0^T \|b(\mathbf{X}_s) - \tilde{b}_{T/m,n}(s, (\mathbf{X}_u)_{u\in[0,T]})\|^2 \mathrm{d}s\right] \qquad (4.93)\\
&\leq \mathbb{E}\left[\int_0^T \|b(\mathbf{X}_s) - b_{T/m}(s, (\mathbf{X}_u)_{u\in[0,T]})\|^2 \mathrm{d}s\right] \\
&\quad + \sum_{i=0}^{m-1} \mathbb{E}\left[\int_{iT/m}^{(i+1)T/m} \|b(\mathbf{X}_{iT/m}) - b_{T/m,n}(\mathbf{X}_{iT/m})\|^2 \mathrm{d}s\right] \\
&\leq \mathbb{E}\left[\int_0^T \|b(\mathbf{X}_s) - b_{T/m}(s, (\mathbf{X}_u)_{u\in[0,T]})\|^2 \mathrm{d}s\right] \\
&\quad + C_1 T^{1+\beta} m^{-\beta} \sup_{s\in[0,T]} \mathbb{E}\left[\|b(\mathbf{X}_s)\|^2\right] .
\end{aligned}
$$

It only remains to show that the first term goes to 0 as $m \to +\infty$. Note that since $(\mathbf{X}_s)_{s\in[0,T]}$ is almost surely continuous and b is continuous on $\mathbb{R}^d$, $\lim_{m\to+\infty} \|b(\mathbf{X}_s) - b_{T/m}(s, (\mathbf{X}_u)_{u\in[0,T]})\|^2 = 0$ for any $s \in [0, T]$ almost surely. Then, using the Lebesgue dominated convergence theorem and the continuity of b, we obtain that for any $M \geq 0$,

$$
\lim_{m\to+\infty} \mathbb{E}\left[\mathbb{1}_{[0,M]}\left(\sup_{s\in[0,T]} \|\mathbf{X}_s\|\right) \int_0^T \|b(\mathbf{X}_s) - b_{T/m}(s, (\mathbf{X}_u)_{u\in[0,T]})\|^2 \mathrm{d}s\right] = 0 . \qquad (4.94)
$$

On the other hand, using Hölder's inequality and the definition of $b_{T/m}$ (4.92), we obtain that for any $M \geq 0$,

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{(M,+\infty)}\left(\sup_{s\in[0,T]} \|\mathbf{X}_s\|\right) \int_0^T \|b(\mathbf{X}_s) - b_{T/m}(s, (\mathbf{X}_s)_{s\in[0,T]})\|^2 \mathrm{d}s\right] \\
&\leq 2\left(\mathbb{P}\left(\sup_{s\in[0,T]} \|\mathbf{X}_s\| > M\right)\right)^{\varepsilon_b/(1+\varepsilon_b)} \\
&\int_0^T \left(\mathbb{E}^{1/(1+\varepsilon_b)}\left[\|b(\mathbf{X}_s)\|^{2(1+\varepsilon_b)}\right] + \mathbb{E}^{1/(1+\varepsilon_b)}\left[\left\|b_{T/m}(s, (\mathbf{X}_u)_{u\in[0,T]})\right\|^{2(1+\varepsilon_b)}\right]\right) \mathrm{d}s \\
&\leq 4T\left(\mathbb{P}\left(\sup_{s\in[0,T]} \|\mathbf{X}_s\| > M\right)\right)^{\varepsilon_b/(1+\varepsilon_b)} \left(\sup_{s\in[0,T]} \mathbb{E}\left[\|b(\mathbf{X}_s)\|^{2(1+\varepsilon_b)}\right]\right)^{1/(1+\varepsilon_b)} .
\end{aligned}
$$

Combining this result, **L**3, and (4.94) in (4.93), we obtain that for any $M \geq 0$,

$$
\begin{aligned}
&\limsup_{m\to+\infty} \mathrm{KL}\left(\delta_x \mathrm{P}_T | \delta_x \tilde{\mathrm{R}}^m_{T/m,n}\right) \\
&\qquad \leq 2T\left(\mathbb{P}\left(\sup_{s\in[0,T]} \|\mathbf{X}_s\| > M\right)\right)^{\varepsilon_b/(1+\varepsilon_b)} \left(\sup_{s\in[0,T]} \mathbb{E}\left[\|b(\mathbf{X}_s)\|^{2(1+\varepsilon_b)}\right]\right)^{1/(1+\varepsilon_b)} .
\end{aligned}
$$

Since $(\mathbf{X}_s)_{s\in[0,T]}$ is a.s. continuous, we get by the monotone convergence theorem and **L**3, taking $M \to +\infty$, that $\lim_{m\to+\infty} \mathrm{KL}(\delta_x \mathrm{P}_T | \delta_x \tilde{\mathrm{R}}^m_{T/m,n}) = 0$, which concludes the proof.

Proof of Proposition 4.1.27 For any $n \in \mathbb{N}$ and $\gamma \in (0, \bar{\gamma}]$, we consider the synchronous Markov coupling $\mathrm{Q}_{\gamma,n}$ for $\mathrm{R}_{\gamma,n}$ and $\tilde{\mathrm{R}}_{\gamma,n}$ defined for any $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\begin{aligned} &\mathrm{Q}_{\gamma,n}((x, y), \mathsf{A}) \\ &= \frac{1}{(2\pi\gamma)^{d/2}} \int_{\mathbb{R}^d} \mathbb{1}_{(\mathrm{Id},\Pi_{\kappa_n})^{\leftarrow}(\mathsf{A})}(\mathcal{T}_\gamma(x) + \sqrt{\gamma}z, \mathcal{T}_\gamma(y) + \sqrt{\gamma}z)\mathrm{e}^{-\|z\|^2/2}\mathrm{d}z \ , \end{aligned} \tag{4.95}$$

with $\mathcal{T}_\gamma(x) = x + \gamma b(x)$. Let $T \geq 0$, $n \in \mathbb{N}$, $m \in \mathbb{N}^\star$ such that $T/m \leq \bar{\gamma}$. Consider $(X_j, \tilde{X}_j)_{j\in\mathbb{N}}$ a Markov chain with Markov kernel $\mathrm{Q}_{T/m,n}$ and started from $X_0 = \tilde{X}_0 = x$ for a fixed $x \in \mathbb{R}^d$. Note that by definition and **L**4, we have that for $k < \tau$, $X_k = \tilde{X}_k$ where $\tau = \inf\{j \in \mathbb{N} : \tilde{X}_j \notin \bar{\mathrm{B}}(0, n)\}$. Using **L**5, $\left(\tilde{V}(\tilde{X}_j) \exp\left[-j \log(\tilde{A})(T/m)(1 + \tilde{E}_n(T/m)^{\tilde{\varepsilon}_n})\right]\right)_{j\in\mathbb{N}}$ is a positive supermartingale. Combining (4.95), the Cauchy-Schwarz inequality, **L**5 and the Doob maximal inequality for positive supermartingale [Nev75, Proposition II-2-7], we get for any $x \in \mathbb{R}^d$

$$\begin{aligned} &\|\delta_x \mathrm{R}^m_{T/m,n} - \delta_x \tilde{\mathrm{R}}^m_{T/m,n}\|_V \leq \mathbb{E}\left[\mathbb{1}_{\Delta^{\mathrm{c}}_{\mathbb{R}^d}}(X_m, \tilde{X}_m)(V(X_m) + V(\tilde{X}_m))/2\right] \\ &\leq (1/2)\mathbb{P}\left(\sup_{j\in\{0,\ldots,m\}} \left\|\tilde{X}_j\right\| \geq n\right)\left(\mathbb{E}\left[V^2(X_m)\right]^{1/2} + \mathbb{E}\left[V^2(\tilde{X}_m)\right]^{1/2}\right) \\ &\leq (1/2)\mathbb{P}\left(\sup_{j\in\{0,\ldots,m\}} \tilde{V}(\tilde{X}_j) \geq n\right)\left(\mathbb{E}\left[V^2(X_m)\right]^{1/2} + \mathbb{E}\left[V^2(\tilde{X}_m)\right]^{1/2}\right) \\ &\leq (2n)^{-1} \exp\left[\log(\tilde{A})(T/m)(1 + \tilde{E}_n(T/m)^{\varepsilon_n})\right] \tilde{V}(x) \\ &\qquad \times \left((\mathrm{R}^m_{T/m,n} V^2(x))^{1/2} + (\tilde{\mathrm{R}}^m_{T/m,n} V^2(x))^{1/2}\right) \ , \end{aligned}$$

which concludes the proof upon taking $m \to +\infty$ then $n \to +\infty$.

Proof of Proposition 4.1.29 Let $p \in \mathbb{N}^\star$ and $V \in \mathrm{C}^2(\mathbb{R}^d, [1, +\infty))$ be defined for any $x \in \mathbb{R}^d$ by $V(x) = 1 + \|x\|^{2p}$. For any $x \in \mathbb{R}^d$, $\nabla V(x) = 2p \|x\|^{2(p-1)} x$ and $\Delta V(x) = (4p(p-1) + 2pd) \|x\|^{2(p-1)}$. Therefore, using **B**2 and the definition of $\mathcal{A}$ we obtain that for any $x \in \mathbb{R}^d$

$$\mathcal{A}V(x) \leq [2p(p-1) + p(d + 2\mathrm{L})] V(x) \ . \tag{4.96}$$

Hence, using (4.96) and [Kha11, Theorem 3.5], we obtain that **L**1 holds. Using that for any $\sup_{x\in\mathbb{R}^d} \|b(x)\|(1 + \|x\|^2)^{-1} < +\infty$, (4.96) and Lemma 4.1.25-(b) we obtain that **L**3 holds.

L2 and **L**4 are trivially satisfied. Finally, using once again **B**2, we have that for any $x \in \mathbb{R}^d$ and $\gamma \in (0, \bar{\gamma}]$ we have

$$\begin{aligned} \mathrm{R}_\gamma(1 + \|x\|^2) &\leq 1 + \|x + \gamma b(x)\|^2 + \gamma d \\ &\leq 1 + \|x\|^2 + 2\gamma \|b(x)\| \|x\| + \gamma^2 \|b(x)\|^2 + \gamma d \\ &\leq 1 + \|x\|^2 + 2\gamma \mathrm{L} \|x\|^2 + \gamma^2 \mathrm{L}^2 \|x\|^2 + \gamma d \\ &\leq (1 + 2\gamma\mathrm{L} + \gamma^2\mathrm{L}^2 + \gamma d)(1 + \|x\|^2) \ , \end{aligned}$$

which implies that **L**5 holds.

Proof of Proposition 4.1.30 Let $p \in \mathbb{N}^\star$ and $V \in \mathrm{C}^2(\mathbb{R}^d, [1, +\infty))$ be defined for any $x \in \mathbb{R}^d$ by $V(x) = 1 + \|x\|^{2p}$. For any $x \in \mathbb{R}^d$, $\nabla V(x) = 2p \|x\|^{2(p-1)} x$ and $\Delta V(x) = (4p(p-1) + 2pd) \|x\|^{2(p-1)}$. Therefore, using **B**3(m) and the definition of $\mathcal{A}$ we obtain that for any $x \in \mathbb{R}^d$

$$\mathcal{A}V(x) \leq [2p(p-1) + p(d - 2\mathtt{m})] V(x) \, . \tag{4.97}$$

Hence, using (4.97) and [Kha11, Theorem 3.5], we obtain that **L**1 holds.

(a) If there exists $\varepsilon_b > 0$ such that $\sup_{x \in \mathbb{R}^d} \|b(x)\|^{2(1+\varepsilon_b)} (1 + \|x\|^{2p})^{-1} < +\infty$, using (4.97) and Lemma 4.1.25-(b) we obtain that **L**3 holds.

(b) If there exists $\varepsilon_b > 0$ such that $\sup_{x \in \mathbb{R}^d} \|b(x)\|^{2(1+\varepsilon_b)} \mathrm{e}^{-\mathtt{m}_2^+ \|x\|^2} < +\infty$, and **C**2 holds, then consider for any $x \in \mathbb{R}^d$, $V(x) = \mathrm{e}^{\mathtt{m}_2^+ \|x\|^2}$. We have for any $x \in \mathbb{R}^d$, $\nabla V(x) = 2\mathtt{m}_2^+ \mathrm{e}^{\mathtt{m}_2^+ \|x\|^2} x$ and $\Delta V(x) = 4\mathtt{m}_2^{+2} \mathrm{e}^{\mathtt{m}_2^+ \|x\|^2} \|x\|^2 + 2\mathtt{m}_2^+ \mathrm{e}^{\mathtt{m}_2^+ \|x\|^2} d$. Therefore, using **C**2 we have for any $x \in \bar{\mathrm{B}}(0, R_2)^{\mathrm{c}}$

$$\mathcal{A}V(x) \leq \mathtt{m}_2^+ \left[d + (4\mathtt{m}_2^+/2 - 2\mathtt{m}_2^+) \|x\|^2 \right] V(x) \leq \mathtt{m}_2^+ d V(x) \, . \tag{4.98}$$

Setting $\zeta = (\mathtt{m}_2^+ d) \vee \sup_{x \in \bar{\mathrm{B}}(0,R_2)} \mathcal{A}V(x)$, we obtain that V satisfies $\mathbf{D}_{\mathrm{c}}(V, \zeta, 0)$. Therefore using (4.98) and Lemma 4.1.25-(b), we obtain that **L**3 holds.

Proof of Proposition 4.1.31 We preface the proof by a preliminary computation. Let $n \in \mathbb{N}$, $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$ and $X = x + \gamma \mathrm{b}_{\gamma,n}(x) + \sqrt{\gamma} Z$, where Z is a d-dimensional Gaussian random variable with zero mean and covariance identity. We have using **B**3(m) and (4.63)

$$\mathbb{E}\left[\|X\|^2 \right] \leq \|x\|^2 - 2\gamma \mathtt{m} \, \Phi_n(x) \|x\|^2 + \gamma^2 \Phi_n(x)^2 \|b(x)\|^2 + \gamma d \, , \tag{4.99}$$

with $\Phi_n(x) = \varphi_n(x) + (1 - \varphi_n(x))(1 + \gamma^\alpha \|b(x)\|)^{-1}$. We recall that

$$\varphi_n(x) \in [0, 1] \quad \text{and} \quad \varphi_n(x) = \begin{cases} 1 & \text{if } x \in \bar{\mathrm{B}}(0, n), \\ 0 & \text{if } x \in \bar{\mathrm{B}}(0, n+1)^{\mathrm{c}} \end{cases} \, . \tag{4.100}$$

Using **B**5 and (4.100), we have

$$\Phi_n(x) \|b(x)\| \leq \mathrm{L}_{n+1} \|x\| + \gamma^{-\alpha} \, . \tag{4.101}$$

Combining (4.99) and (4.101) and since $\Phi_n(x) \leq 1$ by (4.100), we obtain

$$\mathbb{E}\left[1 + \|X\|^2 \right] \leq (1 + \|x\|^2) \left[1 + 2\gamma |\mathtt{m}| + 2\gamma^2 \mathrm{L}_{n+1}^2 \right] + 2\gamma^{2-2\alpha} + \gamma d \, . \tag{4.102}$$

We are now able to complete the proof of Proposition 4.1.31. It is easy to check that **L**2 holds with $\beta = 2\alpha$. It only remains to show that **L**5 holds. Consider for any $x \in \mathbb{R}^d$, $\tilde{V}(x) = 1 + \|x\|$. By (4.102), for any $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$, we have using for any $s \geq \mathbb{R}$, $1 + s \leq \mathrm{e}^s$ we obtain

$$\begin{aligned} \mathrm{R}_{\gamma,n} \tilde{V}(x) &\leq \tilde{V}(x) \left[1 + 2\gamma |\mathtt{m}| + 2\gamma^2 \mathrm{L}_{n+1}^2 + 2\gamma^{2-2\alpha} + \gamma d \right] \\ &\leq \tilde{V}(x) \exp\left[\gamma \left\{ 2 |\mathtt{m}| + d + 2\gamma^{1-2\alpha} (\gamma^{2\alpha} \mathrm{L}_{n+1} + 1) \right\} \right] \\ &\leq \tilde{V}(x) \exp\left[2\gamma \left\{ 2 |\mathtt{m}| + d \right\} \left\{ 1 + \gamma^{1-2\alpha} (\gamma^{2\alpha} \mathrm{L}_{n+1} + 1) \right\} \right] \, . \end{aligned}$$

As a result using that $d \geq 1$, **L**5 holds upon taking $\tilde{A} = \exp(4 |\mathtt{m}| + 2d)$, $\tilde{\varepsilon}_n = 1 - 2\alpha$ and $\tilde{E}_n = 2(\mathrm{L}_{n+1} \bar{\gamma}^{2\alpha} + 1)$.

Proof of Proposition 4.1.33 The proof is similar to the one of Proposition 4.1.34 upon replacing (4.102) by

$$\mathbb{E}\left[1+\|X\|^2\right] \leq (1+\|x\|^2)(1+2\gamma \mathrm{L}+2\gamma^2\mathrm{L}^2)+\gamma d \, .$$

Proof of Proposition 4.1.34 Let $M \geq 0$, $n \in \mathbb{N}$ and $p \geq 1$. Using the Log-Sobolev inequality [BLM13, Theorem 5.5], the fact that ϕ is 1-Lipschitz and that $\Pi_{\bar{\mathrm{B}}(0,n)}$ is non expansive, as well as the Jensen inequality we obtain for any $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$,

$$\begin{aligned}
\mathrm{R}_{\gamma,n} V_M^p(x) &\leq \exp\left[pM\tilde{\mathrm{R}}_{\gamma,n}\phi(x) + (pM)^2\gamma/2\right] \\
&\leq \exp\left[pM\sqrt{\tilde{\mathrm{R}}_{\gamma,n}\phi^2(x)} + (pM)^2\gamma/2\right] \, .
\end{aligned}$$

Using (4.102) and that $\sqrt{1+t} \leq 1+t/2$ for any $t \in (-1, +\infty)$ we get for any $\gamma \in (0, \bar{\gamma}]$ and $x \in \bar{\mathrm{B}}(0,n)$

$$\begin{aligned}
&\mathrm{R}_{\gamma,n} V_M^p(x) \\
&\leq \exp\left[pM\left\{\phi(x)^2(1+2\gamma\left|\mathtt{m}\right| + 2\gamma^2\mathrm{L}_{n+1}^2) + 2\gamma^{2-2\alpha} + \gamma d\right\}^{1/2} + (pM)^2\gamma/2\right] \\
&\leq \exp\left[(1+\gamma\left|\mathtt{m}\right| + \gamma^2\mathrm{L}_{n+1}^2)pM\phi(x)\right] \exp\left[(1+pM)^2\left\{\gamma(d+1)/2 + \gamma^{2-2a}\right\}\right] \\
&\leq V_M^{p(1+C_1\gamma+C_{2,n}\gamma^2)}(x) \exp\left[p^2 C_3\gamma\right] \, ,
\end{aligned}$$

with $C_1 = \left|\mathtt{m}\right|$, $C_{2,n} = \mathrm{L}_{n+1}^2$ and $C_3 = (1+M)^2(d+3)/2$. By recursion, we obtain that for any $m, n \in \mathbb{N}$ with $m^{-1} \in (0, \bar{\gamma}]$, $T \geq 0$ and $x \in \bar{\mathrm{B}}(0,n)$

$$\begin{aligned}
&\mathrm{R}_{T/m,n}^m V_M(x) \\
&\leq V_M(x)^{a_m} \exp\left[TC_3 \sum_{j=0}^{m-1} (1 + TC_1/m + C_{2,n}(T/m)^2)^{2j}/m\right] \\
&\leq V_M(x)^{a_m} \exp\left[TC_3(1 + TC_1/m + C_{2,n}(T/m)^2)^{2m}\right] \, ,
\end{aligned}$$

with $a_m = (1 + TC_1/m + C_{2,n}(T/m)^2)^m$. Since $\lim_{m\to+\infty}(1 + TC_1/m + C_{2,n}(T/m)^2)^{tm} = \exp(tTC_1)$ for any $t, T \geq 0$, we get that for any $n \in \mathbb{N}$, $T \geq 0$ and $x \in \bar{\mathrm{B}}(0,n)$

$$\limsup_{m\to+\infty} \mathrm{R}_{T/m,n}^m V_M(x) \leq \exp(TC_3\exp(2TC_1)) V_M^{\exp(TC_1)}(x) \, . \tag{4.103}$$

We conclude the proof upon remarking that the right-hand side quantity in (4.103) does not depend on n and that the same inequality holds replacing $\mathrm{R}_{T/m,n}$ by $\tilde{\mathrm{R}}_{T/m,n}$ in (4.103).

Proof of Proposition 4.1.35 We have for any $x \in \mathbb{R}^d$,

$$\nabla\phi(x) = x/\phi(x) \, , \qquad \nabla^2\phi(x) = \mathrm{Id}\,/\phi(x) - xx^\top/\phi^2(x) \, ,$$

and therefore since $V_M(x) = \exp(M\phi(x))$,

$$\begin{aligned}
\nabla V_M(x) &= M\nabla\phi(x) V_M(x) \, , \\
\nabla^2 V_M(x) &= \left\{M^2\nabla\phi(x)(\nabla\phi(x))^\top + M\nabla^2\phi(x)\right\} V_M(x) \, .
\end{aligned}$$

Therefore, for any $x \in \mathbb{R}^d$,

$$(\mathcal{A}V_M(x))/V_M(x) \leq \left[M^2 \|x\|^2 / \phi^2(x) + M\left\{d/\phi(x) - \|x\|^2/\phi^2(x)\right\}\right]\Big/ 2 + M \sup_{x\in\mathbb{R}^d} \langle b(x), x\rangle_+ .$$

Hence, for any $x \in \mathbb{R}^d$, $\mathcal{A}V_M(x) \leq \zeta V_M(x)$ with $\zeta = M\{\sup_{x\in\mathbb{R}^d}\langle b(x), x\rangle_+ + d/2\} + M^2$. We conclude using Lemma 4.1.25-(a) and the Doob maximal inequality.

Minorization conditions for functional autoregressive models

In this section, we extend and complete the results of [DM19, Section 6] on functional autoregressive models. Let $\mathsf{X} \in \mathcal{B}(\mathbb{R}^d)$ equipped with its trace σ-field $\mathcal{X} = \{\mathsf{A}\cap\mathsf{X} : \mathsf{A} \in \mathcal{B}(\mathbb{R}^d)\}$. In fact, we consider a slightly more general class of models than [DM19] which is associated with non-homogeneous Markov chains $(X_k^{(\mathrm{a})})_{k\in\mathbb{N}}$ with state space $(\mathsf{X}, \mathcal{X})$ defined for $k \geq 0$ by

$$X_{k+1}^{(\mathrm{a})} = \Pi\left(\mathcal{T}_{k+1}(X_k^{(\mathrm{a})}) + \sigma_{k+1}Z_{k+1}\right) ,$$

where Π is a measurable function from $\mathbb{R}^d$ to X, $(\mathcal{T}_k)_{k\geq 1}$ is a sequence of measurable functions from X to $\mathbb{R}^d$, $(\sigma_k)_{k\geq 1}$ is a sequence of positive real numbers and $(Z_k)_{k\geq 1}$ is a sequence of i.i.d. d dimensional standard Gaussian random variables. We assume that Π satisfies **A**1. We also assume some Lipschitz regularity on the operator $\mathcal{T}_k$ for any $k \in \mathbb{N}^\star$

AR1 (A). *For all $k \geq 1$ there exists $\varpi_k \in \mathbb{R}$ such that for all $(x, y) \in \mathsf{A}$,*

$$\|\mathcal{T}_k(x) - \mathcal{T}_k(y)\|^2 \leq (1 + \varpi_k)\|x - y\|^2 .$$

The sequence $\{X_k^{(\mathrm{a})},\ k \in \mathbb{N}\}$ is an inhomogeneous Markov chain associated with the family of Markov kernels $(P_k^{(\mathrm{a})})_{k\geq 1}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ given for all $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathbb{R}^d$ by

$$P_k^{(\mathrm{a})}(x, \mathsf{A}) = \frac{1}{(2\pi\sigma_k^2)^{d/2}} \int_{\Pi^{-1}(\mathsf{A})} \exp\left(-\|y - \mathcal{T}_k(x)\|^2/(2\sigma_k^2)\right) \mathrm{d}y .$$

We denote for all $n \geq 1$ by $Q_n^{(\mathrm{a})}$ the marginal distribution of $X_n^{(\mathrm{a})}$ given by $Q_n^{(\mathrm{a})} = P_1^{(\mathrm{a})} \cdots P_n^{(\mathrm{a})}$. To obtain an upper bound of $\|\delta_x Q_n^{(\mathrm{a})} - \delta_y Q_n^{(\mathrm{a})}\|_{\mathrm{TV}}$ for any $x, y \in \mathbb{R}^d$, $n \in \mathbb{N}^\star$, we introduce a Markov coupling $(X_k^{(\mathrm{a})}, Y_k^{(\mathrm{a})})_{k\in\mathbb{N}}$ such that for any $n \in \mathbb{N}^\star$, the distribution of $X_n^{(\mathrm{a})}$ and $Y_n^{(\mathrm{a})}$ are $\delta_x Q_n^{(\mathrm{a})}$ and $\delta_x Q_n^{(\mathrm{a})}$ respectively, exactly as we have introduced in the homogeneous setting the Markov coupling with kernel K_γ defined by (4.23) for R_γ defined in (4.20). For completeness and readability, we recall the construction of $(X_k^{(\mathrm{a})}, Y_k^{(\mathrm{a})})_{k\in\mathbb{N}}$. For all $k \in \mathbb{N}^\star$ and $x, y, z \in \mathbb{R}^d$, define

$$\mathrm{e}_k(x, y) = \begin{cases} \mathrm{E}_k(x, y)/\|\mathrm{E}_k(x, y)\| & \text{if } \mathrm{E}_k(x, y) \neq 0 \\ 0 & \text{otherwise} \end{cases} , \qquad \mathrm{E}_k(x, y) = \mathcal{T}_k(y) - \mathcal{T}_k(x) , \tag{4.104}$$

$$\begin{aligned} \mathcal{S}_k(x, y, z) &= \mathcal{T}_k(y) + (\mathrm{Id} - 2\mathrm{e}_k(x, y)\mathrm{e}_k(x, y)^\top)z , \\ p_k(x, y, z) &= 1 \wedge \frac{\boldsymbol{\varphi}_{\sigma_{k+1}^2}(\|\mathrm{E}_k(x, y)\| - \langle \mathrm{e}_k(x, y), z\rangle)}{\boldsymbol{\varphi}_{\sigma_{k+1}^2}(\langle \mathrm{e}_k(x, y), z\rangle)} , \end{aligned} \tag{4.105}$$

where $\varphi_{\sigma_k^2}$ is the one-dimensional zero mean Gaussian distribution function with variance σ_k^2. Let $(U_k)_{k\in\mathbb{N}^\star}$ be a sequence of i.i.d. uniform random variables on $[0,1]$ and define the Markov chain $(X_k^{(\mathrm{a})}, Y_k^{(\mathrm{a})})_{k\in\mathbb{N}}$ starting from $(X_0^{(\mathrm{a})}, Y_0^{(\mathrm{a})}) \in \mathsf{X}^2$ by the recursion: for any $k \geq 0$

$$\begin{aligned}
&\tilde{X}_{k+1}^{(\mathrm{a})} = \mathcal{T}_{k+1}(X_k^{(\mathrm{a})}) + \sigma_{k+1} Z_{k+1} \;, \\
&\tilde{Y}_{k+1}^{(\mathrm{a})} = \begin{cases} \tilde{X}_{k+1}^{(\mathrm{a})} & \text{if } \mathcal{T}_{k+1}(X_k^{(\mathrm{a})}) = \mathcal{T}_{k+1}(Y_k^{(\mathrm{a})}) \;; \\ W_{k+1}^{(\mathrm{a})} \tilde{X}_{k+1}^{(\mathrm{a})} + (1 - W_{k+1}^{(\mathrm{a})}) \mathcal{S}_{k+1}(X_k^{(\mathrm{a})}, Y_k^{(\mathrm{a})}, \sigma_{k+1} Z_{k+1}) & \text{otherwise} \;, \end{cases}
\end{aligned} \tag{4.106}$$

where $W_{k+1}^{(\mathrm{a})} = \mathbb{1}_{(-\infty,0]}(U_{k+1} - p_{k+1}(X_k^{(\mathrm{a})}, Y_k^{(\mathrm{a})}, \sigma_{k+1} Z_{k+1}))$ and finally set

$$(X_{k+1}^{(\mathrm{a})}, Y_{k+1}^{(\mathrm{a})}) = (\Pi(\tilde{X}_{k+1}^{(\mathrm{a})}), \Pi(\tilde{Y}_{k+1}^{(\mathrm{a})})) \;. \tag{4.107}$$

For any $k \in \mathbb{N}^\star$, marginally, the distribution of $X_{k+1}^{(\mathrm{a})}$ given $X_k^{(\mathrm{a})}$ is $P_{k+1}^{(\mathrm{a})}(X_k^{(\mathrm{a})}, \cdot)$, and it is well-know (see *e.g.* [BDJ98, Section 3.3]) that $\tilde{Y}_{k+1}^{(\mathrm{a})}$ and $\mathcal{T}_\gamma(Y_k^{(\mathrm{a})}) + \sigma_{k+1} Z_{k+1}$ have the same distribution given Y_k, and therefore the distribution of Y_{k+1} given Y_k is $P_{k+1}^{(\mathrm{a})}(Y_k, \cdot)$. As a result for any $(x,y) \in \mathsf{X}^2$ and $n \in \mathbb{N}^\star$, $(X_n^{(\mathrm{a})}, Y_n^{(\mathrm{a})})$ with $(X_0^{(\mathrm{a})}, Y_0^{(\mathrm{a})}) = (x,y)$ is a coupling between $\delta_x Q_n^{(\mathrm{a})}$ and $\delta_y Q_n^{(\mathrm{a})}$. Therefore, we obtain that $\|\delta_x Q_n^{(\mathrm{a})} - \delta_y Q_n^{(\mathrm{a})}\|_{\mathrm{TV}} \leq \mathbb{P}(X_n^{(\mathrm{a})} \neq Y_n^{(\mathrm{a})})$. Therefore to get an upper bound on $\|\delta_x Q_n^{(\mathrm{a})} - \delta_y Q_n^{(\mathrm{a})}\|_{\mathrm{TV}}$, it is sufficient to obtain a bound on $\mathbb{P}(X_n^{(\mathrm{a})} \neq Y_n^{(\mathrm{a})})$ which is a simple consequence of the following more general result.

Theorem 4.1.43. *Let $\mathsf{A} \in \mathcal{B}(\mathbb{R}^{2d})$ and assume* **A**1 *and* **AR**1(A). *Let $(X_k^{(\mathrm{a})}, Y_k^{(\mathrm{a})})_{k\in\mathbb{N}}$ be defined by* (4.107), *with $(X_0^{(\mathrm{a})}, Y_0^{(\mathrm{a})}) = (x,y) \in \mathsf{A}$. Then for any $n \in \mathbb{N}^\star$,*

$$\begin{aligned}
&\mathbb{P}\left[X_n^{(\mathrm{a})} \neq Y_n^{(\mathrm{a})} \text{ and for any } k \in \{1, \ldots, n-1\}, (X_k^{(\mathrm{a})}, Y_k^{(\mathrm{a})}) \in \mathsf{A}^2\right] \\
&\qquad\qquad \leq \mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(x,y) \left\{1 - 2\mathbf{\Phi}\left(-\frac{\|x-y\|}{2(\Xi_n^{(\mathrm{a})})^{1/2}}\right)\right\} \;,
\end{aligned}$$

where $\mathbf{\Phi}$ is the cumulative distribution function of the standard normal distribution on $\mathbb{R}$ and $(\Xi_i^{(\mathrm{a})})_{i\geq 1}$ is defined for all $k \geq 1$ by $\Xi_k^{(\mathrm{a})} = \sum_{i=1}^k \{\sigma_i^2 / \prod_{j=1}^i (1+\varpi_j)\}$.

Proof. Let $(\mathcal{F}_k^{(\mathrm{a})})_{k\in\mathbb{N}}$ be the filtration associated to $(X_k^{(\mathrm{a})}, Y_k^{(\mathrm{a})})_{k\in\mathbb{N}}$. Denote for any $k \in \mathbb{N}$,

$$\mathscr{A}_k = \bigcap_{i=0}^k \{(X_i^{(\mathrm{a})}, Y_i^{(\mathrm{a})}) \in \mathsf{A}\} \;, \qquad \mathscr{A}_{-1} = \mathscr{A}_0 \;,$$

and for all $k_1, k_2 \in \mathbb{N}^*$, $k_1 \leq k_2$, $\Xi_{k_1,k_2}^{(\mathrm{a})} = \sum_{i=k_1}^{k_2} \{\sigma_i^2 / \prod_{j=k_1}^i (1+\varpi_j)\}$. Let $n \geq 1$ and $(x,y) \in \mathsf{A}^2$. We show by backward induction that for all $k \in \{0, \cdots, n-1\}$,

$$\mathbb{P}(\{X_n^{(\mathrm{a})} \neq Y_n^{(\mathrm{a})}\} \cap \mathscr{A}_{n-1}) \leq \mathbb{E}\left[\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(X_k^{(\mathrm{a})}, Y_k^{(\mathrm{a})}) \mathbb{1}_{\mathscr{A}_{k-1}} \left[1 - 2\mathbf{\Phi}\left\{-\frac{\left\|X_k^{(\mathrm{a})} - Y_k^{(\mathrm{a})}\right\|}{2\left(\Xi_{k+1,n}^{(\mathrm{a})}\right)^{1/2}}\right\}\right]\right] \;. \tag{4.108}$$

Note that the inequality for $k = 0$ will conclude the proof. Using by (4.106) that $\tilde{X}_n^{(\mathrm{a})} = \tilde{Y}_n^{(\mathrm{a})}$ if $X_{n-1}^{(\mathrm{a})} = Y_{n-1}^{(\mathrm{a})}$ or $W_n = \mathbb{1}_{(-\infty,0]}(U_n - p_n(X_{n-1}^{(\mathrm{a})}, Y_{n-1}^{(\mathrm{a})}, \sigma_n Z_n)) = 1$, where p_n is defined by (4.105), and (U_n, Z_n) are independent random variables independent of $\mathcal{F}_{n-1}^{(\mathrm{a})}$, we obtain on $\{X_{n-1}^{(\mathrm{a})} \neq Y_{n-1}^{(\mathrm{a})}\}$

$$\mathbb{E}\left[\mathbb{1}_{\Delta_{\mathsf{X}}}(\tilde{X}_n^{(\mathrm{a})},\tilde{Y}_n^{(\mathrm{a})})\middle|\mathcal{F}_{n-1}^{(\mathrm{a})}\right]=\mathbb{E}\left[p_n(X_{n-1}^{(\mathrm{a})},Y_{n-1}^{(\mathrm{a})},\sigma_n Z_n)\middle|\mathcal{F}_{n-1}^{(\mathrm{a})}\right]$$
$$=2\Phi\left\{-\left\|(2\sigma_n)^{-1}\mathrm{E}_n(X_{n-1}^{(\mathrm{a})},Y_{n-1}^{(\mathrm{a})})\right\|\right\}\ .$$

Since $\{X_n^{(\mathrm{a})}\neq Y_n^{(\mathrm{a})}\}\subset\{\tilde{X}_n^{(\mathrm{a})}\neq\tilde{Y}_n^{(\mathrm{a})}\}\subset\{X_{n-1}^{(\mathrm{a})}\neq Y_{n-1}^{(\mathrm{a})}\}$ by (4.107) and (4.106), we get

$$\mathbb{P}\left[\{X_n^{(\mathrm{a})}\neq Y_n^{(\mathrm{a})}\}\cap\mathscr{A}_{n-1}\right]\leq\mathbb{E}\left[\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(X_{n-1}^{(\mathrm{a})},Y_{n-1}^{(\mathrm{a})})\mathbb{1}_{\mathscr{A}_{n-1}}\mathbb{E}\left[\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(\tilde{X}_n^{(\mathrm{a})},\tilde{Y}_n^{(\mathrm{a})})\middle|\mathcal{F}_{n-1}^{(\mathrm{a})}\right]\right]$$
$$=\mathbb{E}\left[\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(X_{n-1}^{(\mathrm{a})},Y_{n-1}^{(\mathrm{a})})\mathbb{1}_{\mathscr{A}_{n-1}}\left[1-2\Phi\left\{-\left\|(2\sigma_n)^{-1}\mathrm{E}_n(X_{n-1}^{(\mathrm{a})},Y_{n-1}^{(\mathrm{a})})\right\|\right\}\right]\right]\ ,$$

Using that $(X_{n-1}^{(\mathrm{a})},Y_{n-1}^{(\mathrm{a})})\in\mathsf{A}^2$ on $\mathscr{A}_{n-1}$, **AR**1(A) and (4.104), we obtain that

$$\|\mathrm{E}_n(X_{n-1}^{(\mathrm{a})},Y_{n-1}^{(\mathrm{a})})\|^2\leq(1+\varpi_n)\|X_{n-1}^{(\mathrm{a})}-Y_{n-1}^{(\mathrm{a})}\|^2\ ,$$

showing (4.108) holds for $k=n-1$ since $\mathscr{A}_{n-2}\subset\mathscr{A}_{n-1}$. Assume that (4.108) holds for $k\in\{1,\ldots,n-1\}$. On $\{\tilde{X}_k^{(\mathrm{a})}\neq\tilde{Y}_k^{(\mathrm{a})}\}$, we have

$$\left\|\tilde{X}_k^{(\mathrm{a})}-\tilde{Y}_k^{(\mathrm{a})}\right\|=\left|-\left\|\mathrm{E}_k(X_{k-1}^{(\mathrm{a})},Y_{k-1}^{(\mathrm{a})})\right\|+2\sigma_k\mathrm{e}_k(X_{k-1}^{(\mathrm{a})},Y_{k-1}^{(\mathrm{a})})^{\mathrm{T}}Z_k\right|\ ,$$

which implies by (4.107) and since Π is non expansive by **A**1

$$\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(X_k^{(\mathrm{a})},Y_k^{(\mathrm{a})})\left[1-2\Phi\left\{-\frac{\left\|X_k^{(\mathrm{a})}-Y_k^{(\mathrm{a})}\right\|}{2(\Xi_{k+1,n}^{(\mathrm{a})})^{1/2}}\right\}\right]$$
$$\leq\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(X_k^{(\mathrm{a})},Y_k^{(\mathrm{a})})\left[1-2\Phi\left\{-\frac{\left\|\tilde{X}_k^{(\mathrm{a})}-\tilde{Y}_k^{(\mathrm{a})}\right\|}{2(\Xi_{k+1,n}^{(\mathrm{a})})^{1/2}}\right\}\right]$$
$$\leq\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(X_k^{(\mathrm{a})},Y_k^{(\mathrm{a})})\left[1-2\Phi\left\{-\frac{\left|2\sigma_k\mathrm{e}_k(X_{k-1}^{(\mathrm{a})},Y_{k-1}^{(\mathrm{a})})^{\mathrm{T}}Z_k-\left\|\mathrm{E}_k(X_{k-1}^{(\mathrm{a})},Y_{k-1}^{(\mathrm{a})})\right\|\right|}{2(\Xi_{k+1,n}^{(\mathrm{a})})^{1/2}}\right\}\right]\ .$$

Since Z_k is independent of $\mathcal{F}_k^{(\mathrm{a})}$, $\sigma_k\mathrm{e}_k(X_{k-1}^{(\mathrm{a})},X_{k-1}^{(\mathrm{a})})^{\mathrm{T}}Z_k$ is a real Gaussian random variable with zero mean and variance σ_k^2. Therefore by [DM19, Lemma 20] and since $\mathscr{A}_{k-1}$ is $\mathcal{F}_{k-1}^{(\mathrm{a})}$-measurable, we get

$$\mathbb{E}\left[\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(X_k^{(\mathrm{a})},Y_k^{(\mathrm{a})})\mathbb{1}_{\mathscr{A}_{k-1}}\left[1-2\Phi\left\{-\frac{\left\|X_k^{(\mathrm{a})}-Y_k^{(\mathrm{a})}\right\|}{2(\Xi_{k+1,n}^{(\mathrm{a})})^{1/2}}\right\}\right]\middle|\mathcal{F}_{k-1}^{(\mathrm{a})}\right]$$
$$\leq\mathbb{1}_{\mathscr{A}_{k-1}}\mathbb{1}_{\Delta_{\mathsf{X}}^{\mathrm{c}}}(X_{k-1}^{(\mathrm{a})},Y_{k-1}^{(\mathrm{a})})\left[1-2\Phi\left\{-\frac{\left\|\mathrm{E}_k(X_{k-1}^{(\mathrm{a})},Y_{k-1}^{(\mathrm{a})})\right\|}{2\left(\sigma_k+\Xi_{k+1,n}^{(\mathrm{a})}\right)^{1/2}}\right\}\right]\ .$$

Since **A**2(A) implies that $\|\mathrm{E}_k(X_{k-1}^{(\mathrm{a})},Y_{k-1}^{(\mathrm{a})})\|^2\leq(1+\varpi_{k-1})\left\|X_{k-1}^{(\mathrm{a})}-Y_{k-1}^{(\mathrm{a})}\right\|^2$ on $\mathscr{A}_{k-1}$ and $\mathscr{A}_{k-2}\subset\mathscr{A}_{k-1}$ concludes the induction of (4.108). □

Quantitative convergence results based on [Dou+18; DMR04]

We start by recalling the following lemma from [Dou+18] which is inspired from the results of [DMR04].

Lemma 4.1.44 ([Dou+18, Lemma 19.4.2]). *Let $(\mathsf{Y}, \mathcal{Y})$ be a measurable space and R be a Markov kernel over $(\mathsf{Y}, \mathcal{Y})$. Let Q be a Markov coupling kernel for R. Assume there exist $\mathsf{C} \in \mathcal{Y}^{\otimes 2}$, $M \geq 0$, a measurable function $W : \mathsf{Y} \times \mathsf{Y} \to [1, +\infty)$, $\lambda \in [0, 1)$ and $c \geq 0$ such that for any $x, y \in \mathsf{Y}$,*

$$\mathrm{Q}W(x, y) \leq \lambda W(x, y) \mathbb{1}_{\mathsf{C}^c}(x, y) + c \mathbb{1}_{\mathsf{C}}(x, y) \, .$$

In addition, assume that there exists $\varepsilon > 0$ such that for any $(x, y) \in \mathsf{C}$,

$$\mathrm{Q}((x, y), \Delta_{\mathsf{Y}}^{\mathrm{c}}) \leq 1 - \varepsilon \, ,$$

where $\Delta_{\mathsf{Y}} = \{(y, y) : y \in \mathsf{Y}\}$. Then there exist $\rho \in [0, 1)$ and $C \geq 0$ such that for any $x, y \in \mathcal{Y}$ and $n \in \mathbb{N}^\star$

$$\int_{\mathsf{Y} \times \mathsf{Y}} \mathbb{1}_{\Delta_{\mathsf{Y}}}(\tilde{x}, \tilde{y}) W(\tilde{x}, \tilde{y}) \mathrm{Q}^n((x, y), \mathrm{d}(\tilde{x}, \tilde{y})) \leq C \rho^n W(x, y) \, ,$$

where

$$\begin{aligned} C &= 2(1 + c/\{(1 - \varepsilon)(1 - \lambda)\}) \, , \\ \log(\rho) &= \{\log(1 - \varepsilon) \log(\lambda)\}/\{\log(1 - \varepsilon) + \log(\lambda) - \log(c)\} \, . \end{aligned}$$

Theorem 4.1.45. *Let $(\mathsf{Y}, \mathcal{Y})$ be a measurable space and R be a Markov kernel over $(\mathsf{Y}, \mathcal{Y})$. Let Q be a Markov coupling kernel of R. Assume that there exist $\lambda \in [0, 1)$, $A \geq 0$ and a measurable function $W : \mathsf{Y} \times \mathsf{Y} \to [1, +\infty)$, such that Q satisfies $\mathbf{D}_{\mathrm{d}}(W, \lambda, A, \mathsf{Y})$. In addition, assume that there exist $\ell \in \mathbb{N}^\star$, $\varepsilon > 0$ and $M \geq 1$ such that for any $(x, y) \in \mathsf{C}_M = \{(x, y) \in \mathsf{Y} \times \mathsf{Y}, \ W(x, y) \leq M\}$,*

$$\mathrm{Q}^\ell((x, y), \Delta_{\mathsf{Y}}^{\mathrm{c}}) \leq 1 - \varepsilon \, , \tag{4.109}$$

with $\Delta_{\mathsf{Y}} = \{(x, y) \in \mathsf{Y}^2 : x = y\}$ and $M \geq 2A/(1 - \lambda)$. Then, there exist $\rho \in [0, 1)$ and $C \geq 0$ such that for any $n \in \mathbb{N}$ and $x, y \in \mathsf{Y}$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}^n, \delta_y \mathrm{R}^n) \leq C \rho^{\lfloor n/\ell \rfloor} W(x, y) \, ,$$

with

$$\begin{aligned} C &= 2(1 + A_\ell)(1 + c_\ell/\{(1 - \varepsilon)(1 - \lambda_\ell)\}) \, , \\ \lambda_\ell &= (\lambda^\ell + 1)/2 \, , \quad c_\ell = \lambda^\ell M + A_\ell \, , \quad A_\ell = A(1 - \lambda^\ell)/(1 - \lambda) \, , \\ \log(\rho_\ell) &= \{\log(1 - \varepsilon) \log(\lambda_\ell)\}/\{\log(1 - \varepsilon) + \log(\lambda_\ell) - \log(c_\ell)\} \, . \end{aligned} \tag{4.110}$$

Proof. We first show that for any $(x, y) \in \mathsf{C}_M$,

$$\mathrm{Q}^\ell(x, y) \leq \lambda_\ell W(x, y) \mathbb{1}_{\mathsf{C}_M^{\mathrm{c}}}(x, y) + c_\ell \mathbb{1}_{\mathsf{C}_M}(x, y) \, , \tag{4.111}$$

in order to apply Lemma 4.1.44 to R^ℓ with the Markov coupling kernel Q^ℓ. By a straightforward induction, for any $x, y \in \mathsf{Y}$ we have

$$\mathrm{Q}^\ell W(x, y) \leq \lambda^\ell W(x, y) + A(1 - \lambda^\ell)/(1 - \lambda) \, . \tag{4.112}$$

We distinguish two cases. If $(x, y) \notin \mathsf{C}_M$, using that $A/M \geq (1 - \lambda)/2$ we have that

$$\mathrm{Q}^\ell W(x, y) \leq \lambda^\ell W(x, y) + A(1 - \lambda^\ell) W(x, y)/(M(1 - \lambda)) \leq 2^{-1}(\lambda^\ell + 1) W(x, y) \, .$$

If $(x, y) \in \mathsf{C}_M$, we have

$$\mathrm{Q}^\ell W(x, y) \leq \lambda^\ell M + A(1 - \lambda^\ell)/(1 - \lambda) \, .$$

Therefore (4.111) holds. As a result and since by assumption we have (4.109), we can apply Lemma 4.1.44 to R^ℓ. Then, we obtain that for any $i \in \mathbb{N}$ and $x, y \in \mathsf{Y}$

$$\int_{\mathsf{Y}\times\mathsf{Y}} \mathbb{1}_{\Delta_{\mathsf{Y}}}(\tilde{x},\tilde{y}) W(\tilde{x},\tilde{y}) \mathrm{Q}^{\ell i}((x,y),\mathrm{d}(\tilde{x},\tilde{y})) \leq C_\ell \rho_\ell^{\ell i} W(x,y) \;,$$

with ρ_ℓ defined by (4.110) and $\tilde{C}_\ell = 2\left\{1 + c_\ell \left[(1-\lambda_\ell)(1-\varepsilon)\right]^{-1}\right\}$. In addition, using (4.112), for any $k \in \{0,\ldots,\ell-1\}$ and $x, y \in \mathsf{Y}$, $\mathrm{Q}^k W(x,y) \leq (1+A_\ell) W(x,y)$. Therefore, for any $n \in \mathbb{N}$, since $n = i_n \ell + k_n$ with $i_n = \lfloor n/\ell \rfloor$ and $k_n \in \{0,\ldots,\ell-1\}$, we obtain for any $x, y \in \mathsf{Y}$ that

$$\begin{aligned}\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}^n, \delta_y \mathrm{R}^n) &\leq \tilde{C}_\ell \rho_\ell^{\ell i_n} \int_{\mathsf{Y}\times\mathsf{Y}} \mathbb{1}_{\Delta_{\mathsf{Y}}}(\tilde{x},\tilde{y}) W(\tilde{x},\tilde{y}) \mathrm{Q}^{k_n}((x,y),\mathrm{d}(\tilde{x},\tilde{y})) \\ &\leq (1+A_\ell)\tilde{C}_\ell \rho_\ell^{\lfloor n/\ell \rfloor} W(x,y) \;,\end{aligned}$$

which concludes the proof. □

We now state an important consequence of Theorem 4.1.45. The comparison between Theorem 4.1.46 and Theorem 4.1.8 is conducted in the remarks which follow Theorem 4.1.8.

Theorem 4.1.46. *Assume that there exists a measurable function $W : \mathsf{X}\times\mathsf{X} \to [1,+\infty)$ such that for any $C \geq 0$,*

$$\operatorname{diam}\left\{(x,y) \in \mathsf{X}^2 \,:\, W(x,y) \leq C\right\} < +\infty \;.$$

Assume in addition that there exist $\lambda \in [0,1)$, $A \geq 0$ such that for any $\gamma \in (0,\bar{\gamma}]$, there exists $\tilde{\mathrm{K}}_\gamma$, a Markov coupling kernel for R_γ, satisfying $\mathbf{D}_{\mathrm{d}}(W, \lambda^\gamma, A\gamma, \mathsf{X}^2)$. Further, assume that there exists $\boldsymbol{\Psi} : (0,\bar{\gamma}]\times\mathbb{N}^\star\times\mathbb{R}_+ \to [0,1]$ such that for any $\gamma \in (0,\bar{\gamma}]$, $\ell \in \mathbb{N}^\star$ and $x, y \in \mathsf{X}$, (4.32) is satisfied. Then the following results hold.

(a) For any $\gamma \in (0,\bar{\gamma}]$, $M_{\mathrm{d}} \geq \operatorname{diam}\left(\left\{(x,y) \in \mathsf{X}^2 \,:\, W(x,y) \leq B_{\mathrm{d}}\right\}\right)$ with $B_{\mathrm{d}} = 2A(1+\bar{\gamma})\{1 + \log^{-1}(1/\lambda)\}$, $\ell \in \mathbb{N}^\star$, $x, y \in \mathsf{X}$ and $k \in \mathbb{N}$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq C_\gamma \rho_\gamma^{\left\lfloor k(\ell\lceil 1/\gamma\rceil)^{-1}\right\rfloor} W(x,y) \;, \tag{4.113}$$

where $\mathbf{W}_{\mathbf{c}}$ is the Wasserstein metric associated with $\mathbf{c}$ defined by (4.30),

$$\begin{aligned} C_\gamma &= 2[1+A_\gamma][1 + c_\gamma/\{(1-\bar{\varepsilon}_{\mathrm{d},2})(1-\lambda_\gamma)\}] \;, \\ \log(\rho_\gamma) &= \{\log(1-\bar{\varepsilon}_{\mathrm{d},2})\log(\lambda_\gamma)\}/\{\log(1-\bar{\varepsilon}_{\mathrm{d},2}) + \log(\lambda_\gamma) - \log(c_\gamma)\} < 0 \;, \\ A_\gamma &= A\gamma(1-\lambda^{\gamma\ell\lceil 1/\gamma\rceil})/(1-\lambda^\gamma) \;, \quad c_\gamma = \lambda^{\gamma\ell\lceil 1/\gamma\rceil} A_\gamma + B_{\mathrm{d}} \;, \\ \bar{\varepsilon}_{\mathrm{d},2} &= \inf_{\gamma\in(0,\bar{\gamma}],\ (x,y)\in\Delta_{\mathsf{X},M_{\mathrm{d}}}} \boldsymbol{\Psi}(\gamma,\ell,\|x-y\|) \;, \quad \lambda_\gamma = (\lambda^{\gamma\ell\lceil 1/\gamma\rceil}+1)/2 \;. \end{aligned}$$

(b) For any $\gamma \in (0,\bar{\gamma}]$, $M_{\mathrm{d}} \geq \operatorname{diam}\left(\left\{(x,y) \in \mathsf{X}^2 \,:\, W(x,y) \leq B_{\mathrm{d}}\right\}\right)$ with $B_{\mathrm{d}} = 2A(1+\bar{\gamma})\{1 + \log^{-1}(1/\lambda)\}$ and $\ell \in \mathbb{N}^\star$, it holds that

$$\begin{aligned} C_\gamma &\leq \bar{C}_1 \;, \quad \log(\rho_\gamma) \leq \log(\bar{\rho}_2) \leq 0 \;, \\ \bar{C}_1 &= 2[1+\bar{A}_1][1 + \bar{c}_1/\left\{(1-\bar{\varepsilon}_{\mathrm{d},2})(1-\bar{\lambda}_1)\right\}] \;, \\ \log(\bar{\rho}_2) &= \left\{\log(1-\bar{\varepsilon}_{\mathrm{d},2})\log(\bar{\lambda}_1)\right\}/\left\{\log(1-\bar{\varepsilon}_{\mathrm{d},2}) + \log(\bar{\lambda}_1) - \log(\bar{c}_1)\right\} < 0 \;, \\ \bar{A}_1 &= A(1+\bar{\gamma})\min(\ell, 1+\log^{-1}(1/\lambda)) \;, \quad \bar{c}_1 = \bar{A}_1 + B_{\mathrm{d}} \;, \quad \bar{\lambda}_1 = (\lambda+1)/2 \;, \end{aligned}$$

(c) In addition, if $\bar{\gamma} \leq 1$, $-\log(\lambda) \in [0, \log(2)]$, $A \geq 1$ *and* $0 < \bar{\varepsilon}_{\mathrm{d},2} \leq 1 - \mathrm{e}^{-1}$, *then*

$$\log^{-1}(1/\bar{\rho}_2) \leq 12 \log(2) \log\left[6A\left\{1 + \log^{-1}(1/\lambda)\right\}\right] / (\log(1/\lambda)\bar{\varepsilon}_{\mathrm{d},\bar{\gamma}}) \ . \tag{4.114}$$

Proof. First, note that $1 - \lambda^t = -\int_0^t \log(\lambda)\mathrm{e}^{s\log(\lambda)}\mathrm{d}s \geq -\log(\lambda)\, t\, \mathrm{e}^{t\log(\lambda)}$ for any $t \in (0, \bar{t}]$, for $\bar{t} > 0$, and therefore

$$t/(1-\lambda^t) = t + t\lambda^t/(1-\lambda^t) \leq \bar{t} + \log^{-1}(\lambda^{-1}) \ . \tag{4.115}$$

(a) To establish (4.113), we apply Theorem 4.1.45. For any $x, y \in \mathsf{X}$ such that $W(x,y) \leq B_{\mathrm{d}}$ we have

$$\tilde{\mathrm{K}}_\gamma^{\ell\lceil 1/\gamma\rceil}((x,y), \Delta_{\mathsf{X}}^{\mathrm{c}}) \leq 1 - \bar{\varepsilon}_{\mathrm{d},2} \ .$$

Using that $\tilde{\mathrm{K}}_\gamma$ satisfies $\mathbf{D}_{\mathrm{d}}(W, \lambda^\gamma, A\gamma, \mathsf{X}^2)$, we can apply Theorem 4.1.45 with $M \leftarrow B_{\mathrm{d}} \geq 2A\gamma/(1-\lambda^\gamma)$ by (4.115), which completes the proof of (a).

(b) We now provide upper bounds for C_γ and ρ_γ. These constants are non-decreasing in c_γ and λ_γ. Therefore it suffices to give upper bounds on $c_\gamma, \varepsilon_{\mathrm{d},\gamma}$ and λ_γ. The result is then straightforward using that $(1-\lambda^{\gamma\ell\lceil 1/\gamma\rceil})/(1-\lambda^\gamma) \leq \ell\lceil 1/\gamma\rceil$, $\gamma(1-\lambda^{\gamma\ell\lceil 1/\gamma\rceil})/(1-\lambda^\gamma) \leq \bar{\gamma} + \log^{-1}(1/\lambda)$ and $\lambda^{\gamma\ell\lceil 1/\gamma\rceil} \leq \lambda$.

(c) By assumption on $\bar{\gamma}$, λ and $\bar{\varepsilon}_{\mathrm{d},1}$ we have that $\log((1-\bar{\varepsilon}_{\mathrm{d},2})^{-1}) \leq 1$ and

$$\log(\bar{\lambda}_1^{-1}) \leq \log(\lambda^{-1}) \leq \log(2) \ , \quad \mathrm{e} \leq 2(1 + 1/\log(2)) \leq B_{\mathrm{d}} \leq \bar{c}_1 \ .$$

As a result, we obtain that $\log(\bar{\lambda}_1^{-1})/\log(\bar{c}_1) \leq 1$, $\log((1-\bar{\varepsilon}_{\mathrm{d},2})^{-1})/\log(\bar{c}_1) \leq 1$. Therefore we have

$$\begin{aligned}\log^{-1}(1/\bar{\rho}_2) &= \left[\log(\bar{\lambda}_1^{-1}) + \log((1-\bar{\varepsilon}_{\mathrm{d},2})^{-1}) + \log(\bar{c}_1)\right] \\ &\qquad / \left[\log(\bar{\lambda}_1^{-1})\log((1-\bar{\varepsilon}_{\mathrm{d},2})^{-1})\right] \\ &\leq 3\log[6A(1+\log^{-1}(1/\lambda))] / \left[\log(\bar{\lambda}_1^{-1})\log((1-\bar{\varepsilon}_{\mathrm{d},2})^{-1})\right] \ .\end{aligned}$$

Using that $\log(1-t) \leq -t$ for any $t \in (0,1]$ and the definition of $\bar{\lambda}_1$, we obtain that

$$\log^{-1}(\bar{\rho}_2^{-1}) \leq 6\bar{\varepsilon}_{\mathrm{d},2}^{-1}(1-\lambda)^{-1}\log[6A(1+\log^{-1}(1/\lambda))] \ .$$

Finally, we get (4.114) using that for any $t \in [0, \log(2)]$, $1 - e^{-t} \geq (2\log(2))^{-1}t$.

□

Note that Theorem 4.1.46 gives an upper bound on the rate of convergence $\bar{\rho}_2$ in the worst case scenario for which the minorization constant $\bar{\varepsilon}_{\mathrm{d},2}$ is small and the constant λ in $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, A\gamma, \mathsf{X}^2)$ is close to one.

Some remarks are in order here concerning the bounds obtained in Theorem 4.1.46 and Theorem 4.1.8. Assume that $\ell = 1$, we will see in Section 4.1.4 that the leading term in the upper bound in Theorem 4.1.8, respectively Theorem 4.1.46, is given by $\log(A)/(\log(\lambda^{-1})\bar{\varepsilon}_{\mathrm{d},1})$, respectively $\log(A)/(\log(\lambda^{-1})\bar{\varepsilon}_{\mathrm{d},2})$. In addition, in our applications, $\bar{\varepsilon}_{\mathrm{d},1}$ is larger than $\bar{\varepsilon}_{\mathrm{d},2}$. Therefore, in these cases the bounds provided in Theorem 4.1.8 yield better rates than the ones in Theorem 4.1.46-(c). The main difference between the two results is that in the proof of Theorem 4.1.46 a drift condition on the *iterated* coupling kernel $\tilde{\mathrm{K}}_\gamma^{\lceil 1/\gamma\rceil}$ is required. However, even if such drift conditions can be derived from a drift condition on $\tilde{\mathrm{K}}_\gamma$, the constants obtained using this technique are not sharp in general. On the contrary, the proof of Theorem 4.1.8 uses the iterated minorization condition and a drift condition on the *original* coupling $\tilde{\mathrm{K}}_\gamma$.

Tamed Euler-Maruyama discretization

In this subsection we consider the following assumption.

T1. $\mathsf{X} = \mathbb{R}^d$ *and* $\Pi = \mathrm{Id}$ *and*

$$\mathcal{T}_\gamma(x) = x + \gamma b(x)/(1 + \gamma \|b(x)\|) \text{ for any } \gamma > 0 \text{ and } x \in \mathbb{R}^d \ .$$

Here, we focus on drift b which is no longer assumed to be Lipschitz. Therefore, the ergodicity results obtained in Section 4.1.4 no longer hold since the minorization condition we derived relied heavily on one-sided Lipschitz condition or Lipschitz regularity for b. We now consider the following assumption on b.

T2. *There exists* $\tilde{\mathrm{L}}, \tilde{\ell} \geq 0$ *such that for any* $x, y \in \mathbb{R}^d$

$$\|b(x) - b(y)\| \leq \tilde{\mathrm{L}}(1 + \|x\|^{\tilde{\ell}} + \|y\|^{\tilde{\ell}}) \|x - y\| \ .$$

In addition, assume that $b(0) = 0$ *and* $M_{\tilde{\ell}} = \sup_{x \in \mathbb{R}^d} (1 + \|x\|^{\tilde{\ell}})(1 + \|b(x)\|)^{-1} < +\infty$.

Proposition 4.1.47. *Assume* **T**1 *and* **T**2 *then* **A**2($\mathbb{R}^{2d}$)-(iii) *holds with* $\bar{\gamma} > 0$ *and for any* $\gamma \in (0, \bar{\gamma}]$, $\kappa(\gamma) = 2\tilde{\mathrm{L}}_\gamma + \gamma \tilde{\mathrm{L}}_\gamma^2$ *where*

$$\tilde{\mathrm{L}}_\gamma = 2\gamma^{-1} M_{\tilde{\ell}}(1 + M_{\tilde{\ell}})\tilde{\mathrm{L}} \ .$$

Proof. Let $x, y \in \mathbb{R}^d$ and assume that $\|x\| \geq \|y\|$. We have the following inequalities

$$\begin{aligned}
\left\| \frac{b(x)}{1 + \gamma \|b(x)\|} - \frac{b(y)}{1 + \gamma \|b(y)\|} \right\| &\leq \frac{\|b(x) - b(y)\|}{1 + \gamma \|b(x)\|} + \left| \frac{\|b(y)\|}{1 + \gamma \|b(x)\|} - \frac{\|b(y)\|}{1 + \gamma \|b(y)\|} \right| \\
&\leq \gamma^{-1} 2 M_{\tilde{\ell}} \tilde{\mathrm{L}} \|x - y\| + \gamma \frac{\|b(y)\| \, \|b(x) - b(y)\|}{(1 + \gamma \|b(x)\|)(1 + \gamma \|b(y)\|)} \\
&\leq \gamma^{-1} M_{\tilde{\ell}}(1 + M_{\tilde{\ell}}) \tilde{\mathrm{L}} \|x - y\| \ .
\end{aligned}$$

The same inequality holds with $\|y\| \geq \|x\|$. Therefore, we have

$$\|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\|^2 \leq \left(1 + 2\gamma \tilde{\mathrm{L}}_\gamma + \gamma^2 \tilde{\mathrm{L}}_\gamma^2\right) \|x - y\|^2 \ ,$$

which concludes the proof. □

Proposition 4.1.47 implies that the conclusions of Proposition 4.1.5-(c) hold. Note that contrary to the conclusion of Proposition 4.1.11, we do not get that $\sup_{\gamma \in (0, \bar{\gamma}]} \kappa(\gamma) < +\infty$. Hence we have for any $\tilde{\ell} \in \mathbb{N}^\star$, $\inf_{\gamma \in (0, \bar{\gamma}]} \alpha_+(\kappa, \gamma, \tilde{\ell}) = 0$.

T3. *There exist* $\tilde{R}$ *and* $\tilde{\mathfrak{m}}^+$ *such that for any* $x \in \bar{\mathrm{B}}(0, \tilde{R})^{\mathrm{c}}$,

$$\langle b(x), x \rangle \leq -\tilde{\mathfrak{m}}^+ \|x\| \, \|b(x)\| \ .$$

Under **T**2 and **T**3 it is shown in [Bro+19] that there exists $\bar{\gamma} > 0$, $\lambda \in (0, 1)$ and $A \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, R_γ satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, A\gamma, \mathsf{X})$ with $V(x) = \exp(a(1 + \|x\|^2)^{1/2})$ for some fixed a.

Theorem 4.1.48. *Assume* **T**2 *and* **T**3 *then there exists* $\bar{\gamma} > 0$ *such that for any* $\gamma \in (0, \bar{\gamma}]$ *there exist* $C_\gamma \geq 0$ *and* $\rho_\gamma \in (0, 1)$ *with for any* $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$

$$\|\delta_x \mathrm{R}_\gamma^k - \delta_y \mathrm{R}_\gamma^k\|_V \leq C_\gamma \rho_\gamma^{k\gamma} \{V(x) + V(y)\} \ .$$

Proof. The proof is a direct application of Theorem 4.1.46-(a). □

It is shown in [Bro+19, Theorem 4] that the following result holds: there exists $V : \mathbb{R}^d \to [1, +\infty)$, $\bar{\gamma} > 0$, $C, D \geq 0$ and $\rho \in (0, 1)$ such that for any $k \in \mathbb{N}$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$

$$\|\delta_x \mathrm{R}_\gamma^k - \pi\|_V \leq C\rho^{k\gamma} V(x) + D\sqrt{\gamma} \, ,$$

where π is the invariant measure for the diffusion with drift b and diffusion coefficient Id.

Explicit rates and asymptotics in Theorem 4.1.18

We recall that b satisfies

$$\langle b(x), x\rangle \leq -\mathtt{k}_1 \|x\| \mathbb{1}_{\bar{\mathrm{B}}(0,R_3)^c}(x) - \mathtt{k}_2 \|b(x)\|^2 + \mathtt{a}/2 \, ,$$

with $\mathtt{k}_1, \mathtt{k}_2 > 0$ and $R_3, \mathtt{a} \geq 0$ and that We recall that

$$V(x) = \exp(\mathtt{m}_3^+ \phi(x)) \, , \qquad \phi(x) = \sqrt{1 + \|x\|^2} \, , \qquad \mathtt{m}_3^+ \in (0, \mathtt{k}_1/2] \, . \tag{4.116}$$

Let $W_3(x, y) = (V(x) + V(y))/2$ with $V(x) = \exp[\mathtt{m}_3^+ \sqrt{1 + \|x\|^2}]$ and $\mathtt{m}_3^+ \in (0, \mathtt{k}_1/2]$. Therefore, by Proposition 4.1.17, K_γ satisfies $\mathbf{D}_{\mathrm{d}}(W_3, \lambda^\gamma, A\gamma, \mathsf{X}^2)$ for any $\gamma \in (0, \bar{\gamma}]$ where $\bar{\gamma} \in (0, 2\mathtt{k}_2)$, $R_4 = \max(1, R_3, (d + \mathtt{a})/\mathtt{k}_1)$ and

$$\lambda = e^{-(\mathtt{m}_3^+)^2/2} \, , \quad A = \exp\left[\bar{\gamma}(\mathtt{m}_3^+(d + \mathtt{a}) + (\mathtt{m}_3^+)^2)/2 + \mathtt{m}_3^+(1 + R_4^2)^{1/2}\right] (\mathtt{m}_3^+(d + \mathtt{a})/2 + (\mathtt{m}_3^+)^2) \, ,$$
$$R = \log(2\lambda^{-2\bar{\gamma}} A \log^{-1}(1/\lambda)) \, .$$

Let $\bar{\gamma} \in (0, 2\mathtt{k}_2)$, $\ell \in \mathbb{N}^\star$ specified below, $\lambda_{\bar{\gamma},c}, \rho_{\bar{\gamma},c} \in (0, 1)$ and $D_{\bar{\gamma},1,c}$, $D_{\bar{\gamma},2,c}$, $C_{\bar{\gamma},c} \geq 0$ the constants given by Theorem 4.1.16, such that for any $k \in \mathbb{N}$, $\gamma \in (0, \bar{\gamma}]$ and $x, y \in \mathsf{X}$

$$\mathbf{W}_{\mathbf{c}_3}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \mathrm{K}_\gamma^k \mathbf{c}_3(x, y) \leq \lambda_{\bar{\gamma},c}^{k\gamma/4} [D_{\bar{\gamma},1,c} \mathbf{c}_3(x, y) + D_{\bar{\gamma},2,c} \mathbb{1}_{\Delta_{\mathsf{X}}^c}] + C_{\bar{\gamma},c} \rho_{\bar{\gamma},c}^{k\gamma/4} \, ,$$

with $\mathbf{c}_3(x, y) = \mathbb{1}_{\Delta_{\mathsf{X}}^c}(x, y)\{V(x) + V(y)\}/2$ for any $x, y \in \mathsf{X}$. Note that by (4.116), this result implies that for any $k \in \mathbb{N}$, $\gamma \in (0, \bar{\gamma}]$ and $x, y \in \mathsf{X}$

$$\|\delta_x \mathrm{R}_\gamma^k - \delta_y \mathrm{R}_\gamma^k\|_V \leq \{D_{\bar{\gamma},1,c} + D_{\bar{\gamma},2,c} + C_{\bar{\gamma},c}\} \, \rho_{\bar{\gamma},c}^{k\gamma} \mathbf{c}_3(x, y) \, .$$

Note that using Theorem 4.1.8, we obtain that the following limits exist and do not depend on $\mathtt{L}$

$$\begin{cases} D_{1,c} = \lim_{\bar{\gamma} \to 0} D_{\bar{\gamma},1,c} \, , \quad D_{2,c} = \lim_{\bar{\gamma} \to 0} D_{\bar{\gamma},2,c} \, , \quad C_c = \lim_{\bar{\gamma} \to 0} C_{\bar{\gamma},c} \, , \\ \lambda_c = \lim_{\bar{\gamma} \to 0} \lambda_{\bar{\gamma},c} \, , \quad \rho_c = \lim_{\bar{\gamma} \to 0} \rho_{\bar{\gamma},c} \, . \end{cases}$$

We now discuss the dependency of ρ_b with respect to the introduced parameters, depending on the sign of $\mathtt{m}$ and based on Theorem 4.1.8.

(a) If **B**4 holds, set $\ell = \left\lceil \tilde{M}_{\mathrm{d}}^2 \right\rceil$. Then, if we consider $\mathtt{k}_1, \mathtt{k}_2$ sufficiently small and $\mathtt{a}$ sufficiently large such that the conditions of Theorem 4.1.8 hold, we have

$$\log^{-1}(\rho_c^{-1}) \leq 2 \left[1 + \mathtt{m}_3^+(1 + R)/4 + \log(1 + 2A) + (1 + 4R^2)\mathtt{m}_3^+\right] / \left[\mathtt{m}_3^+ \mathbf{\Phi}(-1/2)\right] \, .$$

Note that the leading term on the right hand side of this equation is of order R^2.

(b) If **B**3($\mathtt{m}$) with $\mathtt{m} \in \mathbb{R}_-$, set $\ell = \left\lceil \tilde{M}_{\mathrm{d}}^2 \right\rceil$. Then, if we consider $\mathtt{k}_1, \mathtt{k}_2$ sufficiently small and $\mathtt{a}$ sufficiently large such that the conditions of Theorem 4.1.8 hold, we have

$$\log^{-1}(\rho_b^{-1}) \leq 2\left[1 + \mathtt{m}_3^+(1+R)/4 + \log(1+2A) + (1+4R^2)\mathtt{m}_3^+\right] \Big/ \left[\mathtt{m}_3^+ \mathbf{\Phi}\{-2(-\mathtt{m})^{1/2}R/(2-2\mathrm{e}^{2\mathtt{m}R^2})^{1/2}\}\right] , \tag{4.117}$$

Note that the right hand side of (4.117) is exponential in $-\mathtt{m}R^2$.

A similar result was already obtained in [DM17, Theorem 10] but the scheme of the proof was different as the authors compared the discretization scheme to the associated diffusion process and used the contraction of the continuous process.

4.2 The SOUL algorithm

4.2.1 Abstract

Maximum likelihood estimation (MLE) is central to modern statistical science. It is a cornerstone of frequentist inference [CB90], and also plays a fundamental role in parametric empirical Bayesian inference [CL00; Cas85]. For simple statistical models, MLE can be performed analytically and exactly. However, for most models, it requires using numerical computation methods, particularly optimization schemes that iteratively seek to maximize the likelihood function and deliver an approximate solution. Following decades of active research in computational statistics and optimization, there are now several computationally efficient methods to perform MLE in a wide range of classes of models [GHM12; BV04].

In this section we consider MLE in models involving incomplete or "missing" data, such as hidden, latent or unobserved variables, and focus on Expectation Maximisation (EM) optimization methods [DLR77], which are the predominant strategy in this setting. While the original EM optimization methodology involved deterministic steps, modern EM methods are mainly stochastic [RC04]. In particular, they typically rely on a Robbins-Monro stochastic approximation (SA) scheme that uses a Monte Carlo stochastic simulation algorithm to approximate the gradients that drive the optimization procedure [RM51; DLM99; KY03; FMP11]. In many cases, SA methods use Markov chain Monte Carlo (MCMC) algorithms, leading to a powerful general methodology which is simple to implement, has a detailed convergence theory [AFM17], and can address a wide range of moderately low-dimensional models. Alternatively, some stochastic EM schemes use a Gibbs sampling algorithm [Cas01], however this requires running several fully converged MCMC chains and can be significantly more computationally expensive as a result.

The expectations and demands on SA methods constantly rise as we seek to address larger problems and provide stronger theoretical guarantees on the solutions delivered. Unfortunately, existing SA methodology and theory do not scale well to large problems. The reasons are twofold. First, the family of MCMC kernels driving the SA scheme needs to satisfy uniform geometric ergodicity conditions that are usually difficult to verify for high-dimensional MCMC kernels. Second, the existing theory requires using asymptotically exact MCMC methods. In practice, these are usually high-dimensional Metropolis-Hastings methods such as the Metropolis-adjusted Langevin algorithm [RT96] or Hamiltonian Monte Carlo [GC11; DMS17], which are difficult to calibrate within the SA scheme to achieve a prescribed acceptance rate. For these reasons, practitioners rarely use SA schemes in high-dimensional settings.

In this section, we propose to address these limitations by using inexact MCMC methods to drive the SA scheme, particularly unadjusted Langevin algorithms, which have easily verifiable geometric

ergodicity conditions, and are easy to calibrate [DM17; Dal17b]. This will allow us to design a high-dimensional stochastic optimization scheme with favourable convergence properties that can be quantified explicitly and easily checked.

Our contributions are structured as follows: Section 4.2.2 formalises the class of MLE problems considered and presents the proposed stochastic optimization method, which is based on a SA approach driven by an unadjusted Langevin algorithm. Detailed theoretical convergence results for the method are reported in Section 4.2.3, which also describes a generalisation of the proposed methodology and theory to other inexact Markov kernels. Section 4.2.4 presents three numerical experiments that demonstrate the proposed methodology in a variety of scenarios. Section 4.2.5 includes additional theoretical results, postponed proofs and some details on computational aspects.

4.2.2 The stochastic optimization via unadjusted Langevin method

The proposed Stochastic Optimization via Unadjusted Langevin (SOUL) method is useful for solving maximum likelihood estimation problems involving intractable likelihood functions. The method is a SA iterative scheme that is driven by an unadjusted Langevin MCMC algorithm. Langevin algorithms are very efficient in high dimensions and lead to an SA scheme that inherits their favourable convergence properties.

Maximum marginal likelihood estimation

Let K be a convex closed set in $\mathbb{R}^p$. The proposed optimization method is well-suited for solving maximum likelihood estimation problems of the form

$$\theta^\star \in \arg\max_{\theta \in \mathsf{K}} \; \log p(y|\theta) - g(\theta) \; , \tag{4.118}$$

where the parameter of interest θ is related to the observed data $y \in \mathsf{Y}$ by a likelihood function $p(y, x|\theta)$ involving an unknown quantity $x \in \mathbb{R}^d$, which is removed from the model by marginalisation. More precisely, we consider problems where the resulting marginal likelihood

$$p(y|\theta) = \int_{\mathbb{R}^d} p(y, x|\theta) \mathrm{d}x \; ,$$

is computationally intractable, and focus on models where the dimension of x is large, making the computation of (4.118) even more difficult. For completeness, we allow the use of a penalty function $g : \mathsf{K} \to \mathbb{R}$, or set $g = 0$ to recover the standard maximum likelihood estimator.

As mentioned previously, the maximum marginal likelihood estimation problem (4.118) arises in problems involving latent or hidden variables [DLR77]. It is also central to parametric empirical Bayes approaches that base their inferences on the pseudo-posterior distribution associated with $p(x|y, \theta^\star) = p(y, x|\theta^\star)/p(y|\theta^\star)$ [CL00]. Moreover, the same optimization problem also arises in hierarchical Bayesian maximum-a-posteriori estimation of θ given y, with marginal posterior $p(\theta|y) \propto p(y|\theta)p(\theta)$ where $p(\theta)$ denotes the prior for θ; in that case $g(\theta) = -\log p(\theta)$ [CB90].

Finally, in this section we assume that $\log p(y, x|\theta)$ is continuously differentiable w.r.t. x and θ, and that g is also continuously differentiable w.r.t. θ.

Stochastic approximation methods

The scheme we propose to solve the optimization problem (4.118) is derived in the SA framework [DLM99], which we recall below.

Starting from any $\theta_0 \in \mathsf{K}$, SA schemes seek to solve (4.118) iteratively by computing a sequence $(\theta_n)_{n\in\mathbb{N}}$ associated with the recursion

$$\theta_{n+1} = \Pi_{\mathsf{K}}[\theta_n + \delta_{n+1}(\Delta_{\theta_n} - \nabla g(\theta_n))] \ , \tag{4.119}$$

where Δ_{θ_n} is some estimator of the intractable gradient $\theta \mapsto \nabla_\theta \log p(y|\theta)$ at θ_n, Π_{K} denotes the projection onto K, and $(\delta_n)_{n\in\mathbb{N}^\star} \in (\mathbb{R}_+^*)^{\mathbb{N}^\star}$ is a sequence of stepsizes. From an optimization viewpoint, iteration (4.119) is a stochastic generalisation of the projected gradient ascent iteration [BV04] for models with intractable gradients. For $n \in \mathbb{N}$, Monte Carlo estimators Δ_{θ_n} for $\nabla_\theta \log p(y|\theta)$ at θ_n are derived from the identity

$$\begin{aligned}\nabla_\theta \log p(y|\theta) &= \int_{\mathbb{R}^d} \frac{\nabla_\theta p(x, y|\theta)}{p(x, y|\theta)} p(x|y, \theta)\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \nabla_\theta \log p(x, y|\theta) p(x|y, \theta)\mathrm{d}x \ ,\end{aligned}$$

which suggests to consider

$$\Delta_{\theta_n} = \frac{1}{m_n} \sum_{k=1}^{m_n} \nabla_\theta \log p(X_k^n, y|\theta_n) \ , \tag{4.120}$$

where $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^\star)^{\mathbb{N}}$ is a sequence of batch sizes and $(X_k^n)_{k\in\{1,\ldots,m_n\}}$ is either an exact Monte Carlo sample from $p(x|y, \theta_n) = p(x, y|\theta_n)/p(y|\theta_n)$, or a sample generated by using a Markov Chain targeting this distribution.

Given a sequence $(\theta_n)_{n=1}^N$ generated by using (4.119), an approximate solution of (4.118) can then be obtained by calculating, for example, the average of the iterates, i.e.,

$$\hat{\theta}_N = \left\{ \sum_{n=1}^N \delta_n \theta_n \right\} \Big/ \left\{ \sum_{n=1}^N \delta_n \right\} \ . \tag{4.121}$$

This estimate converges a.s. to a solution of (4.118) as $N \to \infty$ provided that some conditions on $p(y|\theta)$, g, $p(x|y, \theta)$, $(\delta_n)_{n\in\mathbb{N}}$, and Δ_{θ_n} are fulfilled. Indeed, following three decades of active research efforts in computational statistics and applied probability, we now have a good understanding of how to construct efficient SA schemes, and the conditions under which these schemes converge (see for example [BMP90; FM03; DHS11; AM06; Nem+09; AFM17]).

SA schemes are successfully applied to maximum marginal likelihood estimation problems where the latent variable x has a low or moderately low dimension. However, they are seldomly used when x is high-dimensional because this usually requires using high-dimensional MCMC samplers that, unless carefully calibrated, exhibit poor convergence properties. Unfortunately, calibrating the samplers within a SA scheme is challenging because the target density $p(x|y, \theta_n)$ changes at each iteration. As a result, it is, for example, difficult to use Metropolis-Hastings algorithms that need to achieve a prescribed acceptance probability range. Additionally, the conditions for convergence of MCMC SA schemes are often difficult to verify for high-dimensional samplers. For these reasons, practitioners rarely use SA schemes in high-dimensional settings.

As mentioned previously, we propose to address these difficulties by using modern inexact Langevin MCMC samplers to drive (4.120). These samplers have received a lot of attention in the recent years because they can exhibit excellent large-scale convergence properties and significantly outperform their Metropolised counterparts (see [DMP18] for an extensive comparison in the context of Bayesian imaging models). Stimulated by developments in high-dimensional statistics and machine learning, we now have detailed theory for these algorithms, including explicit and easily verifiable geometric ergodicity

conditions [DM17; Dal17b; EM19], see also Section 4.1. This will allow us to design a stochastic optimization scheme with favourable convergence properties that can be quantified explicitly and easily checked.

Langevin Markov chain Monte Carlo methods

Langevin MCMC schemes to sample from $p(x|y,\theta)$ are based on stochastic continuous dynamics $(\mathbf{X}_t^\theta)_{t\geq 0}$ for which the target distribution $p(x|y,\theta)$ is invariant. Two fundamental examples are the Langevin dynamics solution of the following Stochastic Differential Equation (SDE)

$$\mathrm{d}\mathbf{X}_t^\theta = -\nabla_x \log p(\mathbf{X}_t^\theta|y,\theta)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \ , \tag{4.122}$$

or the kinetic Langevin dynamics solution of

$$\begin{aligned}\mathrm{d}\mathbf{X}_t^\theta &= \mathbf{V}_t^\theta \ , \\ \mathrm{d}\mathbf{V}_t^\theta &= -\nabla_x \log p(\mathbf{X}_t^\theta|y,\theta)\mathrm{d}t - \mathbf{V}_t^\theta \mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \ ,\end{aligned}$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a standard d-dimensional Brownian motion. Under mild assumptions on $p(x|y,\theta)$, these two SDEs admit strong solutions for which $p(x|y,\theta)$ and

$$p(x,v|y,\theta) = p(x|y,\theta)\exp(-\|v\|^2/2)/(2\pi)^{d/2} \ ,$$

are the invariant probability measures. In addition, there are detailed explicit convergence results for $(\mathbf{X}_t^\theta)_{t\geq 0}$ to $p(x|y,\theta)$, and for $(\mathbf{X}_t^\theta, \mathbf{V}_t^\theta)_{t\geq 0}$ to $p(x,v|y,\theta)$, under different metrics [Ebe16; EGZ19].

