Towards a mathematical understanding of deep neural network using a mean field analysis

Arnaud Descours

UNIVERSITÉ Clermont Auvergne

lmbp

Thèse

pour obtenir le grade de

Docteur en mathématiques

de l'Université Clermont Auvergne

Vers une compréhension mathématique des réseaux neuronaux profonds par une analyse champ moyen

présentée et soutenue publiquement par

Arnaud Descours

le vendredi 20 octobre 2023, devant le jury composé de

M. Francis Bach	Directeur de recherche, INRIA, ENS	Rapporteur
M. Alain Durmus	Professeur, École polytechnique	Rapporteur
M. Benjamin Jourdain	Professeur, CERMICS, École des Ponts	Président du jury
Mme Michela Ottobre	Professeur associé, Heriot-Watt University	Examinateur
M. Liming Wu	Professeur, UCA	Examinateur
M. Arnaud Guillin	Professeur, UCA	Directeur de thèse
Mme Manon Michel	Chargé de recherches, CNRS, UCA	Directeur de thèse
M. Boris Nectoux	Professeur assistant, UCA	Directeur de thèse

Laboratoire en Mathématiques Blaise Pascal, UMR 6620, CNRS, Université Clermont Auvergne, Aubière.

À mes grands-parents.

Table des matières

Remerciements

De tous les corps ensemble on ne saurait en faire réussir une petite pensée, cela est impossible et d'un autre ordre. De tous les corps et esprits on n'en saurait tirer un mouvement de vraie charité, cela est impossible et d'un autre ordre, surnaturel.

Pascal, *Pensées.*

Je remercie tout d'abord mes directeurs de thèse : Arnaud Guillin, Manon Michel et Boris Nectoux. Arnaud et Manon, merci de m'avoir fait confiance, de m'avoir guidé pour mon orientation et de m'avoir permis de travailler avec Tom Huix et Éric Moulines. Boris, je te remercie tout particulièrement pour les longues heures que tu as passées à refaire tous mes calculs, à m'avoir expliqué tant de choses, ainsi que pour ton soutien constant et tes encouragements si motivants. Je remercie ensuite mes collaborateurs, Tom Huix et Éric Moulines, avec qui ce fut un plaisir de travailler, notamment lors de mes passages à Paris. Enfin, je remercie mes rapporteurs Francis Bach et Alain Durmus, pour leur lecture attentive de mes travaux, ainsi que mes examinateurs, Benjamin Jourdain, Michela Ottobre et Liming Wu. Je remercie également Aurélien Garivier, grâce à qui j'ai trouvé cette thèse, et qui m'a gentiment invité à présenter mes travaux à l'ENS Lyon, ainsi que Rémi Gribonval pour son accueil.

Trois ans de travail de thèse ne se réalisant pas sans soutien extérieur, je voudrais ici rendre hommage à ceux qui ont, à leur manière – indirectement, mais non moins puissamment – contribué à la réalisation de ce travail. Et tout d'abord mes parents, que je ne saurais trop remercier pour leur soutien indéfectible. Viennent ensuite ces chères amitiés nouées à Clermont-Ferrand. En premier lieu, Élise (certes pour ta proximité géographique, mais aussi pour toutes nos discussions et nos fous rires), Solenne (surtout pour m'avoir fait découvrir les restaurants les plus kitch de la rue des Gras) et Jean-Claude (pour la leçon de droiture et de courage que tu nous as donnée). Merci aussi à Aurélien (pour les rencontres que j'ai pu faire grâce à toi... et le barbecue sous l'orage), à Jean et Chloé (pour les moments simples et chaleureux passés ensemble), aux deux Jérémy, à Denis (nous n'oublierons pas nos déjeuners à Crousti Pain...), à Jean-Luc, à Benoît, à Blanche, à Thémys et tant d'autres que j'oublie. Je remercie également Amandine, pour son soutien tout au long de la deuxième année, ainsi que François et Catherine Hou.

Je tiens à remercier les membres du laboratoire, que j'ai apprécié croiser durant ces trois ans, notamment Thierry Lambre (pour votre gaieté communicative, ainsi que votre conversation, sur l'agrégation et sur Blaise Pascal), Laurent Serlet (pour sa bonne humeur, qui a égayé mes déambulations jusqu'au club), Andrzej Stos (pour son aide précieuse pour les TP de Python), ainsi que Dominique Manchon. Merci également à Valérie Sourlier, pour votre aide dans les démarches administratives, que vous avez souvent faites à ma place – parce que « ça ira plus vite » – après me les avoir pourtant expliquées des dizaines de fois, avec autant de patience et de gentillesse que la première fois. Sans oublier, évidemment, les doctorants : Baptiste Peaucelle, Athina Monemvassitis, Arthur Bottois, Tristan Guyon, Vincent Souveton, Léo Hahn, Sue Claret, Julian Le Clainche, Clément Legrand, Rémi Boutin et Martin Azon. Merci enfin à ceux que j'ai le plus côtoyés (enfin, sauf Nathan, qui devait vite nous laisser pour un TD ou pour « la muscu »)

durant cette thèse : mes cobureaux. Outre Nathan Couchet donc, avec qui nous avons beaucoup ri et beaucoup discuté sur le système éducatif, parti trop tôt à Metz avec Bob, que nous regrettons également, je remercie Mohamed Ayadi (tu auras été un voisin particulièrement agréable, et avec qui j'aurais eu des discussions très profondes) et enfin, bien sûr, merci à Geoffrey Lacour : il me serait difficile de te remercier à la hauteur de ce que je te dois. Scientifiquement d'abord, où ta si vaste culture alliée à ton âme de chercheur ont nourri et orné nos discussions quotidiennes ; ma progression n'aurait pas été la même sans nos échanges. Plus prosaïquement ensuite, merci pour nos innombrables fous rires, pauses cafés, « Call Of », séances de sport, road trip, etc., ainsi bien sûr que pour ton soutien dans les moments difficiles et tes conseils toujours judicieux.

Notations

- $\mathcal{P}(X)$: espace des mesures de probabilité sur X, où X est un espace métrique.
- $\mathcal{P}_p(\mathbf{R}^d)$: espace de Wasserstein d'ordre $p \geq 1$, défini comme le sous-ensemble de $\mathcal{P}(\mathbf{R}^d)$ composé des mesures admettant un moment d'ordre p fini. Voir appendice A.2.
- $\mathcal{P}_c(\mathbf{R}^d)$: espace des mesures de probabilités définies sur $\mathbf{R}^d$ à support compact.
- $\mathcal{C}_b^\infty(\mathbf{R}^d)$: espace des fonctions $f : \mathbf{R}^d \to \mathbf{R}$ de classe $\mathcal{C}^\infty$, bornées, et dont les dérivées de tout ordre sont bornées.
- $\mathcal{D}(I, E)$ où $I = [0, T]$ ou $I = [0, \infty)$: espace des fonctions càdlàg définies sur I et à valeurs dans E (espace métrique). Voir appendice A.1.
- $\mathcal{H}^L(U)$: espace de Sobolev $W^{L,2}(U)$, où U est un ouvert de $\mathbf{R}^n$. Voir appendice A.3.
- $\mathcal{H}^{L,\gamma}(U)$. Espace de Sobolev à poids. Voir section 2.1.2.
- $\mathcal{C}_c^\infty(\mathbf{R}^d)$: espace des fonctions $f : \mathbf{R}^d \to \mathbf{R}$ de classe $\mathcal{C}^\infty$, à support compact.
- $\langle f, \mu \rangle = \int_{\mathbf{R}^d} f(x)\mu(\mathrm{d}x)$ pour $\mu \in \mathcal{P}(\mathbf{R}^d)$ et une fonction μ-intégrable $f : \mathbf{R}^d \to \mathbf{R}$, ou $\mu(f)$ pour $\mu \in \mathcal{H}^{-L,\gamma}(U)$ et $f \in \mathcal{H}^{L,\gamma}(U)$.
- $\lfloor x \rfloor$: partie entière de $x \in \mathbf{R}$, c'est-à-dire le plus grand entier n tel que $n \leq x$.

Chapitre 1

Introduction

Le propre du sage est de mettre de l'ordre.

Aristote, *Métaphysique.*

L'apprentissage automatique, ou *Machine learning*, déjà central dans de nombreuses technologies de nos sociétés (recherches sur internet, reconnaissance et génération d'images, de textes, de sons) est au cœur des recherches visant à développer celles de demain : voiture autonome, aide au diagnostic médical, etc. Un des principaux moyens de réaliser un apprentissage automatique est d'utiliser un réseau de neurones, fonction dont le principe de base est d'appliquer une transformation non linéaire (à l'aide d'une *fonction d'activation*) à une donnée d'entrée. Les capacités de calcul aujourd'hui disponibles permettent d'itérer ces transformations - réaliser de multiples transformations élémentaires - et de réaliser ainsi un apprentissage profond (*deep learning* [GBC16]) ; chaque *couche* du réseau de neurones apprend alors un aspect de l'information contenue dans les données d'entraînement. Cette structure en couches des réseaux de neurones leur confère une grande *expressivité* : ils peuvent approcher de vastes classes de fonctions, et leur structure s'adapte à des jeux de données variés. Les performances exceptionnelles des réseaux de neurones reposent essentiellement sur les capacités de calcul aujourd'hui à notre disposition, en adaptant empiriquement des millions de paramètres qui interagissent *a priori* de manière assez chaotique, et auxquels il est bien difficile de donner une interprétation. Ceci est problématique à plusieurs points de vue. D'une part, d'un point de vue scientifique : outre la soif de compréhension qui anime l'esprit humain, les problèmes auxquels nous faisons face aujourd'hui (par exemple en biologie, en dynamique moléculaire, en astrophysique) requièrent des simulations numériques qui dépassent nos capacités de calcul ; développer de nouveaux algorithmes, plus performants et plus rapides, s'avère nécessaire. D'autre part, les innovations de demain, mentionnées précédemment, permettront d'aider à la prise de décision et même de prendre des décisions de manière entièrement autonome, ce qui posera des questions éthiques et juridiques. Réfléchir sereinement à ces questions et légiférer en conséquence exigera donc une compréhension profonde du fonctionnement des méthodes utilisées.

L'objet de cette thèse est d'étudier théoriquement des problèmes d'apprentissages supervisés, c'est-à-dire des problèmes où l'on cherche à prédire une donnée $y \in \mathcal{Y}$ à partir d'une donnée $x \in \mathcal{X}$. On distingue les problèmes de régression (où y est un réel) et les problèmes de classification (où $y \in \{1, \dots, \mathsf{c}\}$, c étant le nombre de classes[1]). Pour entraîner un réseau de neurones, on se donne un jeu de données d'entraînement $D = \{(x_i, y_i)\}_{i=1}^{M} \subset (\mathcal{X} \times \mathcal{Y})^M$, où les (x_i, y_i) sont tirés idépendamment suivant une loi inconnue $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Étant donné un réseau de neurones f_W, l'objectif est de trouver les meilleurs poids W de sorte que $f_W(x_i) \approx y_i$, pour $i = 1, \dots, M$.

[1] En pratique, si $y = i$, on utilise le *one-hot encoding* : $y = (0, \dots, 0, 1, 0 \dots, 0)$, le 1 étant placé en position i.

Plus précisément, on considère une fonction de coût $\mathsf{L} : \mathbf{R} \times \mathbf{R} \to \mathbf{R}_+$ qui quantifie l'erreur entre $f_W(x_i)$ et y_i. L'objectif est alors de minimiser (sur W) l'erreur $\hat{\mathcal{R}}$ commise sur le jeu de données D (appelée aussi *risque empirique*)

$$\hat{\mathcal{R}}(W, D) = \frac{1}{|D|} \sum_{(x,y)\in D} \mathsf{L}(y, f_W(x)) = \frac{1}{M} \sum_{i=1}^{M} \mathsf{L}(y_i, f_W(x_i)).$$

Trouver

$$W^* = \arg\min_W \hat{\mathcal{R}}(W, D)$$

est un problème délicat, du fait de la grande dimension des poids W et des fonctions non linéaires constituant le réseau de neurones. On cherchera donc à approcher W^* par descente de gradient. Remarquons maintenant que si l'on s'entraîne sur un nombre fini de données, le réseau de neurones risque d'apprendre non pas la véritable distribution des données π, mais la loi empirique $\frac{1}{M}\sum_{i=1}^{M} \delta_{(x_i,y_i)}$. C'est le problème de l'*overfitting*. En fait, la véritable quantité à minimiser est le *risque populationnel*

$$\mathcal{R}(W) = \mathbf{E}_{(X,Y)\sim\pi}[\mathsf{L}(Y, f_W(X))].$$

Cette quantité étant inconnue, la méthode utilisée en pratique est de subdiviser le jeu de données en deux : un jeu de données d'entraînement (grâce auquel on va optimiser les poids W), et un jeu de données de test (sur lequel on n'entraîne pas le réseau de neurones, et grâce auquel on approximera $\mathcal{R}(W)$ par la moyenne empirique $\hat{\mathcal{R}}(W_k, D_{\text{test}})$). Pour entraîner le réseau, on utilise l'algorithme de descente de gradient stochastique[2] : pour $k \geq 0$,

$$\begin{cases} W_{k+1} = W_k - \alpha_k \nabla \hat{\mathcal{R}}(W_k, D_k) \\ W_0 \sim P_{ini} \end{cases} \tag{1.0.1}$$

où $D_k \subset D$ est une partie du jeu de données ; considérer l'ensemble du jeu de données à chaque itération serait trop coûteux, et favoriserait l'*overfitting* (bien que dans le cas de réseaux surparamétrés, on observe un phénomène de double descente [dRBK20]). On appelle *époque* le temps (le nombre d'itérations) au bout duquel l'ensemble du jeu de données a été vu par le le réseau. Subdivisons par exemple le jeu de données en N parties disjointes :

$$D = \cup_{i=0}^{\mathsf{N}-1} D_i.$$

Une fois passée la première époque, l'entraînement se poursuit en évaluant la fonction de coût avec des données déjà utilisées : (1.0.1) s'écrira donc

$$\begin{cases} W_{k+1} = W_k - \alpha_k \nabla \hat{\mathcal{R}}(W_k, D_{k \bmod \mathsf{N}}) \\ W_0 \sim P_{ini} \end{cases}$$

À la fin de chaque époque, on calcule $\hat{\mathcal{R}}(W_k, D_{\text{test}})$. Cette quantité va logiquement décroître, puis se mettre à croître une fois que le réseau apprendra davantage la distribution empirique $\frac{1}{M}\sum_{i=1}^{M} \delta_{(x_i,y_i)}$ que la distribution des données π. Il sera alors temps d'arrêter l'entraînement. Une autre manière d'avoir de bonnes propriétés de généralisation est d'ajouter un terme de pénalisation à la fonction de coût :

$$\hat{\mathcal{R}}(W, D) = \frac{1}{|D|} \sum_{(x,y)\in D} \mathsf{L}(y, f_W(x)) + \mathsf{P}(W)$$

[2]appelé stochastique car on utilise l'estimateur $\hat{\mathcal{R}}(W)$ de $\mathcal{R}(W)$. Notons aussi qu'il existe des variantes de cet algorithme (Adam, AdamW, Lion), que nous ne discuterons pas ici.

où, par exemple, $\mathsf{P}(W) = \|W\|_1$ (lasso) ou $\mathsf{P}(W) = \|W\|_2$ (ridge).

Pour entraîner un réseau de neurones, il a donc fallu faire un certain nombre de choix : le modèle utilisé (nombre de couches, nombre de neurones, fonctions d'activations), la fonction de coût, l'algorithme de minimisation. Ces choix reposent pour une bonne part sur des études expérimentales ; l'objet des études mathématiques est d'une part d'appuyer et d'expliquer ces choix, et d'autre part de proposer de nouveaux modèles et algorithmes, plus performants, moins sensibles aux perturbations et moins coûteux en temps de calcul. Trois grandes pistes de recherche peuvent être identifiées :

(i) les recherches concernant le modèle utilisé, c'est-à-dire les questions relatives à l'expressivité du réseau de neurone considéré. Quelles fonctions peuvent être approchées par ce réseau, et à quelle vitesse ? Par exemple, le théorème d'approximation universelle stipule qu'un réseau de neurones à deux couches avec une fonction d'activation sigmoïdale peut approcher uniformément toute fonction continue sur un compact. Ce résultat, qui correspond au théorème de Weierstrass d'approximation des fonctions continues par des polynômes, n'est cependant pas quantitatif. Or, quand on approche une fonction f^*, par exemple par des polynômes ou des séries de Fourier, les erreurs typiques sont de la forme

$$\|f^* - f_N\|_{\mathcal{H}} \leq \frac{C}{N^{1/d}},$$

où d la dimension des données d'entrées, N est le nombre de paramètres du modèle, et $\mathcal{H}$ est un espace de fonction. Le fait que la dimension d apparaisse dans les bornes obtenues est communément appelé le *fléau de la dimension* ("curse of dimensionality"), puisque les données avec lesquelles on souhaite entraîner les réseaux de neurones sont en grande dimension. Pour revenir à l'exemple du théorème d'approximation universelle, on peut obtenir des bornes ne dépendant pas de la dimension, en considérant les espaces de Barron [Bar93]. Plus généralement, les espaces de fonctions naturellement associés aux réseaux de neurones sont essentiellement des espaces de Sobolev ou de Besov. Pour un travail récent dans ce domaine, nous renvoyons le lecteur à [GKNV22].

(ii) les recherches sur le paysage de la fonction de coût ; en particulier sur le nombre de minima locaux et la taille de leur bassin d'attraction, la nature des points critiques, etc. L'idée générale est que le paysage se simplifie quand la taille du réseau augmente. D'après [MWW$^+$20], on ne dispose pas encore de vision générale sur ce sujet de recherche, bien que des résultats partiels aient été obtenus, par exemple avec des fonctions d'activations linéaires quadratiques [DL18, SJL19].

(iii) les recherches concernant le processus d'apprentissage. Des algorithmes très simples de descente de gradient sont capables d'atteindre des minima de fonctions très complexes, jusqu'à atteindre une erreur nulle sur le jeu de données d'entraînement. L'explication de la convergence de ces algorithmes reste ouverte. De plus, il semble y avoir un phénomène de régularisation implicite lorsque le nombre de paramètres du réseau est largement supérieur au nombre de données d'entraînement, qui fait que les minima locaux atteints par ces algorithmes ont de bonnes performances de généralisation, bien que le réseau ait été entraîné sur un jeu de données fixé, et sans terme de pénalisation [CB20, VRF22].

Dans cette thèse, nous nous intéresserons au processus d'entraînement des réseaux de neurones. Nous utiliserons une approche champ moyen, utilisée classiquement pour les systèmes de particules en interaction.

L'introduction de cette thèse est organisée de la manière suivante. Nous exposerons d'abord, en section 1.1, l'origine et les méthodes des approches champ moyen, ainsi que le type de résultats qu'elles permettent d'obtenir. Ensuite, en section 1.2, nous appliquerons ces méthodes au cas des

réseaux de neurones classiques et présenterons les résultats que nous avons obtenus. Nous ferons ensuite de même en section 1.3 dans le cadre des réseaux de neurones bayésiens.

1.1 Approche champ moyen pour les systèmes de particules

À partir de la fin du XIXè siècle, les modélisations physiques sont de différentes natures :

- déterministes, lorsque la modélisation ne fait pas intervenir d'aléa. Exemple : la mécanique classique.
- stochastiques, lorsque la modélisation fait intervenir l'aléa. Exemple : le mouvement brownien.

Unifier de manière cohérente ces différentes modélisations constitue une partie du sixième problème de Hilbert, énoncé en 1900.

Sixième problème de Hilbert : Traitement mathématique des axiomes en physique. Les investigations sur les fondements de la géométrie suggèrent le problème : *Traiter de la même manière, au moyen d'axiomes, les sciences physiques dans lesquelles les mathématiques jouent déjà aujourd'hui un rôle important ; au premier rang figurent la théorie des probabilités et la mécanique.*

En effet, unifier les différentes théories physiques est un élément important pour traiter mathématiquement les sciences physiques et pouvoir ainsi leur donner une base axiomatique cohérente. Or, les équations de la physique qui découlent des différentes modélisations décrivent le monde à différentes échelles. Les approches champ moyen ont pour objet de relier une description microscopique à une description macroscopique. Considérons par exemple N particules de masse m, de position X^i et de vitesse V^i soumises à une force $\mathbf{F}$ dépendant des positions de chaque particule. La deuxième loi de Newton s'écrit

$$m\frac{\mathrm{d}V_t^i}{\mathrm{d}t} = -\sum_{j=1}^{N} \mathbf{F}(X_t^i - X_t^j).$$

Lorsque le nombre de particules N tend vers l'infini, le terme de force risque de diverger. Il convient donc de réaliser un changement d'échelle, c'est-à-dire de passer d'une description discrète à une description continue, en écrivant par exemple l'équation sous la forme

$$m\frac{\mathrm{d}V_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} \mathbf{F}(X_t^i - X_t^j),$$

forme qui permet d'espérer trouver une limite lorsque $N \to \infty$. Une question naturelle est alors, ayant prescrit la position initiale des particules (ou la loi de ces positions), de savoir quelle est la position des particules à un temps t ultérieur, dans la limite $N \to \infty$. Une autre question naturelle, au regard de considérations thermodynamiques, concerne le chaos moléculaire : si les états initiaux des particules sont indépendants,

$$\mathcal{L}(X_0^1, \dots, X_0^N) = \mathcal{L}(X_0^1) \otimes \cdots \otimes \mathcal{L}(X_0^N),$$

cette propriété reste-t-elle vraie à des temps ultérieurs, dans la limite $N \to \infty$? C'est Kac [Kac56] qui a donné une définition mathématique précise du chaos moléculaire et introduit la notion de propagation du chaos, dans le cadre de l'étude de l'équation de Boltzmann.

Définition 1.1. *Soit E un espace métrique et $P \in \mathcal{P}(E)$. On dit qu'une suite $(P_N)_{N\geq 1}$ de distribution symétriques, telles que $P_N \in \mathcal{P}(E^N)$, est P-chaotique si pour tout entier k, la k-ème marginale de P_N converge faiblement vers $P^{\otimes k}$; autrement dit, si, pour $\phi_1, \dots, \phi_k \in \mathcal{C}_b(E)$,*

$$\lim_{N\to\infty} \int_{E^N} \phi_1(x_1)\dots\phi_k(x_k) P_N(\mathrm{d}x_1, \dots, \mathrm{d}x_N) = \prod_{i=1}^{k} \int_E \phi_i(x) P(\mathrm{d}x).$$

Définition 1.2 (Propagation du chaos)**.** *Considérons un système de particules $(X_t^i)_{i\geq 1, t\geq 0}$ définies sur un espace métrique E, et une famille $(P_t)_{t\geq 0} \subset \mathcal{P}(E)$. Supposons que la loi de $(X_0^1, \dots, X_0^N)$ soit P_0-chaotique. On dit qu'il y a propagation du chaos jusqu'au temps $T > 0$, si, pour tout $t \in [0, T]$, la loi de $(X_t^1, \dots, X_t^N)$ est P_t-chaotique.*

Après le travail de Kac, McKean [MJ69] introduisit une classe de diffusions qui satisfont la propriété de propagation du chaos, donnant ainsi une nature stochastique aux équations cinétiques. Ces idées fécondes ont trouvé des applications non seulement en théorie cinétique [BGSR16, GSRT13, MM13, BFOZ18], mais aussi en biologie et en sciences sociales [NPT10, MT14, ABF$^+$19], en théorie des jeux à champ moyen (où le nombre de joueurs devient grand) [CDLL19, Car10, CD$^+$18] et en optimisation [PTTM17, GP21, CJLZ21]. Les systèmes de particules et leurs limites champ moyen sont devenus des notions centrales en mathématiques appliquées [CD22].

Introduisons maintenant les outils mathématiques propres à l'étude des limites champ moyen. On n'étudiera pas les particules à travers le N-uplet $(X_t^1, \dots, X_t^N)$ pour deux raisons : la dimension de l'espace $(\mathbf{R}^d)^N$ évolue avec N et tend vers l'infini, et l'ordre des particules n'importe pas. On considérera donc plutôt la mesure empirique des particules

$$\mu_t^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i} \in \mathcal{P}(\mathbf{R}^d).$$

À noter que dans le cas de données initiales stochastiques, μ_t^N est une mesure de probabilité aléatoire. La mesure empirique permet d'ailleurs de démontrer des résultats de propagation du chaos.

Proposition 1.3 ([Szn91, Proposition 2.2])**.** *Soit $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$. En notant P^N la loi jointe de $(X_1, \dots, X_N)$, supposée symétrique, on a l'équivalence :*

- *μ^N converge faiblement en loi vers la loi déterministe P.*
- *P^N est P-chaotique.*

On va voir dans un premier temps comment obtenir des limites champ moyen, puis nous nous demanderons comment obtenir des résultats plus fins (trajectoriels ou quantitatifs).

Limite champ moyen à temps fixé. On se donne deux modèles, l'un microscopique (Mi) et l'autre macroscopique (Ma). Dans le cas de données initiales déterministes, on dispose de μ_0^N. Supposons que

$$\mu_0^N \to \mu_0$$

au sens de la convergence faible des mesures de probabilité. Soit alors μ_t la solution de (Ma) avec pour donnée initiale μ_0. A-t-on, pour tout $t > 0$,

$$\mu_t^N \to \mu_t \ ?$$

Dans le cas de données initiales aléatoires, on a par exemple,

$$\mu_0^N \to \mu_0$$

presque sûrement, avec μ_0 déterministe. On se demande alors de même si $\mu_t^N \to \mu_t$, presque sûrement, en probabilité, selon une certaine distance, au sens faible, etc.

Exemple : l'équation de Vlasov. Utilisée en physique des plasma, cette équation décrit l'évolution temporelle de la densité de particules f_t de charge q et de masse m, dans l'espace des phases (position x, vitesse v). La densité f_t satisfait

$$\partial_t f_t = -v \cdot \nabla_x f_t - \frac{1}{m} \overline{\mathbf{F}}_t \cdot \nabla_v f_t, \tag{1.1.1}$$

où

$$\overline{\mathbf{F}}_t(x) = \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} \mathbf{F}(x - x') f_t(x', v') \mathrm{d}x' \mathrm{d}v'.$$

En testant (1.1.1) contre une fonction g lisse à support compact sur l'espace des phases, on obtient, par intégration par parties, en notant μ_t la loi définie par f_t,

$$\partial_t \mu_t(g) = \mu_t(\nabla_x g \cdot v) - \frac{1}{m} \mu_t \times \mu_t(\mathbf{F}(x - x') \cdot \nabla_v g), \tag{1.1.2}$$

où

$$\mu_t \times \mu_t(\mathbf{F}(x - x') \cdot \nabla_v g) = \int_{\mathbf{R}^d \times \mathbf{R}^d} \int_{\mathbf{R}^d \times \mathbf{R}^d} \mathbf{F}(x - x') \cdot \nabla_v g(x, v) f_t(x, v) f_t(x', v') \mathrm{d}x \mathrm{d}v \mathrm{d}x' \mathrm{d}v'.$$

Cette équation admet une unique solution dans l'espace des mesures de probabilité. Considérons un système de N particules satisfaisant aux équations de Newton

$$\begin{cases} \frac{\mathrm{d}X_t^i}{\mathrm{d}t} = V_t^i, \\ m \frac{\mathrm{d}V_t^i}{\mathrm{d}t} = -\frac{1}{N} \sum_{j=1}^N \mathbf{F}(X_t^i - X_t^j), \end{cases} \tag{1.1.3}$$

et à une condition initiale $X_0^i = x_0^i$. Alors

$$\mu_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i} \delta_{V_t^i} \tag{1.1.4}$$

satisfait (1.1.2). On a alors le résultat suivant

Théorème 1.4 ([Spo12, Corollary 5.2])**.** *Soit* $\mathbf{F}$ *une force d'interaction bornée et lipschitzienne, et* x_0^i *des données initiales et* μ *telles que* $d_{BL}(\mu_0^N, \mu) \to 0$*, alors en notant* μ_t *la solution de* (1.1.2) *avec donnée initiale* μ*, on a* $d_{BL}(\mu_t^N, \mu_t) \to 0$ *où* μ_t^N *est donnée par* (1.1.3) *et* (1.1.4)*.*

L'hypothèse sur $\mathbf{F}$ est forte, et ne prend pas en compte des forces coulombienne ou gravitationnelle. On verra par la suite des résultats moins restrictifs sur $\mathbf{F}$ et où la vitesse de convergence sera explicite.

Limite champ moyen trajectorielle. Jusqu'ici, nous avons voulu savoir si, pour tout $t > 0$, $\mu_t^N \to \mu_t$. Cette convergence est-elle uniforme en temps ? Pour répondre à cette question, on peut chercher à montrer des limites champ moyen non plus sur μ_t^N, mais sur

$$t \mapsto \mu_t^N.$$

Ce processus (déterministe ou stochastique), s'il est continu en temps, appartient à l'espace $\mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d))$. Typiquement, on montre alors que la loi de ce processus - qui est donc un élément de $\mathcal{P}(\mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d)))$ - converge vers une mesure de Dirac concentrée sur $t \mapsto \mu_t$, la solution de (Ma). C'est cette approche que nous utilisons dans nos théorèmes 1.10, 1.22 et 1.23. Pour

obtenir plus de précision sur la convergence vers la limite champ moyen, on peut s'intéresser aux fluctuations autour de la limite champ moyen, en cherchant une limite (à temps fixe ou trajectorielle) à

$$\eta_t^N = \sqrt{N}(\mu_t^N - \mu_t).$$

Dans le cas où $\eta_t^N \to \eta_t$, on aura alors un résultat de type théorème central limite :

$$\mu_t^N \approx \mu_t + \frac{1}{\sqrt{N}}\eta_t.$$

Le cadre mathématique convenable pour l'étude des fluctuations η_t^N ne va pas de soi. En effet, on peut voir naturellement η_t^N comme une mesure signée. Or l'espace des mesures signées n'est pas métrisable ; une approche classique [FM97a, JM98b, DLR19b] est donc de considérer η_t^N comme un élément du dual d'un espace de Sobolev $H^{-L}(\mathbf{R}^d) = W^{-L,2}(\mathbf{R}^d)$, où L est suffisamment grand (nous donnons une introduction à ces espaces dans l'appendice A.3). La structure hilbertienne permet d'utiliser les théorèmes classiques de tension de la mesure valable pour des espaces métriques. C'est cette approche que nous suivrons dans le théorème 1.17. On peut également restreindre l'étude au processus $t \mapsto \eta_t^N(f) \in \mathcal{C}(\mathbf{R}_+, \mathbf{R})$, où f est une fonction test (ou plus généralement $t \mapsto \Phi(\eta_t^N) \in \mathcal{C}(\mathbf{R}_+, \mathbf{R})$, voir par exemple [JT21]). Une approche alternative pour obtenir plus de précision sur la convergence $\mu_t^N \to \mu_t$ consiste à quantifier l'écart entre μ_t^N et μ_t. On présente par la suite quelques résultats relatifs à l'équation de Vlasov.

Retour sur l'équation de Vlasov. Considérons de nouveau l'équation (1.1.2) et le système (1.1.3) avec données initiales déterministes. On a le résultat suivant

Théorème 1.5 ([HJ14, Theorem 1])**.** *Si* $\mathbf{F}$ *satisfait, pour tout* $x \in \mathbf{R}^d - \{0\}$*,*

$$|\nabla F(x)| \leq \frac{C}{|x|^\alpha} \text{ et } |\nabla \mathbf{F}(x)| \leq \frac{C}{|x|^{\alpha+1}}, \tag{S_α}$$

pour $C > 0$ *et* $\alpha < 1$*, alors il existe, pour tout* $0 < \gamma < 1$*, un constante* C_0 *telle que pour tout* $T > 0$ *et* N *suffisamment grand,*

$$\forall t \in [0,T],\ \mathsf{W}_1(\mu_t^N, \mu_t) \leq e^{C_0 t}\Big(\mathsf{W}_1(\mu_0^N, \mu_0) + 2N^{-\frac{\gamma}{2d}}\Big),$$

où W_1 *désigne la distance de Wasserstein d'ordre 1 (voir appendice A.2).*

Remarquons également que les hypothèses faites ici sur $\mathbf{F}$ sont plus faibles, bien qu'elles ne prennent pas en compte des forces de Coulomb ou de gravitation, pour des interactions à courte distance.

Citons également un résultat quantitatif de propagation du chaos :

Théorème 1.6 ([HJ14, Theorem 2])**.** *Si* $\mathbf{F}$ *satisfait* (S_α) *et* $(X_0^i, V_0^i)_{1\leq i\leq N} \sim \mu_0^{\otimes N}$ *sont les données initiales de* (1.1.3)*, il existe, pour tout* $T > 0$*, des constantes* $C_0, C_1 > 0$ *telles que pour* N *suffisamment grand,*

$$\mathbf{P}\Big(\exists t \in [0,T],\ \mathsf{W}_1(\mu_t^N, \mu_t) \geq 3e^{C_0 t}N^{-\frac{\gamma}{2d}}\Big) \leq \frac{C_1}{N^s}$$

où $s = \frac{\gamma d - (2-\gamma)\alpha - 2}{3(\alpha+1)}$.

En utilisant la proposition 1.3, ainsi que le fait que la convergence en distance W_1 implique la convergence faible, on obtient qu'il y a propagation du chaos au sens de la définition 1.2. Pour d'autres approches récentes sur l'équation de Vlasov, on pourra consulter par exemple

[LP17]. Une approche probabiliste à l'étude des processus de McKean-Vlasov est développée dans [Szn91, GKM+96]. Pour des systèmes déterministes, on pourra consulter le cours [Gol16].

Dans cette thèse, nous appliquerons les méthodes mathématiques propres aux limites champ moyen trajectorielles au cas des réseaux de neurones. **En effet, les poids du réseau de neurones (ou les paramètres de la famille variationnelle, dans le cas des réseaux bayésiens) peuvent être vus comme des particules en interaction, l'interaction entre les particules ayant lieu à travers l'apprentissage sur des données communes.** Nous formaliserons cela plus précisément dans la section suivante.

1.2 Approche champ moyen pour les réseaux de neurones classiques

L'objectif de cette section est d'appliquer les méthodes présentées à la section précédente dans le cadre d'un réseau de neurones classique à une couche cachée. Nous établissons en section 1.2.1 la modélisation champ moyen pour ces réseaux de neurones et présentons les résultats de limite champ moyen. En section 1.2.2, nous présentons nos résultats concernant les fluctuations autour de la limite champ moyen, sous forme de théorème central limite.

1.2.1 Limite champ moyen : loi des grands nombres

On s'intéresse à un réseau de neurones à une couche cachée avec une paramétrisation champ moyen :

$$g_W^N(x) := \frac{1}{N}\sum_{i=1}^{N} \sigma_*(W^i, x), \tag{1.2.1}$$

où x est la donnée d'entrée, $W = (W^1, \dots, W^N) \in (\mathbf{R}^d)^N$ représente les poids du réseau de neurones, et $g_W^N(x)$ la prédiction faite par le réseau. L'objectif est de minimiser le risque populationnel, c'est-à-dire de trouver

$$W^* \in \arg\min_W \mathcal{R}(W). \tag{$\mathcal{P}_W$}$$

A cause des non linéarités (les fonctions d'activation) constituant le réseau de neurones, la fonction $W \mapsto \mathcal{R}(W)$ n'est pas convexe, ce qui rend le problème difficile à résoudre. On va voir qu'on peut se ramener à un problème convexe, en se plaçant dans l'espace des mesures de probabilité. En notant $\nu = \frac{1}{N}\sum_{i=1}^{N} \delta_{W^i}$, on peut écrire

$$g_W^N(x) = \langle \sigma_*(\cdot, x), \nu \rangle,$$

qui est linéaire en ν. Introduisons maintenant

$$\mathscr{R} : \mu \in \mathcal{P}(\mathbf{R}^d) \mapsto \mathbf{E}_{(X,Y)\sim\pi}[\mathsf{L}(\langle \sigma_*(\cdot, X), \mu \rangle, Y)]. \tag{1.2.2}$$

Si L est une fonction convexe, il en sera de même pour $\mathscr{R}$, et le problème qui consiste à trouver

$$\mu^* \in \arg\min_\mu \mathscr{R}(\mu) \tag{$\mathcal{P}_\mu$}$$

sera donc *a priori* plus simple à résoudre que $(\mathcal{P}_W)$, tout en étant plus général. La résolution de $(\mathcal{P}_W)$ s'effectue par descente de gradient, de type (1.0.1). On considérera un taux d'apprentissage (*learning rate*) fixe au cours de l'entraînement. L'algorithme s'écrit

$$\begin{cases} W_{k+1} = W_k - \alpha_k \nabla \hat{\mathcal{R}}(W_k, B_k) \\ W_0 \sim P_{ini} \end{cases} \tag{1.2.3}$$

L'objectif de l'approche champ moyen pour les réseaux de neurones est de relier (1.0.1) **au problème** $(\mathcal{P}_\mu)$**, et d'obtenir ainsi des informations sur** (1.0.1)**.** Nous allons désormais tâcher de trouver, par des calculs informels, quelle équation sera satisfaite par la dynamique dans la limite $N \to \infty$. On se place dans le cas d'une fonction de coût quadratique

$$\mathsf{L}(x, y) = \frac{1}{2}(x - y)^2. \tag{1.2.4}$$

Quelques calculs heuristiques. Considérons tout d'abord, pour simplifier, une descente de gradient sur le risque populationnel

$$\mathcal{R}(W) = \mathbf{E}_{(X,Y)\sim\pi}[\mathsf{L}(Y, f_W(X))], \tag{1.2.5}$$

au lieu du risque empirique. L'algorithme s'écrit donc

$$W_{k+1} = W_k - \alpha\nabla\mathcal{R}(W_k). \tag{1.2.6}$$

Pour comprendre la dynamique (1.2.6) dans le cas $N = \infty$, on va paramétrer cet algorithme de manière continue, puis chercher une équation différentielle satisfaite par la mesure empirique des poids du réseau. Cherchons d'abord une écriture de cet algorithme en temps continu. Comment relier un paramètre continu t au paramètre discret k ? En utilisant (1.2.5) et (1.2.4), l'algorithme (1.2.6) se réécrit

$$W^i_{k+1} = W^i_k - \frac{\alpha}{N}\mathbf{E}[(g^N_{W_k}(X) - Y)\nabla_W\sigma_*(W^i_k, X)]. \tag{1.2.7}$$

Ici, le facteur $\frac{1}{N}$ provient de la normalisation champ moyen en (1.2.1). Cette écriture montre que la différence entre W_k et W_{k+1} est de l'ordre de $1/N$. Ainsi, si l'on considère N itérations de l'algorithme, les poids W auront varié de l'ordre de $\alpha\mathbf{E}[(g^N_{W_k}(X) - Y)\nabla_W\sigma_*(W^i_k, X)]$, quantité indépendante de N. Prendre en compte cette observation est nécessaire dans le changement d'échelle qui accompagne le passage en temps continu, puisque l'on va chercher une limite $N \to \infty$. Définissons donc

$$\widetilde{W}_t = W_{\lfloor Nt\rfloor},\ t \geq 0,$$

où l'on a relié le nombre de pas de l'algorithme au nombre de neurones N, qui doivent être du même ordre de grandeur. On a, pour $t = k/N$,

$$\widetilde{W}_{t+\frac{1}{N}} = W_{k+1} = W_k - \alpha\nabla\mathcal{R}(W_k) = \widetilde{W}_t - \alpha\nabla\mathcal{R}(\widetilde{W}_t),$$

d'où

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{W}_t \approx N(\widetilde{W}_{t+\frac{1}{N}} - \widetilde{W}_t) = -\alpha N\nabla\mathcal{R}(\widetilde{W}_t)$$

On définira donc la dynamique en temps continu de la manière suivante (on écrira W_t pour $\widetilde{W}_t$ quand il n'y a pas de confusion possible) :

$$\frac{\mathrm{d}}{\mathrm{d}t}W_t = -\alpha N\nabla\mathcal{R}(W_t). \tag{1.2.8}$$

On a alors, pour $i = 1, \ldots, N$,

$$\frac{\mathrm{d}}{\mathrm{d}t}W^i_t = -\alpha\mathbf{E}[(g^N_{W_t}(X) - Y)\nabla_W\sigma_*(W^i_t, X)].$$

Introduisons maintenant la mesure empirique des poids du réseau entraîné par cet algorithme

$$\nu^N_t = \frac{1}{N}\sum_{i=1}^N \delta_{W^i_t}.$$

Cherchons une équation d'évolution satisfaite par la distribution empirique des poids du réseau. Pour une fonction test $f : \mathbf{R}^d \to \mathbf{R}$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle f, \nu_t^N\rangle &= \frac{1}{N}\sum_{i=1}^N \frac{\mathrm{d}}{\mathrm{d}t} f(W_t^i) \\ &= \frac{1}{N}\sum_{i=1}^N \frac{\mathrm{d}}{\mathrm{d}t} W_t^i \cdot \nabla f(W_t^i) = \frac{1}{N}\sum_{i=1}^N \mathbf{E}[\alpha(Y - g_{W_t}^N(X))\nabla_W \sigma_*(W_t^i, X)\cdot \nabla f(W_t^i)] \\ &= \langle \mathbf{E}[\alpha(Y - g_{W_t}^N(X))\nabla_W \sigma_*(\cdot, X)\cdot \nabla f], \nu_t^N\rangle \\ &= \langle \mathbf{E}[\alpha(Y - \langle \sigma_*(\cdot, X), \nu_t^N\rangle)\nabla_W \sigma_*(\cdot, X)\cdot \nabla f], \nu_t^N\rangle. \end{aligned} \tag{1.2.9}$$

Cette équation peut se réécrire en fonction de la variation première de $\mathscr{R}$ (voir (1.2.2)).

Définition 1.7 ([San15, Définition 7.12])**.** *Soit $F : \mathcal{P}(\mathbf{R}^d) \to \mathbf{R}$. On appelle, si elle existe, variation première de F en $\mu \in \mathcal{P}(\mathbf{R}^d)$, toute fonction mesurable $\frac{\delta F}{\delta \mu} : \mathbf{R}^d \to \mathbf{R}$ telle que*

$$\lim_{\varepsilon \to 0} \frac{F(\mu + \varepsilon\chi) - F(\mu)}{\varepsilon} = \int_{\mathbf{R}^d} \frac{\delta F}{\delta \mu}(w)\chi(\mathrm{d}w),$$

pour tout $\chi = \mu - \rho$ où $\rho \in \mathcal{P}(\mathbf{R}^d) \cap L_c^\infty(\mathbf{R}^d)$ (mesure de probabilité qui admet une densité bornée à support compact par rapport à la mesure de Lebesgue).

Notons que la variation première est unique à une constante près. Rappelons la définition de $\mathscr{R}$:

$$\mathscr{R} : \mu \in \mathcal{P}(\mathbf{R}^d) \mapsto \frac{1}{2}\mathbf{E}_{(X,Y)\sim\pi}[(\langle \sigma_*(\cdot, X), \mu\rangle - Y)^2].$$

La variation première de $\mathscr{R}$ en μ est, pour $\mu \in \mathcal{P}(\mathbf{R}^d)$ et $w \in \mathbf{R}^d$,

$$\frac{\delta \mathscr{R}}{\delta \mu}(w) = \mathbf{E}[(\langle \sigma_*(\cdot, X), \mu\rangle - Y)\sigma_*(w, X)].$$

L'équation (1.2.9) signifie que ν_t^N satisfait au sens faible l'équation[3]

$$\partial_t \nu_t^N = -\alpha \mathrm{div}\Big(\nu_t^N \nabla\Big(\frac{\delta \mathscr{R}}{\delta \nu_t^N}\Big)\Big). \tag{1.2.10}$$

A ce stade, plusieurs pistes de recherche se dessinent :

(i) Passer à la limite dans (1.2.10).

(ii) Établir rigoureusement cette limite si l'on remplace ν_t^N par la mesure empirique des poids des algorithmes (1.2.6) ou (1.2.3).

(iii) Relier cette limite au problème $(\mathcal{P}_\mu)$.

C'est à ces problèmes que nous nous sommes intéressés dans cette première partie de la thèse.

État de l'art relatif aux questions *(i)-(iii)*. Dans [RVE22], les auteurs établissent des réponses informelles aux trois questions et appellent la communauté à les établir rigoureusement.

With this in mind, we adopt a presentation style that relies on formal asymptotic arguments to derive our results, though we are confident that providing rigorous proofs to our propositions is achievable.[4]

[3] Nous commenterons cette équation de transport à la suite du théorème 1.10.

[4] Voir le preprint [RVE18b].

Plus précisément, les auteurs s'intéressent au cas d'une fonction de coût quadratique et de données générées par une fonction oracle $g^* : \mathcal{X} \to \mathcal{Y}$, c'est-à-dire que $(x, g^*(x)) \in \mathcal{X} \times \mathcal{Y}$, où g^* et $\pi_x \in \mathcal{P}(\mathcal{X})$, la loi de x, sont inconnues. Ils montrent notamment qu'on peut espérer obtenir

Proposition 1.8 ([RVE22, Proposition 3.5])**.** *Soit ν_t^N la solution de* (1.2.8) *avec pour condition initiale $W_0^i \sim \mu_0^{\otimes N}$ et μ_t la solution de* (1.2.10) *avec condition initiale μ_0. En faisant des hypothèses sur le support et les moments de μ_0 et en supposant que $\cap_{w\in\mathbf{R}^d}\{\sigma_*(w,\cdot)\}^{\perp_{L^2(\pi_x)}} = \{0\}$ et que g^* est représentable (il existe μ telle que $g(x) = \langle\sigma_*(\cdot, x), \mu\rangle\ \forall x \in \mathcal{X}$) on a que si $(\mu_t)_{t\geq 0}$ admet une limite lorsque $t \to \infty$, alors*

$$\lim_{t,N\to\infty} \langle\sigma_*(\cdot, x), \nu_t^N\rangle = g^*(x), \text{ pour } \pi_x\text{-presque partout.}$$

Dans [CB18a], les auteurs se placent dans un cadre légèrement différent de (1.2.8) : ils considèrent un flot gradient de m particules défini sur $\mathscr{R}$, c'est-à-dire que

$$u : \mathbf{R}_+ \to (\mathbf{R}^d)^m$$

est défini par la dynamique

$$u'(t) \in -m\partial\mathscr{R}_m(u(t)),$$

où $\mathscr{R}_m : u \in (\mathbf{R}^d) \mapsto \mathscr{R}(\frac{1}{m}\sum_{i=1}^m \delta_{u_i})$, et $\partial\mathscr{R}_m(u(t))$ désigne la sous-différentielle de $\mathscr{R}_m$ en $u(t)$. En utilisant des méthodes d'analyse de flots gradients développées par [AGS05], les auteurs démontrent le

Théorème 1.9 ([CB18a, Proposition 2.3 et Théorème 3.3])**.** *Sous des hypothèses de différentiabilité de $\mathscr{R}$ et de régularité sur la fonction d'activation σ_*, il existe un unique flot gradient absolument continu u. En notant $\mu_{m,t} = \frac{1}{m}\sum_{i=1}^m \delta_{u_i(t)}$ et en supposant que $\mathsf{W}_2(\mu_{m,0}, \mu_0) \to_{m\to\infty} 0$, on a de plus que $(m_{m,t})_{t\geq 0}$ converge trajectoriellement vers la solution de l'équation* (1.2.10) *avec donnée initiale μ_0. Enfin, si l'on fait en plus des hypothèses sur le support de μ_0, de convexité sur $\mathscr{R}$ et d'homogénéité sur σ_*, on a que si $\mathsf{W}_2(\mu_t, \mu_\infty) \to_{t\to\infty} 0$ pour un certain $\mu_\infty \in \mathcal{P}_2(\mathbf{R}^d)$, alors μ_∞ est un minimum global de $\mathscr{R}$. On a alors, en particulier,*

$$\lim_{m,t\to\infty} \mathscr{R}(\mu_{m,t}) = \min_{\mu\in\mathcal{P}(\mathbf{R}^d)} \mathscr{R}(\mu).$$

Mentionnons également [Woj20] qui considère une fonction d'activation ReLU. Concernant *(i)* et *(ii)*, [MMN18] ont montré une convergence à temps fixé de la mesure empirique des poids du réseau de neurones entraîné par (1.2.3) dans le cas d'un batch de taille 1 ($|B_k| = 1$ dans (1.2.3)), ainsi que dans le cas bruité - nous reparlerons plus loin plus en détails de ce cas bruité. Enfin, [SS20b] ont donné des pistes de démonstration rigoureuse de la convergence trajectorielle de le mesure empirique des poids de (1.2.3) vers la solution de (1.2.10). **Notre objectif sera d'exploiter ces idées et démontrer rigoureusement une limite trajectorielle de la dynamique** (1.2.3) **vers la dynamique** (1.2.10)**, afin d'apporter une solution aux problèmes *(i)* et *(ii)*.**

Cas d'étude, hypothèses et résultats. Détaillons l'algorithme étudié. Nous considèrerons un algorithme plus général que (1.2.3), au sens où un bruit artificiel est ajouté à chaque itération de l'algorithme :

$$\begin{cases} W_{k+1} = W_k - \alpha_k\nabla\hat{\mathcal{R}}(W_k, B_k) + \varepsilon_k, \\ W_0 \sim P_{ini}. \end{cases}$$

Une telle perturbation de l'algorithme a des motivations pratiques (éviter les points-selles, confidentialité différentielle [ACG+16]) et mathématiques, comme nous le verrons plus loin. L'algorithme de descente de gradient considéré est le suivant : pour $k \geq 0$ et $i \in \{1, \dots, N\}$, (pour

simplifier la présentation, on considère, dans l'introduction, un taille de batch fixe B)

$$W_{k+1} = W_k - \frac{\alpha}{B} \sum_{(x,y)\in B_k} \nabla_W \mathsf{L}(g^N_{W_k}(x), y)) + \frac{\varepsilon_k}{N^\beta}, \tag{1.2.11}$$

où

- $\alpha > 0$ est le taux d'apprentissage (*learning rate*), fixe au cours de l'entraînement.
- $B_k \subset \mathcal{X} \times \mathcal{Y}$ est le batch de données sur lequel le réseau s'entraîne à l'itération k. Son cardinal est noté B.
- $\varepsilon_k \sim \mathcal{N}(0, I_{Nd})$ est un bruit ajouté artificiellement.

L'analyse de la dynamique (1.2.11) se fera par l'intermédiaire des mesures empiriques

$$\nu^N_k = \frac{1}{N} \sum_{i=1}^N \delta_{W^i_k}$$

et de sa version continue (avec le changement d'échelle dont nous avons discuté après (1.2.7))

$$\mu^N_t = \nu^N_{\lfloor Nt \rfloor} = \frac{1}{N} \sum_{i=1}^N \delta_{W^i_{\lfloor Nt \rfloor}}, \ t \geq 0.$$

Pour tout $N \geq 1$, la trajectoire

$$t \in \mathbf{R}_+ \mapsto \mu^N_t \tag{1.2.12}$$

est un élément aléatoire de l'espace des fonctions *càdlàg* $\mathcal{D}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d))$, appelé espace de Skorohod (une introduction à ces espaces est donnée dans l'appendice A.1).
Notre objectif sera de montrer que la suite de processus (1.2.12) **converge lorsque** $N \to \infty$ **et de caractériser cette limite.** Pour ce faire, nous ferons les hypothèses **A1.** à **A4.** suivantes. Introduisons $\mathcal{F}^N_k$ la tribu engendrée par les variables aléatoires permettant de connaître la valeur de W_k ; autrement dit, la plus petite tribu telle que W_k soit $\mathcal{F}^N_k$-mesurable. On a donc

$$\mathcal{F}^N_k = \boldsymbol{\sigma}\{W_0, B_j, \varepsilon_j, j = 0, \ldots, k-1\}.$$

A1. *Pour tout $k \in \mathbf{N}$, les données d'entraînement de l'itération k sont indépendantes des données d'entraînement des itérations précédentes :*

$$(B_k, \varepsilon_k) \perp\!\!\!\perp \mathcal{F}^N_k.$$

Par ailleurs B_k et le bruit additionnel ε_k sont indépendants.

Cette hypothèse traduit mathématiquement le fait que les données d'entraînement font partie de l'environnement : elles n'ont pas de lien avec le modèle considéré (le réseau de neurones). Mais elle n'implique pas nécessairement un jeu de données d'entraînement infini. En effet, si $\{x_i, y_i)\}_{i=1}^M$ est un jeu de données fini, on peut considérer B_k comme un sous ensemble aléatoire suivant la loi $\pi = \frac{1}{M} \sum_{i=1}^M \delta_{(x_i,y_i)}$.

A2. *La fonction d'activation est lisse et bornée (ainsi que toutes ses dérivées) :*

$$\sigma_* \in \mathcal{C}^\infty_b(\mathbf{R}^d \times \mathcal{X})$$

Cette hypothèse permet de considérer, entre autres, les fonctions sigmoïde ou tangente hyperbolique, mais pas les fonctions de type ReLU, car non bornées. On pourrait néanmoins adapter nos résultats à une fonction non dérivable en un point, en considérant la sous-différentielle.

A3. *Toutes les données d'entraînement (x, y) sont i.i.d. suivant la loi π, et y a des moments d'ordre suffisamment élevé finis.*

Cette hypothèse est cohérente avec le fait qu'on ne considère pas ici de réseau de neurones récurrents.

A4. *Les poids initiaux W_0 sont initialisés de manière i.i.d. suivant une loi μ_0 qui possède des moments d'ordre suffisamment élevé finis.*

En pratique, on considère souvent des initialisations gaussiennes.

Nous pouvons désormais énoncer notre résultat.

Théorème 1.10. *Supposons $\beta > 1/2$ et* **A1.-A4.** *La suite $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ converge en probabilité vers $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$, caractérisé comme l'unique élément de $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ solution de*

$$\begin{cases} \partial_t \bar{\mu}_t = -\alpha \mathrm{div}\Big(\bar{\mu}_t \nabla\Big(\frac{\delta \mathscr{R}}{\delta \bar{\mu}_t}\Big)\Big), \\ \bar{\mu}_t = \mu_0. \end{cases} \tag{1.2.13}$$

Une illustration de ce théorème est donnée en figure 1.1.

Commentaire sur l'équation (1.2.13). Il s'agit d'une équation de transport sur l'espace des mesures de probabilités, avec un champ de vitesse non local, au sens où le champ de vitesse $\nabla\Big(\frac{\delta \mathscr{R}}{\delta \bar{\mu}_t}\Big) : \mathbf{R}^d \to \mathbf{R}^d$ dépend globalement de la mesure $\bar{\mu}_t$. Dans le cas d'un champ de vitesse local, l'équation correspondante, c'est-à-dire

$$\partial_t \mu_t = -\mathrm{div}(\rho_t v_t),$$

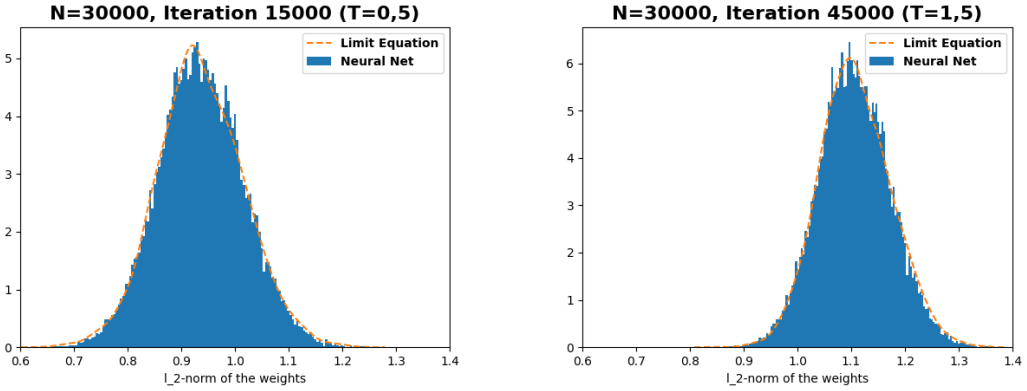
avec $v_t : \mathbf{R}^d \to \mathbf{R}^d$, est bien connue et a été étudiée par de nombreux auteurs (voir par exemple [San15, Section 4.1.2]). L'équation non locale, quant à elle, est plus délicate à étudier, mais a suscité beaucoup d'intérêt ces dernières années ; elle est notamment utilisée pour modéliser des déplacements de foules, où μ_t représente la densité de piétons à l'instant t, [CPT11]. D'un point de vue mathématique, des résultats d'existence sont donnés dans [AG08] sous des conditions de croissance et de continuité du champ de vitesse. Notons que notre limite champ moyen prouve l'existence d'une solution. Pour prouver l'unicité de cette solution, nous utilisons des arguments développés dans [PR13]. Nous donnons également une preuve probabiliste faisant appel aux résultats de [Szn91]. Citons également [PRT15] où les auteurs établissent l'existence d'un contrôle sur l'équation de Cucker-Smale cinétique.

Lorsque le bruit additionnel de (1.2.11) est suffisamment intense, l'équation limite est perturbée, ce qui fait apparaître un terme laplacien.

Théorème 1.11. *Supposons $\beta = 1/2$ et* **A1.-A4.** *La suite $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ converge en probabilité vers $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$, caractérisé comme l'unique élément de $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ solution de*

$$\begin{cases} \partial_t \bar{\mu}_t = -\alpha \mathrm{div}\Big(\bar{\mu}_t \nabla\Big(\frac{\delta \mathscr{R}}{\delta \bar{\mu}_t}\Big) + \nabla \bar{\mu}_t\Big), \\ \bar{\mu}_t = \mu_0. \end{cases} \tag{1.2.14}$$

FIGURE 1.1 : Évolution de la distribution empirique des poids d'un réseau de neurones et distribution prédite par (1.2.13) à différents temps d'apprentissage.



Cette équation correspond à la minimisation de la fonction de coût

$$\widetilde{\mathscr{R}} : \mu \mapsto \mathscr{R}(\mu) - \mathrm{Ent}(\mu),$$

où

$$\mathrm{Ent}(\mu) = \begin{cases} -\int \mu(x)\log(\mu(x))\mathrm{d}x \text{ si } \mu \ll \mathrm{Leb}, \\ -\infty \text{ sinon.} \end{cases}$$

L'équation (1.2.14) est une version régularisée de (1.2.13), grâce à l'ajout d'un terme diffusif, qui rend l'étude de l'unicité de l'équation plus aisée (voir notamment [JMM20, HRSS21]). Notons que [MMN18] ont obtenu le résultat de convergence du théorème 1.11 à temps fixé. Pour une revue des différentes régularisations de (1.2.13), nous renvoyons le lecteur à [FRF22].

La convergence $\mu_t^N \xrightarrow{\mathscr{L}} \bar{\mu}_t$ obtenue dans ces théorèmes permet, par une application directe du théorème de Tanaka-Sznitman ([Got98, Théroème 3.2]), d'obtenir un résultat qualitatif de propagation du chaos, au sens de la définition 1.2.

Corollaire 1.12. *Introduisons, pour $t \geq 0$, la loi jointe*

$$\rho_t^N = \mathrm{Law}(W^i_{\lfloor Nt \rfloor}, i = 1, \ldots, N).$$

Alors, pour tout $t \geq 0$, la suite $(\rho_t^N)_{N\geq 1}$ est $\bar{\mu}_t$-chaotique, au sens de la définition 1.1.

Pour une version quantitative de ce résultat, on pourra se référer à [DBDFS20].

Schéma de preuve des théorèmes 1.10 et 1.11. Nous donnons ci-dessous les étapes des preuves des théorèmes 1.10 et 1.11. La démonstration complète du théorème 1.10 est donnée dans le chapitre 2. Celle du théorème 1.11, similaire, est omise.

(i) Trouver une équation « pré-limite », c'est-à-dire une équation satisfaite par μ_t^N comme nous l'avons montré pour la dynamique (1.2.8) avec l'équation (1.2.10). A noter qu'étudiant une dynamique discrète et avec un coût empirique on n'obtiendra pas exactement l'équation limite, comme ce fut le cas pour l'équation (1.2.10) : on obtiendra aussi des termes additionnels. Cette équation pré-limite s'obtient par un développement limité à l'ordre 1 pour le théorème 1.10, à l'ordre 2 pour le théorème 1.11. Plus précisément, on évalue les différences $\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle$ par un développement limité, faisant donc intervenir la différence $W_{k+1} - W_k$ et un terme de reste. Le terme $W_{k+1} - W_k$ se réécrit grâce à (1.2.11). En sommant ensuite sur $k = 0, \ldots, \lfloor Nt \rfloor$, on obtient une équation intégro-différentielle sur $t \mapsto \langle f, \mu_t^N\rangle$.

(ii) Montrer que les termes additionnels tendent vers 0 lorsque $N \to \infty$.

(iii) Montrer que les moments des poids du réseau sont bornés, uniformément sur un nombre d'itérations du même ordre de grandeur que N. On utilise pour cela nos hypothèses de moments bornés sur les données d'entraînement **A3.** et sur les poids initiaux **A4.**

(iv) Montrer que la suite des lois de $(\mu^N)_{N\geq 1}$ est relativement compacte. En vertu du théorème de Prohorov, ceci est équivalent au fait de montrer que ces lois sont étroites.

Définition 1.13. *Soit $\mathscr{P} \subset \mathcal{P}(S)$ un ensemble de mesure de probabilités définies sur un espace métrique S. On dit que la famille $\mathscr{P}$ est étroite si pour tout $\varepsilon > 0$, il existe un compact $K \subset S$ tel que*

$$\inf_{P\in\mathscr{P}} P(K) \geq 1 - \varepsilon.$$

Ici, $S = \mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. On utilise alors le critère de Jakubowski [Jak86, Théorème 4.6] qui permet de prouver l'étroitesse des lois de $(\mu^N)_{N\geq 1}$ à partir de l'étroitesse des lois de $(\langle f, \mu^N\rangle)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathbf{R})$, où f appartient à une famille de fonctions qui sépare les mesures de probabilité. On utilise crucialement le résultat du point *(iii)*.

(v) Montrer que tout point limite de $(\mu^N)_{N\geq 1}$ satisfait presque sûrement l'équation limite (1.2.13) ou (1.2.14). C'est là un point délicat de la démonstration, qui repose sur l'application du *continuous mapping theorem* à une famille de fonctionnelles. Or ces fonctionnelles ne sont pas toutes continues. Nous appliquons donc ce théorème pour les fonctionnelles continues, et utilisons ensuite des arguments de séparabilité, couplés à d'autres arguments de continuité, pour pouvoir conclure que tout point limite satisfait presque sûrement (1.2.13) ou (1.2.14) au sens faible.

(vi) Montrer que l'équation limite admet une solution dans $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. Pour cela, on doit déjà montrer que tout point limite de $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ est presque sûrement continu. On utilise pour cela le critère de Jacod et Shiryaev [JS87, Chapitre VI, Proposition 3.26] où l'on montre que la probabilité que les sauts du processus soient plus grands que ϵ tend vers 0 lorsque $N \to \infty$, et ce quel que soit $\epsilon > 0$.

(vii) Montrer l'unicité d'une solution de l'équation limite. Pour (1.2.13), on utilise des méthodes développées par [PRT15, PR16]. La preuve de l'unicité de (1.2.14) est faite dans [JMM20, Théorème C.4].

Dans la suite nous allons nous intéresser aux fluctuations de μ^N autour de sa limite $\bar{\mu}$, solution de (1.2.13).

1.2.2 Fluctuations autour de la limite champ moyen : théorème central limite

Dans cette section, nous introduisons les notions mathématiques propres à l'obtention d'un théorème central limite, et présentons nos résultats.

Pour obtenir un théorème central limite sur $(\mu^N)_{N\geq 1}$, on introduit le processus de fluctuation

$$t \mapsto \eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t).$$

Pour tout $t \geq 0$, η_t^N est donc une mesure signée. Il n'est pas aisé de travailler dans cet espace, puisque l'ensemble des mesures signées muni de la topologie de la convergence faible n'est en général pas métrisable. Dès lors, deux approches sont envisageables :

(i) Étudier le processus à valeurs réelles $\langle f, \eta_t^N\rangle$.

(ii) Considérer η_t^N comme un élément du dual d'un espace de Sobolev.

État de l'art. Une dérivation informelle du théorème central limite est donnée, à temps fixé et pour l'algorithme (1.2.8), dans les articles [CRBVE20] et [RVE22] avec des considérations sur les différentes échelles et le temps long. Concernant l'approche *(i)*, [JT21] considère - dans le cadre de l'approximation champ moyen de processus de McKean-Vlasov - une généralisation de cette approche, en considérant le processus $t \mapsto \sqrt{N}(\Phi(\mu_t^N) - \Phi(\bar{\mu}_t)) \in \mathcal{C}(\mathbf{R}_+, \mathbf{R})$, où $\Phi : \mathcal{P}(\mathbf{R}^d) \to \mathbf{R}$ est une fonctionnelle non linéaire.

L'approche *(ii)* a été utilisée dans plusieurs travaux pour l'étude d'équations de type McKean-Vlasov [FM97b, GKM+96, JM98a] ainsi qu'en théorie des jeux à champ moyen [DLR19b]. C'est cette approche que nous allons suivre, en s'inspirant d'idées développées par [SS20a]. **Notre contribution consiste à mettre en œuvre rigoureusement les idées de ce dernier article, en les adaptant au cas d'un mini-batch et d'un bruit (voir Théorème 1.17), ainsi que de mettre en évidence un régime limite lorsque le bruit additionnel est trop intense (voir Théorème 1.18).**

Résultats. Comme nous l'avons dit, nous considérons ici η_t^N comme un élément du dual d'un espace de Sobolev, plus précisément comme un élément de l'espace $\mathcal{H}^{-J}(\mathbf{R}^d)$, dual de l'espace $\mathcal{H}_0^J(\mathbf{R}^d) = W_0^{J,2}(\mathbf{R}^d)$. Justifions que cela est licite. Comme η_t^N fait intervenir des mesures de Dirac (*via* μ_t^N), les fonctions tests associées à η_t^N doivent être définies en chaque point de $\mathbf{R}^d$; or, dès que $J > d/2$, on a l'injection de Sobolev

$$\mathcal{H}_0^J(\mathbf{R}^d) \hookrightarrow \mathcal{C}_b(\mathbf{R}^d),$$

ce qui rend $\langle f, \eta_t^N \rangle_{\mathcal{H}_0^J(\mathbf{R}^d), \mathcal{H}^{-J}(\mathbf{R}^d)}$ bien défini. Nous noterons plus simplement $\langle f, \eta_t^N \rangle$ ce crochet de dualité par la suite.

Introduisons maintenant l'équation différentielle stochastique généralisée qui caractérisera la limite en loi du processus de fluctuations $(\eta^N)_{N \geq 1}$:

$$\langle f, \bar{\eta}_t \rangle - \langle f, \bar{\eta}_0 \rangle = \int_0^t \langle \mathscr{L}^{\bar{\mu}_s} f, \bar{\eta}_s \rangle \mathrm{d}s + \langle f, \mathscr{G}_t \rangle, \ \ f \in \mathcal{C}_c^\infty(\mathbf{R}^d), \tag{1.2.15}$$

où $\mathscr{L}^{\bar{\mu}_s}$ est un opérateur différentiel dépendant de la limite champ moyen obtenue par le théorème 1.10, et $\mathscr{G} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}^d))$ est un processus gaussien en dimension infinie, appelé G-process, et défini ainsi

Définition 1.14 (G-process)**.** *On dit que $\mathscr{G} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}^d))$ est un G-process si pour tout $k \geq 1$, et $f_1, \ldots, f_k \in \mathcal{H}_0^J(\mathbf{R}^d)$, $\{t \mapsto (\langle f_1, \mathscr{G}_t \rangle, \ldots, \langle f_k, \mathscr{G}_t \rangle)^T, t \in \mathbf{R}_+\} \in \mathcal{C}(\mathbf{R}_+, \mathbf{R}^k)$ est un processus à moyenne nulle, à incréments indépendants et de structure de covariance donnée par : pour tous $1 \leq i, j \leq k$ et $0 \leq s \leq t$,*

$$\mathrm{Cov}(\langle f_i, \mathscr{G}_t \rangle, \langle f_j, \mathscr{G}_s \rangle) = \frac{\alpha^2}{B} \int_0^s \mathrm{Cov}_{(x,y) \sim \pi}(\mathrm{Q}_v[f_i](x,y), \mathrm{Q}_v[f_j](x,y)) \mathrm{d}v, \tag{1.2.16}$$

où $\mathrm{Q}_v[f](x,y) = (y - \langle \sigma_*(\cdot, x), \bar{\mu}_v \rangle) \langle \nabla f \cdot \nabla_W \sigma_*(\cdot, x), \bar{\mu}_v \rangle$, où $\bar{\mu}$ est donné par le théorème 1.10.

Nous pouvons maintenant définir les solutions faibles de (1.2.15).

Définition 1.15 (Solution faible)**.** *On dit que $\bar{\eta} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}^d))$ est solution faible de* (1.2.15) *avec donnée initiale $\nu \in \mathcal{H}^{-J}(\mathbf{R}^d)$ s'il existe un* G-process *tel que* (1.2.15) *soit vérifiée presque sûrement et $\eta_0 = \nu$ en loi.*

Ceci conduit à la notion d'unicité faible.

Définition 1.16 (Unicité faible)**.** *On dit qu'il y a unicité faible pour* (1.2.15) *si, étant données deux solutions faibles η^1 et η^2 avec même donnée initiale, possiblement définies sur des espaces probabilisés différents, on a $\eta^1 = \eta^2$ en loi.*

Avant d'énoncer notre résultat, nous devrons faire une hypothèse supplémentaire sur la distribution des poids initiaux.

A5. *Le support de μ_0 est compact.*

Notre théorème central limite est le suivant.

Théorème 1.17. *Supposons $\beta > 3/4$ et* **A1.-A5.** *La suite $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}^d))$ converge en loi vers l'unique solution faible de* (1.2.15) *avec donnée initiale $\nu_0 \in \mathcal{H}^{-J}(\mathbf{R}^d)$ définie par*

$$(\langle f_1, \nu_0\rangle, \dots, \langle f_k, \nu_0\rangle)^T \sim \mathcal{N}(0, \Gamma(f_1, \dots, f_k)),\ f_1, \dots, f_k \in \mathcal{H}_0^J(\mathbf{R}^d),\ k \geq 1,$$

où $\Gamma(f_1, \dots, f_k)$ désigne la matrice de covariance de $(f_1(W_0^1), \dots, f_k(W_0^1))^T$.

Commentaires sur le théorème 1.17 :

- Le théorème 1.17 est valable pour $\beta > 3/4$; la loi des grands nombres du théorème 1.10 est donc vérifiée. Pour que ce théorème central limite soit valable, l'intensité du bruit additionnel doit donc être plus faible.

- le théorème central limite signifie que

$$\mu_t^N \approx \bar{\mu}_t + \frac{1}{\sqrt{N}}\bar{\eta}_t, \tag{1.2.17}$$

  où $\bar{\eta}_t$ satisfait (1.2.15). Un intérêt de cette formule est que si l'on sait échantillonner efficacement la fluctuation $\bar{\eta}_t$ (disons $(\bar{\eta}_t(\omega_i))_{i=1}^m$), alors on peut, à partir d'un estimateur $\mu_t^N(\omega)$ de $\bar{\mu}_t$, obtenir plusieurs estimateurs de la limite champ moyen $\bar{\mu}_t$:

$$\mu_t^N(\omega) + \frac{1}{\sqrt{N}}\bar{\eta}_t(\omega_i),\ i = 1, \dots, m.$$

  Ceci pourrait constituer une méthode pour simuler des réseaux de neurones, et obtenir ainsi des *deep ensembles* [LPB17], sans avoir à entraîner les m réseaux de neurones. Notons toutefois que la simulation de solutions d'équations de type (1.2.15) n'est pas chose aisée, et pas nécessairement plus rapide en temps de calcul que d'entraîner un réseau de neurones (d'autant que les cas d'intérêt sont en grande dimension d).

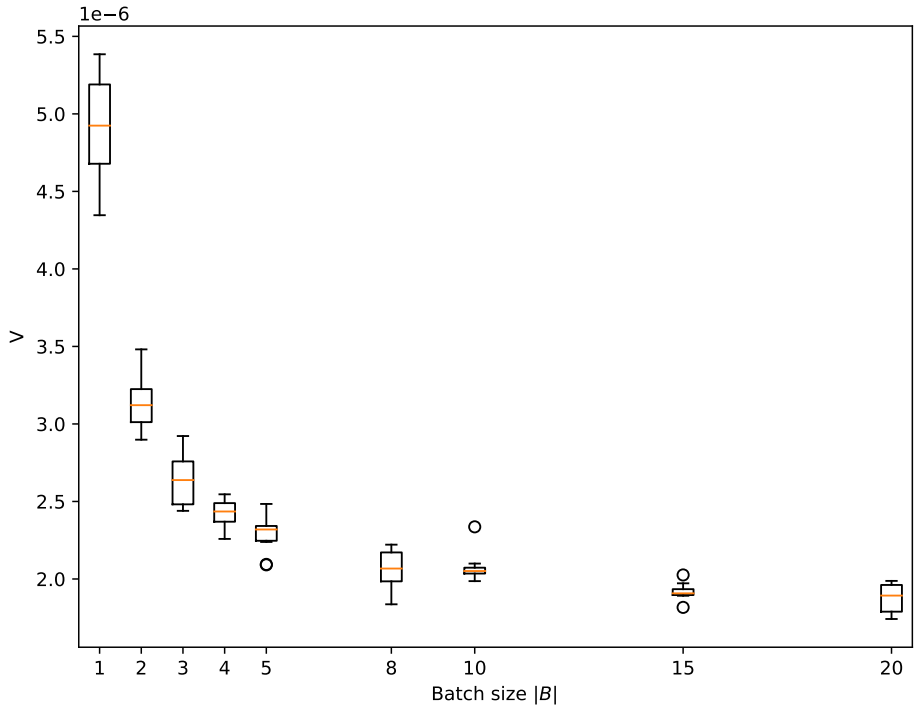
- Pour quantifier la déviation de μ_t^N autour de sa limite champ moyen, la quantité d'intérêt est

$$\mathrm{Var}(\mu_t^N) \approx \frac{1}{N}\mathrm{Var}(\bar{\eta}_t),$$

  l'approximation ayant été obtenue grâce à (1.2.17). L'équation (1.2.16) nous donne l'expression exacte de la variance du G-process, qui montre que celle-ci décroît de façon inversement proportionnelle à la taille du mini-batch B. Cependant, notre équation (1.2.15) ne permet pas de relier directement la variance du G-process à la variance de $\bar{\eta}_t$, puisqu'*a priori* $\mathscr{G}_t$ n'est pas indépendant de $\{\bar{\eta}_s, s \leq t\}$. Notons toutefois qu'une décroissance de la variance en fonction de la taille du mini-batch est attendue, étant donné que l'estimation du risque populationnel par le risque empirique est plus exacte avec un grand mini-batch. Numériquement, nous pouvons constater une décroissance de la variance de $\bar{\eta}_t$ avec la taille du mini-batch, voir figure 1.2.

On n'a *a priori* pas de théorème central limite pour $\beta \leq 3/4$. En effet, dans ce cas, le terme d'erreur $\mathbf{e}_t^N[f]$ diverge avec N (voir (1.2.18) et le point *(v)* dans le schéma de preuve ci-dessous). Les simulations numériques de la figure 1.3 suggèrent toutefois l'existence d'un régime critique

FIGURE 1.2 : Décroissance de la variance de $\langle f, \bar{\eta}_t \rangle$ en fonction de la taille du mini-batch.



dans le cas $\beta = 3/4$. Nous avons pu montrer mathématiquement, dans un cas particulier, que dans ce régime critique, l'équation (1.2.15) est perturbée par un terme additionnel linéaire en temps. Plus précisément, nous montrons, en dimension 1 et avec une fonction test particulière, le résultat suivant.

Théorème 1.18. *Supposons $\beta = 3/4$ et* **A1.-A5.** *La suite $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}))$ est relativement compacte et tout point limite suite la loi de η^*, où η^* satisfait l'équation*

$$\langle f, \eta_t^* \rangle - \langle f, \eta_0^* \rangle = \int_0^t \langle \mathscr{L}^{\bar{\mu}_s} f, \eta_s^* \rangle \mathrm{d}s + \langle f, \mathscr{G}_t \rangle + t\mathbf{E}[f(\varepsilon_1^1)], \;\; f : x \in \mathbf{R} \mapsto x^2.$$

Schéma de preuve du théorème 1.17. Les étapes de la preuve sont les suivantes.

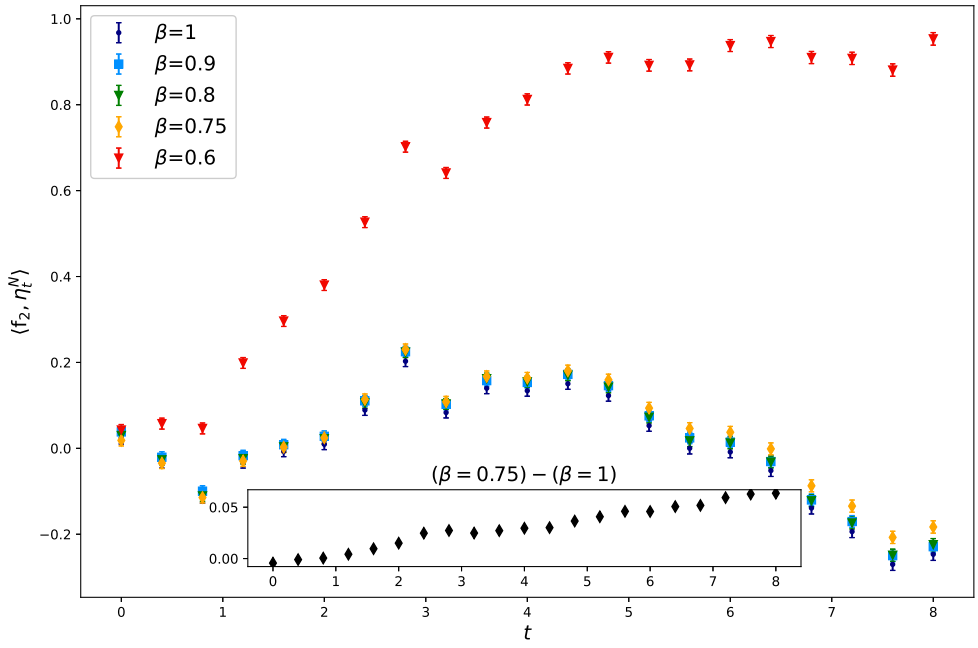
(i) Obtenir une équation « pré-limite » pour le processus de fluctuation en utilisant l'équation « pré-limite » obtenue sur μ_t^N dans la preuve du théorème 1.10 et l'équation (1.2.13) :

$$\langle f, \eta_t^N \rangle - \langle f, \eta_0^N \rangle - \int_0^t \langle \mathscr{L}^{\bar{\mu}_s} f, \eta_s^N \rangle \mathrm{d}s - \langle f, \sqrt{N} M_t^N \rangle = -\mathbf{e}_t^N[f], \tag{1.2.18}$$

où $t \mapsto \langle f, \sqrt{N} M_t^N \rangle$ est une martingale et $\mathbf{e}_t^N[f]$ un terme d'erreur.

(ii) Montrer la relative compacité de $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}^d))$. Pour cela, on travaille dans des espaces de Sobolev à poids de différents ordres, pour utiliser des injections de Hilbert-Schmidt et nos bornes sur les moments. Nous avons pour cela modifié un critère de relative compacité de Kurtz [Kur75, Théorème 4.20] pour pouvoir utiliser nos estimées de moments des poids du réseau de neurones, obtenues au point *(iii)* de la preuve du théorème 1.10.

FIGURE 1.3 : Évolution temporelle du processus de fluctuation pour différentes intensités β du bruit additionnel. On trace aussi la différence entre les courbes $\beta = 0,75$ et $\beta = 1$.



Proposition 1.19. *Soit $\mathcal{H}_1$ et $\mathcal{H}_2$ deux espaces de de Hilbert tels que $\mathcal{H}_1 \hookrightarrow_{\text{H.S.}} \mathcal{H}_2$. Si $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}_2^{-1})$ satisfait les conditions :*

(i) Pour tous $T > 0$ et $\epsilon > 0$, il existe $C > 0$ tel que

$$\sup_{N\geq 1} \mathbf{P}\Big(\sup_{t\in[0,T]} \|\eta_t^N\|_{\mathcal{H}_2^{-1}} > C \Big) \leq \epsilon.$$

(ii) Pour tous $\delta > 0$, $N \geq 1$, $0 \leq t, s \leq T$ avec $|t - s| \leq \delta$, il existe $F_N(\delta) < +\infty$ tel que

$$\mathbf{E}\Big[\|\eta_t^N - \eta_s^N\|^2_{\mathcal{H}_2^{-1}} \Big] \leq F_N(\delta)$$

avec $\lim_{\delta\to 0} \limsup_{N\geq 1} F_N(\delta) = 0$.

Alors la suite $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}_1^{-1})$ est relativement compacte.