However, sampling path solutions for these continuous-time dynamics is not feasible in general. Therefore discretizations have to be used instead. In this section, we mainly focus on the Euler-Maruyama discrete-time approximation of (4.122), known as the Unadjusted Langevin Algorithm (ULA) [RT96], given by

$$X_{k+1} = X_k - \gamma\nabla_x \log p(X_k|y,\theta) + \sqrt{2\gamma}Z_{k+1} \ , \tag{4.123}$$

where $\gamma > 0$ is the discretization time step and $(Z_k)_{k\in\mathbb{N}^*}$ is a i.i.d. sequence of d-dimensional zero-mean Gaussian random variables with covariance matrix identity. We will use this Markov kernel to drive our SA schemes.

Observe that (4.123) does not exactly target $p(x|y,\theta)$ because of the bias introduced by the discrete-time approximation. Computational statistical methods have traditionally addressed this issue by complementing (4.123) with a Metropolis-Hastings correction step to asymptotically remove the bias [RT96]. This correction usually deteriorates the convergence properties of the chain and may lead to poor non-asymptotic estimation results, particularly in very high-dimensional settings (see for example [DMP18]). However, until recently it was considered that using (4.123) without a correction step was too risky. Fortunately, recent works have established detailed theoretical guarantees for (4.123) that do not require using any correction [Dal17b; DM17]. Recently, new explicit convergence bounds have been derived under various assumptions on the target probability [Dal17a; Che+18; CB18; LRG18]. In addition, accelerations and variations of ULA have been studied, both theoretically and experimentally, yielding better ergodic convergence rates [Mad+18; Ma+19; MJ19; DRD18]. However, such extensions are out of the scope of the present work whose main contribution is not to provide new results to the existing Markov chain theory but to use the theoretical guarantees of ULA in order to study SA schemes driven by this efficient but inexact sampler.

Note also that the methodology we propose and analyze in this section is fundamentally different from the Stochastic Gradient Langevin dynamics [VZT16; TTV16; WT11; PT13; ASW14; ABW12], an

MCMC algorithm to sample from $p(x|y,\theta)$ using estimators of $\nabla_x \log p(x|y,\theta)$. Finally it should be highlighted that, in an independent line of work, a similar methodology is studied under a different set of assumptions in [Kar+19]. We discuss the links between [Kar+19, Theorem 2] and Theorem 4.2.4 in Section 4.2.3.

The SOUL algorithm

We are now ready to present the proposed Stochastic Optimization via Unadjusted Langevin (SOUL) methodology. Let $(\delta_n)_{n\in\mathbb{N}^\star} \in (\mathbb{R}_+^*)^{\mathbb{N}^\star}$ and $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^\star)^{\mathbb{N}}$ be the sequences of stepsizes and batch sizes defining the SA scheme (4.119)-(4.120). For any $\theta \in \mathsf{K}$ and $\gamma > 0$, denote by $\mathrm{R}_{\gamma,\theta}$ the Langevin Markov kernel (4.123) to approximately sample from $p(x|y,\theta)$, and by $(\gamma_n)_{n\in\mathbb{N}} \in (\mathbb{R}_+^*)^{\mathbb{N}}$ be the sequence of discrete time steps used.

Formally, starting from some $X_0^0 \in \mathbb{R}^d$ and $\theta_0 \in \mathsf{K}$, for $n \in \mathbb{N}$ and $k \in \{0,\dots,m_n-1\}$, we recursively define $(\{X_k^n : k \in \{0,\dots,m_n\}\},\theta_n)_{n\in\mathbb{N}}$ on a probability space $(\Omega,\mathcal{F},\mathbb{P})$, where $(X_k^n)_{k\in\{0,\dots,m_n\}}$ is a Markov chain with Markov kernel $\mathrm{R}_{\gamma_n,\theta_n}$, $X_0^n = X_{m_{n-1}}^{n-1}$ given $\mathcal{F}_{n-1}$, and

$$\theta_{n+1} = \Pi_{\mathsf{K}}\left[\theta_n - \frac{\delta_{n+1}}{m_n}\sum_{k=1}^{m_n}\Delta_{\theta_n}\right] ,$$

where we recall that Π_{K} is the projection onto K, and for all $n \in \mathbb{N}$

$$\begin{aligned}\mathcal{F}_n &= \sigma\left(\theta_0, \{(X_k^\ell)_{k\in\{0,\dots,m_\ell\}} : \ell \in \{0,\dots,n\}\}\right) ,\\ \mathcal{F}_{-1} &= \sigma(\theta_0)\end{aligned} \tag{4.124}$$

Note that such a construction is always possible by Kolmogorov extension theorem [Kal06, Theorem 5.16], hence for any $n \in \mathbb{N}$, θ_{n+1} is $\mathcal{F}_n$-measurable. Then, as mentioned previously, we compute a sequence of approximate solutions of (4.118) by calculating, for example,

$$\hat{\theta}_N = \left\{\sum_{n=1}^N \delta_n\theta_n\right\} \Big/ \left\{\sum_{n=1}^N \delta_n\right\} . \tag{4.125}$$

The pseudocode associated with the proposed SOUL method is presented in Algorithm 6 below. Observe that, for additional efficiency, instead of generating independent Markov chains at each SA iteration, we warm-start the chains by setting $X_0^n = X_{m_{n-1}}^{n-1}$, for any $n \in \{1,\dots,N\}$.

To conclude, Section 4.2.4 below demonstrates the proposed methodology with three numerical experiments related to high-dimensional logistic regression and statistical audio analysis with sparsity promoting priors. A detailed theoretical analysis of the proposed SOUL method is reported in Section 4.2.3. More precisely, we establish that if the cost function $f(\theta) = g(\theta) - \log p(y|\theta)$ defining (4.118) is convex, and if $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$ go to 0 sufficiently fast, then $\mathbb{E}[f(\hat{\theta}_N)]$ converges to $\min_{\mathsf{K}} f$ and quantify the rate of convergence. Moreover, in the case where $(\gamma_n)_{n\in\mathbb{N}}$ is held fixed, *i.e.* for all $n \in \mathbb{N}$, $\gamma_n = \gamma$, we show convergence to a neighbourhood of the solution, in the sense that there exist explicit $C,\alpha > 0$ such that $\limsup_{N\to+\infty} \mathbb{E}[f(\hat{\theta}_N)] - \min_{\mathsf{K}} f \leq C\gamma^\alpha$. Finally, we also study the important case where f is not convex. In that case, we use the results of [KY03] to establish that $(\theta_n)_{n\in\mathbb{N}}$ converges a.s. to a stationary point of the projected ordinary differential equation associated with ∇f and K. We postpone this result to Section 4.2.5.

Algorithm 6 The Stochastic Optimization via Unadjusted Langevin (SOUL) method

Inputs:
$\theta_0 \in \mathsf{K}$, $X_0^0 \in \mathbb{R}^d$, $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$, $(m_n)_{n\in\mathbb{N}}$, N
for $n \in \{1, \dots, N-1\}$ **do**
  if $n \geq 1$ **then**
    $X_0^n \leftarrow X_{m_{n-1}}^{n-1}$
  end if
  for $k \in \{0, \dots, m_n - 1\}$ **do**
    $Z_{k+1}^n \leftarrow$ sample $\mathrm{N}(0, \mathrm{Id})$
    $X_{k+1}^n \leftarrow X_k^n + \gamma_n \nabla_x \log p(X_k^n|y, \theta_n) + \sqrt{2\gamma_n} Z_{k+1}^n$
  end for
  $\Delta_{\theta_n} \leftarrow \frac{1}{m_n} \sum_{k=1}^{m_n} \nabla_\theta \log p(X_k^n, y|\theta_n)$
  $\theta_{n+1} \leftarrow \Pi_{\mathsf{K}}[\theta_n + \delta_{n+1}(\Delta_{\theta_n} - \nabla g(\theta_n))]$
end for
Outputs:
$\hat{\theta}_N = \left\{\sum_{n=1}^N \delta_n \theta_n\right\} \Big/ \left\{\sum_{n=1}^N \delta_n\right\}$

4.2.3 Theoretical convergence analysis for SOUL, and generalisation to other inexact MCMC kernels (SOUK)

In this section we state our main theoretical results for SOUL. For completeness, we first present the results in a general stochastic optimization setting and by considering a generic inexact MCMC sampler, and then show that our results apply to the specific MLE optimization problem (4.118), and to the specific Langevin algorithm (4.123) used in SOUL. All the proofs are postponed to Section 4.2.5.

Stochastic Optimization with inexact MCMC methods

We consider the problem of minimizing a function $f : \mathsf{K} \to \mathbb{R}$ with $\mathsf{K} \subset \mathbb{R}^p$ under the following assumptions.

A1. *K is a convex compact set and $\mathsf{K} \subset \overline{\mathrm{B}}(0, M_\Theta)$ with $M_\Theta > 0$.*

A2. *There exist an open set $\mathsf{U} \subset \mathbb{R}^p$ and $\mathcal{L} \geq 0$ such that $\mathsf{K} \subset \mathsf{U}$, $f \in \mathrm{C}^1(\mathsf{U}, \mathbb{R})$ and satisfies for any $\theta_1, \theta_2 \in \mathsf{K}$*

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq \mathcal{L}\|\theta_1 - \theta_2\| \ .$$

A3. *For any $\theta \in \mathsf{K}$, there exist $H_\theta : \mathbb{R}^d \to \mathbb{R}^p$ and a probability distribution π_θ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ satisfying that $\pi_\theta(H_\theta) < +\infty$ and for any $\theta \in \mathsf{K}$*

$$\nabla f(\theta) = \int_{\mathbb{R}^d} H_\theta(x) \mathrm{d}\pi_\theta(x) \ .$$

In addition, $(\theta, x) \mapsto H_\theta(x)$ is measurable.

Note that for the maximum marginal likelihood estimation problem (4.118), f corresponds to $\theta \mapsto -\log(p(y|\theta)) + g(\theta)$, for any $\theta \in \mathsf{K}$, $H_\theta : x \mapsto \nabla_\theta \log(p(x, y|\theta))$ and π_θ is the probability distribution with density with respect to the Lebesgue measure $x \mapsto p(x|y, \theta)$.

To minimize the objective function f we suggest the use of a SA strategy which extends the one presented in Section 4.2.2. More precisely, motivated by the methodology described in Section 4.2.2, we

propose a SA scheme which relies on biased estimates of $\nabla f(\theta)$ through a family of Markov kernels $\{\mathrm{K}_{\gamma,\theta}, \gamma \in (0,\bar{\gamma}] \text{ and } \theta \in \mathsf{K}\}$, for $\bar{\gamma} > 0$, such that for any $\theta \in \mathsf{K}$ and $\gamma \in (0,\bar{\gamma}]$, $\mathrm{K}_{\gamma,\theta}$ admits an invariant probability distribution $\pi_{\gamma,\theta}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. In the SOUL method, the Markov kernel $\mathrm{K}_{\gamma,\theta}$ stands for $\mathrm{R}_{\gamma,\theta}$ for any $\gamma \in (0,\bar{\gamma}]$ and $\theta \in \mathsf{K}$, where $\mathrm{R}_{\gamma,\theta}$ is the Markov kernel associated with (4.123). We assume in addition that the bias associated to the use of this family of Markov kernels can be controlled w.r.t. to γ uniformly in θ, *i.e.* for example there exists $C > 0$ such that for all $\gamma \in (0,\bar{\gamma}]$ and $\theta \in \mathsf{K}$, $\|\pi_{\gamma,\theta} - \pi_\theta\|_{\mathrm{TV}} \leq C\gamma^\alpha$ with $\alpha > 0$.

Let now $(\delta_n)_{n\in\mathbb{N}} \in (\mathbb{R}_+^*)^{\mathbb{N}}$ and $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^\star)^{\mathbb{N}}$ be sequences of stepsizes and batch sizes which will be used to define the sequence relatively to the variable θ similarly to (4.119) and (4.120). Let $(\gamma_n)_{n\in\mathbb{N}} \in (\mathbb{R}_+^*)^{\mathbb{N}}$ be a sequence of stepsizes which will be used to get approximate samples from π_{θ_n}, similarly to (4.123). Starting from $X_0^0 \in \mathbb{R}^d$ and $\theta_0 \in \mathsf{K}$, we define on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $(\{X_k^n : k \in \{0,\dots,m_n\}\}, \theta_n)_{n\in\mathbb{N}}$ by the following recursion for $n \in \mathbb{N}$ and $k \in \{0,\dots,m_n - 1\}$

$$\begin{aligned} &(X_k^n)_{k\in\{0,\dots,m_n\}} \text{ is Markov chain with kernel } \mathrm{K}_{\gamma_n,\theta_n} \text{ and } X_0^n = X_{m_{n-1}}^{n-1} \text{ given } \mathcal{F}_{n-1} \ , \\ &\theta_{n+1} = \Pi_{\mathsf{K}}\left[\theta_n - \frac{\delta_{n+1}}{m_n}\sum_{k=1}^{m_n} H_{\theta_n}(X_k^n)\right] \ , \end{aligned} \tag{4.126}$$

where Π_{K} is the projection onto K and $\mathcal{F}_n$ is defined by (4.124). By (4.126), for any $n \in \mathbb{N}$, θ_{n+1} is $\mathcal{F}_n$-measurable and $(\mathcal{F}_n)_{n\in\mathbb{N}}$ given in (4.124). Then the sequence of approximate minimizers of f is given by $(\hat{\theta}_N)_{N\in\mathbb{N}}$, (4.125).

Under different sets of conditions on $f, H, (\delta_n)_{n\in\mathbb{N}}, (\gamma_n)_{n\in\mathbb{N}}$ and $(m_n)_{n\in\mathbb{N}}$ we obtain that $(\theta_n)_{n\in\mathbb{N}}$ converges a.s. to an element of $\arg\min_{\mathsf{K}} f$. In particular in this section we consider the case where f is assumed to be convex. We establish that if $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$ go to 0 sufficiently fast, $\mathbb{E}[f(\hat{\theta}_N)] - \min_{\mathsf{K}} f$ goes to 0 with a quantitative rate of convergence. In the case where $(\gamma_n)_{n\in\mathbb{N}}$ is held fixed, *i.e.* for all $n \in \mathbb{N}$, $\gamma_n = \gamma$, we show that while $\mathbb{E}[f(\hat{\theta}_N)]$ does not converge to 0, there exists $C, \alpha > 0$ such that $\limsup_{N\to+\infty} \mathbb{E}[f(\hat{\theta}_N)] - \min_{\mathsf{K}} f \leq C\gamma^\alpha$. In the case where f is non-convex, we apply some results from stochastic approximation [KY03] which establish that the sequence $(\theta_n)_{n\in\mathbb{N}}$ converges a.s. to a stationary point of the projected ordinary differential equation associated with ∇f and K. We postpone this result to Section 4.2.5. Finally, our upper bounds and convergence results can be extended to the case where ∇f can be expressed as a finite sum of expectations, *i.e.* for any $\theta \in \mathsf{K}$, $\nabla f(\theta) = \sum_{i=1}^N \pi_\theta^{(i)}(H_\theta^{(i)})$ where $N \in \mathbb{N}^\star$ and for all $i \in \{1,\dots,N\}$, $H_\theta^{(i)} : \mathbb{R}^d \to \mathbb{R}^p$ is a measurable function and $\pi_\theta^{(i)}$ is a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. For the sake of clarity we restrict ourselves to the case where $N = 1$.

Main results

We impose a stability condition on the stochastic process $\{(X_k^n)_{k\in\{0,\dots,m_n\}} : n \in \mathbb{N}\}$ defined by (4.126) and that for any $\gamma \in (0,\bar{\gamma}]$ and $\theta \in \mathsf{K}$ the iterates of $\mathrm{K}_{\gamma,\theta}$ are close enough to π_θ after a sufficiently large number of iterations.

H1. *There exists a measurable function $V : \mathbb{R}^d \to [1,+\infty)$ satisfying the following conditions.*

(i) There exists $A_1 \geq 1$ such that for any $n, p \in \mathbb{N}$, $k \in \{0,\dots,m_n\}$

$$\mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n}^p V(X_k^n) \middle| X_0^0\right] \leq A_1 V(X_0^0) \ , \ \mathbb{E}\left[V(X_0^0)\right] < +\infty \ ,$$

where $\{(X_k^\ell)_{k\in\{0,\dots,m_\ell\}} : \ell \in \{0,\dots,n\}\}$ is given by (4.126).

(ii) There exist $A_2, A_3 \geq 1$, $\rho \in [0,1)$ *such that for any* $\gamma \in (0, \bar{\gamma}]$, $\theta \in \mathsf{K}$, $x \in \mathbb{R}^d$ *and* $n \in \mathbb{N}$, $\mathrm{K}_{\gamma,\theta}$ *has a stationary distribution* $\pi_{\gamma,\theta}$ *and*

$$\|\delta_x \mathrm{K}_{\gamma,\theta}^n - \pi_{\gamma,\theta}\|_V \leq A_2 \rho^{n\gamma} V(x) \;, \qquad \pi_{\gamma,\theta}(V) \leq A_3 \;.$$

(iii) There exists $\mathbf{\Psi} : \mathbb{R}_+^\star \to \mathbb{R}_+$ *such that for any* $\gamma \in (0, \bar{\gamma}]$ *and* $\theta \in \mathsf{K}$

$$\|\pi_{\gamma,\theta} - \pi_\theta\|_{V^{1/2}} \leq \mathbf{\Psi}(\gamma) \;.$$

H1-(ii) is an ergodicity condition in V-norm for the Markov kernel $\mathrm{K}_{\gamma,\theta}$ uniform in $\theta \in \mathsf{K}$. There exists an extensive literature on the conditions under which a Markov kernel is ergodic [MT92; Dou+18]. **H**1-(iii) ensures that the distance between the invariant measure $\pi_{\gamma,\theta}$ of the Markov kernel $\mathrm{K}_{\gamma,\theta}$ and π_θ can be controlled uniformly in θ. We show that this condition holds in the case of the Langevin Monte Carlo algorithm in Proposition 4.2.22. We now state our mains results.

Theorem 4.2.1. *Assume* **A**1, **A**2, **A**3 *hold and* f *is convex. Let* $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ *be sequences of non-increasing positive real numbers and* $(m_n)_{n\in\mathbb{N}}$ *be sequences of positive integers satisfying* $\sup_{n\in\mathbb{N}} \delta_n < 1/\mathcal{L}$, $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$ *and*

$$\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty \;, \qquad \sum_{n=0}^{+\infty} \delta_{n+1} \mathbf{\Psi}(\gamma_n) < +\infty \;, \qquad \sum_{n=0}^{+\infty} \delta_{n+1}/(m_n \gamma_n) < +\infty \;. \tag{4.127}$$

Let $\{(X_k^n)_{k\in\{0,\dots,m_n\}} : n \in \mathbb{N}\}$ *and* $(\theta_n)_{n\in\mathbb{N}}$ *be given by* (4.126). *Assume in addition that* **H**1 *is satisfied and that for any* $\theta \in \mathsf{K}$ *and* $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$. *Then, the following statements hold:*

(a) $(\theta_n)_{n\in\mathbb{N}}$ *converges a.s. to some* $\theta^\star \in \arg\min_{\mathsf{K}} f$;

(b) furthermore, a.s. there exists $C \geq 0$ *such that for any* $n \in \mathbb{N}^\star$

$$\left\{ \sum_{k=1}^n \delta_k f(\theta_k) \Big/ \sum_{k=1}^n \delta_k \right\} - \min_{\mathsf{K}} f \leq C \Big/ \left(\sum_{k=1}^n \delta_k \right) \;.$$

Proof. The proof is postponed to Section 4.2.5. □

Note that in (4.126), $X_0^n = X_{m_{n-1}}^{n-1}$ for $n \in \mathbb{N}^\star$. This procedure is referred to as warm-start in the sequel. An inspection of the proof of Theorem 4.2.1 shows that X_0^n could be any random variable independent from $\mathcal{F}_{n-1}$ for any $n \in \mathbb{N}$ with $\sup_{n\in\mathbb{N}^\star} \mathbb{E}\left[V(X_0^n)\right] < +\infty$. It is not an option in the fixed batch size setting of Theorem 4.2.3, where the warm-start procedure is crucial for the convergence to occur.

We extend this theorem to non convex objective function see Theorem 4.2.7. Under the conditions of Theorem 4.2.1 with the additional assumption that $\partial\mathsf{K}$ is a smooth manifold we obtain that $(\theta_n)_{n\in\mathbb{N}}$ converges a.s. to some point θ^* such that $\nabla f(\theta^*) + \mathbf{n} = 0$ with $\mathbf{n} = 0$ if $\theta^* \in \mathrm{int}(\mathsf{K})$ and $\mathbf{n} \in \mathrm{T}(\theta^*, \partial\mathsf{K})^\perp$ if $\theta^* \in \partial\mathsf{K}$, where $\mathrm{T}(\theta, \partial\mathsf{K})$ is the tangent space of $\partial\mathsf{K}$ at point $\theta \in \partial\mathsf{K}$, see [Aub01, Chapter 2].

In the case where $\mathrm{K}_{\gamma,\theta} = \mathrm{R}_{\gamma,\theta}$ is the Markov kernel associated with the Langevin update (4.123), under appropriate conditions Proposition 4.2.22 shows that for any $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$, $\mathbf{\Psi}(\gamma) =$

$\mathcal{O}(\gamma^{1/2})$. In that case, assume then that there exist $a, b, c > 0$ such that for any $n \in \mathbb{N}^\star$, $\delta_n = n^{-a}$, $\gamma_n = n^{-b}$ and $m_n = \lceil n^c \rceil$ then (4.127) is equivalent to

$$a \leq 1 \;, \qquad a + b/2 > 1 \;, \qquad a - b + c > 1 \;. \tag{4.128}$$

Suppose $a \in [0, 1]$ is given, then the previous equation reads

$$b = 2(1-a) + \varsigma_1 \;, \qquad c = 3(1-a) + \varsigma_2 \;, \qquad \varsigma_2 > \varsigma_1 > 0 \;.$$

This illustrates a trade-off between the intrinsic inaccuracy of our algorithm through the family of Markov kernels (4.126) which do not exactly target π_θ and the minimization aim of our scheme. Note also that $(\delta_n)_{n\in\mathbb{N}}$ is allowed to be constant. This case yields $\gamma_n = n^{-2-\varsigma_1}$ and $m_n = \lceil n^{3+\varsigma_2} \rceil$ with $\varsigma_2 > \varsigma_1 > 0$.

In our next result we derive an non-asymptotic upper-bound of $(\mathbb{E}[f(\hat{\theta}_n) - \min_{\mathsf{K}} f])_{n\in\mathbb{N}}$.

Theorem 4.2.2. *Assume* **A**1, **A**2, **A**3 *hold and f is convex. Let $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ be a sequence of positive integers satisfying $\sup_{n\in\mathbb{N}} \delta_n < 1/\mathcal{L}$, $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$. Let $\{(X_k^n)_{k\in\{0,\ldots,m_n\}} : n \in \mathbb{N}\}$ be given by (4.126). Assume in addition that* **H**1 *is satisfied and that for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$. Then, there exists $(E_n)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}^\star$*

$$\mathbb{E}\left[\left\{\sum_{k=1}^n \delta_k f(\theta_k) \Big/ \sum_{k=1}^n \delta_k\right\} - \min_{\mathsf{K}} f\right] \leq E_n \Big/ \left(\sum_{k=1}^n \delta_k\right) \;,$$

with for any $n \in \mathbb{N}^\star$,

$$\begin{aligned} E_n = 2M_\Theta^2 + 2B_1 M_\Theta \mathbb{E}\left[V^{1/2}(X_0^0)\right] \sum_{k=0}^{n-1} \delta_{k+1}/(m_k\gamma_k) \\ + 2M_\Theta \sum_{k=0}^{n-1} \delta_{k+1} \boldsymbol{\Psi}(\gamma_k) + 4B_1^2 \mathbb{E}\left[V(X_0^0)\right] \sum_{k=0}^{n-1} \delta_{k+1}^2/(m_k\gamma_k)^2 \\ + 4\sum_{k=0}^{n-1} \delta_{k+1}^2 \boldsymbol{\Psi}(\gamma_k)^2 + B_2 \sum_{k=0}^{n-1} \delta_{k+1}^2/(m_k\gamma_k)^2 \;, \end{aligned} \tag{4.129}$$

where B_1 and B_2 are given in Lemma 4.2.10 and Lemma 4.2.11.

Proof. The proof is postponed to Section 4.2.5. □

We recall that in the case where $\mathrm{K}_{\gamma,\theta} = \mathrm{R}_{\gamma,\theta}$ is the Markov kernel associated with the Langevin update (4.123), under appropriate conditions Proposition 4.2.22 shows that for any $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$, $\boldsymbol{\Psi}(\gamma) = \mathcal{O}(\gamma^{1/2})$. In that case, if there exist $a, b, c \geq 0$ such that for any $n \in \mathbb{N}^\star$, $\delta_n = n^{-a}$, $\gamma_n = n^{-b}$, $m_n = n^c$ and (4.128) holds, the accuracy, respectively the complexity, of the algorithm are of orders $(\sum_{k=1}^n \delta_k)^{-1} = \mathcal{O}(n^{a-1})$, respectively $\sum_{k=0}^n m_k = \mathcal{O}(n^{3(1-a)+\varsigma_2+1})$ for $\varsigma_2 > 0$. Thus, for a fix target precision $\varepsilon > 0$, it requires that $\varepsilon = \mathcal{O}(n^{a-1})$ and the complexity reads $\mathcal{O}(\varepsilon^{-3} (\log(1/\varepsilon)/(1-a))^{1+\varsigma_2})$. On the other hand, if we fix the complexity budget to N the accuracy is of order $\mathcal{O}(N^{-(3+(1+\varsigma_2)/(1-a))^{-1}})$. These two considerations suggest to set a close to 0. In the special case where $a = 0$, we obtain that the accuracy is of order $\mathcal{O}(n^{-1})$, which is similar to the order identified in the deterministic gradient descent for convex functionals.

A case of interest is the fix stepsize setting, *i.e.* for all $n \in \mathbb{N}$, $\gamma_n = \gamma > 0$. Assume that $(\delta_n)_{n\in\mathbb{N}}$ is non-increasing, $\lim_{n\to+\infty} \delta_n = 0$ and $\lim_{n\to+\infty} m_n = +\infty$. In addition, assume that $\sum_{n\in\mathbb{N}^\star} \delta_n = +\infty$ then, by [PS98, Problem 80, Part I], it holds that

$$\begin{cases} \lim_{n\to+\infty} \left[\left(\sum_{k=1}^n \delta_k/m_k\right) / \left(\sum_{k=1}^n \delta_k\right) \right] = 0 \; ; \\ \lim_{n\to+\infty} \left[\left(\sum_{k=1}^n \delta_k^2\right) / \left(\sum_{k=1}^n \delta_k\right) \right] = 0 \; . \end{cases}$$

Therefore, we obtain that

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\{ \sum_{k=1}^n \delta_k f(\theta_k) \middle/ \sum_{k=1}^n \delta_k \right\} - \min f \right] \leq 2M_\Theta \boldsymbol{\Psi}(\gamma) \; .$$

Similarly, if the stepsize is fixed and the number of Markov chain iterates is fixed, *i.e.* for all $n \in \mathbb{N}$, $\gamma_n = \gamma$ and $m_n = m$ with $\gamma > 0$ and $m \in \mathbb{N}^\star$, we obtain that

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\{ \sum_{k=1}^n \delta_k f(\theta_k) \middle/ \sum_{k=1}^n \delta_k \right\} - \min f \right] \leq \boldsymbol{\Xi}_1(\gamma) \; , \tag{4.130}$$

with

$$\boldsymbol{\Xi}_1(\gamma) = 2B_1 M_\Theta \mathbb{E}\left[V^{1/2}(X_0^0) \right] / \gamma + 2M_\Theta \boldsymbol{\Psi}(\gamma) \; .$$

However if $(m_n)_{n\in\mathbb{N}}$ is constant the convergence cannot be obtained using Theorem 4.2.1. Strengthening the conditions of Theorem 4.2.1 and making use of the warm-start property of the algorithm we can derive the convergence in that case.

We now are interested in the case where the batch size is fixed, *i.e.* $m_n = m_0$ for all $n \in \mathbb{N}$. For ease of exposition we only consider $m_0 = 1$ and let $\tilde{X}_{n+1} = X_1^n$ for any $n \in \mathbb{N}$. However the general case can be adapted from the proof of the result stated below. More precisely we consider the setting where the recursion (4.126) can be written for any $n \in \mathbb{N}$ as

$$\begin{aligned} &X_{n+1} \text{ has distribution } \mathrm{K}_{\gamma_n,\tilde{\theta}_n}(X_n, \cdot) \text{ conditional to } \tilde{\mathcal{F}}_n \; , \\ &\tilde{\theta}_{n+1} = \Pi_{\mathsf{K}}\left[\tilde{\theta}_n - \delta_{n+1} H_{\tilde{\theta}_n}(X_{n+1}) \right] \; , \end{aligned} \tag{4.131}$$

with $\theta_0 \in \mathsf{K}$, $\tilde{X}_0 \in \mathbb{R}^d$ and where $\tilde{\mathcal{F}}_n$ is given by

$$\tilde{\mathcal{F}}_n = \sigma\left(\tilde{\theta}_0, (X_\ell)_{\ell\in\{0,\ldots,n\}} \right) \; . \tag{4.132}$$

We consider the following assumption on the family $\{H_\theta \; : \; \theta \in \mathsf{K}\}$.

A4. *There exists $L_H \geq 0$ such that for any $x \in \mathbb{R}^d$ and $\theta_1, \theta_2 \in \mathsf{K}$,*

$$\|H_{\theta_1}(x) - H_{\theta_2}(x)\| \leq L_H \|\theta_1 - \theta_2\| V^{1/2}(x) \; .$$

We consider a similar property as **A**4 on the family of Markov kernels $\{\mathrm{K}_{\gamma,\theta}, \gamma \in (0, \bar{\gamma}], \theta \in \mathsf{K}\}$, which weakens the assumption [AFM17, H6].

H2. *There exist a measurable function $V : \mathbb{R}^d \to [1, +\infty)$, $\boldsymbol{\Lambda}_1 : (\mathbb{R}_+^*)^2 \to \mathbb{R}_+$ and $\boldsymbol{\Lambda}_2 : (\mathbb{R}_+^*)^2 \to \mathbb{R}_+$ such that for any $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_2 < \gamma_1$, $\theta_1, \theta_2 \in \mathsf{K}$, $x \in \mathbb{R}^d$ and $a \in [1/4, 1/2]$*

$$\|\delta_x \mathrm{K}_{\gamma_1,\theta_1} - \delta_x \mathrm{K}_{\gamma_2,\theta_2}\|_{V^a} \leq \left[\boldsymbol{\Lambda}_1(\gamma_1, \gamma_2) + \boldsymbol{\Lambda}_2(\gamma_1, \gamma_2) \|\theta_1 - \theta_2\| \right] V^{2a}(x) \; .$$

The following theorem ensures convergence properties for $(\theta_n)_{n\in\mathbb{N}}$ similar to the ones of Theorem 4.2.1. The proof of this result is based on a generalization of [FMP11, Lemma 4.2] for inexact MCMC schemes.

Theorem 4.2.3. *Assume* **A**1, **A**2, **A**3, **A**4 *hold and f is convex. Let $\bar{\gamma} > 0$, $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers satisfying $\sup_{n\in\mathbb{N}} \delta_n < 1/\mathcal{L}$, $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$, $\sup_{n\in\mathbb{N}} |\delta_{n+1} - \delta_n|\delta_n^{-2} < +\infty$, $\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty$ and*

$$\begin{aligned} &\sum_{n=0}^{+\infty} \delta_{n+1} \boldsymbol{\Psi}(\gamma_n) < +\infty \;, \qquad \sum_{n=0}^{+\infty} \delta_{n+1}^2 \gamma_n^{-2} < +\infty \;, \\ &\sum_{n=0}^{+\infty} \delta_{n+1} \gamma_{n+1}^{-2} \left[\boldsymbol{\Lambda}_1(\gamma_n, \gamma_{n+1}) + \delta_{n+1} \boldsymbol{\Lambda}_2(\gamma_n, \gamma_{n+1})\right] < +\infty \;. \end{aligned} \tag{4.133}$$

Let $(X_n)_{n\in\mathbb{N}}$ be given by (4.131). Assume in addition that **H**1 *and* **H**2 *are satisfied and that for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/4}(x)$. Then the following statements hold:*

(a) $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ converges a.s. to some $\theta^\star \in \arg\min_{\mathsf{K}} f$;

(b) furthermore, a.s. there exists $C \geq 0$ such that for any $n \in \mathbb{N}^\star$

$$\left\{ \sum_{k=1}^{n} \delta_k f(\tilde{\theta}_k) \Big/ \sum_{k=1}^{n} \delta_k \right\} - \min_{\mathsf{K}} f \leq C \Big/ \left(\sum_{k=1}^{n} \delta_k \right) \;.$$

Proof. The proof is postponed to Section 4.2.5. □

In the case where $\mathrm{K}_{\gamma,\theta} = \mathrm{R}_{\gamma,\theta}$ is the Markov kernel associated with the Langevin update (4.123), under appropriate conditions Proposition 4.2.22 and Proposition 4.2.24 show that for any $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ and $\gamma_1 > \gamma_2$, $\boldsymbol{\Psi}(\gamma_1) = \mathrm{C}_1 \gamma^{1/2}$, $\boldsymbol{\Lambda}_1(\gamma_1, \gamma_2) = \mathrm{C}_2 \gamma_2^{-1} |\gamma_1 - \gamma_2|$ and $\boldsymbol{\Lambda}_2(\gamma_1, \gamma_2) = \mathrm{C}_3 \gamma_2^{1/2}$, for $\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3 \geq 0$. Thus we obtain that the following series should converge

$$\begin{aligned} &\sum_{n=0}^{+\infty} \delta_{n+1} \gamma_n^{1/2} < +\infty \;, \qquad \sum_{n=0}^{+\infty} \delta_{n+1}^2 / \gamma_{n+1}^2 < +\infty \;, \\ &\sum_{n=0}^{+\infty} \delta_{n+1} (\gamma_n - \gamma_{n+1}) / \gamma_{n+1}^3 < +\infty \;. \end{aligned}$$

In addition, assume that $\delta_n = n^{-a}$ and that $\gamma_n = n^{-b}$ with $a, b > 0$. In this case the summability conditions of Theorem 4.2.3 read

$$\begin{aligned} &a \leq 1 \;, \qquad a + b/2 > 1 \;, \\ &2a - 2b > 1 \;, \qquad a + (b+1) - 3b > 1 \;, \end{aligned}$$

Since for any $a \in [0, 1]$, $a - 1/2 \leq a/2$, this condition is equivalent to

$$b \in (2(1-a), a - 1/2) \;, \qquad a \in [0, 1] \;.$$

Note that $2(1-a) < a - 1/2$ as soon as $a > 5/6$. In the special setting where $a = 1$ then the convergence in Theorem 4.2.3 occurs as soon as $b \in (0, 1/2)$. In any case, since $a > a - 1/2$ we obtain that $\lim_{n\to+\infty} (\delta_n / \gamma_n) = 0$. This means that the perturbed gradient descent dynamic associated with $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ moves slower than the Markov chain dynamic associated with $(\tilde{X}_n)_{n\in\mathbb{N}}$.

Theorem 4.2.4. *Assume* **A**1, **A**2, **A**3, **A**4 *hold and f is convex. Let $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ be a sequence of positive integers satisfying $\sup_{n\in\mathbb{N}} \delta_n < 1/\mathcal{L}$ and $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$. Let $(X_n)_{n\in\mathbb{N}}$ be given by (4.131). Assume in addition that* **H**1 *and* **H**2 *are satisfied and that for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/4}(x)$. Then, there exists $(\tilde{E}_n)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}^\star$*

$$\mathbb{E}\left[\left\{\sum_{k=1}^n \delta_k f(\theta_k)\middle/\sum_{k=1}^n \delta_k\right\} - \min_{\mathsf{K}} f\right] \leq \tilde{E}_n \middle/ \left(\sum_{k=1}^n \delta_k\right) ,$$

with for any $n \in \mathbb{N}^\star$,

$$\begin{aligned}\tilde{E}_n &= 2M_\Theta + 2M_\Theta \sum_{k=0}^n \delta_{k+1}\mathbf{\Psi}(\gamma_k) + C_3 \sum_{k=0}^n |\delta_{k+1} - \delta_k|\,\gamma_k^{-1} \\ &+ 2M_\Theta C_2 \sum_{k=0}^n \delta_{k+1}\gamma_{k+1}^{-1}\left[\gamma_{k+1}^{-1}\left\{\mathbf{\Lambda}_1(\gamma_k,\gamma_{k+1})\right.\right. \\ &\left.\left. + \mathbf{\Lambda}_2(\gamma_k,\gamma_{k+1})\delta_{k+1} + \delta_{k+1}\right\} + C_3 \sum_{k=0}^n \delta_{k+1}^2\gamma_{k+1}^{-1}\right] \\ &+ C_3(\delta_{n+1}/\gamma_n - \delta_0/\gamma_0) + C_1 \sum_{k=0}^n \delta_{k+1}^2 .\end{aligned}$$

where C_1, C_2 and C_3 are given in Lemma 4.2.12, Lemma 4.2.15 and Lemma 4.2.14.

Proof. The proof is postponed to Section 4.2.5. □

Theorem 4.2.4 improves the conclusions of Theorem 4.2.2 in the case where $\gamma_n = \gamma > 0$ for any $n \in \mathbb{N}$. Indeed, in that case, similarly to (4.130), assuming that $\lim_{n\to+\infty} \delta_n = 0$, $\sup_{n\in\mathbb{N}} |\delta_{n+1} - \delta_n|\,\delta_n^{-2} < +\infty$, $\mathbf{\Lambda}_1 = 0$, we obtain that for all $n \in \mathbb{N}$

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\{\sum_{k=1}^n \delta_k f(\theta_k)\middle/\sum_{k=1}^n \delta_k\right\} - \min f\right] \leq \mathbf{\Xi}_2(\gamma) ,$$

with

$$\begin{aligned}\mathbf{\Xi}_2(\gamma) &= 2M_\Theta\Psi(\gamma) \\ &\leq \mathbf{\Xi}_1(\gamma) = 2B_1M_\Theta\mathbb{E}\left[V^{1/2}(X_0^0)\right]/\gamma + 2M_\Theta\mathbf{\Psi}(\gamma) .\end{aligned}$$

In the case where $\sup_{\gamma\in(0,\bar{\gamma}]} \mathbf{\Psi}(\gamma) < +\infty$, $\mathbf{\Xi}_2(\gamma)$ is of order $\mathcal{O}(\mathbf{\Psi}(\gamma))$ and $\mathbf{\Xi}_1(\gamma)$ is of order $\mathcal{O}(\gamma^{-1})$. Therefore if $\lim_{\gamma\to 0} \mathbf{\Psi}(\gamma) = 0$, even in the fixed batch size setting, the minimum of the objective function f can be approached with arbitrary precision $\varepsilon > 0$ by choosing γ small enough.

Finally, note that the conclusions of Theorem 4.2.4 are similar to the ones of [Kar+19, Theorem 2]. In [Kar+19] the main result is a bound on $\mathbb{E}[\sum_{k=1}^n \delta_k \|\nabla_\theta f(\theta_k)\|^2 / \sum_{k=1}^n \delta_k]$ and $\nabla f(\theta)$ is not assumed to be convex but only related to a Lyapunov functional [Kar+19, A1-A3]. However, it is assumed that for any $\theta \in \mathsf{K}$ and $\gamma \in (0, \bar{\gamma}]$ the invariant probability distribution of the Markov kernel $\mathrm{K}_{\gamma,\theta}$ is π_θ, *i.e.* $\mathbf{\Psi} = 0$ in **H**1-(iii), which is not the case in our analysis. Plugging this assumption and choosing $\gamma_n = \gamma > 0$ for all $n \in \mathbb{N}$ in Theorem 4.2.4 we obtain that for any $n \in \mathbb{N}$, $\tilde{E}_n \leq C\sum_{k=0}^n \delta_{k+1}^2$ with $C > 0$, which is similar to the upper bound given in [Kar+19, Theorem 2].

Application to SOUL

We now apply our results to the SOUL methodology introduced in Section 4.2.2 where the Markov kernel $\mathrm{R}_{\gamma,\theta}$ with $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \mathsf{K}$ is given by a Langevin Markov kernel and associated with recursion (4.123). Setting for any $\theta \in \mathsf{K}$, $\pi_\theta = p(\cdot|y, \theta)$, we consider the following assumption on the family of probability distributions $(\pi_\theta)_{\theta \in \mathsf{K}}$.

L1. *For any $\theta \in \mathsf{K}$, there exists $U_\theta : \mathbb{R}^d \to \mathbb{R}$ such that π_θ admits a probability density function w.r.t. to the Lebesgue measure proportional to $x \mapsto \exp(-U_\theta(x))$. In addition $(\theta, x) \mapsto U_\theta(x)$ is continuous, $x \mapsto U_\theta(x)$ is differentiable for all $\theta \in \mathsf{K}$ and there exists $\mathtt{L} \geq 0$ such that for any $x, y \in \mathbb{R}^d$,*

$$\sup_{\theta \in \mathsf{K}} \|\nabla_x U_\theta(x) - \nabla_x U_\theta(y)\| \leq \mathtt{L} \|x - y\| \ ,$$

and $\{\|\nabla_x U_\theta(0)\| \ : \ \theta \in \mathsf{K}\}$ is bounded.

In the case where $\mathrm{K}_{\gamma,\theta} = \mathrm{R}_{\gamma,\theta}$ for any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \mathsf{K}$, the first line of (4.126) can be rewritten for any $n \in \mathbb{N}$ and $k \in \{0, \ldots, m_n - 1\}$

$$\begin{aligned} X^n_{k+1} &= X^n_k - \gamma_n \nabla_x U_{\theta_n}(X^n_k) + \sqrt{2\gamma_n} Z^n_{k+1} \ , \\ &\text{with } X^n_0 = X^{n-1}_{m_{n-1}} \text{ if } n \geq 1 \ , \end{aligned} \tag{4.134}$$

given $(\gamma_n)_{n \in \mathbb{N}} \in (0, \bar{\gamma}]^{\mathbb{N}}$, $(m_n)_{n \in \mathbb{N}} \in (\mathbb{N}^\star)^{\mathbb{N}}$ and also $(Z^n_k)_{n \in \mathbb{N}, k \in \{1, \ldots, m_n\}}$ a family of i.i.d. d-dimensional zero-mean Gaussian random variables with covariance matrix identity. In the following propositions, we show that the results above hold by deriving sufficient conditions under which **H**1 and **H**2 are satisfied. Under **L**1, the Langevin diffusion defined by (4.122) admits a unique strong solution for any $\theta \in \mathsf{K}$. Consider now the following additional tail condition on U_θ which ensures geometric ergodicity of $\mathrm{R}_{\gamma,\theta}$ for any $\theta \in \mathsf{K}$ and $\gamma \in (0, \bar{\gamma}]$, with $\bar{\gamma}$ which will be specified below.

L2. *There exist $\mathtt{k}_2 > 0$ and $\mathtt{m}_3^+, \mathtt{c}, R_2 \geq 0$ such that for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$,*

$$\langle \nabla_x U_\theta(x), x \rangle \geq \mathtt{k}_2 \|x\| \mathbb{1}_{\mathrm{B}(0,R_2)^{\mathrm{c}}}(x) + \mathtt{m}_3^+ \|\nabla_x U_\theta(x)\|^2 - \mathtt{c} \ .$$

L3. *There exists $L_U \geq 0$ such that for any $x \in \mathbb{R}^d$ and $\theta_1, \theta_2 \in \mathsf{K}$*

$$\|\nabla_x U_{\theta_1}(x) - \nabla_x U_{\theta_2}(x)\| \leq L_U \|\theta_1 - \theta_2\| V(x)^{1/2} \ .$$

The next theorems assert that under **L**1, **L**2 and **L**3 the SOUL algorithm introduced in Section 4.2.2 satisfy **H**1 and **H**2 and therefore Theorem 4.2.1, Theorem 4.2.2, Theorem 4.2.3 and Theorem 4.2.4 can be applied if in addition **A**1, **A**2, **A**3 and **A**4 hold. Under **L**2 define for any $x \in \mathbb{R}^d$

$$V_{\mathrm{e}}(x) = \exp\left[\mathtt{k}_2 \sqrt{1 + \|x\|^2}/4\right] \ .$$

Theorem 4.2.5. *Assume **L**1 and **L**2. Then, **H**1 holds with $V \leftarrow V_{\mathrm{e}}$, $\bar{\gamma} \leftarrow \min(1, 2\mathtt{m}_3^+)$ and $\mathbf{\Psi}(\gamma) = D_4 \sqrt{\gamma}$ where D_4 is given in Proposition 4.2.22.*

Proof. The proof is postponed to Section 4.2.5. □

Theorem 4.2.6. *Assume **L**1, **L**2, **L**3 and that for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V_{\mathrm{e}}^{1/4}(x)$. **H**2 holds with $V \leftarrow V_{\mathrm{e}}$ and $\bar{\gamma} \leftarrow \min(1, 2\mathtt{m}_3^+)$ and for any $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$, $\gamma_2 < \gamma_1$,*

$$\mathbf{\Lambda}_1(\gamma_1, \gamma_2) = D_5 \gamma_2^{-1} |\gamma_1 - \gamma_2| \ , \quad \mathbf{\Lambda}_2(\gamma_1, \gamma_2) = D_5 \gamma_2^{1/2} \ ,$$

where D_5 is given in Proposition 4.2.24.

Proof. The proof is postponed to Section 4.2.5. □

4.2.4 Numerical results

We now demonstrate the proposed methodology with three experiments that we have chosen to illustrate a variety of scenarios. Section 4.2.4 presents an application to empirical Bayesian logistic regression, where (4.118) can be analytically shown to be a convex optimization problem with an unique solution $\theta^\star$, and where we benchmark our MLE estimate against the solution obtained by calculating the marginal likelihood $p(y|\theta)$ over a θ-grid by using an harmonic mean estimator. Furthermore, Section 4.2.4 presents a challenging application related to statistical audio compressive sensing analysis, where we use SOUL to estimate a regularization parameter that controls the degree of sparsity enforced, and where a main difficulty is the high-dimensionality of the latent space ($d = 2,900$). Finally, Section 4.2.4 presents an application to a high-dimensional empirical Bayesian logistic regression with random effects for which the optimization problem (4.118) is not convex. All experiments were carried out on an Intel i9-8950HK@2.90GHz workstation running Matlab R2018a.

Bayesian Logistic Regression

In this first experiment we illustrate the proposed methodology with an empirical Bayesian logistic regression problem [Wak13; PSW13]. We observe a set of covariates $\{v_i\}_{i=1}^{d_y} \in \mathbb{R}^d$, and binary responses $\{y_i\}_{i=1}^{d_y} \in \{0,1\}$, which we assume to be conditionally independent realisations of a logistic regression model: for any $i \in \{1,\dots,d_y\}$, y_i given β and v_i has distribution $\mathrm{Ber}(s(v_i^\top \beta))$, where $\beta \in \mathbb{R}^d$ is the regression coefficient, $\mathrm{Ber}(\alpha)$ denotes the Bernoulli distribution with parameter $\alpha \in [0,1]$ and $s(u) = \mathrm{e}^u/(1+\mathrm{e}^u)$ is the cumulative distribution function of the standard logistic distribution. The prior for β is set to be $\mathrm{N}(\theta \mathbf{1}_d, \sigma^2 \,\mathrm{Id}_d)$, the d-dimensional Gaussian distribution with mean $\theta \mathbf{1}_d$ and covariance matrix $\sigma^2 \,\mathrm{Id}_d$, where θ is the parameter we seek to estimate, $\mathbf{1}_d = (1,\dots,1) \in \mathbb{R}^d$, $\sigma^2 = 5$ and Id_d is the d-dimensional identity matrix. Following an empirical Bayesian approach, the parameter θ is computed by maximum marginal likelihood estimation using Algorithm 6 with the marginal likelihood $p(y|\theta)$ given by

$$p(y|\theta) = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \left\{ \prod_{i=1}^{d_y} s(v_i^\top \beta)^{y_i} (1 - s(v_i^\top \beta))^{1-y_i} \right\} \times \mathrm{e}^{-\frac{\|\beta - \theta \mathbf{1}_d\|^2}{2\sigma^2}} \mathrm{d}\beta \,. \tag{4.135}$$

Lemma 4.2.25 shows that (4.135) is log-concave with respect to θ. We use the proposed SOUL methodology to estimate $\theta^\star$ for the Wisconsin Diagnostic Breast Cancer dataset[3], for which $d_y = 683$ and $d = 10$, and where we suitably normalise the covariates. In order to assess the quality of our estimation results, we also calculate $p(y|\theta)$ over a grid of values for θ by using a truncated harmonic mean estimator.

To implement Algorithm 6 we derive the log-likelihood function

$$\log p(y|\beta,\theta) = \sum_{i=1}^{d_y} \left\{ y_i v_i^\top \beta - \log(1 + \mathrm{e}^{(v_i^\top \beta)}) \right\} \,,$$

and obtain the following expressions for the gradients used in the MCMC steps (4.123) and SA steps (4.119) respectively

$$\nabla_\beta \log p(\beta|y,\theta) = \sum_{i=1}^{d_y} \left\{ y_i v_i - s(v_i^\top \beta) v_i \right\} - \frac{(\beta - \theta \mathbf{1}_d)}{\sigma^2} \,,$$

[3] Available online: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

$$\nabla_\theta p(\beta, y|\theta) = \langle \mathbf{1}_d, \beta - \theta \mathbf{1}_d \rangle / \sigma^2 \ .$$

For the MCMC steps, we use a fixed stepsize $\gamma_n = 8.34 \times 10^{-5}$, and batch size $m_n = 1$, for any $n \in \mathbb{N}$. On the other hand, we consider for the SA steps, the sequence of stepsizes $\delta_n = 60/n^{0.8}$, $\mathsf{K} = [-100, 100]$ and $\theta_0 = 0$. Finally, we first run 100 burn-in iterations with fixed $\theta_n = \theta_0$ to warm-up the Markov chain, followed by 50 iterations of Algorithm 6 to warm-up the iterates. This procedure is then followed by $N = 10^6$ iterations of Algorithm 6 to compute $\hat{\theta}_N$.

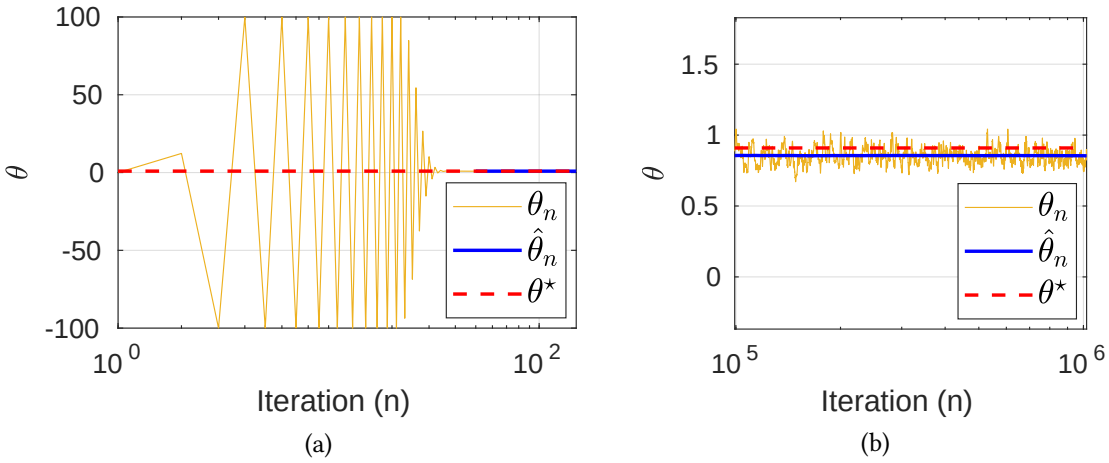


Figure 4.4: Bayesian logistic regression - Evolution of the iterates $\hat{\theta}_n$ and θ_n for the proposed method during (a) burn-in phase and (b) convergence phase. An estimate of $\theta^\star$, the true maximizer of $p(y|\theta)$, is plotted as a reference.

Figure 4.4(a) shows the evolution of the iterates θ_n during the first 100 iterations. Observe that the sequence initially oscillates, and then stabilises close to $\theta^\star$ after approximately 50 iterations. Figure 4.4(b) presents the iterates θ_n for $n = 10^5, \ldots, 10^6$. For completeness, Figure 4.5 shows the histograms corresponding to the marginal posteriors $p(\beta_j|y, v, \hat{\theta}_N)$, for $j = 1, \ldots, 10$, obtained as a by-product of Algorithm 6. In order to verify that the obtained estimate $\hat{\theta}_N$ is close to the true MLE $\theta^\star$ we use a

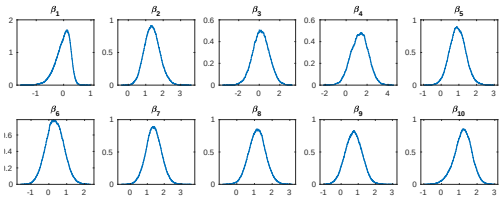


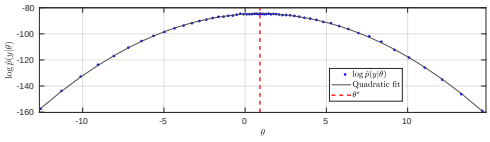
Figure 4.5: Bayesian logistic regression - Normalised histograms of each component of β obtained with 2×10^6 Monte Carlo samples.

truncated harmonic mean estimator (THME) [RW09] to calculate the marginal likelihood $p(y|\theta)$ for a range of values of θ. Although obtaining the THME is usually computationally expensive, it is viable in this particular experiment as β is low-dimensional. More precisely, given n samples $(\beta_i)_{i \in \{1,\ldots,n\}}$ from $p(\beta|y, \theta)$, we obtain an approximation of $p(y|\theta)$ by computing

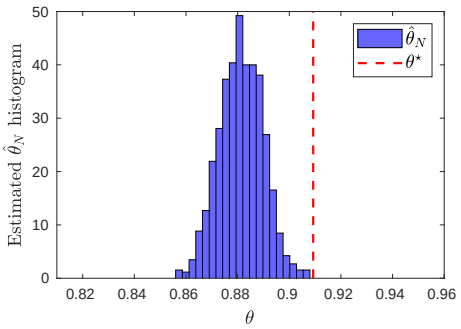
$$\hat{p}(y|\theta) = n \, \mathrm{Vol}(\mathsf{A}) \Bigg/ \left(\sum_{k=1}^{n} \frac{\mathbb{1}_{\mathsf{A}}(\beta_k)}{p(\beta_k, y|\theta)} \right) \ ,$$

where A is a d-dimensional ball centered at the posterior mean $\bar{\beta} = n^{-1} \sum_{k=1}^{n} \beta_k$, and with radius set such that $n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\mathsf{A}}(\beta_i) \approx 0.4$. Using $n = 6 \times 10^5$ samples, we obtain the approximation shown in Figure 4.6(a), where in addition to the estimated points we also display a quadratic fit (corresponding to a Gaussian fit in linear scale), which we use to obtain an estimate of $\theta^\star$ (the obtained log-likelihood values are small because the dataset is large ($d_y = 683$)).

To empirically study the estimation error involved, we replicate the experiment 10^3 times. Figure 4.6 shows the obtained histogram of $\{\hat{\theta}_{N,i}\}_{i=1}^{1000}$, where we observe that all these estimators are very close to the true maximizer $\theta^\star$. Besides, note that the distribution of the estimation error is close to a Gaussian distribution, as expected for a maximum likelihood estimator. Also, there is a small estimation bias of the order of 3%, which can be attributed to the discretization error of SDE (4.122), and potentially to a small error in the estimation of $\theta^\star$. We conclude this experiment by using SOUL to perform a predictive



(a)



(b)

Figure 4.6: Bayesian logistic regression - (a) Estimated points of the marginal log-likelihood $\log \hat{p}(y|\theta)$ with quadratic fit (corresponding to a Gaussian fit in linear scale). (b) Normalised histogram of $\hat{\theta}_N$ for 1000 repetitions of the experiment. An estimate of $\theta^\star$, the maximizer of $\hat{p}(y|\theta)$, is plotted as a reference.

empirical Bayesian analysis on the binary responses. We split the original dataset into an 80% training set $(y^{\text{train}}, v^{\text{train}})$ of size $d_{\text{train}} = 546$, and a 20% test set $(y^{\text{test}}, v^{\text{test}})$ of size $d_{\text{test}} = 137$, and use SOUL to draw samples from the predictive distribution $p(y^{\text{test}}|y^{\text{train}}, v^{\text{train}}, v^{\text{test}}, \hat{\theta}_N)$. More precisely, we use SOUL to simultaneously calculate $\hat{\theta}_N$ and simulate from $p(\beta|y^{\text{train}}, v^{\text{train}}, \hat{\theta}_N)$, followed by simulation from $p(y^{\text{test}}|\beta, y^{\text{train}}, v^{\text{train}}, v^{\text{test}})$. We then estimate the maximum-a-posteriori predictive response $\hat{y}^{\text{test}}$, and measure prediction accuracy against the test dataset by computing the error

$$\epsilon = \|y^{\text{test}} - \hat{y}^{\text{test}}\|_1 / d_{\text{test}} = \sum_{i=1}^{d_{\text{test}}} \left| y_i^{\text{test}} - \hat{y}_i^{\text{test}} \right| / d_{\text{test}} \ ,$$

and obtain $\epsilon = 2.2\%$. For comparison, Figure 4.7 below reports the error ϵ as a function of θ (the discontinuities arise because of the highly non-linear nature of the model). Observe that the estimated $\hat{\theta}_N$ produces a model that has a very good performance in this regard.

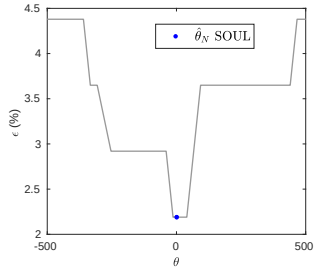


Figure 4.7: Bayesian logistic regression - Percentage of mislabelled binary observations in terms of θ. In blue we show the value of $\hat{\theta}_N$ obtained with Algorithm 6.

Statistical audio compression

Compressive sensing techniques exploit sparsity properties in the data to estimate signals from fewer samples than required by the Nyquist–Shannon sampling theorem [CW08; CW08]. Many real-world data admit a sparse representation on some basis or dictionary. Formally, consider an ℓ-dimensional time-discrete signal $z \in \mathbb{R}^\ell$ that is sparse in some dictionary $\mathbf{\Psi} \in \mathbb{R}^{\ell \times d}$, i.e, there exists a latent vector $x \in \mathbb{R}^d$ such that $z = \mathbf{\Psi} x$ and $\|x\|_0 = \sum_{i=1}^d \mathbb{1}_{\mathbb{R}^*}(x_i) \ll \ell$. This prior assumption can be modelled by using a smoothed-Laplace distribution [LJ12]

$$p(x|\theta) \propto \exp\left(-\theta \sum_{i=1}^d h_\lambda(x_i)\right) , \tag{4.136}$$

where h_λ is the Huber function given for any $u \in \mathbb{R}$ by

$$h_\lambda(u) = \begin{cases} u^2/2 & \text{if } |u| \leq \lambda \ , \\ \lambda(|u| - \lambda/2) & \text{otherwise} \ . \end{cases} \tag{4.137}$$

Acquiring z directly would call for measuring ℓ univariate components. Instead, a carefully designed measurement matrix $\mathbf{M} \in \mathbb{R}^{p \times \ell}$, with $p \ll \ell$, is used to directly observe a "compressed" signal $\mathbf{M}z$, which only requires taking p measurements. In addition, measurements are typically noisy which results in an observation $y \in \mathbb{R}^p$ modeled as $y = \mathbf{M}z + w$ where we assume that the noise w has distribution $\mathrm{N}(0, \sigma^2 \mathrm{Id}_p)$, and therefore the likelihood function is given by

$$p(y|x) \propto \exp\left(-\|y - \mathbf{M}\mathbf{\Psi}x\|_2^2 / (2\sigma^2)\right) ,$$

leading to the posterior distribution

$$p(x|y) \propto \exp\left(-\|y - \mathbf{M}\mathbf{\Psi}x\|_2^2 / (2\sigma^2) - \theta \sum_{i=1}^d h_\lambda(x_i)\right) .$$

To recover z from y, we then compute the maximum-a-posteriori estimate

$$\hat{x}_{\mathrm{MAP}} \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \|y - \mathbf{M}\mathbf{\Psi}x\|_2^2 / 2\sigma^2 + \theta \sum_{i=1}^d h_\lambda(x_i) \right\} , \tag{4.138}$$

and set $\hat{z}_{\mathrm{MAP}} = \mathbf{\Psi}\hat{x}_{\mathrm{MAP}}$.

Following decades of active research, there are now many convex optimization algorithms that can be used to efficiently solve (4.138), even when d is very large [CP16; Mon17]. However, the selection of the value of θ in (4.138) remains a difficult open problem. This parameter controls the degree of sparsity of x and has a strong impact on estimation performance.

A common heuristic within the compressive sensing community is to set $\theta_{\text{cs}} = 0.1 \times \|(\mathbf{M}\mathbf{\Psi})^\intercal y\|_\infty / \sigma^2$, where for any $z \in \mathbb{R}^\ell$, $\|z\|_\infty = \max_{i \in \{1,\dots,\ell\}} |z_i|$, as suggested in [Kim+07] and [FNW07]; however, better results can arguably be obtained by adopting a statistical approach to estimate θ.

The Bayesian framework offers several strategies for estimating θ from the observation y. In this experiment we adopt an empirical Bayesian approach and use SOUL to compute the MLE $\theta^\star$, which is challenging given the high-dimensionality of the latent space.

To illustrate this approach, we consider the audio experiment proposed in [BNE10] for the *"Mary had a little lamb"* song. The MIDI-generated audio file z has $\ell = 319,725$ samples, but we only have access to a noisy observation vector y with $p = 456$ random time points of the audio signal, corrupted by additive white Gaussian noise with $\sigma = 0.015$. The latent signal x has dimension $d = 2,900$ and is related to z by a dictionary matrix $\mathbf{\Psi}$ whose row vectors correspond to different piano notes lasting a quarter-second long [4]. The parameter λ for the prior (4.136) is set to $\lambda = 4 \times 10^{-5}$. We used the heuristic θ_{cs} as the initial value for θ in our algorithm. To solve the optimization problem (4.138) we use the Gradient Projection for Sparse Reconstruction (GPSR) algorithm proposed in [FNW07]. We use this solver because it is the one used in the online MATLAB demonstration of [BNE10], however, more modern algorithms could be used as well. We implemented Algorithm 6 using a fixed stepsize $\gamma_n = 6.9 \times 10^{-6}$, a fixed batch size $m_n = 1$, $\delta_n = 20\, n^{-0.8}/d = 0.0069\, n^{-0.8}$ and 100 burn-in iterations.

The algorithm converged in approximately 500 iterations, which were computed in only 325 milliseconds. Figure 4.8 (left), shows the first 250 iterations of the sequence θ_n and of the weighted average $\hat{\theta}_n$. Again, observe that the iterates oscillate for a few iterations and then quickly stabilise. Finally, to assess the quality of the estimate $\hat{\theta}_N$, Figure 4.8 (right) presents the reconstruction mean squared error as a function of θ. The error is measured with respect to the reconstructed signal and is given by $\text{MSE}(\hat{x}_{\text{MAP}}) = \|z^\star - \mathbf{\Psi}\hat{x}_{\text{MAP}}\|_2^2/\ell$, where $z^\star$ is the true audio signal. Observe that the estimated value $\hat{\theta}_N$ is very close to the value that minimizes the estimation error, and significantly outperforms the heuristic value θ_{cs} commonly used by practitioners.

Sparse Bayesian logistic regression with random effects

Following on from the Bayesian logistic regression in Section 4.2.4, where $p(y|\theta)$ is log-concave and hence $\theta^\star$ unique, we now consider a significantly more challenging sparse Bayesian logistic regression with random effects problem. In this experiment $p(y|\theta)$ is no longer log-concave, so SOUL can potentially get trapped in local maximizers. Furthermore, the dimension of θ in this experiment is very large ($d_\theta = 1001$), making the MLE problem even more challenging. This experiment was previously considered by [AFM17] and we replicate their setup.

Let $\{y_i\}_{i=1}^{d_y} \in \{0,1\}$ be a vector of binary responses which can be modelled as d_y conditionally independent realisations of a random effect logistic regression model,

$$y_i | x \sim \text{Ber}\left(s(v_i^\top \beta + \sigma z_i^\top x)\right) \;, \quad i \in \{1, \dots, d_y\} \;,$$

where $v_i \in \mathbb{R}^p$ are the covariates, $\beta \in \mathbb{R}^p$ is the regression vector, $z_i \in \mathbb{R}^d$ are (known) loading vectors, x are random effects and $\sigma > 0$. In addition, recall that $\text{Ber}(\alpha)$ denotes the Bernoulli distribution with

[4] Each quarter-second sound can have one of 100 possible frequencies and be in 29 different positions in time.

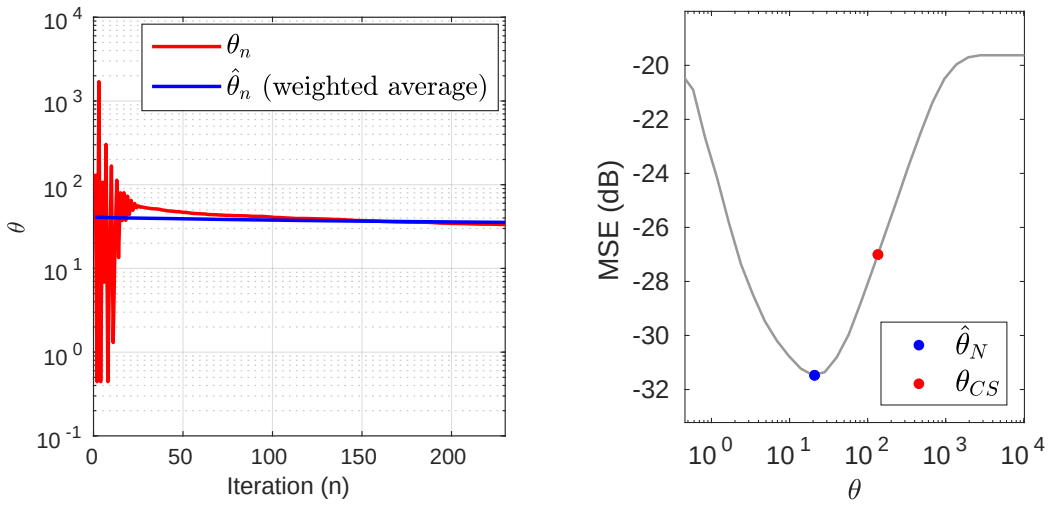


Figure 4.8: Statistical audio compression - Evolution of the the iterate θ_n and $\hat{\theta}_n$ with $\sigma = 0.015$ in log scale (left). Reconstruction mean squared error (MSE) in dB as a function of the θ (right).

parameter $\alpha \in [0,1]$ and $s(u) = \mathrm{e}^u/(1+\mathrm{e}^u)$ is the cumulative distribution function of the standard logistic distribution. The goal is to estimate the unknown parameters $\theta = (\beta, \sigma) \in \mathbb{R}^p \times (0, +\infty)$ directly from $\{y_i\}_{i=1}^{d_y}$, without knowing the value of x, which we assume to follow a standard Gaussian distribution, *i.e.* $p(x) = \exp\{-\|x\|_2^2/2\}/(2\pi)^{d/2}$. We estimate θ by MLE using Algorithm 6 to maximize (4.118), with marginal likelihood given by

$$p(y|\theta) = \int_{\mathbb{R}^d} \prod_{i=1}^{d_y} s(v_i^\top \beta + \sigma z_i^\top x)^{y_i} \times (1 - s(v_i^\top \beta + \sigma z_i^\top x))^{1-y_i} p(x) \mathrm{d}x \ ,$$

and we use the penalty function

$$g(\theta) = \sum_{j=1}^{d} h_\lambda(\beta_j) \ , \tag{4.139}$$

where h_λ is the Huber function defined in (4.137).

We follow the procedure described in [AFM17] to generate the observations $\{y_i\}_{i=1}^{d_y}$, with $d_y = 500$, $p = 1000$ and $d = 5$[5]. The vector of regressors β_{true} is generated from the uniform distribution on $[1,5]$ and 98% of its coefficients are randomly set to zero. The variance σ_{true} of the random effect is set to 0.1, and the projection interval for the estimated σ is $[10^{-5}, +\infty)$. Finally, the parameter λ in (4.139) is set to $\lambda = 30$. We emphasize at this point that θ is high-dimensional in this experiment ($p = 1001$), making the estimation problem particularly challenging.

The conditional log-likelihood function for this model is

$$\log p(y|x,\theta)$$

[5]We renamed some symbols for notation consistency. What we denote by v_i, x, d_y and d, is denoted in [AFM17] by x_i, $\mathbf{U}$, N and q respectively.

$$= \sum_{i=1}^{d_y} \left\{ y_i (v_i^\top \beta + \sigma z_i^\top x) - \log(1 + \mathrm{e}^{v_i^\top \beta + \sigma z_i^\top x}) \right\} .$$

To implement Algorithm 6 we use the gradients

$$\nabla_x \log p(x|y,\theta) = \sum_{i=1}^{d_y} \left\{ \sigma z_i (y_i - s(v_i^\top \beta + \sigma z_i^\top x)) \right\} - x ,$$

$$\nabla_\theta \log p(x,y|\theta) = \sum_{i=1}^{d_y} \left\{ (y_i - s(v_i^\top \beta + \sigma z_i^\top x)) \begin{bmatrix} v_i \\ z_i^\top x \end{bmatrix} \right\} .$$

Finally the gradient of the penalty function is given by

$$\frac{\partial}{\partial \beta_i} g(\theta) = \begin{cases} \beta_i & |\beta_i| \leq \lambda \\ \lambda \ \mathrm{sign}(\beta_i), & |\beta_i| > \lambda \end{cases} , \qquad \frac{\partial}{\partial \sigma} g(\theta) = 0 ,$$

where sign denotes the sign function, *i.e.* for any $s \in \mathbb{R}$, $\mathrm{sign}(s) = |s|/s$ if $s \neq 0$, and $\mathrm{sign}(s) = 0$ otherwise.

We use $\gamma_n = 0.01$, $\delta_n = n^{-0.95}/d = 0.2 \times n^{-0.95}$, a fixed batch size $m_n = 1$, $\beta_0 = \mathbf{1}_p$ and $\sigma_0 = 1$ as initial values. Moreover, we perform 10^4 burn-in iterations with a fixed value of $\theta_0 = (\beta_0, \sigma_0)$ to warm-up the Markov chain, and further 600 iterations of Algorithm 6 to warm-start the iterates. Following on from this, we run $N = 5 \times 10^4$ iterations of Algorithm 6 to compute $\hat{\theta}_N$. Computing this estimates required 25 seconds in total.

Figure 4.9 shows the evolution of the iterates throughout iterations, where we used $\|\hat{\beta}_n\|_0$ as a summary statistic to track the number of active components. Because the Huber penalty (4.137) does not enforce exact sparsity on β, to estimate the number of active components we only consider values that are larger than a threshold τ (we used $\tau = 0.005$).

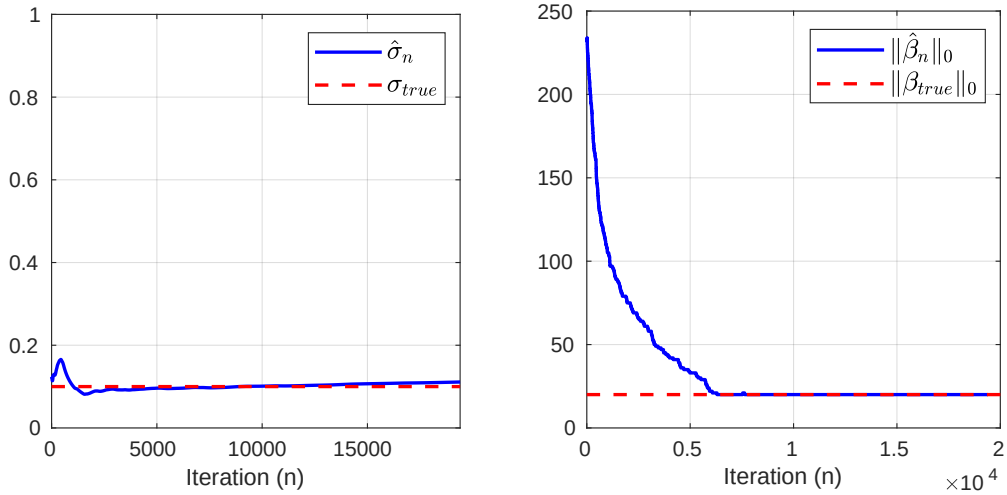


Figure 4.9: Sparse Bayesian logistic regression with random effects - Evolution of the $\|\hat{\beta}_n\|_0$ and of the iterate $\hat{\sigma}_n$ for the proposed method. The true values are plotted in red as a reference.

From Figure 4.9 we observe that $\hat{\sigma}_n$ converges to a value that is very close to σ_{true}, and that the number of active components is also accurately estimated. Moreover, Figure 4.10 shows that most active components were correctly identified. We also observe that $\hat{\beta}_n$ stabilizes after approximately 6300 iterations, which correspond to 6300 Monte Carlo samples as m_n=1. This is in close agreement with the results presented in [AFM17, Figure 5], where they observe stabilization after a similar number of iterations of their highly specialised Polya-Gamma sampler.

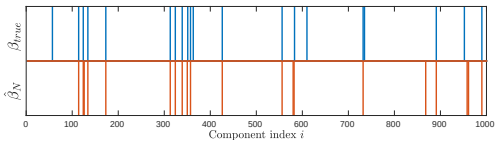


Figure 4.10: Sparse Bayesian logistic regression with random effects - Support of the estimated $\hat{\beta}_N$ compared with the support of β_{true}.

It is worth emphasising at this point that [AFM17] considers the non-smooth penalty $g(\theta) = \lambda\|\beta\|_1$ instead of (4.139). Consequently, instead of using the gradient of g, they resort to the so-called proximal operator of g [CP16]. The generalisation of the SOUL methodology proposed in this paper to models that have non-differentiable terms is addressed in [VP18; Vid+19].

4.2.5 Proofs and additional results

Non-convex objective function

In this section we turn to the case where f is non-convex. We recall that the normal space of a sub-manifold $\mathcal{M} \subset \mathbb{R}^p$ at point θ is given by

$$\mathrm{N}(\theta, \mathcal{M}) = \begin{cases} \mathrm{T}(\theta, \mathcal{M})^{\perp} & \text{if } \theta \in \mathcal{M} \ ; \\ \{0\} & \text{otherwise} \ , \end{cases}$$

where $\mathrm{T}(\theta, \mathcal{M})$ is the tangent space of the sub-manifold $\mathcal{M}$ at point x, see [Aub01].

Theorem 4.2.7. *Assume* **A**1, **A**2, **A**3 *and that* K *is a* p *dimensional connected differentiable manifold with boundary and continuously differentiable outer normal. Let* $\bar{\gamma} > 0$, $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ *be sequences of non-increasing positive real numbers and* $(m_n)_{n\in\mathbb{N}}$ *be a sequence of positive integers such that* $\sup_{n\in\mathbb{N}} \delta_n < 1/\mathcal{L}$, $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$ *and* (4.127) *are satisfied. Let* $\{(X_k^n)_{k\in\{0,\ldots,m_n\}} \ : \ n \in \mathbb{N}\}$ *be given by* (4.126). *Assume in addition that* **H**1 *is satisfied. Then* $(\theta_n)_{n\in\mathbb{N}}$ *defined by* (4.126) *converges a.s. to some* $\theta^\star \in \{\theta \in \mathsf{K} \ : \ \nabla f(\theta) + \mathbf{n} = 0, \ \mathbf{n} \in \mathrm{N}(\theta, \partial\mathsf{K})\}$.

Proof. The proof is an application of [KY03, Chapter 5, Theorem 2.3] using the decomposition of the error term considered in the proof of Theorem 4.2.1 and Theorem 4.2.3. Indeed we decompose the error term η_n defined by (4.140) as $\eta_n = \delta M_n + B_n$, where δM_n is a martingale increment. Then, we only need to show that the following sums converge

$$\sum_{k=0}^{n} \delta_{k+1}^2 \mathbb{E}\left[\|\delta M_k\|^2\right] \ , \qquad \sum_{k=0}^{n} \delta_{k+1} \mathbb{E}\left[\|B_k\|\right] \ ,$$

which is established in Lemma 4.2.10 and Lemma 4.2.11. □

Postponed proofs

We first derive the following technical lemmas.

Lemma 4.2.8. *Let $t \in (0,1)$ and $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ then $\sum_{n\in\mathbb{N}} t^{n\gamma} \leq t^{-\bar{\gamma}} \log^{-1}(1/t)\gamma^{-1}$ and $\sum_{n\in\mathbb{N}} nt^{n\gamma} \leq t^{-\bar{\gamma}} \log^{-2}(1/t)\gamma^{-2}$.*

Proof. Let $t \in (0,1)$ and $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$. Using that $e^u - 1 \leq ue^u$ for all $u \geq 0$, we have

$$\sum_{n\in\mathbb{N}} t^{n\gamma} = -(t^\gamma - 1)^{-1} \leq -\gamma^{-1} \log^{-1}(t) \exp(-\log(t)\gamma) \leq t^{-\bar{\gamma}} \log^{-1}(1/t)\gamma^{-1} ,$$

and

$$\sum_{n\in\mathbb{N}} nt^{n\gamma} = t^\gamma (t^\gamma - 1)^{-2} \leq t^\gamma \{\gamma^{-1} \log^{-1}(t) \exp(-\log(t)\gamma)\}^2 \leq t^{-\bar{\gamma}} \log^{-2}(1/t)\gamma^{-2} ,$$

which completes the proof. □

Lemma 4.2.9. *For any probability measures μ, ν on $\mathcal{B}(\mathbb{R}^d)$, measurable function $V : \mathbb{R}^d \to [1, +\infty)$ such that $\mu(V) + \nu(V) < +\infty$ and $a \in (0,1)$, we have*

$$\|\mu - \nu\|_{V^a} \leq 2\|\mu - \nu\|_V^a .$$

Proof. Let $a \in (0,1]$. The statement is trivial if $\mu = \nu$. We just need to consider the case where $\mu \neq \nu$. Define $\xi = |\mu - \nu| / (|\mu - \nu| (\mathbb{R}^d))$. Using [Dou+18, Definition D.3.1] we get that

$$\begin{aligned} \|\mu - \nu\|_{V^a} &= (1/2)\xi(V^a) \times |\mu - \nu| (\mathbb{R}^d) \\ &\leq (1/2)\xi(V)^a \times |\mu - \nu| (\mathbb{R}^d) \\ &\leq 2^{a-1} \|\mu - \nu\|_V^a \times [|\mu - \nu| (\mathbb{R}^d)]^{1-a} , \end{aligned}$$

which concludes the proof using that $a \leq 1$. □

Jensen's inequality implies that **H**1-(i) holds for $V \leftarrow V^a$ with $a \in (0,1]$ since $A_1 \geq 1$. Lemma 4.2.9 implies that **H**1-(ii) holds replacing V by V^a, ρ by ρ^a and A_2 by $2A_2$. Similarly **H**1-(iii) holds replacing V by V^a and $\mathbf{\Psi}(\gamma)$ by $2\mathbf{\Psi}(\gamma)$.

Proof of Theorem 4.2.1 Consider $(\eta_n)_{n\in\mathbb{N}}$ defined for any $n \in \mathbb{N}$ by

$$\eta_n = m_n^{-1} \sum_{k=1}^{m_n} \{H_{\theta_n}(X_k^n) - \pi_{\theta_n}(H_{\theta_n})\} . \tag{4.140}$$

The proof of Theorem 4.2.1 relies on the two following lemmas. We consider the following decomposition for any $n \in \mathbb{N}$, $\eta_n = \eta_n^{(1)} + \eta_n^{(2)}$, where

$$\eta_n^{(1)} = \mathbb{E}[\eta_n | \mathcal{F}_{n-1}] , \qquad \eta_n^{(2)} = \eta_n - \mathbb{E}[\eta_n | \mathcal{F}_{n-1}] . \tag{4.141}$$

We now give upper bounds on $\mathbb{E}[\|\eta_n^{(1)}\|]$, $\mathbb{E}[\|\eta_n^{(1)}\|^2]$ and $\mathbb{E}[\|\eta_n^{(2)}\|^2]$.