(iii) Montrer que $(\sqrt{N}M^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}^d))$ converge en loi vers un G-process. Pour ce faire, on montre d'abord que la suite $(\sqrt{N}M^N)_{N\geq 1}$ est relativement compacte, en utilisant les techniques du point précédent. Il s'agit ensuite de montrer que la suite $(\sqrt{N}M^N)_{N\geq 1}$ admet une unique valeur d'adhérence. Pour montrer cela, on s'intéresse au processus testé contre des fonctions. Plus précisément, on utilise un théorème central limite pour les martingales (voir [EK09, Théorème 7.1.4]) pour montrer que pour tous $k \geq 1$ et $f_1, \ldots, f_k \in \mathcal{H}_0^J(\mathbf{R}^d)$, la suite

$$\{t \mapsto (\langle f_1, \sqrt{N}M_t^N\rangle, \ldots, \langle f_k, \sqrt{N}M_t^N\rangle), t \in \mathbf{R}_+\}_{N\geq 1}$$

converge en loi vers un processus satisfaisant les conditions de la définition 1.14. Pour conclure, on montre par des considérations de théorie de la mesure que la loi de tout point

limite $\mathscr{G}^*$ de $(\sqrt{N}M^N)_{N\geq 1}$ est entièrement caractérisée par les lois de

$$(\langle f_1, \mathscr{G}^*\rangle, \dots, \langle f_k, \mathscr{G}^*\rangle) \in \mathcal{D}(\mathbf{R}_+, \mathbf{R}^k),$$

où $k \geq 1$ et $(f_k)_{k\geq 1}$ est une base orthonormée de $\mathcal{H}_0^J(\mathbf{R}^d)$. Par unicité du point d'adhérence, toute la suite $(\sqrt{N}M^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}^d))$ converge en loi vers un G-process.

(iv) Montrer l'unicité trajectorielle pour l'équation (1.2.15), c'est-à-dire que si η^1 et η^2 sont deux solutions fortes de (1.2.15) avec le même G-process, alors $\eta^1 = \eta^2$ presque sûrement. On utilise pour cela le lemme de Gronwall.

(v) Montrer que si $(\eta^*, \mathscr{G}^*)$ est une valeur d'adhérence de $(\eta^N, \sqrt{N}M^N)_{N\geq 1}$ (ce couple est relativement compact par les points *(ii)-(iii)*), alors η^* est solution faible de (1.2.15) au sens de la définition 1.15. Pour ce faire, on montre que pour tous t et f fixés :

- la fonctionnelle qui à $(\eta_t^N, \sqrt{N}M_t^N)$ associe le membre de gauche de (1.2.18) est continue.
- le terme d'erreur tend vers 0 : $\mathbf{E}[|\mathbf{e}_t^N[f]|] \to 0$ lorsque $N \to \infty$.

On a ainsi que pour tous t et f, presque sûrement, $(\eta^*, \mathscr{G}^*)$ satisfait l'équation limite. En utilisant un argument de continuité et de séparabilité, on peut conclure que le couple $(\eta^*, \mathscr{G}^*)$ satisfait presque sûrement (1.2.15).

(vi) En considérant deux points limites η^1 et η^2 de $(\eta^N)_{N\geq 1}$, on a, en considérant des sous-suites et en appliquant le point *(v)*, que η^1 et η^2 sont deux solutions faibles de (1.2.15). On utilise ensuite le fait que l'unicité trajectorielle *(iv)* implique l'unicité faible[5] pour conclure que $\eta^1 = \eta^2$ en loi. Ceci conclut la preuve du théorème 1.17.

Dans la section suivante, nous allons nous intéresser au paradigme bayésien, qui offre notamment la possibilité de quantifier l'incertitude des prédictions des réseaux de neurones, ce qui est un sujet d'une importance majeure au regard des innovations technologiques que les réseaux de neurones sont à même de produire.

1.3 Approche champ moyen pour l'inférence variationnelle dans le cadre des réseaux de neurones bayésiens

Cette section est consacrée à la présentation de nos résultats relatifs aux réseaux de neurones bayésiens. Nous commencerons par présenter en section 1.3.1 le formalisme bayésien propre aux réseaux de neurones, et plus particulièrement l'inférence variationnelle. Nous présenterons ensuite en section 1.3.2 notre premier résultat de limite champ moyen, qui a fait l'objet d'une publication dans COLT 2023 [DHG+23]. Enfin, la section 1.3.3 présente un travail en préparation sur les fluctuations autour de la limite champ moyen.

1.3.1 Formalisme bayésien pour les réseaux de neurones

En statistiques bayésiennes, la quantité à approcher - ici, les poids d'un réseau de neurones - est modélisée par une variable aléatoire, appelée variable *latente* et notée w. La loi jointe de w et des données x doit être posée par le statisticien, sous forme d'une loi *a priori* $p(w)$ sur la variable latente et d'une loi conditionnelle des données sachant la variable latente, appelée vraisemblance et notée $p(x|w)$:

$$p(w, x) = p(x|w)p(w).$$

[5] Ce résultat remonte à [YW71] et été utilisé pour un problème similaire au nôtre dans [DLR19b] dans le cadre des jeux à champ moyen.

La loi de la variable latente sachant les données, appelée loi *a posteriori*, s'exprime alors grâce à la formule de Bayes, où $p(x)$ est la loi des données :

$$p(w|x) = \frac{p(x|w)p(w)}{p(x)}.$$

C'est cette loi qui contient toute l'information d'intérêt, comme nous allons le voir plus précisément dans le cadre des réseaux de neurones.

Les réseaux de neurones bayésiens. Une bonne introduction générale aux réseaux de neurones bayésiens et aux questions qui s'y rattachent est donnée dans [JLB+22]. On notera

$$f_w : \mathsf{X} \to \mathbf{R}$$

un réseau de neurones bayésien de poids $w \in \mathscr{W}$, où $\mathsf{X} \times \mathsf{Y}$ est l'ensemble de définition des données d'entraînement. Dans cette écriture w est la variable latente qui suit la distribution *a poteriori* $p(w|D)$, où $D \subset \mathsf{X} \times \mathsf{Y}$ est le jeu de données d'entraînement.
Toute l'information sur les prédictions du réseau (par exemple des informations sur l'incertitude liée à ces prédictions) est alors contenue dans la loi marginale (appelée aussi loi prédictive *a posteriori*) $p(y|x, D)$, qui peut s'exprimer à partir de la loi *a posteriori* et de la vraisemblance $\mathrm{L}(y|x, w')$:

$$p(y|x, D) = \int_{\mathscr{W}} \mathrm{L}(y|x, w')p(w'|D)\mathrm{d}w'. \tag{1.3.1}$$

Comme la vraisemblance fait partie du modèle statistique, on voit que toute l'information s'obtient grâce à la loi *a posteriori*.
Malheureusement, l'espace $\mathscr{W}$ relatif à la variable latente étant typiquement de grande dimension, cette intégrale n'est pas calculable ; la formule (1.3.1) n'est donc pas directement exploitable en pratique. Concrètement, s'il s'agit par exemple de prédire un label $\hat{y}$ à une donnée d'entrée x, on cherchera à échantillonner des poids $(\hat{w}_i)_{i=1,\dots,m}$ de façon i.i.d. suivant la loi *a posteriori* $p(w|D)$, et à utiliser une moyenne empirique :

$$\hat{y} = \frac{1}{m}\sum_{i=1}^{m} f_{\hat{w}_i}(x).$$

On pourra alors calculer une variance empirique pour obtenir une information sur l'incertitude prédictive relative à la donnée x.

Utiliser un réseau de neurones bayésien requiert donc de savoir échantillonner efficacement suivant la loi *a posteriori*. Plus généralement, échantilloner des lois *a posteriori* est un enjeu majeur des statistiques bayésiennes. A cette fin, les algorithmes les plus étudiés et utilisés sont les méthodes de Monte-Carlo par chaînes de Markov [RCC99]. Cependant, ces méthodes peuvent être difficiles à mettre en place lorsque les jeux de données sont grands et les modèles complexes. Dans ce cas, l'inférence variationnelle, avec laquelle on n'échantillonne pas suivant la loi *a posteriori*, constitue une alternative prometteuse.

Inférence variationnelle et *Evidence Lower Bound*. L'idée de l'inférence variationnelle est de transformer un problème d'échantillonnage en un problème d'optimisation [BKM17]. Étant donné un ensemble de densités de probabilités $\mathscr{Q}$, faciles à échantillonner, on cherche celle qui minimise la divergence de Kullback-Leibler par rapport à la distribution *a posteriori* :

$$q^* = \operatorname*{arg\,min}_{q \in \mathscr{Q}} \mathscr{D}_{\mathrm{KL}}(q|p(w|D)), \tag{1.3.2}$$

où la divergence de Kullback-Leibler de P par rapport à Q (où P et Q sont absolument continues par rapport à la mesure de Lebesgue, de densités respectives p et q), est définie par

$$\mathscr{D}_{\mathrm{KL}}(P|Q) = \int \log\Big(\frac{p(x)}{q(x)}\Big)p(x)\mathrm{d}x.$$

Minimiser cette divergence de Kullback-Leibler revient à maximiser une autre quantité, appelée *Evidence Lower Bound*, que nous introduisons maintenant. Partant de (1.3.2), nous avons, pour $q \in \mathscr{Q}$,

$$\begin{aligned}
\mathscr{D}_{\mathrm{KL}}(q|p(w|D)) &= \int_{\mathscr{W}} \log\Big(\frac{q(w)}{p(w|D)}\Big)q(w)\mathrm{d}w \\
&= \int_{\mathscr{W}} \log(q(w))q(w)\mathrm{d}w - \int_{\mathscr{W}} \log(p(w|D))q(w)\mathrm{d}w \\
&= \int_{\mathscr{W}} \log(q(w))q(w)\mathrm{d}w - \int_{\mathscr{W}} \log\Big(\frac{p(w,D)}{p(D)}\Big)q(w)\mathrm{d}w \\
&= \int_{\mathscr{W}} \log(q(w))q(w)\mathrm{d}w - \int_{\mathscr{W}} \log(p(w,D))q(w)\mathrm{d}w + \log(p(D)).
\end{aligned}$$

En utilisant que la loi jointe $p(w, D)$ s'exprime en fonction de la vraisemblance et de la loi *a priori* P,

$$p(w, D) = \mathrm{L}(y|x, w)P(w),$$

il vient

$$\begin{aligned}
\mathscr{D}_{\mathrm{KL}}(q|p(w|D)) &= \int_{\mathscr{W}} \log(q(w))q(w)\mathrm{d}w - \int_{\mathscr{W}} \log(\mathrm{L}(y|x,w))q(w)\mathrm{d}w - \int_{\mathscr{W}} \log(P(w))q(w)\mathrm{d}w + \log(p(D)) \\
&= -\int_{\mathscr{W}} \log(\mathrm{L}(y|x,w))q(w)\mathrm{d}w + \mathscr{D}_{\mathrm{KL}}(q|P) + \log(p(D)). \qquad (1.3.3)
\end{aligned}$$

Le dernier terme ne faisant pas intervenir la distribution q, minimiser $\mathscr{D}_{\mathrm{KL}}(q|p(w|D))$ revient donc à maximiser

$$\mathrm{E}_{\mathrm{lbo}} := \int_{\mathscr{W}} \log(\mathrm{L}(y|x,w))q(w)\mathrm{d}w - \mathscr{D}_{\mathrm{KL}}(q|P). \qquad (1.3.4)$$

Remarquons que par (1.3.3), $\mathrm{E}_{\mathrm{lbo}} \leq \log(p(D))$, d'où le nom *Evidence Lower Bound* (ELBO).

L'écriture (1.3.4) est suggestive. D'une part le premier terme de (1.3.4) est une espérance (conditionnelle) de la fonction de vraisemblance ; la distribution q va donc placer sa masse sur des $w \in \mathscr{W}$ qui maximisent la vraisemblance $\mathrm{L}(y|x, w)$. D'autre part, le deuxième terme est une divergence de Kullback-Leibler par rapport à la distribution *a priori*. La distribution q devra donc être proche de la distribution *a priori*. Maximiser l'$\mathrm{E}_{\mathrm{lbo}}$ revient donc à chercher une distribution q qui :

- permet au réseau de neurones de bien prédire les données observées (premier terme de (1.3.4))
- est proche de la distribution *a priori* (second terme de (1.3.4)).

On retrouve donc, dans ce cas de l'inférence variationnelle pour des réseaux bayésiens, le classique compromis entre risque empirique et terme de pénalisation. Ce lien avec les réseaux de neurones classiques n'est pas propre au cas de l'inférence variationnelle, comme le montre la remarque suivante.

Remarque 1.20 (Lien entre loi *a priori* et pénalisation)**.** *Considérons une loi a priori gaussienne ($\lambda > 0$)*

$$P(w) \propto e^{-\frac{\|w\|_2^2}{2\lambda}},$$

un jeu de données d'apprentissage $D = \{x_i, y_i\}_{i=1}^m$ et une vraisemblance $\mathrm{L}(y|x, w, D) \propto e^{-\mathrm{Loss}(w)}$, avec une fonction de coût $\mathrm{Loss}(w) = \frac{1}{2}\sum_{i=1}^m |y_i - f_w(x_i)|^2$. Par la formule de Bayes, on obtient la loi a posteriori

$$\begin{aligned} p(w|D) &\propto \mathrm{L}(y|x, w, D)P \\ &\propto e^{-\frac{1}{2}\sum_{i=1}^m |y_i - f_w(x_i)|^2} e^{-\frac{\|w\|_2^2}{2\lambda}}. \end{aligned}$$

Échantillonner des poids w suivant la loi a posteriori consiste à échantillonner des w proches de

$$\begin{aligned} \hat{w} &= \arg\max_w e^{-\frac{1}{2}\sum_{i=1}^m |y_i - f_w(x_i)|^2} e^{-\frac{\|w\|_2^2}{2\lambda}} \\ &= \arg\max_w -\frac{1}{2}\sum_{i=1}^m |y_i - f_w(x_i)|^2 - \frac{\|w\|_2^2}{2\lambda} \\ &= \arg\min_w \frac{1}{2}\sum_{i=1}^m |y_i - f_w(x_i)|^2 + \frac{\|w\|_2^2}{2\lambda}. \end{aligned}$$

C'est cette dernière quantité que l'on cherche à approcher dans le cas des réseaux non bayésiens. Ainsi, dans le formalisme bayésien, une loi a priori gaussienne correspond à une pénalisation ridge $\|\cdot\|_2^2$.

Dans la remarque précédente, λ est un hyperparamètre qui quantifie l'importance donnée au terme de pénalisation (ou de la loi a priori). Dans le formalisme bayésien, on peut de même introduire des hyperparamètres, au niveau de la loi *a posteriori*. Il a notamment été montré que des lois *a posteriori* froides ont de meilleures propriétés prédictives.

Remarque 1.21 (Loi *a posteriori* froides (*cold posteriors*))**.** *On peut écrire la loi a posteriori*

$$p(w|D) \propto \mathrm{L}(y|x, w)P(w) = e^{-U(w)}$$

avec $U(w) = -\log(\mathrm{L}(y|x, w)) - \log(P(w))$. Il a été montré empiriquement [WRV+20] qu'échantillonner des poids w suivant

$$e^{-\frac{U(w)}{T}}$$

avec $0 < T < 1$ (loi a posteriori froide) donnait lieu à de meilleures performances prédictives pour des réseaux de neurones sur des tâches de classification.

Dans le cas de l'inférence variationnelle, minimiser la divergence de Kullback-Leibler de q par rapport à $p(w|D) \propto e^{-\frac{U(w)}{T}}$ (voir (1.3.2)) correspond à maximiser l'ELBO (1.3.4) en remplaçant $\mathrm{L}(y|x, w)$ (respectivement $P(w)$) par $\mathrm{L}(y|x, w)^{1/T}$ (respectivement $P(w)^{1/T}$). Cependant, dans le contexte de l'inférence variationnelle, des études expérimentales [ZSDG18, OSK+19, ALMV20] ont suggéré d'affaiblir uniquement le terme de pénalisation (distribution *a priori*) de (1.3.4), et donc d'introduire, pour $\lambda = 1/T > 1$,

$$\mathrm{E_{lbo}}^\lambda = \int_{\mathscr{W}} \log(\mathrm{L}(y|x, w))q(w)\mathrm{d}w - \frac{1}{\lambda}\mathscr{D}_{\mathrm{KL}}(q|P). \tag{1.3.5}$$

Le cas $\lambda = T = 1$ correspond bien sûr à $\mathrm{E_{lbo}}$. Remarquons que $\mathrm{E_{lbo}}^\lambda$ s'obtient par les mêmes calculs que $\mathrm{E_{lbo}}$ avec une loi a posteriori $p(w|D) \propto \mathrm{L}(y|x, w)^\lambda P(w)$. Les études précédemment citées ont suggéré différentes valeurs de λ.

Référence	λ
[ZSDG18]	$\lambda \in \{2, 10\}$
[OSK$^+$19]	$\lambda \in \{5, 10\}$
[ALMV20]	$10^3 \leq \lambda \leq 10^5$

Dans [HMD$^+$22], les auteurs considèrent le cas d'un réseau bayésien à une couche cachée, dans le formalisme champ moyen, et montrent empiriquement que la valeur de λ doit être choisie proportionnellement au nombre de neurones N de la couche cachée.

Notre travail consistera premièrement à justifier rigoureusement que ce rapport entre λ et N est le bon. Nous verrons aussi d'autres conséquences à notre résultat, une fois introduits les algorithmes concernés.

1.3.2 Algorithmes étudiés et loi des grands nombres

Le réseau de neurones que nous considérons est le suivant : pour $N \geq 1$, $\boldsymbol{w} = (w^1, \dots, w^N)$ et $x \in \mathsf{X}$,

$$f^N_{\boldsymbol{w}}(x) = \frac{1}{N} \sum_{i=1}^{N} s(w_i, x),$$

où $s : \mathbf{R}^d \times \mathsf{X} \to \mathbf{R}$ est la fonction d'activation. Pour $N \geq 1$, nous considérons une famille variationnelle de lois gaussiennes indépendantes

$$\mathscr{Q}^N = \{q^N_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in (\mathbf{R}^{d+1})^N\} = \{\Psi_{\theta^1} \# \mathcal{N}(0, I_d) \otimes \cdots \otimes \Psi_{\theta^N} \# \mathcal{N}(0, I_d), (\theta^1, \dots, \theta^N) \in (\mathbf{R}^{d+1})^N\},$$

où, pour $\theta = (m, \sigma) \in \mathbf{R}^d \times \mathbf{R}$ and $z \in \mathbf{R}^d$, $\Psi_\theta(z) = m + \sigma z$, de sorte que

$$\mathsf{Z} \sim \mathcal{N}(0, I_d) \Rightarrow \Psi_\theta(\mathsf{Z}) \sim \mathcal{N}(m, \sigma^2 I_d), \ \theta = (m, \sigma).$$

La quantité à maximiser (sur $\boldsymbol{\theta} \in (\mathbf{R}^{d+1})^N$) est :

$$\mathrm{E_{lbo}}^N = \int_{(\mathbf{R}^d)^N} \log(\mathrm{L}(y|x, \boldsymbol{w})) q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} - \frac{1}{N} \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P). \tag{1.3.6}$$

Comme notre objectif est faire tendre N vers $+\infty$, nous allons nous placer dans un cas où l'intégrale sur $(\mathbf{R}^d)^N$ se simplifie (quand on considère le gradient de (1.3.6) selon $\boldsymbol{\theta}$) , en considérant la vraisemblance

$$\mathrm{L}(y|x, \boldsymbol{w}) = e^{-\frac{1}{2}|y - f^N_{\boldsymbol{w}}(x)|^2} / Z_{\boldsymbol{w}}.$$

ce qui mène à

$$\mathrm{E_{lbo}}^N = -\frac{1}{2} \int_{(\mathbf{R}^d)^N} |y - f^N_{\boldsymbol{w}}(x)|^2 q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} - \frac{1}{N} \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P) - \int_{(\mathbf{R}^d)^N} \log(Z_{\boldsymbol{w}}) q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) \mathrm{d}\boldsymbol{w}. \tag{1.3.7}$$

En pratique, la maximisation de $\mathrm{E_{lbo}}^N$ s'effectue par descente de gradient stochastique, c'est-à-dire qu'on utilise l'algorithme, pour $k \geq 1$,

$$\begin{cases} \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \nabla_{\boldsymbol{\theta}} \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}_k, x_k, y_k), \\ \boldsymbol{\theta}_0 \sim \mu_0^{\otimes N}, \end{cases}$$

où $(x_k, y_k)_{k\geq 0}$ sont les données d'entraînement, η est un taux d'apprentissage fixe. En utilisant la notation $\phi(\theta, z, x) = s(\Psi_\theta(z), x)$, le gradient du premier terme de (1.3.7) se simplifie alors en

$$\begin{aligned}
&-\frac{1}{2}\int_{(\mathbf{R}^d)^N} \nabla_{\theta^i} |y - f^N_{\boldsymbol{w}}(x)|^2 q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} \\
&= -\frac{1}{2}\int_{(\mathbf{R}^d)^N} \nabla_{\theta^i} \Big| y - \frac{1}{N}\sum_{j=1}^N \phi(\theta^j, z^j, x)\Big|^2 \gamma(z^1)\dots\gamma(z^N)\mathrm{d}z^1\dots\mathrm{d}z^N \\
&= \frac{1}{N^2}\sum_{j=1}^N \int_{(\mathbf{R}^d)^N} (y - \phi(\theta^j, z^j, x))\nabla_\theta\phi(\theta^i, z^i, x)\gamma(z^1)\dots\gamma(z^N)\mathrm{d}z^1\dots\mathrm{d}z^N \\
&= \frac{1}{N^2}\Big[\sum_{j=1, j\neq i}^N (y - \langle\phi(\theta^j, \cdot, x), \gamma\rangle)\langle\nabla_\theta\phi(\theta^i, \cdot, x), \gamma\rangle + \langle (y - \phi(\theta^i, \cdot, x))\nabla_\theta\phi(\theta^i, \cdot, x), \gamma\rangle\Big],
\end{aligned}$$

où $\langle\phi(\theta^j, \cdot, x), \gamma\rangle = \int_{\mathbf{R}^d} \phi(\theta^j, z, x)\gamma(z)\mathrm{d}z$. En choisissant une loi *a priori* gaussienne, le deuxième terme de (1.3.7) admet une expression explicite. Quant au troisième terme de (1.3.7), nous ferons le choix, en première approche, de le négliger. Nous obtenons ainsi l'algorithme : pour $k \geq 0$ et $i \in \{1, \dots, N\}$,

$$\begin{cases}
\theta^i_{k+1} = \theta^i_k - \dfrac{\eta}{N^2}\displaystyle\sum_{j=1, j\neq i}^N \Big(\langle\phi(\theta^j_k, \cdot, x_k), \gamma\rangle - y_k\Big)\langle\nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\rangle \\
\qquad\qquad -\dfrac{\eta}{N^2}\Big\langle (\phi(\theta^i_k, \cdot, x_k) - y_k)\nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\Big\rangle - \dfrac{\eta}{N}\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_k}|P^1_0), \\
\theta^i_0 \sim \mu_0.
\end{cases} \tag{Alg-Id}$$

Il s'agit du premier algorithme que nous allons étudier, et que nous appellerons « l'algorithme *idéal* », puisqu'il fait intervenir une intégrale (suivant $\gamma(z)\mathrm{d}z$) incalculable explicitement en général. En pratique, on utilise l'algorithme suivant, appelé *Bayes-by-backprop*, où chaque intégrale est approchée par une réalisation $\mathsf{Z} \sim \mathcal{N}(0, I_d)$,

$$\begin{cases}
\theta^i_{k+1} = \theta^i_k - \dfrac{\eta}{N^2}\displaystyle\sum_{j=1}^N \big(\phi(\theta^j_k, \mathsf{Z}^j_k, x_k) - y_k\big)\nabla_\theta\phi(\theta^i_k, \mathsf{Z}^i_k, x_k) - \dfrac{\eta}{N}\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_k}|P^1_0) \\
\theta^i_0 \sim \mu_0,
\end{cases} \tag{Alg-Bbb}$$

L'évolution des paramètres $(\theta^i_k)_{i=1}^N$ pour $k \geq 0$ s'étudiera là encore à travers l'évolution de la mesure empirique

$$\mu^N_t = \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i_{\lfloor Nt\rfloor}} \in \mathcal{P}(\mathbf{R}^{d+1}),\ t \geq 0.$$

Introduisons la fonction de coût $\mathscr{R}^\gamma : \mathcal{P}(\mathbf{R}^{d+1}) \to \mathbf{R}_+$,

$$\mathscr{R}^\gamma(\mu) = \frac{1}{2}\mathbf{E}_{(X,Y)\sim\pi}[|\langle\phi(\cdot, \cdot, X), \mu\otimes\gamma\rangle - Y|^2]$$

Notre premier résultat est le suivant :

Théorème 1.22. *La suite $(\mu^N)_{N\geq 1}$ définie par* (Alg-Id) *(resp.* (Alg-Bbb)*) converge en probabilité dans $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_c(\mathbf{R}^{d+1}))$ (resp. dans $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$) vers $\bar\mu \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_c(\mathbf{R}^{d+1}))$, définie comme l'unique solution de*

$$\begin{cases}
\partial_t\bar\mu_t = -\eta\mathrm{div}\Big(\bar\mu_t\Big[\nabla\Big(\frac{\delta\mathscr{R}^\gamma}{\delta\bar\mu_t}\Big) + \nabla_\theta\mathscr{D}_{\mathrm{KL}}(q^1_\cdot|P^1_0)\Big]\Big), \\
\bar\mu_0 = \mu_0.
\end{cases} \tag{1.3.8}$$

Commentaires sur le théorème 1.22 :

- Ce résultat justifie que la valeur de λ doit être du même ordre de grandeur que N, afin qu'aucun des deux termes constituant (1.3.5) ne domine l'autre. En effet, si λ était d'un ordre de grandeur plus petit, la suite $(\mu^N)_{N\geq 1}$ divergerait, et si λ était d'un ordre de grandeur plus grand, le terme $\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot|P^1_0)$ n'apparaîtrait pas dans l'équation (1.3.8).

- La limite champ moyen étant la même pour les algorithmes (Alg-Id) et (Alg-Bbb), ce résultat justifie que l'algorithme (Alg-Bbb) est une bonne approximation de l'algorithme idéalisé (Alg-Id).

- Pour simplifier les notations, nous avons considéré un même taux d'apprentissage pour le terme de vraisemblance et le terme de pénalisation ; nous aurions pu prendre deux taux d'apprentissages différents, ce qui n'aurait rien changé mathématiquement.

- C'est pour traiter l'algorithme idéalisé que nous avons fait l'hypothèse d'une fonction de coût quadratique. Dans le cas de l'algorithme (Alg-Bbb), notre analyse pourrait s'adapter au cas d'une autre fonction de coût, moyennant des hypothèses de moments sur le gradient de cette fonction.

L'équation (1.3.8) s'écrit, au sens faible : $\forall f \in \mathcal{C}^\infty_b(\mathbf{R}^{d+1})$ et $\forall t \in \mathbf{R}_+$,

$$\begin{aligned}\langle f, \bar{\mu}_t\rangle - \langle f, \mu_0\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot|P^1_0), \bar{\mu}_s\big\rangle \mathrm{d}s.\end{aligned}$$

Remarquons que cette équation peut se réécrire :

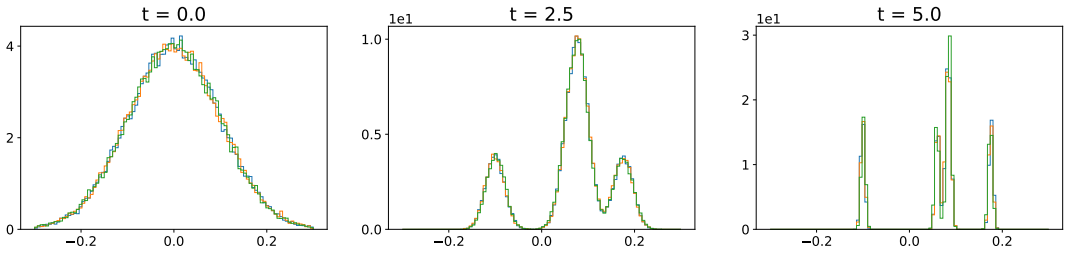
$$\begin{aligned}&\langle f, \bar{\mu}_t\rangle - \langle f, \mu_0\rangle \\ &= -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}\times(\mathbf{R}^d)^2} \big\langle \phi(\cdot, z_1, x) - y, \bar{\mu}_s\big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z_2, x), \bar{\mu}_s\big\rangle \gamma^{\otimes 2}(\mathrm{d}z_1, \mathrm{d}z_2)\pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &\quad - \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot|P^1_0), \bar{\mu}_s\big\rangle \mathrm{d}s.\end{aligned}$$

Ainsi, l'intégration suivant $\gamma^{\otimes 2}$ a le même statut que celle suivant π, c'est-à-dire que les variables z_1 et z_2 peuvent être considérées comme deux données supplémentaires, au même titre que (x, y), et devant être échantillonnées à chaque itération de l'algorithme. En adoptant ce point de vue, nous introduisons un nouvel algorithme :

$$\begin{cases} \theta^i_{k+1} = \theta^i_k - \dfrac{\eta}{N^2} \displaystyle\sum_{j=1}^N \big(\phi(\theta^j_k, \mathsf{Z}^1_k, x_k) - y_k\big) \nabla_\theta \phi(\theta^i_k, \mathsf{Z}^2_k, x_k) - \dfrac{\eta}{N} \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_k}|P^1_0) \\ \theta^i_0 \sim \mu_0, \end{cases} \qquad \text{(Alg-MinVI)}$$

où $\mathsf{Z}^1_k, \mathsf{Z}^2_k \sim \mathcal{N}(0, I_d)$ sont des variables indépendantes. Nous appellerons cet algorithme *Minimal-VI*, car il ne nécessite d'échantillonner que deux variables aléatoires gaussiennes à chaque itération (Z^1_k et Z^2_k) tandis qu'il faut en échantillonner N avec l'algorithme (Alg-Bbb) (Z^j_k, $j = 1, \ldots, N$). Remarquons aussi que ce nouvel algorithme n'est pas une approximation directe de l'algorithme (Alg-Id), puisque chaque intégrale de (Alg-Id) est approchée par une *même* variable aléatoire. Pourtant, comme nous le montrons dans le théorème suivant, ces deux algorithmes ont la même limite champ moyen.

FIGURE 1.4 : Distributions empiriques des paramètres $(\theta^i_{\lfloor Nt \rfloor})_{i=1}^N$ pour les trois algorithmes, pour N grand.



Théorème 1.23. *La suite $(\mu^N)_{N\geq 1}$ définie par* (Alg-MinVI) *est converge en probabilité dans $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ vers $\bar{\mu}$ définie au théorème 1.22.*

Ce résultat est illustré en figure 1.24, où l'on peut voir, à différents temps t, les distributions empiriques μ_t^N des trois algorithmes se concentrer autour $\bar{\mu}_t$, leur limite commune. Ce théorème justifie également que l'algorithme (Alg-MinVI) est une bonne approximation de l'algorithme (Alg-Id), au même titre que l'algorithme (Alg-Bbb), tout en étant moins coûteux en temps de calcul. Remarquons enfin qu'une application directe sur théorème de Tanaka-Sznitman ([Got98, Théroème 3.2]), d'obtenir un résultat qualitatif de propagation du chaos, au sens de la définition 1.2.

Corollaire 1.24. *Introduisons, pour $t \geq 0$, la loi jointe*

$$r_t^N = \mathrm{Law}(\theta^i_{\lfloor Nt \rfloor}, i = 1, \dots, N).$$

Alors, pour tout $t \geq 0$, la suite $(r_t^N)_{N\geq 1}$ est $\bar{\mu}_t$-chaotique, au sens de la définition 1.1.

Remarques sur les preuves des théorèmes 1.22 et 1.23. Les démonstrations complètes de ces deux théorèmes se trouvent au chapitre 3. Les étapes de la preuve sont les mêmes que celles de la preuve du théorème 1.10 : trouver une équation « pré-limite », montrer que la suite $(\mu^N)_{N\geq 1}$ est relativement compacte, passer à la limite dans l'équation « pré-limite » et enfin montrer l'unicité de la solution à l'équation (1.3.8). Néanmoins, la démonstration du théorème 1.22 dans le cas de l'algorithme idéal (Alg-Id) est grandement simplifiée, grâce au lemme suivant.

Lemme 1.25. *Pour tout $T > 0$, il existe $c_T > 0$ tel que pour tous $1 \leq i \leq N$ et $0 \leq k \leq \lfloor NT \rfloor$, presque sûrement,*

$$|\theta_k^i| \leq c_T.$$

Autrement dit, les poids du réseau sont inclus dans le compact

$$\Theta_T = [-c_T, c_T]^{d+1}.$$

Ce résultat de compacité permet d'obtenir des majorations presque sûres, et non plus des majorations sur les moments des différentes variables aléatoires, ce qui rend l'ensemble des calculs beaucoup plus simples. Plus fondamentalement, des simplifications substantielles des preuves sont obtenues ; mentionnons en particulier les deux points suivants.

- La preuve de la relative compacité de $(\mu^N)_{N\geq 1}$. Pour montrer la tension de mesures à valeur dans un espace de Skorohod, il y a toujours deux conditions à vérifier : l'une d'inclusion compacte (par exemple le point *(i)* de la proposition 1.19) et l'autre de régularité (par exemple le point *(ii)* de la proposition 1.19). Ici, μ_t^N étant une mesure à support compact

(grâce au lemme 1.25), la condition d'inclusion compacte est automatiquement vérifiée. La condition de régularité se montre grâce à l'inégalité suivante :

$$|\langle f, \mu_t^N\rangle - \langle f, \mu_s^N\rangle| \leq C\|f\|_\infty\Big(|t-s| + \frac{1}{N}\Big),\ \forall f \in \mathcal{C}^\infty(\Theta_T), 0 \leq t \leq s \leq T, N \geq 1. \quad (1.3.9)$$

- La preuve de la continuité des valeurs d'adhérence de $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d))$. En faisant tendre $t \to s^-$ dans (1.3.9), on obtient que les sauts de $s \mapsto \langle f, \mu_s^N\rangle$ tendent uniformément vers 0 lorsque $N \to \infty$, ce qui montre le lemme suivant.

Lemme 1.26. *Soit $f \in \mathcal{C}^\infty(\mathbf{R}^{d+1})$. Toute valeur d'adhérence de $(\langle f, \mu^N\rangle)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathbf{R})$ appartient presque sûrement à $\mathcal{C}(\mathbf{R}_+, \mathbf{R})$.*

De là, on va pouvoir montrer que toute valeur d'adhérence μ^* de $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d))$ est presque sûrement continue, c'est-à-dire que pour tout $T > 0$,

$$p.s.,\ \forall\ 0 \leq t < T,\ \forall f \in \mathcal{C}(\Theta_T),\ \langle f, \mu_s^*\rangle \to_{s\to t} \langle f, \mu_t^*\rangle. \quad (1.3.10)$$

Cette assertion est plus générale que celle du lemme 1.26, puisque le sous-ensemble de mesure 1 de cette dernière assertion est uniforme sur les fonctions test f. Pour intervertir les symboles « p.s. » et « $\forall f$ », on va utiliser le théorème de Stone-Weierstrass qui fournit une famille $(f_n)_{n\geq 1}$ de polynômes dense dans $\mathcal{C}(\Theta_T)$. Par le lemme 1.26, on a que presque sûrement pour tout $n \geq 1$, $t \mapsto \langle f_n, \mu_t^*\rangle$ est continue. Or, grâce au théorème de Stone-Weierstrass, on peut, pour toute fonction $f \in \mathcal{C}(\Theta_T)$, trouver une suite $(f_{n_m})_{m\geq 1} \subset (f_n)_{n\geq 1}$, qui converge uniformément vers f. En particulier, la suite $\{t \mapsto \langle f_{n_m}, \mu_t^*\rangle\}_{m\geq 1}$ converge uniformément vers $t \mapsto \langle f, \mu_t^*\rangle$, ce qui implique la continuité de cette dernière fonction, et donc l'assertion (1.3.10).

Les preuves pour les algorithmes (Alg-Bbb) et (Alg-MinVI) ne permettent pas de bénéficier du lemme 1.25 puisque les variables $\mathsf{Z}_k^i \sim \mathcal{N}(0, I_d)$ ne sont pas bornées. On doit alors utiliser des estimées sur les moments des variables aléatoires comme dans le théorème 1.10.

Il paraît clair que la convergence de μ^N vers $\bar{\mu}$ va être plus rapide pour les algorithmes (Alg-Id) et (Alg-Bbb) que pour l'algorithme (Alg-MinVI) ; on va dans la prochaine section s'intéresser à la vitesse de convergence de μ^N vers $\bar{\mu}$, par un théorème central limite.

1.3.3 Fluctuations autour de la limite champ moyen

Cette section présente un travail en préparation relatif à la vitesse de convergence de μ^N vers $\bar{\mu}$, pour les trois algorithmes présentés précédemment. Nous avons vu que deux de ces algorithmes sont implémentables : l'algorithme (Alg-Bbb) qui est communément utilisé et l'algorithme (Alg-MinVI), que nous avons introduit, moins coûteux en temps de calcul. Pour vérifier que la suite $(\mu^N)_{N\geq 1}$ relative à ce dernier algorithme converge moins vite vers la limite champ moyen $\bar{\mu}$ que celle relative à l'algorithme (Alg-Bbb), l'étude du comportement de la mesure empirique μ^N autour de $\bar{\mu}$ est nécessaire. Une fois encore, nous introduisons le processus de fluctuations

$$\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t) \in \mathcal{H}^{-J}(\mathbf{R}^{d+1}), \quad (1.3.11)$$

où $J \geq 1$ est suffisamment grand. Nous montrons dans le théorème suivant que la limite en loi de η^N, notée $\bar{\eta}$, satisfait une équation de la forme (1.2.15), avec un opérateur différentiel faisant intervenir la divergence de Kullback-Leibler par rapport à la distribution *a priori*. Nous donnons ici l'expression explicite de cette équation.

Théorème 1.27. *Considérons μ^N définie par les algorithmes* (Alg-Id)*,* (Alg-Bbb) *ou* (Alg-MinVI)*, et $\bar{\mu}$ définie par* (1.3.8)*. Alors la suite $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}^{d+1}))$ définie en* (1.3.11) *converge en loi vers $\bar{\eta} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J}(\mathbf{R}^{d+1}))$, l'unique solution faible de l'équation : $\forall t \geq 0$, $\forall f \in \mathcal{C}_b^\infty(\mathbf{R}^{d+1})$,*

$$\begin{aligned}\langle f, \bar{\eta}_t\rangle - \langle f, \bar{\eta}_0\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\eta}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \bar{\eta}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \left\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\eta}_s \right\rangle \mathrm{d}s + \mathscr{G}_t[f],\end{aligned}$$

où $\bar{\eta}_0$ est définie par

$$(\langle f_1, \bar{\eta}_0\rangle, \ldots, \langle f_k, \bar{\eta}_0\rangle)^T \sim \mathcal{N}(0, \Gamma(f_1, \ldots, f_k)),\ f_1, \ldots, f_k \in \mathcal{H}_0^J(\mathbf{R}^{d+1}),\ k \geq 1,$$

avec $\Gamma(f_1, \ldots, f_k)$ la matrice de covariance de $(f_1(\theta_0^1), \ldots, f_k(\theta_0^1))^T$ et où $\mathscr{G}$ est un G*-process (voir définition 1.14) de structure de covariance :*

- *dans le cas des algorithmes* (Alg-Id) *et* (Alg-Bbb) *: $\forall f, g \in \mathcal{H}_0^J(\mathbf{R}^{d+1}), \forall 0 \leq s \leq t$,*

$$\mathrm{Cov}(\mathscr{G}_t[f], \mathscr{G}_s[g]) = \eta^2 \int_0^s \mathrm{Cov}(\mathrm{Q}[f](x, y, \bar{\mu}_v), \mathrm{Q}[g](x, y, \bar{\mu}_v))\mathrm{d}v,$$

  où $\mathrm{Q}[f](x, y, \bar{\mu}_v) = \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_v \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_v \otimes \gamma\rangle$.

- *dans le cas de l'algorithme* (Alg-MinVI) *: $\forall f, g \in \mathcal{H}_0^J(\mathbf{R}^{d+1}), \forall 0 \leq s \leq t$,*

$$\mathrm{Cov}(\mathscr{G}_t[f], \mathscr{G}_s[g]) = \eta^2 \int_0^s \mathrm{Cov}(\mathrm{Q}[f](x, y, z^1, z^2, \bar{\mu}_v), \mathrm{Q}[g](x, y, z^1, z^2, \bar{\mu}_v))\mathrm{d}v,$$

  où $\mathrm{Q}[f](x, y, z^1, z^2, \bar{\mu}_v) = \langle \phi(\cdot, z^1, x) - y, \bar{\mu}_v\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z^2, x), \bar{\mu}_v\rangle$.

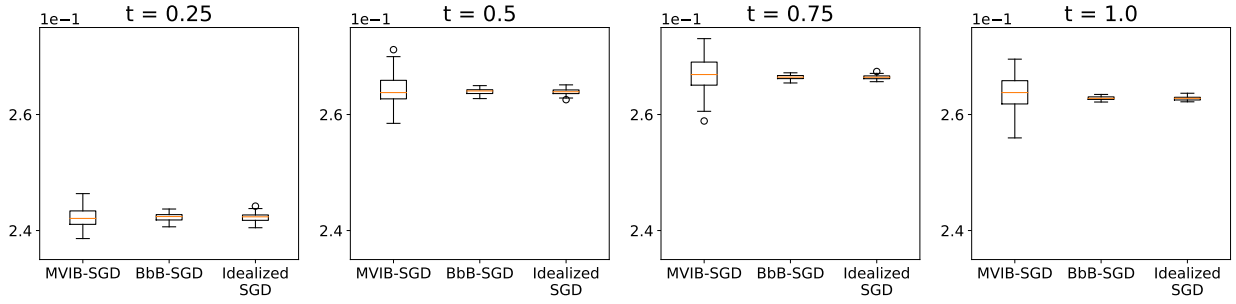
Commentaires sur le théorème 1.27 :

- Remarquons tout d'abord que les algorithmes (Alg-Id) et (Alg-Bbb) vérifient le même théorème central limite. Cela montre donc que l'algorithme utilisé en pratique (Alg-Bbb) est une approximation exacte de l'algorithme (Alg-Id), non seulement au sens de la limite champ moyen (comme nous l'avons montré par le théorème 1.22), mais également au sens des fluctuations autour de cette limite. Il s'agit du premier résultat mathématique en ce sens.

- La structure de covariance du G-process ne fait pas intervenir les variables $\mathsf{Z}^1, \ldots \mathsf{Z}^N$ dans le cas de l'algorithme (Alg-Bbb), tandis qu'elle fait intervenir les variables $\mathsf{Z}^1, \mathsf{Z}^2$ dans le cas de l'algorithme (Alg-MinVI). Ceci met en évidence la différence de statut de ces variables : dans le cas de l'algorithme (Alg-Bbb), les variables $\mathsf{Z}^1, \ldots \mathsf{Z}^N$ font partie du modèle, au même titre que les paramètres θ^i, tandis que dans le cas l'algorithme (Alg-MinVI), les variables $\mathsf{Z}^1, \mathsf{Z}^2$ apparaissent comme extérieures au modèle, comme des données additionnelles, sur lesquelles le modèle s'entraîne, comme nous l'avions déjà évoqué avant d'introduire cet algorithme.

- Le schéma de preuve est identique à celle du théorème 1.17, hormis pour la convergence de $(\sqrt{N}M^N)_{N\geq 1}$ vers un G-process dans le cas de l'algorithme (Alg-Bbb) (voir *(iii)* dans le schéma de preuve du théorème 1.17). En effet, comme dit précédemment, les variables

FIGURE 1.5 : Réalisations de $\langle f, \mu_t^N \rangle$ à différents temps, illustrant la plus lente convergence de μ^N vers $\bar{\mu}$ pour l'algorithme (Alg-MinVI).



$\mathsf{Z}^1, \dots \mathsf{Z}^N$ font dans ce cas partie du modèle, ce qui impose, pour obtenir la structure de covariance du G-process, de considérer la mesure empirique jointe

$$\rho_t^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{(\theta^i_{\lfloor Nt \rfloor}, \mathsf{Z}^i_{\lfloor Nt \rfloor})} \in \mathcal{P}(\mathbf{R}^{d+1} \times \mathbf{R}^d).$$

Remarquons qu'on ne peut pas obtenir de théorème central limite (ni même de loi des grands nombres) directement sur la suite de processus $\{t \mapsto \rho_t^N\}_{N \geq 1}$, puisque celle-ci n'est pas relativement compacte, du fait que la suite $(\mathsf{Z}^i_k)_{k \geq 0}$ est une suite de variables aléatoires indépendantes. Nous avons en revanche démontré le résultat de convergence suivant, qui est suffisant pour appliquer le théorème central limite sur les martingales (déjà utilisé dans le *(iii)* du schéma de preuve du théorème 1.17).

Proposition 1.28. *Si $\mu^N \to \bar{\mu}$ presque sûrement dans $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$, alors presque sûrement, pour tous $0 \leq t \leq T$,*

$$\rho_t^N \xrightarrow[N \to \infty]{} \bar{\mu}_t \otimes \gamma \text{ dans } \mathcal{P}(\mathbf{R}^{d+1} \times \mathbf{R}^d).$$

Nous étudierons précisément la convergence de la martingale vers le G-process au chapitre 4.

- Comme pour le théorème 1.17, $\mathscr{G}_t$ n'est *a priori* pas indépendant de $\{\bar{\eta}_s, s \leq t\}$, ce qui ne permet pas d'obtenir directement une information mathématique sur la différence de variance de $\bar{\eta}_t$ selon que l'on considère les algorithmes (Alg-Bbb) ou (Alg-MinVI).

La dernière remarque montre que l'information fournie par le théorème central limite est limitée, au sens où il ne permet pas de quantifier précisément la plus lente convergence de μ^N vers $\bar{\mu}$, dans le cas de l'algorithme (Alg-MinVI). En pratique on observe bien cette différence de vitesse de convergence, comme on peut le voir sur la figure 1.5. La question naturelle est alors de savoir, pour une précision $\epsilon > 0$ et un temps d'apprentissage t donnés, quel nombre N de neurones il faut considérer pour que μ_t^N soit proche de $\bar{\mu}_t$ à une précision d'ϵ près, le nombre d'itérations de l'algorithme de descente de gradient étant alors, rappelons-le, de $\lfloor Nt \rfloor$. Bien sûr, ce nombre N sera plus grand pour l'algorithme (Alg-MinVI) que pour l'algorithme (Alg-Bbb). On pourra alors calculer le nombre de variables aléatoires Z^i_k simulées par les deux algorithmes, et savoir ainsi si l'algorithme (Alg-MinVI) est plus ou moins coûteux en temps de calcul, *eu égard à sa position autour de sa limite champ moyen*. C'est à cette question que nous comptons répondre par des simulations numériques, qui feront partie de ce travail en préparation.

1.4 Conclusions et perspectives

Le fruit de cette thèse a tout d'abord été d'apporter de la rigueur à des résultats déjà connus, concernant les réseaux de neurones classiques, où nous avons donné une preuve rigoureuse de la loi des grands nombres et du théorème central limite. Nous avons mis en évidence que l'analyse de l'équation satisfaite par la limite champ moyen est délicate, et que peu de résultats à son sujet sont connus à ce jour. En particulier, la question de la convergence en temps long vers un minimum de $\mathscr{R}$ reste ouverte. En outre, le théorème central limite nous a permis de quantifier mathématiquement l'influence du mini-batch sur la réduction de variance, lorsque l'on considère la limite champ moyen ; nous avons également identifié une intensité de bruit additionnel limite, au-delà de laquelle le théorème central limite n'a plus lieu. Il serait toutefois intéressant de quantifier l'erreur avec la limite champ moyen, dans le cas $\beta = 1/2$ où l'équation limite est perturbée. D'une manière générale, les limites de l'approche champ moyen résident dans la complexité des équations obtenues (loi des grands nombres et théorème central limite), qu'il est difficile d'analyser. Une étude plus approfondie des équations aux dérivées partielles obtenues par ces deux théorèmes s'avère donc nécessaire pour pouvoir pleinement tirer profit de cette approche, tant sur le plan théorique que pratique (le théorème central limite pourrait être utilisé pour quantifier l'incertitude de prédiction du réseau, ou pour simuler d'autres réseaux de neurones et obtenir ainsi des *deep ensembles* [LPB17] par exemple).

Pour aller plus loin dans ces résultats asymptotiques, les grandes déviations autour de la limite champ moyen pourraient être étudiées, en utilisant les méthodes similaires à celles employées par [DLR20]. Dans un autre registre, pour obtenir des taux de convergence à N fixé, l'étude d'inégalités de concentration pourrait être menée, afin de compléter les résultats obtenus dans [MMM19] en obtenant des inégalités faisant intervenir les distances de Wasserstein entre la mesure empirique μ_t^N et la mesure limite $\bar{\mu}_t$. À cette fin, les méthodes développées dans [DGW04, FG15] pourraient être mises en œuvre.

Notre analyse champ moyen de l'inférence variationnelle pour les réseaux bayésiens montre plusieurs choses. Premièrement, elle prouve que le terme de pénalisation de l'ELBO doit être tempéré par un facteur $1/N$. Deuxièmement, elle montre que l'algorithme (Alg-Bbb) utilisé en pratique, en utilisant une seule variable aléatoire pour approcher une intégrale, est une bonne approximation de l'algorithme idéal (Alg-Id), tant au niveau de la loi des grands nombres que du théorème central limite. Enfin l'équation satisfaite par la limite champ moyen nous a permis de trouver un nouvel algorithme, qui partage cette même limite champ moyen. Ce nouvel algorithme diffère des deux autres au niveau des fluctuations autour de cette limite (à cause du G-process) ; un travail en cours complètera ces résultats par une analyse numérique.

Enfin, la généralisation de ces résultats au cas de réseaux profonds (avec plusieurs couches), et tout d'abord la limite champ moyen vers une équation de type flot gradient, est un problème difficile et ouvert. En effet, dans le cas d'un réseau à une couche caché, chaque poids W^i est un élément de $\mathbf{R}^d$, où d est la dimension des données d'entrées. Si l'on rajoute une couche, chaque poids de cette nouvelle couche sera un élément de $\mathbf{R}^N$, où N est le nombre de neurones sur la première couche. Passer de une à deux couches cachées nécessite donc de travailler dans de nouveaux espaces, et d'utiliser d'autres techniques d'analyse. Dans le cas d'un entraînement continu, une analyse est menée dans [CVEB22]. D'autres méthodes de généralisation au cas de plusieurs couches ont été proposées dans [AOY19, Ngu19, NP20, SS22].

Chapter 2

Law of large numbers and central limit theorem for wide two-layer neural networks: the mini-batch and noisy case

This chapter corresponds to the preprint [DGMN22]. It is a joint work with Arnaud Guillin, Manon Michel and Boris Nectoux.

Contents

Abstract

In this work, we consider a wide two-layer neural network and study the behavior of its empirical weights under a dynamics set by a stochastic gradient descent along the quadratic loss with mini-batches and noise. Our goal is to prove a trajectorial law of large number as well as a central limit theorem for their evolution. When the noise is scaling as $1/N^\beta$ and $1/2 < \beta \leq \infty$, we rigorously derive and generalize the LLN obtained for example in [CRBVE20, MMM19, SS20b].

When $3/4 < \beta \leq \infty$, we also generalize the CLT (see also [SS20a]) and further exhibit the effect of mini-batching on the asymptotic variance which leads the fluctuations. The case $\beta = 3/4$ is trickier and we give an example showing the divergence with time of the variance thus establishing the instability of the predictions of the neural network in this case. It is illustrated by simple numerical examples.

2.1 Setting and main results

2.1.1 Introduction

Setting and purpose of this work. Thanks to their impressive results, deep learning techniques have nowadays become standard supervised learning methods in various fields of engineering or research [GBC16]. A robust understanding of their behavior and efficiency is however still lacking and a large effort is put towards achieving mathematical foundations of empirical observations. Among this effort, the case of wide two-layer single network, and its connection with mean-field network, has particularly been fruitful, as considered for example in [RVE18a, MMN18, SS20b, SS20a]. In such setting, a convergence towards a limit PDE system can be established when the neuron numbers goes to infinity. The behavior in long time of this limit PDE may then give an easier framework to establish the convergence towards minimizers of the loss function of the neural network. Partial results can be found in this direction [MMN18, CB18a] but as underlined in [E20], a lot still remains to be understood and proved mathematically rigorously. In this context, our work is two-fold. First, we will concern ourselves with the mathematical justification of the law of large numbers and central limit theorems of the trajectory of the empirical measure of the weights, under the optimization by a stochastic gradient descent (SGD), with mini-batching and in the presence of noise with a range of scalings. Mini-batch SGD [BCN18] is widely used in machine learning since it allows for shorter training times thanks to parallelisation, while reducing the variance in SGD estimates. How to choose the optimal mini-batch size, and furthermore with theoretical guarantees, remains an active research line [KMN$^+$17, SL18, GLQ$^+$19]. Introducing noise in SGD, as considered in [MMN18], can lead to better generalisation perfomance thanks to an improved ability to escape saddle points, as shown in [JNG$^+$21]. Note that this differs from the analysis approach consisting in directly modelizing the noise of SGD as for instance done in [WHX$^+$20, SGN$^+$19]. Second we will do so by providing a rigorous framework which could be generalized to study overparametrized limit of other neural networks (e.g. deep ensemble, bayesian neural networks, ...). Thus, the benefit of the overparametrized limit and its convexification of the loss landscape through a non-linear PDE could lead in these different architectures to derivations of theoretical guarantees of convergence, while it remains hard to analyse these landscapes directly in the case of a finite number of neurons, even large.

Let us now precise the framework for this paper. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and $\mathcal{X}$ and $\mathcal{Y}$ be subsets of $\mathbf{R}^n$ ($n \geq 1$) and $\mathbf{R}$ respectively. In this work, we consider the following two-layer neural network

$$g_W^N(x) := \frac{1}{N} \sum_{i=1}^{N} \sigma_*(W^i, x), \tag{2.1.1}$$

where $x \in \mathcal{X}$ denotes the input data, $g_W^N(x) \in \mathbf{R}$ the output returned by the neural network, $\sigma_* : \mathbf{R}^d \times \mathcal{X} \to \mathbf{R}$ the activation function, $N \geq 1$ the number of neurons on the hidden layer, and $W = (W^1, \dots, W^N) \in (\mathbf{R}^d)^N$ are the weights to optimize ($d \geq 1$). In the supervised learning setting, a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is distributed according to $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, where $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denotes the set of probability measures on $\mathcal{X} \times \mathcal{Y}$. Ideally, one chooses the weights $W = (W^1, \dots, W^N)$ as a global minimizer of the risk $\mathbf{E}_\pi[\mathsf{L}(g_W^N(x), y)]$, where $\mathsf{L} : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ is the so-called loss function ($\mathbf{E}_\pi$ stands for the expectation when $(x, y) \sim \pi$). In this work,

we consider the square loss function out of simplicity, but other loss function or classification problem could be considered, namely:

$$\mathsf{L}(g_W^N(x), y) = \frac{1}{2}\big|g_W^N(x) - y\big|^2.$$

Since the risk can not be computed (because π is unknown), the parameters are usually learned by stochastic gradient descent. In this work, we consider the mini-batch setting with weak noise, which is defined as follows. First, for $k \geq 0$, consider $((x_k^n, y_k^n))_{n\geq 1}$ a sequence of random elements on $\mathcal{X} \times \mathcal{Y}$ (each (x_k^n, y_k^n) being distributed according to π), and N_k a random element with values in $\mathbf{N}^* = \{1, 2, 3, \ldots\}$. Then, the mini-batch B_k is defined by:

$$B_k = \{(x_k^1, y_k^1), \ldots, (x_k^{N_k}, y_k^{N_k})\}, \text{ in particular } |B_k| = N_k, \text{ where } |B_k| \text{ denotes the cardinality of } B_k.$$

In addition, at each iteration of SGD, we add a Gaussian noise term, whose variance is scaled according to $N^{-2\beta}$, with $\beta > \frac{1}{2}$, hence qualified *weak*. Note that the case of Gaussian noise with $\beta = 1/2$ is addressed in [MMN18] and could also be considered here in our setting, but with additional assumptions to integrate the noise in the limit process.

Thus, the SGD algorithm we consider is the following : for $k \geq 0$ and $i \in \{1, \ldots, N\}$,

$$\begin{cases} W_{k+1}^i = W_k^i + \dfrac{\alpha}{N|B_k|} \displaystyle\sum_{(x,y)\in B_k} (y - g_{W_k}^N(x)) \nabla_W \sigma_*(W_k^i, x) + \dfrac{\varepsilon_k^i}{N^\beta}, \\ W_0^i \sim \mu_0, \end{cases} \tag{2.1.2}$$

where $\varepsilon_k^i \sim \mathcal{N}(0, I_d)$ and $\mu_0 \in \mathcal{P}(\mathbf{R}^d)$. The evolution of the weights is tracked through their empirical distribution ν_k^N (for $k \geq 0$) and its scaled version μ_t^N (for $t \in \mathbf{R}_+$), which are defined as follows:

$$\nu_k^N := \frac{1}{N}\sum_{i=1}^N \delta_{W_k^i} \quad \text{and} \quad \mu_t^N := \nu_{\lfloor Nt \rfloor}^N.$$

For an element $\mu \in \mathcal{M}_b(\mathbf{R}^d)$ (the space of bounded countably additive measures on $\mathbf{R}^d$), we use the notation

$$\langle f, \mu \rangle_{\mathrm{m}} = \int_{\mathbf{R}^d} f(w)\mu(\mathrm{d}w),$$

for any $f : \mathbf{R}^d \to \mathbf{R}$ such that $\int_{\mathbf{R}^d} f(w)\mu(\mathrm{d}w)$ exists. If no confusion is possible, we simply denote $\langle f, \mu \rangle_{\mathrm{m}}$ by $\langle f, \mu \rangle$. For instance, considering the neural network (2.1.1), we have, for any $x \in \mathcal{X}$,

$$g_{W_k}^N(x) = \frac{1}{N}\sum_{i=1}^N \sigma_*(W_k^i, x) = \langle \sigma_*(\cdot, x), \nu_k^N \rangle, \quad k \geq 0.$$

In this work, we prove that the the whole trajectory of the scaled empirical measures of the weights defined by (2.1.2) (namely $\{t \mapsto \mu_t^N, t \in \mathbf{R}_+\}_{N\geq 1}$) satisfies a law of large numbers and a central limit theorem, see respectively Theorem 2.1 and Theorem 2.8. We also exhibit a particular fluctuation behavior depending on the value of the parameter β ruling the weakness of the added noise.

Related works. Law of large numbers and central limits theorems have been obtained for several kinds of mean-field interacting particle systems in the mathematical literature, see for instance [Szn91, HM86, FM97a, JM98a, DLR19a, DMG99, KX04] and references therein. When considering particle systems arising from the SGD-minimization problem in a two-layer neural network, we refer to [MMN18] for a law of large numbers on the empirical measure at fixed times, see also [MMM19]. We also refer to [RVE18a] where conditions for global convergence of the GD on the ideal loss and of the SGD with mini-batches increasing in size with N, as well

as the scaling of the error with the size of the network, are established from formal asymptotic arguments. Doing so, they also observe with increasing mini-batch size in the SGD the reduction of the variance of the process leading the fluctuations of the empirical measure of the weights (see [RVE18a, Arxiv-V2. Sec 3.3]), until the mini-batches are large enough to recover the situation of the idealized gradient descent (similar to an infinite batch), which leads to other order of fluctuations (see [RVE18a, Arxiv-V2. Prop 2.3]). We also refer to [CRBVE20] for a similar line of work on the GD on the empirical loss. A law of large numbers and a central limit theorem on the whole trajectory of the empirical measure are also obtained in [SS20b, SS20a] for a standard SGD scheme. We also mention the work done in [DBDFS20] on propagation of chaos for SGD with different step-size schemes. In this work, and compared to the existing literature dealing with the SGD minimization problem in two-layer neural networks, we provide a rigorous proof with precise justifications of all steps of the existence of the limit PDE (in particular, uniqueness and relative compactness) in the law of large numbers as well as the limit process for the central limit theorem on the trajectory of the empirical measure. This will be the basis for future works on deep ensembles or overparameterized bayesian neural networks. We furthermore do so in a more general variant of SGD with mini-batching of any size and weak noise (see (2.1.2)). A noisy SGD was also considered in [MMN18], corresponding to $\beta = 1/2$ in our setting, for which they obtain for the LLN a different limit PDE than in the non-noisy case (presence of an additionnal regularizing Laplacian term in the limit equation). While we could recover in a straightforward manner a trajectorial version of [MMN18], we consider here out of concision the range $\beta > 1/2$, showing a single limit PDE for the LLN, and obtain a similar result for $\beta > 3/4$ for the CLT, while showing analytically for $\beta = 3/4$ and numerically for $\beta \leq 3/4$ a particular fluctuation behavior. Furthermore, we analytically show the expected reduction, with the mini-batch size, of the variance of the process leading the fluctuations of the weight empirical measure and numerically display the reduction of the global variance.

2.1.2 Main results

The sequence $\{t \mapsto \mu_t^N, t \in \mathbf{R}_+\}_{N\geq 1}$ is studied as a sequence of processes with values in the dual of some (weighted) Hilbert space on $\mathbf{R}^d$. These Hilbert spaces are introduced in the next section.

Notation and assumptions

Weighted Sobolev spaces. Following [AF03, Chapter 3], we consider, for a function $g \in \mathcal{C}_c^\infty(\mathbf{R}^d)$ (the space of functions $g : \mathbf{R}^d \to \mathbf{R}$ of class $\mathcal{C}^\infty$ with compact support), the following norm, defined for $J \in \mathbf{N}$ and $b \geq 0$:

$$\|g\|_{\mathcal{H}^{J,b}} := \Big(\sum_{|k|\leq J} \int_{\mathbf{R}^d} \frac{|D^k g(x)|^2}{1+|x|^{2b}} \mathrm{d}x \Big)^{1/2}.$$

Let $\mathcal{H}^{J,b}(\mathbf{R}^d)$ be the closure of the set $\mathcal{C}_c^\infty(\mathbf{R}^d)$ for this norm. The space $\mathcal{H}^{J,b}(\mathbf{R}^d)$ is a Hilbert space when endowed with the norm $\|\cdot\|_{\mathcal{H}^{J,b}}$. The associated scalar product on $\mathcal{H}^{J,b}(\mathbf{R}^d)$ will be denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}^{J,b}}$. We denote by $\mathcal{H}^{-J,b}(\mathbf{R}^d)$ its dual space. For an element $\Phi \in \mathcal{H}^{-J,b}(\mathbf{R}^d)$, we use the notation

$$\langle f, \Phi \rangle_{J,b} = \Phi[f], \ f \in \mathcal{H}^{J,b}(\mathbf{R}^d).$$

For ease of notation, and if no confusion is possible, we simply denote $\langle f, \Phi \rangle_{J,b}$ by $\langle f, \Phi \rangle$. Let us now define $\mathcal{C}^{J,b}(\mathbf{R}^d)$ as the space of functions $g : \mathbf{R}^d \to \mathbf{R}$ with continuous partial derivatives up to order $J \in \mathbf{N}$ such that

$$\text{for all } |k| \leq J, \ \lim_{|x|\to\infty} \frac{|D^k g(x)|}{1+|x|^b} = 0.$$

This space is endowed with the norm

$$\|g\|_{\mathcal{C}^{J,b}} := \sum_{|k|\le J} \sup_{x\in\mathbf{R}^d} \frac{|D^k g(x)|}{1+|x|^b}.$$

We also introduce $\mathcal{C}_b(\mathbf{R}^d)$, the space of bounded continuous functions $g : \mathbf{R}^d \to \mathbf{R}$, endowed with the supremum norm. We also denote by $\mathcal{C}_b^\infty(\mathbf{R}^d)$ the space of smooth functions over $\mathbf{R}^d$ whose derivatives of all order are bounded. We have $\mathcal{C}_b^\infty(\mathbf{R}^d) \subset \mathcal{H}^{J,b}(\mathbf{R}^d)$ as soon as $b > d/2$ (more generally $x \in \mathbf{R}^d \mapsto (1-\chi(x))|x|^a \in \mathcal{H}^{J,b}(\mathbf{R}^d)$ if $b - a > d/2$, where $\chi \in \mathcal{C}_c^\infty(\mathbf{R}^d, [0,1])$ equals 1 near 0).

Weighted Sobolev embeddings. We recall that from [FM97a, Section 2],

$$\mathcal{H}^{\ell+j,a}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{j,a+b}(\mathbf{R}^d) \text{ when } \ell > d/2,\ b > d/2, \text{ and } a, j \ge 0 \tag{2.1.3}$$

where $\hookrightarrow_{\text{H.S.}}$ means that the embedding is of Hilbert-Schmidt type, and

$$\mathcal{H}^{\ell+j,a}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{j,a}(\mathbf{R}^d) \text{ when } \ell > d/2, \text{ and } a, j \ge 0. \tag{2.1.4}$$

We set

$$L = \lceil \frac{d}{2} \rceil + 3,\ \gamma = 4\lceil \frac{d}{2} \rceil + 5, \text{ and } \gamma_* := \gamma + 1. \tag{2.1.5}$$

According to (2.1.4) and since $\gamma_* > \gamma$, it holds:

$$\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d). \tag{2.1.6}$$

We set throughout this work, for all $N \ge 1$:

$$\mu^N := \{t \mapsto \mu_t^N, t \in \mathbf{R}_+\}.$$

When E is a metric space, we denote by E' its dual and by $\mathcal{D}(\mathbf{R}_+, E)$ the set of càdlàg functions from $\mathbf{R}_+$ to E. For $b \ge 0$ and for all $N \ge 1$, μ^N is a random element of $\mathcal{D}(\mathbf{R}_+, \mathcal{C}^{0,b}(\mathbf{R}^d)')$, and thus also of $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J,b}(\mathbf{R}^d))$, as soon as $J > d/2$ (by (2.1.4)).

Let for $k \ge 1$,

$$\mathcal{P}_k(\mathbf{R}^d) := \big\{\mu \in \mathcal{P}(\mathbf{R}^d),\ \int_{\mathbf{R}^d} |w|^k \mu(\mathrm{d}w) < +\infty\big\}, \tag{2.1.7}$$

which is endowed with the Wasserstein distance

$$\mathsf{W}_k(\mu,\nu) = \big[\inf\{\mathbf{E}[|X-Y|^k|],\ \mathbf{P}_X = \mu \text{ and } \mathbf{P}_Y = \nu\}\big]^{1/k}.$$

We refer for instance to [San15, Chapter 5] for more about these spaces. We recall that $\mathsf{W}_1(\mu,\nu) \le \mathsf{W}_k(\mu,\nu)$ $(k \ge 1)$ and the dual formula for $\mathsf{W}_1(\mu,\nu)$:

$$\mathsf{W}_1(\mu,\nu) = \sup\Big\{\big|\int_{\mathbf{R}^d} f(w)\mathrm{d}\mu(w) - \int_{\mathbf{R}^d} f(w)\nu(\mathrm{d}w)\big|,\ \|f\|_{\text{Lip}} \le 1\Big\}. \tag{2.1.8}$$

Note also that for all $N \ge 1$, μ^N is a random element of $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_q(\mathbf{R}^d))$, for all $q \ge 0$.

Assumptions. For $N \ge 1$, we introduce the $\boldsymbol{\sigma}$-algebras,

$$\mathcal{F}_0^N = \boldsymbol{\sigma}\big\{\{W_0^i\}_{i=1}^N\big\} \text{ and, for } k \ge 1,\ \mathcal{F}_k^N = \boldsymbol{\sigma}\big\{W_0^i, \{B_j\}_{j=0}^{k-1}, \{\varepsilon_j^i\}_{j=0}^{k-1},\ i \in \{1,\dots,N\}\big\}. \tag{2.1.9}$$

The main assumptions of this work are the following:

A1. For all $k, q \in \mathbf{N}$, $|B_q| \perp\!\!\!\perp ((x_k^n, y_k^n))_{n\ge1}$. In addition, for all $k \in \mathbf{N}$, $\big(|B_k|, ((x_k^n, y_k^n))_{n\ge1}\big) \perp\!\!\!\perp \mathcal{F}_k^N$.

A2. The activation function $\sigma_* : \mathbf{R}^d \times \mathcal{X} \to \mathbf{R}$ belongs to $\mathcal{C}_b^\infty(\mathbf{R}^d \times \mathcal{X})$.

A3. For all $\ell \neq k \in \mathbf{N}$, $((x_\ell^n, y_\ell^n))_{n\geq 1} \perp\!\!\!\perp ((x_k^n, y_k^n))_{n\geq 1}$. In addition, for all $k \in \mathbf{N}$, $((x_k^n, y_k^n))_{n\geq 1}$ is a sequence of i.i.d random variables from $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, and $\mathbf{E}[|y|^{16\gamma_*}]$ is finite.

A4. The randomly initialized parameters $\{W_0^i\}_{i=1}^N$ are i.i.d. with a distribution $\mu_0 \in \mathcal{P}(\mathbf{R}^d)$ such that $\mathbf{E}[|W_0^1|^{8\gamma_*}] < +\infty$.

A5. For all $k \in \mathbf{N}$ and $i \in \{1, \dots, N\}$, $\varepsilon_k^i \sim \mathcal{N}(0, I_d)$ and $\varepsilon_k^i \perp\!\!\!\perp \mathcal{F}_k^N$. In addition, for all $k, l \in \mathbf{N}$ and $i, j \in \{1, \dots, N\}$ such that $(i, k) \neq (j, l)$, $\varepsilon_k^i \perp\!\!\!\perp \varepsilon_l^j$.

Law of large numbers for the empirical measure

Statement of the law of large numbers. The first main result of this work is a law of large numbers for the trajectory of the scaled empirical measures.

Theorem 2.1. *Let $\beta > 1/2$ and assume* **A1**-**A5**. *Then, the sequence $(\mu^N)_{N\geq 1}$ converges in probability to a deterministic element $\bar{\mu}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. In addition, $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ and it is the unique solution in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d)$ of the following measure-valued equation:*

$$\forall f \in \mathcal{C}_b^\infty(\mathbf{R}^d), t \in \mathbf{R}_+,$$
$$\langle f, \bar{\mu}_t \rangle = \langle f, \mu_0 \rangle + \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \bar{\mu}_s \rangle) \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \bar{\mu}_s \rangle \, \pi(\mathrm{d}x, \mathrm{d}y) \, \mathrm{d}s. \tag{2.1.10}$$

Corollary 2.2. *Assume* **A1**-**A5**. *Then, $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-L,\gamma}(\mathbf{R}^d))$. In addition, $\bar{\mu}$ satisfies also* (2.1.10) *for test functions $f \in \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$.*

Proof of Corollary 2.2. Note first that by **A4**, $\mu_0 \in \mathcal{C}^{0,\gamma}(\mathbf{R}^d)' \hookrightarrow \mathcal{H}^{-L,\gamma}(\mathbf{R}^d)$ according to (2.1.6). By (2.1.10), **A2**, and **A3**, it holds for all $f \in \mathcal{C}_c^\infty(\mathbf{R}^d)$ and $0 \leq s \leq t \leq T$,

$$|\langle f, \bar{\mu}_t \rangle - \langle f, \bar{\mu}_s \rangle| \leq C|t - s| \|f\|_{\mathcal{C}^{1,\gamma}} \sup_{u\in[0,T]} |\langle 1 + |\cdot|^\gamma, \bar{\mu}_u \rangle|.$$

Note that $\sup_{u\in[0,T]} |\langle 1 + |\cdot|^\gamma, \bar{\mu}_u \rangle| < +\infty$ since $t \geq 0 \mapsto \langle 1 + |\cdot|^\gamma, \bar{\mu}_u \rangle \in \mathcal{D}(\mathbf{R}_+, \mathbf{R})$ (indeed this follows from the fact that $\bar{\mu} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ and [Vil09, Theorem 6.9]). Thus, using (2.1.6), it holds $\bar{\mu}_t \in \mathcal{H}^{-L,\gamma}(\mathbf{R}^d)$ and $|\langle f, \bar{\mu}_t \rangle - \langle f, \bar{\mu}_s \rangle| \leq C|t - s| \|f\|_{\mathcal{H}^{L,\gamma}}$, proving the first claim in Corollary 2.2. The second claim in Corollary 2.2 is obtained by a density argument and the fact that $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{1,\gamma}(\mathbf{R}^d)$. □

On the proof of Theorem 2.1. Theorem 2.1 is proved in Section 2.2. The proof strategy is the following. We first derive an identity satisfied by $(\mu^N)_{N\geq 1}$, namely the pre-limit equation (2.2.8). This is done in Section 2.2.1. Then, we show in Section 2.2.2 that $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. To this end we use [Jak86, Theorem 4.6]. The compact containment of $(\mu^N)_{N\geq 1}$ relies on a characterization of the compact subsets of $\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1})$ (see Proposition 3.15) and moment estimates on $\{\theta_k^i, i \in \{1, \dots N\}\}_{k=0,\dots,\lfloor NT \rfloor}$ (see Lemma 2.11). We then use the pre-limit equation (2.2.8) to prove that any limit point of the sequence $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ satisfies (2.1.10). This requires to study the continuity property of the involved operator (namely $\Lambda_t[f]$, see Lemma 2.20). This the purpose of Section 2.2.2, and more precisely of Proposition 2.21 there. With rough estimates on the jumps of the function $t \in \mathbf{R}_+ \mapsto \langle f, \mu_t^N \rangle$ (where f is uniformly Lipschitz over $\mathbf{R}^{d+1}$), we also prove in Section 2.2.2 that any limit point $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ belongs a.s. to $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. This is indeed needed since we then prove in Section 2.2.3 that (2.1.10) admits a unique solution in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. To prove that there is at most one solution to (2.1.10), we use arguments of [PRT15] which are based on a representation formula for solution to measure-valued equations [Vil03, Theorem 5.34] together with time estimates in Wasserstein distances between two solutions of (2.1.10) derived in [PR16].

Remark 2.3. *In view of their proofs, Theorem 2.1 and Corollary 2.2 are still valid for $\gamma > \frac{d}{2}$ and $L > d/2 + 1$.*

Remark 2.4. *When $\beta = 1/2$, one can obtain a similar limit equation for $\bar{\mu}$, with an additionnal (regularizing) Laplacian term in the limit equation. To derive it, one should consider a Taylor expansion up to order 3 of the test function in the pre-limit equation* (2.2.8)*. Let us mention that the case $\beta = 1/2$ is studied in [MMN18] but only at fixed t. Straightforward application of our method would lead to a trajectorial version of [MMN18, Theorem 3] which we leave to the reader for the sake of brevity.*

Remark 2.5. *Of course, one important question is the convergence of $\bar{\mu}_t$ in long time. It is not hard to see that the loss function decays (but not strictly a priori) along the training, i.e. with t. This asymptotic behavior of $\bar{\mu}_t$ as $t \to +\infty$ has been studied in [MMN18, Theorem 7] or [CB18a] who give partial results in the case without noise. Roughly speaking, they prove that if it is known that $\bar{\mu}_t$ is converging in Wasserstein distance then it converges to the minimum of the loss function. It is however quite hard to prove such a convergence. We refer also to [E20, MWW+20] for what remains to do in this direction which is clearly a difficult open problem. In the case with noise $\beta = 1/2$ then the situation is different as the limit PDE is a usual McKean-Vlasov diffusion and one can study the free energy and study convergence in long time [MMN18, Theorem 4].*

Central limit theorem for the empirical measure

Fluctuation process and extra assumptions. Assume **A1-A5**. The fluctuation process is the process $\eta^N = \{t \mapsto \eta_t^N, t \in \mathbf{R}_+\}$ defined by:

$$\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t),\ N \geq 1,\ t \in \mathbf{R}_+, \tag{2.1.11}$$

where $\bar{\mu} = \{t \mapsto \bar{\mu}_t, t \in \mathbf{R}_+\}$ is the limit of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ (see Theorem 2.1). Let us introduce the following additional assumptions:

A6. The distribution $\mu_0 \in \mathcal{P}(\mathbf{R}^d)$ is compactly supported.

A7. $|B_k| \to |B_\infty|$ a.s. as $k \to \infty$.

Let

$$J_0 \geq 4\lceil \frac{d}{2} \rceil + 8 \text{ and } j_0 = \lceil \frac{d}{2} \rceil + 2. \tag{2.1.12}$$

For later purpose, we also set

$$J_1 = 2\lceil \frac{d}{2} \rceil + 4,\ j_1 = 3\lceil \frac{d}{2} \rceil + 4,\ J_2 = 3\lceil \frac{d}{2} \rceil + 6, \text{ and } j_2 = 2\lceil \frac{d}{2} \rceil + 3. \tag{2.1.13}$$

By (2.1.3), we have the following embeddings:

$$\mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{J_2,j_2}(\mathbf{R}^d),\ \mathcal{H}^{J_2,j_2}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{J_1+1,j_1}(\mathbf{R}^d),\ \mathcal{H}^{J_1,j_1}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{L,\gamma}(\mathbf{R}^d). \tag{2.1.14}$$

G-process and the limit equation.

Definition 2.6. *We say that $\mathscr{G} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$ is a* G-process *if for all $k \geq 1$ and $f_1 \ldots, f_k \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$, $\{t \mapsto (\langle f_1, \mathscr{G}_t\rangle, \ldots, \langle f_k, \mathscr{G}_t\rangle)^T, t \in \mathbf{R}_+\} \in \mathcal{C}(\mathbf{R}_+, \mathbf{R}^k)$ is a process with zero-mean, independent Gaussian increments (and thus a martingale), and with covariance structure given by: for all $1 \leq i, j \leq k$ and all $0 \leq s \leq t$,*

$$\text{Cov}\left(\langle f_i, \mathscr{G}_t\rangle, \langle f_j, \mathscr{G}_s\rangle\right) = \alpha^2 \mathbf{E}\left[\frac{1}{|B_\infty|}\right] \int_0^s \text{Cov}(\mathrm{Q}_v[f_i](x,y), \mathrm{Q}_v[f_j](x,y))\, \mathrm{d}v, \tag{2.1.15}$$

where $\mathrm{Q}_v[f](x,y) := (y - \langle \sigma_*(\cdot, x), \bar{\mu}_v\rangle)\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \bar{\mu}_v\rangle$ for $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $\bar{\mu}$ is given by Theorem 2.1.

Let us make some comments about Definition 2.6. The first one is that we have decided to call such a process G-process to ease the statement of the results. In addition, notice that $\mathrm{Q}_s[f](x,y)$ is well defined for $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d))$ (indeed for all $k \in \{1,\dots,d\}$, $\partial_{e_k} f \in \mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d) \hookrightarrow \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$) and $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \in \mathcal{H}^{-L,\gamma}(\mathbf{R}^d)))$. Finally, we mention that by Proposition 2.37 below, the law of a process $\mu \in \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J,b}(\mathbf{R}^d))$ is fully determined by the family of laws of the processes $(\langle f_1,\mu\rangle,\dots,\langle f_k,\mu\rangle)^T \in \mathcal{D}(\mathbf{R}_+,\mathbf{R})^k$, $k \geq 1$ and where $\{f_a\}_{a\geq 1}$ is an orthonormal basis $\mathcal{H}^{J,b}(\mathbf{R}^d)$.

For η a $\mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$-valued process and $\mathscr{G} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$ a G-process (see Definition 2.6), define the following equation:

$$\begin{aligned}
&\text{A.s. } \forall f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d), \forall t \in \mathbf{R}_+,\\
&\langle f,\eta_t\rangle - \langle f,\eta_0\rangle = \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot,x),\bar{\mu}_s\rangle)\langle \nabla f\cdot\nabla\sigma_*(\cdot,x),\eta_s\rangle \pi(\mathrm{d}x,\mathrm{d}y)\\
&\qquad\qquad - \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\langle \sigma_*(\cdot,x),\eta_s\rangle\langle \nabla f\cdot\nabla\sigma_*(\cdot,x),\bar{\mu}_s\rangle \pi(\mathrm{d}x,\mathrm{d}y) + \langle f,\mathscr{G}_t\rangle. \qquad (2.1.16)
\end{aligned}$$

Definition 2.7. *Let ν be a $\mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$-valued random variable. We say that a $\mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$-valued process η on a probability space is a* weak solution *of* (2.1.16) *with initial distribution ν if there exist a* G*-process $\mathscr{G} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$ such that* (2.1.16) *holds and $\eta_0 = \nu$ in distribution. In addition, we say that* weak uniqueness *holds if for any weak two solutions η^1 and η^2 of* (2.1.16) *(possibly defined on two different probability spaces) with the same initial distributions, it holds $\eta_1 = \eta_2$ in distribution.*

The second main result of this work is a central limit theorem for the trajectory of the scaled empirical measures.

Theorem 2.8. *Let $\beta > 3/4$. Assume* **A1**-**A7**. *Then:*

1. *(Convergence) The sequence $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$ (see* (2.1.11)*) converges in distribution to a process $\eta^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$.*
2. *(Limit equation) The process η^* has the same distribution as the unique weak solution $\eta^\star$ of* (2.1.16) *with initial distribution ν_0 (see Definition 2.7), where ν_0 is the unique $\mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$-valued random variable such that for all $k \geq 1$ and $f_1 \dots, f_k \in \mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d)$,*

$$(\langle f_1,\eta_0^*\rangle,\dots,\langle f_k,\eta_0^*\rangle)^T \sim \mathcal{N}(0,\Gamma(f_1,\dots,f_k)),$$

where $\Gamma(f_1,\dots,f_k)$ is the covariance matrix of the vector $(f_1(W_0^1),\dots,f_k(W_0^1))^T$.