Lemma 4.2.10. *Assume* **A**1, **A**2, **A**3, **H**1 *and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$. *Then we have for any* $n \in \mathbb{N}$

$$\mathbb{E}\left[\|\eta_n^{(1)}\|\right] \leq B_1 \mathbb{E}\left[V^{1/2}(X_0^0)\right]/(m_n\gamma_n) + \mathbf{\Psi}(\gamma_n) \; ;$$
$$\mathbb{E}\left[\|\eta_n^{(1)}\|^2\right] \leq 2B_1^2 \mathbb{E}\left[V(X_0^0)\right]/(m_n\gamma_n)^2 + 2\mathbf{\Psi}(\gamma_n)^2 \; ,$$

with

$$B_1 = 2A_1A_2\rho^{-\bar{\gamma}}/\log(1/\rho) \; .$$

Proof. Using the definition of $(\mathcal{F}_n)_{n\in\mathbb{N}}$, see (4.124), the Markov property, **H**1-(ii)-(iii), Lemma 4.2.9, Jensen's inequality and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$, we have for any $n \in \mathbb{N}^\star$

$$\begin{aligned}
\|\mathbb{E}\left[\eta_n|\mathcal{F}_{n-1}\right]\| &\leq m_n^{-1}\sum_{k=1}^{m_n}\left\|\mathrm{K}_{\gamma_n,\theta_n}^k H_{\theta_n}(X_0^n) - \pi_{\theta_n}\left(H_{\theta_n}\right)\right\| \\
&\leq m_n^{-1}\sum_{k=1}^{m_n}\left\|\left|\delta_{X_0^n}\mathrm{K}_{\gamma_n,\theta_n}^k - \pi_{\theta_n}\right|\left(H_{\theta_n}\right)\right\| \\
&\leq m_n^{-1}\sum_{k=1}^{m_n}\left|\delta_{X_0^n}\mathrm{K}_{\gamma_n,\theta_n}^k - \pi_{\theta_n}\right|\left(\|H_{\theta_n}\|\right) \\
&\leq m_n^{-1}\sum_{k=1}^{m_n}\left\{\|\delta_{X_0^n}\mathrm{K}_{\gamma_n,\theta_n}^k - \pi_{\gamma_n,\theta_n}\|_{V^{1/2}}\right\} + \|\pi_{\gamma_n,\theta_n} - \pi_{\theta_n}\|_{V^{1/2}} \\
&\leq m_n^{-1}\sum_{k=1}^{m_n}\left\{2A_2\rho^{k\gamma_n}V^{1/2}(X_{m_n}^n) + \mathbf{\Psi}(\gamma_n)\right\} \\
&\leq \frac{2A_2\rho^{-\bar{\gamma}}V^{1/2}(X_{m_n}^n)}{\log(1/\rho)\gamma_n m_n} + \mathbf{\Psi}(\gamma_n) \; ,
\end{aligned}$$

where for the last inequality we have used Lemma 4.2.8. In a similar manner, we have

$$\left\|\mathbb{E}\left[\eta_0|X_0^0\right]\right\| \leq \frac{2A_2\rho^{-\bar{\gamma}}V^{1/2}(X_0^0)}{\log(1/\rho)\gamma_0 m_0} + \mathbf{\Psi}(\gamma_0) \; .$$

We conclude using **H**1-(i) and that $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$. □

Lemma 4.2.11. *Assume* **A**1, **A**2, **A**3, **H**1 *and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$. *Then we have for any* $n \in \mathbb{N}$

$$\mathbb{E}\left[\|\eta_n^{(2)}\|^2\right] \leq B_2 m_n^{-2}\gamma_n^{-1}\left(m_n + \gamma_n^{-1}\mathbb{E}\left[V(X_0^0)\right]\right) \; ,$$

with $B_2 = 2(1+\bar{\gamma})^2\max(B_{2,1}, B_{2,2})$ *and*

$$B_{2,1} = 24A_2^2(1-\rho^{1/2})^{-2}A_3 \; ,$$
$$B_{2,2} = 4A_1\left[1 + 6A_2^2(1-\rho^{1/2})^{-2}\left\{A_2(1-\rho)^{-1} + 2\right\} + A_2^2\log^{-2}(1/\rho) + A_3^2\right] \; .$$

Proof. Let $n \in \mathbb{N}^\star$. We have using the Cauchy-Schwarz inequality

$$\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}\left\{H_{\theta_n}(X_k^n) - \mathbb{E}\left[H_{\theta_n}(X_k^n)|\mathcal{F}_{n-1}\right]\right\}\right\|^2\right]$$

$$
\begin{aligned}
&\leq 2\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}\{H_{\theta_n}(X_k^n)-\pi_{\gamma_n,\theta_n}(H_{\theta_n})\}\right\|^2\right] \\
&+2\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}\{\mathbb{E}\left[H_{\theta_n}(X_k^n)|\mathcal{F}_{n-1}\right]-\pi_{\gamma_n,\theta_n}(H_{\theta_n})\}\right\|^2\right] \qquad (4.142)
\end{aligned}
$$

Using the Markov property, **H**1-(i)-(ii), Lemma 4.2.9, Lemma 4.2.8 and that for any $\theta\in\Theta$ and $x\in\mathbb{R}^d$, $\|H_\theta(x)\|\leq V^{1/2}(x)$ we obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}\{\mathbb{E}\left[H_{\theta_n}(X_k^n)|\mathcal{F}_{n-1}\right]-\pi_{\gamma_n,\theta_n}(H_{\theta_n})\}\right\|^2\right] \\
&\qquad\leq \mathbb{E}\left[\left|\sum_{k=1}^{m_n}\mathbb{E}\left[\|\delta_{X_0^n}\mathrm{R}_{\gamma_n,\theta_n}-\pi_{\gamma_n,\theta_n}\|_{V^{1/2}}\middle|\mathcal{F}_{n-1}\right]\right|^2\right] \\
&\qquad\leq 4A_2^2\mathbb{E}\left[\left|\mathbb{E}\left[V^{1/2}(X_0^n)\middle|\mathcal{F}_{n-1}\right]\sum_{k=1}^{m_n}\rho^{k\gamma_n/2}\right|^2\right] \\
&\qquad\leq 4A_1A_2^2\gamma_n^{-2}\rho^{-2\bar{\gamma}}\log^{-2}(1/\rho)\mathbb{E}\left[V(X_0^0)\right]\ . \qquad (4.143)
\end{aligned}
$$

We now give an upper-bound on the first term in the right-hand side of (4.142). Consider for any $n\in\mathbb{N}$ the Euclidean division of m_n by $\lceil 1/\gamma_n\rceil$ there exist $q_n\in\mathbb{N}$ and $r_n\in\{0,\ldots,\lceil 1/\gamma_n\rceil-1\}$ such that $m_n=q_n\lceil 1/\gamma_n\rceil+r_n$. Therefore using the Cauchy-Schwarz inequality we can derive the following decomposition

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}H_{\theta_n}(X_k^n)-\pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\sum_{j=1}^{r_n}H_{\theta_n}(X^n_{j+q_n\lceil 1/\gamma_n\rceil})-\pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\quad+2\mathbb{E}\left[\left\|\sum_{j=1}^{\lceil 1/\gamma_n\rceil}\sum_{k=0}^{q_n-1}H_{\theta_n}(X^n_{j+k\lceil 1/\gamma_n\rceil})-\pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\sum_{j=1}^{r_n}H_{\theta_n}(\bar{X}^{j,n}_{q_n})-\pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\quad+2\lceil 1/\gamma_n\rceil\sum_{j=1}^{\lceil 1/\gamma_n\rceil}\mathbb{E}\left[\left\|\sum_{k=0}^{q_n-1}H_{\theta_n}(\bar{X}^{j,n}_k)-\pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \qquad (4.144)
\end{aligned}
$$

Setting for any $j\in\{1,\ldots,\lceil 1/\gamma_n\rceil\}$ and $k\in\{0,\ldots,q_n-1\}$, $\bar{X}_k^{j,n}=X^n_{j+k\lceil 1/\gamma_n\rceil}$. We now bound the two terms in the right-hand side. First, using the Cauchy-Schwarz inequality and **H**1-(i)-(iii), the fact

that $r_n \leq \lceil 1/\gamma_n \rceil$ and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$ we have

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{j=1}^{r_n} H_{\theta_n}(\bar{X}_{q_n}^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\ &\leq r_n \sum_{j=1}^{r_n} \mathbb{E}\left[\left\|H_{\theta_n}(\bar{X}_{q_n}^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\ &\leq \lceil 1/\gamma_n \rceil^2 \left(2A_1 \mathbb{E}\left[V(X_0^0)\right] + 2A_3^2\right) . \end{aligned} \tag{4.145}$$

Now consider the solution of the *Poisson equation* [MT93a, Section 17.4.1] associated with $\mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil}$, $x \mapsto \hat{H}_{\gamma_n,\theta_n}(x)$ defined for any $x \in \mathbb{R}^d$ by

$$\hat{H}_{\gamma_n,\theta_n}(x) = \sum_{\ell \in \mathbb{N}} \left(\mathrm{K}_{\gamma_n,\theta_n}^{\ell \lceil 1/\gamma_n \rceil} H_{\theta_n}(x) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right) .$$

Note that by **H**1-(ii), Lemma 4.2.9 and since for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$, we have that for any $x \in \mathbb{R}^d$

$$\left\|\hat{H}_{\gamma_n,\theta_n}(x)\right\| \leq 2A_2(1-\rho^{1/2})^{-1} V^{1/2}(x) , \tag{4.146}$$

and in addition for any $x \in \mathbb{R}^d$

$$\hat{H}_{\gamma_n,\theta_n}(x) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(x) = H_{\theta_n}(x) - \pi_{\gamma_n,\theta_n}(H_{\theta_n}) .$$

Therefore, we have for any $j \in \{1, \ldots, \lceil 1/\gamma_n \rceil\}$

$$\begin{aligned}\sum_{k=0}^{q_n-1} \left(H_{\theta_n}(\bar{X}_k^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right) &= \sum_{k=0}^{q_n-1} \left(\hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n})\right) \\ &= \sum_{k=0}^{q_n-2} \left(\hat{H}_{\gamma_n,\theta_n}(\bar{X}_{k+1}^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n})\right) \\ &\quad + \hat{H}_{\gamma_n,\theta_n}(\bar{X}_0^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_{q_n-1}^{j,n}) . \end{aligned} \tag{4.147}$$

Combining the Cauchy-Schwarz inequality and (4.147) we obtain that

$$\mathbb{E}\left[\left\|\sum_{k=0}^{q_n-1} H_{\theta_n}(\bar{X}_k^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \leq 3(\mathrm{C}_1 + \mathrm{C}_2) , \tag{4.148}$$

with

$$\begin{aligned}\mathrm{C}_1 &= \mathbb{E}\left[\left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}_0^{j,n})\right\|^2 + \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}_{q_n-1}^{j,n})\right\|^2\right] ; \\ \mathrm{C}_2 &= \mathbb{E}\left[\left\|\sum_{k=0}^{q_n-2} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_{k+1}^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n \rceil} \hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n})\right\|^2\right] . \end{aligned}$$

First, using (4.146) and **H**1-(i) we get that

$$\mathrm{C}_1 \leq 4A_2^2(1-\rho^{1/2})^{-2} \left\{\mathbb{E}\left[V(X_j^n)\right] + \mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n} V(X_{q_n+j-1}^n\right]\right\}$$

$$\leq 8A_1A_2^2(1-\rho^{1/2})^{-2}\mathbb{E}\left[V(X_0^0)\right] \ . \tag{4.149}$$

We now give an upper-bound on C_2. For any $j \in \{1,\dots,r_n\}$ let $(\mathcal{G}_{j,k})_{k\in\{0,q_n-2\}}$ generated by $\mathcal{F}_{n-1}$ and the sequence of random variables $X_0^n,\dots,X_{k\lceil 1/\gamma_n\rceil+j}^n$. Using the Markov property we have for any $k \in \{0,\dots,q_n-2\}$ and $j \in \{1,\dots,r_n\}$

$$\mathbb{E}\left[\hat{H}_{\gamma_n,\theta_n}(X_{k+1}^{j,n})\middle|\mathcal{G}_{j,k}\right] = \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}\hat{H}_{\gamma_n,\theta_n}(X_k^{j,n}) \ .$$

Therefore, for any $j \in \{1,\dots,r_n\}$, $\hat{H}_{\gamma_n,\theta_n}(X_{k+1}^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}\hat{H}_{\gamma_n,\theta_n}(X_k^{j,n})$ is a martingale increment with respect to $(\mathcal{G}_{j,k})_{k\in\{0,q_n-2\}}$, Combining this result with the Markov property implies that for any $k \in \{0,\dots,q_n-2\}$ and $j \in \{1,\dots,r_n\}$,

$$\begin{aligned}
\mathrm{C}_2 &= \sum_{k=0}^{q_n-2}\mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}\left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n}) - \mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}\hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n})\right\|^2\right] \\
&= \sum_{k=0}^{q_n-2}\mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}\left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n})\right\|^2 - \left\|\mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}\hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n})\right\|^2\right] \ .
\end{aligned} \tag{4.150}$$

Define for any $x \in \mathbb{R}^d$, $g_n(x) = \|\hat{H}_{\gamma_n,\theta_n}(x)\|^2$. Using (4.150), **H**1-(ii)-(iii) and (4.146) we obtain that

$$\begin{aligned}
\mathrm{C}_2 &= \sum_{k=0}^{q_n-2}\mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}\left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n})\right\|^2 - \left\|\mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}\hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n})\right\|^2\right] \\
&\leq \sum_{k=0}^{q_n-2}\mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n}^{\lceil 1/\gamma_n\rceil}\left\|\hat{H}_{\gamma_n,\theta_n}(\bar{X}_k^{j,n})\right\|^2\right] \\
&\leq \mathbb{E}\left[\sum_{k=0}^{q_n-2}\mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n}^{(k+1)\lceil 1/\gamma_n\rceil}g_n(\bar{X}_0^{j,n}) - \pi_{\gamma_n,\theta_n}(g_n)\middle|\mathcal{G}_{j,0}\right]\right] + \sum_{k=0}^{q_n-2}\pi_{\gamma_n,\theta_n}(g_n) \\
&\leq \frac{4A_2^2}{(1-\rho^{1/2})^2}\left\{\sum_{k=0}^{q_n-2}\mathbb{E}\left[\mathbb{E}\left[\|\delta_{X_j^n}\mathrm{K}_{\gamma_n,\theta_n}^{(k+1)\lceil 1/\gamma_n\rceil} - \pi_{\gamma_n,\theta_n}\|_V\middle|\mathcal{G}_{j,0}\right]\right]\right. \\
&\qquad\qquad \left. + \sum_{k=0}^{q_n-2}\pi_{\gamma_n,\theta_n}(V)\right\} \\
&\leq 4A_2^2(1-\rho^{1/2})^{-2}\left\{A_2(1-\rho)^{-1}\mathbb{E}\left[V(X_j^n)\right] + q_nA_3\right\} \\
&\leq 4A_2^2(1-\rho^{1/2})^{-2}\left\{A_1A_2(1-\rho)^{-1}\mathbb{E}\left[V(X_0^0)\right] + q_nA_3\right\} \ .
\end{aligned} \tag{4.151}$$

Therefore, using (4.149) and (4.151) in (4.148) we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{k=0}^{q_n-1}H_{\theta_n}(\bar{X}_k^{j,n}) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \\
&\leq 12A_2^2(1-\rho^{1/2})^{-2}\left[\left\{A_1A_2(1-\rho)^{-1}\mathbb{E}\left[V(X_0^0)\right] + q_nA_3\right\} + 2\mathbb{E}\left[V(X_0^0)\right]\right] \ .
\end{aligned} \tag{4.152}$$

As a consequence, using (4.145) and (4.152) in (4.144) we get that

$$\mathbb{E}\left[\left\|\sum_{k=1}^{m_n}H_{\theta_n}(X_k^n) - \pi_{\gamma_n,\theta_n}(H_{\theta_n})\right\|^2\right] \leq 4\left\lceil 1/\gamma_n\right\rceil^2\left(A_1\mathbb{E}\left[V(X_0^0)\right] + A_3^2\right)$$

$$
\begin{aligned}
&+ 24 \lceil 1/\gamma_n \rceil^2 A_2^2 (1-\rho^{1/2})^{-2} \left\{ A_1 \mathbb{E}\left[V(X_0^0)\right] (A_2(1-\rho)^{-1}+2) + q_n A_3 \right\} \\
&\leq \left[\gamma_n^{-2} \left(A_1 \mathbb{E}\left[V(X_0^0)\right] \left[24 A_2^2 (1-\rho^{1/2})^{-2} \left\{ A_2(1-\rho)^{-1}+2 \right\} + 4 \right] + 4A_3^2 \right) \right. \\
&\left. + 24 A_2^2 (1-\rho^{1/2})^{-2} A_3 m_n/\gamma_n \right] (1+\bar{\gamma})^2 \qquad (4.153)
\end{aligned}
$$

Combining (4.143) and (4.153) in (4.142) we obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \sum_{k=1}^{m_n} H_{\theta_n}(X_k^n) - \mathbb{E}\left[H_{\theta_n}(X_k^n) \right] \right\|^2 \right] \leq 8 \gamma_n^{-2} A_1 A_2^2 \rho^{-2\bar{\gamma}} \log^{-2}(1/\rho) \mathbb{E}\left[V(X_0^0)\right] \\
&+ 2 \left[\gamma_n^{-2} \left(A_1 \mathbb{E}\left[V(X_0^0)\right] \left[24 A_2^2 (1-\rho^{1/2})^{-2} \left\{ A_2(1-\rho)^{-1}+2 \right\} + 4 \right] + 4A_3^2 \right) \right. \\
&\left. + 24 A_2^2 (1-\rho^{1/2})^{-2} A_3 m_n/\gamma_n \right] (1+\bar{\gamma})^2 \\
&\leq 2(1+\bar{\gamma})^2 \left(A_1 \mathbb{E}\left[V(X_0^0)\right] \left[24 A_2^2 (1-\rho^{1/2})^{-2} \left\{ A_2(1-\rho)^{-1}+2 \right\} \right. \right. \\
&\left. \left. + 4 \left\{ 1 + A_2^2 \log^{-2}(1/\rho) \right\} \right] + 4A_3^2 \right) \gamma_n^{-2} + 48 A_2^2 (1-\rho^{1/2})^{-2} A_3 (1+\bar{\gamma})^2 (m_n/\gamma_n) \, ,
\end{aligned}
$$

which concludes the proof for $n \neq 0$. The same inequality holds in the case where $n = 0$. □

We now turn to the proof of Theorem 4.2.1.

Proof of Theorem 4.2.1. The proof is an application of [AFM17, Theorem 2, Theorem 3].

(a) To apply [AFM17, Theorem 2], it is enough to show that the following series converge a.s.

$$
\sum_{n=0}^{+\infty} \delta_{n+1} \langle \Pi_\Theta(\theta_n - \delta_{n+1} \nabla f(\theta_n)), \eta_n^{(i)} \rangle \, , \quad \sum_{n=0}^{+\infty} \delta_{n+1} \eta_n^{(i)} \, , \quad \sum_{n=0}^{+\infty} \delta_{n+1}^2 \|\eta_n^{(i)}\|^2 \, .
$$

where $i \in \{1,2\}$ and the sequences $(\eta_n^{(1)})_{n\in\mathbb{N}}$ and $(\eta_n^{(2)})_{n\in\mathbb{N}}$ are given in (4.142).

In the case where $i = 1$, since $(\Pi_\Theta(\theta_n - \delta_{n+1}\nabla f(\theta_n)))_{n\in\mathbb{N}}$ is bounded, we are reduced to proving that a.s. $\sum_{n=0}^{+\infty} \delta_{n+1} \|\eta_n^{(1)}\| < +\infty$. Using (4.127), Lemma 4.2.10 and Fubini-Tonelli's theorem we obtain that

$$
\mathbb{E}\left[\sum_{n\in\mathbb{N}} \delta_{n+1} \|\eta_n^{(1)}\| \right] = \sum_{n\in\mathbb{N}} \delta_{n+1} \mathbb{E}\left[\|\eta_n^{(1)}\| \right] < +\infty \, . \qquad (4.154)
$$

We consider the case where $i = 2$. Let $(S_n)_{n\in\mathbb{N}}$ and $(T_n)_{n\in\mathbb{N}}$ be defined for any $n \in \mathbb{N}$ by $S_n = \sum_{k=0}^n \delta_{k+1} \langle \Pi_\Theta(\theta_k - \delta_{k+1}\nabla f(\theta_k)), \eta_k^{(2)} \rangle$ and $T_n = \sum_{k=0}^n \delta_{k+1} \eta_n^{(2)}$ are $(\mathcal{F}_n)_{n\in\mathbb{N}}$-martingale by definition of $(\eta_n^{(2)})_{n\in\mathbb{N}}$ in (4.142) and $(\mathcal{F}_n)_{n\in\mathbb{N}}$ in (4.124). Therefore, using [Wil91, Section 12.5], the Cauchy-Schwarz inequality and that the sequence $(\Pi_\Theta(\theta_n - \delta_{n+1}\nabla f(\theta_n)))_{n\in\mathbb{N}}$ is bounded, it suffices to show that $\sum_{n=0}^{+\infty} \delta_{n+1}^2 \mathbb{E}[\|\eta_n^{(2)}\|^2] < +\infty$. Using Lemma 4.2.11 we get that

$$
\sum_{n=0}^{+\infty} \delta_{n+1}^2 \mathbb{E}[\|\eta_n^{(2)}\|^2] \leq B_2 \left(\sum_{n=0}^{+\infty} \delta_{n+1}^2/(m_n\gamma_n) + \mathbb{E}\left[V(X_0^0)\right] \sum_{n=0}^{+\infty} \delta_{n+1}^2/(m_n\gamma_n)^2 \right) \, .
$$

Combining this result and (4.154) implies the stated convergence applying [AFM17, Theorem 2].

(b) Applying [AFM17, Theorem 3], the Cauchy-Schwarz inequality and using **A**1 we obtain that a.s. for any $n \in \mathbb{N}$

$$\begin{aligned}
&\sum_{k=1}^{n} \delta_k \left\{ f(\theta_k) - \min_{\Theta} f \right\} \qquad (4.155)\\
&\leq \frac{\|\theta_0 - \theta^\star\|^2}{2} - \sum_{k=0}^{n-1} \delta_{k+1} \langle \Pi_\Theta(\theta_k - \delta_{k+1} \nabla f(\theta_k)) - \theta^\star, \eta_k \rangle + \sum_{k=0}^{n-1} \delta_{k+1}^2 \|\eta_k\|^2 \\
&\leq 2M_\Theta^2 - \sum_{i=1}^{2} \sum_{k=0}^{n-1} \delta_{k+1} \langle \Pi_\Theta(\theta_k - \delta_{k+1} \nabla f(\theta_k)) - \theta^\star, \eta_k^{(i)} \rangle + 2 \sum_{=1}^{2} \sum_{k=0}^{n-1} \delta_{k+1}^2 \|\eta_k^{(i)}\|^2 \ .
\end{aligned}$$

which implies by the proof of (a) that $\sup_{n\in\mathbb{N}} [\sum_{k=1}^{n} \delta_k \{f(\theta_k) - \min_\Theta f\}] < +\infty$ a.s.. The proof is then completed upon dividing (4.155) by $\sum_{k=1}^{n} \delta_k$.

□

Proof of Theorem 4.2.2

Proof. Taking the expectation in (4.155) and using that $\eta_n^{(2)}$ is a martingale increment with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, we get that for every $n \in \mathbb{N}$

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{n} \delta_k \left\{ f(\theta_k) - \min_{\Theta} f \right\} \right] &\leq 2M_\Theta^2 + 2M_\Theta \sum_{k=0}^{n-1} \delta_{k+1} \mathbb{E}\left[\left\| \eta_k^{(1)} \right\| \right] \\
&\quad + 2 \sum_{k=0}^{n-1} \delta_{k+1}^2 \mathbb{E}\left[\left\| \eta_k^{(1)} \right\|^2 \right] + 2 \sum_{k=0}^{n-1} \delta_{k+1}^2 \mathbb{E}\left[\left\| \eta_k^{(2)} \right\|^2 \right] \ .
\end{aligned}$$

Combining this result, Lemma 4.2.10 and Lemma 4.2.11 completes the proof. □

Proof of Theorem 4.2.3

We now introduce some tools needed for the proof. By **A**4 and **H**1-(i)-(ii), for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, there exists a function $\hat{H}_{\gamma,\theta} : \mathbb{R}^d \to \mathbb{R}^{d_\theta}$ solution of the *Poisson equation*,

$$(\mathrm{Id} - \mathrm{K}_{\gamma,\theta}) \hat{H}_{\gamma,\theta} = H_\theta - \pi_{\gamma,\theta}(H_\theta) \ , \qquad (4.156)$$

defined for any $x \in \mathbb{R}^d$ by

$$\hat{H}_{\gamma,\theta}(x) = \sum_{j\in\mathbb{N}} \{ \mathrm{K}_{\gamma,\theta}^j H_\theta(x) - \pi_{\gamma,\theta}(H_\theta) \} \ . \qquad (4.157)$$

Note that using **H**1-(ii) and Lemma 4.2.9 we have for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$

$$\left\| \hat{H}_\theta(x) \right\| \leq C_{\hat{H}} \gamma^{-1} V^{1/4}(x) \ , \qquad C_{\hat{H}} = 8A_2 \log^{-1}(1/\rho) \rho^{-\bar{\gamma}/4} \ . \qquad (4.158)$$

Define for any $n \in \mathbb{N}$

$$\tilde{\eta}_n = H_{\theta_n}(\tilde{X}_{n+1}) - \pi_{\tilde{\theta}_n}(H_{\tilde{\theta}_n}) \ . \qquad (4.159)$$

Using (4.156) an alternative expression of $(\tilde{\eta}_n)_{n\in\mathbb{N}}$ is given for any $n \in \mathbb{N}$ by

$$\begin{aligned}\tilde{\eta}_n &= \hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1}) - \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1}) + \pi_{\gamma_n,\tilde{\theta}_n}(H_{\tilde{\theta}_n}) - \pi_{\tilde{\theta}_n}(H_{\tilde{\theta}_n}) \\ &= \tilde{\eta}_n^{(a)} + \tilde{\eta}_n^{(b)} + \tilde{\eta}_n^{(c)} + \tilde{\eta}_n^{(d)} \ ,\end{aligned}$$

where

$$\begin{aligned}\tilde{\eta}_n^{(a)} &= \hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1}) - \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n) \ , \\ \tilde{\eta}_n^{(b)} &= \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n) - \mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1}) \ , \\ \tilde{\eta}_n^{(c)} &= \mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1}) - \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1}) \ , \\ \tilde{\eta}_n^{(d)} &= \pi_{\gamma_n,\tilde{\theta}_n}(H_{\tilde{\theta}_n}) - \pi_{\tilde{\theta}_n}(H_{\tilde{\theta}_n}) \ .\end{aligned} \tag{4.160}$$

To establish Theorem 4.2.3 we need to get estimates on moments of $\left\|\tilde{\eta}_n^{(i)}\right\|$ for $i \in \{a,b,c,d\}$. It is the matter of the following technical results.

Lemma 4.2.12. *Assume* **A**1, **A**2, **A**3, **H**1 *and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x)$. *Then we have for any* $n \in \mathbb{N}$, $\mathbb{E}\left[\|\tilde{\eta}_n\|^2\right] \le C_1$, *with*

$$C_1 = 2A_1\mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right] + 2\sup_{\theta\in\Theta}\|\nabla f(\theta)\|^2 \ .$$

Proof. Using (4.159), that $\|x+y\|^2 \le 2(\|x\|^2 + \|y\|^2)$ for any $x, y \in \mathbb{R}^d$, **A**1, **A**2, **A**3 and **H**1-(i) and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/2}(x)$, we get for any $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\eta}_k\|^2\right] &\le 2\mathbb{E}[\|H_{\tilde{\theta}_k}(\tilde{X}_{k+1})\|^2] + 2\left[\pi_{\tilde{\theta}_k}(\|H_{\tilde{\theta}_k}\|)\right]^2 \\ &\le 2A_1\mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right] + 2\sup_{\theta\in\Theta}\|\nabla f(\theta)\|^2 \ .\end{aligned}$$

□

Lemma 4.2.13. *Assume* **A**1, **A**2, **A**3, **A**4, **H**1, **H**2 *and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x)$. *Then we have for any* $n \in \mathbb{N}$, $\mathbb{E}\left[\left\|\tilde{\eta}_n^{(a)}\right\|^2\right] \le \tilde{C}_1\gamma_n^{-2}$, *with*

$$\tilde{C}_1 = A_1 C_{\hat{H}}^2 \mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right] \ .$$

Proof. By (4.160), using (4.158) and **H**1-(i) we get that for any $n \in \mathbb{N}^\star$

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[\left\|\tilde{\eta}_n^{(a)}\right\|^2\middle|\mathcal{F}_n\right]\right] \\ &\le \mathbb{E}\left[\mathbb{E}\left[\left\|\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_{n+1})\right\|^2\middle|\mathcal{F}_n\right]\right] - \mathbb{E}\left[\left\|\mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n)\right\|^2\right] \\ &\le A_1 C_{\hat{H}}^2 \gamma_n^{-2}\mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right] \ ,\end{aligned}$$

which concludes the proof. □

Lemma 4.2.14. *Assume* **A**1, **A**2, **A**3, **H**1 *and that for any* $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x)$. *Then the following statements hold.*

(a) There exists $C_3 \geq 0$ such that for any $n \in \mathbb{N}$ and $\theta \in \Theta$

$$\mathbb{E}\left[\left\|\sum_{k=0}^{n} \delta_{k+1}\langle a_{k+1}, \tilde{\eta}_k^b\rangle\right\|\right] \leq C_3\left[\sum_{k=0}^{n} |\delta_{k+1}-\delta_k|\,\gamma_k^{-1} + \sum_{k=0}^{n}\delta_{k+1}^2\gamma_k^{-1} + (\delta_{n+1}/\gamma_{n+1} - \delta_1/\gamma_1)\right] .$$

with $a_{k+1} = \Pi_\Theta\left[\tilde{\theta}_k - \delta_{k+1}\nabla f(\tilde{\theta}_k)\right] - \theta^\star$, $\theta^\star \in \arg\min_\Theta f$ *and*

$$C_3 = A_1 C_{\hat{H}}(4M_\Theta + \sup_{\theta\in\Theta}\|\nabla f(\theta)\| + 1 + \delta_1 L_f)\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right] .$$

(b) If (4.133) holds then $\sum_{k=0}^{n}\delta_{k+1}\langle a_{k+1}, \tilde{\eta}_k^{(b)}\rangle$ converges a.s..

Proof. By (4.160) we have for any $n \in \mathbb{N}$ and $\theta \in \Theta$

$$\begin{aligned}
&\sum_{k=0}^{n}\delta_{k+1}\langle a_{k+1}, \tilde{\eta}_k^{(b)}\rangle \\
&= \sum_{k=0}^{n}\langle \delta_{k+1}a_{k+1}, \mathrm{K}_{\gamma_k,\tilde{\theta}_k}\hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_k) - \mathrm{K}_{\gamma_{k+1},\tilde{\theta}_{k+1}}\hat{H}_{\gamma_{k+1},\tilde{\theta}_{k+1}}(\tilde{X}_{k+1})\rangle \\
&= \sum_{k=1}^{n}\langle \delta_{k+1}a_{k+1} - \delta_k a_k, \mathrm{K}_{\gamma_k,\tilde{\theta}_k}\hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_k)\rangle \\
&- \langle \delta_{n+1}a_{n+1}, \mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1})\rangle \\
&+ \langle \delta_1 a_1, \mathrm{K}_{\gamma_0,\tilde{\theta}_0}\hat{H}_{\gamma_0,\tilde{\theta}_0}(\tilde{X}_0)\rangle ,
\end{aligned} \tag{4.161}$$

In addition, we have for any $n \in \mathbb{N}$, $\theta \in \Theta$ using **A**1, **A**2, that Π_Θ is non-expansive, (4.131), **H**1-(i) and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/4}(x)$

$$\begin{aligned}
\|\delta_{n+1}a_{n+1} - \delta_n a_n\| &\leq 2M_\Theta\,|\delta_{n+1}-\delta_n| + \delta_{n+1}\,\|a_{n+1}-a_n\| \\
&\leq 2M_\Theta\,|\delta_{n+1}-\delta_n| + (1+\delta_n L_f)\,\|\theta_{n+1}-\theta_n\| + |\delta_{n+1}-\delta_n|\,\|\nabla f(\theta_{n+1})\| \\
&\leq (2M_\Theta + \sup_{\theta\in\Theta}\|\nabla f(\theta)\|)\,|\delta_{n+1}-\delta_n| + \delta_{n+1}^2(1+\delta_{n+1}L_f)V^{1/4}(\tilde{X}_{n+1}) .
\end{aligned} \tag{4.162}$$

(a) Combining (4.161), (4.162), (4.158), the Cauchy-Schwarz inequality and **H**1-(i) we get that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{k=0}^{n}\delta_{k+1}\langle a_k, \tilde{\eta}_k^{[b)}\rangle\right\|\right] \\
&\quad\leq (2M_\Theta + \sup_{\theta\in\Theta}\|\nabla f(\theta)\|)A_1 C_{\hat{H}}\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right]\sum_{k=0}^{n}|\delta_{k+1}-\delta_k|\,\gamma_k^{-1} \\
&\quad+ A_1 C_{\hat{H}}(1+\delta_1 L_f)\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right]\sum_{k=0}^{n}\delta_{k+1}^2\gamma_k^{-1} \\
&\quad+ 2A_1 M_\Theta C_{\hat{H}}\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right]\{\delta_{n+1}/\gamma_{n+1} + \delta_1/\gamma_1\} ,
\end{aligned}$$

which concludes the proof of Lemma 4.2.14-(a).

(b) Assume now (4.133). We show that a.s. the first term in (4.161) is absolutely convergence and the second term converges to 0.

Using (4.162), (4.158), the Cauchy-Schwarz inequality and (4.133) we get that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{+\infty}\left|\langle \delta_{k+1}a_{k+1}-\delta_k a_k, \mathrm{K}_{\gamma_k,\tilde{\theta}_k}\hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_k)\rangle\right|\right] \\
&\leq (2M_\Theta + \sup_{\theta\in\Theta}\|\nabla f(\theta)\|)A_1 C_{\hat{H}}\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right]\sum_{k=0}^{+\infty}|\delta_{k+1}-\delta_k|\,\gamma_k^{-1} \\
&\quad + A_1 C_{\hat{H}}(1+\delta_1 L_f)\mathbb{E}\left[V^{1/4}(\tilde{X}_0)\right]\sum_{k=0}^{+\infty}\delta_{k+1}^2 \\
&< +\infty\ ,
\end{aligned}$$

which implies that $(\langle \delta_{k+1}a_{k+1}-\delta_k a_k, \mathrm{K}_{\gamma_k,\tilde{\theta}_k}\hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_k)\rangle)_{k\in\mathbb{N}}$ is absolutely convergent a.s.. We have that $\mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\|\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1})\|$ is upper-bounded using (4.158) by $\gamma_{n+1}^{-1}C_{\hat{H}}\mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}V^{1/4}(\tilde{X}_{n+1})$. It follows that we have for any $\theta\in\Theta$, $\varepsilon>0$, using the Markov inequality, the Cauchy-Schwarz inequality, (4.158) and (4.133)

$$\begin{aligned}
&\sum_{n\in\mathbb{N}}\mathbb{P}\left(\|a_{n+1}\|\,\delta_{n+1}\mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}\|\hat{H}_{\gamma_{n+1},\tilde{\theta}_{n+1}}(\tilde{X}_{n+1})\|\geq\varepsilon\right) \\
&\quad\leq\sum_{n\in\mathbb{N}}\mathbb{P}\left(2C_{\hat{H}}M_\Theta\,\delta_{n+1}\,\gamma_{n+1}^{-1}\,\mathrm{K}_{\gamma_{n+1},\tilde{\theta}_{n+1}}V^{1/4}(\tilde{X}_{n+1})\geq\varepsilon\right) \\
&\quad\leq 4\varepsilon^{-2}M_\Theta^2 C_{\hat{H}}^2 A_1\mathbb{E}\left[V^{1/2}(\tilde{X}_0)\right]\sum_{n\in\mathbb{N}}\delta_n^2\gamma_n^{-2}<+\infty\ ,
\end{aligned}$$

Using the Borel-Cantelli lemma, we get $\lim_{n\to+\infty}\langle \delta_n a_n \mathrm{K}_{\gamma_n,\tilde{\theta}_n}\hat{H}_{\gamma_n,\tilde{\theta}_n}(\tilde{X}_n)\rangle = 0$ a.s.. This completes the proof of convergence for any $\theta\in\Theta$ of the series $\sum_{k\in\mathbb{N}}\delta_{k+1}\langle a_{k+1},\tilde{\eta}_k^{(b)}\rangle$.

□

Lemma 4.2.15. *Assume* **A**1, **A**2, **A**3, **A**4, **H**1, **H**2 *and that for any* $\theta\in\Theta$ *and* $x\in\mathbb{R}^d$, $\|H_\theta(x)\|\leq V^{1/4}(x)$. *Then we have for any* $n\in\mathbb{N}$

$$\mathbb{E}\left[\left\|\tilde{\eta}_n^{(c)}\right\|\right]\leq C_2\gamma_{n+1}^{-1}\left[\gamma_{n+1}^{-1}\left\{\mathbf{\Lambda}_1(\gamma_n,\gamma_{n+1})+\mathbf{\Lambda}_2(\gamma_n,\gamma_{n+1})\delta_{n+1}\right\}+\delta_{n+1}\right]\ ,$$

with

$$C_2 = 4A_1A_2\log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\,\max\left[L_H, C_{c,1}+2A_2\log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\right]\ , \tag{4.163}$$

where $C_{c,1}$ *is given by*

$$C_{c,1} = 4A_1A_2\log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\mathbb{E}\left[V(\tilde{X}_0)\right]\ . \tag{4.164}$$

Proof. We start by giving an upper-bound on $\|\pi_{\gamma_1,\theta_1}-\pi_{\gamma_2,\theta_2}\|_{V^{1/2}}$ for $\gamma_1,\gamma_2\in(0,\bar{\gamma}]$ with $\gamma_1>\gamma_2$ and, $\theta_1,\theta_2\in\Theta$. Let $g:\mathbb{R}^d\to\mathbb{R}^{d_\theta}$ be a measurable function satisfying

$$\sup_{x\in\mathbb{R}^d}\left\{\|g(x)\|/V^{1/2}(x)\right\}\leq 1\ .$$

Using **H**1-(i)-(ii), **H**2, Lemma 4.2.8 and Lemma 4.2.9, we get that for any $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $\theta_1, \theta_2 \in \Theta$ and $\ell \in \mathbb{N}^\star$

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathrm{K}^\ell_{\gamma_1,\theta_1} g(\tilde{X}_0) - \mathrm{K}^\ell_{\gamma_2,\theta_2} g(\tilde{X}_0)\right\|\right] \\
&= \left\|\sum_{j=0}^{\ell-1} \mathrm{K}^j_{\gamma_1,\theta_1}(\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2})\left\{\mathrm{K}^{(\ell-1-j)}_{\gamma_2,\theta_2} g(x) - \pi_{\gamma_2,\theta_2}(f)\right\}\right\| \\
&\le 2A_2 \sum_{j=0}^{\ell-1} \rho^{(\ell-1-j)\gamma_2/2}\left\|\mathrm{K}^j_{\gamma_1,\theta_1}(\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) V^{1/2}(x)\right\| \\
&\le 2A_2 \sum_{j=0}^{\ell-1} \rho^{(\ell-1-j)\gamma_2/2}\left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\|\right] \sup_{k\in\mathbb{N}} \mathbb{E}\left[\mathrm{K}^k_{\gamma_1,\theta_1} V(\tilde{X}_0)\right] \\
&\le 4A_1A_2 \log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\gamma_2^{-1}\left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\|\right] \mathbb{E}\left[V(\tilde{X}_0)\right] .
\end{aligned}
$$

Taking $\ell \to +\infty$ and using **H**1-(ii), we obtain that for any $\theta_1, \theta_2 \in \Theta$ and $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$,

$$
\|\pi_{\gamma_1,\theta_1} - \pi_{\gamma_2,\theta_2}\|_{V^{1/2}} \le C_{c,1}\gamma_2^{-1}\left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\|\right] , \tag{4.165}
$$

with $C_{c,1}$ given by(4.164).

In what follows we derive an upper bound $\left\|\mathrm{K}_{\gamma_1,\theta_1}\hat{H}_{\gamma_1,\theta_1}(x) - \mathrm{K}_{\gamma_2,\theta_2}\hat{H}_{\gamma_2,\theta_2}(x)\right\|$ for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$ and $x \in \mathbb{R}^d$. By (4.157) we have for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$ and $x \in \mathbb{R}^d$,

$$
\begin{aligned}
&\left\|\mathrm{K}_{\gamma_1,\theta_1}\hat{H}_{\gamma_1,\theta_1}(x) - \mathrm{K}_{\gamma_2,\theta_2}\hat{H}_{\gamma_2,\theta_2}(x)\right\| \\
&= \left\|\sum_{\ell\in\mathbb{N}^\star}\left\{\mathrm{K}^\ell_{\gamma_1,\theta_1} H_{\theta_1}(x) - \pi_{\gamma_1,\theta_1}(H_{\theta_1})\right\} - \sum_{\ell\in\mathbb{N}^\star}\left\{\mathrm{K}^\ell_{\gamma_2,\theta_2} H_{\theta_2}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_2})\right\}\right\| \\
&\le \sum_{\ell\in\mathbb{N}^\star}\left\|\left\{\mathrm{K}^\ell_{\gamma_1,\theta_1} H_{\theta_1}(x) - \pi_{\gamma_1,\theta_1}(H_{\theta_1})\right\} - \left\{\mathrm{K}^\ell_{\gamma_2,\theta_2} H_{\theta_2}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_2})\right\}\right\| .
\end{aligned}
$$

We now bound each term of the series in the right hand side. For any measurable functions g_1, g_2 with $g_i : \mathbb{R}^d \to \mathbb{R}^{d_\theta}$ and such that $\sup_{x\in\mathbb{R}^d} \|g_i(x)\| / V^{1/4}(x) < +\infty$ with $i \in \{1, 2\}$, $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^\star$, it holds that

$$
\begin{aligned}
&\mathrm{K}^\ell_{\gamma_1,\theta_1} g_1(x) - \mathrm{K}^\ell_{\gamma_2,\theta_2} g_2(x) = \mathrm{K}^\ell_{\gamma_1,\theta_1} g_1(x) - \mathrm{K}^\ell_{\gamma_2,\theta_2} g_1(x) + \mathrm{K}^\ell_{\gamma_2,\theta_2}(g_1(x) - g_2(x)) \\
&= \sum_{j=0}^{\ell-1}\left\{\mathrm{K}^j_{\gamma_1,\theta_1} - \pi_{\gamma_1,\theta_1}\right\}(\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2})\left\{\mathrm{K}^{\ell-1-j}_{\gamma_2,\theta_2} g_1(x) - \pi_{\gamma_2,\theta_2}(g_1)\right\} \\
&\quad + \sum_{j=0}^{\ell-1}\pi_{\gamma_1,\theta_1}\left\{\mathrm{K}^{\ell-1-j}_{\gamma_2,\theta_2} g_1(x) - \mathrm{K}^{\ell-j}_{\gamma_2,\theta_2} g_1(x)\right\} + \mathrm{K}^\ell_{\gamma_2,\theta_2}(g_1(x) - g_2(x)) \\
&= \sum_{j=0}^{\ell-1}\left\{\mathrm{K}^j_{\gamma_1,\theta_1} - \pi_{\gamma_1,\theta_1}\right\}(\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2})\left\{\mathrm{K}^{\ell-1-j}_{\gamma_2,\theta_2} g_1(x) - \pi_{\gamma_2,\theta_2}(g_1)\right\} \\
&\quad - \pi_{\gamma_1,\theta_1}(\mathrm{K}^\ell_{\gamma_2,\theta_2} g_1(x) - g_1(x)) + \mathrm{K}^\ell_{\gamma_2,\theta_2}(g_1(x) - g_2(x)) .
\end{aligned}
$$

Setting $g_1 = H_{\theta_1} - \pi_{\gamma_1,\theta_1}(H_{\theta_1})$ and $g_2 = H_{\theta_2} - \pi_{\gamma_2,\theta_2}(H_{\theta_2})$, we obtain that

$$\begin{aligned}
&\mathrm{K}^{\ell}_{\gamma_1,\theta_1} g_1(x) - \mathrm{K}^{\ell}_{\gamma_2,\theta_2} g_2(x) \\
&= \sum_{j=0}^{\ell-1} \left\{ \mathrm{K}^{j}_{\gamma_1,\theta_1} - \pi_{\gamma_1,\theta_1} \right\} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) \left\{ \mathrm{K}^{\ell-1-j}_{\gamma_2,\theta_2} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1}) \right\} \\
&\quad + \Xi_\ell \ ,
\end{aligned} \tag{4.166}$$

where

$$\begin{aligned}
\Xi_\ell &= -\pi_{\gamma_1,\theta_1}(\mathrm{K}^{\ell}_{\gamma_2,\theta_2} H_{\theta_1}(x) - H_{\theta_1}(x)) \\
&\quad + \mathrm{K}^{\ell}_{\gamma_2,\theta_2}\left[H_{\theta_1}(x) - H_{\theta_2}(x) + \pi_{\gamma_2,\theta_2}(H_{\theta_2}) - \pi_{\gamma_1,\theta_1}(H_{\theta_1}) \right] \\
&= -\pi_{\gamma_1,\theta_1} \mathrm{K}^{\ell}_{\gamma_2,\theta_2} H_{\theta_1}(x) + \mathrm{K}^{\ell}_{\gamma_2,\theta_2}\left[H_{\theta_1}(x) - H_{\theta_2}(x) + \pi_{\gamma_2,\theta_2}(H_{\theta_2}) \right] \\
&= (\pi_{\gamma_2,\theta_2} - \pi_{\gamma_1,\theta_1})(\mathrm{K}^{\ell}_{\gamma_2,\theta_2} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1})) - \pi_{\gamma_2,\theta_2}(H_{\theta_1}) \\
&\qquad\qquad + \mathrm{K}^{\ell}_{\gamma_2,\theta_2}\left[H_{\theta_1}(x) - H_{\theta_2}(x) + \pi_{\gamma_2,\theta_2}(H_{\theta_2}) \right] \\
&= (\pi_{\gamma_2,\theta_2} - \pi_{\gamma_1,\theta_1})(\mathrm{K}^{\ell}_{\gamma_2,\theta_2} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1})) \\
&\qquad\qquad + \mathrm{K}^{\ell}_{\gamma_2,\theta_2}(H_{\theta_1} - H_{\theta_2})(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1} - H_{\theta_2}) \ .
\end{aligned} \tag{4.167}$$

For the first term in (4.166), using **H**1-(ii), **H**2, Lemma 4.2.9 and and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x)$ we obtain for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^\star$

$$\begin{aligned}
&\left\| \sum_{j=0}^{\ell-1} \left\{ \mathrm{K}^{j}_{\gamma_1,\theta_1} - \pi_{\gamma_1,\theta_1} \right\} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) \left\{ \mathrm{K}^{\ell-1-j}_{\gamma_2,\theta_2} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1}) \right\} \right\| \\
&\quad \le 2A_2 \sum_{j=0}^{\ell-1} \rho^{(\ell-1-j)\gamma_1/2} \left\| \left\{ \mathrm{K}^{j}_{\gamma_1,\theta_1} - \pi_{\gamma_1,\theta_1} \right\} (\mathrm{K}_{\gamma_1,\theta_1} - \mathrm{K}_{\gamma_2,\theta_2}) V^{1/2}(x) \right\| \\
&\quad \le 4A_2^2 \left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\| \right] \sum_{j=0}^{\ell-1} \rho^{(j+(\ell-1-j))\gamma_2/2} V^{1/2}(x) \\
&\quad \le 4A_2^2 \left[\mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\| \right] \ell \rho^{(\ell-1)\gamma_2/2} V^{1/2}(x) \ .
\end{aligned} \tag{4.168}$$

For the first term in (4.167), using **H**1-(ii), Lemma 4.2.9, (4.165) and that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \le V^{1/4}(x) \le V^{1/2}(x)$, we obtain for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^\star$

$$\begin{aligned}
&\left\| (\pi_{\gamma_1,\theta_1} - \pi_{\gamma_2,\theta_2})(\mathrm{K}^{\ell}_{\gamma_2,\theta_2} H_{\theta_1}(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1})) \right\| \\
&\qquad \le 2A_2 \rho^{\ell\gamma_2/2} \|\pi_{\gamma_1,\theta_1} - \pi_{\gamma_2,\theta_2}\|_{V^{1/2}} \\
&\qquad \le 2A_2 C_{c,1} \, \rho^{\ell\gamma_2/2} \gamma_2^{-1} \left\{ \mathbf{\Lambda}_1(\gamma_1,\gamma_2) + \mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1 - \theta_2\| \right\} \ .
\end{aligned} \tag{4.169}$$

For the second term in (4.167), using **A**4, **H**1-(ii) and Lemma 4.2.9, we obtain for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^\star$

$$\left\| \mathrm{K}^{\ell}_{\gamma_2,\theta_2}(H_{\theta_1} - H_{\theta_2})(x) - \pi_{\gamma_2,\theta_2}(H_{\theta_1} - H_{\theta_2}) \right\| \le 2A_2 L_H \rho^{\ell\gamma_2/2} \|\theta_1 - \theta_2\| V^{1/2}(x) \ . \tag{4.170}$$

Combining (4.167), (4.168), (4.169), (4.170) in (4.166) and using Lemma 4.2.8, we obtain that for any $\theta_1, \theta_2 \in \Theta$, $\gamma_1, \gamma_2 \in (0, \bar{\gamma}]$ with $\gamma_1 > \gamma_2$, $x \in \mathbb{R}^d$ that

$$\left\| \mathrm{K}_{\gamma_1,\theta_1} \hat{H}_{\gamma_1,\theta_1}(x) - \mathrm{K}_{\gamma_2,\theta_2} \hat{H}_{\gamma_2,\theta_2}(x) \right\|$$

$$\leq C_{c,2}\,\gamma_2^{-1}\left[\gamma_2^{-1}\left\{\mathbf{\Lambda}_1(\gamma_1,\gamma_2)+\mathbf{\Lambda}_2(\gamma_1,\gamma_2)\|\theta_1-\theta_2\|\right\}+\|\theta_1-\theta_2\|\right]V^{1/2}(x)\;,$$

with

$$C_{c,2}\;=4A_2\log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\;\max\left[L_H, C_{c,1}+2A_2\log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\right]\;.$$

Since for any $k\in\mathbb{N}$, $\left\|\tilde{\theta}_{k+1}-\tilde{\theta}_k\right\|\leq\delta_{k+1}V^{1/2}(\tilde{X}_{k+1})$ by (4.131) and the fact that for any $\theta\in\Theta$ and $x\in\mathbb{R}^d$, $\|H_\theta(x)\|\leq V^{1/2}(x)$ and that Π_Θ is non-expansive, we get that for any $k\in\mathbb{N}$,

$$\begin{aligned}&\left\|\mathrm{K}_{\gamma_k,\tilde{\theta}_k}\hat{H}_{\gamma_k,\tilde{\theta}_k}(\tilde{X}_{k+1})-\mathrm{K}_{\gamma_{k+1},\tilde{\theta}_{k+1}}\hat{H}_{\gamma_{k+1},\tilde{\theta}_{k+1}}(\tilde{X}_{k+1})\right\|\\&\quad\leq C_{c,2}\gamma_{k+1}^{-1}\left[\gamma_{k+1}^{-1}\left\{\mathbf{\Lambda}_1(\gamma_k,\gamma_{k+1})+\mathbf{\Lambda}_2(\gamma_k,\gamma_{k+1})\delta_{k+1}\right\}+\delta_{k+1}\right]V(\tilde{X}_{k+1})\;,\end{aligned}$$

which implies by (4.160) and using **H**1-(i) that

$$\mathbb{E}\left[\left\|\tilde{\eta}^{(c)}\right\|\right]\leq C_2\gamma_{k+1}^{-1}\left[\gamma_{k+1}^{-1}\left\{\mathbf{\Lambda}_1(\gamma_k,\gamma_{k+1})+\mathbf{\Lambda}_2(\gamma_k,\gamma_{k+1})\delta_{k+1}\right\}+\delta_{k+1}\right]\;,$$

with C_2 given by (4.163).

□

Lemma 4.2.16. *Assume* **A**1, **A**2, **A**3, **H**1 *and that for any* $\theta\in\Theta$ *and* $x\in\mathbb{R}^d$, $\|H_\theta(x)\|\leq V^{1/4}(x)$. *Then we have for any* $n\in\mathbb{N}$

$$\mathbb{E}\left[\left\|\tilde{\eta}_n^{(d)}\right\|\right]\leq\mathbf{\Psi}(\gamma_n)\;.$$

Proof. By a straightforward application of **H**1-(iii) and by (4.160) along with the fact that for any $\theta\in\Theta$ and $x\in\mathbb{R}^d$, $\|H_\theta(x)\|\leq V^{1/4}(x)$ we have for any $n\in\mathbb{N}$, $\mathbb{E}\left[\left\|\tilde{\eta}_n^{d}\right\|\right]\leq\mathbf{\Psi}(\gamma_n)$. □

We now turn to the proof of Theorem 4.2.3.

Proof. (a) To apply [AFM17, Theorem 2], it is enough to show that the following series converge a.s.

$$\sum_{n=0}^{+\infty}\delta_{n+1}\langle\Pi_\Theta(\theta_n-\delta_{n+1}\nabla f(\theta_n))-\theta^\star,\tilde{\eta}_n^{(i)}\rangle\;,\quad\sum_{n=0}^{+\infty}\delta_{n+1}^2\|\tilde{\eta}_n\|^2\;,$$

with $\theta^\star\in\arg\min_{\theta\in\Theta}f(\theta)$. $\sum_{n=0}^{+\infty}\delta_{n+1}^2\|\tilde{\eta}_n\|^2<+\infty$ a.s. by Lemma 4.2.12 since $\sum_{n\in\mathbb{N}}\delta_{n+1}^2<+\infty$. Since $(\langle\Pi_\Theta(\theta_n-\delta_{n+1}\nabla f(\theta_n))-\theta^\star,\tilde{\eta}_n^{(a)}\rangle)_{n\in\mathbb{N}}$ is a $(\tilde{\mathcal{F}}_n)_{n\in\mathbb{N}}$-martingale increment, see (4.132) and by Lemma 4.2.13 and $\sum_{n\in\mathbb{N}}\delta_{n+1}^2/\gamma_n^2<+\infty$

$$\mathbb{E}\left[\sum_{n=0}^{+\infty}\delta_{n+1}^2\langle\Pi_\Theta(\theta_n-\delta_{n+1}\nabla f(\theta_n))-\theta^\star,\tilde{\eta}_n^{(a)}\rangle^2\right]<+\infty\;,$$

we obtain using [Wil91, Section 12.5] that $\sum_{n=0}^{+\infty}\delta_{n+1}\langle\Pi_\Theta(\theta_n-\delta_{n+1}\nabla f(\theta_n))-\theta^\star,\tilde{\eta}_n^{(a)}\rangle$ converges a.s.. Using **A**1, (4.133) and Lemma 4.2.15 and Lemma 4.2.16 we get that $\sum_{n=0}^{+\infty}\delta_{n+1}\langle\Pi_\Theta(\theta_n-\delta_{n+1}\nabla f(\theta_n))-\theta^\star,\tilde{\eta}_n^{(i)}\rangle$ is absolutely convergent a.s. for $i\in\{c,d\}$. Finally we obtain that $\sum_{n=0}^{+\infty}\delta_{n+1}\langle\Pi_\Theta(\theta_n-\delta_{n+1}\nabla f(\theta_n))-\theta^\star,\tilde{\eta}_n^{(b)}\rangle$ converges a.s. by Lemma 4.2.14-(b).

(b) The proof of is identical to the one of Theorem 4.2.1-(b).

□

Proof of Theorem 4.2.4

The proof is similar to the one of Theorem 4.2.2, using Lemma 4.2.12, Lemma 4.2.14, Lemma 4.2.15, Lemma 4.2.16 and the fact that $\tilde{\eta}_n^{(a)}$ is a $(\tilde{\mathcal{F}}_n)_{n\in\mathbb{N}}$-martingale increment, see (4.132).

Proof of Theorem 4.2.5

In this section, we give the proof of Theorem 4.2.5 by showing that **H**1 holds. First of all in Section 4.2.5, we establish under **L**1 and **L**2 stability results uniform in the parameter $\theta \in \Theta$ for the Langevin diffusion (4.122) and the associated Euler-Maruyama discretization (4.123) based on a Foster-Lyapunov drift condition with constants independent of θ. Then, in Section 4.2.5, we show that the stability conditions that we derive, are sufficient to prove that **H**1 holds. The proof of Theorem 4.2.5 then consists in combining all these results and is presented in Section 4.2.5.

Under **L**1 and **L**2, for any $\theta \in \Theta$, (4.122) defines a Markov semi-group $(\mathrm{P}_{t,\theta})_{t\geq 0}$ for any $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by $\mathrm{P}_{t,\theta}(x,\mathsf{A}) = \mathbb{P}(Y_t^\theta \in \mathsf{A})$ where $(Y_t^\theta)_{t\geq 0}$ is the solution of (4.122) with $Y_0^\theta = x$. Consider now the generator of $(\mathrm{P}_{t,\theta})_{t\geq 0}$ for any $\theta \in \Theta$, defined for any $f \in \mathrm{C}^2(\mathbb{R}^d)$ by

$$\mathcal{A}_\theta f = -\langle \nabla_x f, \nabla_x U_\theta(x)\rangle + \Delta_x f \ . \tag{4.171}$$

We say that a Markov kernel R on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ satisfies a discrete Foster-Lyapunov drift condition $\mathbf{D}_{\mathrm{d}}(V,\lambda,b)$ if there exist $\lambda \in (0,1)$, $b \geq 0$ and a measurable function $V : \mathbb{R}^d \to [1,+\infty)$ such that for all $x \in \mathbb{R}^d$

$$\mathrm{R}V(x) \leq \lambda V(x) + b \ .$$

We say that a Markov semi-group $(\mathrm{P}_t)_{t\geq 0}$ on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ with extended infinitesimal generator $(\mathcal{A}, \mathrm{D}(\mathcal{A}))$ (see *e.g.* [MT93c] for the definition of $(\mathcal{A}, \mathrm{D}(\mathcal{A}))$) satisfies a continuous drift condition $\mathbf{D}_{\mathrm{c}}(V,\zeta,\beta)$ if there exist $\zeta > 0$, $\beta \geq 0$ and a measurable function $V : \mathbb{R}^d \to [1,+\infty)$ with $V \in \mathrm{D}(\mathcal{A})$ such that for all $x \in \mathbb{R}^d$

$$\mathcal{A}V(x) \leq -\zeta V(x) + \beta \ .$$

Foster-Lyapunov drift conditions uniform on θ Define $V_{\mathrm{e}} : \mathbb{R}^d \to [1,+\infty)$ for all $x \in \mathbb{R}^d$ by

$$V_{\mathrm{e}}(x) = \exp(\tilde{\mathtt{m}}_1 \phi(x)) \ , \quad \text{with } \phi(x) = \sqrt{1+\|x\|^2} \text{ and } \tilde{\mathtt{m}}_1 = \mathtt{k}_2/4 \ . \tag{4.172}$$

Proposition 4.2.17. *Assume* **L**1 *and* **L**2. *Let* $\bar{\gamma} < \min(1, 2\mathtt{m}_3^+)$. *Then there exist* $\lambda_{\mathrm{e}} \in (0,1)$ *and* $b_{\mathrm{e}} \geq 0$ *such that for all* $\gamma \in (0,\bar{\gamma}]$ *and* $\theta \in \Theta$ *the Markov kernel* $\mathrm{R}_{\gamma,\theta}$ *associated with the recursion* (4.123) *satisfies the discrete drift condition* $\mathbf{D}_{\mathrm{d}}(V,\lambda^\gamma, b\gamma)$, *i.e. for all* $x \in \mathbb{R}^d$

$$\mathrm{R}_{\gamma,\theta} V_{\mathrm{e}}(x) \leq \lambda_{\mathrm{e}}^\gamma V_{\mathrm{e}}(x) + b_{\mathrm{e}} \gamma \mathbb{1}_{\mathrm{B}(0,r_{\mathrm{e}})}(x) \ , \tag{4.173}$$

with

$$\lambda_{\mathrm{e}} = \mathrm{e}^{-\tilde{\mathtt{m}}_1^2(2^{1/2}-1)} \ , \quad r_{\mathrm{e}} = \max(1, 2(d+\mathtt{c})/\mathtt{k}_2, R_2) \ ,$$
$$b_{\mathrm{e}} = \tilde{\mathtt{m}}_1(d+c+2^{1/2}\tilde{\mathtt{m}}_1)\exp\left[\tilde{\mathtt{m}}_1\left\{(d+c+\tilde{\mathtt{m}}_1)\bar{\gamma} + \sqrt{1+r_{\mathrm{e}}^2}\right\}\right] \ .$$

Proof. Since ϕ is 1-Lipschitz, by the log-Sobolev inequality [BGL14, Proposition 5.4.1], we have for any $x \in \mathbb{R}^d$ and $\theta \in \Theta$,

$$\mathrm{R}_{\gamma,\theta} V_{\mathrm{e}}(x) \leq \exp\left[\tilde{\mathtt{m}}_1 \mathrm{R}_{\gamma,\theta}\phi(x) + \tilde{\mathtt{m}}_1^2 \gamma\right] \tag{4.174}$$

$$\leq \exp\left[\tilde{\mathtt{m}}_1\sqrt{\|x-\gamma\nabla_x U_\theta(x)\|^2+2\gamma d+1}+\tilde{\mathtt{m}}_1^2\gamma\right] ,$$

where we have used Jensen's inequality in the last line. Second, using **L**2 and $\gamma < 2\mathtt{m}_3^+$, we obtain that for any $x \in \mathbb{R}^d$ and $\theta \in \Theta$,

$$\begin{aligned}\|x-\gamma\nabla_x U_\theta(x)\|^2 &\leq \|x\|^2 - 2\gamma\langle x, \nabla_x U_\theta(x)\rangle + \gamma^2\|\nabla_x U_\theta(x)\|^2 \\ &\leq \|x\|^2 - 2\mathtt{k}_2\gamma\|x\|\mathbb{1}_{\mathrm{B}(0,R_2)^{\mathrm{c}}}(x) + \gamma(\gamma-2\mathtt{m}_3^+)\|\nabla_x U_\theta(x)\|^2 + 2\gamma\mathtt{c} \\ &\leq \|x\|^2 - 2\mathtt{k}_2\gamma\|x\|\mathbb{1}_{\mathrm{B}(0,R_2)^{\mathrm{c}}}(x) + 2\gamma\mathtt{c} .\end{aligned}$$

Therefore, using for any $a > 0$, $\sqrt{1+a} - 1 \leq a/2$, we get for any $x \in \mathbb{R}^d$ and $\theta \in \Theta$,

$$\begin{aligned}&\sqrt{\|x-\gamma\nabla_x U_\theta(x)\|^2+2\gamma d+1} - \phi(x) \\ &\leq \phi(x)\left\{\sqrt{1+2\gamma\phi^{-2}(x)(d+\mathtt{c}-\mathtt{k}_2\|x\|\mathbb{1}_{\mathrm{B}(0,R_2)^{\mathrm{c}}}(x))}-1\right\} \\ &\leq \gamma\phi^{-1}(x)(d+\mathtt{c}-\mathtt{k}_2\|x\|\mathbb{1}_{\mathrm{B}(0,R_2)^{\mathrm{c}}}(x)) .\end{aligned} \tag{4.175}$$

Therefore, combining this result with (4.174) and using that for any $\tilde{x} \in \bar{\mathrm{B}}(0,r_{\mathrm{e}})^{\mathrm{c}}$, $\phi(\tilde{x})^2/\|\tilde{x}\|^2 \leq 2$ and $d + c \leq \mathtt{k}_2\|x\|/2$, we obtain for any $x \in \bar{\mathrm{B}}(0,r_{\mathrm{e}})^{\mathrm{c}}$ and $\theta \in \Theta$,

$$\begin{aligned}\mathrm{R}_{\gamma,\theta}V_{\mathrm{e}}(x) &\leq \exp\left[\tilde{\mathtt{m}}_1\phi^{-1}(x)(d+\mathtt{c}-\mathtt{k}_2\|x\|)+\tilde{\mathtt{m}}_1^2\gamma\right]V_{\mathrm{e}}(x) \\ &\leq \exp\left[-2\tilde{\mathtt{m}}_1^2\gamma\phi^{-1}(x)\|x\|+\tilde{\mathtt{m}}_1^2\gamma\right]V_{\mathrm{e}}(x) \leq \lambda_{\mathrm{e}}^\gamma V_{\mathrm{e}}(x) .\end{aligned}$$

Using (4.174), (4.175), and the fact that $\phi(\tilde{x}) \geq 1$ for any $\tilde{x} \in \mathbb{R}^d$, we have for any $x \in \mathrm{B}(0,r_{\mathrm{e}})$ and $\theta \in \Theta$,

$$\mathrm{R}_{\gamma,\theta}V_{\mathrm{e}}(x) \leq \lambda_{\mathrm{e}}^\gamma V_{\mathrm{e}}(x) + \left(\mathrm{e}^{\tilde{\mathtt{m}}_1(d+c+\tilde{\mathtt{m}}_1)\gamma} - \lambda_{\mathrm{e}}^\gamma\right)\exp\left[\tilde{\mathtt{m}}_1\sqrt{1+r_{\mathrm{e}}^2}\right] .$$

The proof of (4.173) for $x \in \mathrm{B}(0,r_{\mathrm{e}})$ and $\theta \in \Theta$ is then completed upon using that $\mathrm{e}^a - \mathrm{e}^b \leq (a-b)\mathrm{e}^a$ for all $a, b \in \mathbb{R}$ with $a \geq b$. □

Proposition 4.2.18. *Assume* **L**1 *and* **L**2. *Then for any* $\theta \in \Theta$, $(\mathrm{P}_{t,\theta})_{t\geq 0}$ *associated with* (4.122) *satisfies the continuous drift condition* $\mathbf{D}_{\mathrm{c}}(V_{\mathrm{e}}, \zeta_{\mathrm{e}}, \beta_{\mathrm{e}})$ *for* V_{e} *defined in* (4.172) *and*

$$\zeta_{\mathrm{e}} = 3\tilde{\mathtt{m}}_1^2/2^{1/2} , \quad \beta_{\mathrm{e}} = \tilde{\mathtt{m}}_1\exp\left[\tilde{\mathtt{m}}_1\sqrt{1+\tilde{r}_{\mathrm{e}}^2}\right](1+\tilde{\mathtt{m}}_1+\mathtt{c}+d) , \quad \tilde{r}_{\mathrm{e}} = \max(1,R_2) .$$

Proof. First, by definition, for any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}\nabla_x V(x) &= \tilde{\mathtt{m}}_1 x V(x)/\phi(x) \\ \Delta_x V(x) &= \{\tilde{\mathtt{m}}_1 V(x)/\phi(x)\}\{\tilde{\mathtt{m}}_1\|x\|^2/\phi(x) + d - \|x\|^2/\phi^2(x)\} .\end{aligned}$$

Therefore, by (4.171) and **L**2, we get for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$,

$$\begin{aligned}\mathcal{A}_\theta V(x) &= \{\tilde{\mathtt{m}}_1 V(x)/\phi(x)\}\left[-\langle\nabla_x U_\theta(x), x\rangle + \tilde{\mathtt{m}}_1\|x\|^2/\phi(x) + d - \|x\|^2/\phi^2(x)\right] \\ &\leq \{\tilde{\mathtt{m}}_1 V(x)/\phi(x)\}\left[-\mathtt{k}_2\|x\|\mathbb{1}_{\mathrm{B}(0,R_2)^{\mathrm{c}}}(x) + \mathtt{c} + \tilde{\mathtt{m}}_1\|x\|^2/\phi(x) + d - \|x\|^2/\phi^2(x)\right] \\ &\leq \{\tilde{\mathtt{m}}_1 V(x)/\phi(x)\}\left[-(3\mathtt{k}_2/4)\|x\|\mathbb{1}_{\mathrm{B}(0,R_2)^{\mathrm{c}}}(x) + \mathtt{c} + \tilde{\mathtt{m}}_1\|x\|\mathbb{1}_{\mathrm{B}(0,R_2)}(x) + d\right] .\end{aligned}$$

The proof is then complete upon using that for any $x \in \mathrm{B}(0,\tilde{r}_{\mathrm{e}})^{\mathrm{c}}$, $\|x\|/\phi(x) \geq 2^{-1/2}$, for any $y \in \mathbb{R}^d$, $\|y\|/\phi(y) \leq 1$.

□

Checking H1

Lemma 4.2.19. *Assume* **L1** *and let* $V : \mathbb{R}^d \to [1, +\infty)$ *measurable and* $\lim_{\|x\|\to+\infty} V(x) = +\infty$ *and* $V \in \mathrm{D}(\mathcal{A}_\theta)$*, for any* $\theta \in \Theta$*, where* $\mathcal{A}_\theta$ *is defined by* (4.171)*. .*

(a) Assume that there exist $\lambda \in (0,1)$*,* $b \geq 0$ *and* $\bar{\gamma} > 0$ *such that for any* $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$*,* $\mathrm{R}_{\gamma,\theta}$ *associated with the recursion* (4.134)*, satisfies* $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$*. Then for any* $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$*,* $\mathrm{R}_{\gamma,\theta}$ *admits an invariant probability measure* $\pi_{\gamma,\theta}$ *on* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *and there exists* $D_3 \geq 0$ *such that for any* $x \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$

$$\delta_x \mathrm{R}^k_{\gamma,\theta} V \leq D_3 + V(x) \;, \qquad \pi_{\gamma,\theta}(V) \leq D_3 \;, \qquad D_3 = b\lambda^{-\bar{\gamma}} / \log(1/\lambda) \;.$$

In addition, for all $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$*,* $\lim_{k\to+\infty} \|\delta_x \mathrm{R}^k_{\gamma,\theta} - \pi_{\gamma,\theta}\|_V = 0$*.*

(b) Assume that there exist $\zeta > 0$ *and* $\beta \geq 0$ *such that for any* $\theta \in \Theta$*,* $(\mathrm{P}_{t,\theta})_{t\geq 0}$ *associated with* (4.122) *satisfies* $\mathbf{D}_{\mathrm{c}}(V, \zeta, \beta)$*. Then for any* $\theta \in \Theta$*, the diffusion is non-explosive,* $\mathcal{A}_\theta$ *admits* π_θ *as an invariant probability measure and*

$$\pi_\theta(V) \leq D_0 \;, \qquad D_0 = \beta/\zeta \;.$$

In addition, for all $\theta \in \Theta$ *and* $x \in \mathbb{R}^d$*,* $\lim_{t\to+\infty} \|\delta_x \mathrm{P}_{t,\theta} - \pi_\theta\|_V = 0$*.*

Proof. (a) for any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$, $\mathrm{R}_{\gamma,\theta}$ is irreducible with respect to the Lebesgue measure on $\mathbb{R}^d$, has the Feller property and satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$ then [MT92, Section 4.4] applies and $\mathrm{R}_{\gamma,\theta}$ admits an invariant probability measure $\pi_{\gamma,\theta}$. The discrete drift condition and [DM17, Lemma 1] give that for any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$

$$\mathrm{R}^k_{\gamma,\theta} V(x) \leq V(x) + b\lambda^{-\bar{\gamma}} / \log(1/\lambda) \;, \qquad \pi_{\gamma,\theta}(V) \leq b\lambda^{-\bar{\gamma}} / \log(1/\lambda) \;.$$

We obtain that for all $\theta \in \Theta$ and $x \in \mathbb{R}^d$, $\lim_{k\to+\infty} \|\delta_x \mathrm{R}^k_{\gamma,\theta} - \pi_{\gamma,\theta}\|_V = 0$ using [MT93a, Theorem 16.0.1].

(b) Using $\mathbf{D}_{\mathrm{c}}(V, \zeta, \beta)$ and [MT93c, Theorem 2.1] we get that the diffusion process is non-explosive and thus $(\mathrm{P}_{t,\theta})_{t\geq 0}$ is defined for any $\theta \in \Theta$ and $t \geq 0$. Using [SV06, Corollary 10.1.4] for any $\theta \in \Theta$, $(\mathrm{P}_{t,\theta})_{t\geq 0}$ is strongly Feller continuous, therefore any compact sets is petite for the Markov kernel $\mathrm{P}_{h,\theta}$, for any $h > 0$ and $\theta \in \Theta$, by [MT93a, Theorem 6.0.1]. Using [RY99, Chapter 7, Proposition 1.5], [EK86, Chapter 4, Theorem 9.17], and the fact that $\pi_\theta(\mathcal{A}_\theta f) = 0$ for any $\theta \in \Theta$ and $f \in \mathrm{C}^2_c(\mathbb{R}^d)$, we obtain that for any $\theta \in \Theta$, π_θ is an invariant measure for $(\mathrm{P}_{t,\theta})_{t\geq 0}$. Using $\mathbf{D}_{\mathrm{c}}(V, \zeta, \beta)$ and [MT93c, Theorem 4.5] we get that for all $\theta \in \Theta$, $\pi_\theta(V) \leq \beta/\zeta$. Finally, the convergence is ensured using [MT93c, Theorem 5.1].

□

As an immediate corollary we obtain that under the conditions of Lemma 4.2.19 for any $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $k \in \mathbb{N}$,

$$\pi_\theta \mathrm{R}^k_{\gamma,\theta} V \leq \beta/\zeta + b\lambda^{-\bar{\gamma}} / \log(1/\lambda) \;. \tag{4.176}$$

Lemma 4.2.20. *Let* $V : \mathbb{R}^d \to [1, +\infty)$*. Assume there exist* $\lambda \in (0,1)$*,* $b \geq 0$ *and* $\bar{\gamma} > 0$ *such that for any* $\theta \in \Theta$ *and* $\gamma \in (0, \bar{\gamma}]$*,* $\mathrm{R}_{\gamma,\theta}$ *associated with the recursion* (4.123) *satisfies* $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$*. Let* $(\gamma_n)_{n\in\mathbb{N}}$*,* $(\delta_n)_{n\in\mathbb{N}}$ *be sequences of non-increasing positive real numbers and* $(m_n)_{n\in\mathbb{N}}$ *be a sequence of positive integers satisfying* $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$*. Then,* $(X^n_k)_{n\in\mathbb{N}, k\in\{0,\ldots,m_n\}}$ *given by* (4.126) *with* $\{\mathrm{K}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\} = \{\mathrm{R}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ *satisfies for all* $p, n \in \mathbb{N}$ *and* $k \in \{0, \ldots, m_n\}$

$$\mathbb{E}\left[\mathrm{R}^p_{\gamma_n,\theta_n} V(X^n_k) \middle| X^0_0\right] \leq D_1 V(X^0_0) \;, \qquad D_1 = 1 + 2b\lambda^{-\bar{\gamma}} / \log(1/\lambda) \;.$$

Proof. By induction we obtain that

$$\mathbb{E}\left[V(X_k^{n+1})\middle|\mathcal{F}_n\right] = \mathrm{R}^k_{\gamma_{n+1},\theta_{n+1}} V(X_0^{n+1}) \tag{4.177}$$
$$\leq \lambda^{k\gamma_{n+1}} V(X_0^{n+1}) + b\gamma_{n+1} \sum_{i=1}^{k} \lambda^{\gamma_{n+1}(k-i)} ,$$

where $(\mathcal{F}_n)_{n\in\mathbb{N}}$ is defined by (4.124). Similarly, we obtain for any $k \in \{0,\dots,m_0\}$,

$$\mathbb{E}\left[V(X_k^0)\middle|X_0^0\right] = \mathrm{R}^k_{\gamma_0,\theta_0} V(X_0^0) \leq \lambda^{k\gamma_0} V(X_0^0) + b\gamma_0 \sum_{i=1}^{k} \lambda^{\gamma_0(k-i)} . \tag{4.178}$$

Define for $k,\ell \in \mathbb{N}$ and $i \in \mathbb{N}^\star$, $\mathrm{q}_{\ell,k} = \sum_{j=0}^{\ell-1} m_j + k$, $\mathrm{q}_n = \mathrm{q}_{\ell,0}$ and $\tilde\gamma_i = \sum_{j=0}^{+\infty} \gamma_j \mathbb{1}_{(\mathrm{q}_j,\mathrm{q}_{j+1}]}(i)$. In addition, consider for any $\mathrm{p},\mathrm{q} \in \mathbb{N}^\star$, $\Gamma_{\mathrm{p},\mathrm{q}} = \sum_{i=\mathrm{p}}^{\mathrm{q}} \tilde\gamma_i$ and $\Gamma_{\mathrm{p}} = \Gamma_{1,\mathrm{p}}$. Combining (4.177), (4.178) and Lemma 4.2.8 we get for any $n \in \mathbb{N}$ and $k \in \{0,\dots,m_n\}$

$$\mathbb{E}\left[\mathrm{R}^p_{\gamma_n,\theta_n} V(X_k^n)\middle|X_0^0\right] \leq \lambda^{\gamma_n p}\mathbb{E}\left[V(X_k^n)\middle|X_0^0\right] + b\log(1/\lambda)\lambda^{-\bar\gamma} \tag{4.179}$$
$$\leq \lambda^{\Gamma_{\mathrm{q}_{n,k}}} V(X_0^0) + b\sum_{i=1}^{\mathrm{q}_{n,k}} \tilde\gamma_i \lambda^{\Gamma_{i+1,\mathrm{q}_{n,k}}} + b\log(1/\lambda)\lambda^{-\bar\gamma} .$$

Since $(\tilde\gamma_i)_{i\in\mathbb{N}}$ is nonincreasing and for all $t \geq 0$, $1-\lambda^t \geq -t\lambda^t\log(\lambda)$, we have for all $\mathrm{q} \in \mathbb{N}^\star$,

$$\begin{aligned}
&\sum_{i=1}^{\mathrm{q}} \tilde\gamma_i \lambda^{\Gamma_{i+1,\mathrm{q}}} \leq \sum_{i=1}^{\mathrm{q}} \tilde\gamma_i \prod_{j=i+1}^{\mathrm{q}} (1+\lambda^{\tilde\gamma_1}\log(\lambda)\tilde\gamma_j) \\
&\leq (-\lambda^{\tilde\gamma_1}\log(\lambda))^{-1} \sum_{i=1}^{\mathrm{q}} \left\{ \prod_{j=i+1}^{\mathrm{q}} (1+\lambda^{\tilde\gamma_1}\log(\lambda)\tilde\gamma_j) - \prod_{j=i}^{\mathrm{q}} (1+\lambda^{\tilde\gamma_1}\log(\lambda)\tilde\gamma_j) \right\} \\
&\leq (-\lambda^{\tilde\gamma_1}\log(\lambda))^{-1} .
\end{aligned}$$

Combining this result and (4.179) completes the proof. □

Lemma 4.2.21. *Let $V : \mathbb{R}^d \to [1,+\infty)$ with $\sup_{x\in\mathbb{R}^d}\left\{(1+\|x\|)^2/V(x)\right\} \leq M_V$ where $M_V \geq 0$ and V measurable. Assume* **L**1 *and that for any $\theta \in \Theta$, $\gamma \in (0,\bar\gamma]$ and $k \in \mathbb{N}$,*

$$\pi_\theta \mathrm{R}^k_{\gamma,\theta}(V) \leq \tilde D_1 , \qquad \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta} V \leq \tilde D_1 , \tag{4.180}$$

with $m_\gamma = \lceil 1/\gamma \rceil$. Then for any $\theta \in \Theta$ and $\gamma \in (0,\bar\gamma]$

$$\|\pi_\theta \mathrm{R}^{m_\gamma}_{\gamma,\theta} - \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\|^2_{V^{1/2}}$$
$$\leq 2\tilde D_1 \mathrm{L}^2 \gamma(1+\bar\gamma)\left\{ d + 2\bar\gamma(\sup_{\theta\in\Theta}\|\nabla_x U_\theta(0)\|^2 + \mathrm{L}^2 M_V \tilde D_1)\right\} ,$$

Proof. The proof follows the lines of [DM17, Theorem 10]. Let $\theta \in \Theta$ and $\gamma \in (0,\bar\gamma]$. We have, using a generalized Pinsker inequality [DM17, Lemma 24], that

$$\begin{aligned}
\|\pi_\theta \mathrm{R}^{m_\gamma}_{\gamma,\theta} - \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\|^2_{V^{1/2}} &\leq 2(\pi_\theta \mathrm{R}^{m_\gamma}_{\gamma,\theta} V + \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta} V)\mathrm{KL}\left(\pi_\theta \mathrm{R}^{m_\gamma}_{\gamma,\theta} | \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\right) . \\
&\leq 4\tilde D_1 \mathrm{KL}\left(\pi_\theta \mathrm{R}^{m_\gamma}_{\gamma,\theta} | \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\right) .
\end{aligned}$$

Using **L**1, [DM17, Equation (15)], [Kul97, Theorem 4.1, Chapter 2], (4.180) and that for any $a, b \in \mathbb{R}$, $(a+b)^2 \leq 2(a^2+b^2)$ we obtain that

$$\begin{aligned}\mathrm{KL}\left(\pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma} | \pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}\right) &\leq \mathtt{L}^2 m_\gamma \gamma^2 (d + \bar{\gamma} \sup_{k\in\mathbb{N}} \pi_\theta \mathrm{R}_{\gamma,\theta}^k \|\nabla_x U_\theta(x)\|^2) \\ &\leq \mathtt{L}^2 (1+\bar{\gamma}) \gamma (d + 2\bar{\gamma} (\sup_{\theta\in\Theta} \|\nabla_x U_\theta(0)\|^2 + \mathtt{L}^2 M_V \tilde{D}_1)) \ ,\end{aligned}$$

which concludes the proof. □

Proposition 4.2.22. *Let $V : \mathbb{R}^d \to [1, +\infty)$ measurable and $M_V \geq 0$ satisfying*

$$\sup_{x\in\mathbb{R}^d} \left\{(1+\|x\|)^2 / V(x)\right\} \leq M_V \ .$$

Assume **L**1 *and that there exist $\lambda \in (0,1)$, $b \geq 0$ and $\bar{\gamma} > 0$ such that for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$ $\mathrm{R}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$. Assume that there exists $D_0 \geq 0$ such that for any $\theta \in \Theta$, $\pi_\theta(V) \leq D_0$. Then there exists $D_4 \geq 0$ such that for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$*

$$\|\pi_{\gamma,\theta} - \pi_\theta\|_{V^{1/2}} \leq D_4 \gamma^{1/2} \ .$$

Proof. Using Lemma 4.2.19 we obtain that for any $\theta \in \Theta$

$$\lim_{k\to+\infty} \|\pi_\theta \mathrm{R}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}} = \|\pi_{\gamma,\theta} - \pi_\theta\|_{V^{1/2}} \ . \tag{4.181}$$

We now give an upper bound on $\|\pi_\theta \mathrm{R}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}}$ for $k = q_\gamma m_\gamma$ with $m_\gamma = \lceil 1/\gamma \rceil$ and $q_\gamma \in \mathbb{N}$. Using Theorem 4.1.8 and that π_θ is invariant for $\mathrm{P}_{t,\theta}$ with $t \geq 0$, see Lemma 4.2.19, we obtain for all $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $k \in \mathbb{N}$

$$\begin{aligned}&\|\pi_\theta \mathrm{R}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}} \\ &\leq \sum_{\ell=0}^{q_\gamma - 1} \|\pi_\theta \mathrm{P}_{\gamma(\ell+1)m_\gamma,\theta} \mathrm{R}_{\gamma,\theta}^{(q_\gamma-(\ell+1))m_\gamma} - \pi_\theta \mathrm{P}_{\gamma\ell m_\gamma,\theta} \mathrm{R}_{\gamma,\theta}^{(q_\gamma-\ell)m_\gamma}\|_{V^{1/2}} \\ &\leq \sum_{\ell=0}^{q_\gamma - 1} C\xi^{\gamma m_\gamma (q_\gamma-(\ell+1))} \|\pi_\theta \mathrm{P}_{\gamma\ell m_\gamma,\theta} \mathrm{P}_{m_\gamma\gamma,\theta} - \pi_\theta \mathrm{P}_{\gamma\ell m_\gamma,\theta} \mathrm{R}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \\ &\leq \|\pi_\theta \mathrm{P}_{m_\gamma\gamma,\theta} - \pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \sum_{\ell=1}^{q_\gamma} C\xi^{\ell\gamma m_\gamma} \ ,\end{aligned} \tag{4.182}$$

where $C \geq 0, \xi \in (0,1)$ are the constants given by Theorem 4.1.8 with minorization condition given by Proposition 4.1.11-(a) with $\mathtt{m} = -\mathtt{L}$ since **L**1 holds and with drift condition $\mathbf{D}_{\mathrm{d}}(V^{1/2}, \lambda^\gamma, b\lambda^{-\bar{\gamma}/2}\gamma/2)$, since for all $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$ we have that $\mathrm{R}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$ and therefore using Jensen's inequality that $\mathrm{R}_{\gamma,\theta}$ satisfies the drift condition $\mathbf{D}_{\mathrm{d}}(V^{1/2}, \lambda^{\gamma/2}, b\lambda^{-\bar{\gamma}/2}\gamma/2)$.