Remark 2.9. *By looking at the definition of the* G*-process and in particular its covariance* (2.1.15)*, one remarks the effect of mini-batching by the $|B_\infty|^{-1}$ prefactor, thus leading to a reduced variance of the* G*-process. Note that this is quite intricate to deduce proper information on the variance of the fluctuation process η, since the terms appearing in* (2.1.16) *are a priori dependent. Nonetheless, it will be shown through the numerical experiments of the next subsection that the variance of fluctuation process reduces when the size of the mini-batches increases (see in particular Figure 2.1).*

Theorem 2.8 is proved in Section 2.3, following inspiration from the previous works [FM97a, JM98a, DLR19a]. The starting point to prove Theorem 2.8, consists in proving, like in the current literature [SS20a], that $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$ is relatively compact (see Propositions 2.27). We then prove that the whole sequence $(\eta^N)_{N\geq 1}$ converges in distribution to the unique weak solution of (2.1.16) in Section 2.3.5.

When $\beta = 3/4$, $(\eta^N)_{N\geq 1}$ is still relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$ (see Proposition 2.27) but the derivation of the limit equation satisfied by its limit points is more tricky. However, in a specific case (when $d = 1$ and the test function is $\mathsf{f}_2 : x \in \mathbf{R} \mapsto |x|^2$), Proposition 2.10 below suggests how the equation (2.1.16) might be perturbed, as shown numerically in Figure 2.2 and more precisely in the inset.

Proposition 2.10. *Let $\beta = 3/4$ and assume that conditions* **A1**-**A7** *hold. Let η be a limit point of $(\eta^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}))$ (see Proposition 2.27). Then, $\eta_0 = \nu_0$ in distribution (see Lemma 2.38), and there exist a $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$-valued process η^* and a* G*-process $\mathscr{G}^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}))$ such that $\eta = \eta_*$ in distribution, and a.s. for every $t \in \mathbf{R}_+$,*

$$\langle \mathsf{f}_2, \eta_t^* \rangle - \langle \mathsf{f}_2, \eta_0^* \rangle = \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \overline{\mu}_s \rangle) \langle \nabla \mathsf{f}_2 \cdot \nabla \sigma_*(\cdot, x), \eta_s^* \rangle \pi(\mathrm{d}x, \mathrm{d}y)$$
$$- \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \langle \sigma_*(\cdot, x), \eta_s^* \rangle \langle \nabla \mathsf{f}_2 \cdot \nabla \sigma_*(\cdot, x), \overline{\mu}_s \rangle \pi(\mathrm{d}x, \mathrm{d}y) + \langle \mathsf{f}_2, \mathscr{G}_t^* \rangle + t\mathbf{E}[\mathsf{f}_2(\varepsilon_1^1)]. \tag{2.1.17}$$

2.1.3 Numerical Experiments

We now illustrate numerically the results derived in the previous sections. First, we consider a regression task on simulated data, based upon an example of [MMN18]. More precisely, we consider (2.1.1) with $\sigma_*(W^i, x) = f(W^i \cdot x)$ where

$$f(t) = \begin{cases} -2.5 \text{ if } t \leq 0.5, \\ 10t - 7.5 \text{ if } 0.5 \leq t \leq 1.5, \\ 7.5 \text{ if } t \geq 1.5. \end{cases}$$

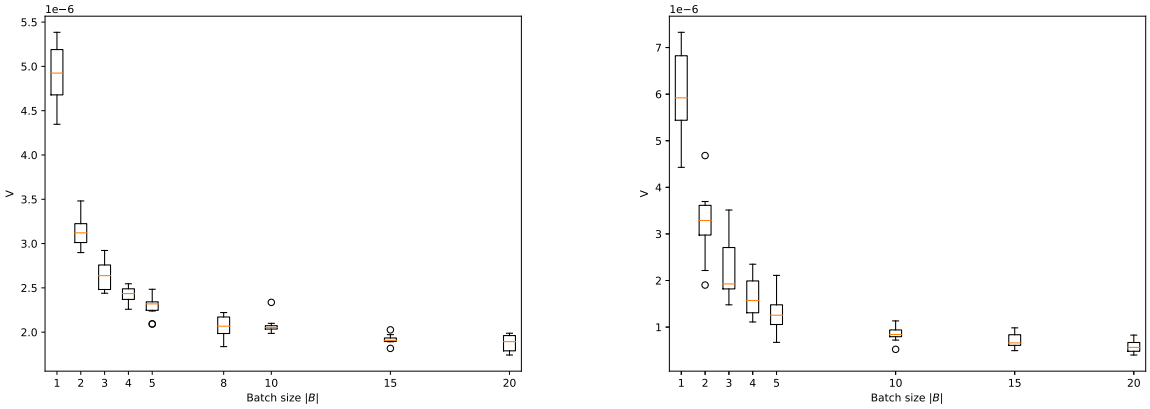
The distribution π of the data is defined as follows: with probability $1/2$, $y = 1$ and $x \sim \mathcal{N}(0, (1 + 0.2)^2 I_d)$ and, with probability $1/2$, $y = -1$ and $x \sim \mathcal{N}(0, (1 - 0.2)^2 I_d)$. This setting satisfies the assumptions of Theorems 2.1 and 2.8, except **A2**, due to the fact that f is not differentiable at $t = 0.5$ and $t = 1.5$ (a smooth modification of f around those points would tackle this problem and would not change the numerical results).

Then, we consider a typical classification task on the MNIST dataset. The neural network we consider is fully connected with one-hidden layer of N neurons and ReLU activation function[1]. The last layer is a softmax layer (we consider one-hot encoding and use Keras and Tensorflow librairies). Given a data $x \in \mathbf{R}^d$ ($d = 784$ here), the neural network returns $\hat{y} = \mathrm{softmax}((W^{\mathbf{o},c} \cdot W^{\mathbf{h}}(x))_{c=0}^9)$ where $W^{\mathbf{h}}(x) = ((W^{\mathbf{h},i} \cdot x)_+)_{i=1}^N$ is the hidden layer ($W^{\mathbf{h},i} \in \mathbf{R}^d$ is the weight of the i-th neuron) and $W^{\mathbf{o},c} \in \mathbf{R}^N$ is the weight of the output layer corresponding to class c. The total number of trainable parameters is thus $dN + 10N$. The neural network is trained with respect to the categorical cross-entropy loss. This case is not covered by our mathematical analysis and the motivation here is to show numerical evidence that the variance reduction derived in Theorem 2.8 is still valid in this case.

Variance Reduction with increasing mini-batch size. We illustrate here that the variance of the limiting fluctuation process decreases with the mini-batch size, even though we only have a mathematical structure of the variance of the G-process (see (2.1.15) together with Remark 2.9). On both experiments, we consider a fixed mini-batch size during the training (i.e. $|B_k| = |B|$ for all $k \in \mathbf{N}$). We first consider the regression task. Consider $\mathsf{L} = 1000$ neural networks (initialized and trained independently) whose $N = 800$ initial neurons are drawn independently according to $\mu_0 = \mathcal{N}(0, \frac{0.8^2}{d} I_d)$. For each neural network, we run $k = 1000$ iterations of the

[1] ReLU function $(\cdot)_+$: $u \in \mathbf{R} \mapsto 0$ if $u < 0$, u if $u \geq 0$.

Figure 2.1: Variance V reduction of the fluctuation process with increasing mini-batch size. **Left:** Regression task on simulated data. V is an empirical estimation from 1000 realisations of the variance of $\langle \|\cdot\|_2, \mu_t^N \rangle$, where $N = 800$ and $t = 1.25$. The other parameters are $d = 40$, $\alpha = 0.1$, $\beta = 1$, and the noise is $\varepsilon_k^i \sim \mathcal{N}(0, 0.01 I_d)$. The boxplots are obtained with 10 samples of V. **Right:** Classification task on MNIST dataset. V is an empirical estimation from 30 realisations of the variance of $\frac{1}{N}\sum_{j=1}^N W_k^{\mathbf{o},0,j}$, where $N = 10000$ and $k = 3000$ ($t = 0.3$). The boxplots are obtained with 10 samples of V.



SGD algorithm (2.1.2) and compute $\mathsf{m}_\ell := \langle \|\cdot\|_2, \mu_t^N \rangle = \frac{1}{N}\sum_{i=1}^N \|W_k^i\|_2$, where $\ell \in \{1, \dots, \mathsf{L}\}$, $t = k/N = 1.25$ and $\|w\|_2 := \sqrt{\sum_{j=1}^d w_j^2}$. Finally, we compute the empirical variance of this quantity, i.e.,

$$\mathsf{V} := \widehat{\mathrm{Var}}(\mathsf{m}_1, \dots, \mathsf{m}_\mathsf{L}) = \frac{1}{\mathsf{L}} \sum_{\ell=1}^{\mathsf{L}} \Big(\mathsf{m}_\ell - \frac{1}{\mathsf{L}} \sum_{\ell'=1}^{\mathsf{L}} \mathsf{m}_{\ell'} \Big)^2 .$$

and display for different mini-batch sizes $|B|$ in Figure 2.1 the obtained boxplots from 10 samples of V. The other parameters are $d = 40$, $\alpha = 0.1$, $\beta = 1$, and the noise is $\varepsilon_k^i \sim \mathcal{N}(0, 0.01 I_d)$.

Second, we turn to the classification task. Consider $\mathsf{L} = 30$ neural networks (initialized and trained independently) with $N = 10000$ neurons on the hidden-layer, until iteration $k = 3000$ of the SGD algorithm ($t = k/N = 0.3$), and compute the mean of the weight of the output layer corresponding to class 0, *i.e.*, for each $\ell = 1, \dots, \mathsf{L}$, we compute $\mathsf{m}_\ell := \frac{1}{N}\sum_{j=1}^N W_k^{\mathbf{o},0,j}$. Finally, we compute the empirical variance of this quantity, i.e., $\mathsf{V} = \widehat{\mathrm{Var}}(\mathsf{m}_1, \dots, \mathsf{m}_\mathsf{L})$ and exhibit for different sizes $|B|$ the boxplots obtained with 10 samples of V in Figure 2.1.

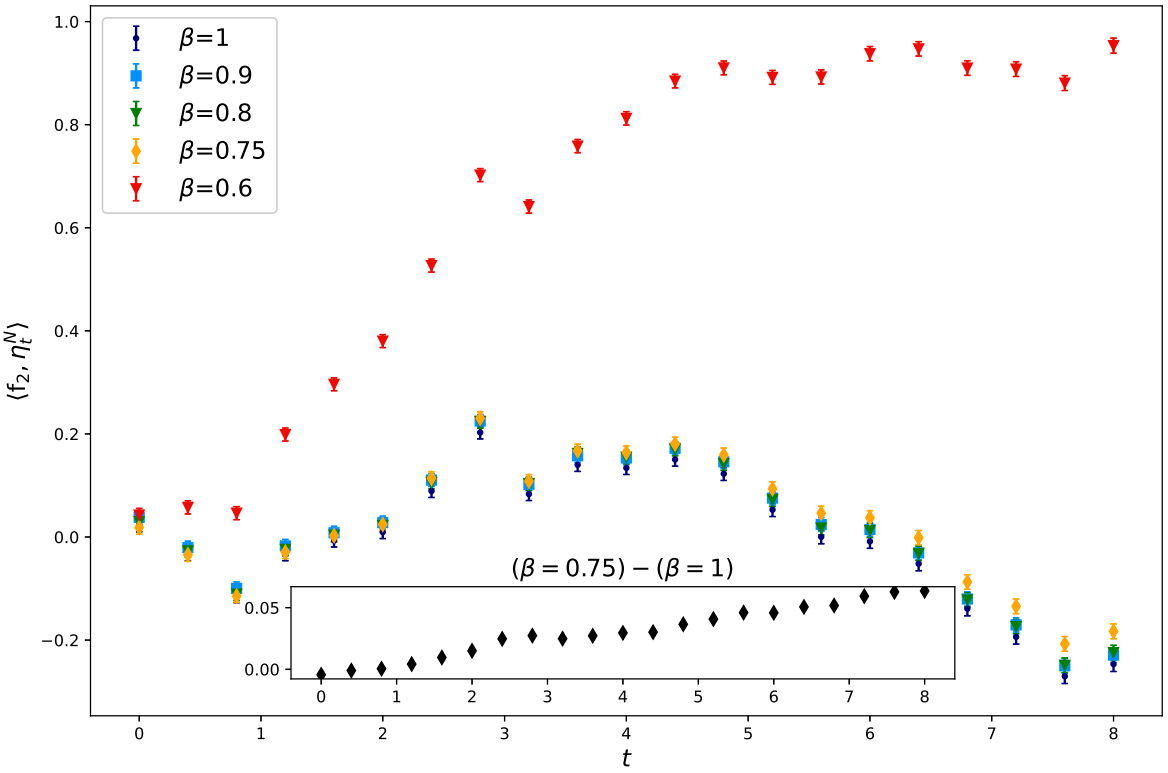
Central Limit Theorem. We focus here on the regression task. For different values of β, we plot in Figure 2.2 $\langle \mathsf{f}_2, \eta_t^N \rangle$ for $0 \le t \le 8$ (recall $\mathsf{f}_2(x) = |x|^2$), to show the agreement of $\langle \mathsf{f}_2, \eta_t^N \rangle$ for different values of $\beta > 3/4$, corresponding to the regime of (2.1.16), and the divergence from it when $\beta \le 3/4$. For $\beta = 3/4$, we also illustrate the regime derived in Proposition 2.10. The parameters chosen are $d = |B| = 1$, $N = 20000$, $\alpha = 0.1$ and $\varepsilon_k^i \sim \mathcal{N}(0, 0.01)$. The procedure to obtain the plots is as follows. We first compute $\langle \mathsf{f}_2, \mu_t^N \rangle$ (we repeat this procedure 20000 times to get confidence intervals). Then, we approximate $\langle \mathsf{f}_2, \bar{\mu}_t \rangle$ by $\langle \mathsf{f}_2, \mu_t^{N'} \rangle$ where $N' = 250000$. On Figure 2.2, we plot $\sqrt{N}(\langle \mathsf{f}_2, \mu_t^N \rangle - \langle \mathsf{f}_2, \mu_t^{N'} \rangle) \simeq \langle \mathsf{f}_2, \eta_t^N \rangle$ as a function of t.

2.2 Proof of Theorem 2.1

2.2.1 Pre-limit equation and remainder terms

In this section, we derive the so-called pre-limit equation (2.2.8). We then show that the remainder terms in this equation are negligible as $N \to +\infty$.

Figure 2.2: Time evolution of the fluctuation process for different values of β on the regression task, with $\mathrm{f}_2 : x \in \mathbf{R} \mapsto |x|^2$, $N = 20000$, $d = |B| = 1$, $\alpha = 0.1$ and $\varepsilon_k^i \sim \mathcal{N}(0, 0.01)$. Confidence intervals are obtained from 20000 realisations. The case $\beta > 3/4$ is driven by (2.1.16). The case $\beta = 3/4$ is driven by (2.1.17). The case $\beta < 3/4$ is not covered by our analysis. The inset exhibits the linear term in time appearing in (2.1.17).



Pre-limit equation

In this section, we introduce several (random) operators acting on $\mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$. Using **A2** and **A3**, it is easy to check that all these operators belong a.s. to the dual of $\mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$. The duality bracket we use in this section then is the one for the duality in $\mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$. Let us consider $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$. The Taylor-Lagrange formula yields, for $N \geq 1$ and $k \in \mathbf{N}$,

$$\begin{aligned}\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle &= \frac{1}{N}\sum_{i=1}^N f(W_{k+1}^i) - f(W_k^i) \\ &= \frac{1}{N}\sum_{i=1}^N \nabla f(W_k^i)\cdot(W_{k+1}^i - W_k^i) + \frac{1}{2N}\sum_{i=1}^N (W_{k+1}^i - W_k^i)^T \nabla^2 f(\widehat{W}_k^i)(W_{k+1}^i - W_k^i),\end{aligned}$$

where, for all $i \in \{1,\dots,N\}$, $\widehat{W}_k^i \in (W_k^i, W_{k+1}^i)$. Using (2.1.2), we have

$$\begin{aligned}\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle &= \frac{1}{N}\sum_{i=1}^N \nabla f(W_k^i)\cdot\left[\frac{\alpha}{N|B_k|}\sum_{(x,y)\in B_k}(y - g_{W^k}^N(x))\nabla_W\sigma_*(W_k^i,x) + \frac{\varepsilon_k^i}{N^\beta}\right] + \langle f, R_k^N\rangle \\ &= \frac{\alpha}{N|B_k|}\sum_{(x,y)\in B_k}(y - \langle\sigma_*(\cdot,x),\nu_k^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\nu_k^N\rangle \\ &\quad + \frac{1}{N^{1+\beta}}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i + \langle f, R_k^N\rangle, \qquad (2.2.1)\end{aligned}$$

where, for $N \geq 1$, $k \in \mathbf{N}$ and $i = 1,\dots,N$,

$$\langle f, R_k^N\rangle := \frac{1}{2N}\sum_{i=1}^N (W_{k+1}^i - W_k^i)^T\nabla^2 f(\widehat{W}_k^i)(W_{k+1}^i - W_k^i). \qquad (2.2.2)$$

For $k \in \mathbf{N}$, we define:

$$\langle f, D_k^N\rangle := \frac{\alpha}{N}\int_{\mathcal{X}\times\mathcal{Y}}(y - \langle\sigma_*(\cdot,x),\nu_k^N\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\nu_k^N\rangle\pi(\mathrm{d}x,\mathrm{d}y), \qquad (2.2.3)$$

$$\langle f, M_k^N\rangle := \frac{\alpha}{N|B_k|}\sum_{(x,y)\in B_k}(y - \langle\sigma_*(\cdot,x),\nu_k^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\nu_k^N\rangle - \langle f, D_k^N\rangle. \qquad (2.2.4)$$

Equation (2.2.1) writes, for $k \in \mathbf{N}$,

$$\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle = \langle f, D_k^N\rangle + \langle f, M_k^N\rangle + \langle f, R_k^N\rangle + \frac{1}{N^{1+\beta}}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i. \qquad (2.2.5)$$

Define for $N \geq 1$ and $t \in \mathbf{R}_+$:

$$\langle f, D_t^N\rangle := \sum_{k=0}^{\lfloor Nt\rfloor - 1}\langle f, D_k^N\rangle \quad\text{and}\quad \langle f, M_t^N\rangle := \sum_{k=0}^{\lfloor Nt\rfloor - 1}\langle f, M_k^N\rangle, \qquad (2.2.6)$$

with the convention that $\sum_0^{-1} = 0$ (which occurs if and only if $0 \leq t < 1/N$). It will be proved later that $\{t \mapsto \langle f, M_t^N\rangle, t \in \mathbf{R}_+\}$ is a martingale (see indeed Lemma 2.24), hence the notation.

One has, for $t \in \mathbf{R}_+$,

$$\begin{aligned}\langle f, D_t^N\rangle &= \sum_{k=0}^{\lfloor Nt\rfloor-1} \int_{\frac{k}{N}}^{\frac{k+1}{N}} \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot,x), \nu_k^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \nu_k^N\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &= \sum_{k=0}^{\lfloor Nt\rfloor-1} \int_{\frac{k}{N}}^{\frac{k+1}{N}} \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot,x), \mu_s^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \mu_s^N\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &= \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot,x), \mu_s^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \mu_s^N\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s + \langle f, V_t^N\rangle,\end{aligned}$$

where $\langle f, V_t^N\rangle$, for $t \in \mathbf{R}_+$:

$$\langle f, V_t^N\rangle := -\int_{\frac{\lfloor Nt\rfloor}{N}}^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot,x), \mu_s^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \mu_s^N\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s. \tag{2.2.7}$$

Therefore, using (2.2.5), we obtain that the scaled empirical measure process $\{t \mapsto \mu_t^N, t \in \mathbf{R}_+\}$ satisfies the following pre-limit equation : for $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$, $N \geq 1$ and $t \in \mathbf{R}_+$,

$$\begin{aligned}\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle &= \sum_{k=0}^{\lfloor Nt\rfloor-1} \langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle \\ &= \langle f, D_t^N\rangle + \langle f, M_t^N\rangle + \sum_{k=0}^{\lfloor Nt\rfloor-1} \langle f, R_k^N\rangle + \frac{1}{N^{1+\beta}} \sum_{k=0}^{\lfloor Nt\rfloor-1} \sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i \\ &= \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot,x), \mu_s^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \mu_s^N\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &\quad + \langle f, M_t^N\rangle + \langle f, V_t^N\rangle + \sum_{k=0}^{\lfloor Nt\rfloor-1} \langle f, R_k^N\rangle + \frac{1}{N^{1+\beta}} \sum_{k=0}^{\lfloor Nt\rfloor-1} \sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i.\end{aligned} \tag{2.2.8}$$

In the next section, we study the four last terms of (2.2.8).

The remainder terms in (2.2.8) are negligible

The aim of this section is to show that the last four terms of (2.2.8) vanish as $N \to +\infty$. This is the purpose of Lemma 2.12. The following result will be used several times in this work.

Lemma 2.11. *Let $\beta \geq 1/2$ and assume* **A1**-**A5**. *Then, for all $T > 0$, there exists a constant $C < +\infty$ such that for all $N \geq 1$, $i \in \{1,\dots,N\}$ and $k \in \{0,\dots,\lfloor NT\rfloor\}$,*

$$\mathbf{E}\left[|W_k^i|^{8\gamma_*}\right] \leq C.$$

Proof. Let us recall the following convexity inequality : for $m, p \geq 1$ and $x_1,\dots,x_p \in \mathbf{R}_+$,

$$\Big(\sum_{l=1}^m x_l\Big)^p \leq m^{p-1}\sum_{l=1}^m x_l^p. \tag{2.2.9}$$

Let $C > 0$ denotes a constant, independent of $i \in \{1,\dots,N\}$ and $0 \leq k \leq \lfloor NT\rfloor$, which can change from one occurence to another. Set $p = 8\gamma_*$. For $i \in \{1,\dots,N\}$ and $1 \leq k \leq \lfloor NT\rfloor$, we have, using (2.1.2) and **A2** :

$$|W_k^i| \leq |W_0^i| + \Big|\sum_{j=0}^{k-1} W_{j+1}^i - W_j^i\Big| \leq |W_0^i| + \frac{C}{N}\sum_{j=0}^{k-1}\frac{1}{|B_j|}\sum_{(x,y)\in B_j}(|y|+C) + \frac{1}{N^\beta}\Big|\sum_{j=0}^{k-1}\varepsilon_j^i\Big|.$$

Thus, by (3.9.3),

$$|W_k^i|^p \le C\Big[|W_0^i|^p + \frac{1}{N}\sum_{j=0}^{k-1}\frac{1}{|B_j|^p}\Big(\sum_{(x,y)\in B_j}(|y|+C)\Big)^p + \frac{1}{N^{p\beta}}\Big|\sum_{j=0}^{k-1}\varepsilon_j^i\Big|^p\Big]$$
$$\le C\Big[|W_0^i|^p + \frac{1}{N}\sum_{j=0}^{k-1}\frac{1}{|B_j|}\sum_{(x,y)\in B_j}(|y|+C)^p + \frac{1}{N^{p\beta}}\Big|\sum_{j=0}^{k-1}\varepsilon_j^i\Big|^p\Big].$$

We have:

$$\mathbf{E}\Big[\frac{1}{|B_j|}\sum_{(x,y)\in B_j}(|y|+C)^p\Big] \le C\Big[\mathbf{E}\Big[\frac{1}{|B_j|}\sum_{n=1}^{|B_j|}|y_j^n|^p\Big]+1\Big],$$

and, using **A1** and (**A3**), it holds for $j \ge 0$:

$$\mathbf{E}\Big[\frac{1}{|B_j|}\sum_{n=1}^{|B_j|}|y_j^n|^p\Big] = \sum_{q=1}^{+\infty}\mathbf{E}\Big[\frac{\mathbf{1}_{|B_j|=q}}{q}\sum_{n=1}^{q}|y_j^n|^p\Big] = \sum_{q=1}^{+\infty}\frac{1}{q}\sum_{n=1}^{q}\mathbf{E}\big[|y_j^n|^p\mathbf{1}_{|B_j|=q}\big]$$
$$= \sum_{q=1}^{+\infty}\frac{1}{q}\sum_{n=1}^{q}\mathbf{E}\big[|y_j^n|^p\big]\,\mathbf{E}\big[\mathbf{1}_{|B_j|=q}\big]$$
$$= \mathbf{E}\big[|y_1^1|^p\big] < +\infty. \tag{2.2.10}$$

Thus, using the two previous inequalities, we deduce that:

$$\mathbf{E}\Big[\frac{1}{N}\sum_{j=0}^{k-1}\frac{1}{|B_j|}\sum_{(x,y)\in B_j}(|y|+C)^p\Big] \le C.$$

By **A4**, $\mathbf{E}\left[|W_0^i|^p\right] \le C$. In addition, we have that, for $i \in \{1,\dots,N\}$,

$$\Big|\sum_{j=0}^{k-1}\varepsilon_j^i\Big|^p \le C\sum_{l=1}^{d}\Big|\sum_{j=0}^{k-1}\varepsilon_j^{i,l}\Big|^p$$

Since we deal with the sum of centered independent Gaussian random variables, we have that, for all $i \in \{1,\dots,N\}$ and $l \in \{1,\dots,d\}$,

$$\mathbf{E}\Big[\Big|\sum_{j=0}^{k-1}\varepsilon_j^{i,l}\Big|^p\Big] \le Ck^{p/2} \le CN^{p/2}.$$

Putting all these inequalities together, we obtain that $\mathbf{E}\left[|W_k^i|^p\right] \le C\left[1+\frac{N^{p/2}}{N^{p\beta}}\right] \le C$ (recall $\beta \ge 1/2$). This concludes the proof of the lemma. □

Lemma 2.12. *Let $\beta \ge 1/2$ and assume* **A1**-**A5**. *Then, for all $T > 0$ there exists $C < \infty$ such that for all $N \ge 1$ and $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$,*

(i) $\max_{0\le k<\lfloor NT\rfloor}\mathbf{E}\left[|\langle f, R_k^N\rangle|\right] \le C\|f\|_{\mathcal{C}^{2,\gamma_*}}\left[\frac{1}{N^2}+\frac{1}{N^{2\beta}}\right]$.

(ii) $\sup_{t\in[0,T]}\mathbf{E}\left[|\langle f, V_t^N\rangle|\right] \le C\|f\|_{\mathcal{C}^{2,\gamma_*}}/N$.

(iii) $\sup_{t\in[0,T]}\mathbf{E}\left[|\langle f, M_t^N\rangle|^2\right] \le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}/N$.

(iv) $\sup_{t\in[0,T]}\mathbf{E}\left[\left|\frac{1}{N^{1+\beta}}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i\right|^2\right] \le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}/N^{2\beta}$.

Proof. Let $T > 0$ and $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$. In what follows, $C > 0$ is a constant, independent of $N \geq 1$, $t \in [0,T]$, f, and $k \in \{0, \dots, \lfloor NT \rfloor - 1\}$, which can change from one line to another.

Proof of item *(i)*. For $k \in \{0, \dots, \lfloor NT \rfloor - 1\}$, by (2.2.2), we have

$$|\langle f, R_k^N \rangle| \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N} \sum_{i=1}^N |W_{k+1}^i - W_k^i|^2 (1 + |\widehat{W}_k^i|^{\gamma_*}). \tag{2.2.11}$$

On the other hand, by (2.1.2), we have:

$$|W_{k+1}^i - W_k^i| \leq \frac{C}{N|B_k|} \sum_{(x,y)\in B_k} \big(|y| + |g_{W_k}^N(x)|\big) + \frac{|\varepsilon_k^i|}{N^\beta}. \tag{2.2.12}$$

By (3.9.3) and the triangle inequality, we deduce

$$|W_{k+1}^i - W_k^i|^2 \leq C\Big[\frac{1}{N^2|B_k|} \sum_{(x,y)\in B_k} (|y|^2 + |g_{W_k}^N(x)|^2) + \frac{|\varepsilon_k^i|^2}{N^{2\beta}}\Big].$$

By definition of $\widehat{W}_k^i$, there exists $\alpha_k^i \in (0,1)$ such that $\widehat{W}_k^i = \alpha_k^i W_k^i + (1-\alpha_k^i) W_{k+1}^i$, leading, by (3.9.3), to $|\widehat{W}_k^i|^{\gamma_*} \leq C\left[|W_k^i|^{\gamma_*} + |W_{k+1}^i|^{\gamma_*}\right]$. Therefore,

$$\begin{aligned}
|W_{k+1}^i - W_k^i|^2 (1 + |\widehat{W}_k^i|^{\gamma_*}) &\leq C\Big[\frac{1}{N^2|B_k|} \sum_{(x,y)\in B_k} (|y|^2 + |g_{W_k}^N(x)|^2) + \frac{|\varepsilon_k^i|^2}{N^{2\beta}}\Big](1 + |W_k^i|^{\gamma_*} + |W_{k+1}^i|^{\gamma_*}) \\
&\leq \frac{C}{N^2}(1 + |W_k^i|^{\gamma_*} + |W_{k+1}^i|^{\gamma_*})^2 + \frac{C}{N^2|B_k|} \sum_{(x,y)\in B_k} (|y|^4 + |g_{W_k}(x)|^4) \\
&\quad + \frac{C}{N^{2\beta}} \left[|\varepsilon_k^i|^4 + (1 + |W_k^i|^{\gamma_*} + |W_{k+1}^i|^{\gamma_*})^2\right].
\end{aligned} \tag{2.2.13}$$

Plugging (2.2.13) in (2.2.11), we obtain

$$\begin{aligned}
|\langle f, R_k^N \rangle| \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N} \sum_{i=1}^N \Big[&\frac{1}{N^2}(1 + |W_k^i|^{\gamma_*} + |W_{k+1}^i|^{\gamma_*})^2 + \frac{1}{N^2|B_k|} \sum_{(x,y)\in B_k} (|y|^4 + C) \\
&+ \frac{1}{N^{2\beta}} \left[|\varepsilon_k^i|^4 + (1 + |W_k^i|^{\gamma_*} + |W_{k+1}^i|^{\gamma_*})^2\right] \Big].
\end{aligned} \tag{2.2.14}$$

Finally, using Lemma 2.11, **A3**, and **A5**, one deduces that $\mathbf{E}\left[|\langle f, R_k^N \rangle|\right] \leq C\|f\|_{\mathcal{C}^{2,\gamma_*}}(1/N^2 + 1/N^{2\beta})$. This proves item *(i)*.

Proof of item *(ii)*. Let $t \in [0,T]$. Since σ_* and all its derivatives are bounded (see **A2**), it holds for all $s \geq 0$:

$$|\langle \sigma_*(\cdot, x), \mu_s^N \rangle| = \Big|\frac{1}{N} \sum_{i=1}^N \sigma_*(W_{\lfloor Ns \rfloor}^i, x)\Big| \leq C, \tag{2.2.15}$$

and

$$|\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \mu_s^N \rangle| = \Big|\frac{1}{N} \sum_{i=1}^N \nabla f(W_{\lfloor Ns \rfloor}^i) \cdot \nabla_W \sigma_*(W_{\lfloor Ns \rfloor}^i, x)\Big| \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N} \sum_{i=1}^N (1 + |W_{\lfloor Ns \rfloor}^i|^{\gamma_*}). \tag{2.2.16}$$

Notice that C above is also independent of $x \in \mathcal{X}$. Since $\mathbf{E}[|y|] < +\infty$ (see (**A3**)), we obtain

$$\Big|\int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \mu_s^N \rangle) \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \mu_s^N \rangle \pi(\mathrm{d}x, \mathrm{d}y)\Big| \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N} \sum_{i=1}^N (1 + |W_{\lfloor Ns \rfloor}^i|^{\gamma_*}). \tag{2.2.17}$$

Noticing that $s \in (\frac{\lfloor Nt \rfloor}{N}, t) \Rightarrow \lfloor Ns \rfloor = \lfloor Nt \rfloor$, we obtain (see (2.2.7))

$$|\langle f, V_t^N \rangle| \leq \left(t - \frac{\lfloor Nt \rfloor}{N} \right) \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N} \sum_{i=1}^{N} (1 + |W^i_{\lfloor Nt \rfloor}|^{\gamma_*}). \tag{2.2.18}$$

Then, by Lemma 2.11, $\mathbf{E}\left[|\langle f, V_t^N \rangle|\right] \leq C\|f\|_{\mathcal{C}^{2,\gamma_*}}/N$. This proves item *(ii)*.

Proof of item *(iii)*. Let $t \in [0,T]$. Recall the definition of $\mathcal{F}_k^N$ in (2.1.9) and $\langle f, M_k^N \rangle$ in (2.2.4).

Step 1. In this step we prove that

$$\mathbf{E}\left[|\langle f, M_k^N \rangle|^2\right] \leq C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}/N^2. \tag{2.2.19}$$

With the same arguments as those used to get (2.2.15) and (2.2.16), we have

$$|\langle \sigma_*(\cdot, x), \nu_k^N \rangle| \leq C \quad \text{and} \quad |\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \nu_k^N \rangle| \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N} \sum_{i=1}^{N} (1 + |W_k^i|^{\gamma_*}). \tag{2.2.20}$$

Note that C above is also independent of $x \in \mathcal{X}$. By (3.9.3) and (2.2.20), we have:

$$\begin{aligned}
|\langle f, M_k^N \rangle|^2 &\leq \frac{C}{N^2|B_k|} \sum_{(x,y)\in B_k} (y - \langle \sigma_*(\cdot, x), \nu_k^N \rangle)^2 \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \nu_k^N \rangle^2 + C|\langle f, D_k^N \rangle|^2 \\
&\leq \frac{C}{N^2|B_k|} \sum_{(x,y)\in B_k} (|y|^2 + C) \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \nu_k^N \rangle^2 + C|\langle f, D_k^N \rangle|^2 \\
&\leq \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^3|B_k|} \sum_{(x,y)\in B_k} (|y|^2 + C) \sum_{i=1}^{N} (1 + |W_k^i|^{\gamma_*})^2 + C|\langle f, D_k^N \rangle|^2.
\end{aligned} \tag{2.2.21}$$

On the other hand, it holds since $(W_k^1, \ldots, W_k^N)$ is $\mathcal{F}_k^N$-measurable and by **A1**:

$$\begin{aligned}
\mathbf{E}\Big[\frac{1}{|B_k|} \sum_{(x,y)\in B_k} (|y|^2 + C) \sum_{i=1}^{N} (1 + |W_k^i|^{\gamma_*})^2\Big] &= \sum_{q\geq 1} \frac{1}{q} \mathbf{E}\Big[\mathbf{1}_{|B_k|=q} \sum_{n=1}^{q} (|y_k^n|^2 + C) \sum_{i=1}^{N} (1 + |W_k^i|^{\gamma_*})^2\Big] \\
&= \sum_{q\geq 1} \frac{1}{q} \mathbf{E}\Big[\mathbf{1}_{|B_k|=q} \sum_{n=1}^{q} (|y_k^n|^2 + C)\Big] \mathbf{E}\Big[\sum_{i=1}^{N} (1 + |W_k^i|^{\gamma_*})^2\Big] \\
&= \sum_{q\geq 1} \frac{1}{q} \mathbf{E}\Big[\mathbf{1}_{|B_k|=q}\Big] \mathbf{E}\Big[\sum_{n=1}^{q} (|y_k^n|^2 + C)\Big] \mathbf{E}\Big[\sum_{i=1}^{N} (1 + |W_k^i|^{\gamma_*})^2\Big] \\
&= \sum_{q\geq 1} \mathbf{E}\Big[\mathbf{1}_{|B_k|=q}\Big] \mathbf{E}\Big[|y_1^1|^2 + C\Big] \mathbf{E}\Big[\sum_{i=1}^{N} (1 + |W_k^i|^{\gamma_*})^2\Big] \\
&= \mathbf{E}\Big[|y_1^1|^2 + C\Big] \mathbf{E}\Big[\sum_{i=1}^{N} (1 + |W_k^i|^{\gamma_*})^2\Big] \leq CN,
\end{aligned} \tag{2.2.22}$$

where we have used Lemma 2.11 and **A3** for the last inequality. Consequently, one has:

$$\mathbf{E}\left[|\langle f, M_k^N \rangle|^2\right] \leq \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^2} + C\mathbf{E}\left[\langle f, D_k^N \rangle^2\right]. \tag{2.2.23}$$

On the other hand, we easily obtain with similar arguments that $\mathbf{E}\left[|\langle f, D_k^N \rangle|^2\right] \leq C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}/N^2$. Together with (2.2.23), this ends the proof of (2.2.19).

Step 2. In this step we prove that for all $k \geq 0$:

$$\mathbf{E}\left[\langle f, M_k^N\rangle | \mathcal{F}_k^N\right] = 0. \tag{2.2.24}$$

For ease of notation, we set

$$\mathrm{Q}^N[f](x, y, \{W_k^i\}_{i=1,\dots,N}) = \big(y - \langle \sigma_*(\cdot, x), \nu_k^N\rangle\big)\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \nu_k^N\rangle. \tag{2.2.25}$$

With this notation, we have (see (2.2.3)) $\langle f, D_k^N\rangle = \frac{\alpha}{N}\int_{\mathcal{X}\times\mathcal{Y}} \mathrm{Q}^N[f](x, y, \{W_k^i\}_{i=1,\dots,N})\pi(\mathrm{d}x, \mathrm{d}y)$ and $\langle f, M_k^N\rangle = \frac{\alpha}{N|B_k|}\sum_{(x,y)\in B_k} \mathrm{Q}^N[f](x, y, \{W_k^i\}_{i=1,\dots,N}) - \langle f, D_k^N\rangle$. It then holds:

$$\begin{aligned}\mathbf{E}\Big[\frac{1}{|B_k|}\sum_{(x,y)\in B_k} \mathrm{Q}^N[f](x, y, \{W_k^i\}_{i=1,\dots,N}) \,\Big|\, \mathcal{F}_k^N\Big] &= \mathbf{E}\Big[\frac{1}{|B_k|}\sum_{n=1}^{|B_k|} \mathrm{Q}^N[f](x_k^n, y_k^n, \{W_k^i\}_{i=1,\dots,N}) \,\Big|\, \mathcal{F}_k^N\Big] \\ &= \sum_{q\geq 1}\frac{1}{q}\mathbf{E}\Big[\mathbf{1}_{|B_k|=q}\sum_{n=1}^{q} \mathrm{Q}^N[f](x_k^n, y_k^n, \{W_k^i\}_{i=1,\dots,N}) \,\Big|\, \mathcal{F}_k^N\Big].\end{aligned}$$

Since $(W_k^1, \dots, W_k^N)$ is $\mathcal{F}_k^N$-measurable, $\big(|B_k|, ((x_k^n, y_k^n))_{n\geq 1}\big) \perp\!\!\!\perp \mathcal{F}_k^N$, and $|B_k| \perp\!\!\!\perp ((x_k^n, y_k^n))_{n\geq 1}$ (see **A1**), we deduce that

$$\begin{aligned}&\mathbf{E}\Big[\mathbf{1}_{|B_k|=q}\sum_{n=1}^{q} \mathrm{Q}^N[f](x_k^n, y_k^n, \{W_k^i\}_{i=1,\dots,N}) \,\Big|\, \mathcal{F}_k^N\Big] \\ &= \mathbf{E}\Big[\mathbf{1}_{|B_k|=q}\sum_{n=1}^{q} \mathrm{Q}^N[f](x_k^n, y_k^n, \{w_k^i\}_{i=1,\dots,N})\Big]\Big|_{\{w_k^i\}_i=\{W_k^i\}_i} \\ &= \mathbf{E}\big[\mathbf{1}_{|B_k|=q}\big]\mathbf{E}\Big[\sum_{n=1}^{q} \mathrm{Q}^N[f](x_k^n, y_k^n, \{w_k^i\}_{i=1,\dots,N})\Big]\Big|_{\{w_k^i\}_i=\{W_k^i\}_i} \\ &= q\mathbf{E}\big[\mathbf{1}_{|B_k|=q}\big]\mathbf{E}\Big[\mathrm{Q}^N[f](x_1^1, y_1^1, \{w_k^i\}_{i=1,\dots,N})\Big]\Big|_{\{w_k^i\}_i=\{W_k^i\}_i} \\ &= q\,\frac{N}{\alpha}\mathbf{E}\big[\mathbf{1}_{|B_k|=q}\big]\,\langle f, D_k^N\rangle,\end{aligned}$$

where we have used **A3** to deduce the last two equalities. We have thus proved that

$$\mathbf{E}\Big[\frac{\alpha}{N|B_k|}\sum_{(x,y)\in B_k} \mathrm{Q}^N[f](x_k^n, y_k^n, \{W_k^i\}_{i=1,\dots,N}) \,\Big|\, \mathcal{F}_k^N\Big] = \langle f, D_k^N\rangle.$$

Therefore, using in addition that $\mathbf{E}[\langle f, D_k^N\rangle | \mathcal{F}_k^N] = \langle f, D_k^N\rangle$ (because $(W_k^1, \dots, W_k^N)$ is $\mathcal{F}_k^N$-measurable), we finally deduce (2.2.24).