We now give an upper bound on error $\|\pi_\theta \mathrm{P}_{m_\gamma\gamma,\theta} - \pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}}$. Indeed, since $\mathcal{A}_\theta$ satisfies a $\mathbf{D}_{\mathrm{c}}(V, \zeta, \beta)$ and $\mathrm{R}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$ for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, we obtain using (4.176) that for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$

$$\pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}(V) \leq D_0 \ , \qquad \pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma}(V) \leq \tilde{D}_1 \ , \qquad \tilde{D}_1 = D_0 + b\lambda^{-\bar{\gamma}} \log(1/\lambda)^{-1} \ , \tag{4.183}$$

Combining this result and Lemma 4.2.21 we have for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$

$$\|\pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta} - \pi_\theta \mathrm{R}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \leq \tilde{D}_2 \gamma^{1/2} \;, \tag{4.184}$$

with

$$\tilde{D}_2 = 2\tilde{D}_1^{1/2}(1+\bar{\gamma})^{1/2} \left\{ d + 2\bar{\gamma}(\mathtt{L}^2 M_V + \sup_{\theta\in\Theta} \|\nabla_x U_\theta(0)\|^2)\tilde{D}_1 \right\}^{1/2} \mathtt{L} \;. \tag{4.185}$$

Combining (4.182) and (4.184) we get for any $k \in \mathbb{N}$, $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$

$$\begin{aligned} \|\pi_\theta \mathrm{R}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}} &\leq C\tilde{D}_2 \sum_{\ell=1}^{q_\gamma} \xi^{\gamma m_\gamma \ell} \gamma^{1/2} \\ &\leq C\tilde{D}_2 (1-\xi)^{-1} \gamma^{1/2} \;, \end{aligned}$$

where we used that $\xi^{\gamma m_\gamma} \leq \xi$. The conclusion follows from this result and (4.181). □

Proof of Theorem 4.2.5 Combining Proposition 4.2.17 and Lemma 4.2.20 we get that **H**1-(i) is satisfied with constant $A_1 \leftarrow D_1$. **L**1, **L**2, Proposition 4.2.17 and Lemma 4.2.19-(a) ensure that **H**1-(ii) is satisfied by Theorem 4.1.18. **H**1-(iii) is satisfied combining Proposition 4.2.17, Proposition 4.2.18 and Proposition 4.2.22 with $\mathbf{\Psi}(\gamma) \leftarrow D_4 \gamma^{1/2}$.

Proof of Theorem 4.2.6

We preface the proof by the following technical lemma. Denote for any $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$, γ_{μ,σ^2} the d-dimensional Gaussian distribution with mean μ and covariance matrix $\sigma^2 \,\mathrm{Id}$.

Lemma 4.2.23. *For any $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\sigma_1, \sigma_2 > 0$,*

$$\mathrm{KL}\left(\gamma_{\mu_1,\sigma_1^2} | \gamma_{\mu_2,\sigma_2^2}\right) = (1/2)\left\{ d\log(\sigma_2^2/\sigma_1^2) + d\sigma_2^{-2}(\sigma_1^2 - \sigma_2^2) + \sigma_2^{-2}\|\mu_1 - \mu_2\|^2 \right\} \;.$$

In addition, if $\sigma_2 < \sigma_1$ then

$$\mathrm{KL}\left(\gamma_{\mu_1,\sigma_1^2} | \gamma_{\mu_2,\sigma_2^2}\right) \leq (1/2)\left\{ d\sigma_2^{-4}(\sigma_1^2 - \sigma_2^2)^2 + \sigma_2^{-2}\|\mu_1 - \mu_2\|^2 \right\} \;.$$

Proof. For any $x \in \mathbb{R}^d$ we have

$$\begin{aligned} &-\sigma_1^{-2}\|x-\mu_1\|^2 + \sigma_2^{-2}\|x-\mu_2\|^2 \\ &\qquad = (\sigma_2^{-2} - \sigma_1^{-2})\|x-\mu_1\|^2 + \sigma_2^{-2}\|\mu_1-\mu_2\|^2 + 2\sigma_2\langle x-\mu_1, \mu_1-\mu_2\rangle \;. \end{aligned} \tag{4.186}$$

Let X be a d-dimensional Gaussian random variable with mean μ_1 and covariance matrix $\sigma_1^2 \,\mathrm{Id}$. Using (4.186) we have

$$\mathbb{E}\left[-\sigma_1^{-2}\|X-\mu_1\|^2 + \sigma_2^{-2}\|X-\mu_2\|^2 \right] = d\sigma_1^2(\sigma_2^{-2} - \sigma_1^{-2}) + \sigma_2^{-2}\|\mu_1-\mu_2\|^2 \;.$$

Using this result we have

$$\begin{aligned} \mathrm{KL}\left(\gamma_{\mu_1,\sigma_1^2} | \gamma_{\mu_2,\sigma_2^2}\right) &= (1/2)\log\left(\frac{(2\pi)^d \sigma_2^{2d}}{(2\pi)^d \sigma_1^{2d}}\right) + (1/2)\mathbb{E}\left[-\sigma_1^{-2}\|X-\mu_1\|^2 + \sigma_2^{-2}\|X-\mu_2\|^2 \right] \\ &= (1/2)\left\{ d\log\left(\sigma_2^2/\sigma_1^2\right) + d\sigma_1^2(\sigma_2^{-2} - \sigma_1^{-2}) + \sigma_2^{-2}\|\mu_1-\mu_2\|^2 \right\} \end{aligned}$$

$$= (1/2)\left\{d\log(\sigma_2^2/\sigma_1^2) + d\sigma_2^{-2}(\sigma_1^2-\sigma_2^2) + \sigma_2^{-2}\|\mu_1-\mu_2\|^2\right\} ,$$

which concludes the first part of the proof. Using this result, the fact that for any $u \geq 0$, $-\log(1+u)+u \leq +u^2$ and $\sigma_1^2/\sigma_2^2 - 1$ we get that

$$\begin{aligned}
\mathrm{KL}\left(\gamma_{\mu_1,\sigma_1^2}|\gamma_{\mu_2,\sigma_2^2}\right) &= (1/2)\left\{d\log\left(\sigma_2^2/\sigma_1^2\right) + d\sigma_1^2(\sigma_2^{-2}-\sigma_1^{-2}) + \sigma_2^{-2}\left\|\mu_1-\mu_2\right\|^2\right\} \\
&= (1/2)\left\{-d\log\left(\sigma_1^2/\sigma_2^2\right) + d(\sigma_1^2/\sigma_2^{-2}-1) + \sigma_2^{-2}\left\|\mu_1-\mu_2\right\|^2\right\} \\
&= (1/2)\left\{-d\log\left(1+\left\{\sigma_1^2/\sigma_2^2-1\right\}\right) + d\left\{\sigma_1^2/\sigma_2^{-2}-1\right\} + \sigma_2^{-2}\left\|\mu_1-\mu_2\right\|^2\right\} \\
&\leq (1/2)\left\{d\sigma_2^{-4}(\sigma_1^2-\sigma_2^2)^2 + \sigma_2^{-2}\|\mu_1-\mu_2\|^2\right\} ,
\end{aligned}$$

which concludes the second part of the proof. □

Proposition 4.2.24. *Let $V : \mathbb{R}^d \to [1,+\infty)$ measurable and $M_{V,4} \geq 0$ satisfying*

$$\sup_{x\in\mathbb{R}^d}\left\{(1+\|x\|^4)/V(x)\right\} \leq M_{V,4} .$$

Assume that there exists $M \geq 1$ such that for any $\theta \in \Theta$, $\gamma \in (0,\bar{\gamma}]$, with $\bar{\gamma} > 0$ and $x \in \mathbb{R}^d$, $\mathrm{R}_{\gamma,\theta}V(x) \leq MV(x)$. Assume **L**1 *and* **L**3*, then we have for any $\theta_1,\theta_2 \in \Theta$, $\gamma_1,\gamma_2 \in (0,\bar{\gamma}]$ with $\gamma_2 < \gamma_1$, $a \in [1/4,1/2]$ and $x \in \mathbb{R}^d$*

$$\|\delta_x\mathrm{R}_{\gamma_1,\theta_1} - \delta_x\mathrm{R}_{\gamma_2,\theta_2}\|_{V^a} \leq D_5\left[\gamma_2^{-1/2}|\gamma_1-\gamma_2| + \gamma_2^{1/2}\|\theta_1-\theta_2\|\right]V^{2a}(x) ,$$

where $\{\mathrm{R}_{\gamma,\theta}, \gamma \in (0,\bar{\gamma}], \theta \in \Theta\}$ is the sequence of Markov kernels associated with the recursion (4.123) and

$$D_5 = \sqrt{2}M^{1/2}\max\left(\sqrt{2}(d+\bar{\gamma})^{1/2}\left[\sup_{\theta\in\Theta}\|\nabla_x U_\theta(0)\|^2 + \mathrm{L}^2 M_{4,V}^{1/2}\right]^{1/2}, L_U\right) .$$

Proof. Let $x \in \mathbb{R}^d$, $\theta_1,\theta_2 \in \Theta$ and $\gamma_1,\gamma_2 \in (0,\bar{\gamma}]$, $\gamma_2 < \gamma_1$. Using [DM17, Lemma 24] we have that

$$\begin{aligned}
&\|\delta_x\mathrm{R}_{\gamma_1,\theta_1} - \delta_x\mathrm{R}_{\gamma_2,\theta_2}\|_{V^a} \\
&\quad\leq \sqrt{2}\left(\mathrm{R}_{\gamma_1,\theta_1}V^{2a}(x) + \mathrm{R}_{\gamma_2,\theta_2}V^{2a}(x)\right)^{1/2}\mathrm{KL}\left(\delta_x\mathrm{R}_{\gamma_1,\theta_1}|\delta_x\mathrm{R}_{\gamma_2,\theta_2}\right)^{1/2} \\
&\quad\leq 2M^a V^a(x)\mathrm{KL}\left(\delta_x\mathrm{R}_{\gamma_1,\theta_1}|\delta_x\mathrm{R}_{\gamma_2,\theta_2}\right)^{1/2}
\end{aligned} \tag{4.187}$$

Using Lemma 4.2.23 we obtain that

$$\mathrm{KL}\left(\delta_x\mathrm{R}_{\gamma_1,\theta_1}|\delta_x\mathrm{R}_{\gamma_2,\theta_2}\right) \leq (1/2)\left\{d\gamma_2^{-2}(\gamma_1-\gamma_2)^2 + (2\gamma_2)^{-1}\Xi\right\} , \tag{4.188}$$

where Ξ is given by

$$\begin{aligned}
\Xi &= \|\gamma_1\nabla_x U_{\theta_1}(x) - \gamma_2\nabla_x U_{\theta_2}(x)\|^2 \\
&= \|\gamma_1\nabla_x U_{\theta_1}(x) - \gamma_2\nabla_x U_{\theta_1}(x) + \gamma_2\nabla_x U_{\theta_1}(x) - \gamma_2\nabla_x U_{\theta_2}(x)\|^2 \\
&\leq 2\|\gamma_1\nabla_x U_{\theta_1}(x) - \gamma_2\nabla_x U_{\theta_1}(x)\|^2 + 2\|\gamma_2\nabla_x U_{\theta_1}(x) - \gamma_2\nabla_x U_{\theta_2}(x)\|^2 \\
&\leq 2(\gamma_1-\gamma_2)^2\|\nabla_x U_{\theta_1}(x)\|^2 + 2\gamma_2^2\|\nabla_x U_{\theta_1}(x) - \nabla_x U_{\theta_2}(x)\|^2 \\
&\leq 2(\gamma_1-\gamma_2)^2\|\nabla_x U_{\theta_1}(x)\|^2 + 2\gamma_2^2 L_U^2\|\theta_1-\theta_2\|^2 V^{2a}(x) ,
\end{aligned} \tag{4.189}$$

using **L**3 for the last inequality. Using **L**3 again and that $\sup_{\theta\in\Theta}\|\nabla_x U_\theta(0)\| < +\infty$ by **L**1, we get for any $a \in [1/4, 1/2]$

$$\|\nabla_x U_{\theta_1}(x)\|^2 \leq 2(\|\nabla_x U_{\theta_1}(x) - \nabla_x U_{\theta_1}(0)\|^2 + \sup_{\theta_1\in\theta_1}\|\nabla_x U_{\theta_1}(0)\|^2) \leq C_\Theta V^{2a}(x) \;,$$

with $C_\Theta = 2\sup_{\theta\in\Theta}\|\nabla_x U_\theta(0)\|^2 + 2\mathrm{L}^2 M_{4,V}^{1/2}$. Combining this result and (4.189) in (4.188), it follows that

$$\begin{aligned}
&\mathrm{KL}\left(\delta_x \mathrm{R}_{\gamma_1,\theta_1} | \delta_x \mathrm{R}_{\gamma_2,\theta_2}\right) \\
&\leq (1/2)\left\{d\gamma_2^{-2}(\gamma_1-\gamma_2)^2 + \gamma_2^{-1}(\gamma_1-\gamma_2)^2 C_\Theta V^{2a}(x) + L_U^2\gamma_2\|\theta_1-\theta_2\|^2 V^{2a}(x)\right\} \\
&\leq (1/2)\left\{(d+\bar\gamma)C_\Theta\gamma_2^{-2}(\gamma_1-\gamma_2)^2 V^{2a}(x) + L_U^2\gamma_2\|\theta_1-\theta_1\|^2 V^{2a}(x)\right\} \\
&\leq (1/2)\max\left((d+\bar\gamma)C_\Theta, L_U^2\right)\left\{\gamma_2^{-2}(\gamma_1-\gamma_2)^2 + \gamma_2\|\theta_1-\theta_2\|^2\right\} \;.
\end{aligned}$$

This result substituted in (4.187) completes the proof with the fact that for any $a, b \in \mathbb{R}_+$, $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$. □

Proof of Theorem 4.2.6. **L**1 and **L**2 ensure a uniform drift condition on $\mathrm{R}_{\gamma,\theta}$, see Proposition 4.2.17 . Note that the Lyapunov function V defined by Proposition 4.2.17 satisfies $\sup_{x\in\mathbb{R}^d}(1+\|x\|^4)/V(x) < +\infty$. **H**2 is then a direct consequence of Proposition 4.2.24 □

Posterior convexity

Lemma 4.2.25. *For any $y \in \{0,1\}^{d_y}$, $\theta \mapsto p(y|\theta)$ given by (4.135) is log-concave.*

Proof. Let $\theta \in \mathbb{R}$, then by (4.135), for any $y \in \mathbb{R}$ we have $p(y|\theta) = \int_{\mathbb{R}^d} p(y,\beta|\theta)\mathrm{d}\beta$ with

$$p(y,\beta|\theta) = (2\pi\sigma^2)^{-d/2}\left\{\prod_{i=1}^{d_y} s(x_i^\top\beta)^{y_i}(1-s(x_i^\top\beta))^{1-y_i}\right\}\mathrm{e}^{-\frac{\|\beta-\theta\mathbf{1}_d\|^2}{2\sigma^2}} \;.$$

Therefore we have using that for any $t \in \mathbb{R}$, $1-s(t) = s(-t)$

$$\begin{aligned}
\log p(y,\beta|\theta) = (-d/2)\log(2\pi\sigma^2) \\
+\left\{\sum_{i=1}^{d_y} y_i \log(s(x_i^\top\beta)) + (1-y_i)\log(s(-x_i^\top\beta))\right\} - \frac{\|\beta-\theta\mathbf{1}_d\|^2}{2\sigma^2} \;.
\end{aligned}$$

Since $y_i \geq 0$, $1-y_i \geq 0$, $(\beta,\theta) \mapsto \|\beta-\theta\mathbf{1}_d\|^2$, $t \mapsto \log(s(t))$ and $t \mapsto \log(s(-t))$ are convex, we obtain that $(\beta,\theta) \mapsto p(y,\beta|\theta)$ is log-concave. Using the Prékopa–Leindler inequality [Gar02, Theorem 7.1] we obtain that $\theta \mapsto p(y|\theta)$ is log-concave which concludes the proof. □

Chapter 5

Entropy-based texture synthesis

5.1 Principle of maximum entropy

5.1.1 Abstract

Understanding texture formation is a crucial step towards a global theory of the human visual system as texture is an important perceptual cue. The more specific problem of exemplar-based texture synthesis arises in computer graphics where it is often desirable to be able to generate new large natural textures which look like an input image. This application highlights the need of a mathematically sound model for texture generation as our only criterion for evaluating the performance of algorithms is via human inspection.

Two main approaches have been proposed in the literature: the patch-based methods [EL99; EF01; Kwa+03; LB06; RDM16; Kwa+05; Han+06; Kas+15; GLR18] and the parametric ones [Wij91; GGM11; GM86; CM88; HB95; ZWM98; GEB15; JAF16; Uly+16; UVL17]. In this section, we are interested in the theoretical and visual properties of information-based parametric models. More precisely, we consider maximum entropy models. Indeed, the maximum entropy approach has the appealing property that the trade-off between innovation (maximizing the entropy) and the visual similarity with the input (geometrical or statistical feature constraints) is explicitly embedded in the model. There exist two main approaches for these maximum entropy formulation, the microcanonical model in which the constraints must be met almost surely and the macrocanonical model in which the constraints must be met in expectation [BM18]. Both share connections with statistical physics. In [BM18] the authors address the convergence of usual sampling scheme for the microcanonical model. In this section, we derive similar results for the macrocanonical model.

Contrary to the microcanonical model, the distribution of any macrocanonical model is a Gibbs measure, *i.e.* the exponential distribution of the features up to a scalar product with some parameters [Jay57]. Our first contribution is to give explicit conditions on the features ensuring the existence of such a macrocanonical model, extending results from information geometry [Csi96; Csi84; Csi75; DE97].

Even if such a Gibbs measure exists we are facing two issues: 1) finding the optimal parameters, 2) sampling from the associated Gibbs measure. The first challenge can in fact be seen as the dual formulation of the maximum entropy problem under constraints and corresponds to the minimization of a convex functional over an open susbet of $\mathbb{R}^p$ with $p \in \mathbb{N}$. Therefore it is natural to consider gradient based method in order to find such parameters and this approach was considered in the seminal work

of [ZWM98] which was the first to consider macrocanonical models in the context of image processing. However, the gradient of this functional is the expectation of the features with respect to the Gibbs measure. In this context, we turn ourselves to the Stochastic Approximation (SA) literature [RM51; CGG87]. More precisely, we use the Markov Chain Monte Carlo (MCMC) SA methodology proposed in Section 5.2.2 and referred as Stochastic Optimization with Unadjusted Langevin Algorithm (SOUL) which relies on the Langevin algorithm. This MCMC method has received a lot of attention in the recent years [Dal17b; Dal17a; DM17] since it exhibits desirable convergence properties and has been extensively used in machine learning applications [WT11; Sim+16; PT13; MCF15; ABW12]. Note that a similar methodology to SOUL was already used in a texture synthesis context in [LZW16].

Our second contribution is to establish the convergence of the methodology proposed in Section 5.2.2 in the context of macrocanonical texture synthesis and improve existing results on the dependency with respect to the hyperparameters in this specific case. In particular, the dependency in the dimension is polynomial even in the non-convex setting. This is in accordance with similar results for the convergence of diffusion processes with respect to the Kantorovitch-Rubinstein distance which are known to be optimal [Ebe16].

Chapter 5 is organized as follows. Related work on maximum entropy methods is discussed in Section 5.1.2 In Section 5.1.2, we give a mathematical presentation of the microcanonical and the macrocanonical models. In Section 5.1.2 we extend results from information geometry to the context of exemplar-based texture synthesis. Two sets of image descriptors are studied in Section 5.1.3: the Gaussian features in Section 5.1.3 and convolutional neural network features in Section 5.1.3. We then turn to the proposed algorithm for sampling macrocanonical models. The SOUL algorithm in the special case of texture synthesis is recalled in Section 5.2.2 and the convergence results applied to our settings are presented in Section 5.2.2. In Section 5.2.2 we investigate the links between the microcanonical model and the macrocanonical one. Numerical experiments are gathered in Section 5.2.3. First, we discuss the convergence of the SOUL algorithm considering a toy circular Gaussian texture synthesis problem in Section 5.2.3. Then, we apply our algorithm for general texture synthesis with features given by convolutional neural network outputs in Section 5.2.3. We study the advantages and the limitations of macrocanonical models and compare our visual results with existing algorithms. Finally an extension of our model to texture style transfer is presented in Section 5.2.3. Proofs and additional results are gathered in Section 5.1.4 and Section 5.2.4.

5.1.2 Statistical texture models

In this work we are interested in sampling probability distributions derived from image models. Let $x_0 \in \mathbb{R}^d$ be an exemplar image and consider a set of constraints associated with some image descriptors $F : \mathbb{R}^d \to \mathbb{R}^p$. Assume that $F(x_0) = 0$. This can always been achieved upon subtracting $F(x_0)$ to the original features. The constraints on the target distribution are then given by $F = 0$ almost surely or in expectation.

Previous work

As emphasized in Section 5.1.1, there have been two main approaches to address the exemplar-based texture synthesis problem. First, non-parametric methods sample an output image from a statistical process, *e.g.* a Markov random field [EL99; CCB85; PR77]. These methods do not require an explicit texture model and most of these algorithms are based on patch information, see the review paper [Raa+17]. Indeed, in order to update the current image, the patches of the input texture are rearranged in order to generate a new element (a pixel or a block of pixels) which is locally coherent with the pre-existing

structure. The seminal work of [EL99] paved the way for the use of such methods and was later extended in [EF01; Kwa+03] to handle blockwise updates instead of pixelwise ones. The statistical model of [EL99] was analyzed in [LB06] in which the authors reformulate the original algorithm as a bootstrap scheme. Since these methods duplicate some part of the input image in order to sample a new one, their innovation capability is limited. In [GLR18], starting from a random microtexture initial image, the patches are rearranged using optimal transport. [Kwa+05] reformulates a patch-based synthesis algorithm as an optimization procedure, therefore yielding a global texture model defined by the patch information. This model was later extended in [Han+06; Kas+15].

For the second type of approaches, *i.e.* parametrics ones, in the early work of [FFC82], textures were described as fractional Brownian motions. It was later noticed in [Wij91] that a large class of textures could be generated using spot noise models whose normalized limit for a large number of spots is a Gaussian random field with a circulant covariance matrix [GGM11]. In this works the underlying image model is Gaussian and the pixel distribution of the output image has, in expectation, the same moments of order 1 and 2 as the input texture. All the images in this class share the property that they do not exhibit salient spatial structures implying that the knowledge of the second-order moments was not enough to reproduce natural images. In [GM86; CM88; HB95; PS00] the authors remark that structured textures could be obtained using hand-selected features. The first texture synthesis methodology based on a maximum entropy approach with constraints in expectation is introduced in [ZWM98]. Replacing wavelet features in [PS00] by convolutional neural network features, [GEB15] obtains striking visual results. In a similar line of work, [JAF16; Uly+16; UVL17] design parametric methods relying on convolutional neural network features. The sampling procedure does not depend on any gradient-based method but instead is performed in a feed-forward manner. More recently, many papers investigate the use of Generative Adversarial Networks (GAN), see [Goo+14], in texture synthesis yielding promising results [JBV16; BJV17; LW16a; Zho+18]. In GANs [JBV16; Goo+14] the structure constraint is encoded in the loss on the generator, *i.e.* the samples must look like the input image. The innovation constraint is encoded in the loss on the discriminator, *i.e.* the samples must be diverse enough for the discrimination task to be hard. Note that in this case, even though the synthesis is performed in a feed-forward manner, the neural network must be retrained when presented a new class of textures. In the following approaches, the generation is not feed-forward but the natural texture distribution is designed so that the same features are generically used for all textures.

Maximum entropy probability measures

We now define microcanonical and macrocanonical models as introduced by [BM18]. Let μ be a probability measure over $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with $d \in \mathbb{N}$. The measure μ will be referred to as the *reference probability measure*. Let $F : \mathbb{R}^d \to \mathbb{R}^p$ with $p \in \mathbb{N}$ be a measurable mapping. This mapping will be referred to as the *statistical constraints* of the model. A discussion on the choice of F was conducted in Section 5.1.3. From now on we assume that we observe an input texture x_0 such that $F(x_0) = 0$. Given a set of features and a target image, we are now interested in the probability distributions which have maximum entropy (innovation constraint) and such that the features are equal to the ones of the target image (structure constraint) a.s.. In order for the model to be well-defined we replace the maximization of the entropy by a minimization of the Kullback-Leibler divergence $\mathrm{KL}(\cdot|\mu)$ where μ is the reference probability measure. This methodology is called the *microcanonical model*, see [BM18].

Definition 5.1.1. *A probability measure π is a* microcanonical model *with respect to the constraints F and the reference measure μ, if $\pi(\{F = 0\}) = 1$ and if for any probability distribution ν which satisfies the previous assumption we have $\mathrm{KL}(\pi|\mu) \leq \mathrm{KL}(\nu|\mu)$.*

This model is considered in [GM86; CM88; HB95; PS00; GEB15; LGX16; BM18]. Unfortunately, the

microcanonical model distribution is untractable for most statistical constraints. Therefore in order to sample from this distribution the following heuristics is used: 1) sample a white noise image; 2) perform a gradient descent on the square norm of the features. We shall see in Section 5.2.2 that this algorithm, while providing satisfying visual results for neural network constraints, does not sample from the microcanonical distribution. We now consider a relaxation of the *microcanonical model*: the *macrocanonical model*, see [BM18]. Given a set of features and a target image, we are now interested in the probability distributions which have maximum entropy (innovation constraint) and such that expectation of the features is equal to the features of the target image (structure constraint).

Definition 5.1.2. *A probability measure π is a* macrocanonical model *with respect to the constraints F and the reference measure μ, if F is π-integrable, $\pi(F) := \int_{\mathbb{R}^d} F(x)\mathrm{d}\pi(x) = 0$ and if for any probability distribution ν which satisfies the previous assumption we have $\mathrm{KL}(\pi|\mu) \leq \mathrm{KL}(\nu|\mu)$.*

This maximum entropy model was introduced by [Jay57] and used in texture synthesis by [ZWM98; LZW16]. In [BM18] conditions under which the macrocanonical and the microcanonical models coincide when the dimension of the image space grows towards infinity are identified.

Existence, uniqueness and dual formulation

Considering the model defined by Definition 5.1.2 it is natural to ask the following questions:

(a) When does a macrocanonical model exist? Can we identify explicit conditions for its existence?

(b) If such a model exists, is it unique?

(c) Can we find closed forms for the probability distribution functions of macrocanonical models?

We will answer positively to (b). In the case where the problem is non degenerate, *i.e.* there exists a probability measure ν such that $\mathrm{KL}(\nu|\mu) < +\infty$ and $\nu(F) = 0$, then the macrocanonical model exists and is given by a parametric measure, answering both (a) and (c). However, checking that the problem is indeed non degenerate can be as hard as finding the macrocanonical model. To show the existence of a macrocanonical model we give a dual, convex and finite dimensional formulation. This problem is then solved under the following conditions on F and μ. Let $\alpha > 0$ and $\beta > 0$:

A1 (α). *F is continuous and there exists $C_\alpha \geq 0$ with $\sup_{x\in\mathbb{R}^d} \left\{\|F(x)\|\left(1+\|x\|^\alpha\right)^{-1}\right\} \leq C_\alpha < +\infty$.*

A2 (β). *There exists $\eta > 0$, such that $\int_{\mathbb{R}^d} \exp[\eta \|x\|^\beta]\mathrm{d}\mu(x) < +\infty$.*

Let $\mathcal{P}_\alpha$ be the set of probability measures over $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $\int_{\mathbb{R}^d} \|x\|^\alpha \mathrm{d}\pi(x) < +\infty$. We define the set of *admissible probability measures* by $\mathcal{P}_\alpha^F = \{\pi \in \mathcal{P}_\alpha : \pi(F) = 0\}$ and consider the following problem:

$$\text{minimize } \mathrm{KL}(\pi|\mu) \text{ subject to } \pi \in \mathcal{P}_\alpha^F \,. \tag{P}$$

We denote $v(\mathrm{P}) = \inf_{\mathcal{P}_\alpha^F} \mathrm{KL}(\pi|\mu)$. First, we assert that if the solution of (P) exists and is non-degenerate, *i.e.* $v(\mathrm{P}) < +\infty$, then it is unique. Let $\pi_1^\star$ and $\pi_2^\star$ be two solutions of (P) with $v(\mathrm{P}) < +\infty$ and $\phi(t) = t\log(t)$, defined on $[0, +\infty)$ with the convention that $\phi(0) = 0$. Since $v(\mathrm{P}) < +\infty$ we have that $\pi_1^\star \ll \mu$ and $\pi_2^\star \ll \mu$. Since $\mathcal{P}_\alpha^F$ is convex, $\pi^\star$ defined for any $x \in \mathbb{R}^d$ by $\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}(x) = \left(\frac{\mathrm{d}\pi_1^\star}{\mathrm{d}\mu}(x) + \frac{\mathrm{d}\pi_2^\star}{\mathrm{d}\mu}(x)\right)/2$ belongs to $\mathcal{P}_\alpha^F$. Using that ϕ is strictly convex we have

$$2v(\mathrm{P}) = \mathrm{KL}(\pi_1^\star|\mu) + \mathrm{KL}(\pi_2^\star|\mu) = \int_{\mathbb{R}^d} \left\{\phi\left(\frac{\mathrm{d}\pi_1^\star}{\mathrm{d}\mu}(x)\right) + \phi\left(\frac{\mathrm{d}\pi_1^\star}{\mathrm{d}\mu}(x)\right)\right\} \mathrm{d}\mu(x) \geq 2\mathrm{KL}(\pi^\star|\mu) \,,$$

with equality if and only if for μ almost every $x \in \mathbb{R}^d$ we have $\frac{\mathrm{d}\pi_1^\star}{\mathrm{d}\mu}(x) = \frac{\mathrm{d}\pi_2^\star}{\mathrm{d}\mu}(x)$. As a consequence, since $\mathrm{KL}\left(\pi^\star|\mu\right) = v(\mathrm{P})$, $\pi_1^\star = \pi_2^\star$. We now introduce another optimization problem.

$$\text{maximize} \inf_{\pi \in \mathcal{P}_\alpha} \left\{ \mathrm{KL}\left(\pi|\mu\right) + \langle \theta, \pi(F) \rangle \right\} \text{ subject to } \theta \in \Theta_F \ , \tag{Q}$$

with

$$\Theta_F = \left\{ \theta \in \mathbb{R}^p \ : \ \int_{\mathbb{R}^d} \exp\left[-\langle \theta, F(x) \rangle \right] \mathrm{d}\mu(x) < +\infty \right\} \ . \tag{5.1}$$

Using Hölder's inequality, one can show that Θ_F is convex. In addition, if **A**1(α) and **A**2(α) hold with $\alpha > 0$ then $\bar{\mathrm{B}}(0, \eta/C_\alpha) \subset \mathrm{int}(\Theta_F)$. Similarly to (P), we denote $v(\mathrm{Q}) = \sup_{\Theta_F} \inf_{\pi \in \mathcal{P}_\alpha} \{\mathrm{KL}\left(\pi|\mu\right) + \langle \theta, \pi(F) \rangle\}$. Introducing the Lagrangian $\mathcal{L}(\pi, \theta) = \mathrm{KL}\left(\pi|\mu\right) + \langle \theta, \pi(F) \rangle$ defined over $\mathcal{P}_\alpha \times \Theta_F$, we have

$$v(\mathrm{Q}) = \sup_{\Theta_F} \inf_{\mathcal{P}_\alpha} \mathcal{L} \leq \inf_{\mathcal{P}_\alpha} \sup_{\Theta_F} \mathcal{L} \leq v(\mathrm{P}) \ . \tag{5.2}$$

We denote $\mathrm{d}(\mathrm{P}, \mathrm{Q})$ the *duality gap* $\mathrm{d}(\mathrm{P}, \mathrm{Q}) = v(\mathrm{P}) - v(\mathrm{Q})$ with the convention that $\infty - \infty = 0$. Let the log-partition function $L : \ \Theta_F \to \mathbb{R}$ be given for any $\theta \in \Theta_F$ by

$$L(\theta) = \log \left[\int_{\mathbb{R}^d} \exp[-\langle \theta, F(x) \rangle] \mathrm{d}\mu(x) \right] \ . \tag{5.3}$$

We also define for any $\theta \in \Theta_F$, the probability measure π_θ whose density with respect to μ is given for any $x \in \mathbb{R}^d$ by

$$\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(x) = \exp[-\langle \theta, F(x) \rangle - L(\theta)] \ . \tag{5.4}$$

Using Proposition 5.1.8, (Q) is equivalent to

$$\text{minimize } L(\theta) \text{ subject to } \theta \in \Theta_F \ . \tag{Q'}$$

More precisely $\theta^\star$ is a solution of (Q) if and only if it is a solution of (Q′). The next proposition is an extension of [Csi75, Theorem 3.1] in the case where F is not bounded. Under **A**2(α') with $\alpha' > \alpha$ the existence of a solution of (P) is ensured as soon as the set of admissible probability measures is not empty.

Proposition 5.1.3. *Assume* **A**1(α) *with* $\alpha \geq 0$. *The following holds.*

(a) If there exists a solution of (P) *such that* $v(\mathrm{P}) < +\infty$ *then there exists* $\theta^\star \in \mathbb{R}^p$ *such that the solution of* (P) *is given by* $\pi_{\theta^\star}$ *defined for* μ *almost any* $x \in \mathbb{R}^d$ *by*

$$\begin{cases} \frac{\mathrm{d}\pi_{\theta^\star}}{\mathrm{d}\mu}(x) = \exp\left[-\langle \theta^\star, F(x) \rangle\right] \Big/ \int_{\mathbb{R}^d \setminus \mathsf{N}} \exp\left[-\langle \theta^\star, F(y) \rangle\right] \mathrm{d}\mu(y) & \text{if } x \notin \mathsf{N} \ , \\ \frac{\mathrm{d}\pi_{\theta^\star}}{\mathrm{d}\mu}(x) = 0 & \text{otherwise} \ , \end{cases}$$

with $\mathsf{N} \in \mathcal{B}(\mathbb{R}^d)$ *such that for all* $\pi \in \mathcal{P}_\alpha^F$ *with* $\mathrm{KL}\left(\pi|\mu\right) < +\infty$, $\pi(\mathsf{N}) = 0$. *If there exists* $\pi \in \mathcal{P}_\alpha^F$ *with* $\mathrm{KL}\left(\pi|\mu\right) < +\infty$ *such that* $\mu \ll \pi$ *then* $\theta^\star$ *is a solution of* (Q), $\pi_{\theta^\star}$ *given by* (5.4) *is a solution of* (P) *and* $v(\mathrm{Q}) = v(\mathrm{P})$, *where* $v(\mathrm{Q})$ *and* $v(\mathrm{P})$ *are given in* (5.2).

(b) Assume **A**2(α') *with* $\alpha' > \alpha$. *There exists* $\pi_{\theta^\star}$ *solution of* (P) *with* $v(\mathrm{P}) < +\infty$ *if and only if there exists* $\pi \in \mathcal{P}_\alpha^F$ *such that* $\mathrm{KL}\left(\pi|\mu\right) < +\infty$.

Proof. The proof is postponed to Section 5.1.4. □

Note that this result was extended to a large class of divergences, see [AN00, Theorem 3.8]. In [Csi75; Csi84; Csi96] different sufficient conditions for solving (P) are derived. In [Top79, Theorem 3], the authors show similar results in the case where the Kullback-Leibler divergence is given by the true entropy. In practice it is difficult to check the conditions of Proposition 5.1.3. Indeed, even if $\mathbf{A2}(\alpha')$ holds with $\alpha' > \alpha$ it is not trivial to find an element $\pi \in \mathcal{P}_\alpha^F$ such that $\mathrm{KL}(\pi|\mu) < +\infty$. We now turn to the dual formulation (Q) which is easier to deal with since it is a finite dimensional and convex optimization problem. Under $\mathbf{A2}(\alpha)$, any stationary point of the log-partition function yields a solution of the primal formulation.

Proposition 5.1.4. *Assume* $\mathbf{A1}(\alpha)$ *and* $\mathbf{A2}(\alpha)$ *with* $\alpha \geq 0$. *Then,* $L \in \mathrm{C}^\infty(\mathrm{int}(\Theta_F))$, *where* L *is given in* (5.3), *and for any* $\theta \in \mathrm{int}(\Theta_F)$, $\nabla L(\theta) = \pi_\theta(F)$ *with* π_θ *given by* (5.4). *In addition, if there exists* $\theta^\star \in \mathrm{int}(\Theta_F)$ *such that* $\nabla L(\theta^\star) = 0$, *then* $\pi_{\theta^\star}$ *is the solution of* (P).

Proof. The proof is postponed to Section 5.1.4. □

A similar result was derived in [JM83], in the case where μ is no longer a probability measure but only sigma-finite. In fact, the log-partition function is a convex function, hence any stationary point is a global minimizer. We exploit this fact in the following proposition.

Proposition 5.1.5. *Assume* $\mathbf{A1}(\alpha)$, $\mathbf{A2}(\alpha')$ *with* $\alpha \geq 0$, $\alpha' > \alpha$.

(a) If for any $\theta \in \mathbb{R}^p$ *with* $\|\theta\| = 1$, *we have* $\mu(\{x \in \mathbb{R}^d \,:\, \langle F(x), \theta\rangle < 0\}) > 0$, *then* $\theta^\star$, *solution of* (Q), *exists and* $\pi_{\theta^\star}$ *given by* (5.4) *is the solution of* (P).

(b) In particular, Proposition 5.1.5-(a) is satisfied if $\mu(\mathsf{A}) > 0$ *for any non-empty open set* $\mathsf{A} \subset \mathbb{R}^d$, F *is continuous and there exists* $x \in F^{-1}(\{0\})$ *such that* F *is differentiable at* x *and we have* $\det(\mathrm{D}F(x)\mathrm{D}F(x)^\top) > 0$.

Proof. The proof is postponed to Section 5.1.4. □

In Section 5.1.3, we apply Proposition 5.1.5 when F is given by a convolutional neural network (CNN). However, note that it does not apply to the case where the non-linearity φ is the rectified linear unit (RELU), *i.e.* for any $t \in \mathbb{R}$, $\varphi(t) = \max(0, t)$. Still we give a similar result, valid for RELU, in Proposition 5.1.7. In [IM05, Theorem 3.5] the authors derive analogous results in the case where μ is the Lebesgue measure, *i.e.* in the case where the Kullback-Leibler is replaced by the true entropy. We are also able to show that the condition derived in Proposition 5.1.5-(a) is almost necessary.

Proposition 5.1.6. *Assume* $\mathbf{A1}(\alpha)$, $\mathbf{A2}(\alpha')$ *with* $\alpha \geq 0$, $\alpha' > \alpha$ *and that there exists* $\theta \in \mathbb{R}^p$ *with* $\|\theta\| = 1$ *such that* $\mu(\{x \in \mathbb{R}^d \,:\, \langle F(x), \theta\rangle \leq 0\}) = 0$. *Then* δ_{x_0} *solves* (P) *and* $v(\mathrm{P}) = +\infty$.

Proof. The proof is postponed to Section 5.1.4. □

We have seen that, under assumptions on the reference distribution μ and the statistical constraints F, the macrocanonical model is the distribution $\pi_{\theta^\star}$ with the following parametric form: for any $x \in \mathbb{R}^d$, $(\mathrm{d}\pi_{\theta^\star}/\mathrm{d}\mu)(x) = \exp\left[-\langle\theta^\star, F(x)\rangle - L(\theta^\star)\right]$, with $\theta^\star$ which satisfies $\theta^\star \in \arg\min_{\theta\in\Theta_F} L(\theta)$. These exponential families can also be retrieved in the following Bayesian framework. Assume that the likelihood of texture images associated with parameters $\theta \in \mathbb{R}^p$ is given for any $x \in \mathbb{R}^d$ by $p(x|\theta) = \exp[-\langle\theta, F(x)\rangle]/\int_{\mathbb{R}^d} \exp[-\langle\theta, F(y)\rangle]\mathrm{d}\mu(y)$ Assume that x_0 is a sample from this distribution and that $F(x_0) = 0$. Using Bayes' formula we obtain that for any $\theta \in \mathbb{R}^p$, $p(\theta|x_0) = p(x_0|\theta)p(\theta)/p(x_0) \propto p(\theta)/\int_{\mathbb{R}^d} \exp[-\langle\theta, F(y)\rangle]\mathrm{d}\mu(y)$. In order to compute the maximum a posteriori estimator θ_{MAP} we

need to set a prior distribution $p(\theta)$. Choosing the non-informative improper prior $p(\theta) = \mathbb{1}_{\Theta_F}(\theta)$ we get that

$$\theta_{\mathrm{MAP}} \in \operatorname*{arg\,min}_{\theta \in \Theta_F} \log \left[\int_{\mathbb{R}^d} \exp\left[-\langle \theta, F(x) \rangle\right] \mathrm{d}\mu(x) \right] ,$$

which corresponds to the dual formulation (Q). However, other prior distributions could be considered, yielding hierarchical Bayesian models [Vid+19].

5.1.3 Image descriptors

In this section, we review some of the popular possibilities for the choice of the function F. In the literature, many different approaches such as Gaussian models, *i.e.* mean and correlation features [Wij91; GGM11], wavelet-based descriptors [HB95; PS00; Pey10; Dui+07] or convolutional neural network features (CNN) [GEB15; Ust+16; JBV16] have been proposed to come up with visually satisfying image descriptors.

In our study we will focus on two sets of features: (i) Gaussian features ; (ii) CNN features. Gaussian features have the mathematical advantage of defining a strongly convex model, therefore allowing for strong convergence results to apply. However Gaussian textures do not exhibit sharp edges and lack long-range structures and as a consequence richer models should be investigated in order to obtain visually satisfying images. Similarly to [GEB15; Ust+16; LZW16] we consider features derived from a pretrained CNN. It has been observed that these features are efficient for describing a large variety of natural images. However, these improvements over the Gaussian model come at a high computational price. First, the features we end up with are no longer convex. Second, the dimension of the associated parameter space is usually high. An experimental investigation of the behavior of our proposed algorithm for these two sets of features is conducted in Section 5.2.3. We now describe precisely these two models.

Gaussian features

Let $x_0 \in \mathbb{R}^d$ and consider $F(x) = x * \check{x} - x_0 * \check{x}_0$. In the Fourier domain, we have for any $i \in \{0, \dots, d-1\}$, $\mathcal{F}(F(x))(i) = |\mathcal{F}(x)(i)|^2 - |\mathcal{F}(x_0)(i)|^2$. Therefore, if $F(x) = 0$, x has same power spectrum, *i.e.* same autocorrelation, as x_0, namely $\mathcal{F}(x)$ has the same modulus as $\mathcal{F}(x_0)$. However the equation $F(x) = 0$ gives no information on the phases of $\mathcal{F}(x)$.

We now study the macrocanonical model associated with F. Let μ, the reference measure, be a Gaussian probability measure with zero mean and diagonal covariance matrix with diagonal coefficient σ^2 with $\sigma > 0$. Then, **A**1(α) and **A**2(α) hold with $\alpha = 2$ and $\eta \in \left(0, (2\sigma^2)^{-1}\right)$. Assume in addition that for any $\ell \in \{0, \dots, d-1\}$, $\mathcal{F}(x_0)(\ell) \neq 0$. Using the Fourier-Plancherel formula we get that for any $x \in \mathbb{R}^d$ and $\theta = (\theta(0), \dots, \theta(d-1)) \in \mathbb{R}^d$,

$$\begin{aligned}
\langle \theta, F(x) \rangle + \|x\|^2 / (2\sigma^2) &= d^{-1} \left[\langle \mathcal{F}(\theta), |\mathcal{F}(x)|^2 - |\mathcal{F}(x_0)|^2 \rangle + \|\mathcal{F}(x)\|^2 / (2\sigma^2) \right] \\
&= \sum_{i=0}^{d-1} \left\{ d^{-1} (\mathcal{F}(\theta)(i) + (2\sigma^2)^{-1}) \, |\mathcal{F}(x)(i)|^2 \right\} - d^{-1} \langle \mathcal{F}(\theta), |\mathcal{F}(x_0)|^2 \rangle .
\end{aligned}$$

This implies that, $\Theta_F = \mathcal{F}^{-1}[\Re^{-1}\{\left(-(2\sigma^2)^{-1}, +\infty\right)^d\}] \cap \mathbb{R}^d$. In addition, for any $\theta \in \Theta_F$ with $\mathcal{F}(\theta) \in \mathbb{R}^d$ and X distributed according to π_θ, we obtain that $\mathcal{F}(X)$ is a d-dimensional complex Gaussian random variable on $\mathcal{F}(\mathbb{R}^d)$ with zero mean and diagonal covariance matrix with diagonal coefficients given by $d(\mathcal{F}(\theta) + (2\sigma^2)^{-1})^{-1}/2$. Similarly, we obtain that X distributed according to π_θ is

a d-dimensional Gaussian random variable with zero mean and invertible circulant covariance matrix $\mathbf{C}_\theta \in \mathrm{S}_d(\mathbb{R})$ whose inverse is given for any $i, j \in \{0, \dots, d-1\}$ by

$$\mathbf{C}_\theta^{-1}(i,j) = 2\theta(i-j) + \sigma^{-2} \; . \tag{5.5}$$

Note that in this case, since $\mathcal{F}(\theta) \in \mathbb{R}^d$, $\theta = \check{\theta}$. Let

$$\mathcal{F}(\theta^\star) = (d|\mathcal{F}(x_0)|^{-2} - \sigma^{-2})/2 \in \mathcal{F}(\Theta_F) \; . \tag{5.6}$$

In this case, for any X distributed according to $\pi_{\theta^\star}$ we obtain that $\mathcal{F}(X)$ is a d-dimensional Gaussian random variable with zero mean and diagonal covariance matrix with diagonal coefficients given by $|\mathcal{F}(x_0)|^2$. Therefore we have $\mathbb{E}[X * \check{X}] = \mathcal{F}^{-1}(\mathbb{E}[|\mathcal{F}(X)|^2]) = \mathcal{F}^{-1}(|\mathcal{F}(x_0)|^2) = x_0 * \check{x}_0$, which implies that $\pi_{\theta^\star}(F) = 0$ and that $\theta^\star = \mathcal{F}^{-1}(d|\mathcal{F}(x_0)|^{-2} - \sigma^{-2})/2$ is a solution of (Q), since $\theta^\star \in \Theta_F$. Using Proposition 5.1.4 we get that $\pi_{\theta^\star}$ is a solution of (P). Therefore the solution of (P) is the Gaussian probability measure with zero mean which satisfies for any $m, n \in \{0, \dots, d-1\}$, $\mathbb{E}\left[X(m)X(n)\right] = d^{-1}(x_0 * \check{x}_0)(m-n)$. This distribution is invariant by spatial translation, *i.e.* denoting $\tau : \mathbb{R}^d \to \mathbb{R}^d$, defined for any $x \in \mathbb{R}^d$ and $i \in \{0, \dots, d-1\}$ by $\tau(x)(i) = x(i+1)$ and $\tau(x)(d-1) = x(0)$, we have for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\pi^\star(\mathsf{A}) = \pi^\star(\tau(\mathsf{A}))$. Note that the distribution of the random variable X is the same as the one of $d^{-1/2}(x_0 * Z)$ with Z a d-dimensional Gaussian random variable with zero mean and covariance matrix identity, see [GGM11].

Neural network features

A Neural network is a series of affine operations (usually convolutions) followed at each step by a pointwise non-linearity. We define

$$(A_j^k)_{j \in \{1,\dots,M\}, k \in \{1,\dots,c_j\}} \in \prod_{j=1}^{M} \mathbb{A}_{n_j, c_{j-1} \times n_{j-1}}(\mathbb{R})^{c_j} \; , \quad (n_j, c_j)_{j \in \{0,\dots,M\}} \in \mathbb{N}^{M+1} \times \mathbb{N}^{M+1} \; , \tag{5.7}$$

with $M \in \mathbb{N}$, $n_0 = d$ and $c_0 = 1$. For each $j \in \{0, \dots, M\}$, n_j is the *dimension* of the j-th layer and c_j is the *number of channels* of the j-th layer, and for any $k \in \{1, \dots, c_j\}$, $A_j^k \in \mathbb{A}_{n_j, c_{j-1} \times n_{j-1}}$. Namely for any layer $j \in \{1, \dots, M\}$ and channel $k \in \{1, \dots, c_j\}$, $A_j^k : \mathbb{R}^{c_{j-1} \times n_{j-1}} \to \mathbb{R}^{n_j}$ is the affine operator which maps the $(j-1)$-th layer to the j-th layer and channel k before the non-linear operation. With our notations, the 0-th layer corresponds to the original image. We recall that for any $j \in \{1, \dots, M\}$ and $k \in \{1, \dots, c_j\}$, $\tilde{A}_j^k$ denotes the linear part of A_j^k. We also define for any $j \in \{1, \dots, M\}$ and $x \in \mathbb{R}^{c_{j-1} \times n_{j-1}}$, $A_j(x) = (A_j^k(x))_{k \in \{1,\dots,c_j\}}$, *i.e.* $A_j \in \mathbb{A}_{c_j \times n_j, c_{j-1} \times n_{j-1}}$ is the affine operator which maps the $j-1$-th layer to the j-th layer before the non-linear operation.

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a measurable function. By a slight abuse of notation we denote for any $d \in \mathbb{N}$ and $x \in \mathbb{R}^d$, $\varphi(x) = (\varphi(x(0)), \dots, \varphi(x(d-1)))$. We assume that φ satisfies the following conditions: (a) for any $d \in \mathbb{N}$, there exists $C_{d,\varphi} \geq 0$ such that for any $x \in \mathbb{R}^d$, $\|\varphi(x)\| \leq C_{d,\varphi}(1 + \|x\|)$, (b) φ is non-decreasing, (c) $\lim_{t \to +\infty} \varphi(t) = +\infty$. We define for any $j \in \{1, \dots, M\}$ and $k \in \{1, \dots, c_j\}$, the (j,k)-th *layer-channel feature* $\mathscr{G}_j^k : \mathbb{R}^d \to \mathbb{R}^{n_j}$, for any $x \in \mathbb{R}^d$, by

$$\mathscr{G}_j^k(x) = \left(\varphi \circ A_j^k \circ \varphi \circ A_{j-1} \circ \cdots \circ \varphi \circ A_1\right)(x) \; , \qquad \mathscr{G}_0(x) = x \; .$$

For any layer $j \in \{1, \dots, M\}$ and channel $k \in \{1, \dots, c_j\}$, $\mathscr{G}_j^k(x)$ is the neural network response of x at layer j and channel k. We also define for any $j \in \{1, \dots, M\}$, the j-th *layer feature* $\mathscr{G}_j : \mathbb{R}^d \to \mathbb{R}^{c_j \times n_j}$, for any $x \in \mathbb{R}^d$, by $\mathscr{G}_j(x) = (\mathscr{G}_j^k(x))_{k \in \{1,\dots,c_j\}}$. Let $\mathbf{j} \subset \{1, \dots, M\}$ then we can define $F(x) \in \mathbb{R}^p$ for

any $x \in \mathbb{R}^d$ by

$$F(x) = \left(\overline{\mathscr{G}}_j^k(x) - \overline{\mathscr{G}}_j^k(x_0)\right)_{j\in\mathbf{j},k\in\{1,\dots,c_j\}} , \qquad \overline{\mathscr{G}}_j^k(x) = n_j^{-1}\sum_{\ell=1}^{n_j}\mathscr{G}_j^k(x)(\ell) , \qquad p = \sum_{j\in\mathbf{j}} c_j . \quad (5.8)$$

A few remarks are in order regarding the dimension of the associated parameter space. In our applications, we will use the `VGG19` convolutional neural network [SZ14], see Section 5.2.4 for details on its structure. Note that since `VGG19` is a convolutional neural network, *i.e.* the linear part of the affine operators is given by a convolutional operator, and since we average the neural network response, the output dimension p is independent of the input dimension d. Selecting the layers $\mathbf{j} = \{1, 3, 6, 8, 11, 13, 15, 24, 26, 31\}$ we have $p = 2688 \approx 10^3$. Usually we will consider images of size at least 512×512 for which $d = 262144 \approx 10^5$. Therefore the features described by (5.8) performs a dimension reduction. In [GEB15] similar image descriptors are considered but Gram matrices are used instead of $(\overline{\mathscr{G}_j})_{j\in\mathbf{j},k\in\{1,\dots,c_j\}}$. This leads to a parameter space with dimension $352256 \approx 10^5$, see [Raa+17].

We now turn to the study of the macrocanonical model associated with F. Let the reference measure μ be a Gaussian probability measure with zero mean and symmetric positive covariance matrix. Then, **A**1(1) and **A**2(α') hold for any $\alpha' \in [0, 2)$. Note that in the case, where φ is differentiable, the results of Proposition 5.1.4 hold assuming that there exists a point $x \in \mathbb{R}^d$ such that $F(x) = F(x_0)$ and $\mathrm{d}F(x)$ is surjective. In the case where for all $t \in \mathbb{R}$, $\varphi(t) = \max(0, t)$ we can define a certificate ensuring the existence of a macrocanonical model.

Before stating Proposition 5.1.7, we introduce some preliminary notations. Let $(A_j)_{j\in\{1,\dots,M\}}$ be given by (5.7) and $(\mathscr{G}_j)_{j\in\{1,\dots,M\}}$ be given by (5.8). For any $j \in \{1, \dots, M\}$, let $A_{j,+} \in \mathbb{A}_{c_j\times n_j, c_{j-1}\times n_{j-1}}(\mathbb{R})$ defined by $A_{j,+} = D_j(x_0)A_j$, with $D_j(x_0) \in \mathbb{R}^{c_j\times n_j} \times \mathbb{R}^{c_j\times n_j}$ a diagonal matrix such that for any $i \in \{1, \dots, c_j \times n_j\}$, $D_j(x_0)(i, i) = \mathbb{1}_{(0,+\infty)}(\mathscr{G}_j(x_0)(i))$. Note that we have for any $j \in \{1, \dots, M\}$, $\tilde{A}_{j,+} = D_j(x_0)\tilde{A}_j$. Let $(\tilde{v}_{j,k})_{j\in\mathbf{j},k\in\{1,\dots,c_j\}}$ such that for any $j, j' \in \mathbf{j}$ and $k, k' \in \{1, \dots, c_j\}$, $\tilde{v}_{j,k} \in \mathbb{R}^{c_j\times n_j}$ and

$$\tilde{v}_{j,c}(j', c') = \begin{cases} n_j^{-1} & \text{if } c' = c , \\ 0 & \text{otherwise} . \end{cases}$$

Note that for any $x \in \mathbb{R}^d$, $F(x) = (\tilde{v}_{j,k}^\top \mathscr{G}_j(x))_{j\in\mathbf{j},k\in\{1,\dots,c_j\}}$. In addition, define for any $j \in \mathbf{j}$ and $k \in \{1, \dots, c_j\}$, $v_{j,k} = \tilde{A}_{1,+}^\top \dots \tilde{A}_{j,+}^\top \tilde{v}_{j,k}$.

Note that for any $j \in \mathbf{j}$ and $k \in \{1, \dots, c_j\}$, $v_{j,k} \in \mathbb{R}^d$. The next proposition highlights the role of $(v_{j,k})_{j\in\mathbf{j},k\in\{1,\dots,c_j\}}$ as a certificate for the existence of the macrocanonical model.

Proposition 5.1.7. *Assume that* **A**2(α) *holds with* $\alpha > 1$, *that* $\mu(\mathsf{A}) > 0$ *for every non-empty open set* $\mathsf{A} \subset \mathbb{R}^d$ *and that* F *is given by* (5.8) *with* $\varphi(t) = \max(0, t)$ *for any* $t \in \mathbb{R}$, $x_0 \in \mathbb{R}^d$ *and* $\mathbf{j} \subset \{1, \dots, M\}$. *Moreover assume that*

$$x_0 \in \bigcap_{j=1}^{M}\bigcap_{k=1}^{c_j\times n_j} \left(\partial\{x \in \mathbb{R}^d \,:\, e_{k,j}^\top A_j \mathscr{G}_{j-1}(x) = 0\}\right)^{\mathrm{c}} , \quad (5.9)$$

where for any $j \in \{1, \dots, M\}$, $(e_{k,j})_{k\in\{1,\dots,c_j\times n_j\}}$ *is the canonical basis of* $\mathbb{R}^{c_j\times n_j}$. *In addition, assume that the family* $(v_{j,c})_{j\in\mathbf{j},c\in\{1,\dots,c_j\}}$ *is linearly independent. Then there exists a solution* $\theta^\star$ *to* (Q) *and* $\pi_{\theta^\star}$ *given by* (5.4) *is the solution of* (P).

Proof. The proof is postponed to Section 5.1.4. □

If for any $j \in \{1, \dots, M\}$ and $k \in \{1, \dots, c_j \times n_j\}$, $e_{k,j}^\top A_j \mathscr{G}_{j-1}(x_0) \neq 0$ then (5.9) is satisfied. Note that since $(v_{j,c})_{j \in \mathbf{j}, c \in \{1,\dots,c_j\}}$ has a closed form, the independence condition in Proposition 5.1.7 can be explicitly checked given a trained neural network. The proof of Proposition 5.1.7 exploits the fact that neural networks are locally linear under mild assumptions. Finally, since the family $(v_{j,c})_{j \in \mathbf{j}, c \in \{1,\dots,c_j\}}$ has cardinality $p = \sum_{j \in \mathbf{j}} c_j$, Proposition 5.1.7 never applies when $p > d$.

5.1.4 Proofs and additional results

We have the following variational formula which is an extension of [DE97, Proposition 1.4.2] to the case where F is not bounded. More precisely, allowing some growth on F, controlled by a parameter α, and restricting the set of probability measures we consider to $\mathcal{P}_\alpha$ we obtain the same equality. The proof is almost identical but is given for completeness.

Proposition 5.1.8. *Assume* **A**1(α) *with* $\alpha > 0$. *Then, for any* $\theta \in \Theta_F$, *with* Θ_F *defined by* (5.1),

$$\inf_{\pi \in \mathcal{P}_\alpha} \{\mathrm{KL}\left(\pi|\mu\right) + \langle \theta, \pi(F) \rangle\} = -\log \left\{ \int_{\mathbb{R}^d} \exp\left[-\langle \theta, F(x) \rangle\right] \mathrm{d}\mu(x) \right\} .$$

Proof. Let $\theta \in \Theta_F$ and $\pi \in \mathcal{P}_\alpha$. Note that under **A**1(α), $\pi(\|F\|) < +\infty$ and $\pi(F)$ is well defined.

If $\mathrm{KL}\left(\pi|\mu\right) = +\infty$, then $\mathrm{KL}\left(\pi|\mu\right) + \langle \theta, \pi(F) \rangle = +\infty$. Consider now the case $\mathrm{KL}\left(\pi|\mu\right) < +\infty$. By definition of Θ_F, we can therefore consider π_θ, the probability measure with density with respect to μ given for any $x \in \mathbb{R}^d$ by

$$\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(x) = \exp[-\langle \theta, F(x) \rangle] \Big/ \int_{\mathbb{R}^d} \exp[-\langle \theta, F(y) \rangle] \mathrm{d}\mu(y) \ .$$

Note that since μ-almost everywhere, $(\mathrm{d}\pi_\theta)/(\mathrm{d}\mu)(x) > 0$, μ and π_θ are equivalent. Since $\mathrm{KL}\left(\pi|\mu\right) < +\infty$, $\pi \ll \mu$ which implies in turn $\pi \ll \pi_\theta$ and we have

$$\begin{aligned} \mathrm{KL}\left(\pi|\mu\right) &= \mathrm{KL}\left(\pi|\pi_\theta\right) + \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(x)\right) \mathrm{d}\pi(x) \\ &= \mathrm{KL}\left(\pi|\pi_\theta\right) - \langle \theta, \pi(F) \rangle - \log \left\{ \int_{\mathbb{R}^d} \exp\left[-\langle \theta, F(x) \rangle\right] \mathrm{d}\mu(x) \right\} , \end{aligned}$$

which concludes the proof, since $\mathrm{KL}\left(\pi|\pi_\theta\right) \geq 0$. □

Proof of Proposition 5.1.3

The proof is divided in two parts:

(a) Assume that there exists $\pi^\star$, solution of (P). Let $\mathcal{E}$ be the convex set defined by $\{\pi \in \mathcal{P}_\alpha^F : \frac{\mathrm{d}\pi}{\mathrm{d}\pi^\star}(x) \leq 2 \text{ for } \pi^\star \text{ almost every } x\}$. For any $\pi_1 \in \mathcal{E}$, consider π_2 with density with respect to $\pi^\star$, $\frac{\mathrm{d}\pi_2}{\mathrm{d}\pi^\star} = 2 - \frac{\mathrm{d}\pi_1}{\mathrm{d}\pi^\star}$ which by definition is an element of $\mathcal{E}$ and $\pi^\star = (\pi_1 + \pi_2)/2$. Hence $\pi^\star$ is an algebraic inner point of $\mathcal{E}$. Therefore using the equality case in [Csi75, Theorem 2.2] we obtain that for any $\pi \in \mathcal{E}$, $\mathrm{KL}\left(\pi|\mu\right) = \mathrm{KL}\left(\pi|\pi^\star\right) + \mathrm{KL}\left(\pi^\star|\mu\right)$. Using that for any $\pi \in \mathcal{E}$, we have $\mathrm{KL}\left(\pi|\pi^\star\right)$ and $\mathrm{KL}\left(\pi^\star|\mu\right) < +\infty$, we get that

$$0 = \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)\right) \mathrm{d}\pi(x) - \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\pi^\star}(x)\right) \mathrm{d}\pi(x) - \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}(x)\right) \mathrm{d}\pi^\star(x) \quad (5.10)$$

$$= \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}(x)\right)\left[\frac{\mathrm{d}\pi}{\mathrm{d}\pi^\star}(x) - 1\right]\mathrm{d}\pi^\star(x) \ .$$

Since, $\mathrm{KL}\left(\pi^\star|\mu\right) < +\infty$ we have that $\log(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}) \in \mathrm{L}^1(\mathbb{R}^d, \pi^\star)$. Let $V = \mathrm{span}\{1, \langle F, e_i\rangle \ : \ i = 1, \ldots, p\}$ where $(e_i)_{i\in\{1,\ldots,p\}}$ is the canonical basis of $\mathbb{R}^p$. V is a finite dimensional (hence closed) subspace of $\mathrm{L}^1(\mathbb{R}^d, \pi^\star)$. Hence, in order to show that $\log(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}) \in V$ it suffices to show that $\log(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}) \in (V^\perp)^\perp$ by [Bre11, Proposition II.12].

We identify the topological dual space of $\mathrm{L}^1(\mathbb{R}^d, \pi^\star)$ and $\mathrm{L}^\infty(\mathbb{R}^d, \pi^\star)$, see [Rud06, Theorem 6.16]. Let $h \in \mathrm{L}^\infty(\mathbb{R}^d, \pi^\star) \cap V^\perp$. Then by definition, $\int_{\mathbb{R}^d} F(x)h(x)\mathrm{d}\pi^\star(x) = 0$ and $\int_{\mathbb{R}^d} h(x)\mathrm{d}\pi^\star(x) = 0$. The same goes for $\tilde{h} = h/\|h\|_\infty$, and we have that π_h defined by $\frac{\mathrm{d}\pi_h}{\mathrm{d}\pi^\star} = 1 + \tilde{h}$ is an element of $\mathcal{E}$. Therefore, by (5.10), we get that $\int_{\mathbb{R}^d} \log(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}(x))h(x) = 0$ and $\log(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}) \in V$. More precisely, there exists $\theta^\star \in \mathbb{R}^p$, $C \in \mathbb{R}$ and $\mathsf{N} \in \mathcal{B}(\mathbb{R}^d)$ with $\pi^\star(\mathsf{N}) = 0$ such that for μ almost any $x \in \mathbb{R}^d \backslash \mathsf{N}$,

$$\log\left(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}(x)\right) = \langle \theta^\star, F(x)\rangle + C \ .$$

We also have that $\pi^\star(\mathsf{N}) = \int_{\mathsf{N}} \frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}(y)\mathrm{d}\mu(y)$ and therefore for μ almost any $x \in \mathsf{N}$, $\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}(x) = 0$. Using [Csi75, Remark 2.14], for any $\pi \in \mathcal{P}_\alpha^F$ such that $\mathrm{KL}\left(\pi|\mu\right) < +\infty$ we have $\pi \ll \pi^\star$ and therefore

$$\pi(\mathsf{N}) = 0 \ . \tag{5.11}$$

Finally, if there exists $\pi \in \mathcal{P}_\alpha^F$ with $\mathrm{KL}\left(\pi|\mu\right) < +\infty$ such that $\mu \ll \pi$ then by (5.11), $\mu(\mathsf{N}) = \pi(\mathsf{N}) = 0$ and we get that $\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu}(x) = \exp\left[-\langle\theta^\star, F(x)\rangle\right] / \int_{\mathbb{R}^d} \exp\left[-\langle\theta^\star, F(y)\rangle\right] \mathrm{d}\mu(y)$ for μ almost every $x \in \mathbb{R}^d$. Then, using Proposition 5.1.8 and that $\theta^\star \in \Theta_F$, we have by definition of (Q), see (5.2),

$$\begin{aligned} v(\mathrm{Q}) \leq v(\mathrm{P}) = \mathrm{KL}\left(\pi^\star|\mu\right) = -\log\left(\int_{\mathbb{R}^d} \exp\left[-\langle\theta^\star, F(x)\rangle\right] \mathrm{d}\mu(x)\right) \\ = \inf_{\pi\in\mathcal{P}_\alpha}\left\{\mathrm{KL}\left(\pi|\mu\right) - \langle\theta^\star, \pi(F)\rangle\right\} \leq v(\mathrm{Q}) \ , \end{aligned}$$

which concludes the first part of the proof.

(b) If there exists $\pi^\star$ solution of (P) with $v(\mathrm{P}) < +\infty$ then $\mathrm{KL}\left(\pi^\star|\mu\right) < +\infty$ and $\pi^\star \in \mathcal{P}_\alpha^F$. Now, assume that there exists $\pi \in \mathcal{P}_\alpha^F$ such that $\mathrm{KL}\left(\pi|\mu\right) < +\infty$. Let $(\pi_n)_{n\in\mathbb{N}}$ be a sequence of probability measures such that for any $n \in \mathbb{N}$, $\pi_n \in \mathcal{P}_\alpha^F$, $\mathrm{KL}\left(\pi_n|\mu\right) < +\infty$ and $\inf_{\mathcal{P}_\alpha^F} \mathrm{KL}\left(\pi|\mu\right) = \lim_{n\to+\infty} \mathrm{KL}\left(\pi_n|\mu\right)$. Using [DV76, Lemma 5.1] we get that $(\pi_n)_{n\in\mathbb{N}}$ is tight. Therefore we can assume, up to extraction, that $(\pi_n)_{n\in\mathbb{N}}$ converges to some probability measure $\pi^\star$ for the weak topology. Since $\pi \mapsto \mathrm{KL}\left(\pi|\mu\right)$ is lower semi-continuous [DE97, Lemma 1.4.3 (b)] we obtain that $\mathrm{KL}\left(\pi^\star|\mu\right) = \inf_{\mathcal{P}_\alpha^F} \mathrm{KL}\left(\pi|\mu\right)$. We recall the Donsker-Varadhan variational formula [DV75, Lemma 2.1] stating that for any continuous, real-valued and bounded mapping ϕ we have for any $n \in \mathbb{N}$

$$\int_{\mathbb{R}^d} \phi(x)\mathrm{d}\pi_n(x) \leq \mathrm{KL}\left(\pi_n|\mu\right) + \log\left(\int_{\mathbb{R}^d} \mathrm{e}^{\phi(x)}\mathrm{d}\mu(x)\right) \ . \tag{5.12}$$

Let $\varphi_M : \mathbb{R}^d \to \mathbb{R}$ defined for any $M \geq 0$ and $x \in \mathbb{R}^d$ by $\varphi_M(x) = \eta \max(\|x\|^{\alpha'}, M)$, with η defined in **A**2(α'). Using (5.12), **A**2(α') and that φ_M is continuous and bounded we get that for any $n \in \mathbb{N}$ and $M \geq 0$

$$\int_{\mathbb{R}^d} \varphi_M(x)\mathrm{d}\pi_n(x) \leq \sup_{n\in\mathbb{N}} \mathrm{KL}\left(\pi_n|\mu\right) + \log\left(\int_{\mathbb{R}^d} \exp[\eta\,\|x\|^{\alpha'}]\mathrm{d}\mu(x)\right) < +\infty \ .$$

Using the monotone convergence theorem we get that $\sup_{n\in\mathbb{N}} \int_{\mathbb{R}^d} \|x\|^{\alpha'} \mathrm{d}\pi_n(x) < +\infty$. By [Kal06, Theorem 5.16] there exist $(X_n)_{n\in\mathbb{N}}$ a sequence of $\mathbb{R}^d$ random variables and X a $\mathbb{R}^d$ random variable such that for any $n \in \mathbb{N}$, X_n is distributed according to π_n and X is distributed according to $\pi^\star$. Since $(\pi_n)_{n\in\mathbb{N}}$ converges weakly to $\pi^\star$, $(X_n)_{n\in\mathbb{N}}$ converges in distribution towards X. Therefore, we get that $(\|X_n\|^\alpha)_{n\in\mathbb{N}}$ converges in distribution to $\|X\|^\alpha$ and $\sup_{n\in\mathbb{N}} \mathbb{E}[\|X\|^{\alpha'}] < \infty$. By [Kal06, Lemma 3.11], we get that $\mathbb{E}[\|X\|^{\alpha'}] = \int_{\mathbb{R}^d} \|x\|^{\alpha'} \mathrm{d}\pi^\star(x) < +\infty$. Hence, $\pi^\star \in \mathcal{P}_\alpha$. In addition, since F is continuous by **A**1, we have that $(F(X_n))_{n\in\mathbb{N}}$ converges in distribution to $F(X)$. By [Wil91, Section 13.3] and **A**1(α), we have that $(F(X_n))_{n\in\mathbb{N}}$ is uniformly integrable. Using [Kal06, Lemma 3.11] and that for any $n \in \mathbb{N}$, $\pi_n(F) = 0$, we get $\lim_{n\to+\infty} \pi_n(F) = \pi^\star(F) = 0$ and $\pi^\star \in \mathcal{P}_\alpha^F$, which concludes the proof.

Proof of Proposition 5.1.4

Let $L : \mathrm{int}(\Theta_F) \to \mathbb{R}$ be the function defined for any $\theta \in \mathrm{int}(\Theta_F)$ by

$$L(\theta) = \log\left\{\int_{\mathbb{R}^d} \exp\left[-\langle\theta, F(x)\rangle\right] \mathrm{d}\mu(x)\right\} .$$

We have $L \in \mathrm{C}^\infty(\mathrm{int}(\Theta_F))$. The proposition is trivial if $\mathrm{int}(\Theta_F) = \emptyset$. Therefore we suppose that $\mathrm{int}(\Theta_F) \neq \emptyset$ and let $\theta_0 \in \mathrm{int}(\Theta_F)$. Since $\mathrm{int}(\Theta_F)$ is open, there exists $a_1 > 1$ such that $a_1\theta_0 \in \mathrm{int}(\Theta_F)$. Let $a_2 > 1$ such that $1/a_1 + 1/a_2 = 1$. Let $R = \eta/(2a_2)$ with η given in **A**2(α). For any $\theta \in \bar{\mathrm{B}}(\theta_0, R)$, using that $t^j \mathrm{e}^{-t} \leq j^j$ for $t \geq 0$ and $j \in \mathbb{N}$, we have for any $x \in \mathbb{R}^d$ and $k \in \mathbb{N}$

$$\begin{aligned}\|F(x)\|^k \exp\left[-\langle\theta, F(x)\rangle\right] &\leq (k/R)^k \exp\left[R\|F(x)\|\right] \exp\left[-\langle\theta, F(x)\rangle\right] \\ &\leq (k/R)^k \exp\left[2R\|F(x)\|\right] \exp\left[-\langle\theta_0, F(x)\rangle\right] ,\end{aligned}$$

The last quantity is independent of θ and μ-integrable using Hölder's inequality, since

$$\begin{aligned}&\int_{\mathbb{R}^d} \exp\left[2R\|F(x)\|\right] \exp\left[-\langle\theta_0, F(x)\rangle\right] \mathrm{d}\mu(x) \\ &\qquad \leq \left(\int_{\mathbb{R}^d} \exp\left[\eta\|x\|^\alpha\right] \mathrm{d}\mu(x)\right)^{1/a_2} \left(\int_{\mathbb{R}^d} \exp\left[-\langle a_1\theta_0, F(x)\rangle\right] \mathrm{d}\mu(x)\right)^{1/a_1} < +\infty .\end{aligned}$$

This result implies that $L \in \mathrm{C}^\infty(\mathrm{int}(\Theta_F))$. Therefore, if $\theta^\star$ is a stationary point, we have

$$\pi_{\theta^\star}(F) \propto \int_{\mathbb{R}^d} F(x) \exp\left[-\langle\theta^\star, F(x)\rangle\right] \mathrm{d}\mu(x) = 0 .$$

Since $\pi_{\theta^\star} \in \mathcal{P}_\alpha$ we have $\pi_{\theta^\star} \in \mathcal{P}_\alpha^F$. Since $\mu \ll \pi_{\theta^\star}$ we have $\pi \ll \pi_{\theta^\star}$ for any $\pi \ll \mu$. Therefore for any $\pi \in \mathcal{P}_\alpha^F$ with $\pi \ll \mu$ we have

$$\mathrm{KL}\left(\pi|\mu\right) = \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)\right) \mathrm{d}\pi(x) = \mathrm{KL}\left(\pi|\pi_{\theta^\star}\right) - L(\theta^\star) = \mathrm{KL}\left(\pi|\pi_{\theta^\star}\right) + \mathrm{KL}\left(\pi_{\theta^\star}|\mu\right) \geq \mathrm{KL}\left(\pi_{\theta^\star}|\mu\right) .$$

If π is not absolutely continuous with respect to μ then $\mathrm{KL}\left(\pi|\mu\right) = +\infty$. Therefore we have that $\pi_{\theta^\star}$ solves (P).

Proof of Proposition 5.1.5

We preface the proof with the following lemma

Lemma 5.1.9. *Let $h : \mathbb{R}^p \to \mathbb{R}$ be a convex function such that h is ray-coercive, i.e. for any $\theta \in \mathbb{R}^p$, with $\|\theta\| = 1$ we have $\lim_{t\to+\infty} h(t\theta) = +\infty$. Then h is coercive, i.e. $\lim_{\|\theta\|\to+\infty} h(\theta) = +\infty$.*

Proof. Assume that h is not coercive. Then there exists a sequence $(\theta_n)_{n\in\mathbb{N}} \in (\mathbb{R}^d \setminus \{0\})^{\mathbb{N}}$ such that $\lim_{n\to+\infty} \|\theta_n\| = +\infty$ and the sequence $(h(\theta_n))_{n\in\mathbb{N}}$ is bounded. Upon extraction we can assume that $\lim_{n\to+\infty} \theta_n / \|\theta_n\| = \tilde{\theta}$. We have for any $t \in \mathbb{R}$

$$h(t\tilde{\theta}) = h(t\tilde{\theta}) - h(t\theta_n/\|\theta_n\|) + h(t\theta_n/\|\theta_n\|) \; . \tag{5.13}$$

Let $t \geq 0$ and $\varepsilon > 0$. Since h is convex, h is continuous and there exists $n_0 \in \mathbb{N}$ such that for any $n \in \mathbb{N}$ with $n \geq n_0$,

$$\left\| h(t\tilde{\theta}) - h(t\theta_n/\|\theta_n\|) \right\| \leq \varepsilon \; , \qquad h(t\theta_n/\|\theta_n\|) \leq t/\|\theta_n\| \, h(\theta_n) + (1 - t/\|\theta_n\|) h(0) \; . \tag{5.14}$$

Combining (5.13) and (5.14) we obtain that for any $t \geq 0$,

$$h(t\tilde{\theta}) \leq \varepsilon + \|h(0)\| + \sup_{n\in\mathbb{N}} \|h(\theta_n)\| < +\infty \; ,$$

which is absurd. Hence, h is coercive. □

We now turn to the proof of Proposition 5.1.5. We divide the proof in two parts.

(a) Using that $\Theta_F = \mathbb{R}^p$ and the first part of Proposition 5.1.4 we have that $L : \Theta_F \to \mathbb{R}$ is continuously differentiable over $\mathbb{R}^p$. In addition, for any $\theta_1, \theta_2 \in \mathbb{R}^p$ and $\theta \in (0,1)$ we have

$$\int_{\mathbb{R}^d} \exp\left[-\langle \theta\theta_1 + (1-\theta)\theta_2, F(x)\rangle\right] \mathrm{d}\mu(x) \\ = \int_{\mathbb{R}^d} \exp\left[-\theta\langle \theta_1, F(x)\rangle\right] \exp\left[(1-\theta)\langle \theta_2, F(x)\rangle\right] \mathrm{d}\mu(x) \; .$$

Applying Hölder's inequality we get that

$$L(\theta\theta_1 + (1-\theta)\theta_2) = \log\left\{ \int_{\mathbb{R}^d} \exp\left[-\langle \theta\theta_1 + (1-\theta)\theta_2, F(x)\rangle\right] \mathrm{d}\mu(x) \right\} \\ \leq \theta \log\left\{ \int_{\mathbb{R}^d} \exp\left[-\langle \theta_1, F(x)\rangle\right] \mathrm{d}\mu(x) \right\} + (1-\theta) \log\left\{ \int_{\mathbb{R}^d} \exp\left[\langle \theta_2, F(x)\rangle\right] \mathrm{d}\mu(x) \right\} \; ,$$

hence L is convex. Using the monotone convergence theorem we have that for any $\theta \in \mathbb{R}^p$ with $\|\theta\| = 1$,

$$\lim_{t\to+\infty} L(t\theta) = \lim_{t\to+\infty} \log\left\{ \int_{\mathbb{R}^d} \exp\left[-t\langle \theta, F(x)\rangle\right] \mathrm{d}\mu(x) \right\} \\ \geq \lim_{t\to+\infty} \log\left\{ \int_{\{x\in\mathbb{R}^d \, : \, \langle \theta, F(x)\rangle < 0\}} \exp\left[-t\langle \theta, F(x)\rangle\right] \mathrm{d}\mu(x) \right\} = \infty \; ,$$

which implies that L is ray-coercive. Combining this result, the fact that L is convex and Lemma 5.1.9 we get that L is coercive. Since L is continuous and coercive it admits a minimizer $\theta^\star$ and therefore $\nabla L(\theta^\star) = 0$. Applying the second part of Proposition 5.1.4 concludes the first part of the proof.

(b) Let $x \in F^{-1}(\{0\})$ such that $\det(\mathrm{D}F(x)\mathrm{D}F(x)^\top) > 0$. We obtain that $\mathrm{Ker}(\mathrm{D}F(x)^\top) = \mathrm{Ker}(\mathrm{D}F(x)\mathrm{D}F(x)^\top) = \{0\}$. Hence, $\mathrm{rank}(\mathrm{d}F(x)) = \mathrm{rank}(\mathrm{d}F(x)^\top) = p$ and $\mathrm{d}F(x)$ is surjective. Using the submersion theorem, there exists $G : \mathsf{U} \to \mathbb{R}^d$ with U an open neighborhood of $0 \in \mathbb{R}^p$ such that for any $\zeta \in \mathsf{U}$, $F(G(\zeta)) = \zeta$. Therefore, for any $\theta \in \mathbb{R}^p$ with $\|\theta\| = 1$ there exists ζ_θ such that $\langle \theta, F(G(\zeta_\theta))\rangle = \langle \theta, \zeta_\theta\rangle < 0$. Hence, since F is continuous, there exists an open set V in $\mathbb{R}^d$ such that for any $y \in \mathsf{V}$, $\langle F(y), \theta\rangle < 0$. Combining this result with the fact that for any A open and $\mathsf{A} \neq \emptyset$, $\mu(\mathsf{A}) > 0$ we conclude the proof.

Proof of Proposition 5.1.6

Let $\theta \in \mathbb{R}^p$ such that $\|\theta\| = 1$ and such that $\mu(\{x \in \mathbb{R}^d \, : \, \langle F(x), \theta\rangle \leq 0\}) = 0$. Then, we have using the dominated convergence theorem

$$\lim_{t\to+\infty} \int_{\mathbb{R}^d} \exp[-t\langle \theta, F(x)\rangle]\mathrm{d}\mu(x) = \lim_{t\to+\infty} \int_{\{x\in\mathbb{R}^d \, : \, \langle F(x),\theta\rangle>0\}} \exp[-t\langle \theta, F(x)\rangle]\mathrm{d}\mu(x) = 0 \; . \quad (5.15)$$

Recall that **A**1(α) and **A**2(α') imply that $\Theta_F = \mathbb{R}^p$. Therefore, using (5.15), we have $v(\mathrm{Q}) = -\inf_{\theta\in\mathbb{R}^p} L(\theta) = +\infty$. Since $v(\mathrm{P}) \geq v(\mathrm{Q})$, see (5.2), we have $v(\mathrm{P}) = +\infty$. Hence, for any $\pi \in \mathcal{P}_\alpha^F$, $\mathrm{KL}\,(\pi|\mu) = +\infty$ and any $\pi \in \mathcal{P}_\alpha^F$ solves (P). We conclude upon remarking that $\delta_{x_0} \in \mathcal{P}_\alpha^F$.

Proof of Proposition 5.1.7

First we recall that for any $j \in \{1, \ldots, M\}$, we consider $A_{j,+} \in \mathbb{A}_{c_j\times n_j, c_{j-1}\times n_{j-1}}(\mathbb{R})$ defined by

$$A_{j,+} = D_j(x_0)A_j \; , \quad (5.16)$$

with $D_j(x_0) \in \mathbb{R}^{c_j\times n_j}\times\mathbb{R}^{c_j\times n_j}$ a diagonal matrix such that for any $i \in \{1, \ldots, c_j \times n_j\}$, $D_j(x_0)(i,i) = \mathbb{1}_{(0,+\infty)}(\mathscr{G}_j(x_0)(i))$. Note that we have for any $j \in \{1, \ldots, M\}$, $\tilde{A}_{j,+} = D_j(x_0)\tilde{A}_j$.

We start to show that for any $j \in \{1, \ldots, M\}$ there exists $\varepsilon_j > 0$ such that for any $x \in \bar{\mathrm{B}}(x_0, \varepsilon_j)$

$$\mathscr{G}_j(x) = A_{j,+} \ldots A_{1,+}(x) \; . \quad (5.17)$$

Namely, for any $j \in \{1, \ldots, M\}$, $\mathscr{G}_j$ is locally affine around x_0. First, for any $k \in \{1, \ldots, c_1 \times n_1\}$, either (i) $x_0 \in \mathrm{int}(\mathrm{Ker}(e_{k,1}^\top A_1))$ or (ii) $e_{k,1}^\top A_1 x_0 \neq 0$, where $(e_{k,1})_{k\in\{1,\ldots,c_1\times n_1\}}$ is the canonical basis of $\mathbb{R}^{c_1\times n_1}$. Let $k \in \{1, \ldots, n_1 \times c_1\}$. If (i) holds then there exists $\varepsilon_{1,k} > 0$ such that for any $x \in \mathrm{B}(x_0, \varepsilon_{1,k})$, $\mathscr{G}_1(x_0)(k) = \mathscr{G}_1(x)(k) = 0$. If (ii) holds $A_1(x_0)(k) > 0$ or $A_1(x_0)(k) < 0$. Therefore, since A_1 is continuous, there exists $\varepsilon_{1,k} > 0$ such that for any $x \in \bar{\mathrm{B}}(x_0, \varepsilon_{1,k})$, $A_1(x)(k)A_1(x_0)(k) > 0$. Combining the two previous cases, there exists $\varepsilon_1 > 0$ such that for any $x \in \mathrm{B}(x_0, \varepsilon_1)$ and $k \in \{1, \ldots, n_1 \times c_1\}$, $\mathrm{sign}(\mathscr{G}_1(x)(k)) = \mathrm{sign}(\mathscr{G}_1(x_0)(k))$. Hence, we have for any $x \in \mathrm{B}(x_0, \varepsilon_1)$,

$$\mathscr{G}_1(x) = \varphi(A_1(x)) = D_1(x)A_1(x) = D_1(x_0)A_1(x) \; ,$$

where $D_{\ell+1}$ is given in (5.16). Now assume that (5.17) is true for $j \in \{1, \ldots, \ell\}$ with $\ell \in \{1, \ldots, M-1\}$. There exists $\varepsilon_\ell > 0$ such that for any $x \in \bar{\mathrm{B}}(x_0, \varepsilon_\ell)$

$$\mathscr{G}_{\ell+1}(x) = \varphi(A_{\ell+1}\mathscr{G}_\ell(x)) = \varphi(A_{\ell+1}A_{\ell,+} \ldots A_{1,+}(x)) \; .$$

Then, for any $k \in \{1, \ldots, c_{\ell+1}\times n_{\ell+1}\}$, either (i) $x_0 \in \mathrm{int}(\mathrm{Ker}(e_{k,\ell+1}^\top A_{\ell+1}A_{\ell,+} \ldots A_{1,+}))$ or (ii) $e_{k,\ell+1}^\top A_{\ell+1}A_{\ell,+} \ldots A_{1,-}$, where $(e_{k,\ell+1})_{k\in\{\mathbb{1},\ldots,c_1\times n_1\}}$ is the canonical basis of $\mathbb{R}^{c_{\ell+1}\times n_{\ell+1}}$. Let $k \in \{1, \ldots, n_{\ell+1} \times c_{\ell+1}\}$. If (i) holds then there exists $\varepsilon_{\ell+1,k} > 0$ such that for any $x \in \mathrm{B}(x_0, \varepsilon_{\ell+1,k})$, $\mathscr{G}_{\ell+1}(x_0)(k) = \mathscr{G}_{\ell+1}(x)(k) =$

0. If (ii) holds $A_{\ell+1}A_{\ell,+}\dots A_{1,+}(x_0)(k) > 0$ or $A_{\ell+1}A_{\ell,+}\dots A_{1,+}(x_0)(k) < 0$. Therefore, since $A_{\ell+1}A_{\ell,+}\dots A_{1,+}$ is continuous, there exists $\varepsilon_{\ell+1,k} > 0$ such that for any $x \in \bar{\mathrm{B}}(x_0, \varepsilon_{\ell+1,k})$,

$$A_{\ell+1}A_{\ell,+}\dots A_{1,+}(x)(k)A_{\ell+1}A_{\ell,+}\dots A_{1,+}(x_0)(k) > 0\,.$$

Combining the two previous cases, there exists $\varepsilon_{\ell+1} > 0$ such that for any $x \in \mathrm{B}(x_0, \varepsilon_1)$ and $k \in \{1, \dots, n_{\ell+1} \times c_{\ell+1}\}$, $\mathrm{sign}(\mathscr{G}_{\ell+1}(x)(k)) = \mathrm{sign}(\mathscr{G}_{\ell+1}(x_0)(k))$. Hence, we have for any $x \in \mathrm{B}(x_0, \varepsilon_{\ell+1})$

$$\begin{aligned}\mathscr{G}_{\ell+1}(x) = \varphi(A_{\ell+1}A_{\ell,+}\dots A_{1,+}(x)) &= D_{\ell+1}(x)A_{\ell+1}A_{\ell,+}\dots A_{1,+}(x)\\ &= D_{\ell+1}(x_0)A_{\ell+1}A_{\ell,+}\dots A_{1,+}(x) = A_{\ell+1,+}A_{\ell,+}\dots A_{1,+}(x)\,,\end{aligned}$$

where $D_{\ell+1}$ is given in (5.16), which concludes the recursion. Let $\theta \in \mathbb{R}^p$ with $\|\theta\| = 1$. We have for any $x \in \bar{\mathrm{B}}(x_0, \varepsilon_M)$

$$\begin{aligned}\langle \theta, F(x)\rangle &= \sum_{j\in\mathbf{j}}\sum_{k=1}^{c_j}\theta_{j,k}\tilde{v}_{j,k}^\top(\mathscr{G}_j(x) - \mathscr{G}_j(x_0))\\ &= \sum_{j\in\mathbf{j}}\sum_{k=1}^{c_j}\theta_{j,k}\tilde{v}_{j,k}^\top\{A_{j,+}\dots A_{1,+}(x) - A_{j,+}\dots A_{1,+}(x_0)\}\\ &= \sum_{j\in\mathbf{j}}\sum_{k=1}^{c_j}\theta_{j,k}\tilde{v}_{j,k}^\top\tilde{A}_{j,+}\dots\tilde{A}_{1,+}(x-x_0) = \left\langle\sum_{j\in\mathbf{j}}\sum_{k=1}^{c_j}\theta_{j,k}v_{j,k}, x - x_0\right\rangle\,.\end{aligned}$$

Since, $(v_{j,c})_{j\in\mathbf{j},c\in\{1,\dots,c_j\}}$ is assumed to be linearly independent we have that $v = \sum_{j\in\mathbf{j}}\sum_{c=1}^{c_j}\theta_{j,c}v_{j,c}$ is non zero and therefore setting $x = x_0 - \varepsilon_M v/\|v\|$ we get that $\langle\theta, F(x)\rangle < 0$. Since F is continuous and $\mu(\mathsf{A}) > 0$ for every non-empty open set A we have that for $\theta \in \mathbb{R}^p$ with $\|\theta\| = 1$, $\mu(\{x \in \mathbb{R}^d : \langle F(x), \theta\rangle < 0\}) > 0$, which concludes the proof upon using Proposition 5.1.5-(a).

5.2 A texture synthesis algorithm

5.2.1 A brief history of texture synthesis

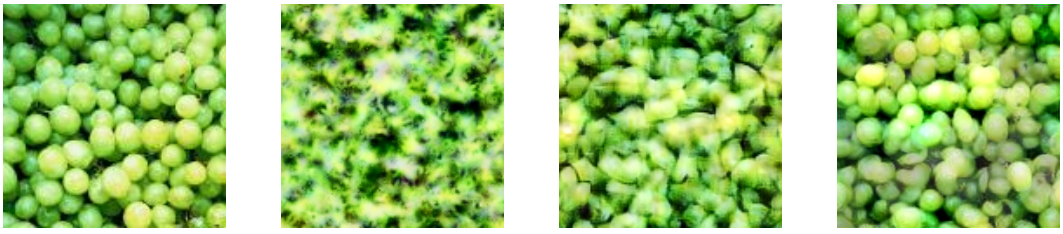
We recall that the aim of exemplar-based texture synthesis is, given a texture image x_0, to find a way to synthesize new images which look like x_0 but are not verbatim copies of x_0. Henceforth any algorithm tackling must met two requirements: structure (or geometrical) constraints and innovation constraints.

There have been two main approaches to deal with this problem: the *parametric* (or model-based) approach and the *non-parametric* (or patch-based) approach, see [Raa+17] for an extensive review of these methods. We highlight that most of these algorithms do not meet the requirements of computer graphics efficiency (procedural and parallel synthesis [Lag+10]). Instead, they aim at reproducing a large class of complex texture images.

Non parametric methods use the structural and perceptual information contained in the patches of the original image x_0, see [Wei+09]. In most cases, the patches of x_0 are extracted an rearranged. This rearrangement is often random so that the innovation constraint is met. In [EL99; WL00; Ash01; Har01] each pixel is updated using its local information. At each iteration a pixel is selected in the unknown region of the synthesized image. Its new value is estimated using the information contained in the nearby pixels which have already been synthesized. This procedure has then been extended to patch updates [Lia+01; EF01; Kwa+03] allowing for fast implementations of these algorithms. Under the

assumption that x_0 is a realization of some random field, the synthesizing process described in [EF01] can be interpreted as some bootstrap sampling scheme [LB06]. In this case the patch marginal of the image are asymptotically correct as the size of the images grow towards infinity. It is also of interest to consider the distribution of the patches in the example image x_0. In [GLR18] the authors approximate the optimal transport between a Gaussian distribution of patches and the empirical distribution of the patches in the target image x_0. The image is then obtained using patch aggregation [SDDB19]. By reformulating the texture synthesis problem as an optimization problem [Kwa+05] proposes an image model based on patch information.

This later model draws a link between non-parametric and parametric methods as the parametric ones aim at producing samples from a distribution which is given by x_0 and some spatial constraints. These models allow for more innovation than their non-parametric counterparts. In addition, since the distribution the images we synthesize using non-parametric methods is unknown it is not possible to infer or impose some structural or spatial properties. Fractional Brownian motions [FFC82], random functions [Per85] and reaction-diffusion [Tur91; WK91] were among the first attempts at modeling textures. Using spot-noise methods which rely on point processes [Wij91] provides an algorithm which can sample accurately a large class of microtextures. As emphasized in the previous section, when the number of point in the point process goes to infinity [GGM11] identify a Gaussian limiting process and study its properties from an image processing point of view. Any Gaussian random field is fully determined by its first and second moments and therefore by its first and second order statistics as defined by Diaconis and Freedman [DF81]. Since [JGV78; DF81] provided counterexamples to the Julesz conjecture [Jul62; Jul81] it is known that there exist textures with identical first and second order statistics which can be visually discriminated. Therefore it is necessary to go beyond the Gaussian model in order to be able to synthesize textures outside of the class of microtextures. [GM86; CM88; HB95; DB97; PS00; Pey10] proposed to sample from random fields associated with various statistical constraints. For instance, in the work of [PS00] the constraints are given by wavelet transforms. By replacing the wavelets used in [PS00] by convolutional neural networks, [GEB15] obtains striking visual results. Since then other methods based on pretrained convolutional neural networks have been developed [JAF16; UVL17; Uly+16]. In Figure 5.1 we show that the quality of the synthesis process highly depends on the choice of constraints.



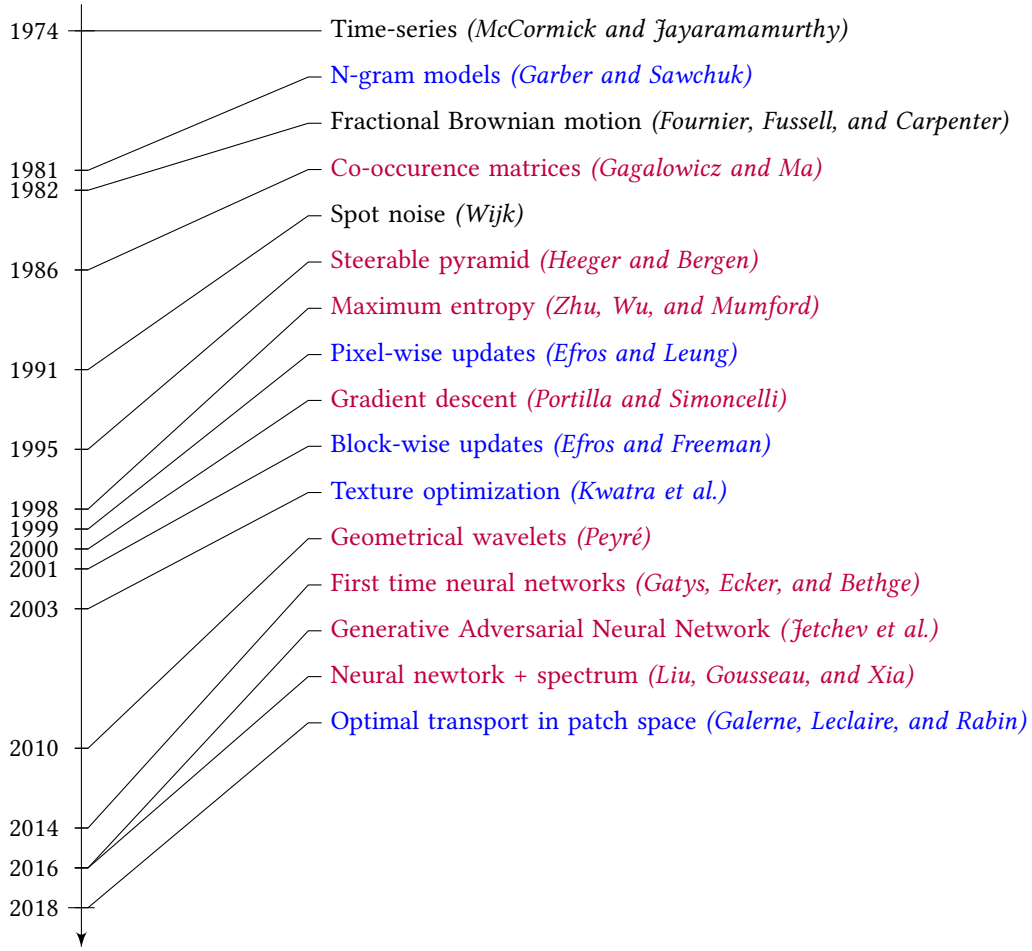
(a) (b) (c) (d)

Figure 5.1: In (a) we present the exemplar texture. In (b) we display the result we obtain using [HB95] pyramid synthesis. In (c) we show the result we obtain using [PS00] where the constraints are given by wavelet features. Finally in (d) we show the result of the method [GEB15]. Note that only this method, which is based on convolutional neural networks, is able to synthesize the complex structures observed in the image.

Even more recently, methods based on *Generative Adversarial Networks* (GAN) were used with promising results [JBV16; LW16a; Zho+18; BJV17]. However, this new algorithms cannot be cast as parametric methods since they do not sample from given distribution but rather learn a distribution

using x_0 and adversarial examples.

We gather some of advances in the domain exemplar-based texture synthesis in the following chronology. The non-parametric methods are given in **blue** whereas the parametric methods are given in **red**.



5.2.2 Sampling from macrocanonical models

In this section, our objective is twofold. First, we want to find a sequence $(\theta_n)_{n\in\mathbb{N}}$ which converges a.s. to $\theta^\star$, the solution of (Q). Second, we aim at sampling from the macrocanonical model $\pi_{\theta^\star}$ defined by (5.4). We present a Stochastic Approximation (SA) algorithm addressing simultaneously these two problems in Section 5.2.2. Our main result are summarized in Section 5.2.2. In Section 5.2.2 we draw a qualitative link between macrocanonical and microcanonical models.

The Stochastic Optimization via Unadjusted Langevin method

Stochastic approximation Let $\mathsf{K} \subset \mathrm{int}(\Theta_F)$ be a non-empty compact convex set such that $\mathsf{K} \cap \arg\min_{\Theta_F} L \neq \emptyset$ with L the log-partition given in (5.3). Since $\theta \mapsto L(\theta)$ is a convex mapping we

obtain that the sequence $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ defined by $\tilde{\theta}_0 \in \mathsf{K}$ and for any $n \in \mathbb{N}$, $\tilde{\theta}_{n+1} = \Pi_{\mathsf{K}}[\tilde{\theta}_n - \delta\nabla L(\tilde{\theta}_n)]$ where $\delta > 0$ is a stepsize and Π_{K} is the projection onto K, converges under mild assumptions to $\theta^\star \in \arg\min_{\Theta_F} L$, since L is convex, see [Nes04]. However, for any $\theta \in \Theta_F$, $\nabla L(\theta) = \pi_\theta(F)$ and evaluating this quantity is generally unfeasible. In what follows, we rely on Monte-Carlo approximations for which we derive explicit upper-bounds on the bias. More precisely, assuming that it is possible to sample from π_θ then $\nabla L(\theta)$ can be approximated by $N^{-1}\sum_{k=1}^N F(X_k)$, where $(X_k)_{k\in\{1,\dots,N\}}$ are independently sampled from π_θ. In most of our applications it is not feasible to sample directly from π_θ, but we can construct Markov chains for which π_θ is an invariant probability measure. Then, under assumptions and using classical Markov chain theory, it can be shown that $N^{-1}\sum_{k=1}^N F(X_k)$ converges a.s. to $\pi_\theta(F)$, [Dou+18, Theorem 11.3.1]. Such examples of Markov chains include the Metropolis-Hastings algorithm [Has70], which uses a rejection step. In a high-dimensional setting, the acceptance ratio can be extremely low and the proposed new iterate is then always discarded. Hence, we focus on Markov chains without rejection step. In this scenario, π_θ is not an invariant measure of the Markov chain in general. However, for an appropriate choice of Markov chain, the bias between its actual invariant probability measure and the target probability measure π_θ can be explicitly controlled.

SOUL algorithm First, we consider some regularity assumption on the measure μ with respect to the Lebesgue measure.

B1. *$\mu \ll \lambda$ and its Radon-Nikodym density w.r.t. to the Lebesgue measure is given for almost every $x \in \mathbb{R}^d$ by $\exp[-r(x)]/\int_{\mathbb{R}^d}\exp[-r(y)]\mathrm{d}y$, where $r : \mathbb{R}^d \to \mathbb{R}$ is measurable.*

Let $\theta = (\theta(0),\dots,\theta(p-1)) \in \mathbb{R}^p$ and consider the overdamped unadjusted Langevin algorithm, called ULA in [RT96], defined by $(\tilde{X}_n)_{n\in\mathbb{N}}$ with $\tilde{X}_0 = x \in \mathbb{R}^d$ and for any $n \in \mathbb{N}$

$$\tilde{X}_{n+1} = \tilde{X}_n - \gamma\left(\sum_{i=1}^p \theta(i)\nabla F_i(\tilde{X}_n) + \nabla r(\tilde{X}_n)\right) + \sqrt{2\gamma}\mathrm{Z}_{n+1}\ , \tag{5.18}$$

where $\gamma > 0$ is a stepsize and $(\mathrm{Z}_n)_{n\in\mathbb{N}}$ is a sequence of independent d-dimensional Gaussian random variables with zero mean and identity covariance matrix. This algorithm is the Euler-Maruyama discretization of the overdamped Langevin stochastic differential equation [Dur+17] for which π_θ is the invariant probability measure. The study of the geometric convergence of this Markov chain under various metrics was conducted in [DM17; Dur+17; Dal17b]. In Section 5.2.2 a SA scheme, the Stochastic Optimization with Unadjusted Langevin (SOUL) Algorithm, is proposed in order to construct a sequence $(\theta_n)_{n\in\mathbb{N}}$ such that $(\theta_n)_{n\in\mathbb{N}}$ converges a.s. and in L^1 to some $\theta^\star \in \arg\min_{\Theta_F\cap\mathsf{K}} L$. Let $\theta_0 \in \mathsf{K}$ and $X_0^0 \in \mathbb{R}^d$. For any $n \in \mathbb{N}$ and $k \in \{0,\dots,m_n-1\}$ we define

$$\begin{aligned} X_{k+1}^n &= X_k^n - \gamma_n\left(\sum_{i=1}^p \theta_n(i)\nabla F_i(X_k^n) + \nabla r(X_k^n)\right) + \sqrt{2\gamma_n}\mathrm{Z}_{k+1}^n \quad \text{and } X_0^n = X_{n-1}^{m_{n-1}}\ ; \\ \theta_{n+1} &= \Pi_{\mathsf{K}}\left[\theta_n + \delta_{n+1}m_n^{-1}\sum_{k=1}^{m_n} F(X_k^n)\right]\ , \end{aligned} \tag{5.19}$$

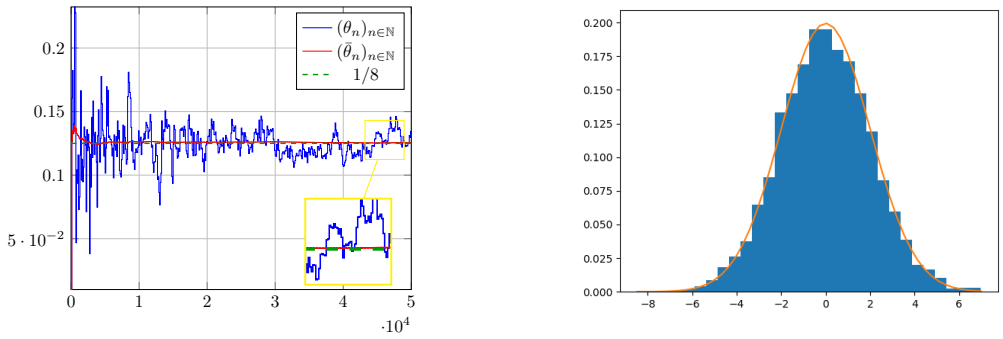
where $(\delta_n)_{n\in\mathbb{N}^\star}$ and $(\gamma_n)_{n\in\mathbb{N}}$ are sequence of positive stepsizes and the sequence $(\mathrm{Z}_k^n)_{n\in\mathbb{N},k\in\{1,\dots,m_n\}}$ is a sequence of independent d-dimensional Gaussian random variables with zero mean and covariance identity. By convention, $X_{-1}^{m_{-1}} = X_0^0$. The condition $X_0^n = X_{n-1}^{m_n-1}$ for all $n \in \mathbb{N}$ is referred to as a warm-start condition.

To illustrate the expected behavior of the proposed SOUL algorithm (5.19), we consider the toy example where $x_0 \in \mathbb{R}$, $F(x) = x^2 - 4$ and $F(x_0) = 0$. In this case the maximum entropy distribution

is given by the Gaussian distribution with zero mean and variance 4, see [MD10, Section 4.6.2]. One shows that the optimal weight $\theta^\star$ is given by $\theta^\star = 1/8$. In Figure 5.2, we experimentally check the convergence of $(\theta_n)_{n\in\mathbb{N}}$. We set $r(x) = 0$ and observe that the sequence $(\theta_n)_{n\in\mathbb{N}}$ as well as the sequence $(\bar{\theta}_n)_{n\in\mathbb{N}}$ defined for any $n \in \mathbb{N}$ by

$$\bar{\theta}_n = 0 \text{ if } n < N \ , \qquad \bar{\theta}_n = \sum_{k=N}^{n} \delta_k \theta_k \Big/ \sum_{k=N}^{n} \delta_k \ \text{ otherwise} \ , \tag{5.20}$$

where $N \in \mathbb{N}$ is a fixed parameter, converge to $\theta^\star$. We now state our main results on the dependency



(a) (b)

Figure 5.2: **Variance estimation** In (a), the sequence of parameters $(\theta_n)_{n\in\mathbb{N}}$ (blue curve) and the sequence of average parameters $(\bar{\theta}_n)_{n\in\mathbb{N}}$ (red curve) converge to the optimal value $\theta^\star = 1/8$. In (b) we illustrate empirically the convergence of the sequence $(X_0^n)_{n\in\mathbb{N}}$ to the Gaussian distribution with zero mean and variance 4 (orange curve) by plotting its histogram. In this example $N = 0$, $\delta_n = 0.1 \times n^{-0.7}$, $\gamma_n = 0.1 \times n^{-0.3}$ and $m_n = 10 \times \lceil n^{0.6} \rceil$.

on the dimension in the explicit error in SOUL.

Main results

In Section 5.2.2, the convergence of the sequence $(\theta_n)_{n\in\mathbb{N}}$ is studied under general assumptions. In this section, we complement these results in our setting. In particular, we show that the error bound in L^1 norm between $L(\theta_n)$ and $L(\theta^\star)$ is upper bounded by a constant which depends polynomially in the dimension d. Let $\alpha \geq 1$.

B2 (α). *There exists* $\mathsf{K} \subset \mathbb{R}^p$ *such that:*

(a) K *is a non-empty convex compact set with* $\mathsf{K} \subset \mathrm{int}(\Theta_F)$ *and we denote* $M_\Theta > 0$ *such that* $\mathsf{K} \subset \bar{\mathrm{B}}(0, M_\Theta)$ *;*

(b) F *is differentiable and there exists* $\mathtt{M} \geq 0$ *such that for any* $x, y \in \mathbb{R}^d$, $\|F(x) - F(y)\| \leq \mathtt{M}(1 + \|x\|^{\alpha-1} + \|y\|^{\alpha-1}) \|x - y\|$ *;*

(c) there exists $\theta^\star \in \mathrm{int}(\mathsf{K})$ *solution of* (Q).

Under **B**2(α) and **A**2(α) with $\alpha \geq 1$, a solution of (P) exists and is given by (5.4), see Proposition 5.1.4. In addition, L is differentiable on K and we show in Proposition 5.1.4 that $\nabla L \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R}^d)$, hence Lipschitz continuous over K with constant $\sup_{\theta\in\mathsf{K}} \left\|\nabla^2 L(\theta)\right\|$. Conditions under which **B**2(α)-(c) is satisfied are given in Proposition 5.1.5.

Condition **B**1 implies that the density of π_θ with respect to the Lebesgue measure, is given for any $x \in \mathbb{R}^d$ by $(\mathrm{d}\pi_\theta/\mathrm{d}\lambda)(x) = \exp[-U(\theta,x)]/\int_{\mathbb{R}^d}\exp[-U(\theta,y)]\mathrm{d}y$ with U defined for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$ by $U(\theta,x) = \langle\theta, F(x)\rangle + r(x)$. The mapping $U : \mathsf{K}\times\mathbb{R}^d \to \mathbb{R}$ is referred to as the potential function. Consider the following assumption on U.

B3. *There exist $U_i : \mathsf{K}\times\mathbb{R}^d \to \mathbb{R}$ with $i \in \{1,2\}$ such that for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$ $U(\theta,x) = U_1(\theta,x) + U_2(\theta,x)$. In addition,*

(a) there exists $\mathtt{L} \geq 0$ such that for any $i \in \{1,2\}$, $x \mapsto U_i(\theta,x)$ is continuously differentiable and for any $x, y \in \mathbb{R}^d$, $\|\nabla_x U_i(\theta,x) - \nabla_x U_i(\theta,y)\| \leq \mathtt{L}\,\|x-y\|$;

(b) there exists $\mathtt{m}_1 > 0$ and $x^\star \in \mathbb{R}^d$ such that for any $\theta \in \mathsf{K}$, $U_1(\theta,\cdot)$ is $\mathtt{m}_1$-strongly convex and $x^\star \in \arg\min_{x\in\mathbb{R}^d} U_1(\theta,x)$;

(c) there exists $\mathcal{M} \geq 0$ such that for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$, $\|\nabla_x U_2(\theta,x)\| \leq \mathcal{M}$;

We can relax the assumption that for any $\theta \in \mathsf{K}$, $x^\star \in \arg\min_{x\in\mathbb{R}^d} U_1(\theta,x)$ by the following: there exists $R \geq 0$ such that for any $\theta \in \mathsf{K}$, there exists $x_\theta^\star \in \arg\min_{x\in\mathbb{R}^d} U_1(\theta,x)$ and $x_\theta^\star \in \bar{\mathrm{B}}(0,R)$. But for the sake of simplicity we do not consider this assumption. The general assumption **B**3 is satisfied for both the Gaussian features and the CNN features introduced in Section 5.1.3. Indeed, if the features are Gaussian and the reference measure is Gaussian we recall that $\Theta_F = \mathcal{F}^{-1}[\mathfrak{R}^{-1}\{\left(-(2\sigma^2)^{-1},+\infty\right)^d\}] \cap \mathbb{R}^d$ containing $\theta^\star$ with $\theta^\star$ given in (5.6). Then, $x \mapsto U(\theta,x)$ is a definite positive quadratic form associated with the symmetric matrix $\mathbf{C}_\theta$, see (5.5). Setting $\mathtt{L}$ and $\mathtt{m}$ respectively the largest and lowest eigenvalues of $\mathbf{C}_\theta$ over K we obtain that **B**3 is satisfied with $U_1 = U$ and $U_2 = 0$.

In the case of CNN features, if the reference measure is a Gaussian distribution with zero mean and invertible covariance matrix $\mathbf{C}$, we obtain that for any $\theta \in \mathbb{R}^p$ and $x \in \mathbb{R}^d$, $U(\theta,x) = \langle\theta, F(x)\rangle + x^\top\mathbf{C}^{-1}x/2$. If in addition, φ is differentiable with Lipschitz derivative and for any $t \in \mathbb{R}$, $\sup_{t\in\mathbb{R}}|\varphi'(t)| < +\infty$, we have that **B**3 is satisfied with for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$, $U_1(\theta,x) = x^\top\mathbf{C}^{-1}x/2$ and $U_2(\theta,x) = \langle\theta, F(x)\rangle$. In particular the fact that U_2 is gradient-Lipschitz and Lipschitz is ensured by a straightforward recursion since for any $f \in \mathrm{C}^1(\mathbb{R}^{d_3},\mathbb{R}^{d_2})$ and $g \in \mathrm{C}^1(\mathbb{R}^{d_2},\mathbb{R}^{d_1})$, $x \mapsto \mathrm{d}(g\circ f)(x)$ and $g\circ f$ Lipschitz if $f, g, \mathrm{d}f$ and $\mathrm{d}g$ are Lipschitz. Note that the differentiability assumption is not met in classical convolutional neural networks such as `VGG19` . Therefore, in all of our experiments we replace the max-pooling operator by a mean-pooling operator and the RELU function by a Continuously Differentiable Exponential Linear Unit (CELU), see [Bar17]. We now state our main results in the case where U is a strongly convex potential, *i.e.* $U_2 = 0$.

Theorem 5.2.1. *Let $\alpha \geq 1$. Assume* **A**2(α), **B**1, **B**2(α), **B**3 *with $U_2 = 0$. Let $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ a sequence of positive integers satisfying $\delta_n < 1/(\sup_{\theta\in\mathsf{K}}\|\nabla^2 L(\theta)\|)$ and $\gamma_n < \min(\mathtt{k}_1/(2\mathtt{L}^2), 1/2)$ for any $n \in \mathbb{N}$. Then, there exists $(E_n)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}^\star$*

$$\mathbb{E}\left[\left\{\sum_{k=1}^n \delta_k L(\theta_k)\Big/\sum_{k=1}^n \delta_k\right\} - \min_{\mathsf{K}} L\right] \leq E_n\Big/\left(\sum_{k=1}^n \delta_k\right) \;,$$

with for any $n \in \mathbb{N}^\star$,

(a) if $m_n = m_0$ for all $n \in \mathbb{N}$ and $\sup_{n\in\mathbb{N}}|\delta_{n+1}-\delta_n|\,\delta_n^{-2} < +\infty$

$$E_n = C_1(1+d^{\varpi})\left(1 + \sum_{k=0}^{n-1}\delta_{k+1}\gamma_k^{1/2} + \sum_{k=0}^{n-1}\delta_{k+1}\gamma_{k+1}^{-5/2}(\gamma_k-\gamma_{k+1})^{1/2} + \sum_{k=0}^{n-1}\delta_{k+1}^2/\gamma_k^{3/2} + \delta_n/\gamma_n\right) \;;$$

(b) otherwise

$$E_n = C_2(1+d^{\varpi})\left(1+\sum_{k=0}^{n-1}\delta_{k+1}\gamma_k^{1/2}+\sum_{k=0}^{n-1}\delta_{k+1}/(m_k\gamma_k)+\sum_{k=0}^{n-1}\delta_{k+1}^2\gamma_k+\sum_{k=0}^{n-1}\delta_{k+1}^2/(m_k\gamma_k)^2\right) ,$$

with $C_1, C_2, \varpi \geq 0$ which do not depend on the dimension d.

Proof. The proof is postponed to Section 5.2.4. □

In the case where for any $n \in \mathbb{N}$, $m_n = m_0$, $\gamma_n = \gamma_0$ and $\lim_{n\to+\infty}\delta_n = 0$ with $\sum_{k=0}^{+\infty}\delta_n = +\infty$, we obtain using [PS98, Problem 80, Part I] that, $\lim_{n\to+\infty}\sum_{k=0}^{n}\delta_k^2/\sum_{k=0}^{n}\delta_k = 0$. Therefore, using Theorem 5.2.1-(a) we get that

$$\limsup_{n\to+\infty}\mathbb{E}\left[L(\bar{\theta}_n)\right] - \min_{\mathsf{K}} L \leq C_1(1+d^{\varpi})\gamma_0^{1/2} .$$

We now state our main results in the case where the potential is not convex anymore. We consider the following additional regularity assumption on F.

B4. *$F \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R}^p)$ and there exists $\mathtt{B} \geq 0$ such that for any $x, y \in \mathbb{R}^d$, $\|\mathrm{d}F(x) - \mathrm{d}F(y)\| \leq \mathtt{B}\|x-y\|$.*

Theorem 5.2.2. *Assume* **A**2(1), **B**1, **B**2(1), **B**3 *and* **B**5. *Let $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ a sequence of positive integers satisfying $\delta_n < 1/(\sup_{\theta\in\mathsf{K}}\|\nabla^2 L(\theta)\|)$ and $\gamma_n < \min(\mathtt{m}_1/(8\mathtt{L}^2), 1/2)$ for any $n \in \mathbb{N}$. Then, there exists $(E_n)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}^\star$,*

$$\mathbb{E}\left[\left\{\sum_{k=1}^{n}\delta_k L(\theta_k)\right\}\Big/\sum_{k=1}^{n}\delta_k\right] - \min_{\mathsf{K}} L \leq E_n \Big/ \left(\sum_{k=1}^{n}\delta_k\right) ,$$

with for any $n \in \mathbb{N}^\star$,

(a) if $m_n = m_0$, $\gamma_n = \gamma_0$ for all $n \in \mathbb{N}$ and $\sup_{n\in\mathbb{N}}|\delta_{n+1}-\delta_n|\,\delta_n^{-2} < +\infty$

$$E_n = C_1(1+d^{\varpi})\left(1+\sum_{k=0}^{n-1}\delta_{k+1}\gamma_0^{1/2}+\sum_{k=0}^{n-1}\delta_{k+1}^2/\gamma_0+\delta_n/\gamma_0\right) ;$$

(b) else

$$E_n = C_2(1+d^{\varpi})\left(1+\sum_{k=0}^{n-1}\delta_{k+1}\gamma_k^{1/2}+\sum_{k=0}^{n-1}\delta_{k+1}/(m_k\gamma_k)+\sum_{k=0}^{n-1}\delta_{k+1}^2\right) ,$$

with $C_1, C_2, \varpi \geq 0$ which do not depend on the dimension d.

Proof. The proof is postponed to Section 5.2.4. □

The discussion conducted after Theorem 5.2.1 is still valid in this case. Theorem 5.2.2 follows from more general results derived in Theorem 5.2.10 and Theorem 5.2.14.

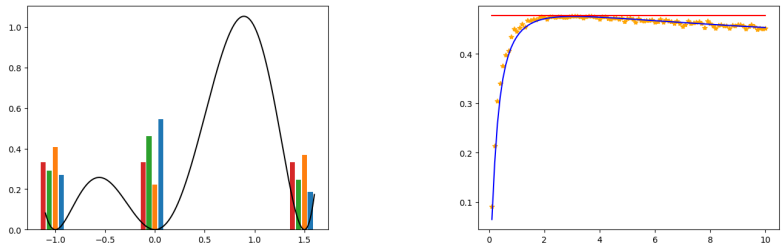


Figure 5.3: **Microcanonical sampling scheme** In this one-dimensional toy example the features are given by $F(x) = (x+1)^2x^2(x-1.5)^2$ (black curve). The microcanonical model associated with these features is the uniform distribution over $\{-1, 0, 1.5\}$ (red bars). In (a), we plot, the distribution $(\Phi_\infty)_\sharp(\nu_0)$ for different initial distributions ν_0, standard Gaussian (blue bars), uniform over $[-3, 3]$ (orange bars) and standard Cauchy (green bars). The distribution $(\Phi_\infty)_\sharp(\nu_0)$ is approximated by sampling 10^3 points according to ν_0 and performing the recursion associated with (5.21) for these points for 10^4 iterations. None of the initial distribution yields a distribution $(\Phi_\infty)_\sharp(\nu_0)$ which is the uniform distribution. Let ν_0 be a Gaussian distribution with zero mean and variance σ^2 with $\sigma > 0$. In (b), we show the dependency of the entropy of $(\Phi_\infty)_\sharp(\nu_0)$ with respect to σ^2 (orange points). The distribution $(\Phi_\infty)_\sharp(\nu_0)$ is approximated by sampling 10^3 points according to ν_0 and performing the recursion associated with (5.21) for these points for 10^3 iterations. Then, we compute its entropy and show that it is close to the one given by numerical integration (blue curve). We also plot the entropy upper-bound $\log_{10}(3)$ (red curve) given by the uniform distribution on $\{-1, 0, 1.5\}$.

Links with microcanonical models

In this section, we present qualitative results on the microcanonical model and the asymptotic behavior of the macrocanonical model for specific geometrical constraints.

Let $\nu_0 \ll \lambda$ be an initial probability measure. We consider the sequence of probability measures $(\nu_n)_{n\in\mathbb{N}}$ defined by the following recursion: for any $n \in \mathbb{N}$,

$$\nu_{n+1} = \Phi_\sharp(\nu_n) \ , \tag{5.21}$$

where $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ is defined for any $x \in \mathbb{R}^d$ by $\Phi(x) = x - \gamma \mathrm{d}F(x)^\top F(x)$, with $\gamma > 0$ a stepsize, *i.e.* for all $n \in \mathbb{N}$, ν_n is the pushforward measure of ν_0 by n steps of the gradient descent for the the loss function $x \mapsto \|F(x)\|^2$. Under some assumptions on F, [BM18, Theorem 3.7] implies that $\nu_\infty = \lim_n \nu_n$ exists and its support is $\mathsf{A}_0 = F^{-1}(\{0\})$.

If A_0 is compact, the microcanonical model, see Definition 5.1.1, associated with the reference measure λ and the constraints F, is given by the uniform distribution over A_0, denoted ν_{A_0}. If ν_∞ were the microcanonical model associated with F then we should have $\nu_\infty = \nu_{\mathsf{A}_0}$. However, as illustrated in Figure 5.3, ν_∞ strongly depends on the initial probability measure ν_0.

In the next result, we prove that considering specific constraint functions f_ε, there exists an explicit probability measure π_∞ such that its support is included in $F^{-1}(\{0\})$ and π_∞ is the limit of macrocanonical models associated with f_ε and some reference probability measure μ. Let $\varepsilon > 0$. We define $f_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ such that for any $x \in \mathbb{R}^d$, $f_\varepsilon(x) = \|F(x)\|^2 - \varepsilon$. We denote π_ε the macrocanonical model, see Definition 5.1.2, associated with f_ε when it exists.

Proposition 5.2.3. *Assume* **A**2(2) *and that for any non-empty open set* $\mathsf{A} \subset \mathbb{R}^d$, $\mu(\mathsf{A}) > 0$. *Let* F *be given by* (5.8), *assume that* $1 \in \mathbf{j}$ *and that there exists* $k \in \{1, \dots, c_1\}$ *such that for any* $x \in \mathbb{R}^d$ *with* $x \neq 0$

there exists $\ell \in \{1, \dots, n_1\}$ *with* $e_\ell^\top \tilde{A}_1^k x > 0$. *Then there exists* $\varepsilon_0 > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0)$, π_ε *exists. In addition, assume that* $\mu(F^{-1}(\{0\})) > 0$ *then* $\lim_{\varepsilon \to 0} \pi_\varepsilon = \pi_\infty$, *with for any* $x \in \mathbb{R}^d$, $(\mathrm{d}\pi_\infty / \mathrm{d}\mu)(x) = \mathbb{1}_{F^{-1}(\{0\})}(x) / \mu(F^{-1}(\{0\}))$

Proof. The proof is postponed to Section 5.2.4. □

Note that under other assumptions on $F^{-1}(\{0\})$ other explicit measures π_∞ are derived in Proposition 5.2.29.

5.2.3 Experiments

In this section, we assess the computational efficiency of SOUL algorithm (5.19) for texture synthesis. Variants of the original methodology are presented in Section 5.2.4.

Periodic Gaussian model

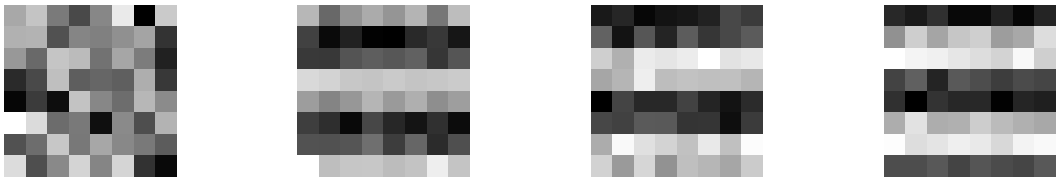
First, we consider the toy problem of periodic Gaussian texture synthesis, see Section 5.1.3 for details. Note that the extension of our findings to two dimensional signals is straightforward upon replacing the one dimensional Fourier transform by its two dimensional counterpart. We recall that the macrocanonical model is explicit and given by a measure $\pi_{\theta^\star}$ which is the probability distribution of $X = d^{-1/2}(x_0 * \mathbb{Z})$ where $\mathbb{Z}$ is a standard d-dimensional Gaussian random variable, see Section 5.1.3.

Empirical convergence We consider a 8×8 image, denoted x_0, corrupted by some noise, so that $\mathcal{F}(x_0)$ is non-zero everywhere on the 8×8 grid. The reference measure μ is a Gaussian distribution with zero mean and diagonal covariance matrix with diagonal coefficients given by σ^2. In this setting, $d = p = 64$ and using (5.6) we have $\theta^\star = \mathcal{F}^{-1}(d|\mathcal{F}(x_0)|^{-2} - \sigma^{-2})/2$. Using the spatial translation invariance property of $\pi_{\theta^\star}$, see Section 5.1.3, we identify four configurations which are equally likely to be sampled by $\pi_{\theta^\star}$, see Figure 5.4.

Figure 5.4: **Exemplar images and parameters** The exemplar image x_0 is shown in (a). Translated versions of this image, which are equally likely to be sampled by $\pi_{\theta^\star}$ are presented in (b), (c) and (d). The target parameter $\theta^\star$ is shown in (d).

The images $(X_0^n)_{n \in \mathbb{N}}$ generated by the SOUL algorithm (5.19) are approximate samples of $\pi_{\theta^\star}$ for n large enough. The configurations identified in Figure 5.4 are recovered during one run of the algorithm, see Figure 5.5. A video of the evolution of the sequence $(X_0^n)_{n \in \mathbb{N}}$ is available at https://vdeborto.github.io/publication/texture_soul/.

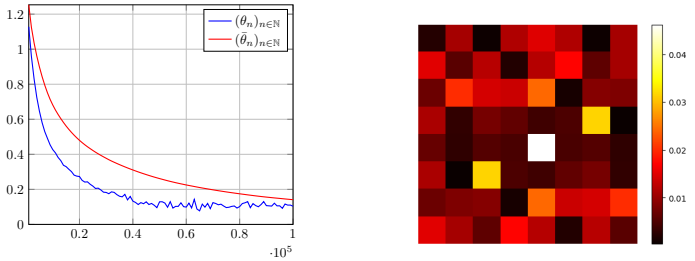
The main theoretical results in Theorem 5.2.1 deal with the error between $L(\bar{\theta}_n)_{n \in \mathbb{N}}$ and $\arg\min_{\theta \in \Theta} L(\theta)$, where $(\bar{\theta}_n)_{n \in \mathbb{N}}$ is given by (5.20). Selecting fixed parameters, $\gamma_n = 10^{-4}$, $\delta_n = 10^{-1}$ and $m_n = 1$, we observe the convergence of the sequence $(\theta_n)_{n \in \mathbb{N}}$ towards a biased estimate of $\theta^\star$. The Normalized Root



(a) $n = 0$ (b) $n = 58000$ (c) $n = 74000$ (d) $n = 100000$

Figure 5.5: **Sequence of images** The initialization (a) of the algorithm is some white noise, *i.e.* the realization of a standard Gaussian random variable on the 8×8 grid. We then show some selected samples (b)-(d) of the sequence generated with fixed parameters $\delta_n = 10^{-1}$, $\gamma_n = 10^{-4}$ and $m_n = 1$. Note that these samples are visually close to the ones presented in Figure 5.4.

Mean Square Error (NRMSE) defined for any $n \in \mathbb{N}$ by $\mathrm{N}(\theta_n) = \|\theta_n - \theta^\star\|_2 / \|\theta^\star\|_2$, is upper bounded by 0.2 for $n \geq 4 \times 10^4$, see Figure 5.6. In Figure 5.7, we show that this error level yields satisfactory parameters from a visual point of view. We highlight that $\mathcal{F}(\theta^\star)$ corresponds to the precision matrix (up to a constant factor) of the Gaussian model under study.



(a) (b)

Figure 5.6: **Convergence of the parameters** We recall that the parameters are initialized with $\theta_0 = 0$ and that $\gamma_n = 10^{-4}$, $\delta_n = 10^{-1}$ and $m_n = 1$. The NRMSE error in (a) rapidly decreases before oscillating around 0.1 (blue curve). A similar smoothed behaviour can be observed for the averaged sequence $(\bar{\theta}_n)_{n\in\mathbb{N}}$ (red curve). The heatmap of the NRMSE, *i.e.* a pixel $i \in \{0, \dots, 7\}^2$ in (b) corresponds to $(\theta_n(i) - \theta^\star(i))^2 / \|\theta^\star\|^2$.

The previous experiment suggests to set $\gamma_n = \gamma > 0$, $\delta_n = \delta > 0$ and $m_n = m \in \mathbb{N}^\star$, at least for a burnin period. When the behavior of the sequence $(\theta_n)_{n\in\mathbb{N}}$ becomes oscillatory, the setting can be changed in order to obtain a better approximation of $\theta^\star$. We investigate the long-time behavior after a burnin period of $N = 5 \times 10^4$ iterations with $m_n = 1$, $\gamma_n = 10^{-4}$ and $\delta_n = 10^{-1}$. After this period we set $m_n = \lceil \bar{n}^a \rceil$, $\gamma_n = \bar{n}^{-b}$ and $\delta_n = \bar{n}^{-c}$ with $\bar{n} = n - N + 1$, $a, b, c > 0$. We observe that the NRMSE error decreases from 0.1 to 0.06 for appropriate choices of rates, see Figure 5.8. Nevertheless, this improvement comes at a cost since the number of Markov chain iterations is no longer equal to the number of iterations n and grows as $\lceil \bar{n}^a \rceil$.

The previous comments along with Figure 5.8 suggest to set fix hyperparameters with $m_n = 1$ for all $n \in \mathbb{N}$. This is a good strategy to obtain acceptable approximations of the target parameter $\theta^\star$ in reasonable time. However, the sampled images move slowly between the acceptable configurations of Figure 5.4. Increasing the fixed batch size, *i.e.* increasing m_n, for instance setting $m_n = 10^2$ for all $n \in \mathbb{N}$ we obtain more innovation in the chain $(X_0^n)_{n\in\mathbb{N}}$. Indeed, for the same number of Markov chain iterations the chain $(X_0^n)_{n\in\mathbb{N}}$ visits more different acceptable configurations for $m_n = 10^2$ than for

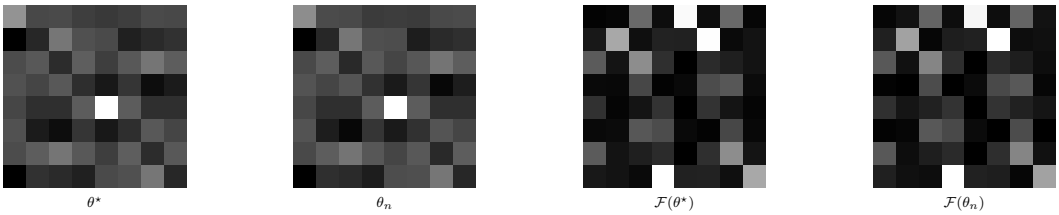


Figure 5.7: **Visual evaluation of parameters** We display the target parameters $\theta^\star$ and the parameters obtained after 10^6 iterations of the algorithm. Similarly we display the discrete Fourier transform of the target parameters $\mathcal{F}(\theta^\star)$ and the Fourier transform of the parameters after 10^6 iterations. There is no visual difference between $\theta^\star$ and θ_n.

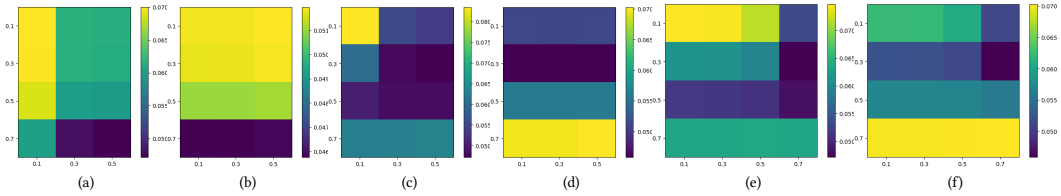


Figure 5.8: **Evolution of the error** In (a) and (b) we present the heatmap of the NRMSE error between $\theta_{2\times10^5}$ and $\theta^\star$ in (a) and $\bar{\theta}_{2\times10^5}$ and $\theta^\star$ in (b), given different values of $b, c > 0$ where $\gamma_n = 10^{-4} \times \bar{n}^{-b}$ and $m_n = \lceil \bar{n}^c \rceil$ with $\delta_n = 10^{-1} \times \bar{n}^{-0.3}$ and $\bar{n} = n - N + 1$ with $N = 5 \times 10^4$. On the y-axis in (a) and (b) we represent the different values for parameter b and on the x-axis the different values for parameter c. Similarly, in (c) and (d) we present the heatmap of the NRMSE error between $\theta_{2\times10^5}$ and $\theta^\star$ in (c) and $\bar{\theta}_{2\times10^5}$ and $\theta^\star$ in (d), given different values of $a, c > 0$ where $\delta_n = 10^{-1} \times \bar{n}^{-a}$ and $m_n = \lceil \bar{n}^c \rceil$ with $\gamma_n = 10^{-4} \times \bar{n}^{-0.7}$. On the y-axis in (c) and (d) we represent the different values for parameter a and on the x-axis the different values for parameter c. In (e) and (f) we present the heatmap of the NRMSE error between $\theta_{2\times10^5}$ and $\theta^\star$ in (e) and $\bar{\theta}_{2\times10^5}$ and $\theta^\star$ in (f), given different values of $b, c > 0$ where $\delta_n = 10^{-1}\bar{n}^{-a}$ and $\gamma_n = \lceil \bar{n}^{-b} \rceil$ with $m_n = \lceil \bar{n}^{0.5} \rceil$. On the y-axis in (e) and (f) we represent the different values for parameter a and on the x-axis the different values for parameter b. A plot of the NRMSE for the averaged sequence is presented in (g) with $a = 0.3$ and $c = 0.7$.

$m_n = 1$, see Figure 5.9. However, if $m_n = 10^2$, the NRMSE error of the sequence $(\theta_n)_{n\in\mathbb{N}}$ has a lower decrease rate than if $m_n = 1$.

Therefore, the hyperparameters of the algorithm should be adapted for the problem at hand. If we are interested in finding $\theta^\star$ then fixed settings for a burnin period followed by an eventual run of the algorithm with increasing batch size and decreasing stepsizes is recommended. However, if we are concerned with the innovation of the sequence $(X_0^n)_{n\in\mathbb{N}}$ then larger batch sizes, not necessarily increasing, are recommended.

Neural network features

Spatially averaged CNN features We now investigate the case where the features are given by a convolutional neural network, see Section 5.1.3. In our experiments we fix $\mathsf{K} = \left[-10^4, 10^4\right]^d$.

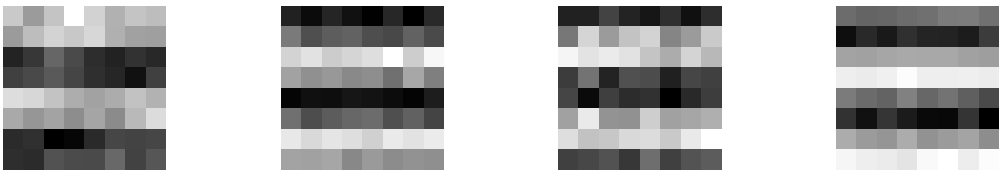


Figure 5.9: **Larger batch sizes improve visual quality** The algorithm with $\delta_n = 10^{-1}$, $\gamma_n = 10^{-4}$ and $m_n = 10^2$ produces more diverse samples than the ones obtained with the same algorithm and $m_n = 1$, see Figure 5.4. Note that the NRMSE errors $0.73, 0.63, 0.60$ and 0.47 are still high.

Model hyperparameters In the proposed model a few hyperparameters must be selected. First, a convolutional neural network architecture is to be chosen. We use the `VGG19` model since it has been highlighted by [Ust+16; GEB15] that such an architecture is well suited for the task of texture synthesis. In [GEB15] the neural network is pretrained on a classification task, see [SZ14]. The importance of the pretraining is highlighted in Section 5.2.4.

Another hyperparameter of the model is the set $\mathbf{j}$ of layers we consider to build our features, in (5.8). We consider three settings: (i) shallow network; (ii) deep network; (iii) full network. The structure of the network is given in Section 5.2.4. In (i) we set $\mathbf{j} = \{1, 3, 6, 8, 11, 13\}$, in (ii) we set $\mathbf{j} = \{15, 24, 26, 31\}$ and in (iii) we set $\mathbf{j} = \{1, 3, 6, 8, 11, 13, 15, 24, 26, 31\}$. Note that in the restricted models (i) and (ii) the dimension of the parameter space is reduced to $p = 896$ respectively $p = 1792$, whereas in the full model $p = 2688$. The influence of $\mathbf{j}$ is visually investigated in Figure 5.10. In what follows we consider the full CNN model given by (iii) in order to be able to synthesize a wide variety of texture images. In Section 5.2.4, it is shown that considering color statistics and neural network features improve the visual results. Finally, we assess in Section 5.2.4 that contrary to the algorithm proposed in [LZW16], our implementation can produce images of arbitrary dimensions from one input image.

Behavior of the parameter sequence We now study the behavior of the sequence $(\theta_n)_{n\in\mathbb{N}}$. In Figure 5.11 we present the evolution of $(\theta_n)_{n\in\mathbb{N}}$ for some layers in $\mathbf{j}$ and three channels for each layer. The sequence $(\theta_n)_{n\in\mathbb{N}}$ does not converge, even though we observe some stabilization of the averaged sequences. The reasons for the failure of the convergence are twofold. First, in all our settings we fix the hyperparameters as follows: $\delta_n = 10^{-3}$, $\gamma_n = 10^{-5}$ and $m_n = 1$ but run only 10^5 iterations. Considering a continuous Langevin dynamics, the images we observe correspond to a time $T = 10^5 \times \gamma_n = 1$ of the evolution. Increasing the stepsize γ_n is not an option since it yields diverging sequences of images. Second, the chain is slowly mixing and therefore it is hard to produce entirely different, yet visually coherent, samples with one run of SOUL.

It appears that the algorithm produces good visual results even though the parameter sequence is not stable. Increasing the number of Langevin iterations m_n generates images which are noisier but also increases the innovation of the algorithm, see Section 5.2.3. This is in accordance with Figure 5.9.

Comparison with existing methods In this section we compare the proposed algorithm with several examplar-based texture synthesis methods. We set $\delta_n = 10^{-3}$, $\gamma_n = 10^{-5}$ and $m_n = 1$. The algorithm is run for 10^4 iterations for each image. For each comparison we systematically include the results obtained with the methodology proposed in [GEB15], which is a microcanonical methodology using Gram matrices computed on neural network outputs as features.

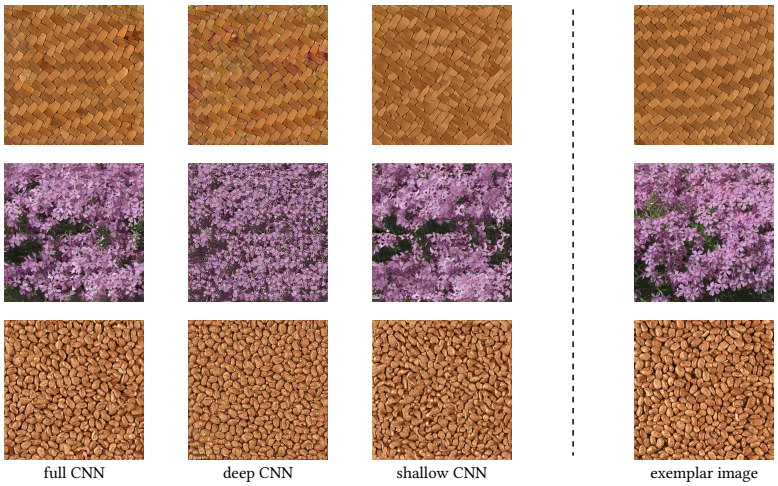


Figure 5.10: **Influence of j** As expected the best visual results of SOUL after 10^4 iterations are obtained with the full CNN setting. The local structure and some details (the petals of the flowers, the form of the beans) are lost when using the shallow CNN setting. On the other hand, using only the deep part of the CNN is not suitable for texture with strong low frequency components. For instance in the flower image, almost no grass is retrieved when using the deep CNN setting. The hyperparameters are fixed as follows: $\delta_n = 10^{-3}$, $\gamma_n = 10^{-5}$ and $m_n = 1$.

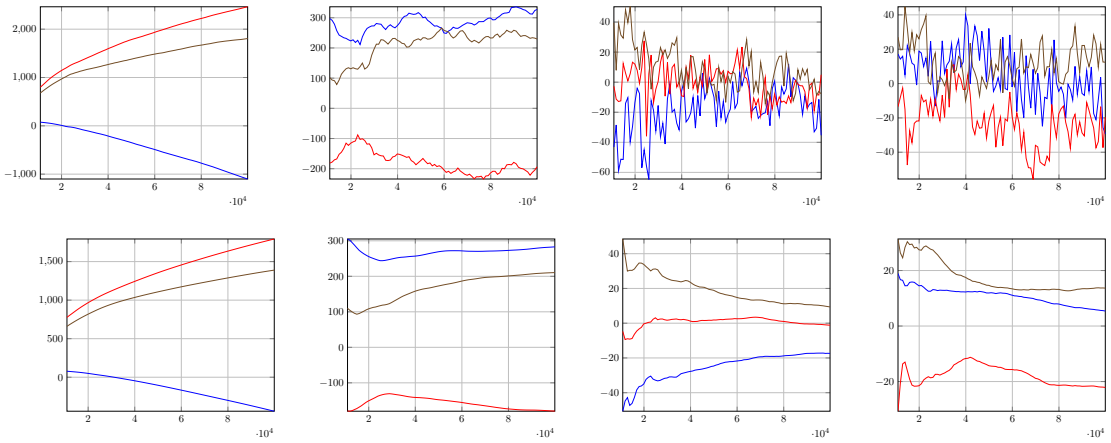


Figure 5.11: **Non convergence of the weights** For the layers corresponding to $j = 3, 8, 24$ and 31 we study, on three channels ($k = 10$, 20 and 30), the behavior of the sequence $(\theta_n(i_{k,j}))_{n\in\mathbb{N}}$ (first row) and the averaged sequence $(\bar{\theta}_n(i_{k,j}))_{n\in\mathbb{N}}$ (second row), where $i_{k,j}$ is the index corresponding to layer j and channel k. These sequences have not converged yet, although the averaged sequence seems to stabilize for some layers, *i.e.* some values of j.

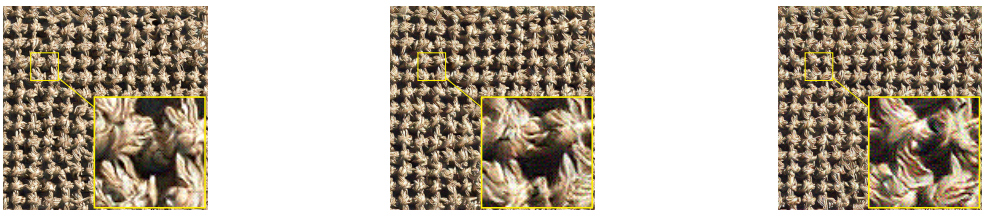


Figure 5.12: **Langevin iterations and noise** Running the SOUL algorithm with different m_n yield visually coherent images. However the higher m_n is, the noisier the images are. The other hyperparameters are fixed as follow: $\delta_n = 10^{-3}$ and $\gamma_n = 10^{-5}$.

First, we consider the Portilla-Simoncelli algorithm [PS00], see Section 5.2.3, which is a microcanonical based methodology and does not rely on neural network features, see Section 5.2.3. Our algorithm and the one from [GEB15] provide visually satisfying results, whereas the method from [PS00] fails to produce realistic images.

Second, compare our algorithm to the one of [LZW16] in Section 5.2.3. In [LZW16], the authors propose a similar macrocanonical methodology but do not consider more than one convolutional neural network layer to build their features.

In Section 5.2.4, we also test our algorithm on texture images which do not exhibit salient spatial structures. It was already noted in [Raa+17, Figure 26] that the generative model [JBV16] fails to produce high quality image in this case. On the other hand, our algorithm and the one from [GEB15] yield good visual results.

Another experiment on highly regular textures, comparing our algorithm with the ones of [LGX16] and [GGL19], is presented in Section 5.2.4.

Texture style transfer

We conclude this experimental part by considering other applications than texture synthesis and assess that the proposed algorithm can be used for the task of style transfer. Indeed given one content image x_{content}, a style image x_{style}, not necessarily of the same size, and $\mathbf{j}_{\text{content}} \subset \mathbf{j}$ we consider the same CNN feature as before but x_0 is replaced by x_{content} for $j \in \mathbf{j}_{\text{content}}$ in (5.8). In the rest of the neural network features, x_0 is replaced by x_{style} in (5.8), *i.e.* $F(x) = \left(\overline{\mathscr{G}}^k_j(x) - \overline{\mathscr{G}}^k_j(x_0^j)\right)_{j\in\mathbf{j},k\in\{1,\dots,c_j\}}$, with $x_0^j = x_{\text{content}}$ if $j \in \mathbf{j}_{\text{content}}$ and x_{style} otherwise. These new features are well-suited to perform a style transfer task as illustrated in Figure 5.15 with $\mathbf{j}_{\text{content}} = \{1, 3, 6, 8, 11\}$ and $\mathbf{j} = \{1, 3, 6, 8, 11, 13, 15, 24, 26, 31\}$.

5.2.4 Proofs and additional results

Proofs of Section 5.2.2

We start by introducing some notations. Let $V : \mathbb{R}^d \to [1, +\infty)$ be a measurable function. For $f \in \mathbb{F}(\mathbb{R}^d)$, the V-norm of f is given by $\|f\|_V = \|f/V\|_\infty$. Let ξ be a finite signed measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

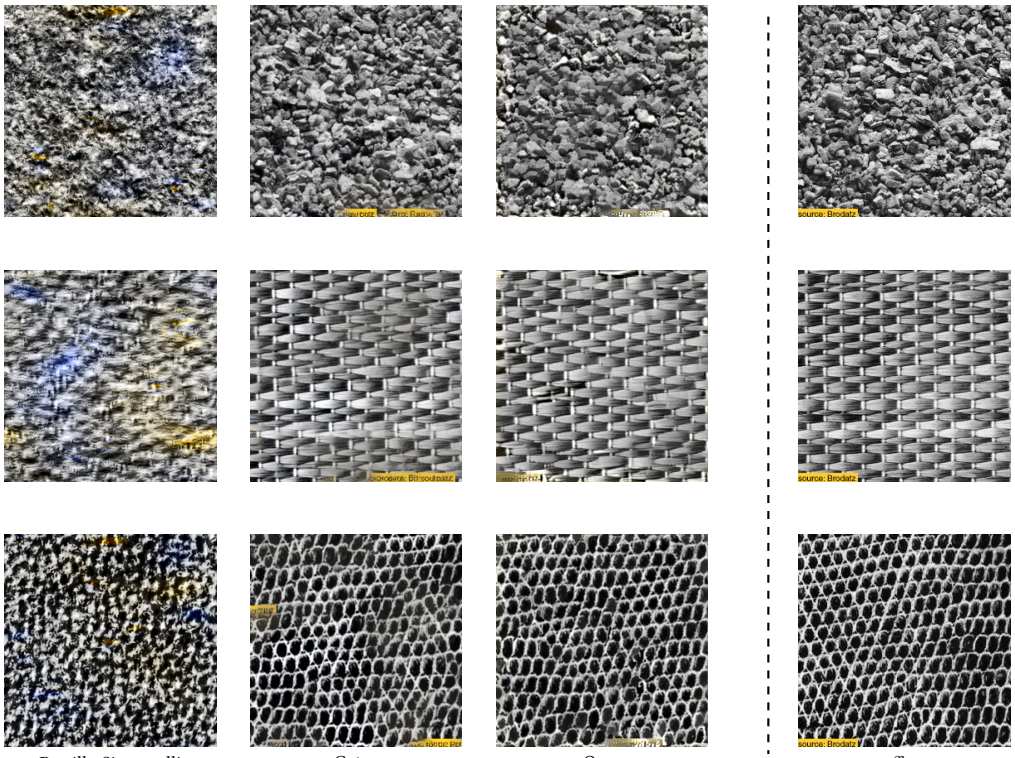


Figure 5.13: **Comparison with [PS00]** The images presented in the column "Portilla-Simoncelli" are synthesized with [PS00], the ones presented in the column "Gatys" are generated with [GEB15], the third column contains our results.

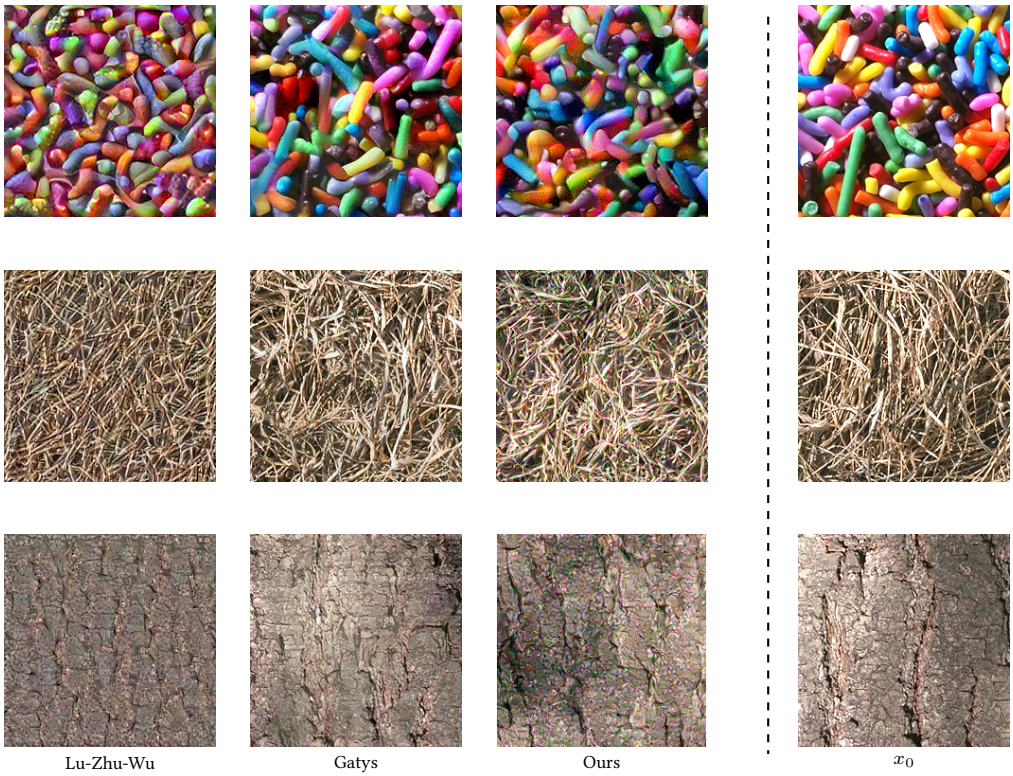
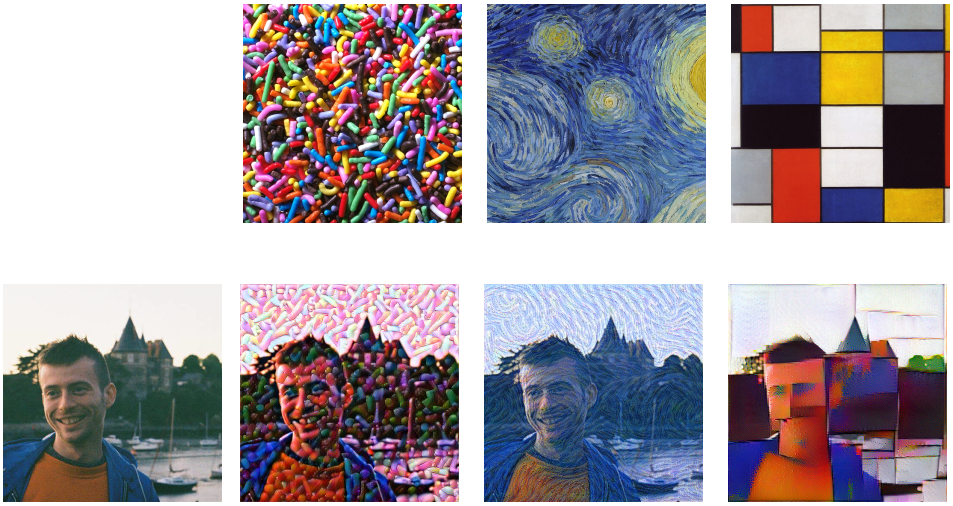


Figure 5.14: **Comparison with [LZW16]** The images presented in the column "Lu-Zhu-Wu" are synthesized with the algorithm introduced in [LZW16], the ones presented in the column "Gatys" are generated with [GEB15] and the third column contains our results.



exemplar image x_0 (a) (b) (c)

Figure 5.15: **Style transfer** In (a), (b) and (c) we present the outputs of the SOUL algorithm with an exemplar content given in the leftmost column and exemplar style given by the first row. See Section 5.2.3 for more details.

The V-total variation norm of ξ is defined as

$$\|\xi\|_V = \sup_{f \in \mathbb{F}(\mathbb{R}^d), \|f\|_V \leq 1} \left| \int_{\mathbb{R}^d} f(x) \mathrm{d}\xi(x) \right| .$$

If $V \equiv 1$, then $\|\cdot\|_V$ is the total variation norm denoted by $\|\cdot\|_{\mathrm{TV}}$.

Let $c : \mathbb{R}^d \times \mathbb{R}^d \to (0, +\infty]$ be defined for any $x, y \in \mathbb{R}^d$ by $c(x, y) = \mathbb{1}_{\Delta_{\mathbb{R}^d}}(x, y) W(x, y)$ where $W : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$ is a lower semi-continuous function such that for any $x, y, z \in \mathbb{R}^d$, $W(x, y) = W(y, x)$ and $W(x, z) \leq W(x, y) + W(y, z)$. Then for any probability measures μ and ν such that there exist $x_\mu, x_\nu \in \mathbb{R}^d$ satisfying $\mu(W(\cdot, x_\mu)) < \infty$ and $\nu(W(\cdot, x_\nu)) < \infty$, we define the Wasserstein extended distance associated with cost c between μ and ν by

$$\mathrm{d}_{W_R}(\mu, \nu) = \sup_{g \in \mathbb{G}_{\mu,W}} \left| \int_{\mathbb{R}^d} g(x) \mathrm{d}\mu(x) - \int_{\mathbb{R}^d} g(y) \mathrm{d}\nu(y) \right| , \tag{5.22}$$

with $\mathbb{G}_{\mu,W} = \{g \in \mathbb{F}(\mathbb{R}^d) \, : \, \|g(x) - g(y)\| \leq W(x, y) \text{ , for all } x, y \in \mathbb{R}^d\}$.

Proof of Theorem 5.2.1

This proof is an application of Theorem 4.2.2 and Theorem 4.2.4. Therefore, we are reduced to checking that **H**1 and **H**2 in Section 4.2.3 hold. More precisely, we study the geometric ergodicity of the Langevin Markov chain under **A**2(α), **B**1, **B**2(α) and **B**3 with $U_2 = 0$ and $\alpha \geq 1$ as well as its discretization error. Foster-Lyapunov conditions are derived in Lemma 5.2.5 and we check that **H**1-(i) in Theorem 4.2.2 holds in Lemma 5.2.6. In Theorem 5.2.7 we show that **H**1-(ii) in Theorem 4.2.2 is satisfied. We check that **H** 1-(iii) in Theorem 4.2.2 is satisfied in Lemma 5.2.8 and Proposition 5.2.9.

We denote by $\mathrm{K}_{\gamma,\theta}$ the Markov kernel associated with the following Langevin recursion

$$\tilde{X}_{n+1} = \tilde{X}_n - \gamma \left(\sum_{i=1}^{p} \theta(i) \nabla F_i(\tilde{X}_n) + \nabla r(\tilde{X}_n) \right) + \sqrt{2\gamma} \mathrm{Z}_{n+1} \ , \tag{5.23}$$

This kernel is given for any $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$

$$\mathrm{K}_{\gamma,\theta}(x, \mathsf{A}) = (2\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left[-(2\gamma)^{-1} \|y - x + \gamma \nabla_x U(\theta, x)\|^2 \right] \mathrm{d}y \ , \tag{5.24}$$

with for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$ by $U(\theta, x) = \langle \theta, F(x) \rangle + r(x)$ as in Section 5.2.2. Note that (5.24) is well-defined under **B**1 and **B**2(α) with $\alpha \geq 1$. We say that a Markov kernel K on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ satisfies a discrete Foster-Lyapunov drift condition $\mathbf{D}_{\mathrm{d}}(V, \lambda, b)$ if there exist $\lambda \in (0,1)$, $b \geq 0$ and a measurable function $V : \mathbb{R}^d \to [1, +\infty)$ such that for all $x \in \mathbb{R}^d$

$$\mathrm{K}V(x) \leq \lambda V(x) + b \ .$$

First, we state the following technical lemma.

Lemma 5.2.4. *Let $p \in \mathbb{N}^\star$. Then for any $u, v > 0$ and $t \geq 0$,*

$$u(1+t)^{2p-1} - v t^{2p} \leq \Upsilon_p(u, v)$$

with

$$\Upsilon_p(u, v) = 2^{(4p-2)p} \max\left\{ u, u^{2p}/v^{2p-1} \right\} \ .$$

Proof. Let $p \in \mathbb{N}^\star$, $\tilde{u}, \tilde{v} > 0$ and $\tilde{f}(t) = \tilde{u} t^{2p-1} - \tilde{v} t^{2p}$. We have for any $t \in \mathbb{R}$, $\tilde{f}'(t) = (2p-1)\tilde{u} t^{2p-2} - 2p\tilde{v} t^{2p-1}$. Since $\lim_{|t| \to +\infty} \tilde{f}(t) = -\infty$ and $\tilde{f}$ is continuous, the maximum is attained at some point t_0 which satisfies

$$\tilde{f}'(t_0) = (2p-1)\tilde{u} t_0^{2p-2} - 2p\tilde{v} t_0^{2p-1} = 0 \ ,$$

and therefore $t_0 = (2p-1)\tilde{u}/(2p\tilde{v})$. We have for any $t \geq 0$

$$\tilde{u} t^{2p-1} - \tilde{v} t^{2p} \leq \tilde{u} t_0^{2p-1} \leq \tilde{u} (\tilde{u}/\tilde{v})^{2p-1} \leq \tilde{u}^{2p}/\tilde{v}^{2p-1} \ . \tag{5.25}$$

If $t \geq 1$ then $u(1+t)^{2p-1} - v t^{2p} \leq 2^{2p-1} u t^{2p-1} - v t^{2p}$ and using (5.25) we have $u(1+t)^{2p-1} - v t^{2p} \leq 2^{(4p-2)p} u^{2p}/v^{2p-1}$. If $t \leq 1$ then $u(1+t)^{2p-1} - v t^{2p} \leq 2^{2p-1} u$, which concludes the proof. □

Lemma 5.2.5. *Assume* **B**1, **B**2(α) *and* **B**3 *with $U_2 = 0$ and $\alpha \geq 1$. Let $p \in \mathbb{N}^\star$, $\theta \in \mathsf{K}$ and $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} < \min(\mathtt{m}_1/\mathtt{L}^2, 1/2)$. Then $\mathrm{K}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, \tilde{b}\gamma)$ with*

$$V(x) = 1 + \|x - x^\star\|^{2p} \ , \qquad \lambda = \exp[-\mathtt{m}_1 + \bar{\gamma}\mathtt{L}^2] \ , \qquad \tilde{b}_p = \Upsilon_p(2^{2p+1} d^p \mathbf{\Gamma}(p + 1/2), \mathtt{m}_1) + \mathtt{m}_1 \ , \tag{5.26}$$

where for any $t \geq 0$, $\mathbf{\Gamma}(t) = \int_0^{+\infty} u^{t-1} \mathrm{e}^{-u} \mathrm{d}u$ and Υ_p is given in Lemma 5.2.4. In addition, $\mathrm{K}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b_p(1 + d^{\varpi_{0,p}})\gamma)$ with λ given in (5.26) and $b_p, \varpi_{0,p} \geq 0$ independent of the dimension d.