Step 3. We now end the proof of item *(iii)*. If $j > k$, $\langle f, M_k^N\rangle$ is $\mathcal{F}_j^N$-measurable (because $\langle f, M_k^N\rangle$ is $\mathcal{F}_{k+1}^N$-measurable). Then, using also (2.2.24), one obtains that for $j > k$:

$$\mathbf{E}\left[\langle f, M_k^N\rangle\langle f, M_j^N\rangle\right] = \mathbf{E}\left[\langle f, M_k^N\rangle\mathbf{E}\left[\langle f, M_j^N\rangle | \mathcal{F}_j^N\right]\right] = \mathbf{E}\left[\langle f, M_k^N\rangle \times 0\right] = 0. \tag{2.2.26}$$

We then have (see (2.2.6)):

$$\mathbf{E}\left[|\langle f, M_t^N\rangle|^2\right] = \sum_{k=0}^{\lfloor Nt\rfloor - 1} \mathbf{E}\left[|\langle f, M_k^N\rangle|^2\right]. \tag{2.2.27}$$

Plugging (2.2.19) in (2.2.27) implies item *(iii)*.

Proof of item *(iv)*. Let $t \in [0,T]$. By Lemma 2.11, $\nabla f(W_k^i)\cdot\varepsilon_k^i$ is square-integrable for all $k \in \mathbf{N}$ and $i \in \{1,\dots,N\}$. From the equality

$$\mathbf{E}\Big[\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big] = \sum_{j,k=0}^{\lfloor Nt\rfloor-1}\sum_{i,\ell=1}^{N}\mathbf{E}\big[\nabla f(W_k^i)\cdot\varepsilon_k^i\,\nabla f(W_j^\ell)\cdot\varepsilon_j^\ell\big].$$

Recall that W_a^b is $\mathcal{F}_a^N$-measurable for all $a \in \mathbf{N}$ and $b \in \{1,\dots,N\}$, and that $\varepsilon_a^b \perp\!\!\!\perp \mathcal{F}_a^N$ (see **A5**). Let e_q denotes the q-th element of the canonical basis of $\mathbf{R}^d$ ($q \in \{1,\dots,d\}$). Assume that $0 \le j < k \le \lfloor Nt\rfloor - 1$. Then, ε_j^ℓ is $\mathcal{F}_k^N$-measurable, and it holds for all $i,\ell \in \{1,\dots,N\}$:

$$\mathbf{E}\big[\nabla f(W_k^i)\cdot\varepsilon_k^i\,\nabla f(W_j^\ell)\cdot\varepsilon_j^\ell\big] = \sum_{n,m=1}^{d}\mathbf{E}\Big[\partial_{e_n}f(W_k^i)\,\partial_{e_m}f(W_j^\ell)\,\varepsilon_j^\ell\cdot e_m\Big]\mathbf{E}\big[\varepsilon_k^i\cdot e_n\big] = 0. \tag{2.2.28}$$

because $\varepsilon_k^i \sim \mathcal{N}(0,I_d)$ (see **A5**). On the other hand, using **A5**, we have for all $0 \le k \le \lfloor Nt\rfloor - 1$ and when $i \neq \ell \in \{1,\dots,N\}$:

$$\mathbf{E}\big[\nabla f(W_k^i)\cdot\varepsilon_k^i\,\nabla f(W_k^\ell)\cdot\varepsilon_k^\ell\big] = \sum_{n,m=1}^{d}\mathbf{E}\Big[\,\partial_{e_n}f(W_k^i)\,\partial_{e_m}f(W_k^\ell)\,\Big]\mathbf{E}\big[\varepsilon_k^i\cdot e_n\big]\mathbf{E}\big[\varepsilon_k^\ell\cdot e_m\big] = 0. \tag{2.2.29}$$

Consequently, we have:

$$\mathbf{E}\Big[\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big] = \sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\mathbf{E}\big[|\nabla f(W_k^i)\cdot\varepsilon_k^i|^2\big].$$

Using the Cauchy-Schwarz inequality, we deduce, using also Lemma 2.11 and **A5**, that:

$$\begin{aligned}\mathbf{E}\Big[\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big] &\le \sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\mathbf{E}\left[|\nabla f(W_k^i)|^2\right]\mathbf{E}\left[|\varepsilon_k^i|^2\right]\\ &\le C\|f\|_{\mathcal{C}^{2,\gamma_*}}^2\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\mathbf{E}\left[(1+|W_k^i|^{\gamma_*})^2\right] \le C\|f\|_{\mathcal{C}^{2,\gamma_*}}^2 N^2. \end{aligned}\tag{2.2.30}$$

This proves *(iv)*. The proof of Lemma 2.12 is complete. □

We now want to pass to the limit in (2.2.8). To this end, we first prove that $(\mu^N)_{N\ge1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+,\mathcal{H}^{-L,\gamma}(\mathbf{R}^d))$. This is the purpose of the following section.

2.2.2 Relative compactness in $\mathcal{D}(\mathbf{R}_+,\mathcal{P}_\gamma(\mathbf{R}^d))$ and convergence to the limit equation

In this section, we show that $(\mu^N)_{N\ge1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+,\mathcal{P}_\gamma(\mathbf{R}^d))$. Then, we prove that any limit point of $(\mu^N)_{N\ge1}$ satisfies a.s. (2.1.10).

Relative compactness in $\mathcal{D}(\mathbf{R}_+,\mathcal{P}_\gamma(\mathbf{R}^d))$

In this section we prove the following result.

Proposition 2.13. *Let $\beta \ge 1/2$ and assume that the conditions* **A1**-**A5** *hold. Then, $(\mu^N)_{N\ge1}$ is relatively compact in in $\mathcal{D}(\mathbf{R}_+,\mathcal{P}_\gamma(\mathbf{R}^d))$.*

We first recall the following standard result.

Proposition 2.14. *Let $q > p \geq 1$ and $C > 0$. The set $\mathscr{K}_C^q := \{\mu \in \mathcal{P}_p(\mathbf{R}^d), \int_{\mathbf{R}^d} |x|^q \mu(\mathrm{d}x) \leq C\}$ is compact.*

We have the following result.

Lemma 2.15. *Let $\beta \geq 1/2$ and assume that* **A1**-**A5** *hold. Then, for every $T > 0$, there exists $C > 0$ such that for all $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$,*

$$\sup_{N \geq 1} \mathbf{E}\Big[\sup_{t \in [0,T]} \langle f, \mu_t^N \rangle^2 \Big] \leq C \|f\|^2_{\mathcal{C}^{2,\gamma_*}}. \tag{2.2.31}$$

Proof. Let $T > 0$ and $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$. All along the proof, $C < \infty$ denotes a constant independent of $t \in [0,T]$, $N \geq 1$, $k \in \{0, \ldots, \lfloor Nt \rfloor\}$, $i \in \{1, \ldots, N\}$, and $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$, which can change from one occurence to another. From (2.2.8), we have:

$$\begin{aligned}
\sup_{t \in [0,T]} \langle f, \mu_t^N \rangle^2 \leq C \Big[\langle f, \mu_0^N \rangle^2 + \int_0^T \int_{\mathcal{X} \times \mathcal{Y}} (y - \langle \sigma_*(\cdot, x), \mu_s^N \rangle)^2 \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \mu_s^N \rangle^2 \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \\
+ \sup_{t \in [0,T]} \langle f, M_t^N \rangle^2 + \sup_{t \in [0,T]} |\langle f, V_t^N \rangle|^2 + \sup_{t \in [0,T]} \Big| \sum_{k=0}^{\lfloor Nt \rfloor - 1} \langle f, R_k^N \rangle \Big|^2 \\
+ \frac{1}{N^{2+2\beta}} \sup_{t \in [0,T]} \Big| \sum_{k=0}^{\lfloor Nt \rfloor - 1} \sum_{i=1}^N \nabla f(W_k^i) \cdot \varepsilon_k^i \Big|^2 \Big].
\end{aligned} \tag{2.2.32}$$

We now study each term of the right-hand side of (2.2.32). Let us deal with the first term in the right-hand side of (2.2.32). Using **A4** and (3.9.3), it holds:

$$\begin{aligned}
\mathbf{E}\left[\langle f, \mu_0^N \rangle^2\right] = \mathbf{E}\left[\langle f, \nu_0^N \rangle^2\right] = \mathbf{E}\Big[\Big| \frac{1}{N} \sum_{i=1}^N f(W_0^i) \Big|^2\Big] \leq \frac{1}{N} \sum_{i=1}^N \mathbf{E}\left[|f(W_0^i)|^2\right] \leq \frac{\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N} \sum_{i=1}^N \mathbf{E}\left[(1 + |W_0^i|^{\gamma_*})^2\right] \\
\leq C \|f\|^2_{\mathcal{C}^{2,\gamma_*}}.
\end{aligned}$$

For the second term in the right-hand side of (2.2.32), we have since $\mathbf{E}[|y|^2] < +\infty$ (see (**A3**)) and using (2.2.15), (2.2.16), and Lemma 2.11:

$$\mathbf{E}\left[\int_0^T \int_{\mathcal{X} \times \mathcal{Y}} (y - \langle \sigma_*(\cdot, x), \mu_s^N \rangle)^2 \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \mu_s^N \rangle^2 \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \right] \leq C \|f\|^2_{\mathcal{C}^{2,\gamma_*}}.$$

Let us deal with the third term in the right-hand side of (2.2.32). By (2.2.6) and (3.9.3), we have, for $t \in [0,T]$,

$$\sup_{t \in [0,T]} |\langle f, M_t^N \rangle|^2 \leq \lfloor NT \rfloor \sum_{k=0}^{\lfloor NT \rfloor - 1} \langle f, M_k^N \rangle^2.$$

Hence, using (2.2.19), we obtain that $\mathbf{E}\big[\sup_{t \in [0,T]} \langle f, M_t^N \rangle^2\big] \leq C \|f\|^2_{\mathcal{C}^{2,\gamma_*}}$. Let us deal with the fourth term in the right-hand side of (2.2.32). From (3.9.3) and (2.2.18),

$$\sup_{t \in [0,T]} |\langle f, V_t^N \rangle|^2 \leq \frac{C \|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^3} \sum_{i=1}^N \max_{0 \leq k \leq \lfloor NT \rfloor} (1 + |W_k^i|^{\gamma_*})^2,$$

which leads to

$$\mathbf{E}\Big[\sup_{t\in[0,T]}|\langle f,V_t^N\rangle|^2\Big] \le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^3}\sum_{i=1}^N \mathbf{E}\left[\max_{0\le k\le \lfloor NT\rfloor}(1+|W_k^i|^{\gamma_*})^2\right]$$
$$\le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^3}\sum_{i=1}^N \sqrt{\sum_{k=0}^{\lfloor NT\rfloor}\mathbf{E}\left[(1+|W_k^i|^{\gamma_*})^4\right]} \le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^{3/2}}. \tag{2.2.33}$$

Let us now consider the fifth term in the right-hand side of (2.2.32). From (2.2.14) and (3.9.3), we have

$$|\langle f,R_k^N\rangle|^2 \le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N}\sum_{i=1}^N\Big[\frac{1}{N^4}(1+|W_k^i|^{\gamma_*}+|W_{k+1}^i|^{\gamma_*})^4+\frac{1}{N^4|B_k|}\sum_{(x,y)\in B_k}(|y|^4+C)^2$$
$$+\frac{1}{N^{4\beta}}\left[|\varepsilon_k^i|^8+(1+|W_k^i|^{\gamma_*}+|W_{k+1}^i|^{\gamma_*})^4\right]\Big].$$

Then, by **A3** and **A5** together with Lemma 2.11 and (2.2.10), we obtain

$$\mathbf{E}\left[|\langle f,R_k^N\rangle|^2\right] \le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}\Big[\frac{1}{N^4}+\frac{1}{N^{4\beta}}\Big]. \tag{2.2.34}$$

Therefore, using also (3.9.3), it holds:

$$\mathbf{E}\Big[\sup_{t\in[0,T]}\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,R_k^N\rangle\Big|^2\Big] \le \lfloor NT\rfloor\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbf{E}\left[|\langle f,R_k^N\rangle|^2\right] \le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}\Big[\frac{1}{N^2}+\frac{N^2}{N^{4\beta}}\Big]. \tag{2.2.35}$$

Let us deal with the last term in the right-hand side of (2.2.32). Using the same arguments leading to (2.2.30) together with (3.9.3) and (2.2.29) we have

$$\frac{1}{N^{2+2\beta}}\mathbf{E}\Big[\sup_{t\in[0,T]}\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big] \le \frac{C}{N^{2+2\beta}}\mathbf{E}\Big[\sup_{t\in[0,T]}\lfloor Nt\rfloor\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big|\sum_{i=1}^N\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big]$$
$$\le \frac{C\lfloor NT\rfloor}{N^{2+2\beta}}\mathbf{E}\Big[\sum_{k=0}^{\lfloor NT\rfloor-1}\Big|\sum_{i=1}^N\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big]$$
$$\le \frac{C}{N^{1+2\beta}}\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbf{E}\Big[\Big|\sum_{i=1}^N\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big]$$
$$\le \frac{C}{N^{1+2\beta}}\sum_{k=0}^{\lfloor NT\rfloor-1}\sum_{i=1}^N\mathbf{E}\big[|\nabla f(W_k^i)\cdot\varepsilon_k^i|^2\big] \le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^{2\beta-1}}. \tag{2.2.36}$$

Plugging all these previous bounds in (2.2.32), we obtain (recall that $\beta\ge 1/2$), for all $f\in\mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$, $\mathbf{E}[\sup_{t\in[0,T]}\langle f,\mu_t^N\rangle^2]\le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}$. This proves (2.2.31) and ends the proof of the lemma. □

Lemma 2.15 provides the following compact containment for $(\mu^N)_{N\ge1}$ in $\mathcal{D}(\mathbf{R}_+,\mathcal{P}_\gamma(\mathbf{R}^d))$.

Corollary 2.16. *Assume $\beta\ge1/2$ and* **A1**-**A5**. *Let $0<\epsilon<1$. For every $T>0$,*

$$\sup_{N\ge1}\mathbf{E}\Big[\sup_{t\in[0,T]}\int_{\mathbf{R}^d}|w|^{\gamma+\epsilon}\mu_t^N(\mathrm{d}w)\Big] < +\infty. \tag{2.2.37}$$

Proof. Recall $\gamma_* - \gamma = 1$. Thus, it holds $f : w \mapsto (1 - \chi(w))|w|^{\gamma+\epsilon} \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$ since $\gamma_* > \gamma + \epsilon$. The result follows from Lemma 2.15. □

The following result will also be needed.

Lemma 2.17. *Assume* **A1**-**A5** *and* $\beta \geq 1/2$. *For all* $T > 0$, *there exists* $C > 0$ *such that for all* $\delta > 0$ *and* $0 \leq r < t \leq T$ *such that* $t - r \leq \delta$, *one has for all* $N \geq 1$ *and* $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$:

$$\mathbf{E}\left[|\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle|^2\right] \leq C\|f\|^2_{\mathcal{C}^{2,\gamma_*}} \left[\delta^2 + \frac{1}{N} + \frac{1}{N^2} + (N\delta + 1)\left[\frac{1}{N^4} + \frac{1}{N^{4\beta}}\right] + \frac{1}{N^{2\beta}}\right]. \tag{2.2.38}$$

Proof. Let $\delta > 0$ and $0 \leq r < t \leq T$ such that $t - r \leq \delta$. Let $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$. In the following, $C > 0$ is a constant independent of t, r, δ, N, and f, which can change from one occurence to another. From (2.2.8), we have

$$\begin{aligned}\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle = &\int_r^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \mu_s^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot, x), \mu_s^N\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &+ \langle f, M_t^N\rangle - \langle f, M_r^N\rangle + \langle f, V_t^N\rangle - \langle f, V_r^N\rangle + \sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor - 1} \langle f, R_k^N\rangle \\ &+ \frac{1}{N^{1+\beta}} \sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor - 1} \sum_{i=1}^N \nabla f(W_k^i) \cdot \varepsilon_k^i.\end{aligned}$$

Jensen's inequality provides

$$\begin{aligned}|\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle|^2 \leq C\Big[&(t - r)\int_r^t \int_{\mathcal{X}\times\mathcal{Y}} \left|(y - \langle \sigma_*(\cdot, x), \mu_s^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot, x), \mu_s^N\rangle\right|^2 \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &+ \left|\langle f, M_t^N\rangle - \langle f, M_r^N\rangle\right|^2 + \left|\langle f, V_t^N\rangle - \langle f, V_r^N\rangle\right|^2 + \left|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor - 1} \langle f, R_k^N\rangle\right|^2 \\ &+ \frac{1}{N^{2+2\beta}}\left|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor - 1} \sum_{i=1}^N \nabla f(W_k^i) \cdot \varepsilon_k^i\right|^2\Big].\end{aligned} \tag{2.2.39}$$

We now study each term of the right-hand side of (2.2.39). Let us consider the first term in the right-hand side of (2.2.39). From (2.2.15), (2.2.16) and (3.9.3), we have:

$$\mathbf{E}\left[|y - \langle \sigma_*(\cdot, x), \mu_s^N\rangle|^2 |\langle \nabla f \cdot \nabla\sigma_*(\cdot, x), \mu_s^N\rangle|^2\right] \leq \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N} \mathbf{E}\Big[(|y|^2 + C)\sum_{i=1}^N (1 + |W^i_{\lfloor Ns\rfloor}|^{\gamma_*})^2\Big] \leq C\|f\|^2_{\mathcal{C}^{2,\gamma_*}},$$

where the last inequality follows from **A3** and Lemma 2.11. We then have:

$$\begin{aligned}\mathbf{E}\left[(t - r)\int_r^t \int_{\mathcal{X}\times\mathcal{Y}} \left((y - \langle \sigma_*(\cdot, x), \mu_s^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot, x), \mu_s^N\rangle\right)^2 \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s\right] &\leq C(t - r)^2\|f\|^2_{\mathcal{C}^{2,\gamma_*}} \\ &\leq C\delta^2\|f\|^2_{\mathcal{C}^{2,\gamma_*}}.\end{aligned} \tag{2.2.40}$$

Let us consider the second term in the right-hand side of (2.2.39). From item *(iii)* of Lemma 2.12, we have

$$\mathbf{E}\left[\left(\langle f, M_t^N\rangle - \langle f, M_r^N\rangle\right)^2\right] \leq 2\mathbf{E}\left[\langle f, M_t^N\rangle^2 + \langle f, M_r^N\rangle^2\right] \leq \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N}. \tag{2.2.41}$$

Let us consider the third term in the right-hand side of (2.2.39). From (2.2.18) and (3.9.3), we have

$$|\langle f, V_t^N\rangle|^2 \le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^3}\sum_{i=1}^N (1+|W^i_{\lfloor Nt\rfloor}|^{\gamma_*})^2.$$

Therefore, by Lemma 2.11, we obtain that:

$$\mathbf{E}\big[\big|\langle f, V_t^N\rangle - \langle f, V_r^N\rangle\big|^2\big] \le 2\mathbf{E}\left[|\langle f, V_t^N\rangle|^2 + |\langle f, V_r^N\rangle|^2\right] \le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^2}. \tag{2.2.42}$$

Let us consider the fourth term in the right-hand side of (2.2.39). By (2.2.34),

$$\begin{aligned}\mathbf{E}\Big[\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\langle f, R_k^N\rangle\Big|^2\Big] &\le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}(\lfloor Nt\rfloor - \lfloor Nr\rfloor)\left[\frac{1}{N^4}+\frac{1}{N^{4\beta}}\right]\\ &\le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}(N\delta+1)\left[\frac{1}{N^4}+\frac{1}{N^{4\beta}}\right].\end{aligned} \tag{2.2.43}$$

Let us consider the last term in the right-hand side of (2.2.39). By item *(iv)* in Lemma 2.12,

$$\begin{aligned}&\frac{1}{N^{2+2\beta}}\mathbf{E}\Big[\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big]\\ &\le \frac{2}{N^{2+2\beta}}\mathbf{E}\Big[\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2 + \Big|\sum_{k=0}^{\lfloor Nr\rfloor-1}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big] \le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^{2\beta}}.\end{aligned} \tag{2.2.44}$$

Using (2.2.40), (2.2.41), (2.2.42), (2.2.43), (2.2.44), and (2.2.39), we deduce (2.2.38). □

We now collect the results of the previous lemmata to prove Proposition 2.13.

Proof of Proposition 2.13. To prove Proposition 2.13, we apply [Jak86, Theorem 4.6] with $E = \mathcal{P}_\gamma(\mathbf{R}^d)$ and $\mathbb{F} = \{\mathsf{V}_f, f \in \mathcal{C}_c^\infty(\mathbf{R}^d)\}$ where

$$\mathsf{V}_f : \nu \in \mathcal{P}_\gamma(\mathbf{R}^d) \mapsto \langle f, \nu\rangle.$$

The set $\mathbb{F}$ on $\mathcal{P}_\gamma(\mathbf{R}^d)$ satisfies Conditions [Jak86, (3.1) and (3.2) in Theorem 3.1]. Condition (4.8) there is a consequence of Proposition 3.15, Corollary 3.19, together with Markov's inequality. We now prove that [Jak86, Condition (4.9)] is verified, i.e. let us show that all $f \in \mathcal{C}_c^\infty(\mathbf{R}^d)$, the sequence $(\langle f, \mu^N\rangle)_{N\ge1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathbf{R})$. To do so, it suffices to use Lemma 2.17 and Proposition 2.41 below (with $\mathcal{H}_1 = \mathcal{H}_2 = \mathbf{R}$ there). In conclusion, according to [Jak86, Theorem 4.6], $(\mu^N)_{N\ge1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. □

Limit points in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ are continuous in time

In this section we show that any limit point of $(\mu^N)_{N\ge1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ belongs a.s. to $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$.

Proposition 2.18. *Let $\beta > 1/2$ and assume* **A1-A5**. *Consider $\mu^* \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ a limit point of $(\mu^N)_{N\ge1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. Then, a.s. $\mu^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$.*

Proof. Let N' be a subsequence such that in distribution $\mu^{N'} \to \mu^*$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. Because $\mathsf{W}_1 \le \mathsf{W}_\gamma$, $\mu^{N'} \to \mu^*$ in distribution also in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. By [JS87, Proposition 3.26 in Chapter VI], $\mu^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ a.s. if for all $T > 0$, $\lim_{N\to+\infty} \mathbf{E}\big[\sup_{t\in[0,T]} \mathsf{W}_1(\mu^N_{t-}, \mu^N_t)\big] = 0$. According to the duality formula (3.8.13), this is equivalent to

$$\lim_{N\to+\infty} \mathbf{E}\Big[\sup_{t\in[0,T]} \sup_{\|f\|_{\mathrm{Lip}}\le 1} |\langle f, \mu^N_{t-}\rangle - \langle f, \mu^N_t\rangle|\Big] = 0. \tag{2.2.45}$$

Let $T > 0$ and consider a Lipschitz function $f : \mathbf{R}^d \to \mathbf{R}$ such that $\|f\|_{\mathrm{Lip}} \le 1$. One has that $\langle f, \mu^N_t\rangle = \langle f, \mu^N_0\rangle + \sum_{k=0}^{\lfloor Nt\rfloor -1} \langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle$ (with the convention $\sum_0^{-1} = 0$). Therefore, the discontinuity points of $t \in [0,T] \mapsto \langle f, \mu^N_t\rangle$ are exactly $\{1/N, 2/N, \dots, \lfloor NT\rfloor/N\}$ and for all $t \in [0,T]$,

$$|\langle f, \mu^N_{t-}\rangle - \langle f, \mu^N_t\rangle| \le \max_{k=0,\dots,\lfloor NT\rfloor -1} |\langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle|. \tag{2.2.46}$$

Let $k \in \{0, \dots, \lfloor NT\rfloor - 1\}$. We have using (2.1.2) and **A2**:

$$|\langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle| \le \frac{1}{N}\sum_{i=1}^N |W^i_{k+1} - W^i_k| \le \frac{C}{N}\sum_{i=1}^N \Big[\frac{1}{N|B_k|}\sum_{(x,y)\in B_k} (|y|+1) + \frac{|\varepsilon^i_k|}{N^\beta}\Big] =: \beta^N_k \tag{2.2.47}$$

Then, one deduces that

$$|\beta^N_k|^2 \le \frac{C}{N}\sum_{i=1}^N \Big[\frac{1}{N^2|B_k|}\sum_{(x,y)\in B_k} (|y|^2+1) + \frac{|\varepsilon^i_k|^2}{N^{2\beta}}\Big],$$

and hence that $\mathbf{E}[|\beta^N_k|^2] \le C(1/N^2 + 1/N^{2\beta})$ where $C > 0$ is independent of $N \ge 1$ and $k = 0, \dots, \lfloor NT\rfloor - 1$. Then, using (3.9.20) and(3.9.21),

$$\begin{aligned}
\mathbf{E}\Big[\sup_{t\in[0,T]} \sup_{\|f\|_{\mathrm{Lip}}\le 1} |\langle f, \mu^N_{t-}\rangle - \langle f, \mu^N_t\rangle|\Big] &\le \mathbf{E}\Big[\sup_{\|f\|_{\mathrm{Lip}}\le 1} \max_{k=0,\dots,\lfloor NT\rfloor -1} |\langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle|\Big] \\
&\le \mathbf{E}\Big[\max_{k=0,\dots,\lfloor NT\rfloor -1} \beta^N_k\Big] \\
&\le \mathbf{E}\Big[\sqrt{\sum_{k=0}^{\lfloor NT\rfloor -1} |\beta^N_k|^2}\Big] \le \sqrt{\mathbf{E}\Big[\sum_{k=0}^{\lfloor NT\rfloor -1} |\beta^N_k|^2\Big]} \le C\big[1/\sqrt{N} + \sqrt{N/N^{2\beta}}\big].
\end{aligned}$$

This proves (3.9.19) since $\beta > 1/2$. The proof of Proposition 3.23 is complete. □

We end this section with the following result which will be used later in the proof of Theorem 2.8.

Lemma 2.19. *Let $\beta \ge 1/2$ and assume* **A1**-**A5**. *Then, for all $T > 0$ there exists $C > 0$,*

$$\mathbf{E}\Big[\sup_{t\in[0,T]} \langle f, \mu^N_t - \mu^N_{t-}\rangle^2\Big] \le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}\Big[\frac{1}{N^{3/2}} + \sqrt{\frac{1}{N^7} + \frac{1}{N^{8\beta-1}}} + \frac{\sqrt{N}}{N^{2\beta}}\Big],\ \forall f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d). \tag{2.2.48}$$

Proof. The arguments used in the proof of Proposition 3.23 are not sufficient to prove (2.2.48). We will rather use (2.2.8). Let $T > 0$ and $f \in \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$. In what follows, $C > 0$ is a constant, independent of $N \ge 1$, $k = 0, \dots \lfloor NT\rfloor - 1$, and f, which can change from one occurence to another. Recall that $t \in [0,T] \mapsto \langle f, \mu^N_t\rangle \in \mathbf{R}$ has $\lfloor NT\rfloor$ discontinuities, located at the points

$\frac{1}{N}, \frac{2}{N}, \dots, \frac{\lfloor NT \rfloor}{N}$. In addition, from (2.2.8), (2.2.7), and (2.2.6), for $k \in \{1, \dots, \lfloor NT \rfloor\}$, its k-th discontinuity is equal to

$$\begin{aligned}\mathbf{d}_k^N[f] :=& \langle f, M_{k-1}^N \rangle + \int_{\frac{k-1}{N}}^{\frac{k}{N}} \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \mu_s^N \rangle) \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \mu_s^N \rangle \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \\ &+ \langle f, R_{k-1}^N \rangle + \frac{1}{N^{1+\beta}} \sum_{i=1}^N \nabla f(W_{k-1}^i) \cdot \varepsilon_{k-1}^i. \end{aligned} \tag{2.2.49}$$

Thus,

$$\sup_{t\in[0,T]} |\langle f, \mu_t^N - \mu_{t^-}^N \rangle|^2 \leq \max\left\{ |\mathbf{d}_{k+1}^N[f]|^2,\ 0 \leq k < \lfloor NT \rfloor \right\}. \tag{2.2.50}$$

By (2.2.21) and (3.9.3), it holds:

$$|\langle f, M_k^N \rangle|^4 \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}^4}{N^5 |B_k|} \sum_{(x,y)\in B_k} \sum_{i=1}^N (y^4 + C)(1 + |W_k^i|^{\gamma_*})^4.$$

Then, using Lemma 2.11, we have with the same computations as the one made in (2.2.22):

$$\mathbf{E}\left[|\langle f, M_k^N \rangle|^4 \right] \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}^4}{N^5} \sum_{i=1}^N \mathbf{E}\left[y^4 + C \right] \mathbf{E}\left[(1 + |W_k^i|^{\gamma_*})^4 \right] \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}^4}{N^4}. \tag{2.2.51}$$

Consequently, one has:

$$\mathbf{E}\left[\max_{0\leq k < \lfloor NT \rfloor} \langle f, M_k^N \rangle^2 \right] \leq \Big| \sum_{k=0}^{\lfloor NT \rfloor - 1} \mathbf{E}\left[\langle f, M_k^N \rangle^4 \right] \Big|^{1/2} \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}^2}{N^{3/2}}. \tag{2.2.52}$$

By (2.2.15), (2.2.16) and since $\lfloor Ns \rfloor = k$ when $s \in [k/N, (k+1)/N]$, we have

$$\begin{aligned} &\Big| \int_{\frac{k}{N}}^{\frac{k+1}{N}} \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \mu_s^N \rangle) \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \mu_s^N \rangle \, \pi(\mathrm{d}x, \mathrm{d}y) \, \mathrm{d}s \Big| \\ &\leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N} \int_{\frac{k}{N}}^{\frac{k+1}{N}} \int_{\mathcal{X}\times\mathcal{Y}} (|y| + C) \sum_{i=1}^N (1 + |W_k^i|^{\gamma_*}) \, \pi(\mathrm{d}x, \mathrm{d}y) \, \mathrm{d}s \\ &= \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N^2} \mathbf{E}[|y| + C] \sum_{i=1}^N (1 + |W_k^i|^{\gamma_*}) \\ &\leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N^2} \sum_{i=1}^N (1 + |W_k^i|^{\gamma_*}). \end{aligned}$$

By (3.9.3) and Lemma 2.11, it then holds:

$$\begin{aligned} &\mathbf{E}\Big[\Big| \int_{\frac{k}{N}}^{\frac{k+1}{N}} \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \mu_s^N \rangle) \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \mu_s^N \rangle \, \pi(\mathrm{d}x, \mathrm{d}y) \, \mathrm{d}s \Big|^4 \Big] \\ &\leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}^4 N^3}{N^8} \mathbf{E}\Big[\sum_{i=1}^N (1 + |W_k^i|^{\gamma_*})^4 \Big] \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}^4}{N^4}. \end{aligned}$$

Thus, one has:

$$\mathbf{E}\Big[\max_{0\le k<\lfloor NT\rfloor}\Big|\int_{\frac{k}{N}}^{\frac{k+1}{N}}\int_{\mathcal{X}\times\mathcal{Y}}\alpha(y-\langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\mu_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|^2\Big]$$
$$\le\Big|\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbf{E}\Big[\Big|\int_{\frac{k}{N}}^{\frac{k+1}{N}}\int_{\mathcal{X}\times\mathcal{Y}}\alpha(y-\langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\mu_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|^4\Big]\Big|^{1/2}\le\frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}}{N^{3/2}}.$$

On the other hand, from (2.2.14) and (3.9.3), we have

$$|\langle f,R_k^N\rangle|^4\le\frac{C\|f\|^4_{\mathcal{C}^{2,\gamma_*}}}{N}\sum_{i=1}^N\Bigg[\frac{1}{N^8}(1+|W_k^i|^{\gamma_*}+|W_{k+1}^i|^{\gamma_*})^8+\frac{1}{N^8|B_k|}\sum_{(x,y)\in B_k}(|y|^4+C)^4$$
$$+\frac{C}{N^{8\beta}}\left[|\varepsilon_k^i|^{16}+(1+|W_k^i|^{\gamma}+|W_{k+1}^i|^{\gamma_*})^8\right]\Bigg].$$

Using Lemma 2.11, **A3**, and the same computations as those made in (2.2.10), we deduce that:

$$\mathbf{E}\left[|\langle f,R_k^N\rangle|^4\right]\le C\|f\|^4_{\mathcal{C}^{2,\gamma}}\big(1/N^8+1/N^{8\beta}\big).$$

Then, it holds:

$$\mathbf{E}\Big[\max_{0\le k<\lfloor NT\rfloor}|\langle f,R_k^N\rangle|^2\Big]\le\Big|\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbf{E}\left[|\langle f,R_k^N\rangle|^4\right]\Big|^{1/2}\le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}\Big[\frac{1}{N^7}+\frac{1}{N^{8\beta-1}}\Big]^{1/2}.$$

By (3.9.3), **A5**, and Lemma 2.11,

$$\mathbf{E}\Big[\Big|\sum_{i=1}^N\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^4\Big]\le N^3\sum_{i=1}^N\mathbf{E}\left[|\nabla f(W_k^i)\cdot\varepsilon_k^i|^4\right]$$
$$\le N^3\|f\|^4_{\mathcal{C}^{2,\gamma_*}}\sum_{i=1}^N\mathbf{E}\left[(1+|W_k^i|^{\gamma_*})^4\right]\mathbf{E}\left[|\varepsilon_k^i|^4\right]\le C\|f\|^4_{\mathcal{C}^{2,\gamma_*}}N^4.$$

Thus, one deduces that

$$\mathbf{E}\Big[\max_{0\le k<\lfloor NT\rfloor}\Big|\frac{1}{N^{1+\beta}}\sum_{i=1}^N\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big]\le\Big|\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbf{E}\Big[\Big|\frac{1}{N^{1+\beta}}\sum_{i=1}^N\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^4\Big]\Big|^{1/2}$$
$$\le\Big|\frac{1}{N^{4+4\beta}}C\|f\|^4_{\mathcal{C}^{2,\gamma_*}}N^5\Big|^{1/2}\le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}\frac{\sqrt{N}}{N^{2\beta}}.$$

Plugging all these previous bounds in (2.2.50) implies (2.2.48). □

Convergence to the limit equation (2.1.10)

This section is devoted to prove Proposition 2.21 where we show that any limit point of $(\mu^N)_{N\ge1}$ in $\mathcal{D}(\mathbf{R}_+,\mathcal{P}_\gamma(\mathbf{R}^d))$ satisfies a.s. (2.1.10).

For $t\in\mathbf{R}_+$ and $f\in\mathcal{C}^{1,\gamma}(\mathbf{R}^d)$, we introduce the function $\mathbf{\Lambda}_t[f]:\mathcal{D}(\mathbf{R}_+,\mathcal{P}_\gamma(\mathbf{R}^d))\to\mathbf{R}$ defined by

$$\mathbf{\Lambda}_t[f]:m\mapsto\Big|\langle f,m_t\rangle-\langle f,\mu_0\rangle-\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha(y-\langle\sigma_*(\cdot,x),m_s\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),m_s\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big|.$$

To prove that any limit point of the sequence $(\mu^N)_{N\ge1}$ in the space $\mathcal{D}(\mathbf{R}_+,\mathcal{P}_\gamma(\mathbf{R}^d))$ satisfies (2.1.10), we study the continuity of the function $\mathbf{\Lambda}_t[f]$. This is the purpose of Lemma 2.20.

Lemma 2.20. *For any $t \in \mathbf{R}_+$ and $f \in \mathcal{C}^{1,\gamma}(\mathbf{R}^d)$, the function $\mathbf{\Lambda}_t[f]$ is well defined. In addition, let $(m^N)_{N\geq 1}$ be such that $m^N \to m$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. Then, for all continuity points $t \in \mathbf{R}_+$ of m, $\mathbf{\Lambda}_t[f](m^N) \to \mathbf{\Lambda}_t[f](m)$ as $N \to +\infty$.*

Proof. In the following $C > 0$ is a constant independent of $f \in \mathcal{C}^{1,\gamma}(\mathbf{R}^d)$, $s \in [0,t]$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, which can change from one occurence to another. By **A4**

$$|\langle f, \mu_0\rangle| = \Big| \int_{\mathbf{R}^d} \frac{f(w)}{1+|w|^\gamma}(1+|w|^\gamma)\mu_0(\mathrm{d}w)\Big| \leq (1+\mathbf{E}[|W_0^1|^\gamma])\|f\|_{\mathcal{C}^{0,\gamma}}. \tag{2.2.53}$$

The following result [Vil09, Theorem 6.9] will be used many times in the following:

$$\mu_n \to \mu \text{ in } \mathcal{P}_\gamma(\mathbf{R}^d) \text{ iff } \langle g, \mu_n\rangle \to \langle g, \mu\rangle \text{ for all } g : \mathbf{R}^d \to \mathbf{R} \text{ continuous s.t. } \frac{g}{1+|\cdot|^\gamma} \text{ is bounded.} \tag{2.2.54}$$

In particular $u \geq 0 \mapsto \langle 1+|\cdot|^\gamma, m_u\rangle \in \mathcal{D}(\mathbf{R}_+, \mathbf{R})$ and thus $\sup_{u\in[0,t]} |\langle 1+|\cdot|^\gamma, m_u\rangle| < +\infty$ for all $t \geq 0$. Define the function $\phi_s^{x,y}(m) = \alpha(y - \langle \sigma_*(\cdot, x), m_s\rangle)\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), m_s\rangle$. Using **A2**, one has for $s \in [0,t]$:

$$|\phi_s^{x,y}(m)| \leq C(1+|y|) \sup_{u\in[0,t]} |\langle 1+|\cdot|^\gamma, m_u\rangle| \|f\|_{\mathcal{C}^{1,\gamma}}. \tag{2.2.55}$$

Using also **A3**, this proves that $\mathbf{\Lambda}_t[f]$ is well defined.

Let us now consider $(m^N)_{N\geq 1}$ such that $m^N \to m$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. Denote by $\mathcal{C}(m) \subset \mathbf{R}_+$ the set of continuity points of m. From [EK09, Proposition 5.2 in Chapter 3], we have that for all $t \in \mathcal{C}(m)$, $m_t^N \to m_t$ in $\mathcal{P}_\gamma(\mathbf{R}^d)$, and thus, for all $t \in \mathcal{C}(m)$, according to (2.2.54),

$$\langle f, m_t^N\rangle \underset{N\to\infty}{\longrightarrow} \langle f, m_t\rangle.$$

For the same reasons, for all $s \in [0,t] \cap \mathcal{C}(m)$ and $x \in \mathcal{X}$,

$$\langle \sigma_*(\cdot, x), m_s^N\rangle \underset{N\to\infty}{\longrightarrow} \langle \sigma_*(\cdot, x), m_s\rangle \quad \text{and} \quad \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), m_s^N\rangle \underset{N\to\infty}{\longrightarrow} \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), m_s\rangle.$$

Since $\mathbf{R}_+ \backslash \mathcal{C}(m)$ is at most countable (see [EK09, Lemma 5.1 in Chapter 3]), it holds a.e. on $[0,t] \times \mathcal{X} \times \mathcal{Y}$, $\phi_s^{x,y}(m^N) \to \phi_s^{x,y}(m)$. Note that using [EK09, Item (b) in Proposition 5.3 in Chapter 3] together with the triangular inequality:

$$\langle |\cdot|^\gamma, m_u^N\rangle = \mathsf{W}_\gamma(\delta_0, m_u^N) \leq [\mathsf{W}_\gamma(\delta_0, m_{\lambda_u^N}) + \mathsf{W}_\gamma(m_{\lambda_u^N}, m_u^N)]^\gamma,\ \lambda^N : \mathbf{R}_+ \to \mathbf{R}_+,\ u \geq 0,$$

one deduces that there exists $C > 0$, for all $N \geq 1$ and $s \in [0,t]$, $|\langle 1+|\cdot|^\gamma, m_s^N\rangle| \leq C$. Together with (2.2.55), one has using the dominated convergence theorem, $\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \phi_s^{x,y}(m^N)\pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \to \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \phi_s^{x,y}(m)\pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s$. This proves the desired result. □

We are now in position to prove that any limit point of the sequence $(\mu^N)_{N\geq 1}$ in the space $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ satisfies (2.1.10).

Proposition 2.21. *Let $\beta > 1/2$ and assume **A1**-**A5**. Let μ^* be a limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. Then, a.s., μ^* satisfies* (2.1.10).

Proof. Up to extracting a subsequence, we assume that in distribution $\mu^N \to \mu^*$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$.