Proof. Let $x \in \mathbb{R}^d$, $p \in \mathbb{N}^\star$, $\theta \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$ and $\mathbb{Z}$ a d-dimensional Gaussian random variable with zero mean and identity covariance matrix. First, denoting $\mathbb{Z} = (z_1, \ldots, z_d)$ we have using Holder's inequality

$$\mathbb{E}\left[\|\mathbb{Z}\|^{2p} \right] = \sum_{i_1=1}^{d} \cdots \sum_{i_p=1}^{d} \mathbb{E}\left[\prod_{j=1}^{p} z_{i_j}^2 \right] \leq \sum_{i_1=1}^{d} \cdots \sum_{i_p=1}^{d} \mathbb{E}\left[z_1^{2p} \right] \leq (2d)^p \mathbf{\Gamma}(p + 1/2) \ . \tag{5.27}$$

Let $\mathcal{T}_\gamma(x) = x - x^\star - \gamma \nabla_x U(\theta, x)$. Using **B**3 we get

$$\begin{aligned}\|\mathcal{T}_\gamma(x)\|^2 &\leq \|x - x^\star\|^2 - 2\gamma\langle \nabla_x U(\theta, x) - \nabla_x U(\theta, x^\star), x - x^\star\rangle + \|\nabla_x U(\theta, x) - \nabla_x U(\theta, x^\star)\|^2 \\ &\leq (1 - 2\gamma \mathtt{m}_1 + \gamma^2 \mathtt{L}^2) \|x - x^\star\|^2 \ .\end{aligned}$$

Hence, we obtain

$$\begin{aligned}\mathbb{E}\left[\|X - x^\star\|^{2p}\right] &= \mathbb{E}\left[\sum_{k=0}^{p}\sum_{j=0}^{k} \binom{p}{k}\binom{k}{j} \|\mathcal{T}_\gamma(x)\|^{2(p-k)} (2\gamma)^{j/2} \langle \mathcal{T}_\gamma(x), \mathbb{Z}\rangle^j (2\gamma)^{k-j} \|\mathbb{Z}\|^{2(k-j)}\right] \\ &\leq (1 - \gamma(\mathtt{m}_1 - 2\mathtt{L}^2\bar{\gamma})) \|x - x^\star\|^{2p} + \gamma C_p(x - x^\star) \ , \end{aligned} \tag{5.28}$$

where we have, using that $\|\mathcal{T}_\gamma(x)\| \leq \|x - x^\star\|$, (5.27), $2\gamma \leq 1$, the Isserlis' formula [Iss18] and the Cauchy-Schwarz inequality

$$\begin{aligned}\gamma C_p(x - x^\star) &= \sum_{k=1}^{p}\sum_{j=0}^{k} \binom{p}{k}\binom{k}{j} \|x - x^\star\|^{2(p-k)} (2\gamma)^{k-j/2} \mathbb{E}\left[\langle \mathcal{T}_\gamma(x), \mathbb{Z}\rangle^j \|\mathbb{Z}\|^{2(k-j)}\right] \\ &= \sum_{k=1}^{p}\sum_{j=0}^{\lfloor k/2\rfloor} \binom{p}{k}\binom{k}{2j} \|x - x^\star\|^{2(p-k)} (2\gamma)^{k-j/2} \mathbb{E}\left[\langle \mathcal{T}_\gamma(x), \mathbb{Z}\rangle^{2j} \|\mathbb{Z}\|^{2(k-2j)}\right] \\ &\leq 2\gamma \sum_{k=1}^{p}\sum_{j=0}^{\lfloor k/2\rfloor} \binom{p}{k}\binom{k}{2j} \|x - x^\star\|^{2(p-k+j)} \mathbb{E}\left[\|\mathbb{Z}\|^{2(k-j)}\right] \\ &\leq 2\gamma(1 + \|x - x^\star\|)^{2p-1} (2d)^p \mathbf{\Gamma}(p + 1/2) \sum_{k=1}^{p}\sum_{j=0}^{\lfloor k/2\rfloor} \binom{p}{k}\binom{k}{2j} \\ &\leq 2^{2p+1}\gamma d^p \mathbf{\Gamma}(p + 1/2)(1 + \|x - x^\star\|)^{2p-1} \ . \end{aligned} \tag{5.29}$$

Combining (5.28) and (5.29) we get that

$$\begin{aligned}\mathrm{K}_{\gamma,\theta}(\|x - x^\star\|^{2p}) \leq (1 - \gamma(\mathtt{m}_1 - \mathtt{L}^2\bar{\gamma})) \|x - x^\star\|^{2p} \\ + 2^{2p+1}\gamma d^p \mathbf{\Gamma}(p + 1/2)(1 + \|x - x^\star\|)^{2p-1} - \gamma \mathtt{m}_1 \|x - x^\star\|^{2p} \ . \end{aligned} \tag{5.30}$$

Using Lemma 5.2.4, we have

$$2^{2p+1} d^p \mathbf{\Gamma}(p + 1/2)(1 + \|x - x^\star\|)^{2p-1} - \mathtt{m}_1 \|x - x^\star\|^{2p} \leq \Upsilon_p(2^{2p+1} d^p \mathbf{\Gamma}(p + 1/2), \mathtt{m}_1) \ .$$

Combining this result with (5.30) we get,

$$\mathrm{K}_{\gamma,\theta}(\|x - x^\star\|^{2p}) \leq (1 - \gamma(\mathtt{m}_1 - \bar{\gamma}\mathtt{L}^2)) \|x - x^\star\|^{2p} + \gamma \Upsilon_p(2^{2p+1} d^p \mathbf{\Gamma}(p + 1/2), \mathtt{m}_1) \ .$$

Therefore we obtain

$$\begin{aligned}\mathrm{K}_{\gamma,\theta}(1 + \|x - x^\star\|^{2p}) \leq (1 - \gamma(\mathtt{m}_1 - \bar{\gamma}\mathtt{L}^2))(1 + \|x - x^\star\|^{2p}) \\ + \gamma \left\{\Upsilon_p(2^{2p+1} d^p \mathbf{\Gamma}(p + 1/2), \mathtt{m}_1) + \mathtt{m}_1\right\} \ , \end{aligned}$$

which concludes the proof upon noting that $\tilde{b}_p$ is a polynomial in the dimension d. □

Lemma 5.2.6. *Assume* **B**1, **B**2(α) *and* **B**3 *with* $U_2 = 0$, $\alpha \geq 1$ *and let* $(X_k^n)_{n\in\mathbb{N},k\in\{0,\dots,m_n}$ *be given by* (5.19) *with* $\bar{\gamma} < \min(\mathtt{k}_1/\mathtt{L}^2, 1/2)$. *Let* $p \in \mathbb{N}^\star$, *then there exist* $A_{1,p} \geq 1$ *and* $\varpi_{1,p} \geq 0$ *such that for any* $n, p \in \mathbb{N}$ *and* $k \in \{0, \dots, m_n\}$

$$\mathbb{E}\left[\mathrm{K}_{\gamma_n,\theta_n}^p V(X_k^n)\middle| X_0^0\right] \leq A_{1,p}(1 + d^{\varpi_{1,p}})V(X_0^0) \;, \qquad \mathbb{E}\left[V(X_0^0)\right] < +\infty \;,$$

with $V(x) = 1 + \|x - x^\star\|^{2p}$ *and* $A_{1,p}, \varpi_{1,p} \geq 0$ *which do not depend on the dimension* d.

Proof. Combining Lemma 4.2.20 and Lemma 5.2.5 conclude the proof. □

Theorem 5.2.7. *Assume* **B**1, **B**2(α) *and* **B**3 *with* $U_2 = 0$ *and* $\alpha \geq 1$. *Then for any* $p \in \mathbb{N}^\star$ *there exist* $A_{2,p}, \varpi_{2,p} \geq 0$ *and* $\rho_p \in (0,1)$ *such that for any* $\theta \in \mathsf{K}$ *and* $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < \min(\mathtt{m}_1/\mathtt{L}^2, 1/2)$, $\mathrm{K}_{\gamma,\theta}$ *admits an invariant probability measure* $\pi_{\gamma,\theta}$ *and for any* $x, y \in \mathbb{R}^d$ *and* $n \in \mathbb{N}$

$$\begin{aligned}
\|\delta_x \mathrm{K}_{\gamma,\theta}^n - \pi_{\gamma,\theta}\|_V &\leq A_{2,p}(1 + d^{\varpi_{2,p}}) \exp[-n\kappa_p\gamma/\log^2(1 + d^{\varpi_{2,p}})]V(x) \;, \\
\|\delta_x \mathrm{K}_{\gamma,\theta}^n - \delta_y \mathrm{K}_{\gamma,\theta}^n\|_V &\leq A_{2,p}(1 + d^{\varpi_{2,p}}) \exp[-n\kappa_p\gamma/\log^2(1 + d^{\varpi_{2,p}})]\left\{V(x) + V(y)\right\} \;,
\end{aligned}$$

with $V(x) = 1 + \|x - x^\star\|^{2p}$ *and* $A_{2,p}, \varpi_{2,p} \geq 0$ *and* $\kappa_p > 0$ *which do not depend on the dimension* d.

Proof. For any $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \mathsf{K}$, $\mathrm{K}_{\gamma,\theta}$ has the Feller property and satisfies $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b\gamma)$ then [Dou+18, Theorem 12.3.3] applies and $\mathrm{K}_{\gamma,\theta}$ admits an invariant probability measure $\pi_{\gamma,\theta}$.

Let $p \in \mathbb{N}^\star$, $\theta \in \mathsf{K}$ and $\gamma \in (0, \bar{\gamma}]$. Using Proposition 4.1.5 in Section 4.1.3 with $\mathsf{A} \leftarrow \mathbb{R}^d$, we have for any $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$

$$\delta_{(x,y)}\tilde{\mathrm{K}}_{\gamma,\theta}^{n\lceil 1/\gamma\rceil}(\Delta_{\mathbb{R}^d}^{\mathrm{c}}) \leq 1 - 2\mathbf{\Phi}\left\{-\alpha^{-1/2}(n)\left\|x - y\right\|/(2\sqrt{2})\right\} \;,$$

where for any $x, y \in \mathbb{R}^d$, $\tilde{\mathrm{K}}_{\gamma,\theta}((x,y),\cdot)$ is the reflexive coupling between $\mathrm{K}_{\gamma,\theta}(x,\cdot)$ and $\mathrm{K}_{\gamma,\theta}(y,\cdot)$, see (4.23) in Section 4.1.3. In addition, we have for any $n \in \mathbb{N}$

$$\alpha^{-1}(n) = (2\mathtt{m}_1 - \mathtt{L}^2\bar{\gamma})/\left\{\exp((2\mathtt{m}_1 - \mathtt{L}^2\bar{\gamma})n) - 1\right\} \leq 2\mathtt{m}_1/(\lambda^{-2n} - 1) \leq 2\mathtt{m}_1\lambda^{2n}/(1-\lambda) \;.$$

and λ is given by (5.26). Therefore, we get that for any $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$

$$\delta_{(x,y)}\tilde{\mathrm{K}}_{\gamma,\theta}^{n\lceil 1/\gamma\rceil}(\Delta_{\mathbb{R}^d}^{\mathrm{c}}) \leq 1 - 2\mathbf{\Phi}\left\{-\lambda^n \mathtt{m}_1^{1/2}\left\|x - y\right\|/(1-\lambda)^{1/2}\right\} \;.$$

For any $x, y \in \mathbb{R}^d$ let

$$\begin{aligned}
W(x,y) &= 1 + (\|x - x^\star\|^{2p} + \|y - x^\star\|^{2p})/2 \;, \\
K_{\mathrm{d}} &= 2b_p(1 + d^{\varpi_{0,m}})(1 + \bar{\gamma})(1 + \log^{-1}(1/\lambda)) \;, \qquad M_{\mathrm{d}} = 2K_{\mathrm{d}}^{1/p} \;.
\end{aligned}$$

Note that for any $x, y \in \mathbb{R}^d$ such that $\|x - y\| \geq M_{\mathrm{d}}$, $W(x,y) \geq K_{\mathrm{d}}$. In addition, let $\mathrm{n}_0 = \max\{\lceil -\log(\mathtt{m}_1^{1/2}M_{\mathrm{d}}/(1-\lambda)^{1/2})\log^{-1}(1/\lambda)\rceil, 0\}$. We have for any $x, y \in \mathbb{R}^d$ with $\|x - y\| \leq M_{\mathrm{d}}$,

$$\delta_{(x,y)}\tilde{\mathrm{K}}_{\gamma,\theta}^{n\lceil 1/\gamma\rceil}(\Delta_{\mathbb{R}^d}^{\mathrm{c}}) \leq 1 - 2\mathbf{\Phi}(-1) \;.$$

Then, applying Theorem 4.1.8 in Section 4.1.3, with $\tilde{\mathrm{K}} \leftarrow \tilde{\mathrm{K}}_{\gamma,\theta}$, $p \leftarrow \mathrm{n}_0$, $\varepsilon_{1,\mathrm{d}} \leftarrow 2\mathbf{\Phi}(-1)$, $\bar{C}_1 \leftarrow C$, $\bar{\rho}_1 \leftarrow \rho$, $\bar{A}_1 \leftarrow A$ and $\bar{c}_1 \leftarrow c$, we obtain that for any $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$

$$\|\delta_x \mathrm{K}_{\gamma,\theta}^n - \delta_y \mathrm{K}_{\gamma,\theta}^n\|_V \leq C\rho^{\lfloor k/(\mathrm{n}_0\lceil 1/\gamma\rceil)\rfloor}\left\{V(x) + V(y)\right\}/2 \;,$$

where

$$C = 2\left[1+A\right]\left[1+(A+K_{\mathrm{d}})/\left\{\mathbf{\Phi}(-1)(1-\lambda)\right\}\right] ,$$
$$A = b_p(1+d^{\varpi_{0,p}})(1+\bar{\gamma})(1+\log^{-1}(1/\lambda))) ,$$
$$\log^{-1}(1/\rho) = \log^{-1}(1/(2\mathbf{\Phi}(-1))) + \log^{-1}(2/(1+\lambda)) + \log(A+K_{\mathrm{d}})\log^{-1}(1/(2\mathbf{\Phi}(-1)))\log^{-1}(2/(1+\lambda)) .$$

Since $\lfloor n/(\mathrm{n}_0\lceil 1/\gamma\rceil)\rfloor \geq n\gamma/(\mathrm{n}_0(1+\bar{\gamma})) - 1$, setting $\tilde{A}_{2,p} = C\rho^{-1}/2$ and $\rho_p = \rho^{1/(\mathrm{n}_0(1+\bar{\gamma}))}$, we get that for any $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$

$$\|\delta_x \mathrm{K}^n_{\gamma,\theta} - \delta_y \mathrm{K}_{\gamma,\theta}\|_V \leq \tilde{A}_{2,p}\rho_p^{\gamma n}\left\{V(x)+V(y)\right\} . \tag{5.31}$$

Using Lemma 5.2.5 we have that $\pi_{\gamma,\theta}(V) \leq b_p(1+d^{\varpi_{0,m}})(1+\log^{-1}(1/\lambda))V(x)$. Combining this result with (5.31) we get that for any $x \in \mathbb{R}^d$ and $n \in \mathbb{N}$

$$\|\delta_x \mathrm{K}^n_{\gamma,\theta} - \pi_{\gamma,\theta}\|_V \leq \tilde{A}_{2,p}\left\{1 + b_p(1+d^{\varpi_{0,m}})(1+\log^{-1}(1/\lambda))\right\}\tilde{\rho}_p^{\gamma n}V(x) .$$

Since in Lemma 5.2.5, λ and b_p do not depend on the dimension d we get that K_{d} is upper-bounded by a polynomial in the dimension d. Hence, there exists $\varpi^{(a)}_{2,p} > 0$ which does not depend on the dimension such that $\tilde{A}_{2,p}\{1 + b_p(1+d^{\varpi_{0,m}})(1+\log^{-1}(1/\lambda))\} \leq A_{2,p}(1+d^{\varpi^{(a)}_{2,p}})$ with $A_{2,p} \geq 0$ which does not depend on the dimension d. Similarly, there exists $\varpi^{(b)}_{2,p} > 0$ independent of d such that $\sup_{d\in\mathbb{N}}[\{\log^{-1}(\rho)+\mathrm{n}_0\}/\log(1+d^{\varpi^{(b)}_{2,p}})^{-1}] < +\infty$ which implies that $\log^{-1}(1/\rho_m) \leq \kappa_p^{-1}\log^2(1+d^{\varpi^{(b)}_{2,p}})$ with $\kappa_p > 0$ which do not depend on the dimension d. We conclude the proof upon setting $\varpi_{2,p} = \max(\varpi^{(a)}_{2,p}, \varpi^{(b)}_{2,p})$. □

Similarly to the discrete setting, we say that a Markov semi-group $(\mathrm{P}_t)_{t\geq 0}$ on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ with extended infinitesimal generator $(\mathcal{A}, \mathrm{D}(\mathcal{A}))$ (see *e.g.* [MT93c] for the definition of $(\mathcal{A}, \mathrm{D}(\mathcal{A}))$) satisfies a continuous drift condition $\mathbf{D}_{\mathrm{c}}(V,\zeta,\beta)$ if there exist $\zeta > 0$, $\beta \geq 0$ and a measurable function $V : \mathbb{R}^d \to [1,+\infty)$ with $V \in \mathrm{D}(\mathcal{A})$ such that for all $x \in \mathbb{R}^d$

$$\mathcal{A}V(x) \leq -\zeta V(x) + \beta .$$

Let $\theta \in \mathsf{K}$ and $(\mathrm{P}_{t,\theta})_{t\geq 0}$ be the Markov semi-group associated with the Langevin diffusion

$$\mathrm{d}\mathbf{X}_t = -\left(\sum_{i=1}^p \theta(i)\nabla F_i(\mathbf{X}_t) + \nabla r(\mathbf{X}_t)\right)\mathrm{d}t + \mathrm{d}\mathbf{B}_t ,$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion. Consider now the generator $\mathcal{A}_\theta$ of $(\mathrm{P}_{t,\theta})_{t\geq 0}$ for any $\theta \in \mathsf{K}$, defined for any $f \in \mathrm{C}^2(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ by

$$\mathcal{A}_\theta f(x) = -\left\langle \nabla f(x), \sum_{i=1}^p \theta(i)\nabla F_i(x) + \nabla r(x)\right\rangle + \Delta f(x) .$$

Using Lemma 4.2.19 we have that π_θ is an invariant probability measure for $(\mathrm{P}_{t,\theta})_{t\geq 0}$.

Lemma 5.2.8. *Assume* **B**1, **B**2(α) *and* **B**3 *with* $U_2 = 0$ *and* $\alpha \geq 1$. *Then for any* $p \in \mathbb{N}^\star$ *there exist* $\zeta > 0$ *and* $\beta \geq 0$ *such that for any* $\theta \in \mathsf{K}$, $\mathcal{A}_\theta$ *satisfies* $\mathbf{D}_{\mathrm{c}}(V,\zeta,\tilde{\beta}_p)$ *with*

$$V(x) = 1 + \|x - x^\star\|^{2p} , \qquad \zeta = -\mathtt{m}_1 p , \qquad \tilde{\beta}_p = 2p\Upsilon_p(2(p-1)+d, \mathtt{k}_1/2) + 2\mathtt{m}_1 p , \tag{5.32}$$

with Υ_m *given in Lemma 5.2.4. In addition,* $\mathcal{A}_\theta$ *satisfies* $\mathbf{D}_{\mathrm{c}}(V,\zeta,\beta_p(1+d^{\varpi'_{0,p}})\gamma)$ *with* ζ *given in* (5.32) *and* $\beta_p, \varpi'_{0,p} \geq 0$ *independent of the dimension* d.

Proof. Let $\theta \in \mathsf{K}$ and $p \in \mathbb{N}^\star$. Then, we have for any $x \in \mathbb{R}^d$

$$\begin{aligned} V(x) &= 1 + \|x - x^\star\|^{2p} \ , \\ \nabla V(x) &= 2p \|x - x^\star\|^{2(p-1)} (x - x^\star) \ , \\ \nabla^2 V(x) &= 4p(p-1) \|x - x^\star\|^{2(p-2)} (x - x^\star)(x - x^\star)^\top + 2p \|x - x^\star\|^{2(p-1)} \mathrm{Id} \ . \end{aligned} \tag{5.33}$$

Hence, for any $x \in \mathbb{R}^d$, $\Delta V(x) = 4p(p-1) \|x - x^\star\|^{2(p-1)} + 2pd \|x - x^\star\|^{2(p-1)}$. Using **B**3, (5.33) and Lemma 5.2.4 we get that for any $x \in \mathbb{R}^d$

$$\begin{aligned} &\mathcal{A}_\theta V(x) = -2p \|x - x^\star\|^{2(p-1)} \langle \nabla_x U(\theta, x), x - x^\star \rangle + 2p \left(2(p-1) + d\right) \|x - x^\star\|^{2(p-1)} \\ &= -2p \|x - x^\star\|^{2(p-1)} \langle \nabla_x U(\theta, x) - \nabla_x U(\theta, x^\star), x - x^\star \rangle + 2p \left(2(p-1) + d\right) \|x - x^\star\|^{2(p-1)} \\ &\leq -2\mathtt{m}_1 p V(x) + 2p \left(2(p-1) + d\right) \|x - x^\star\|^{2(p-1)} + 2\mathtt{m}_1 p \\ &\leq -\mathtt{m}_1 p V(x) + 2p \left(2(p-1) + d - \mathtt{m}_1 \|x - x^\star\|^2 / 2\right) \|x - x^\star\|^{2(p-1)} + 2\mathtt{m}_1 p \\ &\leq -\mathtt{m}_1 p V(x) + 2p \Upsilon_p(2(p-1) + d, \mathtt{k}_1/2) + 2\mathtt{m}_1 p \ , \end{aligned}$$

which concludes the proof. □

Proposition 5.2.9. *Assume* **B**1, **B**2(α) *and* **B**3 *with* $U_2 = 0$ *and* $\alpha \geq 1$. *Then for any* $p \in \mathbb{N}^\star$, *there exist* $A_{3,p}, \varpi_{3,p} \geq 0$ *such that for any* $\theta \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < \min(\mathtt{m}_1/\mathtt{L}^2, 1/2)$,

$$\|\pi_{\gamma,\theta} - \pi_\theta\|_{V^{1/2}} \leq A_{3,p}(1 + d^{\varpi_{3,p}})\gamma^{1/2} \ ,$$

with $V(x) = 1 + \|x - x^\star\|^{2p}$ *and* $A_{3,p}, \varpi_{3,p} \geq 0$ *which do not depend on the dimension* d.

The proof is similar to the one of Proposition 4.2.22 except that in this presentation we explicit the constants appearing in the proof and track the dependency of the constants with respect to the dimension d.

Proof. Let $p \in \mathbb{N}^\star$, $\theta \in \mathsf{K}$ and $\gamma \in (0, \bar{\gamma}]$. Since π_θ is an invariant probability measure for $(\mathrm{P}_{t,\theta})_{t \geq 0}$ we have using Theorem 5.2.7 that

$$\lim_{k \to +\infty} \|\pi_\theta \mathrm{K}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}} = \|\pi_{\gamma,\theta} - \pi_\theta\|_{V^{1/2}} \ .$$

We now give an upper bound on $\|\pi_\theta \mathrm{K}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}}$ for $k = q_\gamma m_\gamma$ with $m_\gamma = \lceil 1/\gamma \rceil$ and $q_\gamma \in \mathbb{N}$. We obtain using Theorem 5.2.7

$$\begin{aligned} &\|\pi_\theta \mathrm{K}_{\gamma,\theta}^k - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}} \\ &\leq \sum_{j=0}^{q_\gamma - 1} \|\pi_\theta \mathrm{P}_{\gamma(j+1)m_\gamma,\theta} \mathrm{K}_{\gamma,\theta}^{(q_\gamma - (j+1))m_\gamma} - \pi_\theta \mathrm{P}_{\gamma j m_\gamma,\theta} \mathrm{K}_{\gamma,\theta}^{(q_\gamma - j)m_\gamma}\|_{V^{1/2}} \\ &\leq \sum_{j=0}^{q_\gamma - 1} \Big\{ A_{2,p}(1 + d^{\varpi_{2,p}}) \exp[-\kappa_p \gamma m_\gamma (q_\gamma - (j+1)) / \log^2(1 + d^{\varpi_{2,p}})] \\ &\qquad \times \|\pi_\theta \mathrm{P}_{\gamma j m_\gamma,\theta} \mathrm{P}_{m_\gamma \gamma,\theta} - \pi_\theta \mathrm{P}_{\gamma j m_\gamma,\theta} \mathrm{K}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \Big\} \\ &\leq \|\pi_\theta \mathrm{P}_{m_\gamma \gamma,\theta} - \pi_\theta \mathrm{K}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \sum_{j=1}^{q_\gamma} A_{2,p}(1 + d^{\varpi_{2,p}}) \exp[-\kappa_p \gamma j m_\gamma / \log^2(1 + d^{\varpi_{2,p}})] \end{aligned}$$

$$\leq A_{2,p}(1+d^{\varpi_{2,p}})(1+\log^2(1+d^{\varpi_{2,p}})/\kappa_p)\|\pi_\theta \mathrm{P}_{m_\gamma\gamma,\theta} - \pi_\theta \mathrm{K}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \;, \tag{5.34}$$

We now give an upper bound on $\|\pi_\theta \mathrm{P}_{m_\gamma\gamma,\theta} - \pi_\theta \mathrm{K}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}}$. Indeed, since $\mathcal{A}_\theta$ satisfies $\mathbf{D}_{\mathrm{c}}(V,\zeta,\beta)$ by Lemma 5.2.8 and $\mathrm{K}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(V,\lambda^\gamma,b\gamma)$ for any $\theta \in \mathsf{K}$ and $\gamma \in (0,\bar{\gamma}]$ by Lemma 5.2.5 and, we obtain that

$$\pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta}(V) \leq D_0 \;, \qquad \pi_\theta \mathrm{K}_{\gamma,\theta}^{m_\gamma}(V) \leq D_1 \;,$$
$$D_0 = \beta_p(1+d^{\varpi'_{0,p}})/\zeta \;, \qquad D_1 = D_0 + b_p(1+d^{\varpi_{0,p}})(\bar{\gamma}+\log^{-1}(1/\lambda)) \;.$$

Combining this result and Lemma 4.2.21 we have for any $\theta \in \mathsf{K}$ and $\gamma \in (0,\bar{\gamma}]$

$$\|\pi_\theta \mathrm{P}_{\gamma m_\gamma,\theta} - \pi_\theta \mathrm{K}_{\gamma,\theta}^{m_\gamma}\|_{V^{1/2}} \leq D_2 \gamma^{1/2} \;, \tag{5.35}$$

with

$$D_2 = 2D_1^{1/2}(1+\bar{\gamma})^{1/2}\left\{d + 2\bar{\gamma}(2\mathtt{L}^2 + \sup_{\theta\in\mathsf{K}}\|\nabla_x U_1(\theta,0)\|^2)D_1\right\}^{1/2}\mathtt{L} \;.$$

Combining (5.34) and (5.35) we get

$$\|\pi_\theta \mathrm{K}_{\gamma,\theta}^{k} - \pi_\theta \mathrm{P}_{\gamma k,\theta}\|_{V^{1/2}} \leq D_2 A_{2,p}(1+d^{\varpi_{2,p}})(1+\log^2(1+d^{\varpi_{2,p}})/\kappa_p)\gamma^{1/2} \;.$$

Combining (4.185), (4.183) and Lemma 5.2.5, we get that D_0, D_1, D_2 are upper-bounded by polynomials in the dimension d, which concludes the proof. □

We now turn to the proof of Theorem 5.2.1.

Proof. Let $p = \lceil 2\alpha \rceil$ and $V(x) = 1 + \|x - x^\star\|^{2p}$. Lemma 5.2.6 implies **H**1-(i) in Theorem 4.2.2 with $A_{1,p} \leftarrow A_{1,p}(1+d^{\varpi_{1,p}})$, Theorem 5.2.7 implies **H**1-(ii) in Theorem 4.2.2 with $A_{2,p} \leftarrow A_{2,p}(1+d^{\varpi_{2,p}})$ and $\rho \leftarrow \exp[-\kappa_p/\log^2(1+d^{\varpi_{2,p}})]$. In addition, Proposition 5.2.9 implies **H**1-(iii) in Theorem 4.2.2 with $\boldsymbol{\Psi}(\gamma) = A_{3,p}(1+d^{\varpi_{3,p}})\gamma^{1/2}$. Using Proposition 5.1.4 we have that **A**1, **A**2 and **A**3 in Theorem 4.2.2 hold. Since $H_\theta \leftarrow F$ in (5.19) we get that for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$, $\|H_\theta(x)\| \leq V^{1/2}(x)$. We can apply Theorem 4.2.2 and we get that for any $n \in \mathbb{N}$

$$\mathbb{E}\left[\left\{\sum_{k=1}^n \delta_k L(\theta_k)\right/\sum_{k=1}^n \delta_k\right\} - \min_{\mathsf{K}} L\right] \leq E_n \Big/ \left(\sum_{k=1}^n \delta_k\right) \;,$$

with,

$$\begin{aligned} E_n &= 2M_\Theta^2 + 2B_{1,p}M_\Theta \mathbb{E}\left[V^{1/2}(X_0^0)\right]\sum_{k=0}^{n-1}\delta_{k+1}/(m_k\gamma_k) \\ &\quad + 2M_\Theta A_{3,p}(1+d^{\varpi_{3,p}})\sum_{k=0}^{n-1}\delta_{k+1}\gamma_k^{1/2} + 4B_{1,p}^2\mathbb{E}\left[V(X_0^0)\right]\sum_{k=0}^{n-1}\delta_{k+1}^2/(m_k\gamma_k)^2 \\ &\quad + 4A_{3,p}^2(1+d^{\varpi_{3,p}})^2\sum_{k=0}^{n-1}\delta_{k+1}^2\gamma_k + B_{2,p}\sum_{k=0}^{n-1}\delta_{k+1}^2/(m_k\gamma_k)^2 \;, \end{aligned} \tag{5.36}$$

and

$$\begin{aligned} B_{1,p} &= 2(1+d^{\varpi_p})^2\log^2(1+d^{\varpi_p})A_{1,p}A_{2,p}\exp[-\bar{\gamma}\kappa_p/\log^2(1+d^{\varpi_p})]/\kappa_p \;; \\ B_{2,p} &= 2(1+\bar{\gamma})^2\max(B_{2,p}^{(a)}, B_{2,p}^{(b)}) \;; \\ B_{2,p}^{(a)} &= 24(1+d^{\varpi_p})^3A_{2,p}^2(1-\exp[-\kappa_p/(2\log^2(1+d^{\varpi_p}))])^{-2}A_{3,p} \;; \\ B_{2,p}^{(b)} &= 4(1+d^{\varpi_p})^3A_{1,p}\left[1+6A_{2,p}^2(1-\exp[-\kappa_p/(2\log^2(1+d^{\varpi_p}))])^{-2}\right. \\ &\qquad \left.\times\left\{A_{2,p}(1-\exp[-\kappa_p/\log^2(1+d^{\varpi_p})])^{-1}+2\right\} + A_{2,p}^2\log^4(1+d^{\varpi_p})/\kappa_p^2 + A_{3,p}^2\right] \;, \end{aligned}$$

which concludes the proof of Theorem 5.2.1-(b), upon setting $\varpi_p = \max_{i\in\{1,2,3\}}(\varpi_{i,p})$.

We have that for any $x \in \mathbb{R}^d$, $\|H_\theta(x)\| = \|F(x)\| \leq V^{1/4}(x)$. Since H_θ does not depend on θ we get that **A**4 in Section 4.2.3 is satisfied. In addition Proposition 4.2.24 in Section 4.2.5 implies that **H**2 is satisfied with

$$\mathbf{\Lambda}_1(\gamma_1,\gamma_2) = A_{4,p}(1+d^{\varpi_{4,p}})\gamma_2^{-1/2}\left|\gamma_1-\gamma_2\right| \ , \quad \mathbf{\Lambda}_2(\gamma_1,\gamma_2) = A_{4,p}(1+d^{\varpi_{4,p}})\gamma_2^{1/2} \ ,$$

with $A_{4,p} \geq 0$ which does not depend on the dimension d. As a consequence we can apply Theorem 4.2.4 and we get that for any $n \in \mathbb{N}^\star$

$$\mathbb{E}\left[\left\{\sum_{k=1}^n \delta_k L(\theta_k)\Big/\sum_{k=1}^n \delta_k\right\} - \min_{\mathsf{K}} L\right] \leq \tilde{E}_n \Big/ \left(\sum_{k=1}^n \delta_k\right) \ ,$$

with,

$$\begin{aligned}
\tilde{E}_n &= 2M_\Theta + 2M_\Theta \sum_{k=0}^n \delta_{k+1}\mathbf{\Psi}(\gamma_k) + C_{3,p}\sum_{k=0}^n \left|\delta_{k+1}-\delta_k\right|\gamma_k^{-1} \\
&+ 2M_\Theta C_{2,p}\sum_{k=0}^n \delta_{k+1}\gamma_{k+1}^{-2}\left[\mathbf{\Lambda}_1(\gamma_k,\gamma_{k+1}) + \mathbf{\Lambda}_2(\gamma_k,\gamma_{k+1})\delta_{k+1} + \delta_{k+1}\gamma_{k+1}\right] \\
&+ C_{3,p}\sum_{k=0}^n \delta_{k+1}^2\gamma_{k+1}^{-1} + C_{3,p}(\delta_{n+1}/\gamma_n - \delta_0/\gamma_0) + C_{1,p}\sum_{k=0}^n \delta_{k+1}^2 \ ,
\end{aligned}$$

□

with

$$\begin{aligned}
C_{1,p} &= 2A_{1,p}(1+d^{\varpi_p})\mathbb{E}\left[V(X_0^0)\right] + 2\sup_{\theta\in\mathsf{K}}\|\nabla L(\theta)\|^2 \ , \\
C_{2,p} &= 8(1+d^{\varpi_p})^4\log^4(1+d^{\varpi_p})A_{1,p}A_{2,p}^2 \\
&\qquad \times \exp[-2\bar{\gamma}\kappa_p/\log^2(1+d^{\varpi_p})](1+2A_{1,p}\mathbb{E}\left[V(X_0^0)\right])/\kappa_p \ , \\
C_{3,p} &= (1+d^{\varpi_p})A_{1,p}C_H(4M_\Theta + \sup_{\theta\in\mathsf{K}}\|\nabla L(\theta)\| + 1 + \delta_1\mathtt{B})\mathbb{E}\left[V(X_0^0)^{1/4}\right] \ , \\
C_H &= 8(1+d^{\varpi_p})\log^2(1+d^{\varpi_p})A_{2,p}\exp[-\bar{\gamma}\kappa_p/(4\log^2(1+d^{\varpi_p}))]/\kappa_p \ .
\end{aligned}$$

Similarly to Theorem 5.2.1-(b) since $A_{1,p}$, $A_{2,p}$, $A_{3,p}$, $A_{4,p}$ and κ_p are independent of the dimension d. Setting $\varpi_p = \max_{i\in\{1,2,3,4\}}(\varpi_{i,p})$ concludes the proof of Theorem 5.2.1-(a)

Proof of Theorem 5.2.2

In this section, we give alternative results to Theorem 4.2.2 and Theorem 4.2.4. The main results of Section 4.2.3 are stated in V-norm or total variation. However, our particular framework allows us to use a Wasserstein distance with an appropriate cost function which implies that the constants appearing in our results scale polynomially in the dimension d even if the potential is non convex. The increasing batch size case is considered in Section 5.2.4 and the fixed batch size case in Section 5.2.4. We check that the main assumptions **H**1 and **H**2 below are satisfied in the setting of Theorem 5.2.2 in Section 5.2.4 and conclude in Section 5.2.4.

Increasing batch size In this section, we give an alternative result to Theorem 4.2.2 in the case where (a) the controls on the family of Markov kernels $\{\mathrm{K}_{\gamma,\theta} : \gamma \in (0,\bar{\gamma}], \theta \in \mathsf{K}\}$ are obtained with respect to an appropriate Wasserstein distance (b) the stochastic gradient does not depend on θ. First, we show that under **B**2(1), $\mu \mapsto \mu(F)$ is Lipschitz with respect to the considered Wasserstein distance in Lemma 5.2.11. Then, we control the error in the perturbed gradient scheme in Lemma 5.2.12 and Lemma 5.2.13. Our main result is stated in Theorem 5.2.10.

Let $c : \mathbb{R}^d \times \mathbb{R}^d \to [0,+\infty)$, defined for any $x, y \in \mathbb{R}^d$ by $c(x,y) = \mathbb{1}_{\Delta_{\mathbb{R}^d}}(x,y)(1 + \|x-y\|/R)$ where $R \geq 0$. Consider also the function $W_R : \mathbb{R}^{2d} \to \mathbb{R}_+$, and $V_R : \mathbb{R}^d \to \mathbb{R}_+$ given for $x, y \in \mathbb{R}^d$ by

$$W_R(x,y) = 1 + \|x-y\|/R\,, \qquad V_R(x) = 1 + \|x\|/R\,. \tag{5.37}$$

We also define for any $p \in \mathbb{N}$, $\mathscr{V}_p : \mathbb{R}^d \to [1,+\infty)$ given for any $x \in \mathbb{R}^d$ by

$$\mathscr{V}_p(x) = 1 + \|x\|^{2p}\,. \tag{5.38}$$

We recall that $\mathrm{K}_{\gamma,\theta}$ is the Markov kernel associated with the Langevin recursion (5.23) and expression given by (5.24). This kernel is well-defined under **B**1 and **B**2(α) with $\alpha \geq 1$. Consider the following assumption.

H1. *(i) There exists $A_1 \geq 1$ such that for any $a \in [1,3]$, $n, p \in \mathbb{N}$, $k \in \{0,\dots,m_n\}$*

$$\mathbb{E}\left[\mathrm{K}^p_{\gamma_n,\theta_n} V_R^a(X_k^n) \middle| X_0^0\right] \leq A_1 V_R^a(X_0^0)\,, \qquad \mathbb{E}\left[V_R^a(X_0^0)\right] < +\infty\,.$$

with $\{(X_k^\ell)_{k\in\{0,\dots,m_\ell\}} : \ell \in \{0,\dots,n\}\}$ given by (5.19).

(ii) There exist $A_2, A_3 \geq 1$, $\rho \in [0,1)$ such that for any $\gamma \in (0,\bar{\gamma}]$, $\theta \in \mathsf{K}$, $x, y \in \mathbb{R}^d$ and $n \in \mathbb{N}$, $\mathrm{K}_{\gamma,\theta}$ has a stationary distribution $\pi_{\gamma,\theta}$ and

$$\mathrm{d}_{W_R}(\delta_x \mathrm{K}^n_{\gamma,\theta}, \delta_y \mathrm{K}^n_{\gamma,\theta}) \leq A_2 \rho^{\gamma n} W_R(x,y)\,, \quad \mathrm{d}_{W_R}(\delta_x \mathrm{K}^n_{\gamma,\theta}, \pi_{\gamma,\theta}) \leq A_2 \rho^{\gamma n} V_R(x)\,, \quad \pi_{\gamma,\theta}(V_R) \leq A_3\,.$$

(iii) There exists $\boldsymbol{\Psi} : \mathbb{R}_+^\star \to \mathbb{R}_+$ such that for any $\gamma \in (0,\bar{\gamma}]$ and $\theta \in \mathsf{K}$, $\mathrm{d}_{W_R}(\pi_{\gamma,\theta}, \pi_\theta) \leq \boldsymbol{\Psi}(\gamma)$.

Theorem 5.2.10. *Assume **A**2(1), **B**1, **B**2(1) and **H**1. Let $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ be sequences of non-increasing positive real numbers and $(m_n)_{n\in\mathbb{N}}$ a sequence of positive integers satisfying $\delta_n < \left(\sup_{\theta\in\mathsf{K}} \|\nabla^2 L(\theta)\|\right)^{-1}$ and $\gamma_n < \bar{\gamma}$. Then, there exists $(E_n)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}^\star$*

$$\mathbb{E}\left[\left\{\sum_{k=1}^n \delta_k L(\theta_k) \Big/ \sum_{k=1}^n \delta_k\right\} - \min_{\mathsf{K}} L\right] \leq E_n \Big/ \left(\sum_{k=1}^n \delta_k\right)\,,$$

with $(\theta_k)_{k\in\mathbb{N}}$ and L are defined in (5.19) and (5.3) respectively, and for any $n \in \mathbb{N}^\star$

$$E_n = 2M_\Theta^2 + 6M_\Theta R \mathtt{M} p A_1 A_2 (\rho^{-\bar{\gamma}}/\log(1/\rho) + 1) \sum_{k=0}^{n-1} \delta_{k+1} \{1/(m_k\gamma_k) + \boldsymbol{\Psi}(\gamma_k)\}$$
$$+ \left(2A_1(\|F(0)\| + 3R\mathtt{M})^2 \mathbb{E}\left[V_R^2(X_0)\right] + 2\sup_{\theta\in\mathsf{K}} L(\theta)^2\right) \sum_{k=0}^{n-1} \delta_{k+1}^2\,.$$

The proof of this result is a simple adaptation to the one of Theorem 4.2.2. However it is given for completeness.

Let $(\eta_n)_{n\in\mathbb{N}}$ be defined for any $n \in \mathbb{N}$ by

$$\eta_n = m_n^{-1} \sum_{k=1}^{m_n} \{F(X_k^n) - \pi_{\theta_n}(F)\} \ . \tag{5.39}$$

We consider the following decomposition for any $n \in \mathbb{N}$,

$$\eta_n = \eta_n^{(1)} + \eta_n^{(2)} \ , \qquad \eta_n^{(1)} = \mathbb{E}\left[\eta_n | \mathcal{F}_{n-1}\right] \ , \qquad \eta_n^{(2)} = \eta_n - \mathbb{E}\left[\eta_n | \mathcal{F}_{n-1}\right] \ , \tag{5.40}$$

and $(\mathcal{F}_n)_{n\in\mathbb{N}\cup\{-1\}}$ is defined for all $n \in \mathbb{N}$ by

$$\mathcal{F}_n = \sigma\left(\theta_0, \{(X_k^\ell)_{k\in\{0,\ldots,m_\ell\}} \ : \ \ell \in \{0,\ldots,n\}\}\right) \ , \qquad \mathcal{F}_{-1} = \sigma(\theta_0) \tag{5.41}$$

We start with the following technical lemmas.

Lemma 5.2.11. *Assume* **B**2(1) *and* **H**1. *Then for any probability measures μ, ν on $\mathcal{B}(\mathbb{R}^d)$ such that $\mu(\|\cdot\|) + \nu(\|\cdot\|) < +\infty$,*

$$\|\mu(F) - \nu(F)\| \leq 3R\mathtt{M}p\mathrm{d}_{W_R}(\mu,\nu) \ ,$$

with W_R given in (5.37).

Proof. Using **B**2(1) we have that for any $i \in \{1,\ldots,p\}$ and $x, y \in \mathbb{R}^d$, $|F_i(x) - F_i(y)| \leq 3\mathtt{M}\|x-y\| \leq 3R\mathtt{M}W_R(x,y)$. Let μ and ν on $\mathcal{B}(\mathbb{R}^d)$ such that $\mu(\|\cdot\|) + \nu(\|\cdot\|) < +\infty$. Using the definition of the Wasserstein distance (5.22), we have $\|\mu(F) - \nu(F)\| \leq \sum_{i=1}^p |\mu(F_i) - \nu(F_i)| \leq 3R\mathtt{M}p\mathrm{d}_{W_R}(\mu,\nu)$. □

Lemma 5.2.12. *Assume* **B**1, **B**2(1) *and* **H**1. *Then we have for any $n \in \mathbb{N}$*

$$\mathbb{E}\left[\|\eta_n^{(1)}\|\right] \leq B_1 \left\{\mathbb{E}\left[V_R(X_0^0)\right]/(m_n\gamma_n) + \mathbf{\Psi}(\gamma_n)\right\} \ ,$$

with $B_1 = 3R\mathtt{M}pA_1A_2(\rho^{-\bar{\gamma}}/\log(1/\rho)+1)$.

Proof. Using the definition of $(\mathcal{F}_n)_{n\in\mathbb{N}}$, see (5.41), the Markov property, **H**1-(ii)-(iii), Lemma 5.2.11 and that for any $\theta \in \mathsf{K}$, we have for any $n \in \mathbb{N}^\star$

$$\begin{aligned}
\|\mathbb{E}\left[\eta_n|\mathcal{F}_{n-1}\right]\| &\leq m_n^{-1} \sum_{k=1}^{m_n} \left\|\mathrm{K}_{\gamma_n,\theta_n}^k F(X_0^n) - \pi_{\theta_n}(F)\right\| \\
&\leq 3R\mathtt{M}pm_n^{-1} \sum_{k=1}^{m_n} \left\{\mathrm{d}_{W_R}(\delta_{X_0^n}\mathrm{K}_{\gamma_n,\theta_n}^k, \pi_{\gamma_n,\theta_n})\right\} + 3R\mathtt{M}p\mathrm{d}_{W_R}(\pi_{\gamma_n,\theta_n}, \pi_{\theta_n}) \\
&\leq 3R\mathtt{M}pm_n^{-1} \sum_{k=1}^{m_n} \left\{A_2\rho^{\gamma_n k} V_R(X_{m_n}^n)\right\} + 3R\mathtt{M}p\mathbf{\Psi}(\gamma_n) \leq \frac{3R\mathtt{M}pA_2\rho^{-\bar{\gamma}}V_R(X_{m_n}^n)}{\log(1/\rho)\gamma_n m_n} + 3R\mathtt{M}p\mathbf{\Psi}(\gamma_n) \ .
\end{aligned}$$

In a similar manner, we have

$$\left\|\mathbb{E}\left[\eta_0 \middle| X_0^0\right]\right\| \leq \frac{3R\mathtt{M}pA_2\rho^{-\bar{\gamma}}V_R(X_0^0)}{\log(1/\rho)\gamma_0 m_0} + 3R\mathtt{M}p\mathbf{\Psi}(\gamma_0) \ .$$

We conclude using **H**1-(i). □

Lemma 5.2.13. *Assume* **A**2(1), **B**1, **B**2(1) *and* **H**1. *Then we have for any $n \in \mathbb{N}$, $\mathbb{E}\left[\|\eta_n\|^2\right] \leq B_2$, with*

$$B_2 = 2A_1(\|F(0)\| + 3R\mathtt{M})^2\mathbb{E}\left[V_R^2(X_0^0)\right] + 2\sup_{\theta\in\mathsf{K}}\|\nabla L(\theta)\|^2 \ .$$

Proof. Using that $\|x+y\|^2 \leq 2(\|x\|^2 + \|y\|^2)$ for any $x, y \in \mathbb{R}^d$, the Cauchy-Schwarz inequality **H**1-(i) and Proposition 5.1.4, we get for any $n \in \mathbb{N}$,

$$\mathbb{E}\left[\|\eta_n\|^2\right] \leq 2m_n^{-1} \sum_{k=1}^{m_n} \|F(X_k^n)\|^2 + 2\|\nabla L(\theta_n)\|^2 \ .$$

We conclude using that for any $x \in \mathbb{R}^d$, $\|F(x)\| \leq (\|F(0)\| + 3R\mathtt{M})V_R(x)$ and the fact that $\sup_{\theta \in \mathsf{K}} \|\nabla L(\theta)\| < +\infty$. □

Proof of Theorem 5.2.10. Taking the expectation in [AFM17, Theorem 3, Equation (8)], using the Cauchy-Schwarz inequality and the fact that $(\eta_n^{(2)})_{n\in\mathbb{N}}$ is a martingale increment with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, we get that for every $n \in \mathbb{N}$

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{n} \delta_k \left\{L(\theta_k) - \min_{\mathsf{K}} L\right\}\right] \\
&\quad \leq \mathbb{E}\left[2M_\Theta^2 - \sum_{k=0}^{n-1} \delta_{k+1} \langle \Pi_{\mathsf{K}}(\theta_k - \delta_{k+1}\nabla L(\theta_k)) - \theta^\star, \eta_k \rangle + \sum_{k=0}^{n-1} \delta_{k+1}^2 \|\eta_k\|^2\right] \\
&\quad \leq 2M_\Theta^2 + 2M_\Theta \sum_{k=0}^{n-1} \delta_{k+1} \mathbb{E}\left[\|\eta_k^{(1)}\|\right] + 2\sum_{k=0}^{n-1} \delta_{k+1}^2 \mathbb{E}\left[\|\eta_k\|^2\right] \ .
\end{aligned}$$

Combining this result, Lemma 5.2.12 and Lemma 5.2.13 completes the proof. □

Fixed batch size In this section, we give an alternative result to Theorem 4.2.4, in the case where $m_n = 1$ and $\gamma_n = \gamma_0$ for all $n \in \mathbb{N}$. We consider the following additional assumption on the family of kernels $\{\mathrm{K}_{\gamma,\theta} \ : \ \theta \in \mathsf{K}, \gamma \in (0, \bar{\gamma}]\}$.

H2. *There exists* $\mathbf{\Lambda} : \mathbb{R}_+^* \to \mathbb{R}_+$ *such that for any* $\gamma \in (0, \bar{\gamma}]$, $\theta_1, \theta_2 \in \mathsf{K}$, $x \in \mathbb{R}^d$

$$\|\delta_x \mathrm{K}_{\gamma,\theta_1} - \delta_x \mathrm{K}_{\gamma,\theta_2}\|_{V_R} \leq \mathbf{\Lambda}(\gamma)\|\theta_1 - \theta_2\| V_R^2(x) \ .$$

We also recall the following assumption.

B5. $F \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R}^p)$ *and there exists* $\mathtt{B} \geq 0$ *such that for any* $x, y \in \mathbb{R}^d$

$$\|\mathrm{d}F(x) - \mathrm{d}F(y)\| \leq \mathtt{B}\,\|x - y\| \ .$$

Theorem 5.2.14. *Assume* **A**2(1), **B**1, **B**2(1), **B**3, **B**5, **H**1 *and* **H**2. *Let* $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ *be sequences of non-increasing positive real numbers and* $(m_n)_{n\in\mathbb{N}}$ *a sequence of positive integers satisfying* $\delta_n < 1/(\sup_{\theta\in\mathsf{K}} \|\nabla^2 L(\theta)\|)$ *and for any* $n \in \mathbb{N}$, $\gamma_n = \gamma < \bar{\gamma}$, $m_n = m_0$ *and* $\sup_{n\in\mathbb{N}} |\delta_{n+1} - \delta_n| \, \delta_n^{-2} < +\infty$. *Then, there exists* $(\tilde{E}_n)_{n\in\mathbb{N}}$ *such that for any* $n \in \mathbb{N}^\star$

$$\mathbb{E}\left[\left\{\sum_{k=1}^{n} \delta_k L(\theta_k) \Big/ \sum_{k=1}^{n} \delta_k\right\} - \min_{\mathsf{K}} L\right] \leq \tilde{E}_n \Big/ \left(\sum_{k=1}^{n} \delta_k\right) \ ,$$

with $(\theta_k)_{k\in\mathbb{N}}$ *and* L *are defined in* (5.19) *and* (5.3) *respectively, and*

$$E_n = D\left\{1 + \sum_{k=0}^{n-1} \delta_{k+1}^2/\gamma + \sum_{k=0}^{n-1} \delta_{k+1}^2 \mathbf{\Lambda}(\gamma)/\gamma^2 + \sum_{k=0}^{n-1} \delta_{k+1} \mathbf{\Psi}(\gamma) + \delta_{n+1}/\gamma\right\} \ ,$$

and $D = 2M_\Theta^2 + 6M_\Theta R\mathtt{M}p + 2M_\Theta \tilde{B}_2 + 3\tilde{B}_1 + B_2$ *where* $\tilde{B}_2$ *is given in Lemma 5.2.13,* $\tilde{B}_1$ *in Lemma 5.2.15 and* $\tilde{B}_2$ *in Lemma 5.2.18.*

The proof of this result is an adaptation to the one of Theorem 4.2.4. The main difference in the proof consists in a refinement of Lemma 5.2.18 which can be established in our setting and is given Lemma 5.2.18.

Similarly to the proof of Theorem 5.2.10, we need to analyze the error η_n for $n \in \mathbb{N}$ defined by (5.39), but the decomposition (5.40) has to be improved. For that purpose, we introduce *Poisson solutions* associated with F. Under **H**1 for any $\theta \in \mathsf{K}$ and $\gamma \in (0, \bar{\gamma}]$, consider $\hat{F}_{\gamma,\theta} : \mathbb{R}^d \to \mathbb{R}^p$ solution of the *Poisson equation*,

$$(\mathrm{Id} - \mathrm{K}_{\gamma,\theta})\hat{F}_{\gamma,\theta} = F - \pi_{\gamma,\theta}(F) \; . \tag{5.42}$$

Note that by **H**1-(ii), $\hat{F}_{\gamma,\theta}$ is well defined and is given for any $x \in \mathbb{R}^d$ by

$$\hat{F}_{\gamma,\theta}(x) = \sum_{j\in\mathbb{N}} \{\mathrm{K}^j_{\gamma,\theta} F(x) - \pi_{\gamma,\theta}(F)\} \; . \tag{5.43}$$

In addition, by Lemma 5.2.11 and **H**1-(ii), we have for any $\theta \in \mathsf{K}$ and $x \in \mathbb{R}^d$

$$\begin{aligned} \left\|\hat{F}_{\gamma,\theta}(x)\right\| &\le \left\|\sum_{j\in\mathbb{N}} \mathrm{K}^j_{\gamma,\theta} F(x) - \pi_{\gamma,\theta}(F)\right\| \le 3R\mathtt{M}p \sum_{j\in\mathbb{N}} \mathrm{d}_{W_R}(\mathrm{K}^j_{\gamma,\theta}, \pi_{\gamma,\theta}) \\ &\le 3R\mathtt{M}pA_2 \sum_{j\in\mathbb{N}} \rho^{\gamma j} V_R(x) \le 3R\mathtt{M}pA_2 \log^{-1}(1/\rho)\rho^{-\bar{\gamma}}\gamma^{-1} V_R(x) \le C_F \gamma^{-1} V_R(x) \; , \end{aligned} \tag{5.44}$$

with $C_F = 3R\mathtt{M}pA_2 \log^{-1}(1/\rho)\rho^{-\bar{\gamma}}$. We now denote for any $n \in \mathbb{N}$, $\tilde{X}_{n+1} = X^n_1$ and therefore η_n defined by (5.39) is given for any $n \in \mathbb{N}$ by $\eta_n = F(\tilde{X}_{n+1}) - \pi_{\theta_n}(F)$. Using (5.42) an alternative expression of $(\eta_n)_{n\in\mathbb{N}}$ is given for any $n \in \mathbb{N}$ by

$$\eta_n = \hat{F}_{\gamma,\theta_n}(\tilde{X}_{n+1}) - \mathrm{K}_{\gamma,\theta_n}\hat{F}_{\gamma,\theta_n}(\tilde{X}_{n+1}) + \pi_{\gamma,\theta_n}(F) - \pi_{\theta_n}(F) = \eta^{(a)}_n + \eta^{(b)}_n + \eta^{(c)}_n + \eta^{(d)}_n \; ,$$

where

$$\begin{aligned} \eta^{(a)}_n &= \hat{F}_{\gamma,\theta_n}(\tilde{X}_{n+1}) - \mathrm{K}_{\gamma,\theta_n}\hat{F}_{\gamma,\theta_n}(\tilde{X}_n) \; , \\ \eta^{(b)}_n &= \mathrm{K}_{\gamma,\theta_n}\hat{F}_{\gamma,\theta_n}(\tilde{X}_n) - \mathrm{K}_{\gamma,\theta_{n+1}}\hat{F}_{\gamma,\theta_{n+1}}(\tilde{X}_{n+1}) \; , \\ \eta^{(c)}_n &= \mathrm{K}_{\gamma,\theta_{n+1}}\hat{F}_{\gamma,\theta_{n+1}}(\tilde{X}_{n+1}) - \mathrm{K}_{\gamma,\theta_n}\hat{F}_{\gamma,\theta_n}(\tilde{X}_{n+1}) \; , \\ \eta^{(d)}_n &= \pi_{\gamma,\theta_n}(F) - \pi_{\theta_n}(F) \; . \end{aligned} \tag{5.45}$$

In the next results, we analyze each term in this decomposition separately, except for $(\eta^{(a)}_n)_{n\in\mathbb{N}}$ which is a sequence of martingale increments with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.

Lemma 5.2.15. *Assume* **A**2(1), **B**1, **B**2(1) *and* **H**1. *Then, for any* $n \in \mathbb{N}$

$$\begin{aligned} &\mathbb{E}\left[\left\|\sum_{k=0}^{n} \delta_{k+1}\langle a_{k+1}, \eta^{(b)}_k\rangle\right\|\right] \\ &\qquad \le \tilde{B}_1 \left[\sum_{k=0}^{n} |\delta_{k+1} - \delta_k|\,\gamma^{-1} + \sum_{k=0}^{n} \delta_k^2 \gamma^{-1} + (\delta_{n+1}/\gamma - \delta_1/\gamma)\right] \; . \end{aligned}$$

with $(\eta^{(b)}_n)_{n\in\mathbb{N}}$ *defined in* (5.45), $a_{k+1} = \Pi_{\mathsf{K}}\left[\theta_k - \delta_{k+1}\nabla L(\theta_k)\right] - \theta^\star$, $\theta^\star \in \arg\min_{\mathsf{K}} L$ *and*

$$\begin{aligned} \tilde{B}_1 &= A_1 C_F (2M_\Theta + \sup_{\theta\in\mathsf{K}} \|\nabla L(\theta)\|)\mathbb{E}\left[V_R(\tilde{X}_0)\right] \\ &\quad + A_1 C_F (1 + \delta_1 \sup_{\theta\in\mathsf{K}} \|\nabla^2 L(\theta)\|)(\|F(0)\| + R\mathtt{M})\mathbb{E}\left[V^2_R(\tilde{X}_0)\right] + 4A_1 C_F M_\Theta \mathbb{E}\left[V_R(\tilde{X}_0)\right] \; . \end{aligned}$$

Proof. By (5.45) we have for any $n \in \mathbb{N}$ and $\theta \in \mathsf{K}$

$$\sum_{k=0}^{n} \delta_{k+1} \langle a_{k+1}, \eta_k^{(b)} \rangle = \sum_{k=1}^{n} \langle \delta_{k+1} a_{k+1} - \delta_k a_k, \mathrm{K}_{\gamma,\theta_k} \hat{F}_{\gamma,\theta_k}(\tilde{X}_k) \rangle \\ - \langle \delta_{n+1} a_{n+1}, \mathrm{K}_{\gamma,\theta_{n+1}} \hat{F}_{\gamma,\theta_{n+1}}(\tilde{X}_{n+1}) \rangle + \langle \delta_1 a_1, \mathrm{K}_{\gamma,\theta_0} \hat{F}_{\gamma,\theta_0}(\tilde{X}_0) \rangle \ . \tag{5.46}$$

In addition, using Proposition 5.1.4, we have for any $n \in \mathbb{N}$ and $\theta \in \mathsf{K}$,

$$\begin{aligned} &\|\delta_{n+1} a_{n+1} - \delta_n a_n\| \le 2M_\Theta \, |\delta_{n+1} - \delta_n| + \delta_{n+1} \, \|a_{n+1} - a_n\| \\ &\le 2M_\Theta \, |\delta_{n+1} - \delta_n| + \delta_{n+1} (1 + \delta_n \sup_{\theta \in \mathsf{K}} \|\nabla^2 L(\theta)\|) \, \|\theta_n - \theta_{n-1}\| + \delta_{n+1} \, |\delta_{n+1} - \delta_n| \, \|\nabla L(\theta_n)\| \\ &\le (2M_\Theta + \delta_1 \sup_{\theta \in \mathsf{K}} \|\nabla L(\theta)\|) \, |\delta_{n+1} - \delta_n| \\ &\qquad + \delta_n^2 (1 + \delta_n \sup_{\theta \in \mathsf{K}} \|\nabla^2 L(\theta)\|)(\|F(0)\| + 3R\mathtt{M}) V_R(\tilde{X}_{n+1}) \ , \end{aligned} \tag{5.47}$$

where we have used in the last inequality that Π_{K} is non-expansive, **B**2(1) and **H**1-(i) and Proposition 5.1.4 again. Combining (5.46), (5.47), (5.44), the Cauchy-Schwarz inequality and **H**1-(i) we get that

$$\begin{aligned} \mathbb{E}\left[\left\| \sum_{k=0}^{n} \delta_{k+1} \langle a_k, \eta_k^{(b)} \rangle \right\| \right] &\le (2M_\Theta + \delta_1 \sup_{\theta \in \mathsf{K}} \|\nabla L(\theta)\|) A_1 C_F \mathbb{E}\left[V_R(\tilde{X}_0) \right] \sum_{k=0}^{n} |\delta_{k+1} - \delta_k| \, \gamma^{-1} \\ &\quad + A_1 C_F (\|F(0)\| + 3R\mathtt{M})(1 + \delta_1 \sup_{\theta \in \mathsf{K}} \|\nabla^2 L(\theta)\|) \mathbb{E}\left[V_R^2(\tilde{X}_0) \right] \sum_{k=0}^{n} \delta_k^2 \gamma^{-1} \\ &\quad + 2A_1 M_\Theta C_F \mathbb{E}\left[V_R(\tilde{X}_0) \right] \{ \delta_{n+1}/\gamma + \delta_1/\gamma \} \ , \end{aligned}$$

which concludes the proof of Lemma 5.2.15. □

We now upper bound $\mathbb{E}\left[\|\eta_n^{(c)}\| \right]$ for $n \in \mathbb{N}$ using the two following lemmas.

Lemma 5.2.16. *Assume* **B**1, **B**2(1), **H**1 *and* **H**2. *Then, for any* $\gamma \in (0, \bar{\gamma}]$, $\theta_1, \theta_2 \in \mathsf{K}$ *and* $x \in \mathbb{R}^d$

$$\mathrm{d}_{W_R}(\pi_{\gamma,\theta_1}, \pi_{\gamma,\theta_2}) \le A_1 A_2 \rho^{-\bar{\gamma}} \log^{-1}(1/\rho) \mathbf{\Lambda}(\gamma) \|\theta_1 - \theta_2\| V_R^2(x) \gamma^{-1} \ .$$

Proof. Let $\gamma \in (0, \bar{\gamma}]$, $\theta_1, \theta_2 \in \mathsf{K}$, $\ell \in \mathbb{N}^\star$, $j \in \mathbb{N}$ with $\ell \ge j + 1$ and $g : \mathbb{R}^d \to \mathbb{R}$ measurable such that for any $y, z \in \mathbb{R}^d$, $|g(y) - g(z)| \le W_R(y, z)$. Using **H**1-(ii) we have

$$\left| \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} g(x) - \pi_{\gamma,\theta_2}(g) \right| \le A_2 \rho^{(\ell-1-j)\gamma} V_R(x) \ .$$

Combining this result and **H**2 we have that

$$\left| (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} g(x) \right| \le A_2 \rho^{\gamma(\ell-1-j)} \mathbf{\Lambda}(\gamma) \|\theta_1 - \theta_2\| V_R^2(x) \ . \tag{5.48}$$

Using **H**1-(i) in (5.48), we get

$$\left| \mathrm{K}_{\gamma,\theta_1}^{j} (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} g(x) \right| \le A_1 A_2 \rho^{\gamma(\ell-1-j)} \mathbf{\Lambda}(\gamma) \|\theta_1 - \theta_2\| V_R^2(x) \ .$$

Combining this result and the triangular inequality we obtain

$$\left| \mathrm{K}_{\gamma,\theta_1}^{\ell} g(x) - \mathrm{K}_{\gamma,\theta_2}^{\ell} g(x) \right| \le \sum_{j=0}^{\ell-1} \left| \mathrm{K}_{\gamma,\theta_1}^{j+1} \mathrm{K}_{\gamma,\theta_2}^{\ell-j-1} g(x) - \mathrm{K}_{\gamma,\theta_1}^{j} \mathrm{K}_{\gamma,\theta_2}^{\ell-j} g(x) \right|$$

$$
\begin{aligned}
&\leq \sum_{j=0}^{\ell-1} \left| \mathrm{K}_{\gamma,\theta_1}^{j} (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} g(x) \right| \\
&\leq A_1 A_2 \mathbf{\Lambda}(\gamma) \|\theta_1 - \theta_2\| V_R^2(x) \sum_{j=0}^{\ell-1} \rho^{\gamma(\ell-1-j)} \\
&\leq A_1 A_2 \rho^{-\bar{\gamma}} \log^{-1}(1/\rho) \mathbf{\Lambda}(\gamma) \|\theta_1 - \theta_2\| V_R^2(x) \gamma^{-1} \, .
\end{aligned}
$$

Taking the limit $\ell \to +\infty$ and using **H**1-(ii) concludes the proof. □

Lemma 5.2.17. *Assume* **B**1, **B**2(1), **B**5 *and* **H**1. *Then, for any* $\gamma \in (0, \bar{\gamma}]$, $\theta_1, \theta_2 \in \mathsf{K}$, $\ell \in \mathbb{N}^\star$, $j \in \mathbb{N}$ *with* $\ell \geq j+1$ *and* $x \in \mathbb{R}^d$

$$
\left\| \left\{ \mathrm{K}_{\gamma,\theta_1}^{j} - \pi_{\gamma,\theta_1} \right\} (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \left\{ \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} F(x) - \pi_{\gamma,\theta_2}(F) \right\} \right\| \leq D_F \, \|\theta_1 - \theta_2\| \, V_R(x) \gamma \rho^{\gamma \ell} \, ,
$$

with $D_F = A_2^2 \mathtt{BM}(1 + 2M_\Theta \mathtt{B}\bar{\gamma})R + 2\mathtt{M}^2$.

Proof. Let $\gamma \in (0, \bar{\gamma}]$, $\theta_1, \theta_2 \in \mathsf{K}$, $\ell \in \mathbb{N}^\star$, $j \in \mathbb{N}$ with $\ell \geq j+1$ and $x, y \in \mathbb{R}^d$. First, we have

$$
\begin{aligned}
&(\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} \{F(x) - \pi_{\gamma,\theta_2}(F)\} - (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} \{F(x) - \pi_{\gamma,\theta_2}(F)\} \\
&\quad = (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} F(x) - (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} F(y) \\
&\quad = \mathrm{K}_{\gamma,\theta_1} \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} F(x) - \mathrm{K}_{\gamma,\theta_2}^{\ell-j} F(x) - \mathrm{K}_{\gamma,\theta_1} \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} F(y) + \mathrm{K}_{\gamma,\theta_2}^{\ell-j} F(y) \\
&\quad = \mathrm{K}_{\gamma,\theta_2}^{\ell-j} (F(x + \Delta_\gamma(x)) - F(x) - F(y + \Delta_\gamma(y)) + F(y)) \\
&\quad = \mathrm{K}_{\gamma,\theta_2}^{\ell-j} G(x) - \mathrm{K}_{\gamma,\theta_2}^{\ell-j} G(y) \, ,
\end{aligned} \tag{5.49}
$$

with $\Delta_\gamma(x) = \gamma(\nabla_x U(\theta_1, x) - \nabla_x U(\theta_2, x)) = \gamma \sum_{i=1}^p (\theta_1^i - \theta_2^i) \nabla F_i(x)$ and $G : \mathbb{R}^d \to \mathbb{R}^p$ defined for any $z \in \mathbb{R}^d$ by

$$
G(z) = F(z + \Delta_\gamma(z)) - F(z) \, . \tag{5.50}
$$

Using **B**2(1) and **B**5 we have that for any $x, y \in \mathbb{R}^d$,

$$
\|\Delta_\gamma(x)\| \leq \mathtt{M}\gamma \, \|\theta_1 - \theta_2\| \, , \qquad \|\Delta_\gamma(x) - \Delta_\gamma(y)\| \leq M_\Theta \mathtt{B}\gamma \, \|x - y\| \, . \tag{5.51}
$$

Using (5.50), (5.51), we have for any $x, y \in \mathbb{R}^d$ with $x \neq y$

$$
\begin{aligned}
\|G(x) - G(y)\| &= \left\| \int_0^1 \{ \mathrm{d}F(x + t\Delta_\gamma(x))(\Delta_\gamma(x)) - \mathrm{d}F(y + t\Delta_\gamma(y))(\Delta_\gamma(y)) \} \, \mathrm{d}t \right\| \\
&\leq \int_0^1 \|\mathrm{d}F(x + t\Delta_\gamma(x)) - \mathrm{d}F(y + t\Delta_\gamma(y))\| \, \|\Delta_\gamma(x)\| \, \mathrm{d}t \\
&\quad + \int_0^1 \|\mathrm{d}F(y + t\Delta_\gamma(y))\| \, (\|\Delta_\gamma(x)\| + \|\Delta_\gamma(y)\|) \mathrm{d}t \\
&\leq \mathtt{B}(\|x - y\| + \|\Delta_\gamma(x) - \Delta_\gamma(y)\|) \, \|\Delta_\gamma(x)\| + 2\mathtt{M}^2 \gamma \, \|\theta_1 - \theta_2\| \\
&\leq \mathtt{BM}(1 + 2M_\Theta \mathtt{B}\bar{\gamma}) \gamma \, \|\theta_1 - \theta_2\| \, \|x - y\| + 2\mathtt{M}^2 \gamma \, \|\theta_1 - \theta_2\| \\
&\leq \tilde{D}_F \gamma \, \|\theta_1 - \theta_2\| \, (1 + \|x - y\| / R) \, ,
\end{aligned}
$$

with $\tilde{D}_F = \mathtt{BM}(1 + 2M_\Theta \mathtt{B}\bar{\gamma})R + 2\mathtt{M}^2$. Therefore, for any $x, y \in \mathbb{R}^d$,

$$
\|G(x) - G(y)\| \leq \tilde{D}_F \gamma \, \|\theta_1 - \theta_2\| \, W_R(x, y) \, . \tag{5.52}
$$

Combining (5.49), (5.52) and **H**1-(ii) we obtain that

$$\left\|(\mathrm{K}_{\gamma,\theta_1}-\mathrm{K}_{\gamma,\theta_2})\mathrm{K}_{\gamma,\theta_2}^{\ell-1-j}\left\{F(x)-\pi_{\gamma,\theta_2}(F)\right\}-(\mathrm{K}_{\gamma,\theta_1}-\mathrm{K}_{\gamma,\theta_2})\mathrm{K}_{\gamma,\theta_2}^{\ell-1-j}\left\{F(x)-\pi_{\gamma,\theta_2}(F)\right\}\right\| \\ \leq A_2\tilde{D}_F\left\|\theta_1-\theta_2\right\|\gamma\rho^{\gamma(\ell-j)}W_R(x,y)\ .$$

Therefore, using **H**1-(ii) we get

$$\left\|\left\{\mathrm{K}_{\gamma,\theta_1}^{j}-\pi_{\gamma,\theta_1}\right\}(\mathrm{K}_{\gamma,\theta_1}-\mathrm{K}_{\gamma,\theta_2})\left\{\mathrm{K}_{\gamma,\theta_2}^{\ell-1-j}F(x)-\pi_{\gamma,\theta_2}(F)\right\}\right\| \leq A_2^2\tilde{D}_F\left\|\theta_1-\theta_2\right\|V_R(x)\gamma\rho^{\gamma\ell}\ ,$$

which concludes the proof. □

Lemma 5.2.18. *Assume* **B**1, **B**2(1), **B**5, **H**1 *and* **H**2. *Then we have for any* $n\in\mathbb{N}$

$$\mathbb{E}\left[\left\|\eta_n^{(c)}\right\|\right]\leq\tilde{B}_2\delta_{n+1}\gamma^{-2}\left(\mathbf{\Lambda}(\gamma)+\gamma\right)\ ,$$

with

$$\tilde{B}_2=A_1(\|F(0)\|+3R\mathtt{M})\rho^{-2\bar{\gamma}}\log^{-2}(1/\rho)\max\left\{D_F,E_F\right\}\ , \tag{5.53}$$

with $(\eta_n^{(c)})_{n\in\mathbb{N}}$ *defined in* (5.45), $E_F=3R\mathtt{M}pA_1A_2^2$ *and* D_F *in Lemma 5.2.17.*

Proof. We first give an upper bound on $\left\|\mathrm{K}_{\gamma,\theta_1}\hat{H}_{\gamma,\theta_1}(x)-\mathrm{K}_{\gamma,\theta_2}\hat{H}_{\gamma,\theta_2}(x)\right\|$ for any $\theta_1,\theta_2\in\mathsf{K}$, $\gamma\in(0,\bar{\gamma}]$ and $x\in\mathbb{R}^d$. By (5.43) we have for any $\theta_1,\theta_2\in\mathsf{K}$, $\gamma\in(0,\bar{\gamma}]$ and $x\in\mathbb{R}^d$,

$$\begin{aligned}&\left\|\mathrm{K}_{\gamma,\theta_1}\hat{H}_{\gamma,\theta_1}(x)-\mathrm{K}_{\gamma,\theta_2}\hat{H}_{\gamma,\theta_2}(x)\right\| \\ &=\left\|\sum_{\ell\in\mathbb{N}^\star}\left\{\mathrm{K}_{\gamma,\theta_1}^{\ell}F(x)-\pi_{\gamma,\theta_1}(F)\right\}-\sum_{\ell\in\mathbb{N}^\star}\left\{\mathrm{K}_{\gamma,\theta_2}^{\ell}F(x)-\pi_{\gamma,\theta_2}(F)\right\}\right\| \\ &\leq\sum_{\ell\in\mathbb{N}^\star}\left\|\mathrm{K}_{\gamma,\theta_1}^{\ell}F(x)-\pi_{\gamma,\theta_1}(F)-\mathrm{K}_{\gamma,\theta_2}^{\ell}F(x)-\pi_{\gamma,\theta_2}(F)\right\|\ .\end{aligned} \tag{5.54}$$

We now bound each term of the series in the right hand side. For any measurable functions g_1,g_2 with $g_i:\mathbb{R}^d\to\mathbb{R}^p$ and such that $\sup_{x\in\mathbb{R}^d}\left\|g_i(x)\right\|/V_R(x)<+\infty$ with $i\in\{1,2\}$, $\theta_1,\theta_2\in\mathsf{K}$, $\gamma\in(0,\bar{\gamma}]$, $x\in\mathbb{R}^d$ and $\ell\in\mathbb{N}^\star$, using that π_{γ,θ_1} is invariant for $\mathrm{K}_{\gamma,\theta_1}$, it holds that

$$\begin{aligned}\mathrm{K}_{\gamma,\theta_1}^{\ell}g_1(x)-\mathrm{K}_{\gamma,\theta_2}^{\ell}g_2(x)&=\mathrm{K}_{\gamma,\theta_1}^{\ell}g_1(x)-\mathrm{K}_{\gamma,\theta_2}^{\ell}g_1(x)+\mathrm{K}_{\gamma,\theta_2}^{\ell}(g_1(x)-g_2(x)) \\ &=\sum_{j=0}^{\ell-1}\left\{\mathrm{K}_{\gamma,\theta_1}^{j}-\pi_{\gamma,\theta_1}\right\}(\mathrm{K}_{\gamma,\theta_1}-\mathrm{K}_{\gamma,\theta_2})\left\{\mathrm{K}_{\gamma,\theta_2}^{\ell-1-j}g_1(x)-\pi_{\gamma,\theta_2}(g_1)\right\} \\ &\quad+\sum_{j=0}^{\ell-1}\pi_{\gamma,\theta_1}\left\{\mathrm{K}_{\gamma,\theta_2}^{\ell-1-j}g_1(x)-\mathrm{K}_{\gamma,\theta_2}^{\ell-j}g_1(x)\right\}+\mathrm{K}_{\gamma,\theta_2}^{\ell}(g_1(x)-g_2(x)) \\ &=\sum_{j=0}^{\ell-1}\left\{\mathrm{K}_{\gamma,\theta_1}^{j}-\pi_{\gamma,\theta_1}\right\}(\mathrm{K}_{\gamma,\theta_1}-\mathrm{K}_{\gamma,\theta_2})\left\{\mathrm{K}_{\gamma,\theta_2}^{\ell-1-j}g_1(x)-\pi_{\gamma,\theta_2}(g_1)\right\} \\ &\quad-\pi_{\gamma,\theta_1}(\mathrm{K}_{\gamma,\theta_2}^{\ell}g_1(x)-g_1(x))+\mathrm{K}_{\gamma,\theta_2}^{\ell}(g_1(x)-g_2(x))\ .\end{aligned}$$

Setting $g_1=F-\pi_{\gamma,\theta_1}(F)$ and $g_2=F-\pi_{\gamma,\theta_2}(F)$, we obtain that

$$\mathrm{K}_{\gamma,\theta_1}^{\ell}F(x)-\pi_{\gamma,\theta_1}(F)-\mathrm{K}_{\gamma,\theta_2}^{\ell}F(x)-\pi_{\gamma,\theta_2}(F)$$

$$= \sum_{j=0}^{\ell-1} \left\{ \mathrm{K}_{\gamma,\theta_1}^j - \pi_{\gamma,\theta_1} \right\} (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \left\{ \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} F(x) - \pi_{\gamma,\theta_2}(F) \right\} + \Xi_\ell \;, \tag{5.55}$$

where, using that π_{γ,θ_2} is invariant for $\mathrm{K}_{\gamma,\theta_2}$, we have

$$\begin{aligned}\Xi_\ell &= -\pi_{\gamma,\theta_1}(\mathrm{K}_{\gamma,\theta_2}^\ell F(x) - F(x)) + \mathrm{K}_{\gamma,\theta_2}^\ell \left[\pi_{\gamma,\theta_2}(F) - \pi_{\gamma,\theta_1}(F)\right] \\ &= (\pi_{\gamma,\theta_2} - \pi_{\gamma,\theta_1}) \mathrm{K}_{\gamma,\theta_2}^\ell F(x) \;.\end{aligned}$$

Using Lemma 5.2.17 we obtain for any $\theta_1, \theta_2 \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^\star$

$$\begin{aligned}&\left\| \sum_{j=0}^{\ell-1} \left\{ \mathrm{K}_{\gamma,\theta_1}^j - \pi_{\gamma,\theta_1} \right\} (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \left\{ \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j} F(x) - \pi_{\gamma,\theta_2}(F) \right\} \right\| \\ &\quad \le \sum_{j=0}^{\ell-1} \left\| \left\{ \mathrm{K}_{\gamma,\theta_1}^j - \pi_{\gamma,\theta_1} \right\} (\mathrm{K}_{\gamma,\theta_1} - \mathrm{K}_{\gamma,\theta_2}) \left\{ \mathrm{K}_{\gamma,\theta_2}^{\ell-1-j}(F) - \pi_{\gamma,\theta_2}(F) \right\} \right\| \\ &\quad \le \sum_{j=0}^{\ell-1} D_F \gamma \left\| \theta_1 - \theta_2 \right\| \rho^{\gamma\ell} V_R(x) \le D_F \gamma V_R(x) \left\| \theta_1 - \theta_2 \right\| \ell \rho^{\gamma\ell} \;.\end{aligned} \tag{5.56}$$

Using **H**1-(ii), Lemma 5.2.11 and Lemma 5.2.16, we obtain for any $\theta_1, \theta_2 \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$ and $\ell \in \mathbb{N}^\star$

$$\begin{aligned}\left\| (\pi_{\gamma,\theta_1} - \pi_{\gamma,\theta_2}) \mathrm{K}_{\gamma,\theta_2}^\ell F(x) \right\| &\le 3R\mathtt{M}pA_2 \rho^{\gamma\ell} \mathrm{d}_{W_R}(\pi_{\gamma,\theta_1}, \pi_{\gamma,\theta_2}) \\ &\le E_F \rho^{-\bar{\gamma}} \log^{-1}(1/\rho) \mathbf{\Lambda}(\gamma) \|\theta_1 - \theta_2\| V_R^2(x) \gamma^{-1} \rho^{\gamma\ell} \;.\end{aligned} \tag{5.57}$$

with $E_F = 3R\mathtt{M}pA_1A_2^2$. Combining (5.56) and (5.57) in (5.55), we obtain that for any $\theta_1, \theta_2 \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$ that

$$\begin{aligned}&\mathrm{K}_{\gamma,\theta_1}^\ell F(x) - \pi_{\gamma,\theta_1}(F) - \mathrm{K}_{\gamma,\theta_2}^\ell F(x) - \pi_{\gamma,\theta_2}(F) \\ &\quad \le D_F \left\| \theta_1 - \theta_2 \right\| V_R(x) \gamma \ell \rho^{\gamma\ell} + E_F \rho^{-\bar{\gamma}} \log^{-1}(1/\rho) \|\theta_1 - \theta_2\| V_R^2(x) \mathbf{\Lambda}(\gamma) \gamma^{-1} \rho^{\gamma\ell} \;.\end{aligned}$$

Using this result in (5.54) and that for any $t \in (-1, 1)$ and $a > 0$, $\sum_{k\in\mathbb{N}} k t^{ak} = t(1-t^a)^{-2} \le a^{-2} t^{-a} \log^{-2}(1/t)$, we get that

$$\begin{aligned}\left\| \mathrm{K}_{\gamma,\theta_1} \hat{H}_{\gamma,\theta_1}(x) - \mathrm{K}_{\gamma,\theta_2} \hat{H}_{\gamma,\theta_2}(x) \right\| &\le D_F \rho^{-2\bar{\gamma}} \log^{-2}(1/\rho) \left\| \theta_1 - \theta_2 \right\| V_R(x) \gamma^{-1} \\ &\quad + E_F \rho^{-2\bar{\gamma}} \log^{-2}(1/\rho) \|\theta_1 - \theta_2\| V_R^2(x) \mathbf{\Lambda}(\gamma) \|\gamma^{-2} \\ &\le C_c \, \gamma^{-2} \left(\mathbf{\Lambda}(\gamma) \|\theta_1 - \theta_2\| + \gamma \|\theta_1 - \theta_2\| \right) V_R^2(x) \;,\end{aligned} \tag{5.58}$$

with $C_c = \rho^{-2\bar{\gamma}} \log^{-2}(1/\rho) \max\{D_F, E_F\}$. Note that for any $k \in \mathbb{N}$, by **B**2(1) and the fact that Π_{K} is non-expansive we have $\|\theta_{k+1} - \theta_k\| \le \delta_{k+1}(\|F(0)\| + 3R\mathtt{M}) V_R(\tilde{X}_{k+1})$. Therefore, plugging this result in (5.58), we get for any $k \in \mathbb{N}$,

$$\begin{aligned}&\left\| \mathrm{K}_{\gamma,\theta_k} \hat{H}_{\gamma,\theta_k}(\tilde{X}_{k+1}) - \mathrm{K}_{\gamma,\theta_{k+1}} \hat{H}_{\gamma,\theta_{k+1}}(\tilde{X}_{k+1}) \right\| \\ &\quad \le C_{c,2} (\|F(0)\| + 3R\mathtt{M}) \delta_{k+1} \gamma^{-2} \left(\mathbf{\Lambda}(\gamma) + \gamma \right) V_R^3(\tilde{X}_{k+1}) \;.\end{aligned} \tag{5.59}$$

Therefore by definition of $(\eta_k^{(c)})_{k\in\mathbb{N}}$, see (5.45), and using **H**1-(i) in (5.59) we get that for any $k \in \mathbb{N}$

$$\mathbb{E}\left[\left\| \eta_k^{(c)} \right\| \right] \le \tilde{B}_2 \delta_{k+1} \gamma^{-2} \left(\mathbf{\Lambda}(\gamma) + \gamma \right) \;,$$

with $\tilde{B}_2$ given by (5.53). □

Lemma 5.2.19. *Assume* **B**1, **B**2(1) *and* **H**1. *Then we have for any* $n \in \mathbb{N}$

$$\mathbb{E}\left[\left\|\eta_n^{(d)}\right\|\right] \leq 3R\mathtt{M}p\mathbf{\Psi}(\gamma) \ ,$$

with $(\eta_n^{(d)})_{n\in\mathbb{N}}$ *defined in* (5.45).