Let $t \in \mathbf{R}_+$ and $f \in \mathcal{C}_c^\infty(\mathbf{R}^d)$. By (2.2.8) and Lemma 2.12, we have:

$$\begin{aligned}
&\mathbf{E}\left[\mathbf{\Lambda}_t[f](\mu^N)\right] \\
&= \mathbf{E}\Big[\Big|\langle f,\mu_0^N\rangle - \langle f,\mu_0\rangle + \langle f,M_t^N\rangle + \langle f,V_t^N\rangle + \sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,R_k^N\rangle + \frac{1}{N^{1+\beta}}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|\Big] \\
&\le \mathbf{E}[\,|\langle f,\mu_0^N\rangle - \langle f,\mu_0\rangle|\,] + \mathbf{E}[|\langle f,V_t^N\rangle|] + \sqrt{\mathbf{E}[\langle f,M_t^N\rangle^2]} + \mathbf{E}\Big[\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,R_k^N\rangle\Big|\Big] \\
&\quad + \sqrt{\mathbf{E}\Big[\Big|\frac{1}{N^{1+\beta}}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big]} \le C\|f\|_{\mathcal{C}^{2,\gamma_*}}\Big[\frac{1}{\sqrt{N}} + \frac{1}{N} + \frac{1}{N^{2\beta-1}} + \frac{1}{N^\beta}\Big],
\end{aligned}$$

where the bound $\mathbf{E}[|\langle f,\mu_0^N\rangle - \langle f,\mu_0\rangle|] \le C\|f\|_{\mathcal{C}^{2,\gamma_*}}/\sqrt{N}$ follows from (2.2.53) and the fact that the initial coefficients are i.i.d. (see **A4**). Therefore, since $\beta > 1/2$, for all $t \in \mathbf{R}_+$ and $f \in \mathcal{C}_c^\infty(\mathbf{R}^d)$,

$$\lim_{N\to+\infty} \mathbf{E}\left[\mathbf{\Lambda}_t[f](\mu^N)\right] = 0. \tag{2.2.56}$$

By [EK09, Lemma 7.7 in Chapter 3], the complementary of the set

$$\mathcal{C}(\mu^*) = \{t \ge 0,\, \mathbf{P}(\mu^*_{t^-} = \mu^*_t) = 1\}$$

is at most countable. Let $t_* \in \mathcal{C}(\mu^*)$. Denoting by $\mathsf{D}(\mathbf{\Lambda}_{t_*}[f])$ the set of discontinuity points of $\mathbf{\Lambda}_{t_*}[f]$, we recall that from Lemma 2.20, $m \notin \mathsf{D}(\mathbf{\Lambda}_{t_*}[f])$ if m is continuous at t_*. Then, we have:

$$\mathbf{P}\big(\mu^* \in \mathsf{D}(\mathbf{\Lambda}_{t_*}[f])\big) = 0.$$

By [Bil99, Theorem 2.7], it then holds:

$$\lim_{N'\to+\infty} \mathbf{\Lambda}_{t_*}[f](\mu^N) = \mathbf{\Lambda}_{t_*}[f](\mu^*) \text{ in distribution, } \forall t_* \in \mathcal{C}(\mu^*). \tag{2.2.57}$$

By uniqueness of the limit in distribution, (2.2.56) and (2.2.57) imply that for all $t_* \in \mathcal{C}(\mu^*)$ and $f \in \mathcal{C}_c^\infty(\mathbf{R}^d)$, a.s. $\mathbf{\Lambda}_{t_*}[f](\mu^*) = 0$. It then remains to show that a.s. for all $t \in \mathbf{R}_+$ and $f \in \mathcal{C}_c^\infty(\mathbf{R}^d)$, $\mathbf{\Lambda}_t[f](\mu^*) = 0$. To do so we use a standard continuity argument.

First of all, for all $m \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ and $f \in \mathcal{C}_c^\infty(\mathbf{R}^d)$, the function $t \in \mathbf{R}_+ \mapsto \mathbf{\Lambda}_t[f](m)$ is right-continuous. Moreover, there exists a countable subset $\mathcal{R}_{\mu^*}$ of $\mathcal{C}(\mu^*)$ such that for all $t \ge 0$ and $\epsilon > 0$, there exists $s \in \mathcal{R}_{\mu^*}$, $s \in [t, t+\varepsilon]$. Thus, for all $f \in \mathcal{C}_c^\infty(\mathbf{R}^d)$, it holds a.s. for all $t \in \mathbf{R}_+$ $\mathbf{\Lambda}_t[f](\mu^*) = 0$.

Secondly, for all $m \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ and $t \ge 0$, using the dominated convergence theorem, the function $f \in \mathcal{H}^{L,\gamma}(\mathbf{R}^d) \mapsto \mathbf{\Lambda}_t[f](m)$ is continuous (because $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{1,\gamma}(\mathbf{R}^d)$, by (2.1.6)). Furthermore, $\mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ admits a dense and countable subset of elements in $\mathcal{C}_c^\infty(\mathbf{R}^d)$. Thus, a.s. for all $f \in \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ and all $t \in \mathbf{R}_+$ $\mathbf{\Lambda}_t[f](\mu^*) = 0$.

Note also that $\mathcal{C}_b^\infty(\mathbf{R}^d) \subset \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ since $2\gamma > d$. This ends the proof of the proposition. □

Note that we have not used Proposition 3.23 in the proof of Proposition 2.21.

2.2.3 Uniqueness of the limit equation in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ and proof of Theorem 2.1

Uniqueness of the limit equation in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$

Proposition 2.22. *There exists a unique solution to* (2.1.10) *in the space* $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$.

Proof. We have already proved the existence. Let us now prove that there is at most one solution to (2.1.10) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. The proof of the uniqueness of (2.1.10) relies on arguments developed in [PR16, PRT15] and is divided into several steps.

Step 1. Preliminary considerations.

If $\mu^\star$ is solution to (2.1.10), then for all $f \in \mathcal{C}_b^\infty(\mathbf{R}^d)$, $s \geq 0 \mapsto \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \mu_s^\star\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot, x), \mu_s^\star\rangle \pi(\mathrm{d}x, \mathrm{d}y)$ is continuous (by the dominated convergence theorem). This implies that for all $f \in \mathcal{C}_b^\infty(\mathbf{R}^d)$ and $t \in \mathbf{R}_+$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle f, \mu_t^\star\rangle = \int_{\mathbf{R}^d} \nabla f(w) \cdot \mathbf{V}[\mu_t^\star](w)\mu_t^\star(\mathrm{d}w),$$

where $\mathbf{V} : \mu \in \mathcal{P}(\mathbf{R}^d) \mapsto \mathbf{V}[\mu]$ is defined by

$$\mathbf{V}[\mu] : w \in \mathbf{R}^d \mapsto \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \mu\rangle)\nabla\sigma_*(w, x)\pi(\mathrm{d}x, \mathrm{d}y) \in \mathbf{R}^d.$$

Adopting the terminology of [San15, Section 4.1.2], $\mu^\star$ is thus a *weak solution*[2] of the measure-valued equation

$$\begin{cases} \partial_t \mu_t^\star = \operatorname{div}\left(\mathbf{V}[\mu_t^\star]\mu_t^\star\right) \\ \mu_0^\star = \mu_0. \end{cases} \tag{2.2.58}$$

Therefore, to prove the uniqueness result in Proposition 2.22, it is enough to show that (2.2.58) has a unique weak solution in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. To this end, we consider two solutions $\mu^1 = \{t \mapsto \mu_t^1, t \geq 0\}$ and $\mu^2 = \{t \mapsto \mu_t^2, t \geq 0\}$ of (2.2.58) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$, and we introduce the following mappings

$$\mathbf{v}^1 : (t, x) \in \mathbf{R}_+ \times \mathbf{R}^d \mapsto \mathbf{V}[\mu_t^1](x) \quad \text{and} \quad \mathbf{v}^2 : (t, x) \in \mathbf{R}_+ \times \mathbf{R}^d \mapsto \mathbf{V}[\mu_t^2](x).$$

Step 2. In this step, we prove some basic regularity properties of $\mathbf{V}$, $\mathbf{v}^1$, and $\mathbf{v}^2$.

Let us first prove that the velocity fields $\mathbf{v}^1$ and $\mathbf{v}^2$ are globally Lipschitz continuous over $\mathbf{R}_+ \times \mathbf{R}^d$. Let $\mu \in \{\mu^1, \mu^2\}$ and set $v(t, x) = \mathbf{V}[\mu_t](x)$. For $0 \leq s \leq t$ and $w_1, w_2 \in \mathbf{R}^d$, we have

$$|v(t, w_1) - v(s, w_2)| \leq |v(t, w_1) - v(t, w_2)| + |v(t, w_2) - v(s, w_2)|.$$

By **A2**, the function $w \mapsto \mathbf{V}[\mu](w)$ is smooth and $|\nabla\mathbf{V}[\mu](w)| \leq C$ for some $C > 0$ independent of μ and w. Thus, it holds

$$|v(t, w_1) - v(t, w_2)| = |\mathbf{V}[\mu_t](w_1) - \mathbf{V}[\mu_t](w_2)| \leq C|w_1 - w_2|,$$

for some $C > 0$ independent of t, w_1, and w_2. Secondly, for any $x \in \mathcal{X}$, considering (2.1.10) with $f = \sigma_*(\cdot, x)$, we obtain

$$\begin{aligned} |\langle \sigma_*(\cdot, x), \mu_s - \mu_t\rangle| &\leq \int_s^t \int_{\mathcal{X}\times\mathcal{Y}} \left|\alpha(y - \langle \sigma_*(\cdot, x'), \mu_r\rangle)\langle \nabla\sigma_*(\cdot, x) \cdot \nabla\sigma_*(\cdot, x'), \mu_r\rangle\right| \pi(\mathrm{d}x', \mathrm{d}y)\mathrm{d}r \\ &\leq C|t - s|, \end{aligned}$$

leading to

$$|v(t, w_2) - v(s, w_2)| = \left|\int_{\mathcal{X}\times\mathcal{Y}} \alpha\langle \sigma_*(\cdot, x), \mu_s - \mu_t\rangle \nabla\sigma_*(w_2, x)\pi(\mathrm{d}x, \mathrm{d}y)\right| \leq C|t - s|.$$

[2] We mention that according to [San15, Proposition 4.2], the two notions of solutions of (2.2.58) (namely the weak solution and the *distributional* solution) are equivalent.

Thus, there exists $C > 0$ such that for $0 \leq s \leq t$ and $w_1, w_2 \in \mathbf{R}^d$, $|v(t, w_1) - v(s, w_2)| \leq C(|t-s| + |w_1 - w_2|)$, which proves that v is globally Lipschitz. Now we claim that there exists $L' > 0$ such that for every $\mu, \nu \in \mathcal{P}_1(\mathbf{R}^d)$,

$$\|\mathbf{V}[\mu] - \mathbf{V}[\nu]\|_\infty := \sup_{w \in \mathbf{R}^d} |\mathbf{V}[\mu](w) - \mathbf{V}[\nu](w)| \leq L' \mathsf{W}_1(\mu, \nu). \tag{2.2.59}$$

By **A2**, there exists $C > 0$ such that for all $\mu, \nu \in \mathcal{P}_1(\mathbf{R}^d)$ and all $w \in \mathbf{R}^d$,

$$\begin{aligned} |\mathbf{V}[\mu](w) - \mathbf{V}[\nu](w)| &= \left| \int_{\mathcal{X} \times \mathcal{Y}} \alpha(\langle \sigma_*(\cdot, x), \nu \rangle - \langle \sigma_*(\cdot, x), \mu \rangle) \nabla_W \sigma_*(w, x) \pi(\mathrm{d}x, \mathrm{d}y) \right| \\ &\leq C \int_{\mathcal{X} \times \mathcal{Y}} |\langle \sigma_*(\cdot, x), \nu \rangle - \langle \sigma_*(\cdot, x), \mu \rangle| \, \pi(\mathrm{d}x, \mathrm{d}y) \leq C \mathsf{W}_1(\mu, \nu), \end{aligned}$$

where the last inequality is obtained by the Lipschitz continuity of $\sigma_*(\cdot, x)$ (which is uniform in $x \in \mathcal{X}$).

Step 3. End of the proof of Proposition 2.22.

Since v is globally Lipschitz, we can introduce the flows $(\phi_t^1)_{t \in [0,T]}$ and $(\phi_t^2)_{t \in [0,T]}$ with respect to μ^1 and μ^2. By [Vil03, Theorem 5.34], one has

$$\mu_t^1 = \phi_t^1 \# \mu_0, \quad \mu_t^2 = \phi_t^2 \# \mu_0, \quad \forall t \geq 0. \tag{2.2.60}$$

The symbol $\#$ stands for the pushforward of a measure. Let $L > 0$ be a constant such that $|v_t^i(w_1) - v_t^i(w_2)| \leq L|w_1 - w_2|$ for all $i = 1, 2$, $t \in \mathbf{R}_+$ and $w_1, w_2 \in \mathbf{R}^d$ (which exists by the previous step). Then by [PR16, Proposition 4], it holds for all $\mu, \nu \in \mathcal{P}_1(\mathbf{R}^d)$,

$$\mathsf{W}_1(\phi_t^1 \# \mu, \phi_t^2 \# \nu) \leq e^{Lt} \mathsf{W}_1(\mu, \nu) + \frac{e^{Lt} - 1}{L} \sup_{0 \leq s \leq t} \|v_s^1 - v_s^2\|_\infty. \tag{2.2.61}$$

We are now in position to prove that $\mu^1 = \mu^2$. We use the techniques introduced in [PRT15]. Let us now consider $T > 0$, and introduce

$$t_0 := \inf\{t \in [0, T], \, \mathsf{W}_1(\mu_t^1, \mu_t^2) \neq 0\}.$$

We shall prove that $t_0 = T$. Assume that $t_0 < T$. By (2.2.60) and (2.2.61), we have, for $0 \leq s \leq T - t_0$,

$$\mathsf{W}_1(\mu_{t_0+s}^1, \mu_{t_0+s}^2) \leq e^{Ls} \mathsf{W}_1(\mu_{t_0}^1, \mu_{t_0}^2) + \frac{e^{Ls} - 1}{L} \sup_{t_0 \leq \tau \leq t_0 + s} \|v_\tau^1 - v_\tau^2\|_\infty$$

By continuity, $\mathsf{W}_1(\mu_{t_0}^1, \mu_{t_0}^2) = 0$. For s small enough such that $e^{Ls} - 1 < 2Ls$, we obtain, using (2.2.59),

$$\mathsf{W}_1(\mu_{t_0+s}^1, \mu_{t_0+s}^2) \leq 2sL' \sup_{t_0 \leq \tau \leq t_0 + s} \mathsf{W}_1(\mu_\tau^1, \mu_\tau^2).$$

Then, for $0 \leq s' \leq s < \min(1/2L', T - t_0)$, applying the last inequality for s' gives

$$\mathsf{W}_1(\mu_{t_0+s'}^1, \mu_{t_0+s'}^2) < \sup_{t_0 \leq \tau \leq t_0 + s} \mathsf{W}_1(\mu_\tau^1, \mu_\tau^2),$$

which is not possible. Hence, $t_0 = T$, and again, by continuity, we conclude that $\mathsf{W}_1(\mu_t^1, \mu_t^2) = 0$, $\forall t \in [0, T]$. Therefore, $\mu^1 = \mu^2$. We have thus proved that (2.1.10) admits a unique solution in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$, which is the desired result. □

We end this section by giving an alternative proof of Proposition 2.22 using probabilistic arguments developed by [Szn91].

Alternative proof of Proposition 2.22. Introduce, for $\hat{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d))$, the system

$$\begin{cases} \mathrm{d}X_t = \alpha \int_{\mathcal{X}\times\mathcal{Y}} (y - \langle \sigma_*(\cdot, x), \hat{\mu}_t \rangle) \nabla_W \sigma_*(X_t, x) \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}t, \\ X_0 \sim \mu_0. \end{cases} \tag{$S_{\hat{\mu}}$}$$

Introduce the function

$$\begin{aligned} F : \mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d)) &\to \mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d)) \\ \hat{\mu} &\mapsto (t \mapsto \hat{P}_t) \end{aligned}$$

where $\hat{P} \in \mathcal{P}(\mathcal{C}(\mathbf{R}_+, \mathbf{R}^d))$ is the weak solution of $(S_{\hat{\mu}})$. Let us justify why $t \mapsto \hat{P}_t$ defines an element of $\mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d))$. Considering a test function $g \in \mathcal{C}_c^\infty(\mathbf{R}^d)$ and $0 \leq s < t$, we have, using **A2**,

$$\Big| \int_{\mathbf{R}^d} g(w) \hat{P}_t(\mathrm{d}w) - \int_{\mathbf{R}^d} g(w) \hat{P}_s(\mathrm{d}w) \Big| \leq \mathbf{E}[|g(X_t) - g(X_s)|] \leq C\mathbf{E}[|X_t - X_s|] \leq C(t-s)$$

Hence, by [AGS05, Remark 5.1.6], the function $t \mapsto \hat{P}_t$ defines an element of $\mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d))$. Given a vector field $v : \mathbf{R}^d \to \mathbf{R}^d$, introduce also

$$\begin{cases} \partial_t \bar{\mu} = \mathrm{div}(v\bar{\mu}_t), \\ \bar{\mu}_0 = \mu_0, \end{cases} \tag{2.2.62}$$

We are now in position to prove the uniqueness of (2.1.10) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. Let $\mu^1, \mu^2 \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ be two solutions of (2.1.10). On the one hand, for $i \in \{1, 2\}$, μ^i is a solution of (2.2.62) with $v^i = \mathbf{V}[\mu^i]$. On the other hand, taking the expectancy of

$$g(X_t) = g(X_0) + \int_0^t \nabla g(X_s) \cdot \frac{\mathrm{d}X_s}{\mathrm{d}s} \mathrm{d}s,$$

where $g \in \mathcal{C}_c^1(\mathbf{R}^d)$, shows that $F(\mu^i)$ is a solution of (2.2.62) with $v^i = \mathbf{V}[\mu^i]$. Since (2.2.62) admits a unique solution in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^d))$ (see [Vil03, Theorem 5.34]), it holds $\mu^i = F(\mu^i)$. By [Szn91, Theorem 1.1], F admits a unique fixed point. Hence $\mu^1 = \mu^2$, which concludes the proof. □

End of the proof of Theorem 2.1

We can now prove Theorem 2.1.

Proof of Theorem 2.1. By Proposition 2.13, the sequence $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. Let $\bar{\mu}^1$ and $\bar{\mu}^2$ be two limit points of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. Let $j \in \{1, 2\}$. By Lemma 3.23, a.s. $\bar{\mu}^j \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. According to Proposition 2.21, $\bar{\mu}^j$ satisfies a.s. (2.1.10). Let $\mu^\star \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$ be the unique solution of (2.1.10) (see Proposition 2.22). Therefore, one has that a.s. $\bar{\mu}^j = \mu^\star$ in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. Note that this implies also that $\mu^\star \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ and then that a.s. $\bar{\mu}^j = \mu^\star$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. Therefore, $(\mu^N)_{N\geq 1}$ converges in distribution to $\mu^\star$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$ (and then the convergence holds in probability). This ends the proof of Theorem 2.1. □

2.3 Proof of Theorem 2.8

In this section, we prove Theorem 2.8. Recall that $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-L,\gamma}(\mathbf{R}^d))$ is given by Corollary 2.2 and that the fluctuation process is defined by (see (2.1.11)):

$$\eta^N = \sqrt{N}(\mu^N - \bar{\mu}),\ N \geq 1.$$

Throughout this section, we assume that **A1**-**A7** hold.

2.3.1 Relative compactness of $(\eta^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$

To prove the relative compactness of $(\eta^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$, we will use Proposition 2.41 with $\mathcal{H}_1 = \mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d)$ and $\mathcal{H}_2 = \mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$. Mimicking the proof of [Szn91, Theorem 1.1], there exists a unique, trajectorial and in law, solution of

$$\begin{cases} \mathrm{d}X_t = \alpha \int_{\mathcal{X}\times\mathcal{Y}} (y - \langle \sigma_*(\cdot, x), \hat{\mu}_t \rangle) \nabla_W \sigma_*(X_t, x) \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}t, \\ X_0 \sim \mu_0, \quad \hat{\mu}_t = \mathrm{Law}(X_t). \end{cases} \tag{2.3.1}$$

Denote by $\hat{\mu} \in \mathcal{P}(\mathcal{C}(\mathbf{R}_+, \mathbf{R}^d))$ this solution. The mapping $t \geq 0 \mapsto \hat{\mu}_t$ satisfies Equation (2.1.10). In addition, using **A2**, it is straightforward to show that the function $t \mapsto \widetilde{\mu}_t$ lies in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^d))$. Since $\bar{\mu}$ is the unique solution of (2.1.10) (see Proposition 2.22), $\hat{\mu} = \bar{\mu}$. Therefore, we introduce, as it is customary, the particle system defined as follows: for $N \geq 1$, let $\bar{X}^i = \{t \mapsto \bar{X}^i_t, t \in \mathbf{R}_+\}$ ($i \in \{1, \ldots, N\}$) be the N independent processes satisfying:

$$\begin{cases} \bar{X}^i_t = W^i_0 + \int_0^t \alpha \int_{\mathcal{X}\times\mathcal{Y}} (y - \langle \sigma_*(\cdot, x), \bar{\mu}_s \rangle) \nabla_W \sigma_*(\bar{X}^i_s, x) \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s,\ t \in \mathbf{R}_+ \\ \bar{\mu}_t = \mathrm{Law}(\bar{X}^i_t). \end{cases} \tag{2.3.2}$$

We then introduce its empirical measure:

$$\bar{\mu}^N_t = \frac{1}{N} \sum_{i=1}^N \delta_{\bar{X}^i_t},\ N \geq 1,\ t \in \mathbf{R}_+. \tag{2.3.3}$$

By **A2**, there exists $C_0 > 0$ such that a.s. for all $0 \leq s \leq t$, it holds for all $i \in \{1, \ldots, N\}$:

$$|\bar{X}^i_t - \bar{X}^i_s| \leq C_0(t - s). \tag{2.3.4}$$

In particular, $t \in \mathbf{R}_+ \mapsto \bar{X}^i_t \in \mathbf{R}^d$ is a.s. continuous for all $i \in \{1, \ldots, N\}$. Because μ_0 is compactly supported (see indeed **A6**), one deduces that there exists $C > 0$ such that a.s. for all $T > 0$ and for all $i \in \{1, \ldots, N\}$:

$$\sup_{t\in[0,T]} |\bar{X}^i_t| \leq C(1 + T). \tag{2.3.5}$$

Thus, for any $\beta \geq 0$, a.s. $\bar{\mu}^N \in \mathcal{C}(\mathbf{R}_+, \mathcal{C}^{1,\beta}(\mathbf{R}^d)')$. We now define, for $N \geq 1$,

$$\Upsilon^N := \sqrt{N}(\mu^N - \bar{\mu}^N) \ \text{ and } \ \Theta^N := \sqrt{N}(\bar{\mu}^N - \bar{\mu}). \tag{2.3.6}$$

It then holds:

$$\eta^N = \Upsilon^N + \Theta^N. \tag{2.3.7}$$

For all $N \geq 1$ and for any $\beta \geq 0$, since a.s. $\mu^N \in \mathcal{D}(\mathbf{R}_+, \mathcal{C}^{0,\beta}(\mathbf{R}^d)')$, one has

$$\text{a.s. } \Upsilon^N \in \mathcal{D}(\mathbf{R}_+, \mathcal{C}^{1,\beta}(\mathbf{R}^d)').$$

In particular a.s. $\Upsilon^N \in \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{1-J_1,j_1}(\mathbf{R}^d))$ because $\mathcal{H}^{J_1-1,j_1}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{1,j_1}(\mathbf{R}^d)$ (according to (2.1.4) and since $J_1 - 1 > d/2 + 1$). On the other hand, since $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-L,\gamma}(\mathbf{R}^d))$ (see Corollary 2.2) and since a.s. $\bar{\mu}^N \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-L,\gamma}(\mathbf{R}^d))$, it holds for all $N \geq 1$:

$$\text{a.s. } \Theta^N \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-L,\gamma}(\mathbf{R}^d)). \tag{2.3.8}$$

We start with the following lemma.

Lemma 2.23. *Let $\beta \geq 3/4$ and assume* **A1**-**A7**. *Then, for all $T > 0$, we have*

$$\sup_{N\geq 1}\sup_{t\in[0,T]} \mathbf{E}\big[\|\Theta_t^N\|^2_{\mathcal{H}^{-J_1,j_1}} + \|\Upsilon_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}\big] < +\infty.$$

In particular, $\sup_{N\geq 1}\sup_{t\in[0,T]} \mathbf{E}[\|\eta_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}] < +\infty$.

Proof. Let $T > 0$. In all this proof, $f \in \mathcal{H}^{J_1,j_1}(\mathbf{R}^d)$ and $\{f_a\}_{a\geq 1}$ is an orthonormal basis of $\mathcal{H}^{J_1,j_1}(\mathbf{R}^d)$. In the following, C denotes a constant independent of $N \geq 1$, $t \in [0,T]$, $(x,y) \in \mathcal{X}\times\mathcal{Y}$, and the test function f, which can change from one occurence to another. The proof is divided into two steps.

Step 1. Upper bound on $\mathbf{E}\left[\|\Theta_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}\right]$.

Since $\bar{X}_s^1,\dots,\bar{X}_s^N$ are i.i.d. with law $\bar{\mu}_s$ (see (2.3.2)), using (2.3.5), the fact that $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-L,\gamma}(\mathbf{R}^d))$ (see Corollary 2.2), and $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{0,\gamma}(\mathbf{R}^d)$, it holds for all $t \in [0,T]$:

$$\begin{aligned}
\mathbf{E}\left[\langle f, \Theta_t^N\rangle^2\right] = \mathbf{E}\Big[\Big\langle f, \sqrt{N}(\bar{\mu}_t^N - \bar{\mu}_t)\Big\rangle^2\Big] &= \mathbf{E}\Big[\Big|\frac{1}{\sqrt{N}}\sum_{i=1}^N \big[f(\bar{X}_t^i) - \langle f, \bar{\mu}_t\rangle\big]\Big|^2\Big] \\
&= \frac{1}{N}\sum_{i=1}^N \mathbf{E}\Big[\Big|f(\bar{X}_t^i) - \langle f, \bar{\mu}_t\rangle\Big|^2\Big] \\
&\leq \frac{2}{N}\sum_{i=1}^N \mathbf{E}\big[|f(\bar{X}_t^i)|^2\big] + |\langle f, \bar{\mu}_t\rangle|^2 \leq C\|f\|^2_{\mathcal{H}^{L,\gamma}}.
\end{aligned}$$

Taking $f = f_a$ in the previous inequality, summing over $a \in \mathbf{N}^*$, and using the fact that $\mathcal{H}^{J_1,j_1}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$, one deduces that:

$$\sup_{N\geq 1}\sup_{t\in[0,T]} \mathbf{E}\left[\|\Theta_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}\right] \leq C. \tag{2.3.9}$$

Step 2. Upper bound on $\mathbf{E}\left[\|\Upsilon_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}\right]$.

Recall that a.s. $\Upsilon^N \in \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{1-J_1,j_1}(\mathbf{R}^d))$. Thus, a.s. $\Upsilon^N \in \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_1,j_1}(\mathbf{R}^d))$. We then have for all $t \in \mathbf{R}_+$,

$$\|\Upsilon_t^N\|^2_{\mathcal{H}^{-J_1,j_1}} = \sum_{a=1}^{+\infty}\langle f_a, \Upsilon_t^N\rangle^2. \tag{2.3.10}$$

For all $i \in \{1,\dots,N\}$, a.s. $t \in \mathbf{R}_+ \mapsto f(\bar{X}_t^i) \in \mathcal{C}^1(\mathbf{R}_+)$. Indeed, $f \in \mathcal{C}^1(\mathbf{R}^d)$ and a.s. $t \in \mathbf{R}_+ \mapsto \bar{X}_t^i$ (see (2.3.2)) is $\mathcal{C}^1$ (because a.s. $s \in \mathbf{R}_+ \mapsto \alpha\int_{\mathcal{X}\times\mathcal{Y}}(y - \langle\sigma_*(\cdot,x),\bar{\mu}_s\rangle)\nabla_W\sigma_*(\bar{X}_s^i,x)\pi(\mathrm{d}x,\mathrm{d}y)$ is continuous by the dominated convergence theorem). Therefore, it holds $\langle f, \bar{\mu}_t^N\rangle = \langle f, \bar{\mu}_0^N\rangle + \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\langle f, \bar{\mu}_s^N\rangle \mathrm{d}s$ and therefore,

$$\langle f, \bar{\mu}_t^N\rangle = \langle f, \bar{\mu}_0^N\rangle + \int_0^t\int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle\sigma_*(\cdot,x),\bar{\mu}_s\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x), \bar{\mu}_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s. \tag{2.3.11}$$

Thus, by definition of Υ_t^N (see (2.3.6)) and using also (2.2.8), we have:

$$\begin{aligned}
\langle f, \Upsilon_t^N\rangle = \sqrt{N}\underbrace{\langle f, (\mu_0^N - \bar{\mu}_0^N)\rangle}_{=0} &+ \sqrt{N}\int_0^t\int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\mu_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&- \sqrt{N}\int_0^t\int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle\sigma_*(\cdot,x),\bar{\mu}_s\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar{\mu}_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&+ \sqrt{N}\langle f, M_t^N\rangle + \sqrt{N}\langle f, V_t^N\rangle + \sqrt{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f, R_k^N\rangle + \frac{\sqrt{N}}{N^{1+\beta}}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i.
\end{aligned} \tag{2.3.12}$$

Furthermore, it holds,

$$\begin{aligned}
&\sqrt{N}(y-\langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\mu_s^N\rangle\\
&=(y-\langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\Upsilon_s^N\rangle-\langle\sigma_*(\cdot,x),\Upsilon_s^N\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle\\
&\quad+\sqrt{N}(y-\langle\sigma_*(\cdot,x),\bar\mu_s\rangle-\langle\sigma_*(\cdot,x),\bar\mu_s^N-\bar\mu_s\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle. \qquad (2.3.13)
\end{aligned}$$

Consequently, plugging (2.3.13) in (2.3.12), we obtain for $t\in\mathbf{R}_+$:

$$\begin{aligned}
\langle f,\Upsilon_t^N\rangle&=\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha(y-\langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\Upsilon_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad-\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha\langle\sigma_*(\cdot,x),\Upsilon_s^N\rangle\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad-\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha\langle\sigma_*(\cdot,x),\sqrt{N}(\bar\mu_s^N-\bar\mu_s)\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad+\sqrt{N}\langle f,M_t^N\rangle+\sqrt{N}\langle f,V_t^N\rangle+\sqrt{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,R_k^N\rangle+\frac{\sqrt{N}}{N^{1+\beta}}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i. \qquad (2.3.14)
\end{aligned}$$

By Lemma 2.44 (in Section 2.5, see also Remark 2.45), one then has for all $t\in\mathbf{R}_+$:

$$\langle f,\Upsilon_t^N\rangle^2\le\mathbf{A}_t^N[f]+\mathbf{B}_t^N[f], \qquad (2.3.15)$$

where

$$\begin{aligned}
\mathbf{A}_t^N[f]&=2\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha\langle f,\Upsilon_s^N\rangle(y-\langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\Upsilon_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad-2\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha\langle f,\Upsilon_s^N\rangle\langle\sigma_*(\cdot,x),\Upsilon_s^N\rangle\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad-2\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha\langle f,\Upsilon_s^N\rangle\langle\sigma_*(\cdot,x),\sqrt{N}(\bar\mu_s^N-\bar\mu_s)\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&=:\mathbf{I}_t^N[f]+\mathbf{J}_t^N[f]+\mathbf{K}_t^N[f],
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{B}_t^N[f]&=\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big[2\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle\sqrt{N}\langle f,M_k^N\rangle+4N\langle f,M_k^N\rangle^2\Big]+\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big[2\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle\sqrt{N}\langle f,R_k^N\rangle+4N|\langle f,R_k^N\rangle|^2\Big]\\
&\quad+\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big[\frac{2\sqrt{N}}{N^{1+\beta}}\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i+\frac{4}{N^{1+2\beta}}\big|\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i\big|^2\Big]\\
&\quad+\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big[2\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle\mathbf{a}_k^N[f]+4|\mathbf{a}_k^N[f]|^2\Big]-2\sqrt{N}\int_0^t\langle f,\Upsilon_s^N\rangle\mathbf{L}_s^N[f]\mathrm{d}s,
\end{aligned}$$

with

$$\mathbf{L}_s^N[f]=\int_{\mathcal{X}\times\mathcal{Y}}\alpha(y-\langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\mu_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\ \text{ and }\ \mathbf{a}_k^N[f]=\sqrt{N}\int_{\frac{k}{N}}^{\frac{k+1}{N}}\mathbf{L}_s^N[f]\mathrm{d}s. \qquad (2.3.16)$$

Using (2.3.15) and (2.3.10),

$$\|\Upsilon_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}=\sum_{a=1}^{+\infty}\mathbf{A}_t^N[f_a]+\sum_{a=1}^{+\infty}\mathbf{B}_t^N[f_a]. \qquad (2.3.17)$$

By Lemma 2.42, detailed in Section 2.5, and since $\beta \geq 3/4$, one has for all $f \in \mathcal{H}^{J_1,j_1}(\mathbf{R}^d)$ and $t \in \mathbf{R}_+$,

$$\mathbf{E}\big[\mathbf{B}_t^N[f]\big] \leq C\big(\|f\|^2_{\mathcal{H}^{L,\gamma}} + \mathbf{E}\left[\int_0^t \langle f, \Upsilon_s^N\rangle^2 \mathrm{d}s\right]\big).$$

Thus, recalling $\mathcal{H}^{J_1,j_1}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ and (2.3.10), we obtain,

$$\sum_{a=1}^{+\infty} \mathbf{E}[\mathbf{B}_t^N[f_a]] \leq C + C\int_0^t \mathbf{E}\big[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}\big]\mathrm{d}s. \tag{2.3.18}$$

Let us now provide a similar upper bound on $\sum_{a=1}^{+\infty} \mathbf{A}_t^N[f_a]$. By (2.3.5) and because $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$ (see (2.1.6)),

$$|\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \bar{\mu}_s^N\rangle| = \Big|\frac{1}{N}\sum_{i=1}^N \nabla f(\bar{X}_s^i)\cdot\nabla\sigma_*(\bar{X}_s^i, x)\Big| \leq \frac{C\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N}\sum_{i=1}^N (1 + |\bar{X}_s^i|^{\gamma_*}) \leq C\|f\|_{\mathcal{H}^{L,\gamma}}. \tag{2.3.19}$$

Then, using **A2** and since $j_1 > d/2$, it holds: we have using (2.3.19):

$$\begin{aligned}
&-2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\langle f, \Upsilon_s^N\rangle\langle\sigma_*(\cdot,x), \Upsilon_s^N\rangle\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x), \bar{\mu}_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\leq 4\alpha\int_0^t\int_{\mathcal{X}\times\mathcal{Y}} \big(\langle f, \Upsilon_s^N\rangle^2 + C\|f\|^2_{\mathcal{H}^{L,\gamma}}\|\sigma_*(W,x)\|^2_{\mathcal{H}^{J_1,j_1}}\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}\big)\,\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\leq C\int_0^t \big(\langle f, \Upsilon_s^N\rangle^2 + C\|f\|^2_{\mathcal{H}^{L,\gamma}}\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}\big)\,\mathrm{d}s.
\end{aligned}$$

Thus, $\sum_{a=1}^{+\infty}\mathbf{E}[\mathbf{J}_t^N[f_a]] \leq C\int_0^t \mathbf{E}\big[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}\big]\mathrm{d}s$. Let us now study $\sum_{a=1}^{+\infty}\mathbf{K}_t^N[f_a]$. Since $\bar{X}_s^1,\ldots,\bar{X}_s^N$ are i.i.d. with law $\bar{\mu}_s$ (see (2.3.2)) and because σ_* is bounded (see **A2**), we have:

$$\begin{aligned}
\mathbf{E}\left[\big\langle\sigma_*(\cdot,x), \bar{\mu}_s^N - \bar{\mu}_s\big\rangle^2\right] &= \mathbf{E}\Big[\Big|\frac{1}{N}\sum_{i=1}^N \sigma_*(\bar{X}_s^i, x) - \langle\sigma_*(\cdot,x),\bar{\mu}_s\rangle\Big|^2\Big] \\
&= \frac{1}{N^2}\sum_{i=1}^N \mathbf{E}\left[(\sigma_*(\bar{X}_s^i,x) - \langle\sigma_*(\cdot,x),\bar{\mu}_s\rangle)^2\right] \leq \frac{C}{N}.
\end{aligned} \tag{2.3.20}$$

Hence, $\mathbf{E}[\langle\sigma_*(\cdot,x), \sqrt{N}(\bar{\mu}_s^N - \bar{\mu}_s)\rangle^2] \leq C$. Thus, using in addition (2.3.19), one deduces that

$$\begin{aligned}
&\mathbf{E}\left[-2\int_0^t\int_{\mathcal{X}\times\mathcal{Y}} \alpha\langle f,\Upsilon_s^N\rangle\langle\sigma_*(\cdot,x),\sqrt{N}(\bar{\mu}_s^N-\bar{\mu}_s)\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar{\mu}_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\right] \\
&\leq C\int_0^t\int_{\mathcal{X}\times\mathcal{Y}} \Big\{\mathbf{E}[\langle f,\Upsilon_s^N\rangle^2] + \mathbf{E}\left[\langle\sigma_*(\cdot,x),\sqrt{N}(\bar{\mu}_s^N-\bar{\mu}_s)\rangle^2\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar{\mu}_s^N\rangle^2\right]\Big\}\,\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\leq C\int_0^t\int_{\mathcal{X}\times\mathcal{Y}} \Big\{\mathbf{E}[\langle f,\Upsilon_s^N\rangle^2] + \mathbf{E}\left[\langle\sigma_*(\cdot,x),\sqrt{N}(\bar{\mu}_s^N-\bar{\mu}_s)\rangle^2\right]\|f\|^2_{\mathcal{H}^{L,\gamma}}\Big\}\,\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\leq C\int_0^t \big\{\mathbf{E}[\langle f,\Upsilon_s^N\rangle^2] + \|f\|^2_{\mathcal{H}^{L,\gamma}}\big\}\,\mathrm{d}s.
\end{aligned}$$

Therefore, $\sum_{a=1}^{+\infty}\mathbf{E}[\mathbf{K}_t^N[f_a]] \leq C + C\int_0^t \mathbf{E}\big[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}\big]\mathrm{d}s$. It remains to study $\sum_{a=1}^{+\infty}\mathbf{I}_t^N[f_a]$. To this end, for $x \in \mathcal{X}$, introduce the bounded linear operator

$$\mathsf{T}_x : f \in \mathcal{H}^{J_1,j_1}(\mathbf{R}^d) \mapsto \nabla f\cdot\nabla\sigma_*(\cdot,x) \in \mathcal{H}^{J_1-1,j_1}(\mathbf{R}^d). \tag{2.3.21}$$

Then, one has:

$$\sum_{a=1}^{+\infty} \mathbf{I}_t^N[f_a] = \sum_{a=1}^{+\infty} 2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\langle f_a, \Upsilon_s^N\rangle(y - \langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle \mathsf{T}_x f_a, \Upsilon_s^N\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$= 2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle\sigma_*(\cdot,x),\mu_s^N\rangle) \sum_{a=1}^{+\infty}\langle f_a, \Upsilon_s^N\rangle\langle \mathsf{T}_x f_a, \Upsilon_s^N\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$= 2\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle\sigma_*(\cdot,x),\mu_s^N\rangle) \left\langle \Upsilon_s^N, \mathsf{T}_x^*\Upsilon_s^N\right\rangle_{-J_1,j_1} \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s.$$

Since σ_* and $\mathcal{Y}$ are bounded, this implies that:

$$\sum_{a=1}^{+\infty} \mathbf{E}[\mathbf{I}_t^N[f_a]] \le C\int_0^t \mathbf{E}\Big[\int_{\mathcal{X}\times\mathcal{Y}} |y - \langle\sigma_*(\cdot,x),\mu_s^N\rangle|\,\big|\langle \Upsilon_s^N, \mathsf{T}_x^*\Upsilon_s^N\rangle_{-J_1,j_1}\big|\pi(\mathrm{d}x,\mathrm{d}y)\Big]\mathrm{d}s$$
$$\le C\int_0^t \mathbf{E}\Big[\int_{\mathcal{X}\times\mathcal{Y}} (|y| + C)\,|\langle \Upsilon_s^N, \mathsf{T}_x^*\Upsilon_s^N\rangle_{-J_1,j_1}|\pi(\mathrm{d}x,\mathrm{d}y)\Big]\mathrm{d}s$$

By Lemma 2.43, detailed in Section 2.5, and since a.s. $\Upsilon^N \in \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{1-J_1,j_1}(\mathbf{R}^d))$, there exists $C>0$ such that for all $x\in\mathcal{X}$, $|\langle \Upsilon_s^N, \mathsf{T}_x^*\Upsilon_s^N\rangle_{\mathcal{H}^{-J_1,j_1}}| \le C\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}$. Hence, since $\mathbf{E}[|y|]<+\infty$, we deduce that:

$$\sum_{a=1}^{+\infty} \mathbf{E}[\mathbf{I}_t^N[f_a]] \le C\int_0^t \mathbf{E}\left[\left\|\Upsilon_s^N\right\|^2_{\mathcal{H}^{-J_1,j_1}}\right]\mathrm{d}s.$$