Proof. The proof is a direct consequence of Lemma 5.2.11 and **H**1-(iii). □

Proof of Theorem 5.2.14. Taking the expectation in [AFM17, Theorem 3, Equation (8)], using the Cauchy-Schwarz inequality, the decomposition of the error (5.45) and the fact that $(\eta_n^{(a)})_{n\in\mathbb{N}}$ is a martingale increment with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, we get that for every $n \in \mathbb{N}$

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{n} \delta_k \left\{L(\theta_k) - \min_{\mathsf{K}} L\right\}\right] \\
&\leq \mathbb{E}\left[2M_\Theta^2 - \sum_{k=0}^{n-1} \delta_{k+1}\langle \Pi_{\mathsf{K}}(\theta_k - \delta_{k+1}\nabla L(\theta_k)) - \theta^\star, \eta_k\rangle + \sum_{k=0}^{n-1} \delta_{k+1}^2 \|\eta_k\|^2\right] \\
&\leq 2M_\Theta^2 + 2M_\Theta \sum_{k=0}^{n-1} \delta_{k+1}\mathbb{E}\left[\left\|\eta_k^{(c)}\right\| + \left\|\eta_k^{(d)}\right\|\right] + \mathbb{E}\left[\sum_{k=0}^{n-1} \delta_{k+1}\langle a_{k+1}, \eta_k^{(b)}\rangle\right] + \sum_{k=0}^{n-1} \delta_{k+1}^2 \mathbb{E}\left[\|\eta_k\|^2\right] \ .
\end{aligned}$$

Combining this result, Lemma 5.2.15, Lemma 5.2.18, Lemma 5.2.19 and Lemma 5.2.13 completes the proof. □

Proof of Theorem 5.2.2 In this section, we check that **H**1 and **H**2 are satisfied in order to apply Theorem 5.2.14. More precisely, we study the geometric ergodicity of the Langevin Markov chain under **B**1, **B**2(1) and **B**3 as well as its discretization error. We begin with the following technical lemma

Lemma 5.2.20. *Assume* **B**1, **B**2(1) *and* **B**3. *Let* $\mathtt{m} = \mathtt{m}_1/2, \tilde{\mathtt{L}} = 2\mathtt{L}$, $R = 4\mathcal{M}/\mathtt{m}_1$ *and* $\upsilon = \sup_{\theta\in\mathsf{K}} \|\nabla_x U(\theta, 0)\|$. *In addition, for any* $\theta \in \mathsf{K}$, $\gamma > 0$ *and* $x \in \mathbb{R}^d$, *let*

$$\mathcal{T}_\gamma(x) = \|x - \gamma\nabla_x U(\theta, x)\|^2 \ . \tag{5.60}$$

Then for any $\theta \in \mathsf{K}$ *and* $x, y \in \mathbb{R}^d$

(a) $\|\nabla_x U(\theta, x) - \nabla_x U(\theta, y)\| \leq \tilde{\mathtt{L}}\|x - y\|$;

(b) if $\|x - y\| \geq R$, $\langle \nabla_x U(\theta, x) - \nabla U_x(\theta, y), x - y\rangle \geq \mathtt{m}\|x - y\|^2$,

(c) we have

$$\|\mathcal{T}_\gamma(x)\| \leq (1 + \gamma\tilde{\mathtt{L}})\|x\| + \gamma\upsilon \ ,$$

(d) if $\|x\| \geq \max(R, 2\upsilon/\mathtt{m})$ *and* $\gamma \leq \mathtt{m}/(2\tilde{N}^2)$

$$\|\mathcal{T}_\gamma(x)\| \leq (1 - \gamma\mathtt{m}/2 + \gamma^2\tilde{\mathtt{L}}^2/2)\|x\| \ .$$

Proof. Let $\theta \in \mathsf{K}$. The proof of (a) is straightforward. Let $\mathtt{m} = \mathtt{k}_1/2$ and $x, y \in \mathbb{R}^d$ such that $\|x - y\| \geq R$ with $R = 4\mathcal{M}/\mathtt{k}_1$. Using **B**3-(b)-(c) we have

$$\langle \nabla_x U(\theta, x) - \nabla_x U(\theta, y), x - y\rangle$$

$$
\begin{aligned}
&= \langle \nabla_x U_1(\theta, x) - \nabla_x U_1(\theta, y), x - y \rangle + \langle \nabla_x U_2(\theta, x) - \nabla_x U_2(\theta, y), x - y \rangle \\
&\geq \mathtt{k}_1 \left\| x - y \right\|^2 - 2\mathcal{M} \left\| x - y \right\| \geq (\mathtt{k}_1 - 2\mathcal{M}/R) \left\| x - y \right\|^2 \\
&\geq (\mathtt{k}_1/2) \left\| x - y \right\|^2 \ ,
\end{aligned}
$$

which concludes the proof of (b). For any $x \in \mathbb{R}^d$,

$$
\left\| \mathcal{T}_\gamma(x) \right\| \leq \left\| \mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(0) \right\| + \left\| \mathcal{T}_\gamma(0) \right\| \leq (1 + \gamma \tilde{\mathtt{L}}) \left\| x \right\| + \gamma \upsilon \leq (1 + \gamma(\tilde{\mathtt{L}} + \upsilon))(1 + \left\| x \right\|) - 1 \ ,
$$

and therefore (c) holds. Finally, let $\|x\| \geq \max(R, 2\upsilon/\mathtt{m})$ and $\gamma \leq \mathtt{m}/(2\tilde{N}^2)$. Using that for any $t \geq 0$, $\sqrt{1+t} \leq 1 + t/2$,

$$
\begin{aligned}
\left\| \mathcal{T}_\gamma(x) \right\| &\leq \left\| \mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(0) \right\| + \gamma \left\| \nabla_x U(\theta, 0) \right\| \\
&\leq (1 - \gamma \mathtt{m} + \gamma^2 \tilde{\mathtt{L}}^2/2) \left\| x \right\| + \gamma \upsilon \leq (1 - \gamma \mathtt{m}/2 + \gamma^2 \tilde{\mathtt{L}}^2/2) \left\| x \right\| \ ,
\end{aligned}
$$

which concludes the proof of (d). □

Lemma 5.2.21. *Assume* **B**1, **B**2(1) *and* **B**3. *Let* $\mathtt{m}, \tilde{\mathtt{L}}$ *and* R *be given by Lemma 5.2.20. Then for any* $p \in \mathbb{N}^\star$, $\theta \in \mathsf{K}$ *and* $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < \min(\mathtt{m}/(2\tilde{\mathtt{L}}^2), 1/2)$, $\mathrm{K}_{\gamma,\theta}$ *satisfies* $\mathbf{D}_{\mathrm{d}}(\mathscr{V}_p, \lambda^\gamma, \tilde{b}_p \gamma)$ *with* $\mathscr{V}_p$ *given in* (5.38) *and*

$$
\begin{aligned}
&\lambda = \exp[-\mathtt{m}/4 + \bar{\gamma}\tilde{\mathtt{L}}^2/2] \ , \\
&\tilde{b}_p = \Upsilon_p(2^{2p+1} d^p (1 + \bar{\gamma}\tilde{\mathtt{L}})^{2p-1} \mathbf{\Gamma}(p + 1/2), \mathtt{m}/4) + \mathtt{m}/4 + \mathrm{e}^{\kappa\bar{\gamma}}(\kappa + \log(1/\lambda)) \mathscr{V}_p(\tilde{R}) + C_p(\tilde{R}) \ , \\
&\kappa = \left\{ (1 + \bar{\gamma}\tilde{\mathtt{L}}) \max(\tilde{R}, 1) + \upsilon \right\}^{2p} \ , \\
&\upsilon = \sup_{\theta \in \mathsf{K}} \left\| \nabla_x U(\theta, 0) \right\| \ , \qquad \tilde{R} = \max(R, 2\upsilon/\mathtt{m}) \ , \\
&C_p(\tilde{R}) = 2^{2p+1} d^p \left\{ 1 + \bar{\gamma}(\tilde{\mathtt{L}} + \upsilon) \right\}^{2p-1} \mathbf{\Gamma}(p + 1/2)(1 + \tilde{R})^{2p-1} \ ,
\end{aligned}
\tag{5.61}
$$

where for any $t \geq 0$, $\mathbf{\Gamma}(t) = \int_0^{+\infty} u^{t-1} \mathrm{e}^{-u} \mathrm{d}u$ *and* Υ_p *is given in Lemma 5.2.4. In addition,* $\mathrm{K}_{\gamma,\theta}$ *satisfies* $\mathbf{D}_{\mathrm{d}}(V, \lambda^\gamma, b_p(1 + d^{\varpi_{0,p}})\gamma)$ *with* λ *given in* (5.61) *and* $b_p, \varpi_{0,p} \geq 0$ *independent of the dimension* d.

Proof. First, note that using **B**2(1)-(a) we get

$$
\upsilon \leq \left\| \sum_{i=1}^p \theta_i \nabla F_i(0) + \nabla r(0) \right\| \leq \left\| \nabla r(0) \right\| + p M_\Theta \sup_{i \in \{1, \ldots, p\}} \left\| \nabla F_i(0) \right\| < +\infty \ .
$$

Let $p \in \mathbb{N}^\star$, $\theta \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$. Similarly to Lemma 5.2.5, we obtain that

$$
\begin{aligned}
\int_{\mathbb{R}^d} \left\| y \right\|^{2p} \mathrm{K}_{\gamma,\theta}(x, \mathrm{d}y) &\leq \left\| \mathcal{T}_\gamma(x) \right\|^{2p} + \gamma 2^{2p+1} \gamma d^p \mathbf{\Gamma}(p + 1/2)(1 + \left\| \mathcal{T}_\gamma(x) \right\|)^{2p-1} \\
&\leq \left\| \mathcal{T}_\gamma(x) \right\|^{2p} + \gamma 2^{2p+1} \gamma C_p(x) \ ,
\end{aligned}
\tag{5.62}
$$

with $\mathcal{T}_\gamma$ defined in (5.60) and

$$
C_p(x) = 2^{2p+1} d^p \left\{ 1 + \bar{\gamma}(\tilde{\mathtt{L}} + \upsilon) \right\}^{2p-1} \mathbf{\Gamma}(p + 1/2)(1 + \left\| x \right\|)^{2p-1} \ ,
\tag{5.63}
$$

where we have used Lemma 5.2.20-(c). Let $\tilde{R} = \max(R, 2\upsilon/\mathtt{m})$. We divide the rest of the proof in two parts:

(a) Let $\|x\| \geq \tilde{R}$. We have using Lemma 5.2.20-(d),

$$\|\mathcal{T}_\gamma(x)\| \leq (1 - \gamma\mathtt{m}/2 + \gamma^2\tilde{\mathtt{L}}^2/2)\,\|x\| \ .$$

Hence, $\|\mathcal{T}_\gamma(x)\|^{2p} \leq (1 - \gamma\mathtt{m}/4 + \gamma^2\tilde{\mathtt{L}}^2/2)\,\|x\|^{2p} - \gamma\mathtt{m}\,\|x\|^{2p}/4$ and we have using Lemma 5.2.4 and (5.63) in (5.62)

$$\begin{aligned}
\mathrm{K}_{\gamma,\theta}\mathscr{V}_p(x) &\leq (1 - \gamma\mathtt{m}/4 + \gamma^2\tilde{\mathtt{L}}^2/2)(1 + \|x\|^{2p}) \\
&\quad + \gamma\left[2^{2p+1}d^p\left\{1 + \bar{\gamma}(\tilde{\mathtt{L}} + \upsilon)\right\}^{2p-1}\mathbf{\Gamma}(p + 1/2)(1 + \|x\|)^{2p-1} - \mathtt{m}\,\|x\|^{2p}/4 + \mathtt{m}/4\right] \\
&\leq (1 - \gamma\mathtt{m}/4 + \gamma^2\tilde{\mathtt{L}}^2/2)(1 + \|x\|^{2p}) \\
&\quad + \gamma\left[\Upsilon_p(2^{2p+1}d^p\left\{1 + \bar{\gamma}(\tilde{\mathtt{L}} + \upsilon)\right\}^{2p-1}\mathbf{\Gamma}(p + 1/2, \mathtt{m}/4) + \mathtt{m}/4\right] \ .
\end{aligned}$$

(b) Now assume that $\|x\| \leq \tilde{R}$. Let $\kappa = \{(1 + \tilde{\mathtt{L}})\max(1, \tilde{R}) + \upsilon\}^{2p}$. We have, using that $\gamma \leq 1$,

$$(1 + \gamma\tilde{\mathtt{L}})^{2p} \leq 1 + \gamma\sum_{k=1}^{2p}\binom{2p}{k}\mathtt{L}^k \leq 1 + \gamma(1 + \tilde{\mathtt{L}})^{2p} \leq 1 + \gamma\kappa \ .$$

Combining this result with Lemma 5.2.20-(c) and the fact that $\bar{\gamma} \leq 1$, we get

$$\begin{aligned}
1 + \|\mathcal{T}_\gamma(x)\|^{2p} &\leq 1 + \left[(1 + \gamma\tilde{\mathtt{L}})\,\|x\| + \gamma\upsilon\right]^{2p} \\
&\leq 1 + (1 + \gamma\tilde{\mathtt{L}})^{2p}\,\|x\|^{2p} + \gamma\sum_{k=1}^{2p}\binom{2p}{k}(1 + \bar{\gamma}\tilde{\mathtt{L}})^{2p-k}\tilde{R}^{2p-k}\upsilon^k \\
&\leq 1 + (1 + \gamma\kappa)\,\|x\|^{2p} + \gamma\kappa \leq (1 + \gamma\kappa)(1 + \|x\|^{2p}) \ . \qquad (5.64)
\end{aligned}$$

Let

$$C_p(\tilde{R}) = 2^{2p+1}d^p\left\{1 + \bar{\gamma}(\tilde{\mathtt{L}} + \upsilon)\right\}^{2p-1}\mathbf{\Gamma}(p + 1/2)(1 + \tilde{R})^{2p-1} \ .$$

Using (5.64) in (5.62) and that for any $a \geq b$, $\mathrm{e}^a - \mathrm{e}^b \leq (a - b)\mathrm{e}^a$ we have

$$\begin{aligned}
\mathrm{K}_{\gamma,\theta}(1 + \|x\|^{2p}) &\leq 1 + \|\mathcal{T}_\gamma(x)\|^{2p} + \gamma C_m(\tilde{R}) \\
&\leq \lambda^\gamma\mathscr{V}_p(x) + \gamma\mathrm{e}^{\kappa\bar{\gamma}}(\kappa + \log(1/\lambda))\mathscr{V}_p(\tilde{R}) + \gamma C_m(\tilde{R}) \ ,
\end{aligned}$$

which concludes the proof upon noting that $\tilde{b}_p$ is a polynomial in the dimension d.

□

Lemma 5.2.22. *Assume* **B**1, **B**2(1), **B**3 *and let* $(X_k^n)_{n\in\mathbb{N},k\in\{0,\ldots,m_n}$ *be given by* (5.19) *with* $\bar{\gamma} < \min(\mathtt{m}/(2\tilde{\mathtt{L}}^2), 1/2)$. *There exist* $A_1 \geq 1$ *and* $\varpi_1 \geq 0$ *such that for any* $a \in [1, 3]$, $n, p \in \mathbb{N}$ *and* $k \in \{0, \ldots, m_n\}$

$$\mathbb{E}\left[\mathrm{K}^p_{\gamma_n,\theta_n}V_R^a(X_k^n)\middle|X_0^0\right] \leq A_1V_R^a(X_0^0) \ , \qquad \mathbb{E}\left[V_R^a(X_0^0)\right] < +\infty \ ,$$

with V_R *given in* (5.37) *and* A_1, ϖ_1 *which do not depend on the dimension* d.

Proof. Using Jensen's inequality it suffices to prove the result for $a = 3$. Using Lemma 5.2.21, there exist $\lambda \in (0, 1)$ and $b \geq 0$ such that for any $p \in \{1, 2, 3\}$, $\theta \in \mathsf{K}$ and $\gamma \in (0, \bar{\gamma}]$, $\mathrm{K}_{\gamma,\theta}$ satisfies

$\mathbf{D}_{\mathrm{d}}(\mathscr{V}_p, \lambda^\gamma, b_p(1+d^{\varpi_{0,p}})\gamma)$ with $\mathscr{V}_p$ given in (5.38). Hence, since λ and b_p do not depend on the dimension d, using Lemma 4.2.20, there exists $\tilde{A}_1, \varpi_1 \geq 0$ such that for any $p \in \{1,2,3\}$

$$\mathbb{E}\left[\mathrm{K}^p_{\gamma_n,\theta_n}\mathscr{V}_p(X^n_k)\middle|X^0_0\right] \leq \tilde{A}_1(1+d^{\varpi_1})\mathscr{V}_p(X^0_0)\;, \qquad \mathbb{E}\left[\mathscr{V}_p(X^0_0)\right] < +\infty\;,$$

with $\tilde{A}_1$ and ϖ_1 which do not depend on the dimension d. Combining this result and Jensen's inequality we obtain that

$$\begin{aligned}
\mathbb{E}\left[\mathrm{K}^p_{\gamma_n,\theta_n}V^3_R(X^n_k)\middle|X^0_0\right] &\leq R^{-p}\sum_{p=0}^{3}\binom{3}{p}\mathbb{E}\left[\mathrm{K}^p_{\gamma_n,\theta_n}\mathscr{V}_p(X^n_k)\middle|X^0_0\right]^{1/2} \\
&\leq R^{-p}\sum_{p=0}^{3}\binom{3}{p}\tilde{A}_1^{1/2}(1+d^{\varpi_1})^{1/2}(1+\|X^0_0\|^{2p})^{1/2} \\
&\leq R^{-p}\tilde{A}_1^{1/2}(1+d^{\varpi_1})^{1/2}\sum_{p=0}^{3}\binom{3}{p}(1+\|X^0_0\|^p) \\
&\leq 9\tilde{A}_1^{1/2}(1+d^{\varpi_1})^{1/2}(1+\|X^0_0\|/R)^3\;,
\end{aligned}$$

which concludes the proof. □

Theorem 5.2.23. *Assume* **B**1, **B**2(1) *and* **B**3. *Then there exist* $A_2, \varpi_2 \geq 0$ *and* $\rho \in (0,1)$ *such that for any* $\theta \in \mathsf{K}$ *and* $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < \min(\mathtt{m}/(2\tilde{\mathrm{L}}^2), 1/2)$, $\mathrm{K}_{\gamma,\theta}$ *admits an invariant probability measure* $\pi_{\gamma,\theta}$ *and for any* $n \in \mathbb{N}$ *and* $x \in \mathbb{R}^d$

$$\mathrm{d}_{W_R}(\delta_x\mathrm{K}^n_{\gamma,\theta}, \pi_{\gamma,\theta}) \leq A_2(1+d^{\varpi_2})\rho^{\gamma n}V_R(x)\;, \qquad \mathrm{d}_{W_R}(\delta_x\mathrm{K}^n_{\gamma,\theta}, \delta_y\mathrm{K}^n_{\gamma,\theta}) \leq A_2(1+d^{\varpi_2})\rho^{\gamma n}W_R(x,y)\;,$$

with V_R, W_R *given in* (5.37) *and* $A_2, \varpi_2 \geq 0$ *and* $\rho \in (0,1)$ *which do not depend on the dimension* d.

Proof. Let $\theta \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^d$ Applying , we obtain that there exist $\tilde{A}_2 \geq 0$ and $\rho \in (0,1)$ independent of the dimension d such that

$$\mathrm{d}_{W_R}(\delta_x\mathrm{K}^n_{\gamma,\theta}, \delta_y\mathrm{K}^n_{\gamma,\theta}) \leq \tilde{A}_2\rho^{\gamma n}W_R(x,y)\;.$$

In addition, $\mathrm{K}_{\gamma,\theta}$ admits an invariant probability measure $\pi_{\gamma,\theta}$. Hence, we have

$$\mathrm{d}_{W_R}(\delta_x\mathrm{K}^n_{\gamma,\theta}, \pi_{\gamma,\theta}) \leq \tilde{A}_2\rho^{\gamma n}\int_{\mathbb{R}^d} W_R(x,y)\mathrm{d}\pi_{\gamma,\theta}(y)\;. \tag{5.65}$$

By Lemma 5.2.21 we get that $\mathrm{K}_{\gamma,\theta}$ satisfies $\mathbf{D}_{\mathrm{d}}(\mathscr{V}_2, \lambda^\gamma, b_2(1+d^{\varpi_{0,2}})\gamma)$ where $\mathscr{V}_2(x)$ is given by (5.38) with $m=2$, λ, b_2 and $\varpi_{0,2}$ are independent of the dimension d. Hence, using Jensen's inequality and that $\sqrt{1+t} \leq 1 + t/2$ for $t \geq 0$ we get

$$\pi_{\gamma,\theta}(\mathscr{V}_2^{1/2}) \leq \lambda^{\gamma/2}\pi_{\gamma,\theta}(\mathscr{V}_2^{1/2}) + b_2(1+d^{\varpi_{0,2}})\lambda^{-\bar{\gamma}/2}\gamma/2\;.$$

Therefore we obtain that

$$\pi_{\gamma,\theta}(\mathscr{V}_2^{1/2}) \leq b_2(1+d^{\varpi_{0,2}})\lambda^{-\bar{\gamma}/2}\gamma/(2-2\lambda^{\gamma/2}) \leq b_2(1+d^{\varpi_{0,2}})\lambda^{-\bar{\gamma}}\log^{-1}(1/\lambda)\;. \tag{5.66}$$

In addition, for any $y \in \mathbb{R}^d$,

$$1 + \|y\|/R \leq 2^{1/2}(1+\|y\|^2/R^2)^{1/2} \leq 2^{1/2}(1+1/R^2)^{1/2}\mathscr{V}_2^{1/2}(y)\;. \tag{5.67}$$

Combining (5.65), (5.66), (5.67) and Jensen's inequality we get that

$$\begin{aligned}
\mathrm{d}_{W_R}(\delta_x \mathrm{K}_{\gamma,\theta}^n, \pi_{\gamma,\theta}) &\leq \tilde{A}_2 \rho^{\gamma n} \int_{\mathbb{R}^d} W_R(x,y) \mathrm{d}\pi_{\gamma,\theta}(y) \\
&\leq \tilde{A}_2 \rho^{\gamma n} \left\{ \|x\| / R + \pi_{\gamma,\theta}(V_R) \right\} \\
&\leq \tilde{A}_2 \rho^{\gamma n} \left\{ \|x\| / R + 2^{1/2} (1 + 1/R^2)^{1/2} b_2 (1 + d^{\varpi_{0,2}}) \lambda^{-\bar{\gamma}} / \log(1/\lambda) \right\} \\
&\leq \tilde{A}_2 \left\{ 1 + 2^{1/2} (1 + 1/R^2)^{1/2} b_2 (1 + d^{\varpi_{0,2}}) \lambda^{-\bar{\gamma}} / \log(1/\lambda) \right\} V_R(x) \ ,
\end{aligned}$$

which concludes the proof. □

Lemma 5.2.24. *Assume* **B**1, **B**2(1) *and* **B**3. *Then there exists* $\tilde{A}_3, \varpi_3' \geq 0$ *such that for any* $\theta \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < \min(\mathtt{m}/(2\tilde{\mathrm{L}}^2), 1/2)$ *and* $k \in \mathbb{N}$

$$\|\pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma \rceil, \theta} \mathrm{P}_{\gamma\lceil 1/\gamma \rceil, \theta} - \pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma \rceil, \theta} \mathrm{K}_{\gamma,\theta}^{\lceil 1/\gamma \rceil}\|_{V_R} \leq \tilde{A}_3 (1 + d^{\varpi_3'}) \gamma^{1/2} \ ,$$

with V_R *given in* (5.37) *and* $\tilde{A}_3, \varpi_3' \geq 0$ *which do not depend on the dimension* d.

Proof. Let $\theta \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$. First, we show that $(\mathrm{P}_{t,\theta})_{t \geq 0}$ satisfies a drift condition $\mathbf{D}_{\mathrm{c}}(\mathscr{V}_2, \zeta, \beta_p(1 + d^{\varpi_{0,p}'}))$, with $\mathscr{V}_2(x) = 1 + \|x\|^2$, $\zeta > 0$ and $\beta_p, \varpi_{0,p}' \geq 0$ independent of the dimension d. We have that for any $x \in \mathbb{R}^d$, $\nabla \mathscr{V}_2(x) = 2x$ and $\Delta \mathscr{V}_2(x) = 2d$. Hence, for any $x \in \mathbb{R}^d$

$$\mathcal{A}_\theta \mathscr{V}_2(x) = -\langle \nabla_x U(\theta, x), \nabla \mathscr{V}_2(x) \rangle + \Delta \mathscr{V}_2(x) = -2 \langle \nabla_x U(\theta, x), x \rangle + 2d \ .$$

We now distinguish two cases.

(a) If $\|x\| \geq R$, using Lemma 5.2.20-(b) we have

$$\begin{aligned}
\mathcal{A}_\theta \mathscr{V}_2(x) &\leq -2\mathtt{m} \|x\|^2 + 2d + 2 \sup_{\theta \in \mathsf{K}} \|\nabla_x U(\theta, 0)\| \|x\| \\
&\leq -\mathtt{m} \mathscr{V}_2(x) + 2\{ d + \sup_{\theta \in \mathsf{K}} \|\nabla_x U(\theta, 0)\| \|x\| - \mathtt{m} \|x\|^2 / 2 + \mathtt{m} \} \\
&\leq -\mathtt{m} \mathscr{V}_2(x) + 2\{ d + \sup_{\theta \in \mathsf{K}} \|\nabla_x U(\theta, 0)\|^2 / (2\mathtt{m}) + \mathtt{m} \} \ .
\end{aligned}$$

(b) If $\|x\| \leq R$, using Lemma 5.2.20-(a) we have

$$\begin{aligned}
\mathcal{A}_\theta \mathscr{V}_2(x) &\leq 2(\tilde{\mathrm{L}} \|x\| + \sup_\theta \|\nabla_x U(\theta, x)\|) \|x\| + 2d \\
&\leq -\mathtt{m} \mathscr{V}_2(x) + 2(\tilde{\mathrm{L}} R + \sup_{\theta \in \mathsf{K}} \|\nabla_x U(\theta, x)\|) R + 2d + \mathtt{m} \mathscr{V}_2(R) \ .
\end{aligned}$$

Hence, there exists $\zeta > 0$ and $\beta_p, \varpi_{0,p}' \geq 0$ such that $(\mathrm{P}_{t,\theta})_{t \geq 0}$ satisfies $\mathbf{D}_{\mathrm{c}}(\mathscr{V}_2, \zeta, \beta_p(1 + d^{\varpi_{0,p}}))$, with ζ, β_p and $\varpi_{0,p}$ independent of the dimension d. This implies by [MT93c, Theorem 4.5]

$$\pi_\theta(\mathscr{V}_2) \leq \beta_p (1 + d^{\varpi_{0,p}}) / \zeta \ . \tag{5.68}$$

Using a generalized Pinsker inequality [DM17, Lemma 24], [DM17, Equation 15] and that for any $y \in \mathbb{R}^d$, $V_R(y) \leq (1 + 1/R^2)^{1/2} \mathscr{V}_2(y)$, we get that

$$\|\pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma \rceil, \theta} \mathrm{P}_{\gamma\lceil 1/\gamma \rceil, \theta} - \pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma \rceil, \theta} \mathrm{K}_{\gamma,\theta}^{\lceil 1/\gamma \rceil}\|_{V_R}$$

$$
\begin{aligned}
&\leq 2(1+1/R^2)^{1/2}(\pi_\theta \mathrm{P}_{(k+1)\gamma\lceil 1/\gamma\rceil}\mathscr{V}_2 + \pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma\rceil}\mathrm{K}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}\mathscr{V}_2)^{1/2} \\
&\qquad \times \mathrm{KL}\left(\pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma\rceil,\theta}\mathrm{K}_{\gamma,\theta}^{\lceil 1/\gamma\rceil} | \pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma\rceil,\theta}\mathrm{P}_{\gamma\lceil 1/\gamma\rceil,\theta}\right)^{1/2} \\
&\leq (1+1/R^2)^{1/2}(\pi_\theta(\mathscr{V}_2) + \pi_\theta \mathrm{K}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}\mathscr{V}_2)^{1/2} \\
&\qquad \times \tilde{\mathrm{L}}\left(2\tilde{\mathrm{L}}\bar{\gamma}\sup_{j\in\mathbb{N}}\left\{\pi_\theta \mathrm{K}_{\gamma,\theta}^{\lceil 1/\gamma\rceil j}\mathscr{V}_2\right\} + 2\bar{\gamma}\sup_{\theta\in\mathsf{K}}\|\nabla_x U(\theta,0)\|^2 + d\right)^{1/2} .
\end{aligned}
$$

Combining this result, (5.68) and Lemma 5.2.21 completes the proof. □

Proposition 5.2.25. *Assume* **B**1, **B**2(1) *and* **B**3. *Then there exist* $A_3, \varpi_3 \geq 0$ *such that for any* $\theta \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < \min(\mathtt{m}/(2\tilde{\mathrm{L}}^2), 1/2)$,

$$
\mathrm{d}_{W_R}(\pi_{\gamma,\theta}, \pi_\theta) \leq A_3(1+d^{\varpi_3})\gamma^{1/2} ,
$$

with W_R *given in* (5.37) *and* $A_3, \varpi_3 \geq 0$ *which do not depend on the dimension* d.

Proof. Let $\theta \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$. Using Theorem 4.1.13, we get that

$$
\mathrm{d}_{W_R}(\pi_{\gamma,\theta}, \pi_\theta) = \lim_{\mathrm{n}\to+\infty} \mathrm{d}_{W_R}(\pi_\theta \mathrm{K}_{\gamma,\theta}^{n\lceil 1/\gamma\rceil}, \pi_\theta) .
$$

By Theorem 5.2.23, Lemma 5.2.24 and that for any $\theta \in \mathsf{K}$, π_θ is an invariant probability measure for $(\mathrm{P}_{t,\theta})_{t\geq 0}$, we get for any $n \in \mathbb{N}$

$$
\begin{aligned}
\mathrm{d}_{W_R}(\pi_\theta \mathrm{K}_{\gamma,\theta}^{n\lceil 1/\gamma\rceil}, \pi_\theta \mathrm{P}_{n\gamma\lceil 1/\gamma\rceil,\theta}) &\leq \sum_{k=0}^{n-1} \mathrm{d}_{W_R}(\pi_\theta \mathrm{P}_{(k+1)\gamma\lceil 1/\gamma\rceil,\theta}\mathrm{K}_{\gamma,\theta}^{(n-k-1)\lceil 1/\gamma\rceil}, \pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma\rceil,\theta}\mathrm{K}_{\gamma,\theta}^{(n-k)\lceil 1/\gamma\rceil}) \\
&\leq A_2(1+d^{\varpi_2})\sum_{k=0}^{n-1}\rho^{n-k-1}\mathrm{d}_{W_R}(\pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma\rceil,\theta}\mathrm{P}_{\gamma\lceil 1/\gamma\rceil,\theta}, \pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma\rceil,\theta}\mathrm{K}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}) \\
&\leq A_2(1+d^{\varpi_2})\sum_{k=0}^{n-1}\rho^{n-k-1}\|\pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma\rceil,\theta}\mathrm{P}_{\gamma\lceil 1/\gamma\rceil,\theta} - \pi_\theta \mathrm{P}_{k\gamma\lceil 1/\gamma\rceil}\mathrm{K}_{\gamma,\theta}^{\lceil 1/\gamma\rceil}\|_{V_R} \\
&\leq \gamma^{1/2}A_2\tilde{A}_3(1+d^{\varpi_2})^2/\log(1/\rho) ,
\end{aligned}
$$

which concludes the proof since A_2, $\tilde{A}_3$ and ρ do not depend on the dimension d. □

Lemma 5.2.26. *There exist* $A_4, \varpi_4 \geq 0$ *such that for any* $\theta \in \mathsf{K}$, $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} < \min(\mathtt{m}/(2\tilde{\mathrm{L}}^2), 1/2)$,

$$
\|\delta_x \mathrm{K}_{\gamma_1,\theta_1} - \delta_x \mathrm{K}_{\gamma_2,\theta_2}\|_{V_R} \leq A_4(1+d^{\varpi_4})\left[\gamma_2^{-1/2}|\gamma_1-\gamma_2| + \gamma_2^{1/2}\|\theta_1-\theta_2\|\right] V_R^2(x) ,
$$

with V_R *given in* (5.37) *and* $A_4, \varpi_4 \geq 0$ *which do not depend on the dimension* d.

Proof. The proof is similar to the one of Proposition 4.2.24. □

We now turn to the proof of Theorem 5.2.2.

Proof. Combining Lemma 5.2.22, Theorem 5.2.23 and Proposition 5.2.25 we obtain that **H**1 is satisfied. Lemma 5.2.26 implies that **H**2 holds. Therefore Theorem 5.2.10 and Theorem 5.2.14 can be applied. □

Proof of Proposition 5.2.3

We recall that for any $\varepsilon > 0$ and $x \in \mathbb{R}^d$ we define $f_\varepsilon(x) = \|F(x)\|^2 - \varepsilon$.

Proposition 5.2.27. *Assume* **A**1(α) *and* **A**2(2α) *with* $\alpha > 0$. *In addition, assume that* F *is continuous,* $F^{-1}(\{0\}) \neq \emptyset$, $F^{-1}(\{0\}^c) \neq \emptyset$ *and that for every open set* $\mathsf{A} \neq \emptyset$, $\mu(\mathsf{A}) > 0$. *Then there exists* $\varepsilon_0 > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0]$, *the macrocanonical model* π_ε *associated with* f_ε *and the reference measure* μ, *solution of* (P), *exists and is given by* $(\mathrm{d}\pi_\varepsilon/\mathrm{d}\mu)(x) \propto \exp\left[-\vartheta_\varepsilon f_\varepsilon(x)\right]$, *with* $\vartheta_\varepsilon > 0$. *In addition* $\lim_{\varepsilon \to 0} \vartheta_\varepsilon = +\infty$.

Proof. Let $\varepsilon_0 = \mu(\|F\|^2)/2 > 0$, since $\mu(F^{-1}(\{0\}^c)) > 0$. Let $\varepsilon \in (0, \varepsilon_0]$. Since **A**2($2\alpha$) holds, for any $\vartheta > -\eta/C_\alpha^2$, with C_α given in **A**1(α) and η given in **A**2(2α), we have $\int_{\mathbb{R}^d} \exp[-\vartheta f_\varepsilon(x)]\mathrm{d}\mu(x) < +\infty$. Let $\mathsf{I} = \left(-\eta/C_\alpha^2, +\infty\right)$ and $L_\varepsilon : \mathsf{I} \to \mathbb{R}$ such that for any $\vartheta \in \mathsf{I}$, $L(\vartheta) = \log\left\{\int_{\mathbb{R}^d} \exp\left[-\vartheta f_\varepsilon(x)\right] \mathrm{d}\mu(x)\right\}$. By Proposition 5.1.4, we have that L is continuously differentiable on I. Since $F^{-1}(\{0\}) \neq \emptyset$ we have that there exists a non-empty open set I_ε such that for any $x \in \mathsf{I}_\varepsilon$, $f_\varepsilon(x) < 0$. Therefore

$$\lim_{\vartheta \to +\infty} L(\vartheta) \geq \lim_{\vartheta \to +\infty} \log\left\{\int_{\mathsf{I}_\varepsilon} \exp\left[-\vartheta f_\varepsilon(x)\right] \mathrm{d}\mu(x)\right\} = +\infty \;,$$

where we used the monotone convergence theorem in the last inequality. Since L is continuous we obtain that there exists $\vartheta_\varepsilon \in [0, +\infty)$ such that $L(\vartheta_\varepsilon) = \min_{[0,+\infty)} L(\vartheta)$. We have that $L'(0) \leq \varepsilon - \mu(\|F\|^2) < 0$, therefore $\vartheta_\varepsilon \in (0, +\infty)$ and $L'(\vartheta_\varepsilon) = 0$. Applying Proposition 5.1.4, we obtain that $\pi_{\vartheta_\varepsilon}$ is a solution of (P). We denote π_ε this solution.

Assume that there exists a sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ with $\varepsilon_n > 0$ such that $(\vartheta_{\varepsilon_n})_{n\in\mathbb{N}}$ is bounded. Then, up to extraction, there exists $\vartheta^\star \geq 0$ such that $\lim_n \vartheta_{\varepsilon_n} = \vartheta^\star$. Using the dominated convergence theorem we obtain that

$$0 = \lim_n \varepsilon_n = \lim_n \pi_{\varepsilon_n}(f_{\varepsilon_n}) = \pi_{\vartheta^\star}(\|F\|^2) > 0 \;,$$

which is a contradiction. Therefore, $\lim_{\varepsilon \to 0} \vartheta_\varepsilon = +\infty$. □

We now turn to the study of the tightness of the sequence $(\pi_\varepsilon)_{\varepsilon>0}$ in the special case where F is given by (5.8). Under the assumptions of Proposition 5.2.28, for each sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ such that $\lim_n \varepsilon_n = 0$, up to extraction, we have that $(\pi_{\varepsilon_n})_{n\in\mathbb{N}}$ converges to a probability measure π_∞ which concentrates on $F^{-1}(\{0\})$.

Proposition 5.2.28. *Assume* **A**2(2) *and that for any non-empty open set* $\mathsf{A} \subset \mathbb{R}^d$, $\mu(\mathsf{A}) > 0$. *Let* F *be given by* (5.8) *assume that* $1 \in \mathbf{j}$ *and that there exists* $k \in \{1, \ldots, c_1\}$ *such that for any* $x \in \mathbb{R}^d$ *with* $x \neq 0$, *there exists* $\ell \in \{1, \ldots, n_1\}$ *with* $e_\ell^\top \tilde{A}_1^k x > 0$. *Then for any sequence* $(\varepsilon_n)_{n\in\mathbb{N}}$ *with* $\lim_n \varepsilon_n = 0$, $(\pi_{\varepsilon_n})_{n\in\mathbb{N}}$ *is tight.*

Proof. Let F be given by (5.8). Then for any $x \in \mathbb{R}^d$, $f_\varepsilon(x) = \|F(x)\|^2 - \varepsilon$. We show that the level sets of $x \mapsto \|F(x)\|^2$ are compact. Let $\mathbb{S}^{d-1}$ be the sphere in $\mathbb{R}^d$ and define $\mathrm{f} : \mathbb{S}^{d-1} \to (0, +\infty)$ for any $x \in \mathbb{S}^{d-1}$ by

$$\mathrm{f}(x) = \max_{\ell \in \{1,\ldots,c_1\}} \left\{e_\ell^\top A_1^k x\right\} \;.$$

f is continuous and since $\mathbb{S}^{d-1}$ is compact, f reaches its minimum f_0 and therefore $\mathrm{f}_0 > 0$. Let $x \in \mathbb{R}^d$, using that φ is non-increasing we have for any $k \in \{1, \ldots, c_1\}$

$$n_1^{-1} \sum_{\ell=1}^{n_1} \mathscr{G}_1^k(x)(\ell) = n_1^{-1} \sum_{\ell=1}^{n_1} \varphi(e_\ell^\top A_1^k x)$$

$$\geq n_1^{-1} \sum_{\ell=1}^{n_1} \varphi(e_\ell^\top \tilde{A}_1^k x + e_\ell^\top A_1^k 0)$$
$$\geq n_1^{-1} \varphi(\mathrm{f}_0 \|x\| + \min_{\ell \in \{1,\dots,n_1\}} e_\ell^\top b_1^k) \; .$$

This result combined with the fact that $\lim_{t\to+\infty} \varphi(t) = +\infty$ implies that

$$\lim_{\|x\|\to+\infty} \|F(x)\|^2 = +\infty \; .$$

Therefore, $F^{-1}(\{0\}^{\mathrm{c}}) \neq \emptyset$. $F^{-1}(\{0\}) \neq \emptyset$ since $F(x_0) = 0$. F is continuous and **A**1(1) is satisfied. Therefore, Proposition 5.2.27 applies and $(\pi_{\varepsilon_n})_{n\in\mathbb{N}}$ is well-defined for any sequence $\lim_{n\to+\infty} \varepsilon_n = 0$. In addition, $\lim_{n\to+\infty} \vartheta_{\varepsilon_n} = +\infty$.

Since $x \mapsto \|F(x)\|^2$ is continuous and coercive, the level sets of f_ε are compact for any $\varepsilon > 0$. We conclude using [Hwa80, Proposition 2.3]. □

We now turn to the proof of Proposition 5.2.3.

Proof. The proof is then a direct consequence of the tightness of any sequence $(\pi_{\varepsilon_n})_{n\in\mathbb{N}}$ and that $\lim_{n\to+\infty} \vartheta_{\varepsilon_n} = +\infty$ combined with [Hwa80, Proposition 2.2]. □

Under other assumptions on $F^{-1}(\{0\})$ other explicit measures π_∞ are obtained.

Proposition 5.2.29. *Assume* **A**2(2) *and that for any non-empty open set* $\mathsf{A} \subset \mathbb{R}^d$, $\mu(\mathsf{A}) > 0$. *Let* F *be given by* (5.8), *assume that* $1 \in \mathbf{j}$ *and that there exists* $k \in \{1,\dots,c_1\}$ *such that for any* $x \in \mathbb{R}^d$ *with* $x \neq 0$ *there exists* $\ell \in \{1,\dots,n_1\}$ *with* $e_\ell^\top \tilde{A}_1^k x > 0$. *Then there exists* $\varepsilon_0 > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0)$, π_ε *exists. In addition, the following propositions hold:*

(a) Assume that $F^{-1}(\{0\}) = \{x_1,\dots,x_K\}$ *with* $(x_i)_{i\in\{1,\dots,K\}} \in (\mathbb{R}^d)^K$, $K \in \mathbb{N}^\star$, $\varphi \in \mathrm{C}^3(\mathbb{R})$, $x \mapsto (\mathrm{d}\mu/\mathrm{d}\lambda)(x)$ *is continuous and* $(\mathrm{d}\mu/\mathrm{d}\lambda)^{-1}(F^{-1}(\{0\})) \neq \{0\}$. *Let* $H(x) = \nabla^2(\|F(\cdot)\|^2)(x)$ *and assume that for any* $x \in F^{-1}(\{0\})$, $\det H(x) \neq 0$. *Then* $\lim_{\varepsilon\to 0} \pi_\varepsilon = \pi_\infty$ *with*

$$\pi_\infty = \sum_{i=1}^K \frac{\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x_i) \det(H(x_i))}{\sum_{j=1}^K \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x_j) \det(H(x_j))} \delta_{x_i} \; .$$

(b) Assume that $F^{-1}(\{0\})$ *is a smooth compact manifold,* $\varphi \in \mathrm{C}^3(\mathbb{R})$, $x \mapsto (\mathrm{d}\mu/\mathrm{d}\lambda)(x)$ *is continuous and* $(\mathrm{d}\mu/\mathrm{d}\lambda)^{-1}(F^{-1}(\{0\})) \neq 0$. *Let* $H(x) = \nabla^2(\|F(\cdot)\|^2)(x)$ *and assume that for any* $x \in F^{-1}(\{0\})$, $\det H(x) \neq 0$. *Then* $\lim_{\varepsilon\to 0} \pi_\varepsilon = \pi_\infty$ *with for any* $x \in \mathbb{R}^d$

$$\frac{\mathrm{d}\pi_\infty}{\mathrm{d}\mathcal{H}}(x) = \frac{\mathbb{1}_{F^{-1}(\{0\})}(x) \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x) \det H(x)}{\int_{F^{-1}(\{0\})} \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(y) \det H(y) \mathrm{d}\mathcal{H}(y)} \; ,$$

where $\mathcal{H}$ *is the intrisic measure on* $F^{-1}(\{0\})$, *see [Boo86, Chapter 6].*

Proof. The proof is then a direct consequence of the tightness of any sequence $(\pi_{\varepsilon_n})_{n\in\mathbb{N}}$ and that $\lim_{n\to+\infty} \vartheta_{\varepsilon_n} = +\infty$ combined with [Hwa80, Theorem 2.1, Theorem 3.1]. □

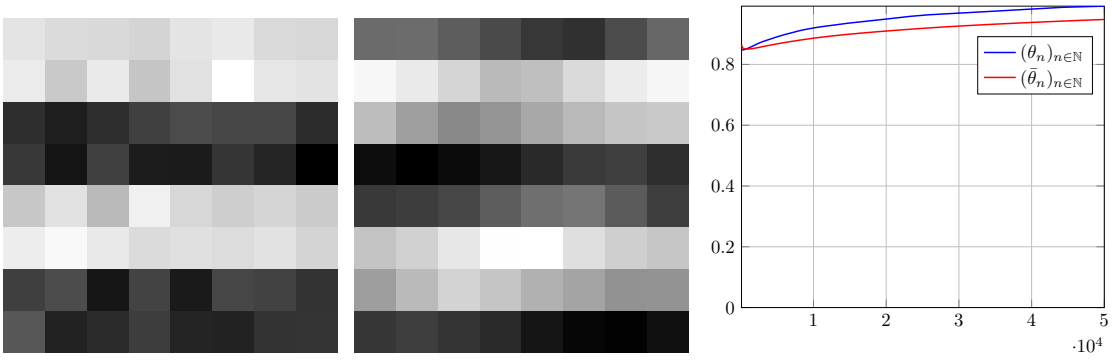
It should be noted that Proposition 5.2.3 is merely a strengthening of Proposition 5.2.28 under additional assumptions on the form of $F^{-1}(\{0\})$.

Additional experiments

Accelerations and noiseless versions of SOUL First we investigate the following discrete dynamics

$$\begin{cases} \tilde{X}_{k+1}^n = \tilde{X}_k^n - \gamma_n \left(\sum_{i=1}^p \tilde{\theta}_n^i \nabla F_i(\tilde{X}_k^n) + \nabla r(\tilde{X}_k^n) \right) & \text{and } \tilde{X}_0^n = \tilde{X}_{n-1}^{m_{n-1}} \; ; \\ \tilde{\theta}_{n+1} = \Pi_{\mathsf{K}} \left[\tilde{\theta}_n - \delta_{n+1} m_n^{-1} \sum_{j=1}^{m_n} F(\tilde{X}_j^n) \right] \; , \end{cases} \tag{5.69}$$

which corresponds to the one of SOUL (5.19) without the Gaussian noise term in the Langevin update. We refer to this algorithm as noiseless SOUL. Note that the families $\{\tilde{\theta}_n \; : \; n \in \mathbb{N}\}$ and $\{\tilde{X}_k^n \; : \; n \in \mathbb{N}, k \in \{0, \dots, m_n\}\}$ are deterministic up to their initialization. In the setting (5.69), the sequence $(\tilde{X}_0^n)_{n\in\mathbb{N}}$ seems to converge to one of the configurations presented in Figure 5.4, whereas the sequence $(\theta_n)_{n\in\mathbb{N}}$ does not converge towards the optimal parameters, see Figure 5.16. This experiment highlights that the use of a Markov kernel in the SOUL dynamics cannot be avoided in order to obtain the convergence of $(\theta_n)_{n\in\mathbb{N}}$ towards $\theta^\star$.



(a) (b) (c)

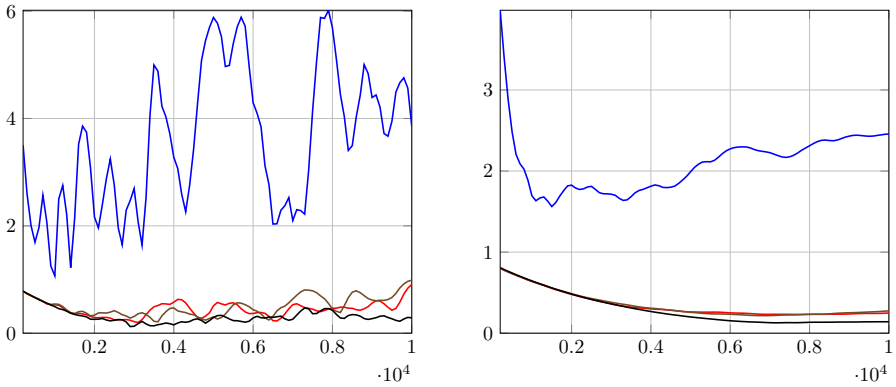
Figure 5.16: **Noiseless SOUL** The original target image is recalled in (a) and the limiting configuration obtained with the noiseless SOUL algorithm (5.69) is given in (b), whereas the non-convergence of the error towards 0 can be observed in (c). The blue curve is the NRMSE of the sequence $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ and the red curve is the NRMSE of the associated averaged sequence.

Another modification of the SOUL algorithm can be considered replacing the gradient descent step in (5.19) by another optimization methodology. Here, we focus on a popular extrapolation technique: the Nesterov acceleration. The accelerated SOUL algorithm is then given by the following recursion

$$\begin{cases} \tilde{X}_{k+1}^n = \tilde{X}_k^n - \gamma_n \left(\sum_{i=1}^p \tilde{\theta}_n^i \nabla F_i(\tilde{X}_k^n) + \nabla r(\tilde{X}_k^n) \right) + \sqrt{2\gamma_n} \mathbb{Z}_{k+1}^n & \text{and } \tilde{X}_0^n = \tilde{X}_{n-1}^{m_{n-1}} \; ; \\ \tilde{\theta}_{n+1/2} = \Pi_{\mathsf{K}} \left[\tilde{\theta}_n - \delta_{n+1} m_n^{-1} \sum_{j=1}^{m_n} F(\tilde{X}_j^n) \right] \; ; \\ \tilde{\theta}_{n+1} = \tilde{\theta}_{n+1} + \frac{n-2}{n+1} \left\{ \tilde{\theta}_{n+1/2} - \tilde{\theta}_{n-1/2} \right\} \; , \end{cases} \tag{5.70}$$

where the sequence $(\mathrm{Z}_k^n)_{n\in\mathbb{N},k\in\{1,\dots,m_n\}}$ is a sequence of independent d-dimensional zero mean Gaussian random variables with covariance identity. This algorithm is not a descent algorithm but reaches the optimal convergence rate $\mathcal{O}(1/n^2)$ for convex functions in a deterministic setting, see [Nes04; Nes83]. The perturbed gradient case is treated in [AC15] in a general framework and in [For+18; Auj+19] when the perturbation is given by a Monte Carlo approximation of the gradient. Recall that for any $n \in \mathbb{N}$ we define $\eta_n = \nabla L(\theta_n) - m_n^{-1} \sum_{k=1}^{m_n} F(X_k^n)$. The assumption on the summability of the sequence of

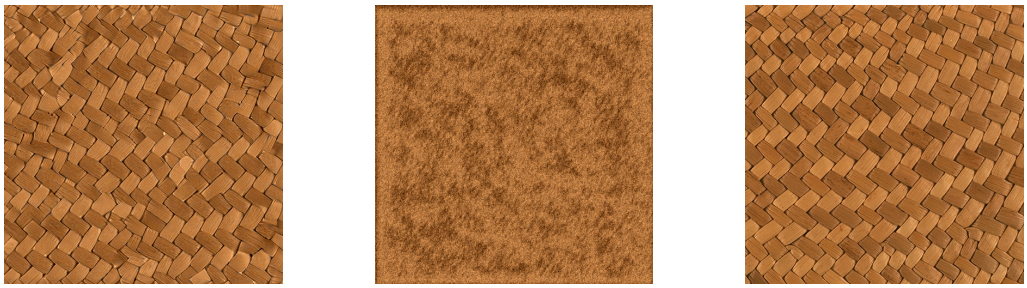
perturbations $(\eta_n)_{n\in\mathbb{N}}$ is of the form $\sum_{n\in\mathbb{N}} n\,\|\eta_n\| < +\infty$ in [AC15, Theorem 5.1]. This is a stronger requirement than $\sum_{n\in\mathbb{N}} \|\eta_n\| < +\infty$ which is a common assumption for the convergence of the perturbed gradient descent, see [KY03, Section 5.2.1]. In this accelerated setting (5.70), letting $m_n = 1$, generates oscillatory sequences $(\theta_n)_{n\in\mathbb{N}}$ which do not reduce the NRMSE. However this oscillatory effect can be counterbalanced with the use of a larger batch size, *e.g.* $m_n = 10$, see Figure 5.17.



(a) (b)

Figure 5.17: **Nesterov acceleration** The Nesterov accelerated version of SOUL (5.70) does not yield a satisfactory sequence $(\theta_n)_{n\in\mathbb{N}}$ in terms of NRMSE (blue curve in (a)) nor a satisfactory averaged sequence $(\bar{\theta}_n)_{n\in\mathbb{N}}$ (blue curve in (b)) with parameters $\delta_n = 10^{-1}$, $\gamma_n = 10^{-4}$ and $m_n = 1$. If $\delta_n = 10^{-2}$ then the results are improved (red curves) or $\delta_n = 10^{-2} \times n^{-0.5}$ (brown curve). The best results are obtained if $\delta_n = 10^{-2} \times n^{-0.5}$ and $m_n = \lceil n^{0.5} \rceil$ (black curve).

Pretraining We first assess that this pretraining is a crucial step in our model in Figure 5.18. Indeed, if for each convolutional layer ℓ and channel c, the pretrained filters are replaced by filters with weights given by a Gaussian random variable which has same mean and same variance as the pretrained filters then no visually satisfying results are obtained.



output (with pretraining) output (no pretraining) exemplar image

Figure 5.18: **Influence of the pretraining** The exemplar image on the right is a 512×512 color image. We present the output of the SOUL algorithm on this image after 10^4 iterations. The hyperparameters are fixed as follows: $\delta_n = 10^{-3}$, $\gamma_n = 10^{-5}$ and $m_n = 1$.

Color statistics It has been observed, in the case of microcanonical model, that using only CNN based features is not sufficient to describe all the textures. For instance in [LGX16], the authors propose to add spectrum constraints in order to reimpose some spatial arrangement in the images. Similarly we can combine our neural network features with pixel-based features. In order to impose some color statistics we set $F^{\mathrm{m}}_{\mathrm{color}} : \mathbb{R}^d \to \mathbb{R}^3$ and $F^{\mathrm{cov}}_{\mathrm{color}} : \mathbb{R}^d \to \mathcal{S}_3(\mathbb{R})$ defined for any $i \in \{1,2,3\}$ by

$$\begin{aligned}
&\tilde{F}^{\mathrm{m}}_{\mathrm{color}}(x)(i) = D^{-1}\sum_{k=1}^{D} x_i(k)\ ,\\
&\tilde{F}^{\mathrm{cov}}_{\mathrm{color}}(x) = D^{-1}\begin{pmatrix} x_1 - \tilde{F}^{\mathrm{m}}_{\mathrm{color}}(x)(1)\\ x_2 - \tilde{F}^{\mathrm{m}}_{\mathrm{color}}(x)(2)\\ x_3 - \tilde{F}^{\mathrm{m}}_{\mathrm{color}}(x)(3)\end{pmatrix}\begin{pmatrix} x_1 - \tilde{F}^{\mathrm{m}}_{\mathrm{color}}(x)(1)\\ x_2 - \tilde{F}^{\mathrm{m}}_{\mathrm{color}}(x)(2)\\ x_3 - \tilde{F}^{\mathrm{m}}_{\mathrm{color}}(x)(3)\end{pmatrix}^{\top}\ ,\\
&F^{\mathrm{m}}_{\mathrm{color}}(x) = \tilde{F}^{\mathrm{m}}_{\mathrm{color}}(x) - \tilde{F}^{\mathrm{m}}_{\mathrm{color}}(x_0)\ , \qquad F^{\mathrm{cov}}_{\mathrm{color}}(x) = \tilde{F}^{\mathrm{cov}}_{\mathrm{color}}(x) - \tilde{F}^{\mathrm{cov}}_{\mathrm{color}}(x_0)\ .
\end{aligned} \tag{5.71}$$

where $d = 3D$ and $x = (x_1, x_2, x_3)$ where x_i corresponds to the i-th color channel of x. These features add 9 more parameters to the model. We refer to this model as the CNN + color features. Doing so the color statistics are imposed in expectation. It is also natural to ask that all the produced images have exactly the same color statistics as the exemplar image, *i.e.* that the equality holds a.s.. This procedure can be implemented by reimposing at each Langevin step the mean and the color covariance matrix of the images. We call this model CNN + color projection. The effect of imposing, in expectation or a.s., the color constraints is investigated in Figure 5.19 and we observe that the proposed modifications do reimpose the color statistics of order 1 and 2.

Arbitrary size texture synthesis

Arbitrary size synthesis We assess in Figure 5.20 that contrary to the algorithm proposed in [LZW16], our implementation can produce arbitrary large images from one input. Indeed, if for any $j \in \{1,\dots,M\}$ and $k \in \{1,\dots,c_j\}$, $\tilde{A}^k_j$ in (5.7) is given by a convolutional operator, (5.8) can be defined for any $d \in \mathbb{N}$. The number of features does not depend on the size of the image but only on the number of layers selected in the network, since we average the neural network response in (5.8).

Highly regular textures In Section 5.2.4 we perform the comparison on highly regular textures. On these textures our algorithm and the one of [GEB15] fail at reproducing visually satisfying images (with the notable exception of the brick image for which [GEB15] yields excellent results). Adding spectral constraints, as in [LGX16], yields more regular images although the results are still not satisfactory. A solution is proposed in [GGL19] where autocorrelation features are considered at each layers. This method yields the best visual results but the parameter space is much larger than the initial image space.

Comparison with [JBV16]

Structure of `VGG19`

The layers of the `VGG19` network [SZ14] are given as follows (for each convolutional layer we indicate $(c_j, n_j) \to (c_{j+1}, n_{j+1})$):

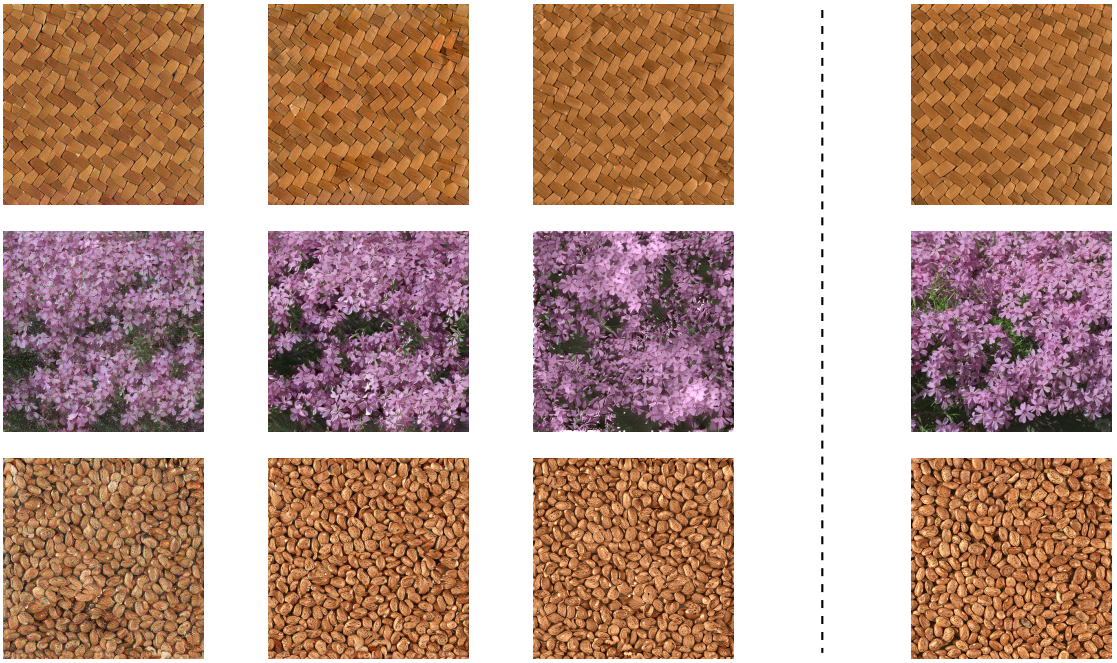


Figure 5.19: **Color models** The SOUL algorithm with CNN features yields images with less contrast than the exemplar images. To address this issue we either introduce color features in the model (CNN + color), see (5.71), or reimpose the mean and color covariance of the image after each Langevin iteration (CNN + color projection). The results are similar for both methods. The hyperparameters are fixed as follows: $\delta_n = 10^{-3}$, $\gamma_n = 10^{-5}$ and $m_n = 1$.

0. Convolutional layer, $(3, n_0) \to (64, n_0)$
1. ReLU layer
2. Convolutional layer, $(64, n_0) \to (64, n_0)$
3. ReLU layer
4. Max-pooling layer
5. Convolutional layer, $(64, n_0/2) \to (128, n_0/2)$
6. ReLU layer
7. Convolutional layer, $(128, n_0/2) \to (128, n_0/2)$
8. ReLU layer
9. Max-pooling layer
10. Convolutional layer, $(128, n_0/4) \to (256, n_0/4)$
11. ReLU layer
12. Convolutional layer, $(256, n_0/4) \to (256, n_0/4)$
13. ReLU layer
14. Convolutional layer, $(256, n_0/4) \to (256, n_0/4)$
15. ReLU layer
16. Convolutional layer, $(256, n_0/4) \to (256, n_0/4)$
17. ReLU layer
18. Max-pooling layer
19. Convolutional layer, $(256, n_0/8) \to (512, n_0/8)$
20. ReLU layer
21. Convolutional layer, $(512, n_0/8) \to (512, n_0/8)$
22. ReLU layer
23. Convolutional layer, $(512, n_0/8) \to (512, n_0/8)$
24. ReLU layer
25. Convolutional layer, $(512, n_0/8) \to (512, n_0/8)$
26. ReLU layer
27. Max-pooling layer
28. Convolutional layer, $(512, n_0/16) \to (512, n_0/16)$
29. ReLU layer
30. Convolutional layer, $(512, n_0/16) \to (512, n_0/16)$

(a) (b)

Figure 5.20: **Arbitrary size synthesis** a texture of size 1024×1024 (a) is generated from an exemplar texture of size 512×512 (b). The hyperparameters are fixed as follows: $\delta_n = 10^{-3}$, $\gamma_n = 10^{-5}$ and $m_n = 1$.

31. ReLU layer
32. Convolutional layer, $(512, n_0/16) \to (512, n_0/16)$
33. ReLU layer
34. Convolutional layer, $(512, n_0/16) \to (512, n_0/16)$
35. ReLU layer
36. Max-pooling layer

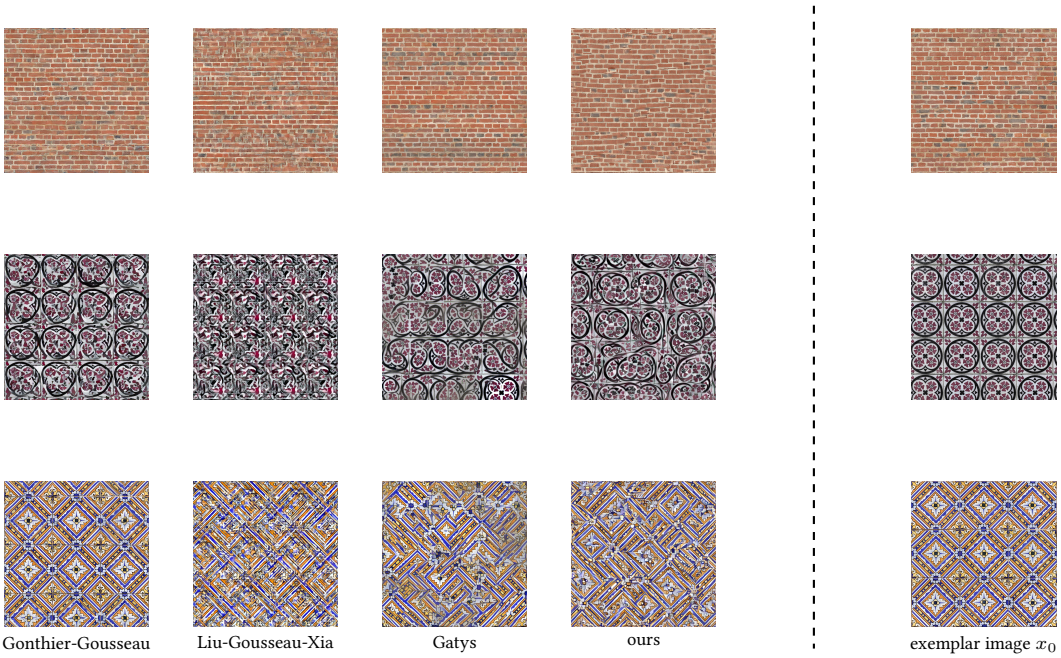


Figure 5.21: **Comparison with [LGX16]** The images presented in the column "Gonthier-Gousseau" corresponds to the features described in [GGL19] "Liu-Gousseau-Xia" are synthesized with the features considered in [LGX16], the ones presented in the column "Gatys" are generated with [GEB15] and the fourth column contains our results.

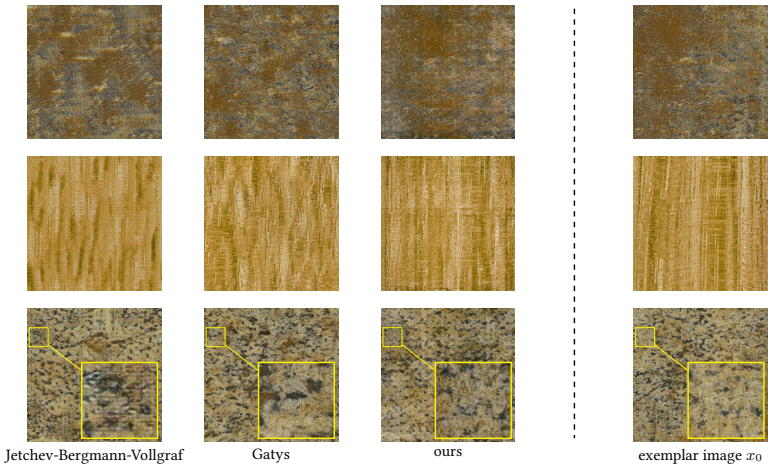


Figure 5.22: **Comparison with [JBV16]** The images presented in the column "Jetchev-Bergmann-Vollgraf" are synthesized with the algorithm introduced in [JBV16], the ones presented in the column "Gatys" are generated with [GEB15] and the third column contains our results.

Appendix A

Markov chains in general state spaces

In this section, we present a compilation of some general results in the theory on Markov chains in general state spaces. We follow closely the presentation of [Dou+18], see also [MT93a; MT92] and the references therein. We consider the following functional autoregressive model: $X_0 \in \mathbb{R}^d$ and for any $n \in \mathbb{N}$

$$X_{n+1} = F(X_n) + Z_{n+1} \ , \tag{A.1}$$

where $(Z_n)_{n\in\mathbb{N}}$ is a family of d-dimensional independent and identically distributed random variables and $F : \mathbb{R}^d \to \mathbb{R}^d$ is measurable. We assume that Z_0 admits a density with respect to the Lebesgue measure $h : \mathbb{R}^d \to [0, +\infty)$. We consider $\mathrm{Q} : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$ the Markov kernel associated with (A.1) and given for any $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\mathrm{Q}(x, \mathsf{A}) = \int_{\mathbb{R}^d} \mathbb{1}_{\mathsf{A}}(F(x) + z)h(z)\mathrm{d}z = \int_{\mathsf{A}} h(z - F(x))\mathrm{d}z \ . \tag{A.2}$$

In what follows, we specify the general results we gather for these Markov chains. For simplicity, all the results are stated on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ but most of them can be extended to any Polish space, *i.e.* any complete and separable metric space. In what follows, we consider a Markov kernel $\mathrm{K} : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$.

A.1 Small, petite and Doeblin sets

Definition A.1.1. $\mathsf{B} \in \mathcal{B}(\mathbb{R}^d)$ *is small if there exists a non-zero measure* μ *such that for any* $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ *and* $m \in \mathbb{N}^\star$ *such that for any* $x \in \mathsf{B}$ *we have* $\mathrm{K}^m(x, \mathsf{A}) \geq \mu(\mathsf{A})$.

This notion relaxes the atomic condition used, in combination with a renewal approach, to prove the ergodicity of the Markov kernel [Dou+18, Chapter 8]. If $F \in \mathrm{C}(\mathbb{R}^d, \mathbb{R}^d)$ and h is lower semi-continuous and positive then any compact set is small with measure $\mu = \varepsilon\lambda$, for some $\varepsilon > 0$.

Definition A.1.2. $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ *is accessible if for any* $x \in \mathbb{R}^d$*, there exists* $m \in \mathbb{N}$ *such that* $\mathrm{K}^m(x, \mathsf{A}) > 0$.

Definition A.1.3. K *is irreducible if it admits an accessible small set.*

Definition A.1.4. *A non-zero measure ν is said to be an irreducibility measure for* K *if for any* $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\nu(\mathsf{A}) > 0$ *implies that* A *is accessible. A non-zero measure ν is said to be a maximal irreducibility measure if it is an irreducibility measure such that for any* $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, *if* A *is accessible then* $\nu(\mathsf{A}) > 0$

Maximal irreducibility measures always exist provided that irreducibility measures exist.

Proposition A.1.5. *if there exists an irreducibility measure for* K *then there exists a maximal irreducibility measure for* K. *In addition, all maximal irreducibility measures are equivalent.*

Assume that $h > 0$, lower semi-continuous and $F \in \mathrm{C}(\mathbb{R}^d, \mathbb{R}^d)$ then [Dou+18, Proposition 9.1.9] implies that λ is an irreducibility measure for Q defined by (A.2) and one can show that λ is maximal. In fact, irreducibility measures fully characterize the irreducibility of K.

Theorem A.1.6. K *is irreducible if and only if it admits an irreducibility measure.*

Irreducibility gives information on the invariant probability measure π (if it exists).

Theorem A.1.7. *Assume that there exists π an invariant probability measure for* K *and that* K *is irreducible. Then the following holds.*

(a) π is a maximal irreducibility measure.

(b) π is unique.

Therefore, under the previous conditions on h and F we obtain that, if the invariant probability measure exists, it is unique and equivalent to the Lebesgue measure. We now turn to generalizations of the notion of small sets : the petite sets and the Doeblin sets.

Definition A.1.8. *Let a be probability distribution on $\mathbb{N}$. Let K_a be the a-sampled kernel given for any $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by*

$$\mathrm{K}_a(x, \mathsf{A}) = \sum_{n \in \mathbb{N}} a(n) \mathrm{K}^n(x, \mathsf{A}) \ .$$

Definition A.1.9. $\mathsf{C} \in \mathcal{B}(\mathbb{R}^d)$ *is petite if there exists a non-zero measure μ such that for any* $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ *and a probability distribution on $\mathbb{N}$, a, such that for any $x \in \mathsf{C}$ we have* $\mathrm{K}_a(x, \mathsf{A}) \geq \mu(\mathsf{A})$.

It is clear that any small set is petite. The converse is not true in general. Another interesting property of the petite sets is that, contrary to the small sets, they are stable under union.

Definition A.1.10. *We say that* K *is strongly aperiodic if there exists* B *and a non-zero measure μ such that for any $x \in \mathsf{B}$ and* $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\mathrm{K}(x, \mathsf{A}) \geq \mu(\mathsf{A})$ *and* $\mu(\mathsf{B}) > 0$

Under the previous assumptions on h and F we obtain that K is strongly aperiodic.

Theorem A.1.11. *If* K *is irreducible and strongly aperiodic then any petite set is small.*

The strongly aperiodicity assumption can be relaxed into an aperiodicity assumption [Dou+18, Definition 9.3.1]. In addition, we have the following properties

Proposition A.1.12. *Assume that* K *is irreducible then the following hold.*

(a) $\mathbb{R}^d = \bigcup_{n \in \mathbb{N}} \mathsf{B}_n$ with B_n small for any $n \in \mathbb{N}$;

(b) $\mathbb{R}^d = \bigcup_{n\in\mathbb{N}} \mathsf{C}_n$ *with* C_n *petite and* $\mathsf{C}_n \subset \mathsf{C}_{n+1}$ *for any* $n \in \mathbb{N}$.

We introduce our last condition (which is in fact the minimum requirement in our ergodicity results for the total variation).

Definition A.1.13. $\mathsf{D} \in \mathcal{B}(\mathbb{R}^d)$ *is a Doeblin set if there exists* $m \in \mathbb{N}$ *and* $\varepsilon > 0$ *such that for any* $x, y \in \mathsf{D}$,

$$\|\mathrm{K}^m(x,\cdot) - \mathrm{K}^m(y,\cdot)\|_{\mathrm{TV}} \leq 1 - \varepsilon \; .$$

This last condition is the one we consider in Section 4.1 in order to derive explicit convergence rates for the Markov kernel K. Note that [DM19] obtain sharp Doeblin set conditions in the case of autoregressive models in the case where F is Lipschitz and $(Z_n)_{n\in\mathbb{N}}$ is a sequence of Gaussian random variable. We can relate the Doeblin sets to the small sets.

Proposition A.1.14. *The following hold.*

(a) any small set is a Doeblin set ;

(b) if K *is irreducible and strongly aperiodic then any Doeblin set is a small set.*

Once again, the strong aperiodicity assumption can be relaxed into an aperiodicity assumption [Dou+18, Lemma 18.2.7].

A.2 Recurrence and Harris recurrence

Definition A.2.1. $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ *is recurrent if for any* $x \in \mathsf{A}$, $\mathbb{E}_x[N_{\mathsf{A}}] = \sum_{n\in\mathbb{N}} \mathrm{K}^n(x, \mathsf{A}) = +\infty$, *where* N_{A} *is the number of visits to* A.

Definition A.2.2. $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ *is Harris recurrent if for any* $x \in \mathsf{A}$, $\mathbb{P}(N_{\mathsf{A}} = +\infty) = 1$

Note that for atomic chains the definitions coincide.

Definition A.2.3. K *is recurrent (or Harris recurrent) if any accessible set is recurrent (or Harris recurrent).*

For general chain it might be difficult to prove directly that the chain is recurrent or Harris recurrent. The following Foster-Lyapunov drift condition gives an easy criterion.

Proposition A.2.4. *Let* $V : \mathbb{R}^d \to [0, +\infty)$. *Assume that for any* $M \geq 0$, $\{x \in \mathbb{R}^d : V(x) \leq M\}$ *is petite and that there exists a petite set* C *and* $b \geq 0$ *such that for any* $x \in \mathbb{R}^d$,

$$\mathrm{K}V(x) \leq V(x) + b\mathbb{1}_{\mathsf{C}}(x) \; .$$

Then K *is Harris recurrent.*

We now establish some consequence of the recurrence and Harris-recurrence of a Markov kernel.

Theorem A.2.5. *Assume that* K *is irreducible and recurrent. Then,* K *admits a non-zero invariant measure* λ, *unique up to a positive constant.*

Theorem A.2.6. *Assume that* K *is irreducible, aperiodic, Harris-recurrent and admits an invariant probability measure* π, *then for any* $\xi \in \mathscr{P}(\mathbb{R}^d)$,

$$\lim_{n\to+\infty} \|\xi\mathrm{K}^n - \pi\|_{\mathrm{TV}} = 0 \; .$$

A.3 Feller kernels

Definition A.3.1 (Feller kernel)**.** K *is Feller if for any* $f \in \mathrm{C}_b(\mathbb{R}^d, \mathbb{R})$, $\mathrm{K}f \in \mathrm{C}_b(\mathbb{R}^d, \mathbb{R})$. K *is strongly Feller if for* f *measurable and bounded,* $\mathrm{K}f \in \mathrm{C}_b(\mathbb{R}^d, \mathbb{R})$.

By [Dou+18, Lemma 12.1.5], Q defined by (A.2) is a Feller kernel if $F \in \mathrm{C}(\mathbb{R}^d, \mathbb{R}^d)$. In addition, if $h \in \mathrm{C}_b(\mathbb{R}^d, (0, +\infty))$, then Q is strongly Feller.

Definition A.3.2 (T-kernel)**.** K *is a* T*-kernel if there exists* a *be a probability distribution on* $\mathbb{N}$ *and a kernel* T *such that:*

(a) for any $x \in \mathbb{R}^d$, $\mathrm{T}(x, \mathbb{R}^d) > 0$,

(b) for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $T(\cdot, \mathsf{A})$ *is lower semi-continuous,*

(c) for any $x \in \mathbb{R}^d$ *and* $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\mathrm{K}_a(x, \mathsf{A}) \geq \mathrm{T}(x, \mathsf{A})$.

Theorem A.3.3. *Let* K *be an irreducible kernel. Then, every compact set is petite if and only if* K *is a* T*-kernel.*

The last theorem allows us to ensure the existence of an invariant probability measure for Feller kernels under a mild Foster-Lyapunov drift condition.

Theorem A.3.4. *Assume that* K *is a Feller kernel and that there exist* $V : \mathbb{R}^d \to [0, +\infty)$, $f : \mathbb{R}^d \to [0, +\infty)$ *with* $\lim_{\|x\| \to +\infty} V(x) = +\infty$ *and* $b \geq 0$ *such that for any* $x \in \mathbb{R}^d$,

$$\mathrm{K}V(x) + f(x) \leq V(x) + b \, .$$

Then K *admits an invariant probability measure.*

Appendix B

Stochastic Differential Equations

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space. In this section, we present some basic results on Stochastic Differential Equations. Most of the results are extracted from [IW89; SV06; EK86; KS91].

In what follows we let $d \in \mathbb{N}$, $b \in \mathrm{C}([0,+\infty) \times \mathbb{R}^d, \mathbb{R}^d)$ and $S \in \mathrm{C}([0,+\infty) \times \mathbb{R}^d, \mathcal{M}_d(\mathbb{R}))$. We consider the following Stochastic Differential Equation

$$\mathrm{d}\mathbf{X}_t = b(t, \mathbf{X}_t)\mathrm{d}t + S(t, \mathbf{X}_t)\mathrm{d}\mathbf{B}_t \ . \tag{B.1}$$

We denote by $\bar{\mathbb{R}}^d = \mathbb{R}^d \cup \{\infty\}$ the one-point compactification of $\mathbb{R}^d$, see [Kel75, p.149]. This compactification will allow us to define explosive solutions of SDEs, similarly to [IW89]. There exists a topology τ on $\bar{\mathbb{R}}^d$ such that (i) $\bar{\mathbb{R}}^d$ is compact, (ii) the usual topology of $\mathbb{R}^d$ is contained in the one of τ. Since $\bar{\mathbb{R}}^d$ is a topological space we can define its associated Borel sigma-field $\mathcal{B}(\bar{\mathbb{R}}^d)$.

We extend the classical notion of Wiener space in order to take into account the explosion phenomenon. For any $w : [0,+\infty) \to \bar{\mathbb{R}}^d$, we define

$$e_w = \inf\{t \in [0,+\infty) \ : \ w(t) = \infty\} \ .$$

Note that for any $t \leq e_w$, $w|_{[0,t]} \in \mathrm{C}([0,t], \mathbb{R}^d)$. We denote $\bar{W}^d([0,+\infty))$ the extended Wiener space defined as follows

$$\bar{W}^d([0,+\infty)) = \left\{w \in \mathrm{C}([0,+\infty), \bar{\mathbb{R}}^d) \ : \ \text{for any } t \geq e_w,\, w(t) = \infty\right\} \ .$$

We define the following sigma-field $\mathcal{B}(\bar{W}^d)$ on $\bar{W}^d$,

$$\mathcal{B}(\bar{W}^d) = \sigma\left(\left\{w \in \bar{W}^d \ : \ n \in \mathbb{N},\ (t_i)_{i\in\{1,\dots,n\}} \in [0,+\infty)^n, \right.\right.$$
$$\left.\left.(\mathsf{A}_i)_{i\in\{1,\dots,n\}} \in \mathcal{B}(\bar{\mathbb{R}}^d),\ \text{for any } i \in \{1,\dots,n\},\ w(t_i) \in \mathsf{A}_i\right\}\right) \ .$$

For any $(\bar{W}^d, \mathcal{B}(\bar{W}^d))$-valued random variable $\mathbf{X}$, we associate $e_{\mathbf{X}}$ given for any $\omega \in \Omega$ by, $e_{\mathbf{X}}(\omega) = e_{\mathbf{X}(\omega)}$. Note that $e_{\mathbf{X}}$ is $[0,+\infty]$-valued random variable.

B.1 Existence of solutions

Definition B.1.1. *We say that (B.1) admits a weak solution, if there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a filtration $(\mathcal{F}_t)_{t\geq 0}$, a $(\mathcal{F}_t)_{t\geq 0}$-Brownian motion $(\mathbf{B}_t)_{t\geq 0}$ and $\mathbf{X}$ such that*

(a) $\mathbf{X}:\ (\Omega,\mathcal{F})\to(\bar{W}^d,\mathcal{B}(\bar{W}^d))$,

(b) $\mathbf{X}$ *is* $(\mathcal{F}_t)_{t\geq 0}$*-adapted, i.e. for any* $t\geq 0$, $\mathbf{X}(t):\ (\Omega,\mathcal{F}_t)\to(\bar{\mathbb{R}}^d,\mathcal{B}(\bar{\mathbb{R}}^d))$,

(c) there exists $\mathsf{A}\in\mathcal{F}$ *such that* $\mathbb{P}(\mathsf{A})=1$ *and for any* $\omega\in\mathsf{A}$, *and* $t\leq e_{\mathbf{X}}(\omega)$,

$$\mathbf{X}(t)=\mathbf{X}(0)+\int_0^t b(s,\mathbf{X}_s)\mathrm{d}s+\int_0^t S(s,\mathbf{X}_s)\mathrm{d}\mathbf{B}_s\ .$$

Note that in the previous definition, the probability space, the filtration and the Brownian motion are not fixed. We denote $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\geq 0},\mathbf{B},\mathbf{X})$ a weak solution of (B.1). In the following definition, we consider the problem of the existence of solutions of (B.1) in the case the probability space, the filtrations and the Brownian motion are given beforehand.

Definition B.1.2. *We say that* (B.1) *admits a strong solution, if for any probability space* $(\Omega,\mathcal{F},\mathbb{P})$, *filtration* $(\mathcal{F}_t)_{t\geq 0}$, *and* $(\mathcal{F}_t)_{t\geq 0}$*-Brownian motion* $(\mathbf{B}_t)_{t\geq 0}$ *there exists* $\mathbf{X}$ *such that*

(a) $\mathbf{X}:\ (\Omega,\mathcal{F})\to(\bar{W}^d,\mathcal{B}(\bar{W}^d))$,

(b) $\mathbf{X}$ *is* $(\mathcal{F}_t)_{t\geq 0}$*-adapted, i.e. for any* $t\geq 0$, $\mathbf{X}(t):\ (\Omega,\mathcal{F}_t)\to(\bar{\mathbb{R}}^d,\mathcal{B}(\bar{\mathbb{R}}^d))$,

(c) there exists $\mathsf{A}\in\mathcal{F}$ *such that* $\mathbb{P}(\mathsf{A})=1$ *and for any* $\omega\in\mathsf{A}$, *and* $t\leq e_{\mathbf{X}}(\omega)$,

$$\mathbf{X}(t)=\mathbf{X}(0)+\int_0^t b(s,\mathbf{X}_s)\mathrm{d}s+\int_0^t S(s,\mathbf{X}_s)\mathrm{d}\mathbf{B}_s\ .$$

Definition B.1.1 allows for a martingale formulation of the SDE, see [SV06, Chapter 6]. Using this martingale formulation, we obtain weak solutions under mild conditions on the coefficients b and S. However, the weak formulation is not suitable if one wants to compare solutions of SDEs. Indeed, assume that the drift term b depends on some parameter $\eta\in\mathbb{R}$, b_η and denote $\mathbf{X}^\eta$ the solution of (B.1). If (B.1) only admits weak solutions then $\mathbb{E}\left[\|\mathbf{X}^{\eta_1}(t)-\mathbf{X}^{\eta_2}(t)\|^2\right]$, with $\eta_1,\eta_2\in\mathbb{R}$ and $t\geq 0$, is not defined, since the solutions may not be defined on the same probability spaces. In this case we need to consider strong solutions as in Definition B.1.2.

The following result ensures the existence of weak solutions. Recall that $b\in\mathrm{C}([0,+\infty)\times\mathbb{R}^d,\mathbb{R}^d)$ and $S\in\mathrm{C}([0,+\infty)\times\mathbb{R}^d,\mathcal{M}_d(\mathbb{R}))$.

Theorem B.1.3. *For any* $x\in\mathbb{R}^d$, *there exists a weak solution* $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\geq 0},\mathbf{X})$ *such that* $\mathbf{X}(0)=x$ *a.s.*.

If there exists $\mathtt{L}\geq 0$ such that for any $t\geq 0$ and $x,y\in\mathbb{R}^d$

$$\|b(t,x)-b(t,y)\|\leq\mathtt{L}\|x-y\|\ ,\qquad\|S(t,x)-S(t,y)\|\leq\mathtt{L}\|x-y\|\ ,$$

then strong existence results can be derived, see [IW89, p.166], using a proof similar to the one of the Cauchy-Peano existence theorem for ordinary differential equations. In the following section, we remark that one may construct strong solutions from weak solutions using some uniqueness results.

B.2 Uniqueness and from weak to strong solutions

We start by introducing some definition of uniqueness for SDEs.

Definition B.2.1. *We say that the pathwise uniqueness of solutions holds for* (B.1) *if for any solutions* $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{B}, \mathbf{X})$ *and* $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \tilde{\mathbf{B}}, \tilde{\mathbf{X}})$ *such that* $\mathbf{X}(0) = \tilde{\mathbf{X}}(0)$ *a.s., then* $\mathbf{X} = \tilde{\mathbf{X}}$ *a.s..*

If there exist a weak solution to (B.1) and the pathwise uniqueness of solutions holds then we can go from weak to strong solutions.

Proposition B.2.2. *Assume that* (B.1) *admits a weak solution and that the pathwise uniqueness of solutions holds. Then* (B.1) *admits a strong solution.*

In the following proposition, we give conditions on the SDE coefficients so that the pathwise uniqueness holds.

Proposition B.2.3. *Assume that for any* $t \geq 0$ *and* $n \in \mathbb{N}$*, there exists* $\mathtt{L}_n \geq 0$ *such that for any* $x, y \in \bar{\mathrm{B}}(0, n)$*,*

$$\|b(t,x) - b(t,y)\| + \|S(t,x) - S(t,y)\| \leq \mathtt{L}_n \|x - y\| \ .$$

Then, the pathwise uniqueness of solutions holds for (B.1)*. Hence, using Proposition B.2.2 and Theorem B.1.3,* (B.1) *admits a pathwise unique, strong solution.*

This existence and uniqueness condition on the coefficient on the SDE is similar to the ones we impose to ordinary differential equations in order for the Picard-Lindelöf existence and uniqueness theorem to hold. However, the last proposition does not give information on the explosion time $e_{\mathbf{X}}$. In most of our work, we need the strong solution to be global, *i.e.* for any $(\Omega, \mathcal{F})$, $(\mathcal{F}_t)_{t\geq 0}$ and $(\mathbf{B}_t)_{t\geq 0}$, $e_{\mathbf{X}} = \infty$.

B.3 Global solutions

The globality of the strong solution of (B.1) is controlled by the growth of the coefficients. We give a first, easy to check, criterion.

Proposition B.3.1. *Assume that* (B.1) *admits a strong solution and that there exists* $C \geq 0$ *such that for any* $x \in \mathbb{R}^d$ *and* $t \geq 0$

$$\|b(t,x)\| + \|S(t,x)\| \leq C(1 + \|x\|) \ .$$

Then each strong solution is global.

However, it may happen that this last sub-linearity condition is too restrictive for our purposes. In this case we turn to the stability theorem of SDE. The following theorem is a slight modification of [Kha11, Theorem 3.5].

Proposition B.3.2. *Assume that for any* $t \geq 0$ *and* $n \in \mathbb{N}$*, there exists* $\mathtt{L}_n \geq 0$ *such that for any* $x, y \in \bar{\mathrm{B}}(0, n)$*,*

$$\|b(t,x) - b(t,y)\| + \|S(t,x) - S(t,y)\| \leq \mathtt{L}_n \|x - y\| \ .$$

In addition, assume that there exists $V \in \mathrm{C}^2(\mathbb{R}^d, [0, +\infty))$ *such that* $\lim_{\|x\|\to+\infty} V(x) = +\infty$ *and that there exists* $\mathtt{m} \in \mathbb{R}$ *such that for any* $t \geq 0$ *and* $x \in \mathbb{R}^d$

$$\langle \nabla V(x), b(t,x)\rangle + (1/2)\langle \nabla^2 V(x), SS^\top(t,x)\rangle \leq \mathtt{m} V(x) \ .$$

Then (B.1) *admits a global, pathwise unique, strong solution.*

B.4 Invariant probability measures

In this section we assume that (B.1) is time-homogeneous, *i.e.* we consider

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + S(\mathbf{X}_t)\mathrm{d}\mathbf{B}_t \ ,$$

with $b \in \mathrm{C}(\mathbb{R}^d, \mathbb{R}^d)$ and $S \in \mathrm{C}(\mathbb{R}^d, \mathcal{M}_d(\mathbb{R}))$. In addition, we assume that for any $x \in \mathbb{R}^d$, (B.1) admits a global, pathwise unique, strong solution $\mathbf{X}^x$ such that $\mathbf{X}_0^x = x$. In this case, we consider $(\mathrm{P}_t)_{t\geq 0}$ such that for any $t \in \mathbb{R}^d$, $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$

$$\mathrm{P}_t(x, \mathsf{A}) = \mathbb{P}(\mathbf{X}_t^x \in \mathsf{A}) \ . \tag{B.2}$$

For any $f \in \mathrm{C}_b(\mathbb{R}^d, \mathbb{R})$, $\pi \in \mathscr{P}(\mathbb{R}^d)$, $x \in \mathbb{R}^d$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ and $t \geq 0$ we define

$$\mathrm{P}_t f(x) = \int_{\mathbb{R}^d} f(y)\mathrm{P}_t(x, \mathrm{d}y) \ , \qquad \pi\mathrm{P}(\mathsf{A}) = \int_{\mathbb{R}^d} \mathrm{P}(y, \mathsf{A})\mathrm{d}\pi(y) \ .$$

We say that π is an invariant probability measure for $(\mathrm{P}_t)_{t\geq 0}$ if for any $t \geq 0$, $\pi\mathrm{P}_t = \pi$. We now aim at deriving an invariant probability measure for $(\mathrm{P}_t)_{t\geq 0}$ given the coefficients b and S. In order to do so, most authors rely on the theory of semi-groups on Banach spaces, see [EK86, Chapter 1]. We recall that $(\mathrm{C}_0(\mathbb{R}^d, \mathbb{R}), \|\cdot\|_\infty)$ is the Banach space of vanishing continuous functions, *i.e.* $f \in \mathrm{C}(\mathbb{R}^d, \mathbb{R})$ and $\lim_{\|x\|\to+\infty} f(x) = 0$. For any $f \in \mathrm{C}_0(\mathbb{R}^d, \mathbb{R})$, we denote $\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)|$.

Definition B.4.1 (Semi-group and Feller semi-group)**.** *A semi-group on a Banach space* E *is a family of bounded linear operators* $(A_t)_{t\geq 0}$ *such that for any* $t, s \geq 0$, $A_{t+s} = A_t A_s$, $A_0 = \mathrm{Id}$ *and for any* $f \in \mathsf{E}$, $\lim_{t\to 0}\{A_t(f) - f\} = 0$. *A semi-group is a contraction semi-group if for any* $t \geq 0$, $\|A_t\| \leq 1$. *A Feller semi-group is a contraction semi-group for the Banach space* $\mathrm{C}_0(\mathbb{R}^d, \mathbb{R})$.

In order to derive useful results on the invariant probability of the family of Markov kernels $(\mathrm{P}_t)_{t\geq 0}$ defined by (B.2) we must assume that $(\mathrm{P}_t)_{t\geq 0}$ is a Feller semi-group. The following proposition gives conditions on the coefficients b and S under which this is the case, see [RY99, Chapter IX, Theorem 2.5].

Proposition B.4.2. *If there exists* $\mathtt{L} \geq 0$ *such that for any* $x, y \in \mathbb{R}^d$

$$\|b(x)\| + \|S(x)\| \leq \mathtt{L} \ , \qquad \|b(x) - b(y)\| + \|S(x) - S(y)\| \leq \mathtt{L}\,\|x - y\| \ ,$$

then $(\mathrm{P}_t)_{t\geq 0}$ *is a Feller semi-group.*

We also introduce the infinitesimal operator. Let $(\mathrm{P}_t)_{t\geq 0}$ be given by (B.2) and assume that for any $t \geq 0$, P_t is an operator on $\mathrm{C}_0(\mathbb{R}^d, \mathbb{R})$. For any $f \in \mathrm{C}_0(\mathbb{R}^d, \mathbb{R})$ such that the limit is defined, we let $\mathcal{A}f = \lim_{t\to 0}(1/t)(\mathrm{P}_t f - f)$. We denote by $\mathscr{D}(\mathcal{A})$ the subspace of $\mathrm{C}_0(\mathbb{R}^d, \mathbb{R})$ such that the limit exists. The following result, extracted from [Dur16], allows to explicitly compute the invariant probability measure of the family of Markov kernels $(\mathrm{P}_t)_{t\geq 0}$.

Proposition B.4.3. *Let* $\pi \in \mathscr{P}(\mathbb{R}^d)$. *Assume that* $(\mathrm{P}_t)_{t\geq 0}$ *is a Feller semi-group such that* $\mathscr{D}(\mathcal{A})$ *is an algebra and is dense in* $\mathrm{C}_0(\mathbb{R}^d, \mathbb{R})$. *In addition, assume that for any* $f \in \mathscr{D}(\mathcal{A})$,

$$\int_{\mathbb{R}^d} (\mathcal{A}f)(x)\mathrm{d}\pi(x) = 0 \ .$$

Then, π *is an invariant probability measure for* $(\mathrm{P}_t)_{t\geq 0}$.

Proof. Using [RY99, Proposition 1.5], we have that $(\mathrm{P}_t)_{t\geq 0}$ satisfies the positive maximum principle, *i.e.* for any $f \in \mathscr{D}(\mathcal{A})$ and for any $x_0 \in \mathbb{R}^d$ such that $f(x_0) = \sup_{x\in\mathbb{R}^d} f(x) \geq 0$, we have $\mathcal{A}f(x_0) \leq 0$. Then the proof follows from [EK86, Theorem 9.17]. □

Bibliography

[ABW12] Sungjin Ahn, Anoop Korattikara Balan, and Max Welling. "Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring". In: *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

[AC15] Hedy Attouch and Zaki Chbani. *Fast inertial dynamics and FISTA algorithms in convex optimization. Perturbation aspects*. 2015. arXiv: 1507.01367.

[Adl81] Robert Adler. *The geometry of random fields*. John Wiley & Sons, Ltd., Chichester, 1981, pp. xi+280. ISBN: 0-471-27844-0.

[ADV03] Andrés Almansa, Agnès Desolneux, and Sébastien Vamech. "Vanishing Point Detection without Any A Priori Information". In: *IEEE Transactions on Pattern Analysis Machine Intelligence* 25.4 (2003), pp. 502–507.

[AFM17] Yves Atchadé, Gersende Fort, and Eric Moulines. "On perturbed proximal gradient algorithms". In: *Journal of Machine Learning Research (JMLR)* 18 (2017), Paper No. 10, 33. ISSN: 1532-4435.

[AM06] Christophe Andrieu and Éric. Moulines. "On the ergodicity properties of some adaptive MCMC algorithms". In: *The Annals of Applied Probability* 16.3 (2006), pp. 1462–1505. ISSN: 1050-5164.

[AN00] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. Vol. 191. Translations of Mathematical Monographs. 2000, pp. x+206. ISBN: 0-8218-0531-2.

[Ash01] Michael Ashikhmin. "Synthesizing natural textures". In: *Proceedings of the 2001 Symposium on Interactive 3D Graphics, SI3D 2001, Chapel Hill, NC, USA, March 26-29, 2001*. Ed. by John F. Hughes and Carlo H. Séquin. ACM, 2001, pp. 217–226.

[AST10] Robert Adler, Gennady Samorodnitsky, and Jonathan Taylor. "Excursion sets of three classes of stable random fields". In: *Advances in Applied Probability* 42.2 (2010), pp. 293–318. ISSN: 0001-8678.

[ASW14] Sungjin Ahn, Babak Shahbaba, and Max Welling. "Distributed Stochastic Gradient MCMC". In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, 2014, pp. 1044–1052.

[Aub01] Thierry Aubin. *A course in differential geometry*. Vol. 27. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2001, pp. xii+184. ISBN: 0-8218-2709-X.

[Auj+19] Jean-François Aujol, Charles Dossal, Gersende Fort, and Éric Moulines. *Rates of Convergence of Perturbed FISTA-based algorithms*. working paper or preprint. July 2019.

[AW06] Suyash Awate and Ross Whitaker. "Unsupervised, Information-Theoretic, Adaptive Image Filtering for Image Restoration". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.3 (2006), pp. 364–376.

[BA16] Dean Bodenham and Niall Adams. "A comparison of efficient approximations for a weighted sum of chi-squared random variables". In: *Statistics and Computing* 26.4 (2016), pp. 917–928. ISSN: 0960-3174.

[Bad13] Adrian Baddeley. "Spatial point patterns: models and statistics". In: *Stochastic geometry, spatial statistics and random fields*. Vol. 2068. Lecture Notes in Math. Springer, Heidelberg, 2013, pp. 49–114.

[Bak+08] Dominique Bakry, Franck Barthe, Patrick Cattiaux, and Arnaud Guillin. "A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case". In: *Electronic Communications in Probability* 13 (2008), pp. 60–66.

[Bak+19] Jack Baker, Paul Fearnhead, Emily Fox, and Christopher Nemeth. "Control variates for stochastic gradient MCMC". In: *Statistics and Computing* 29.3 (2019), pp. 599–615. ISSN: 0960-3174.

[Bar17] Jonathan Barron. *Continuously Differentiable Exponential Linear Units*. 2017. arXiv: 1704.07483.

[BC11] Heinz Bauschke and Patrick Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. 1st ed. CMS Books in Mathematics. Springer-Verlag New York, 2011. ISBN: 1441994661.

[BCM05] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. "A Non-Local Algorithm for Image Denoising". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*. 2005, pp. 60–65.

[BCM11] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. "Non-Local Means Denoising". In: *Image Processing On Line* 1 (2011), pp. 208–212.

[BD16] Hermine Biermé and Agnès Desolneux. "On the perimeter of excursion sets of shot noise random fields". In: *The Annals of Probability* 44.1 (2016), pp. 521–543. ISSN: 0091-1798.

[BDJ98] Russ Bubley, Martin Dyer, and Mark Jerrum. "An elementary analysis of a procedure for sampling points in a convex body". In: *Random Structures & Algorithms* 12.3 (1998), pp. 213–235. ISSN: 1042-9832.

[BE88] Michael Buckley and G.K. Eagleson. "An approximation to the distribution of quadratic forms in normal random variables". In: *Australian Journal of Statistics* 30A.1 (1988), pp. 150–159.

[BEL15] Sébastien. Bubeck, Ronen Eldan, and Joeseph Lehec. "Finite-time Analysis of Projected Langevin Monte Carlo". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 1243–1251.

[Ben96] Michel Benaim. "A dynamical system approach to stochastic approximations". In: *SIAM Journal on Control and Optimization* 34.2 (1996), pp. 437–472. ISSN: 0363-0129.

[Ber18] Espen Bernton. "Langevin Monte Carlo and JKO splitting". In: *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1777–1798.

[Bes86] Julian Besag. "On the statistical analysis of dirty pictures". In: *Journal of the Royal Statistical Society. Series B. Methodological* 48.3 (1986), pp. 259–302. ISSN: 0035-9246.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552. ISBN: 978-3-319-00226-2; 978-3-319-00227-9.

[Bil95] Patrick Billingsley. *Probability and measure*. Third. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1995, pp. xiv+593. ISBN: 0-471-00710-2.

[BJV17] Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. "Learning Texture Manifolds with the Periodic Spatial GAN". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. 2017, pp. 469–477.

[BLC05] Léon Bottou and Yann Le Cun. "On-line learning for very large data sets". In: *Applied Stochastic Models in Business and Industry* 21.2 (2005), pp. 137–151.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013, pp. x+481. ISBN: 978-0-19-953525-5.

[BLN96] Jonathan. Borwein, Adrian. Lewis, and Dominikus Noll. "Maximum entropy reconstruction using derivative information. I. Fisher information and convex duality". In: *Mathematics of Operations Research* 21.2 (1996), pp. 442–468. ISSN: 0364-765X.

[BM11] Francis Bach and Éric Moulines. "Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning". In: *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*. 2011, pp. 451–459.

[BM18] Joan Bruna and Stephane Mallat. "Multiscale Sparse Microcanonical Models". In: *European Mathematical Society* 1 (2018), pp. 257–315.

[BMP90] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*. Vol. 22. Applications of Mathematics (New York). Springer-Verlag, Berlin, 1990, pp. xii+365. ISBN: 3-540-52894-6.

[BNE10] Laura Balzano, Robert Nowak, and J Ellenberg. *Compressed sensing audio demonstration*. 2010. URL: https://web.eecs.umich.edu/~girasole/csaudio/.

[Boo86] William Boothby. *An introduction to differentiable manifolds and Riemannian geometry*. Second. Vol. 120. Pure and Applied Mathematics. 1986, pp. xvi+430. ISBN: 0-12-116052-1; 0-12-116053-X.

[Bor+19] Valentin De Bortoli, Agnès Desolneux, Bruno Galerne, and Arthur Leclaire. "Patch Redundancy in Images: A Statistical Testing Framework and Some Applications". In: *SIAM Journal on Imaging Sciences* 12.2 (2019), pp. 893–926.

[Bre11] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011, pp. xiv+599. ISBN: 978-0-387-70913-0.

[Bre99] Pierre Bremaud. *Markov chains*. Vol. 31. Texts in Applied Mathematics. Springer-Verlag, New York, 1999, pp. xviii+444. ISBN: 0-387-98509-3.

[Bro+18] Nicolas Brosse, Alain Durmus, Sean Meyn, Eric Moulines, and Anand Radhakrishnan. *Diffusion approximations and control variates for MCMC*. 2018. arXiv: 1808.01665.

[Bro+19] Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. "The tamed unadjusted Langevin algorithm". In: *Stochastic Processes and their Applications* 129.10 (2019), pp. 3638–3663. ISSN: 0304-4149.

[But14] Oleg Butkovsky. "Subgeometric rates of convergence of Markov processes in the Wasserstein metric". In: *The Annals of Applied Probability* 24.2 (2014), pp. 526–552. ISSN: 1050-5164.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004, pp. xiv+716. ISBN: 0-521-83378-7.

[Cao04] Frédéric Cao. "Application of the Gestalt principles to the detection of good continuations and corners in image level lines". In: *Computing and Visualization in Science* 7.1 (2004), pp. 3–13. ISSN: 1432-9360.

[Cas01] George Casella. "Empirical Bayes Gibbs sampling". In: *Biostatistics* 2.4 (Dec. 2001), pp. 485–500. ISSN: 1465-4644.

[Cas85] George Casella. "An introduction to empirical Bayes data analysis". In: *The American Statistician* 39.2 (1985), pp. 83–87. ISSN: 0003-1305.

[CB18] Xiang Cheng and Peter L. Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Algorithmic Learning Theory, ALT 2018, 7-9 April 2018, Lanzarote, Canary Islands, Spain*. Ed. by Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan. Vol. 83. Proceedings of Machine Learning Research. PMLR, 2018, pp. 186–211.

[CB90] George Casella and Roger Berger. *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990, pp. xviii+650. ISBN: 0-534-11958-1.

[CCB85] Rama Chellappa, Sangit Chatterjee, and Richard Bagdazian. "Texture synthesis and compression using Gaussian-Markov random field models". In: *IEEE Transactions on Systems, Man, and Cybernetics* 15.2 (1985), pp. 298–303.

[CDY05] David Coupier, Agnès Desolneux, and Bernard Ycart. "Image Denoising by Statistical Area Thresholding". In: *Journal of Mathematical Imaging and Vision* 22.2-3 (2005), pp. 183–197.

[CGG87] Han-Fu Chen, Lei Guo, and Ai-Jun Gao. "Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds". In: *Stochastic Processes and their Applications* 27 (1987), pp. 217–231.

[CGG99] Imre Csiszár, Fabrice Gamboa, and Elisabeth Gassiat. "MEM pixel correlated solutions for generalized moment and interpolation problems". In: *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* 45.7 (1999), pp. 2253–2270. ISSN: 0018-9448.

[CH80] Richard Conners and Charles Harlow. "Toward a structural textural analyzer based on statistical methods". In: *Computer Graphics and Image Processing* 12.3 (1980), pp. 224–256.

[Cha+18] Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, and Michael Jordan. "On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo". In: *International Conference on Machine Learning*. 2018, pp. 763–772.

[Cha93] Kung-Sik Chan. "Asymptotic behavior of the Gibbs sampler". In: *Journal of the American Statistical Association* 88.421 (1993), pp. 320–326. ISSN: 0162-1459.

[Che+18] Xiang Cheng, Niladri Chatterji, Yasin Abbasi-Yadkori, Peter Bartlett, and Michael Jordan. *Sharp Convergence Rates for Langevin Dynamics in the Nonconvex Setting*. 2018. arXiv: 1805.01648.

[Chi+13] Sung Nok Chiu, Dietrich Stoyan, Wilfrid Kendall, and Joseph Mecke. *Stochastic geometry and its applications*. Third. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2013, pp. xxvi+544. ISBN: 978-0-470-66481-0.

[Cia82] Philippe Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Collection Mathématiques Appliquées pour la Maîtrise. Masson, Paris, 1982, pp. xii+279. ISBN: 2-225-68893-1.

[CJ83] George Cross and Anil Jain. “Markov Random Field Texture Models”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5.1 (1983), pp. 25–39.

[CJY91] Raymond Chan, Xiao-Qing Jin, and Man-Chung Yeung. “The circulant operator in the Banach algebra of matrices”. In: *Linear Algebra and its Applications* 149 (1991), pp. 41–53. ISSN: 0024-3795.

[CL00] Bradley P. Carlin and Thomas A. Louis. “Empirical Bayes: past, present and future”. In: *Journal of the American Statistical Association* 95.452 (2000), pp. 1286–1289. ISSN: 0162-1459.

[CL04] Harald Cramér and M. R. Leadbetter. *Stationary and related stochastic processes*. 2004, pp. xiv+348. ISBN: 0-486-43827-9.

[CM88] D. Cano and T.Ha Minh. “Texture synthesis using hierarchical linear transforms”. In: *Signal Processing* (1988). ISSN: 0165-1684.

[CP16] Antonin Chambolle and Thomas Pock. “An introduction to continuous optimization for imaging”. In: *Acta Numerica* 25 (2016), pp. 161–319.

[CPT04] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. “Region filling and object removal by exemplar-based image inpainting”. In: *IEEE Transactions on Image Processing* 13.9 (2004), pp. 1200–1212.

[Cra99] Harald Cramér. *Mathematical methods of statistics*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1999, pp. xvi+575. ISBN: 0-691-00547-8.

[Csi75] Imre Csiszár. “I-divergence geometry of probability distributions and minimization problems”. In: *The Annals of Probability* 3 (1975), pp. 146–158.

[Csi84] Imre Csiszár. “Sanov property, generalized I-projection and a conditional limit theorem”. In: *The Annals of Probability* 12.3 (1984), pp. 768–793. ISSN: 0091-1798.

[Csi96] Imre Csiszár. “Maxent, mathematics, and information theory”. In: *Maximum entropy and Bayesian methods (Santa Fe, NM, 1995)*. Vol. 79. Fund. Theories Phys. Kluwer Acad. Publ., Dordrecht, 1996, pp. 35–50.

[CT91] Rong. Chen and Ruey Tsay. “On the ergodicity of $\mathrm{TAR}(1)$ processes”. In: *The Annals of Applied Probability* 1.4 (1991), pp. 613–634. ISSN: 1050-5164.

[CW08] Emmanuel Candes and Michael Wakin. “An Introduction To Compressive Sampling”. In: *IEEE Signal Processing Magazine* 25.2 (2008), pp. 21–30.

[CW36] Harald Cramér and Herman Wold. “Some Theorems on Distribution Functions”. In: *The Journal of the London Mathematical Society* 11.4 (1936), pp. 290–294. ISSN: 0024-6107.

[CW95] Mu-Fa Chen and Feng-Yu Wang. “Estimation of the first eigenvalue of second order elliptic operators”. In: *Journal of Functional Analysis* 131.2 (1995), pp. 345–363. ISSN: 0022-1236.

[CW97] Mu-Fa Chen and Feng-Yu Wang. “Estimation of spectral gap for elliptic operators”. In: *Transactions of the American Mathematical Society* 349.3 (1997), pp. 1239–1267. ISSN: 0002-9947.

[Dab+07] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. “Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering”. In: *IEEE Transactions on Image Processing* 16.8 (2007), pp. 2080–2095.

[DAG10] Vincent Duval, Jean-François Aujol, and Yann Gousseau. “On the parameter choice for the Non-Local Means”. Mar. 2010. URL: https://hal.archives-ouvertes.fr/hal-00468856.

[DAG11] Vincent Duval, Jean-François Aujol, and Yann Gousseau. “A Bias-Variance Approach for the Nonlocal Means”. In: *SIAM Journal on Imaging Sciences* 4.2 (2011), pp. 760–788.

[Dal17a] Arnak Dalalyan. “Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent”. In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, 2017, pp. 678–689.

[Dal17b] Arnak Dalalyan. “Theoretical guarantees for approximate sampling from smooth and log-concave densities”. In: *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 79.3 (2017), pp. 651–676. ISSN: 1369-7412.

[Dal71] D. J. Daley. “The definition of a multi-dimensional generalization of shot noise”. In: *Journal of Applied Probability* 8 (1971), pp. 128–135. ISSN: 0021-9002.

[Dau85] John Daugman. “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters”. In: *Journal of the Optical Society of America A* 2.7 (1985), pp. 1160–1169.

[Dav+18] Axel Davy, Thibaud Ehret, Jean-Michel Morel, and Mauricio Delbracio. “Reducing Anomaly Detection in Images to Detection in Noise”. In: *2018 IEEE International Conference on Image Processing, ICIP 2018, Athens, Greece, October 7-10, 2018*. 2018, pp. 1058–1062.

[Dav77] A. W. Davis. “A differential equation approach to linear combinations of independent chi-squares”. In: *Journal of the American Statistical Association* 72.357 (1977), pp. 212–214. ISSN: 0162-1459.

[DB+19] Valentin De Bortoli, Alain Durmus, Marcelo Pereyra, and Ana Vidal. *Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. Application to maximum marginal likelihood and empirical Bayesian estimation*. 2019. arXiv: 1906.12281.

[DB97] Jeremy De Bonet. “Multiresolution sampling procedure for analysis and synthesis of texture images”. In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1997, Los Angeles, CA, USA, August 3-8, 1997*. 1997, pp. 361–368.

[DD13] Julie Delon and Agnès Desolneux. “A Patch-Based Approach for Removing Impulse or Mixed Gaussian-Impulse Noise”. In: *SIAM Journal on Imaging Sciences* 6.2 (2013), pp. 1140–1174.

[DDT09] Charles-Alban Deledalle, Loïc Denis, and Florence Tupin. “Iterative Weighted Maximum Likelihood Denoising With Probabilistic Patch-Based Weights”. In: *IEEE Transactions on Image Processing* 18.12 (2009), pp. 2661–2672.

[DDT12] Charles-Alban Deledalle, Loïc Denis, and Florence Tupin. “How to compare noisy patches? Patch similarity beyond Gaussian noise”. In: *International Journal of Computer Vision* 99.1 (2012), pp. 86–102. ISSN: 0920-5691.

[DE97] Paul Dupuis and Richard Ellis. *A weak convergence approach to the theory of large deviations*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1997, pp. xviii+479. ISBN: 0-471-07672-4.

[Del96] Bernard Delyon. “General results on the convergence of stochastic algorithms”. In: *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 41.9 (1996), pp. 1245–1255. ISSN: 0018-9286.

[DF81] Persi Diaconis and David Freedman. “On the statistics of vision: The Julesz conjecture”. In: *Journal of Mathematical Psychology* 24.2 (1981), pp. 112 –138. ISSN: 0022-2496.

[DFG09] Randall Douc, Gersende Fort, and Arnaud Guillin. “Subgeometric rates of convergence of f-ergodic strong Markov processes”. In: *Stochastic Processes and their Applications* 119.3 (2009), pp. 897–923. ISSN: 0304-4149.

[DFM16] Alain Durmus, Gersende Fort, and Éric Moulines. “Subgeometric rates of convergence in Wasserstein distance for Markov chains”. In: *Annales de l’Institut Henri Poincare Probabilites et Statistiques* 52.4 (2016), pp. 1799–1822. ISSN: 0246-0203.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. “Adaptive subgradient methods for online learning and stochastic optimization”. In: *Journal of Machine Learning Research (JMLR)* 12 (2011), pp. 2121–2159. ISSN: 1532-4435.

[DJ93] Bernard Delyon and Anatoli Juditsky. “Accelerated stochastic approximation”. In: *SIAM Journal on Optimization* 3.4 (1993), pp. 868–881. ISSN: 1052-6234.

[DK19] Arnak Dalalyan and Avetik Karagulyan. “User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient”. In: *Stochastic Processes and their Applications* 129.12 (2019), pp. 5278–5311. ISSN: 0304-4149.

[DKM17] Adithya Devraj, Ioannis Kontoyiannis, and Sean Meyn. *Geometric Ergodicity in a Weighted Sobolev Space*. 2017. arXiv: 1711.03652.

[DLM99] Bernard Delyon, Marc Lavielle, and Éric Moulines. “Convergence of a stochastic approximation version of the EM algorithm”. In: *The Annals of Statistics* 27.1 (1999), pp. 94–128. ISSN: 0090-5364.

[DLR77] Arthur Dempster, Nan Laird, and Donald Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society. Series B. Methodological* 39.1 (1977), pp. 1–38. ISSN: 0035-9246.

[DM15] Alain Durmus and Éric Moulines. “Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis Adjusted Langevin Algorithm”. In: *Statistics and Computing* 25.1 (2015), pp. 5–19.

[DM17] Alain Durmus and Éric Moulines. “Nonasymptotic convergence analysis for the unadjusted Langevin algorithm”. In: *The Annals of Applied Probability* 27.3 (2017), pp. 1551–1587. ISSN: 1050-5164.

[DM19] Alain Durmus and Éric Moulines. “High-dimensional Bayesian inference via the unadjusted Langevin algorithm”. In: *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability* 25.4A (2019), pp. 2854–2882. ISSN: 1350-7265.

[DMM00] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. “Meaningful Alignments”. In: *International Journal of Computer Vision* 40.1 (2000), pp. 7–23.

[DMM01] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. “Edge Detection by Helmholtz Principle”. In: *Journal of Mathematical Imaging and Vision* 14.3 (2001), pp. 271–284.

[DMM08] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. *From Gestalt theory to image analysis*. Vol. 34. Interdisciplinary Applied Mathematics. Springer, New York, 2008, pp. xii+273. ISBN: 978-0-387-72635-9.

[DMM19] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. “Analysis of Langevin Monte Carlo via convex optimization”. In: *Journal of Machine Learning Research (JMLR)* 20 (2019), Paper No. 73, 46. ISSN: 1532-4435.

[DMP18] Alain Durmus, Éric Moulines, and Marcelo Pereyra. “Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau”. In: *SIAM Journal on Imaging Sciences* 11.1 (2018), pp. 473–506.

[DMR04] Randall Douc, Éric Moulines, and Jeffrey Rosenthal. “Quantitative bounds on convergence of time-inhomogeneous Markov chains”. In: *The Annals of Applied Probability* 14.4 (2004), pp. 1643–1665. ISSN: 1050-5164.

[DMS17] Alain Durmus, Éric Moulines, and Eero Saksman. *On the convergence of Hamiltonian Monte Carlo*. 2017. arXiv: 1705.00166.

[DMT95] Douglas Down, Sean Meyn, and Richard Tweedie. "Exponential and uniform ergodicity of Markov processes". In: *The Annals of Probability* 23.4 (1995), pp. 1671–1691. ISSN: 0091-1798.

[Dou+04] Randall Douc, Gersende Fort, Éric Moulines, and Philippe Soulier. "Practical drift conditions for subgeometric rates of convergence". In: *The Annals of Applied Probability* 14.3 (2004), pp. 1353–1377. ISSN: 1050-5164.

[Dou+18] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018, pp. xviii+757. ISBN: 978-3-319-97703-4; 978-3-319-97704-1.

[DRD18] Arnak Dalalyan and Lionel Riou-Durand. *On sampling from a log-concave density using kinetic Langevin diffusions*. 2018. arXiv: 1807.09382.

[DT12] Arnak Dalalyan and Alexandre Tsybakov. "Sparse regression learning by aggregation and Langevin Monte-Carlo". In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1423–1443. ISSN: 0022-0000.

[Dui+07] Remco Duits, Michael Felsberg, Gösta Granlund, and Bart ter Haar Romeny. "Image analysis and reconstruction using a wavelet transform constructed from a reducible representation of the Euclidean motion group". In: *International Journal of Computer Vision* 72.1 (2007), pp. 79–102.

[Dur+17] Alain Durmus, Gareth Roberts, Gilles Vilmart, and Konstantinos Zygalakis. "Fast Langevin based algorithm for MCMC in high dimensions". In: *The Annals of Applied Probability* 27.4 (2017), pp. 2195–2237.

[Dur16] Alain Durmus. "High dimensional Markov chain Monte Carlo methods: theory, methods and applications". PhD thesis. Paris Saclay, 2016.

[DV75] Monroe Donsker and Srinivasa Varadhan. "Asymptotic evaluation of certain Markov process expectations for large time. I". In: *Communications on Pure and Applied Mathematics* 28 (1975), 1–47; ibid. 28 (1975), 279–301. ISSN: 0010-3640.

[DV76] Monroe Donsker and Srinivasa Varadhan. "Asymptotic evaluation of certain Markov process expectations for large time. III". In: *Communications on Pure and Applied Mathematics* 29.4 (1976), pp. 389–461. ISSN: 0010-3640.

[Ebe11] Andreas Eberle. "Reflection coupling and Wasserstein contractivity without convexity". In: *Comptes Rendus Mathématique. Académie des Sciences. Paris* 349.19-20 (2011), pp. 1101–1104. ISSN: 1631-073X.

[Ebe16] Andreas Eberle. "Reflection couplings and contraction rates for diffusions". In: *Probability Theory and Related Fields* 166.3-4 (2016), pp. 851–886. ISSN: 0178-8051.

[EF01] Alexei Efros and William Freeman. "Image quilting for texture synthesis and transfer". In: *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001, Los Angeles, California, USA, August 12-17, 2001*. 2001, pp. 341–346.

[EGZ18] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. "Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes". In: *Transactions of the American Mathematical Society* (2018).

[EGZ19] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. "Couplings and quantitative contraction rates for Langevin dynamics". In: *The Annals of Probability* 47.4 (2019), pp. 1982–2010. ISSN: 0091-1798.

[EK86] Stewart Ethier and Thomas Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986, pp. x+534. ISBN: 0-471-08186-8.

[EL99] Alexei A. Efros and Thomas K. Leung. "Texture Synthesis by Non-parametric Sampling". In: *Proceedings of the International Conference on Computer Vision, Kerkyra, Corfu, Greece, September 20-25, 1999*. IEEE Computer Society, 1999, pp. 1033–1038.

[EM19] Andreas Eberle and Mateusz Majka. "Quantitative contraction rates for Markov chains on general state spaces". In: *Electronic Journal of Probability* 24 (2019), Paper No. 26, 36.

[FAI05] Ronan Fablet, J.-M Augustin, and Alexandru Isar. "Speckle denoising using a variational multi-wavelet approach". In: *Oceans 2005 - Europe*. Vol. 1. July 2005, 539 –544 Vol. 1. ISBN: 0-7803-9103-9.

[FFC82] Alain Fournier, Donald Fussell, and Loren Carpenter. "Computer Rendering of Stochastic Models". In: *Commun. ACM* 25.6 (1982), pp. 371–384.

[FG19] Wei Fang and Michael Giles. "Multilevel Monte Carlo method for ergodic SDEs without contractivity". In: *Journal of Mathematical Analysis and Applications* 476.1 (2019), pp. 149–176. ISSN: 0022-247X.

[FM03] Gersende Fort and Éric Moulines. "Convergence of the Monte Carlo expectation maximization for curved exponential families". In: *The Annals of Statistics* 31.4 (2003), pp. 1220–1259. ISSN: 0090-5364.

[FMP11] Gersende Fort, Éric Moulines, and Pierre Priouret. "Convergence of adaptive and interacting Markov chain Monte Carlo algorithms". In: *The Annals of Statistics* 39.6 (2011), pp. 3262–3289. ISSN: 0090-5364.

[FNW07] Mário Figueiredo, Robert Nowak, and Stephen Wright. "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems". In: *IEEE Journal of selected topics in signal processing* 1.4 (2007), pp. 586–597.

[For+18] Gersende Fort, Laurent Risser, Yves Atchadé, and Éric Moulines. "Stochastic FISTA Algorithms: So Fast?" In: *2018 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE. 2018, pp. 796–800.

[For02] Gersende Fort. "Computable bounds for V-geometric ergodicity of Markov transition kernels". In: *Rapport de Recherche, Université Joseph Fourier* (2002).

[Fos53] Gordon Foster. "On the stochastic matrices associated with certain queuing processes". In: *Annals of Mathematical Statistics* 24 (1953), pp. 355–360. ISSN: 0003-4851.

[FR05] Gersende Fort and Gareth Roberts. "Subgeometric ergodicity of strong Markov processes". In: *The Annals of Applied Probability* 15.2 (2005), pp. 1565–1589. ISSN: 1050-5164.

[Gal16] Bruno Galerne. "Random fields of bounded variation and computation of their variation intensity". In: *Advances in Applied Probability* 48.4 (2016), pp. 947–971. ISSN: 0001-8678.

[Gar02] Richard Gardner. "The Brunn-Minkowski inequality". In: *American Mathematical Society. Bulletin. New Series* 39.3 (2002), pp. 355–405. ISSN: 0273-0979.

[GC11] Mark Girolami and Ben Calderhead. "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.2 (2011), pp. 123–214.

[GD03] Andrzej Granas and James Dugundji. *Fixed point theory*. Springer Monographs in Mathematics. Springer-Verlag, New York, 2003, pp. xvi+690. ISBN: 0-387-00173-5.

[GD78] Stephen Gull and Geoff Daniell. “Image reconstruction from incomplete and noisy data”. In: *Nature* 272.5655 (1978), pp. 686–690.

[GEB15] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. “Texture Synthesis Using Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. 2015, pp. 262–270.

[GGL19] Nicolas Gonthier, Yann Gousseau, and Saïd Ladjal. “High resolution neural texture synthesis with long range constraints”. working paper or preprint. 2019.

[GGM11] Bruno Galerne, Yann Gousseau, and Jean-Michel Morel. “Random Phase Textures: Theory and Synthesis”. In: *IEEE Transactions on Image Processing* 20.1 (2011), pp. 257–267.

[GHM12] James Gentle, Wolfgang Karl Härdle, and Yuichi Mori. *Handbook of computational statistics: concepts and methods*. Springer Science & Business Media, 2012.

[GL17] Bruno. Galerne and Arthur. Leclaire. “Texture Inpainting Using Efficient Gaussian Conditional Simulation”. In: *SIAM Journal on Imaging Sciences* 10.3 (2017), pp. 1446–1474.

[GLM14] Bruno Galerne, Arthur Leclaire, and Lionel Moisan. “A texton for fast and flexible Gaussian texture synthesis”. In: *22nd European Signal Processing Conference, EUSIPCO 2014, Lisbon, Portugal, September 1-5, 2014*. 2014, pp. 1686–1690.

[GLR18] Bruno Galerne, Arthur Leclaire, and Julien Rabin. “A Texture Synthesis Model Based on Semi-Discrete Optimal Transport in Patch Space”. In: *SIAM Journal on Imaging Sciences* 11.4 (2018), pp. 2456–2493.

[GM06] Benjami Goldys and Bohdan Maslowski. “Lower estimates of transition densities and bounds on exponential ergodicity for stochastic PDE’s”. In: *The Annals of Probability* 34.4 (2006), pp. 1451–1496. ISSN: 0091-1798.

[GM86] André Gagalowicz and Song De Ma. “Model driven synthesis of natural textures for 3-D scenes”. In: *Computers & Graphics* (1986).

[GM94] Ulf Grenander and Michael I Miller. “Representations of knowledge in complex systems”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.4 (1994), pp. 549–581.

[Goo+08] Bart Goossens, Hiep Luong, Aleksandra Pizurica, and Wilfried Philips. “An improved non-local denoising algorithm”. eng. In: *Local and Non-Local Approximation in Image Processing, International Workshop, Proceedings*. Ed. by Jaakko Astola, Karen Egiazarian, and Vladimir Katkovnik. Lausanne, Switzerland, 2008, pp. 143–156.

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 2014, pp. 2672–2680.

[GS81] David Garber and Alexander Sawchuk. “Computational models for texture analysis and synthesis”. In: *Techniques and Applications of Image Understanding*. Vol. 281. International Society for Optics and Photonics. 1981, pp. 254–274.

[GS91] Roger Ghanem and Pol Spanos. *Stochastic finite elements: a spectral approach*. Springer-Verlag, New York, 1991, pp. x+214. ISBN: 0-387-97456-3.

[GZW11] Sven Grewenig, Sebastian Zimmer, and Joachim Weickert. “Rotationally invariant similarity measures for nonlocal image denoising”. In: *Journal on Visual Communication and Image Representation* 22.2 (2011), pp. 117–130.

[Hal83] Peter Hall. “Chi squared approximations to the distribution of a sum of independent random variables”. In: *The Annals of Probability* 11.4 (1983), pp. 1028–1036. ISSN: 0091-1798.

[Han+06] Jianwei Han, Kun Zhou, Li-Yi Wei, Minmin Gong, Hujun Bao, Xinming Zhang, and Baining Guo. “Fast example-based surface texture synthesis via discrete optimization”. In: *The Visual Computer* 22.9-11 (2006), pp. 918–925.

[Har01] Paul Harrison. “A Non-Hierarchical Procedure for Re-Synthesis of Complex Textures”. In: *The 9-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2001, WSCG 2001, University of West Bohemia, Campus Bory, Plzen-Bory, Czech Republic, February 5-9, 2001*. 2001, pp. 190–197.

[Has70] Wilfried Hastings. “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57.1 (1970), pp. 97–109. ISSN: 0006-3444.

[HB95] David Heeger and James Bergen. “Pyramid-based texture analysis/synthesis”. In: *Proceedings 1995 International Conference on Image Processing, Washington, DC, USA, October 23-26, 1995*. IEEE Computer Society, 1995, pp. 648–651.

[HBD17] Antoine Houdard, Charles Bouveyron, and Julie Delon. “High-Dimensional Mixture Models For Unsupervised Image Denoising (HDMI)”. working paper or preprint. June 2017.

[Hel+95] Isaac Held, Raymond Pierrehumbert, Stephen Garner, and Kyle Swanson. “Surface quasi-geostrophic dynamics”. In: *Journal of Fluid Mechanics* 282 (1995), pp. 1–20. ISSN: 0022-1120.

[Hel25] Hermann von Helmholtz. “Treatise on physiological optics”. In: *Rochester, NY: Optical Society of America* (1925).

[HM11] Martin Hairer and Jonathan Mattingly. “Yet another look at Harris' ergodic theorem for Markov chains”. In: *Seminar on Stochastic Analysis, Random Fields and Applications VI*. Springer. 2011, pp. 109–117.

[HMS11] Martin Hairer, Jonathan Mattingly, and Michael Scheutzow. “Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations”. In: *Probability Theory and Related Fields* 149.1-2 (2011), pp. 223–259. ISSN: 0178-8051.

[HS14] Kaiming He and Jian Sun. “Image Completion Approaches Using the Statistics of Similar Patches”. In: *IEEE Transactions Pattern Analysis and Machine Intelligence* 36.12 (2014), pp. 2423–2435.

[HSD73] Robert Haralick, Sam Shanmugam, and Its'hak Dinstein. “Textural Features for Image Classification”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 3.6 (1973), pp. 610–621.

[HSV14] Martin Hairer, Andrew Stuart, and Sebastian Vollmer. “Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions”. In: *The Annals of Applied Probability* 24.6 (2014), pp. 2455–2490. ISSN: 1050-5164.

[HW59] David Hubel and Torsten Wiesel. “Receptive fields of single neurones in the cat's striate cortex”. In: *The Journal of physiology* 148.3 (1959), pp. 574–591.

[Hwa80] Chii-Ruey Hwang. “Laplace's method revisited: weak convergence of probability measures”. In: *The Annals of Probability* 8.6 (1980), pp. 1177–1182. ISSN: 0091-1798.

[IL89] Alexander Ivanov and Nikolai Leonenko. *Statistical analysis of random fields*. Vol. 28. Mathematics and its Applications (Soviet Series). Kluwer Academic Publishers Group, Dordrecht, 1989, pp. x+244. ISBN: 90-277-2800-3.

[IM05] Prakash Ishwar and Pierre Moulin. “On the existence and characterization of the maxent distribution under general moment inequality constraints”. In: *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* 51.9 (2005), pp. 3322–3333. ISSN: 0018-9448.

[Imh61] J. P. Imhof. “Computing the distribution of quadratic forms in normal variables”. In: *Biometrika* 48 (1961), pp. 419–426. ISSN: 0006-3444.

[Iss18] Leon Isserlis. “On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables”. In: *Biometrika* 12.1-2 (1918), pp. 134–139.

[IW89] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*. Second. Vol. 24. North-Holland Mathematical Library. 1989, pp. xvi+555. ISBN: 0-444-87378-3.

[JAF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. “Perceptual Losses for Real-Time Style Transfer and Super-Resolution”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*. 2016, pp. 694–711.

[Jan88] Svante Janson. “Normal convergence by higher semi-invariants with applications to sums of dependent random variables and random graphs”. In: *The Annals of Probability* 16.1 (1988), pp. 305–312. ISSN: 0091-1798.

[Jay57] Edwin Jaynes. “Information theory and statistical mechanics”. In: *Physical Review. Series II* 106 (1957), pp. 620–630. ISSN: 0031-899X.

[Jay79] Edwin Jaynes. “Where do we stand on maximum entropy?” In: *Maximum entropy formalism (Conf., Mass. Inst. Tech., Cambridge, Mass., 1978)*. MIT Press, Cambridge, Mass.-London, 1979, pp. 15–118.

[JBV16] Nikolay Jetchev, Urs Bergmann, and Roland Vollgraf. *Texture Synthesis with Spatial Generative Adversarial Networks*. 2016. arXiv: 1611.08207.

[JGV78] Bela Julesz, Edgar Gilbert, and Jonathan Victor. “Visual discrimination of textures with identical third-order statistics”. In: *Biological Cybernetics* 31.3 (1978), pp. 137–140.

[JH01] Galin. Jones and James Hobert. “Honest exploration of intractable probability distributions via Markov chain Monte Carlo”. In: *Statistical Science. A Review Journal of the Institute of Mathematical Statistics* 16.4 (2001), pp. 312–334. ISSN: 0883-4237.

[JH04] Galin Jones and James Hobert. “Sufficient burn-in for Gibbs samplers for a hierarchical random effects model”. In: *The Annals of Statistics* 32.2 (2004), pp. 784–817. ISSN: 0090-5364.

[JJ81] Jaswant Jain and Anil Jain. “Displacement Measurement and Its Application in Interframe Image Coding”. In: *IEEE Transactions on Communications* 29.12 (1981), pp. 1799–1808. ISSN: 0090-6778.

[JM17] James Johndrow and Jonathan Mattingly. *Error bounds for Approximations of Markov chains used in Bayesian Sampling*. 2017. arXiv: 1711.05382.

[JM83] Peter Jupp and Kanti Mardia. “A note on the maximum-entropy principle”. In: *Scandinavian Journal of Statistics* (1983), pp. 45–47.

[JO10] Ald’eric Joulin and Yann Ollivier. “Curvature, concentration and error estimates for Markov chain Monte Carlo”. In: *The Annals of Probability* 38.6 (2010), pp. 2418–2442.

[JR02] Soren Jarner and Gareth Roberts. “Polynomial convergence rates of Markov chains”. In: *The Annals of Applied Probability* 12.1 (2002), pp. 224–247. ISSN: 1050-5164.

[JT01] Soren Jarner and Richard Tweedie. “Locally contracting iterated functions and stability of Markov chains”. In: *Journal of Applied Probability* 38.2 (2001), pp. 494–507. ISSN: 0021-9002.

[Jul62] Béla Julesz. “Visual Pattern Discrimination”. In: *IRE Transactions on Information Theory* 8.2 (1962), pp. 84–92.

[Jul81] Béla Julesz. "Textons, the elements of texture perception, and their interactions". In: *Nature* 290.5802 (1981), p. 91.

[Kal06] Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

[Kar+19] Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. *Non-asymptotic Analysis of Biased Stochastic Approximation Scheme*. 2019. arXiv: 1902.00629.

[Kas+15] Alexandre Kaspar, Boris Neubert, Dani Lischinski, Mark Pauly, and Johannes Kopf. "Self Tuning Texture Optimization". In: *Computer Graphics Forum* 34.2 (2015), pp. 349–359.

[KB08] Charles Kervrann and Jérôme Boulanger. "Local Adaptivity to Variable Smoothness for Exemplar-Based Image Regularization and Representation". In: *International Journal of Computer Vision* 79.1 (2008), pp. 45–69.

[Kel75] John Kelley. *General topology*. Springer-Verlag, New York-Berlin, 1975, pp. xiv+298.

[Kha11] Rafail Khasminskii. *Stochastic stability of differential equations*. Vol. 66. Springer Science & Business Media, 2011.

[Kim+07] Seung-Jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky. "A method for large-scale l1-regularized least squares". In: *IEEE Journal on Selected Topics in Signal Processing* 1.4 (2007), pp. 606–617.

[KJB67] Samuel Kotz, N. L. Johnson, and D. W. Boyd. "Series representations of distributions of quadratic forms in normal variables. II. Non-central case". In: *Annals of Mathematical Statistics* 38 (1967), pp. 838–848. ISSN: 0003-4851.

[KM17] Ionnais Kontoyiannis and Sean Meyn. "Approximating a diffusion by a finite-state hidden Markov model". In: *Stochastic Processes and their Applications* 127.8 (2017), pp. 2482–2507. ISSN: 0304-4149.

[Kof13] Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 2013.

[KS91] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*. Second. Vol. 113. Graduate Texts in Mathematics. Springer-Verlag, New York, 1991, pp. xxiv+470. ISBN: 0-387-97655-8.

[KT11] Alex Kulesza and Ben Taskar. "Learning Determinantal Point Processes". In: *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*. 2011, pp. 419–427.

[Kul97] Solomon Kullback. *Information theory and statistics*. 1997, pp. xvi+399. ISBN: 0-486-69684-7.

[KW52] Jack Kiefer and Jacob Wolfowitz. "Stochastic estimation of the maximum of a regression function". In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 462–466.

[Kwa+03] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. "Graphcut textures: image and video synthesis using graph cuts". In: *ACM Transactions on Graphics* 22.3 (2003), pp. 277–286.

[Kwa+05] Vivek Kwatra, Irfan A. Essa, Aaron F. Bobick, and Nipun Kwatra. "Texture optimization for example-based synthesis". In: *ACM Transactions on Graphics* 24.3 (2005), pp. 795–802.

[KY03] Harold Kushner and George Yin. *Stochastic approximation and recursive algorithms and applications*. Second. Vol. 35. Applications of Mathematics (New York). Springer-Verlag, New York, 2003, pp. xxii+474. ISBN: 0-387-00894-2.

[Köh92] Wolfgang Köhler. *Gestalt psychology: The definitive statement of the Gestalt theory*. H. Liveright, 1992.

[Ĺ08] Christian Léonard. “Minimization of entropy functionals”. In: *Journal of Mathematical Analysis and Applications* 346.1 (2008), pp. 183–204. ISSN: 0022-247X.

[Lag+10] Ares Lagae, Sylvain Lefebvre, Robert Cook, Tony DeRose, George Drettakis, David Ebert, John Lewis, Ken Perlin, and Matthias Zwicker. “A Survey of Procedural Noise Functions”. In: *Computer Graphics Forum* 29.8 (2010), pp. 2579–2600.

[Lan08] Paul Langevin. “Sur la théorie du mouvement brownien”. In: *Comptes Rendus. Académie des Sciences. Paris* 146 (1908), pp. 530–533.

[LB06] Elizaveta Levina and Peter Bickel. “Texture synthesis and nonparametric resampling of random fields”. In: *The Annals of Statistics* 34.4 (2006), pp. 1751–1773. ISSN: 0090-5364.

[LBM13] Marc Lebrun, Antoni Buades, and Jean-Michel Morel. “A Nonlocal Bayesian Image Denoising Algorithm”. In: *SIAM Journal on Imaging Sciences* 6.3 (2013), pp. 1665–1688.

[LCT03] Yanxi Liu, Robert T. Collins, and Yanghai Tsin. “A Computational Model for Periodic Pattern Perception Based on Frieze and Wallpaper Groups”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.3 (2003), pp. 354–371.

[Leb+12] Marc Lebrun, Miguel Colom, Antoni Buades, and Jean-Michel Morel. “Secrets of image denoising cuisine”. In: *Acta Numer.* 21 (2012), pp. 475–576.

[Leb12] Marc Lebrun. “An Analysis and Implementation of the BM3D Image Denoising Method”. In: *Image Processing On Line* 2 (2012), pp. 175–213.

[Lec15] Arthur Leclaire. “Champs à phase aléatoire et champs gaussiens pour la mesure de netteté d’images et la synthèse rapide de textures”. PhD thesis. Université Paris Descartes, 2015.

[LGX16] Gang Liu, Yann Gousseau, and Gui-Song Xia. “Texture synthesis through convolutional neural networks and spectrum constraints”. In: *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*. IEEE, 2016, pp. 3234–3239.

[Lia+01] Lin Liang, Ce Liu, Ying-Qing Xu, Baining Guo, and Heung-Yeung Shum. “Real-time texture synthesis by patch-based sampling”. In: *ACM Transactions on Graphics* 20.3 (2001), pp. 127–150.

[Lin13] Georg Lindgren. *Stationary stochastic processes*. Chapman & Hall/CRC Texts in Statistical Science Series. CRC Press, Boca Raton, FL, 2013, pp. xxviii+347. ISBN: 978-1-4665-5779-6.

[Lin97] Michael Lindenbaum. “An integrated model for evaluating the amount of data required for reliable recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.11 (1997), pp. 1251–1264.

[LJ12] Sajan Goud Lingala and Mathews Jacob. “A blind compressive sensing frame work for accelerated dynamic MRI”. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2012, pp. 1060–1063.

[LK81] Bruce Lucas and Takeo Kanade. “An Iterative Image Registration Technique with an Application to Stereo Vision”. In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI ’81, Vancouver, BC, Canada, August 24-28, 1981*. 1981, pp. 674–679.

[Low12] David Lowe. *Perceptual organization and visual recognition*. Vol. 5. Springer Science & Business Media, 2012.

[LR86] Torgny Lindvall and L. C. G. Rogers. “Coupling of multidimensional diffusions by reflection”. In: *The Annals of Probability* 14.3 (1986), pp. 860–872. ISSN: 0091-1798.

[LRG18] Holden Lee, Andrej Risteski, and Rong Ge. “Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering Langevin Monte Carlo”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 7847–7856.

[LS01] Robert Liptser and Albert Shiryaev. *Statistics of random processes. I.* expanded. Vol. 5. Applications of Mathematics (New York). Springer-Verlag, Berlin, 2001, pp. xvi+427. ISBN: 3-540-63929-2.

[LT96] Robert Lund and Richard Tweedie. "Geometric convergence rates for stochastically ordered Markov chains". In: *Mathematics of Operations Research* 21.1 (1996), pp. 182–194. ISSN: 0364-765X.

[LTZ09] Huan Liu, Yongqiang Tang, and Hao Helen Zhang. "A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables". In: *Computational Statistics & Data Analysis* 53.4 (2009), pp. 853–856.

[LW16a] Chuan Li and Michael Wand. "Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks". In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*. 2016, pp. 702–716.

[LW16b] Dejun Luo and Jian Wang. "Exponential convergence in L^p-Wasserstein distance for diffusion processes without uniformly dissipative drift". In: *Mathematische Nachrichten* 289.14-15 (2016), pp. 1909–1926. ISSN: 0025-584X.

[LWY97] Hsin-Chih Lin, Ling-Ling Wang, and Shi-Nine Yang. "Extracting periodicity of a regular texture based on autocorrelation functions". In: *Pattern Recognition Letters* 18.5 (1997), pp. 433–443.

[LZW16] Yang Lu, Song-Chun Zhu, and Ying Nian Wu. "Learning FRAME Models Using CNN Filters". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 2016, pp. 1902–1910.

[Ma+19] Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I. Jordan. *Is There an Analog of Nesterov Acceleration for MCMC?* 2019. arXiv: 1902.00996.

[Mad+18] Chris J. Maddison, Daniel Paulin, Yee Whye Teh, Brendan O'Donoghue, and Arnaud Doucet. *Hamiltonian Descent Methods*. 2018. arXiv: 1809.05042.

[MB15] Niklas Mevenkamp and Benjamin Berkels. "Unsupervised and Accurate Extraction of Primitive Unit Cells from Crystal Images". In: *Pattern Recognition - 37th German Conference, GCPR 2015, Aachen, Germany, October 7-10, 2015, Proceedings*. 2015, pp. 105–116.

[MCF15] Yi-An Ma, Tianqi Chen, and Emily Fox. "A complete recipe for stochastic gradient MCMC". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2917–2925.

[MD10] David Mumford and Agnès Desolneux. *Pattern theory: The stochastic analysis of real-world signals*. Applying Mathematics. 2010, pp. xii+407. ISBN: 978-1-56881-579-4.

[Met+53] Nicholas Metropolis, Arianna Rosenbluth, Marshall Rosenbluth, Augusta Teller, and Edward Teller. "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[Mey67] Philip Meyers. "A converse to Banach's contraction theorem". In: *Journal of Research of the National Bureau of Standards* 71B (1967), pp. 73–76. ISSN: 0160-1741.

[Mic01] Daniele Micciancio. "The shortest vector in a lattice is hard to approximate to within some constant". In: *SIAM Journal on Computing* 30.6 (2001), pp. 2008–2035. ISSN: 0097-5397.

[Mil95] Grigori Milstein. *Numerical integration of stochastic differential equations*. Vol. 313. Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht, 1995, pp. viii+169. ISBN: 0-7923-3213-X.

[MJ19] Michael Muehlebach and Michael Jordan. “A Dynamical Systems Perspective on Nesterov Acceleration”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 4656–4662.

[MJ74] Bruce McCormick and Sadali Jayaramamurthy. “Time series model for texture synthesis”. In: *International Journal of Parallel Programming* 3.4 (1974), pp. 329–343.

[MMN83] Takashi Matsuyama, Shu-Ichi Miura, and Makoto Nagao. “Structural analysis of natural textures by Fourier transformation”. In: *Computer Vision, Graphics, and Image Processing* 24.3 (1983), pp. 347–362.

[MMS18] Mateusz B. Majka, Aleksandar Mijatović, and Lukasz Szpruch. *Non-asymptotic bounds for sampling algorithms without log-concavity*. 2018. arXiv: 1808.07105.

[Mon17] Vishal Monga. “Introduction”. In: *Handbook of Convex Optimization Methods in Imaging Science*. Ed. by Vishal Monga. Springer, 2017, pp. 1–13.

[MP84] Michel Métivier and Pierre Priouret. “Applications of a Kushner and Clark lemma to general classes of stochastic algorithms”. In: *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* 30.2, part 1 (1984), pp. 140–151. ISSN: 0018-9448.

[MST08] Ferenc Móricz, Ulrich Stadtmuller, and Monika Thalmaier. “Strong laws for blockwise M-dependent random fields”. In: *Journal of Theoretical Probability* 21.3 (2008), pp. 660–671. ISSN: 0894-9840.

[MT92] Sean Meyn and Richard Tweedie. “Stability of Markovian processes. I. Criteria for discrete-time chains”. In: *Advances in Applied Probability* 24.3 (1992), pp. 542–574. ISSN: 0001-8678.

[MT93a] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993, pp. xvi+ 548. ISBN: 3-540-19832-6.

[MT93b] Sean Meyn and Richard Tweedie. “Stability of Markovian processes. II. Continuous-time processes and sampled chains”. In: *Advances in Applied Probability* 25.3 (1993), pp. 487–517. ISSN: 0001-8678.

[MT93c] Sean Meyn and Richard Tweedie. “Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes”. In: *Advances in Applied Probability* 25.3 (1993), pp. 518–548. ISSN: 0001-8678.

[MT94] Sean Meyn and Richard Tweedie. “Computable bounds for geometric convergence rates of Markov chains”. In: *The Annals of Applied Probability* 4.4 (1994), pp. 981–1011. ISSN: 1050-5164.

[Nem+09] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. “Robust stochastic approximation approach to stochastic programming”. In: *SIAM Journal on optimization* 19.4 (2009), pp. 1574–1609.

[Nes04] Yuri Nesterov. *Introductory lectures on convex optimization*. Vol. 87. Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004, pp. xviii+236. ISBN: 1-4020-7553-7.

[Nes83] Yuri Nesterov. “A method for solving the convex programming problem with convergence rate $O(1/k^2)$”. In: *Doklady Akademii Nauk SSSR* 269.3 (1983), pp. 543–547. ISSN: 0002-3264.

[Nev75] Jacques Neveu. *Discrete-parameter martingales*. Revised. 1975, pp. viii+236.

[New+18] Alasdair Newson, Andrés Almansa, Yann Gousseau, and Saïd Ladjal. “Taking Apart Autoencoders: How do They Encode Geometric Shapes ?” working paper or preprint. Jan. 2018.

[NT78] Esa Nummelin and Richard Tweedie. “Geometric ergodicity and R-positivity for general Markov chains”. In: *The Annals of Probability* 6.3 (1978), pp. 404–420.

[NT82] Esa Nummelin and Pekka Tuominen. “Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory”. In: *Stochastic Processes and their Applications* 12.2 (1982), pp. 187–202. ISSN: 0304-4149.

[NT83] Esa Nummelin and Pekka Tuominen. “The rate of convergence in Orey’s theorem for Harris recurrent Markov chains with applications to renewal theory”. In: *Stochastic Processes and their Applications* 15.3 (1983), pp. 295–311. ISSN: 0304-4149.

[Oll09] Yann Ollivier. “Ricci curvature of Markov chains on metric spaces”. In: *Journal of Functional Analysis* 256.3 (2009), pp. 810–864.

[OLS99] Gyuhwan Oh, Seungyong Lee, and Sung Yong Shin. “Fast determination of textural periodicity using distance matching function”. In: *Pattern Recognition Letters* 20.2 (1999), pp. 191–197.

[OZ81] Samuel Oman and Samuel Zacks. “A mixture approximation to the distribution of a weighted sum of chi-squared variables”. In: *Journal of Statistical Computation and Simulation* 13.3-4 (1981), pp. 215–224.

[Par+09] Minwoo Park, Kyle Brocklehurst, Robert T. Collins, and Yanxi Liu. “Deformed Lattice Detection in Real-World Images Using Mean-Shift Belief Propagation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.10 (2009), pp. 1804–1816.

[Pau16] Daniel Paulin. “Mixing and concentration by Ricci curvature”. In: *Journal of Functional Analysis* 270.5 (2016), pp. 1623–1662. ISSN: 0022-1236.

[PB14] Neal Parikh and Stephen Boyd. “Proximal Algorithms”. In: *Foundations and Trends in Optimization* 1.3 (2014), pp. 127–239.

[Per85] Ken Perlin. “An image synthesizer”. In: *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1985, San Francisco, California, USA, July 22-26, 1985*. 1985, pp. 287–296.

[Pes73] Peter Peskun. “Optimum monte-carlo sampling using markov chains”. In: *Biometrika* 60.3 (1973), pp. 607–612.

[Pey10] Gabriel Peyré. “Texture Synthesis with Grouplets”. In: *IEEE Transactions on Pattern Analysis Machine Intelligence* (2010).

[Pon+07] Nikolay Ponomarenko, Vladimir Lukin, Mikhail Zriakhov, Arto Kaarna, and Jaakko Astola. “An automatic approach to lossy compression of AVIRIS images”. In: *IEEE International Geoscience & Remote Sensing Symposium, IGARSS 2007, July 23-28, 2007, Barcelona, Spain, Proceedings*. 2007, pp. 472–475.

[Pop77] N. N. Popov. “Geometric ergodicity conditions for countable Markov chains”. In: *Doklady Akademii Nauk SSSR* 234.2 (1977), pp. 316–319. ISSN: 0002-3264.

[Pot09] Jürgen Potthoff. “Sample properties of random fields. II. Continuity”. In: *Communications on Stochastic Analysis* 3.3 (2009), pp. 331–348. ISSN: 0973-9599.

[PR77] Stephen Purks and Whitman Richards. “Visual texture discrimination using random-dot patterns”. In: *J. Opt. Soc. Am.* 67.6 (1977), pp. 765–771.

[PS00] Javier Portilla and Eero Simoncelli. “A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients”. In: *International Journal of Computer Vision* 40.1 (2000), pp. 49–70.

[PS14] Natesh Pillai and Aaron Smith. *Ergodicity of Approximate MCMC Chains with Applications to Large Data Sets*. 2014. eprint: 1405.0182.

[PS98] George Pólya and Gabor Szegő. *Problems and theorems in analysis. I.* Classics in Mathematics. Springer-Verlag, Berlin, 1998, pp. xx+389. ISBN: 3-540-63640-4.

[PSW13] Nicholas Polson, James Scott, and Jesse Windle. "Bayesian inference for logistic models using Pólya–Gamma latent variables". In: *Journal of the American statistical Association* 108.504 (2013), pp. 1339–1349.

[PT13] Sam Patterson and Yee Whye Teh. "Stochastic gradient Riemannian Langevin dynamics on the probability simplex". In: *Advances in neural information processing systems*. 2013, pp. 3102–3110.

[QH18] Qian Qin and James Hobert. *Wasserstein-based methods for convergence complexity analysis of MCMC with applications*. 2018. arXiv: 1810.08826.

[QH19] Qian Qin and James Hobert. *Geometric convergence bounds for Markov chains in Wasserstein distance based on generalized drift and contraction conditions*. 2019. arXiv: 1902.02964.

[Raa+17] Lara Raad, Axel Davy, Agnès Desolneux, and Jean-Michel Morel. *A survey of exemplar-based texture synthesis*. 2017. arXiv: 1707.07184.

[RC04] Christian Robert and George Casella. *Monte Carlo Statistical Methods (2nd ed.)* New York: Springer-Verlag, 2004.

[RDM16] Lara Raad, Agnès Desolneux, and Jean-Michel Morel. "A Conditional Multiscale Locally Gaussian Texture Synthesis Algorithm". In: *Journal of Mathematical Imaging and Vision* 56.2 (2016), pp. 260–279.

[Ric44] Stephen Rice. "Mathematical analysis of random noise". In: *Bell System Technical Journal* 23.3 (1944), pp. 282–332.

[Ric77] John Rice. "On generalized shot noise". In: *Advances in Applied Probability* 9.3 (1977), pp. 553–565. ISSN: 0001-8678.

[RM51] Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *The annals of mathematical statistics* (1951), pp. 400–407.

[Ros02] Jeffrey Rosenthal. "Quantitative convergence rates of Markov chains: a simple account". In: *Electronic Communications in Probability* 7 (2002), pp. 123–128. ISSN: 1083-589X.

[Ros95] Jeffrey Rosenthal. "Minorization conditions and convergence rates for Markov chain Monte Carlo". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 558–566. ISSN: 0162-1459.

[RP94] Gareth Roberts and Nicholas Polson. "On the Geometric Convergence of the Gibbs Sampler". In: *Journal of the Royal Statistical Society, Series B* 56 (1994), pp. 377–384.

[RR96] Gareth Roberts and Jeffrey Rosenthal. "Quantitative bounds for convergence rates of continuous time Markov processes". In: *Electronic Journal of Probability* 1 (1996), no. 9, approx. 21 pp. ISSN: 1083-6489.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. *Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis*. 2017. arXiv: 1702.03849.

[RS18] Daniel Rudolf and Nikolaus Schweizer. "Perturbation theory for Markov chains via Wasserstein distance". In: *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability* 24.4A (2018), pp. 2610–2639. ISSN: 1350-7265.

[RT96] Gareth Roberts and Richard Tweedie. “Exponential convergence of Langevin distributions and their discrete discrete approximations”. In: *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability* 2.4 (1996), pp. 341–363. ISSN: 1350-7265.

[RT99] Gareth Roberts and Richard Tweedie. “Bounds on regeneration times and convergence rates for Markov chains”. In: *Stochastic Processes and their Applications* 80.2 (1999), pp. 211–229. ISSN: 0304-4149.

[Rud06] Walter Rudin. *Real and complex analysis*. Tata McGraw-hill education, 2006.

[RW09] Christian Robert and Darren Wraith. “Computational methods for Bayesian model choice”. In: *Aip conference proceedings*. Vol. 1193. 1. AIP. 2009, pp. 251–262.

[RY99] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Third. Vol. 293. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1999, pp. xiv+602. ISBN: 3-540-64325-7.

[Sal10] Joseph Salmon. “On Two Parameters for Denoising With Non-Local Means”. In: *IEEE Signal Process. Lett.* 17.3 (2010), pp. 269–272.

[Sar42] Arthur Sard. “The measure of the critical values of differentiable maps”. In: *Bulletin of the American Mathematical Society* 48 (1942), pp. 883–890. ISSN: 0002-9904.

[SB84] John Skilling and RK Bryan. “Maximum entropy image reconstruction-general algorithm”. In: *Monthly notices of the royal astronomical society* 211 (1984), p. 111.

[SDDB19] Alexandre Saint-Dizier, Julie Delon, and Charles Bouveyron. “A unified view on patch aggregation”. In: *Journal of Mathematical Imaging and Vision* (2019), pp. 1–20.

[Sha48] Claude Shannon. “A mathematical theory of communication”. In: *The Bell System Technical Journal* 27 (1948), pp. 379–423, 623–656. ISSN: 0005-8580.

[Sim+16] Umut Simsekli, Roland Badeau, A. Taylan Cemgil, and Gaël Richard. “Stochastic Quasi-Newton Langevin Monte Carlo”. In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by Maria-Florina Balcan and Kilian Q. Weinberger. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 642–651.

[SL14] Xiahan Sang and James LeBeau. “Revolving scanning transmission electron microscopy: Correcting sample drift distortion without prior knowledge”. In: *Ultramicroscopy* 138 (2014), pp. 28 –35. ISSN: 0304-3991.

[SM13] Laurent Sifre and Stéphane Mallat. “Rotation, Scaling and Deformation Invariant Scattering for Texture Discrimination”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. 2013, pp. 1233–1240.

[ST11] Ulrich Stadtmuller and Le Van Thanh. “On the strong limit theorems for double arrays of blockwise M-dependent random variables”. In: *Acta Mathematica Sinica (English Series)* 27.10 (2011), pp. 1923–1934. ISSN: 1439-8516.

[Sur15] Frédéric Sur. “An a-contrario approach to quasi-periodic noise removal”. In: *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*. 2015, pp. 3841–3845.

[SV06] Daniel Stroock and Srinivasa Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin, 2006, pp. xii+338. ISBN: 978-3-540-28998-2; 3-540-28998-4.

[SW08] Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2008, pp. xii+693. ISBN: 978-3-540-78858-4.

[SZ13] Ohad Shamir and Tong Zhang. "Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes". In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. Vol. 28. JMLR Workshop and Conference Proceedings. JMLR.org, 2013, pp. 71–79.

[SZ14] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014. arXiv: 1409.1556.

[Top79] Flemming Topsøe. "Information-theoretical optimization techniques". In: *Kybernetika (Prague)* 15.1 (1979), pp. 8–27. ISSN: 0023-5954.

[TT90] Denis Talay and Luciano Tubaro. "Expansion of the global error for numerical schemes solving stochastic differential equations". In: *Stochastic Analysis and Applications* 8.4 (1990), 483–509 (1991). ISSN: 0736-2994.

[TT94] Pekka Tuominen and Richard Tweedie. "Subgeometric rates of convergence of f-ergodic Markov chains". In: *Advances in Applied Probability* 26.3 (1994), pp. 775–798. ISSN: 0001-8678.

[TTV16] Yee Whye Teh, Alexandre Thiery, and Sebastian Vollmer. "Consistency and fluctuations for stochastic gradient Langevin dynamics". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 193–225.

[Tur91] Greg Turk. "Generating textures on arbitrary surfaces using reaction-diffusion". In: *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1991, Providence, RI, USA, April 27-30, 1991*. 1991, pp. 289–298.

[TV93] Marc Teboulle and Igor Vajda. "Convergence of best ϕ-entropy estimates". In: *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* 39.1 (1993), pp. 297–301. ISSN: 0018-9448.

[Uly+16] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. "Texture Networks: Feed-forward Synthesis of Textures and Stylized Images". In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. 2016, pp. 1349–1357.

[Ust+16] Ivan Ustyuzhaninov, Wieland Brendel, Leon A. Gatys, and Matthias Bethge. *Texture Synthesis Using Shallow Convolutional Networks with Random Filters*. 2016. arXiv: 1606.00021.

[UTY95] Michael Unser, Philippe Thévenaz, and Leonid Yaroslavsky. "Convolution-based interpolation for fast, high-quality rotation of images". In: *IEEE Transactions on Image Processing* 4.10 (1995), pp. 1371–1381.

[UVL17] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. "Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2017, pp. 4105–4113.

[Ver97] A. Yu. Veretennikov. "On polynomial mixing bounds for stochastic differential equations". In: *Stochastic Processes and their Applications* 70.1 (1997), pp. 115–127. ISSN: 0304-4149.

[VG+08] Rafael Grompone Von Gioi, Jeremie Jakubowicz, Jean-Michel Morel, and Gregory Randall. "LSD: A fast line segment detector with a false detection control". In: *IEEE transactions on pattern analysis and machine intelligence* 32.4 (2008), pp. 722–732.

[Vid+19] Ana Vidal, Valentin De Bortoli, Marcelo Pereyra, and Alain Durmus. *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach*. 2019. arXiv: 1911.11709 [stat.ME].

[Vil09] Cédric Villani. *Optimal transport*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 2009, pp. xxii+973. ISBN: 978-3-540-71049-3.

[VK04] A. Yu. Veretennikov and S. A. Klokov. "On the subexponential rate of mixing for Markov processes". In: *Teoriya Veroyatnostei i ee Primeneniya* 49.1 (2004), pp. 21–35. ISSN: 0040-361X.

[VP18] Ana Fernandez Vidal and Marcelo Pereyra. "Maximum Likelihood Estimation of Regularisation Parameters". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 1742–1746.

[VZT16] Sebastian Vollmer, Konstantinos Zygalakis, and Yee Whye Teh. "Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 5504–5548.

[Wak13] Jon Wakefield. *Bayesian and frequentist regression methods*. Springer Series in Statistics. Springer, New York, 2013, pp. xx+697. ISBN: 978-1-4419-0924-4; 978-1-4419-0925-1.

[Wan+04] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[Wan94] Feng-Yu Wang. "Application of coupling methods to the Neumann eigenvalue problem". In: *Probability Theory and Related Fields* 98.3 (1994), pp. 299–306. ISSN: 0178-8051.

[Wei+09] Li-Yi Wei, Sylvain Lefebvre, Vivek Kwatra, and Greg Turk. "State of the art in example-based texture synthesis". In: *Eurographics 2009, State of the Art Report, EG-STAR*. Eurographics Association. 2009, pp. 93–117.

[Wer+77] Stephen Wernecke et al. "Maximum entropy image reconstruction". In: *IEEE Transactions on Computers* 4 (1977), pp. 351–364.

[Wer23] Max Wertheimer. "Untersuchungen zur Lehre von der Gestalt. II". In: *Psychologische forschung* 4.1 (1923), pp. 301–350.

[Wij91] Jarke van Wijk. "Spot noise texture synthesis for data visualization". In: *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1991, Providence, RI, USA, April 27-30, 1991*. 1991, pp. 309–318.

[Wil91] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991, pp. xvi+251. ISBN: 0-521-40455-X; 0-521-40605-6.

[WK91] Andrew Witkin and Michael Kass. "Reaction-diffusion textures". In: *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1991, Providence, RI, USA, April 27-30, 1991*. 1991, pp. 299–308.

[WL00] Li-Yi Wei and Marc Levoy. "Fast texture synthesis using tree-structured vector quantization". In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co. 2000, pp. 479–488.

[WM13] Yi-Qing Wang and Jean-Michel Morel. "SURE Guided Gaussian Mixture Image Denoising". In: *SIAM Journal on Imaging Sciences* 6.2 (2013), pp. 999–1034.

[Woo89] Andrew Wood. "An F Approximation to the Distribution of a Linear Combination of Chi-squared Variables." In: *Communications in Statistics - Simulation and Computation* 18.4 (1989), pp. 1439–1456.

[WSB03] Z. Wang, E. P. Simoncelli, and A. C. Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*. Vol. 2. 2003, 1398–1402 Vol.2.

[WT11] Max Welling and Yee Whye Teh. “Bayesian Learning via Stochastic Gradient Langevin Dynamics”. In: *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. Ed. by Lise Getoor and Tobias Scheffer. Omnipress, 2011, pp. 681–688.

[Wu+13] Yue Wu, Brian Tracey, Premkumar Natarajan, and Joseph Noonan. “Probabilistic Non-Local Means”. In: *IEEE Signal Process. Lett.* 20.8 (2013), pp. 763–766.

[WW59] Max Wertheimer and Michael Wertheimer. *Productive thinking*. Harper New York, 1959.

[Xia+14] Gui-Song Xia, Sira Ferradans, Gabriel Peyré, and Jean-François Aujol. “Synthesizing and Mixing Stationary Gaussian Texture Models”. In: *SIAM Journal on Imaging Sciences* 7.1 (2014), pp. 476–508.

[Yel93] John Yellott. “Implications of triple correlation uniqueness for texture statistics and the Julesz conjecture”. In: *Journal of the Optical Society of America* 10.5 (1993), pp. 777–793.

[Zho+18] Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. “Non-stationary texture synthesis by adversarial expansion”. In: *ACM Transactions on Graphics* 37.4 (2018), 49:1–49:13.

[Zhu99] Song Chun Zhu. “Embedding Gestalt Laws in Markov Random Fields”. In: *IEEE Transactions on Pattern Analysis Machine Intelligence* 21.11 (1999), pp. 1170–1187.

[ZT80] Steven Zucker and Demetri Terzopoulos. “Finding structure in co-occurrence matrices for texture analysis”. In: *Computer Graphics and Image Processing* 12.3 (1980), pp. 286 –308. ISSN: 0146-664X.

[ZWM98] Song Chun Zhu, Ying Nian Wu, and David Mumford. “Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling”. In: *International Journal of Computer Vision* 27.2 (1998), pp. 107–126.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre: Statistiques non locales dans les images : modélisation, estimation et échantillonnage

Mots clés: A contrario, redondance spatiale, maximum d'entropie, optimisation stochastique, algorithme de Langevin, convergence de chaîne de Markov

Résumé: Dans cette thèse, on étudie d'un point de vue probabiliste deux statistiques non locales dans les images : la redondance spatiale et les moments de certaines couches de réseaux de neurones convolutionnels. Plus particulièrement, on s'intéresse à l'estimation et à la détection de la redondance spatiale dans les images naturelles et à l'échantillonnage de modèles d'images sous contraintes de moments de sorties de réseaux de neurones.

On commence par proposer une définition de la redondance spatiale dans les images naturelles. Celle-ci repose sur une analyse Gestaltiste de la notion de similarité ainsi que sur un cadre statistique pour le test d'hypothèses via la méthode a contrario. On développe un algorithme pour identifier cette redondance dans les images naturelles. Celui-ci permet d'identifier les patchs similaires dans une image. On utilise cette information pour proposer de nouveaux algorithmes de traitement d'image (débruitage, analyse de périodicité).

Le reste de cette thèse est consacré à la modélisation et à l'échantillonnage d'images sous contraintes non locales. Les modèles d'images considérés sont obtenus via le principe de maximum d'entropie. On peut alors déterminer la distribution cible sur les images via une procédure de minimisation. On aborde ce problème en utilisant des outils issus de l'optimisation stochastique.

Plus précisément, on propose et analyse un nouvel algorithme pour l'optimisation stochastique : l'algorithme SOUL (Stochastic Optimization with Unadjusted Langevin). Dans cette méthodologie, le gradient est estimé par une méthode de Monte Carlo par chaîne de Markov (ici l'algorithme de Langevin non ajusté). Les performances de cet algorithme repose sur les propriétés de convergence ergodiques des noyaux de Markov associés aux chaînes de Markov utilisées. On s'intéresse donc aux propriétés de convergence géométrique de certaines classes de modèles fonctionnels autorégressifs. On caractérise précisément la dépendance des taux de convergence de ces modèles vis à vis des constantes du modèle (dimension, régularité, convexité...).

Enfin, on applique l'algorithme SOUL au problème de synthèse de texture par maximum d'entropie. On étudie les liens qu'entretient cette approche avec d'autres modèles de maximisation d'entropie (modèles macrocanoniques, modèles microcanoniques). En utilisant des statistiques de moments de sorties de réseaux de neurones convolutionnels on obtient des résultats visuels comparables à ceux de l'état de l'art.

Title: Non-local statistics in images: modelisation, estimation and sampling

Keywords: A contrario, spatial redundancy, maximum of entropy, stochastic optimization, Langevin algorithm, Markov chain convergence

Abstract: In this thesis we study two non-local statistics in images from a probabilistic point of view: spatial redundancy and convolutional neural network features. More precisely, we are interested in the estimation and detection of spatial redundancy in natural images. We also aim at sampling images with neural network constraints.

We start by giving a definition of spatial redundancy in natural images. This definition relies on two concepts: a Gestalt analysis of the notion of similarity in images, and a hypothesis testing framework (the a contrario method). We propose an algorithm to identify this redundancy in natural images. Using this methodology we can detect similar patches in images and, with this information, we propose new algorithms for diverse image processing tasks (denoising, periodicity analysis).

The rest of this thesis deals with sampling images with non-local constraints. The image models we consider are obtained via the maximum entropy principle. The target distribution is then obtained by minimizing an energy functional. We use tools from stochastic optimization to tackle this problem.

More precisely, we propose and analyze a new algorithm: the SOUL (Stochastic Optimization with Unadjusted Langevin) algorithm. In this methodology, the gradient is estimated using Monte Carlo Markov Chains methods. In the case of the SOUL algorithm we use an unadjusted Langevin algorithm. The efficiency of the SOUL algorithm is related to the ergodic properties of the underlying Markov chains. Therefore we are interested in the convergence properties of certain class of functional autoregressive models. We characterize precisely the dependency of the convergence rates of these models with respect to their parameters (dimension, smoothness, convexity).

Finally, we apply the SOUL algorithm to the problem of examplar-based texture synthesis with a maximum entropy approach. We draw links between our model and other entropy maximization procedures (macrocanonical models, microcanonical models). Using convolutional neural network constraints we obtain state-of-the art visual results.