We have thus proved that $\sum_{a=1}^{+\infty}\mathbf{E}[\mathbf{A}_t^N[f_a]] \le C + C\int_0^t \mathbf{E}[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}]\mathrm{d}s$. In conclusion, using also (2.3.18) and (2.3.17), we have $\mathbf{E}[\|\Upsilon_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}] \le C + C\int_0^t \mathbf{E}[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}]\mathrm{d}s$. By Gronwall's Lemma, we get:

$$\sup_{N\ge1}\sup_{t\in[0,T]} \mathbf{E}[\|\Upsilon_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}] < +\infty.$$

Together with the first step, this ends the proof of the Lemma (recall the decomposition $\eta^N = \Upsilon^N + \Theta^N$, see (2.3.7)). □

Lemma 2.24. *Assume* **A1**-**A7**. *Introduce the following* $\boldsymbol{\sigma}$*-algebra (see* (2.1.9)*):*

$$\mathfrak{F}_t^N := \mathcal{F}_{\lfloor Nt\rfloor}^N,\ t\in\mathbf{R}_+.$$

Then, for all $f\in\mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$*, the two processes*

$$\Big\{t\mapsto \sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N} \nabla f(W_k^i)\cdot\varepsilon_k^i,\ t\in\mathbf{R}_+\Big\} \text{ and } \big\{t\mapsto \langle f, M_t^N\rangle, t\in\mathbf{R}_+\big\} \text{ are } \mathfrak{F}_t^N\text{-martingale.} \quad (2.3.22)$$

Proof. Recall that by Lemma 2.11, the first process in (2.3.22) is integrable. By (2.2.19) and (2.2.6), the second process in (2.3.22) is integrable. For $0\le s<t$, we write

$$\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N} \nabla f(W_k^i)\cdot\varepsilon_k^i|\mathfrak{F}_s^N\Big] = \sum_{k=0}^{\lfloor Ns\rfloor-1}\sum_{i=1}^{N} \mathbf{E}\big[\nabla f(W_k^i)\cdot\varepsilon_k^i|\mathcal{F}_{\lfloor Ns\rfloor}^N\big] + \mathfrak{R}_{t,s}^N[f], \quad (2.3.23)$$

where

$$\text{if } \lfloor Nt\rfloor = \lfloor Ns\rfloor\text{: } \mathfrak{R}_{t,s}^N[f]=0, \text{ and if } \lfloor Nt\rfloor > \lfloor Ns\rfloor\text{: } \mathfrak{R}_{t,s}^N[f] = \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N} \mathbf{E}[\nabla f(W_k^i)\cdot\varepsilon_k^i|\mathcal{F}_{\lfloor Ns\rfloor}^N].$$

When $\lfloor Nt\rfloor > \lfloor Ns\rfloor$, since the W_k^i's are $\mathcal{F}_k^N$-measurable (see (2.1.2) and (2.1.9)) and the ε_k^i's are centered and independent of $\mathcal{F}_k^N$ (see **A5**), we have, for $k \geq \lfloor Ns\rfloor$ and $i \in \{1,\dots,N\}$:

$$\mathbf{E}\big[\nabla f(W_k^i)\cdot\varepsilon_k^i|\mathcal{F}_{\lfloor Ns\rfloor}^N\big] = \mathbf{E}\Big[\nabla f(W_k^i)\cdot\mathbf{E}[\varepsilon_k^i|\mathcal{F}_k^N]\Big|\mathcal{F}_{\lfloor Ns\rfloor}^N\Big] = 0.$$

Hence, for all $0 \leq s < t$, $\mathfrak{R}_{t,s}^N[f] = 0$. Furthermore, for $0 \leq k \leq \lfloor Ns\rfloor - 1$ and $i \in \{1,\dots,N\}$, $\nabla f(W_k^i)\cdot\varepsilon_k^i$ is $\mathcal{F}_{k+1}^N$-measurable and thus is $\mathcal{F}_{\lfloor Ns\rfloor}^N$-measurable. Therefore, (2.3.23) reduces to

$$\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i\Big|\mathfrak{F}_s^N\Big] = \sum_{k=0}^{\lfloor Ns\rfloor-1}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i.$$

This proves that $\{t\mapsto\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i,\ t\in\mathbf{R}_+\}$ is a $\mathfrak{F}_t^N$-martingale. Let us now prove that the process $\big\{t\mapsto\langle f, M_t^N\rangle, t\in\mathbf{R}_+\big\}$ is a $\mathfrak{F}_t^N$-martingale (see (2.2.6)). We have, for $0\leq s<t$,

$$\mathbf{E}\big[\langle f, M_t^N\rangle|\mathfrak{F}_s^N\big] = \sum_{k=0}^{\lfloor Ns\rfloor-1}\mathbf{E}\big[\langle f, M_k^N\rangle\big|\mathcal{F}_{\lfloor Ns\rfloor}^N\big] + \mathfrak{E}_{t,s}^N[f],$$

where

$$\text{if } \lfloor Nt\rfloor = \lfloor Ns\rfloor\text{: } \mathfrak{E}_{t,s}^N[f]=0, \text{ and if } \lfloor Nt\rfloor > \lfloor Ns\rfloor\text{: } \mathfrak{E}_{t,s}^N[f] = \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\mathbf{E}\big[\langle f, M_k^N\rangle|\mathcal{F}_{\lfloor Ns\rfloor}^N\big].$$

When $\lfloor Nt\rfloor > \lfloor Ns\rfloor$, we have for $k\geq\lfloor Ns\rfloor$, by (2.2.24):

$$\mathbf{E}\big[\langle f, M_k^N\rangle|\mathcal{F}_{\lfloor Ns\rfloor}^N\big] = \mathbf{E}\Big[\mathbf{E}\big[\langle f, M_k^N\rangle|\mathcal{F}_k^N\big]\Big|\mathcal{F}_{\lfloor Ns\rfloor}^N\Big] = 0.$$

Hence, for all $0\leq s<t$, $\mathfrak{E}_{t,s}^N[f]=0$. In addition, for $k\leq\lfloor Ns\rfloor-1$, $\langle f, M_k^N\rangle$ is $\mathcal{F}_{k+1}^N$-measurable and thus is $\mathcal{F}_{\lfloor Ns\rfloor}^N$-measurable. In conclusion, $\big\{t\mapsto\langle f, M_t^N\rangle, t\in\mathbf{R}_+\big\}$ is a $\mathfrak{F}_t^N$-martingale. This ends the proof of Lemma 2.24. □

The following lemma provides the compact containment condition needed to prove that $(\eta^N)_{N\geq1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$.

Lemma 2.25. *Let $\beta\geq 3/4$ and assume* **A1**-**A7**. *Then, for all $T>0$,*

$$\sup_{N\geq1}\mathbf{E}\Big[\sup_{t\in[0,T]}\|\eta_t^N\|_{\mathcal{H}^{-J_2,j_2}}^2\Big] < +\infty. \tag{2.3.24}$$

Proof. Let $T>0$. All along the proof, $C>0$ denotes a constant indendent of $t\in[0,T]$ and $N\geq1$, which can change from one occurence to another. Recall that $\eta^N = \Upsilon^N + \Theta^N$, see (2.3.7). By (2.3.14) and Jensen's inequality, it holds for $f\in\mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$ (recall $\mathcal{H}^{J_2,j_2}(\mathbf{R}^d)\hookrightarrow\mathcal{H}^{J_1,j_1}(\mathbf{R}^d)$,

see (2.1.14)),

$$\sup_{t\in[0,T]}\langle f,\Upsilon_t^N\rangle^2 \le C\Bigg[\int_0^T\int_{\mathcal{X}\times\mathcal{Y}}\left|(y-\langle\sigma_*(\cdot,x),\mu_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\Upsilon_s^N\rangle\right|^2\pi(\mathrm{d}s,\mathrm{d}y)\mathrm{d}s$$
$$+\int_0^T\int_{\mathcal{X}\times\mathcal{Y}}\left|\langle\sigma_*(\cdot,x),\Upsilon_s^N\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle\right|^2\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$+\int_0^T\int_{\mathcal{X}\times\mathcal{Y}}\left|\langle\sigma_*(\cdot,x),\sqrt{N}(\bar\mu_s^N-\bar\mu_s)\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle\right|^2\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$+N\sup_{t\in[0,T]}|\langle f,M_t^N\rangle|^2+N\sup_{t\in[0,T]}\left|\langle f,V_t^N\rangle\right|^2+N\sup_{t\in[0,T]}\left|\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,R_k^N\rangle\right|^2$$
$$+\frac{1}{N^{1+2\beta}}\sup_{t\in[0,T]}\left|\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N\nabla f(W_k^i)\cdot\varepsilon_k^i\right|^2\Bigg]. \tag{2.3.25}$$

We now consider successively each term in the right-hand-side of (2.3.25). By Lemma 2.23, for $0\le s\le T$, we have using also **A2** and $\mathbf{E}[|y|^2]<+\infty$ (see **A3**):

$$\begin{aligned}&\mathbf{E}\Big[\int_{\mathcal{X}\times\mathcal{Y}}|y-\langle\sigma_*(\cdot,x),\mu_s^N\rangle|^2\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\Upsilon_s^N\rangle^2\pi(\mathrm{d}x,\mathrm{d}y)\Big]\\&\le C\mathbf{E}\Big[\int_{\mathcal{X}\times\mathcal{Y}}(|y|^2+1)\|\nabla f\cdot\nabla\sigma_*(\cdot,x)\|_{\mathcal{H}^{J_1,j_1}}\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}\pi(\mathrm{d}x,\mathrm{d}y)\Big]\\&\le C\|f\|^2_{\mathcal{H}^{J_1+1,j_1}}\mathbf{E}[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}]\le C\|f\|^2_{\mathcal{H}^{J_1+1,j_1}}.\end{aligned} \tag{2.3.26}$$

Let us now study the second term in (2.3.25). By (2.3.19) and since $\sup_{x\in\mathcal{X}}\|\sigma_*(\cdot,x)\|<+\infty$ (because $j_1>d/2$ and $\sigma_*\in\mathcal{C}_b^\infty(\mathbf{R}^d\times\mathcal{X})$ by **A2**), for $0\le s\le T$,

$$\mathbf{E}\left[\left|\langle\sigma_*(\cdot,x),\Upsilon_s^N\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle\right|^2\right]\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}\mathbf{E}\left[\|\Upsilon_s^N\|^2_{\mathcal{H}^{-J_1,j_1}}\right]\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}. \tag{2.3.27}$$

Let us now consider the third term in (2.3.25). By (2.3.19) and (2.3.20), we have for $0\le s\le T$,

$$\mathbf{E}\Big[\langle\sigma_*(\cdot,x),\sqrt{N}(\bar\mu_s^N-\bar\mu_s)\rangle^2\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\bar\mu_s^N\rangle^2\Big]\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}. \tag{2.3.28}$$

Let us now deal with the fourth term in (2.3.25). By Lemma 2.24, we have using Doob's maximal inequality, $\mathbf{E}[\sup_{t\in[0,T]}\langle f,M_t^N\rangle^2]\le C\mathbf{E}[\langle f,M_T^N\rangle^2]$. Then by Lemma 2.12 and since $\mathcal{H}^{L,\gamma}(\mathbf{R}^d)\hookrightarrow\mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$ (recall indeed (2.1.6)), we obtain

$$N\mathbf{E}\Big[\sup_{t\in[0,T]}\langle f,M_t^N\rangle^2\Big]\le C\|f\|^2_{\mathcal{C}^{2,\gamma_*}}\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}. \tag{2.3.29}$$

Using (2.2.33) and again (2.1.6), the fifth term in (2.3.25) satisfies:

$$N\mathbf{E}\Big[\sup_{t\in[0,T]}|\langle f,V_t^N\rangle|^2\Big]\le CN^{-1/2}\|f\|^2_{\mathcal{H}^{L,\gamma}}. \tag{2.3.30}$$

Using (2.2.35), the sixth term in (2.3.25) satisfies:

$$\mathbf{E}\Big[N\sup_{t\in[0,T]}\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,R_k^N\rangle\Big|^2\Big]\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}(1/N+N^3/N^{4\beta}).$$

Let us deal with the last term in the right-hand side of (2.3.25) for which we need a more accurate upper bound than (2.2.36). By Lemma 2.24 and Doob's maximal inequality, we have using (2.2.30) and $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$,

$$\frac{1}{N^{1+2\beta}}\mathbf{E}\Big[\sup_{t\in[0,T]}\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big] \le \frac{C}{N^{1+2\beta}}\mathbf{E}\Big[\Big|\sum_{k=0}^{\lfloor NT\rfloor-1}\sum_{i=1}^{N}\nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big] \le C\|f\|^2_{\mathcal{H}^{L,\gamma}}N^{1-2\beta}.$$

Collecting these bounds, we obtain, from (2.3.25), for $f \in \mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$,

$$\mathbf{E}\Big[\sup_{t\in[0,T]}\langle f,\Upsilon_t^N\rangle^2\Big] \le C\left(\|f\|^2_{\mathcal{H}^{J_1+1,j_1}} + \|f\|^2_{\mathcal{H}^{L,\gamma}}\right). \tag{2.3.31}$$

We now turn to the study of $\mathbf{E}[\sup_{t\in[0,T]}\langle f,\Theta_t^N\rangle^2]$ for $f \in \mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$. By (2.3.11) and Corollary 2.2, we have, for all $t \in [0,T]$ (recall that $\Theta_t^N = \sqrt{N}(\bar\mu_t^N - \bar\mu_t)$, see (2.3.6)),

$$\langle f,\Theta_t^N\rangle = \langle f,\Theta_0^N\rangle + \int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\alpha(y-\langle\sigma_*(\cdot,x),\bar\mu_s\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\Theta_s^N\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s. \tag{2.3.32}$$

By Jensen's inequality together with **A2** and (2.3.9), one has:

$$\begin{aligned}\mathbf{E}\Big[\sup_{t\in[0,T]}\langle f,\Theta_t^N\rangle^2\Big] &\le C\mathbf{E}\big[|\langle f,\Theta_0^N\rangle|^2\big] + C\int_0^T\int_{\mathcal{X}\times\mathcal{Y}}(|y|^2+1)\mathbf{E}\left[\langle\nabla f\cdot\nabla\sigma_*(\cdot,x),\Theta_s^N\rangle^2\right]\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\ &\le C\|f\|^2_{\mathcal{H}^{J_1,j_1}}\mathbf{E}\big[\|\Theta_0^N\|^2_{\mathcal{H}^{-J_1,j_1}}\big]\\ &\quad + C\int_0^T\int_{\mathcal{X}\times\mathcal{Y}}(|y|^2+1)\,\|\nabla f\cdot\nabla\sigma_*(\cdot,x)\|^2_{\mathcal{H}^{J_1,j_1}}\,\mathbf{E}\left[\|\Theta_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}\right]\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\ &\le C\|f\|^2_{\mathcal{H}^{J_1+1,j_1}}.\end{aligned} \tag{2.3.33}$$

Let $\{f_a\}_{a\ge1}$ be an orthonormal basis of $\mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$. Let us recall that (see (2.1.14)) $\mathcal{H}^{J_2,j_2}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{J_1+1,j_1}(\mathbf{R}^d)$ and $\mathcal{H}^{J_1+1,j_1}(\mathbf{R}^d) \hookrightarrow \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$. Then, by (2.3.31) and (2.3.33), we obtain, since $\beta \ge 3/4$,

$$\begin{aligned}\mathbf{E}\Big[\sup_{t\in[0,T]}\|\eta_t^N\|^2_{\mathcal{H}^{-J_2,j_2}}\Big] &= \mathbf{E}\Big[\sup_{t\in[0,T]}\sum_{a\ge1}\langle f_a,\eta_t^N\rangle^2\Big] \le \sum_{a\ge1}\mathbf{E}\Big[\sup_{t\in[0,T]}\langle f_a,\eta_t^N\rangle^2\Big]\\ &\le 2\sum_{a\ge1}\Big(\mathbf{E}\Big[\sup_{t\in[0,T]}\langle f_a,\Upsilon_t^N\rangle^2\Big] + \mathbf{E}\Big[\sup_{t\in[0,T]}\langle f_a,\Theta_t^N\rangle^2\Big]\Big)\\ &\le C\sum_{a\ge1}\left(\|f_a\|^2_{\mathcal{H}^{J_1+1,j_1}} + \|f_a\|^2_{\mathcal{H}^{L,\gamma}}\right) \le C.\end{aligned}$$

This concludes the proof of the lemma. □

The following result provides the regularity condition needed to prove that $(\eta^N)_{N\ge1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+,\mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$.

Lemma 2.26. *Let $\beta \ge 3/4$ and assume* **A1**-**A7**. *Let $T>0$. Then, there exists $C>0$ such that for all $\delta>0$ and $0\le r<t\le T$ such that $t-r\le\delta$, one has*

$$\mathbf{E}\left[\left\|\eta_t^N-\eta_r^N\right\|^2_{\mathcal{H}^{-J_2,j_2}}\right] \le C\left[\delta^2 + \frac{N\delta+1}{N} + \frac{1}{\sqrt{N}} + (N\delta+1)^2\Big(\frac{1}{N^3}+\frac{1}{N^{4\beta-1}}\Big) + \frac{N\delta+1}{N^{2\beta}}\right]. \tag{2.3.34}$$

Proof. Let $T > 0$. In the following, $C > 0$ is a constant independent of $\delta > 0$, $0 \le r < t \le T$, $N \ge 1$, and $f \in \mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$ which can change from one occurence to another. In what follows $t - r \le \delta$. Recall $\eta^N = \Upsilon^N + \Theta^N$, see (2.3.7). Using (2.3.14) and the Jensen's inequality, one has:

$$\begin{aligned}\left|\langle f, \Upsilon_t^N\rangle - \langle f, \Upsilon_r^N\rangle\right|^2 \le C\Big[&(t-r)\int_r^t \int_{\mathcal{X}\times\mathcal{Y}} \left|(y - \langle \sigma_*(\cdot,x), \mu_s^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \Upsilon_s^N\rangle\right|^2 \pi(\mathrm{d}s,\mathrm{d}y)\mathrm{d}s \\ &+ (t-r)\int_r^t \int_{\mathcal{X}\times\mathcal{Y}} \left|\langle \sigma_*(\cdot,x), \Upsilon_s^N\rangle\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \bar{\mu}_s^N\rangle\right|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &+ (t-r)\int_r^t \int_{\mathcal{X}\times\mathcal{Y}} \left|\langle \sigma_*(\cdot,x), \sqrt{N}(\bar{\mu}_s^N - \bar{\mu}_s)\rangle\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \bar{\mu}_s^N\rangle\right|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &+ N|\langle f, M_t^N - M_r^N\rangle|^2 + N\left|\langle f, V_t^N\rangle - \langle f, V_r^N\rangle\right|^2 + N\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor - 1}\langle f, R_k^N\rangle\Big|^2 \\ &+ \frac{1}{N^{1+2\beta}}\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big]. \end{aligned} \tag{2.3.35}$$

We now study each term of the right-hand side of (2.3.35). By (2.3.26), we bound the first term in (2.3.35) as follows:

$$\mathbf{E}\Big[(t-r)\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\left|(y - \langle \sigma_*(\cdot,x), \mu_s^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \Upsilon_s^N\rangle\right|^2 \pi(\mathrm{d}s,\mathrm{d}y)\mathrm{d}s\Big] \le C\delta^2\|f\|^2_{\mathcal{H}^{J_1+1,j_1}}.$$

Using (2.3.27), we bound the second term of (2.3.35) as follows:

$$\mathbf{E}\Big[(t-r)\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\left|\langle \sigma_*(\cdot,x), \Upsilon_s^N\rangle\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \bar{\mu}_s^N\rangle\right|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big] \le C\delta^2\|f\|^2_{\mathcal{H}^{L,\gamma}}.$$

Using (2.3.28), we have the following bound on the third term of (2.3.35):

$$\mathbf{E}\Big[(t-r)\int_r^t\int_{\mathcal{X}\times\mathcal{Y}}\left|\langle \sigma_*(\cdot,x), \sqrt{N}(\bar{\mu}_s^N - \bar{\mu}_s)\rangle\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \bar{\mu}_s^N\rangle\right|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\Big] \le C\delta^2\|f\|^2_{\mathcal{H}^{L,\gamma}}.$$

In addition, we have, using (2.2.26), (2.2.19), and $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$ (see (2.1.6)),

$$\begin{aligned} N\mathbf{E}[|\langle f, M_t^N - M_r^N\rangle|^2] &= N\mathbf{E}\Big[\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\langle f, M_k^N\rangle\Big|^2\Big] = N\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\mathbf{E}\Big[\langle f, M_k^N\rangle^2\Big] \\ &\le CN(\lfloor Nt\rfloor - \lfloor Nr\rfloor)\|f\|^2_{\mathcal{C}^{2,\gamma_*}}/N^2 \le C(N\delta+1)\|f\|^2_{\mathcal{H}^{L,\gamma}}/N. \end{aligned} \tag{2.3.36}$$

The fifth term of (2.3.35) is bounded as follows using (2.3.30):

$$\mathbf{E}\left[N|\langle f, V_t^N\rangle - \langle f, V_r^N\rangle|^2\right] \le 2N\mathbf{E}[|\langle f, V_t^N\rangle|^2] + 2N\mathbf{E}[|V_r^N[f]|^2] \le C\|f\|^2_{\mathcal{H}^{L,\gamma}}/\sqrt{N}.$$

Let us consider the sixth term in the right-hand side of (2.3.35). By (3.9.3), (2.2.34), and because $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$, we have that:

$$\mathbf{E}\Big[N\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\langle f, R_k^N\rangle\Big|^2\Big] \le N(\lfloor Nt\rfloor - \lfloor Nr\rfloor)\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\mathbf{E}\left[|\langle f, R_k^N\rangle|^2\right] \le CN(N\delta+1)^2\|f\|^2_{\mathcal{H}^{L,\gamma}}(1/N^4 + 1/N^{4\beta}).$$

Let us consider the last term in the right-hand side of (2.3.35). Using (2.2.28) and (2.2.29), we have:

$$\begin{aligned}\mathbf{E}\Big[\frac{1}{N^{1+2\beta}}\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N \nabla f(W_k^i)\cdot\varepsilon_k^i\Big|^2\Big] &= \frac{1}{N^{1+2\beta}}\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\sum_{i=1}^N \mathbf{E}\left[|\nabla f(W_k^i)\cdot\varepsilon_k^i|^2\right] \\ &\le \frac{1}{N^{1+2\beta}} \times C\|f\|^2_{\mathcal{H}^{L,\gamma}}N(N\delta+1) = \frac{C\|f\|^2_{\mathcal{H}^{L,\gamma}}}{N^{2\beta}}(N\delta+1).\end{aligned}$$

Let $\{f_a\}_{a\geq 1}$ be an orthonormal basis of $\mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$. Gathering the previous bounds, we obtain, using also (2.1.14),

$$\begin{aligned}\mathbf{E}\big[\|\Upsilon_t^N - \Upsilon_r^N\|^2_{\mathcal{H}^{-J_2,j_2}}\big] &= \sum_{a=1}^{+\infty} \mathbf{E}\left[\big|\langle f_a, \Upsilon_t^N\rangle - \langle f_a, \Upsilon_r^N\rangle\big|^2\right] \\ &\leq C\left[\delta^2 + \frac{N\delta+1}{N} + \frac{1}{\sqrt{N}} + (N\delta+1)^2\Big(\frac{1}{N^3} + \frac{1}{N^{4\beta-1}}\Big) + \frac{N\delta+1}{N^{2\beta}}\right].\end{aligned} \tag{2.3.37}$$

By (2.3.32) and using the same arguments leading to (2.3.33), we obtain for $f \in \mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$,

$$\mathbf{E}\left[|\langle f, \Theta_t^N\rangle - \langle f, \Theta_r^N\rangle|^2\right] \leq C\delta \int_r^t \mathbf{E}\left[\|\nabla f\|^2_{\mathcal{H}^{J_1,j_1}} \|\Theta_t^N\|^2_{\mathcal{H}^{-J_1,j_1}}\right] \mathrm{d}s \leq C\delta^2 \|f\|^2_{\mathcal{H}^{J_1+1,j_1}}.$$

Considering an orthonormal basis of $\mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$, and using the fact that $\mathcal{H}^{J_2,j_2}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{J_1+1,j_1}(\mathbf{R}^d)$, we obtain $\mathbf{E}[\|\Theta_t^N - \Theta_r^N\|^2_{\mathcal{H}^{-J_2,j_2}}] \leq C\delta^2$. Combining this result with (2.3.37), we obtain (2.3.34). This concludes the proof of the lemma. □

Now, we collect the results of Lemmata 2.25 and 2.26 to prove the following result.

Proposition 2.27. *Let $\beta \geq 3/4$ and assume* **A1**-**A7**. *Then, the sequence $(\eta^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$.*

Proof. Recall $\mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{J_2,j_2}(\mathbf{R}^d)$ (by (2.1.14)). Using Markov's inequality, Lemma 2.25 implies item 1 in Proposition 2.41. In addition, according to Lemma 2.26, item 2 in Proposition 2.41 is satisfied. Consequently, $(\eta^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$. The result follows from Proposition 2.41. □

To prove Proposition 2.27, we mention that one could also have used [Jak86, Theorem 4.6] with $E = \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$ and $\mathbb{F} = \{\mathsf{L}_f, f \in \mathcal{C}_c^\infty(\mathbf{R}^d)\}$ where $\mathsf{L}_f : \Phi \in \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d) \mapsto \langle f, \Phi\rangle_{J_0-1,j_0}$.

2.3.2 Relative compactness of $(\sqrt{N}M^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$

We begin this section with the compact containment condition on the sequence $\{t \mapsto \sqrt{N}M_t^N, t \in \mathbf{R}_+\}_{N\geq 1}$ (see (2.2.4) and (2.2.6)).

Lemma 2.28. *Let $\beta \geq 3/4$ and assume* **A1**-**A7**. *Then, for all $T > 0$,*

$$\sup_{N\geq 1} \mathbf{E}\Big[\sup_{t\in[0,T]} \big\|\sqrt{N}M_t^N\big\|^2_{\mathcal{H}^{-J_1,j_1}}\Big] < +\infty.$$

Proof. Let $T > 0$ and $f \in \mathcal{H}^{J_1,j_1}(\mathbf{R}^d)$. Then, according to (2.3.29), we have:

$$\mathbf{E}\Big[\sup_{t\in[0,T]} \langle f, \sqrt{N}M_t^N\rangle^2\Big] \leq C\|f\|^2_{\mathcal{H}^{L,\gamma}}.$$

The proof of the lemma is complete considering an orthonormal basis $\{f_a\}_{a\geq 1}$ of $\mathcal{H}^{J_1,j_1}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ (see (2.1.14)). □

Let us now turn to the regularity condition on the process $\{t \mapsto \sqrt{N}M_t^N, t \in \mathbf{R}_+\}_{N\geq 1}$.

Lemma 2.29. *Let $\beta \geq 3/4$ and assume* **A1**-**A7**. *Fix $T > 0$. Then, there exists $C > 0$ such that for all $\delta > 0$ and $0 \leq r < t \leq T$ such that $t - r \leq \delta$, one has*

$$\mathbf{E}\Big[\Big\|\sqrt{N}M_t^N - \sqrt{N}M_r^N\Big\|^2_{\mathcal{H}^{-J_1,j_1}}\Big] \leq C\frac{N\delta+1}{N}.$$

Proof. This lemma is a direct consequence of (2.3.36) (which also holds for $f \in \mathcal{H}^{J_1,j_1}(\mathbf{R}^d)$) together with the embedding $\mathcal{H}^{J_1,j_1}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ (see (2.1.14)). □

Proposition 2.30. *Let $\beta \geq 3/4$ and assume* **A1**-**A7**. *Then, the sequence $\{t \mapsto \sqrt{N}M_t^N, t \in \mathbf{R}_+\}_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$.*

Proof. It is a direct consequence of Proposition 2.41, Lemmata 2.28 and 2.29, together with the embedding $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{J_1,j_1}(\mathbf{R}^d)$ (see (2.1.14)). □

2.3.3 Regularity of the limit points

Lemma 2.31. *Let $\beta > 3/4$ and assume* **A1**-**A7**. *Then, for all $T > 0$,*

$$\lim_{N\to+\infty} \mathbf{E}\Big[\sup_{t\in[0,T]} \big\|\eta_t^N - \eta_{t^-}^N\big\|^2_{\mathcal{H}^{-J_0+1,j_0}}\Big] = 0. \tag{2.3.38}$$

In particular, any limit point of $(\eta^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$ belongs a.s. to $\mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$.

Proof. Let $T > 0$ and $f \in \mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d)$. We have (see (2.3.7)):

$$\sup_{t\in[0,T]} \|\eta_t^N - \eta_{t^-}^N\|^2_{\mathcal{H}^{-J_0+1,j_0}} \leq 2 \sup_{t\in[0,T]} \|\Upsilon_t^N - \Upsilon_{t^-}^N\|^2_{\mathcal{H}^{-J_0+1,j_0}} + 2 \sup_{t\in[0,T]} \|\Theta_t^N - \Theta_{t^-}^N\|^2_{\mathcal{H}^{-J_0+1,j_0}}.$$

According to (2.3.8) and since $\mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d) \hookrightarrow \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ (see (2.1.14)), one has a.s. for all $t \in \mathbf{R}_+$ and $N \geq 1$,

$$\|\Theta_t^N - \Theta_{t^-}^N\|_{\mathcal{H}^{-J_0+1,j_0}} = 0.$$

Since a.s. $\bar{\mu}^N \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-L,\gamma}(\mathbf{R}^d))$, by definition of Υ^N (see (2.3.6)), it follows that a.s. for all $N \geq 1$,

$$\sup_{t\in[0,T]} \langle f, \Upsilon_t^N - \Upsilon_{t^-}^N\rangle^2 = N \sup_{t\in[0,T]} \langle f, \mu_t^N - \mu_{t^-}^N\rangle^2$$

From (2.2.48) and the fact that $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$, we obtain

$$\mathbf{E}\Big[\sup_{t\in[0,T]} \langle f, \Upsilon_t^N - \Upsilon_{t^-}^N\rangle^2\Big] \leq C\|f\|^2_{\mathcal{H}^{L,\gamma}}\Big[\frac{1}{\sqrt{N}} + \sqrt{\frac{1}{N^5} + \frac{1}{N^{8\beta-3}}} + N^{\frac{3}{2}-2\beta}\Big].$$

Using $\mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ (see (2.1.14)), we deduce that

$$\mathbf{E}\Big[\sup_{t\in[0,T]} \|\Upsilon_t^N - \Upsilon_{t^-}^N\|^2_{\mathcal{H}^{-J_0+1,j_0}}\Big] \leq C\Big[\frac{1}{\sqrt{N}} + \sqrt{\frac{1}{N^5} + \frac{1}{N^{8\beta-3}}} + N^{\frac{3}{2}-2\beta}\Big].$$

Because $\beta > 3/4$, this ends the proof of (2.3.38). The second statement in Lemma 2.31 follows from Proposition 2.27, (2.3.38), and [JS87, Condition 3.28 in Proposition 3.26]. The proof of Lemma 2.31 is complete. □

Lemma 2.32. *Let $\beta > 3/4$ and assume* **A1**-**A7**. *Then, for all $T > 0$:*

$$\lim_{N\to+\infty} \mathbf{E}\Big[\sup_{t\in[0,T]} \|\sqrt{N}M_t^N - \sqrt{N}M_{t^-}^N\|^2_{\mathcal{H}^{-J_0,j_0}}\Big] = 0. \tag{2.3.39}$$

In particular, any limit point of $(\sqrt{N}M^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$ belongs a.s. to $\mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$.

Proof. Let $T > 0$ and $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$. The function $t \in [0,T] \mapsto \langle f, \sqrt{N} M_t^N \rangle \in \mathbf{R}$ has $\lfloor NT \rfloor$ discontinuities, located at the times $\frac{1}{N}, \frac{2}{N}, \dots, \frac{\lfloor NT \rfloor}{N}$. For $k \in \{1, \dots, \lfloor NT \rfloor\}$, its k-th discontinuity is equal to $\sqrt{N}\langle f, M_{k-1}^N \rangle$. Thus,

$$\sup_{t\in[0,T]} \langle f, \sqrt{N} M_{t^-}^N - \sqrt{N} M_t^N \rangle^2 = N \max_{0 \le k < \lfloor NT \rfloor} \langle f, M_k^N \rangle^2.$$

Then, using (2.2.52) and because $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$ (see indeed (2.1.6)),

$$\mathbf{E}[\sup_{t\in[0,T]} \langle f, \sqrt{N} M_{t^-}^N - \sqrt{N} M_t^N \rangle^2] \le C \|f\|^2_{\mathcal{C}^{2,\gamma_*}} / \sqrt{N} \le C \|f\|^2_{\mathcal{H}^{L,\gamma}} / \sqrt{N}. \tag{2.3.40}$$

Considering an orthonormal basis of $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ (by (2.1.14)), we obtain

$$\mathbf{E}\Big[\sup_{t\in[0,T]} \| \sqrt{N} M_t^N - \sqrt{N} M_{t^-}^N \|^2_{\mathcal{H}^{-J_0,j_0}} \Big] \le C/\sqrt{N},$$

for some $C > 0$ independent of $N \ge 1$ and f. This proves (2.3.39). The second statement in Lemma 2.32 is a consequence of Proposition 2.30, (2.3.39), and condition 3.28 of [JS87, Proposition 3.26]. The proof of Lemma 2.32 is complete. □

2.3.4 Convergence of $(\sqrt{N} M^N)_{N\ge1}$ to a G-process

The aim of this section is to prove Proposition 2.35 below which states that $\{t \mapsto \sqrt{N} M_t^N, t \in \mathbf{R}_+\}_{N\ge1}$ (see (2.2.4) and (2.2.6)) converges towards a G-process (see Definition 2.6). To this end, we first show the convergence of this process against test functions.

Proposition 2.33. *Let $\beta > 3/4$ and assume* **A1**-**A7**. *Then, for every $f \in \mathcal{C}^{2,\gamma}(\mathbf{R}^d)$ the sequence $\{t \mapsto \sqrt{N}\langle f, M_t^N \rangle, t \in \mathbf{R}_+\}_{N\ge1}$ (see (2.2.6)) converges in distribution in $\mathcal{D}(\mathbf{R}_+, \mathbf{R})$ towards a process $X^f \in \mathcal{C}(\mathbf{R}_+, \mathbf{R})$ that has independent Gaussian increments. Moreover, for all $t \in \mathbf{R}_+$,*

$$\mathbf{E}[X_t^f] = 0 \quad \textit{and} \quad \mathrm{Var}(X_t^f) = \alpha^2 \mathbf{E}\left[\frac{1}{|B_\infty|}\right] \int_0^t \mathrm{Var}(\mathrm{Q}_s[f](x,y))\mathrm{d}s,$$

where we recall $\mathrm{Q}_s[f](x,y) = (y - \langle \sigma_*(\cdot,x), \bar{\mu}_s \rangle)\langle \nabla f \cdot \nabla \sigma_*(\cdot,x), \bar{\mu}_s \rangle$ (see Definition 2.6).

Proof. Let $f \in \mathcal{C}^{2,\gamma}(\mathbf{R}^d)$. Set for ease of notation

$$\mathfrak{m}_t^N[f] = \sqrt{N}\langle f, M_t^N \rangle.$$

To prove Proposition 2.34 we apply the martingale central limit theorem [EK09, Theorem 7.1.4] to the sequence $\{t \mapsto \mathfrak{m}_t^N[f], t \in \mathbf{R}_+\}_{N\ge1}$. To this end, let $T > 0$. Let us first show that Condition (a) in [EK09, Theorem 7.1.4] holds. First of all, by [EK09, Remark 7.1.5] and (2.2.6), the covariation matrix of $\mathfrak{m}_t^N[f]$ is $\mathfrak{a}_t^N[f] = N \sum_{k=0}^{\lfloor Nt \rfloor - 1} \langle f, M_k^N \rangle^2$ and therefore $\mathfrak{a}_t^N[f] - \mathfrak{a}_s^N[f] \ge 0$ if $t \ge s$. On the other hand, by (2.3.40) and (2.1.6), we have:

$$\lim_{N\to\infty} \mathbf{E}\Big[\sup_{t\in[0,T]} \big| \mathfrak{m}_t^N[f] - \mathfrak{m}_{t^-}^N[f] \big| \Big] = 0. \tag{2.3.41}$$

Thus Condition (a) in [EK09, Theorem 7.1.4] holds. Let us now prove the last required condition in [EK09, Theorem 7.1.4], namely that for all $t \in \mathbf{R}_+$, $\mathfrak{a}_t^N[f] \xrightarrow{p} \mathfrak{c}_t[f]$ where $\mathfrak{c}$ satisfies the assumptions of [EK09, Theorem 7.1.1], i.e., $t \in \mathbf{R}_+ \mapsto \mathfrak{c}_t[f]$ continuous, $\mathfrak{c}_0[f] = 0$, and $\mathfrak{c}_t[f] - \mathfrak{c}_s[f] \ge 0$ if $t \ge s$. Recall the definition of the $\boldsymbol{\sigma}$-algebra $\mathcal{F}_k^N$ in (2.1.9). For $t \in \mathbf{R}_+$,

$$\mathfrak{a}_t^N[f] = N \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbf{E}[\langle f, M_k^N \rangle^2 | \mathcal{F}_k^N] + N \sum_{k=0}^{\lfloor Nt \rfloor - 1} \left(\langle f, M_k^N \rangle^2 - \mathbf{E}[\langle f, M_k^N \rangle^2 | \mathcal{F}_k^N] \right). \tag{2.3.42}$$

We start by studying the first term in the right-hand side of (2.3.42). Recall that (see (2.2.25))

$$\mathrm{Q}^N[f](x,y,\{W_k^i\}_i) = (y - \langle \sigma_*(\cdot,x), \nu_k^N\rangle)\langle \nabla f \cdot \nabla\sigma_*(\cdot,x), \nu_k^N\rangle,$$

and set (see (2.2.3))

$$\bar{\mathrm{Q}}^N[f](\{W_k^i\}_i) := \int_{\mathcal{X}\times\mathcal{Y}} \mathrm{Q}^N[f](x,y,\{W_k^i\}_i)\pi(\mathrm{d}x,\mathrm{d}y) = \frac{N}{\alpha}\langle f, D_k^N\rangle.$$

Using that $\big(|B_k|, ((x_k^n,y_k^n))_{n\geq 1}\big) \perp\!\!\!\perp \mathcal{F}_k^N$, $|B_k| \perp\!\!\!\perp ((x_k^n,y_k^n))_{n\geq 1}$ (see **A1**), and $(W_k^1,\dots,W_k^N)$ is $\mathcal{F}_k^N$-measurable, it holds:

$$\begin{aligned}
&\mathbf{E}\Big[\frac{1}{|B_k|^2}\sum_{1\leq n<m\leq |B_k|}\Big(\mathrm{Q}^N[f](x_k^n,y_k^n,\{W_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{W_k^i\}_i)\Big)\Big(\mathrm{Q}^N[f](x_k^m,y_k^m,\{W_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{W_k^i\}_i)\Big)\Big|\mathcal{F}_k^N\Big]\\
&= \sum_{q\geq 1}\frac{1}{q^2}\sum_{1\leq n<m\leq q}\mathbf{E}\Big[\mathbf{1}_{|B_k|=q}\Big(\mathrm{Q}_k^N[f](x_k^n,y_k^n,\{W_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{W_k^i\}_i)\Big)\\
&\quad\times\Big(\mathrm{Q}^N[f](x_k^m,y_k^m,\{W_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{W_k^i\}_i)\Big)\Big|\mathcal{F}_k^N\Big]\\
&= \sum_{q\geq 1}\frac{1}{q^2}\sum_{1\leq n<m\leq q}\mathbf{E}\Big[\mathbf{1}_{|B_k|=q}\Big(\mathrm{Q}^N[f](x_k^n,y_k^n,\{w_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{w_k^i\}_i)\Big)\\
&\quad\times\Big(\mathrm{Q}^N[f](x_k^m,y_k^m,\{w_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{w_k^i\}_i)\Big)\Big]\Big|_{\{w_k^i\}_i=\{W_k^i\}_i}\\
&= \sum_{q\geq 1}\frac{1}{q^2}\sum_{1\leq n<m\leq q}\mathbf{E}[\mathbf{1}_{|B_k|=q}]\\
&\quad\times\mathbf{E}\Big[\Big(\mathrm{Q}^N[f](x_k^n,y_k^n,\{w_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{w_k^i\}_i)\Big)\Big(\mathrm{Q}^N[f](x_k^m,y_k^m,\{w_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{w_k^i\}_i)\Big)\Big]\Big|_{\{w_k^i\}_i=\{W_k^i\}_i}\\
&= \sum_{q\geq 1}\frac{1}{q^2}\sum_{1\leq n<m\leq q}\mathbf{E}[\mathbf{1}_{|B_k|=q}]\\
&\quad\times\mathbf{E}\Big[\mathrm{Q}^N[f](x_k^n,y_k^n,\{w_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{w_k^i\}_i)\Big]\Big|_{\{w_k^i\}_i=\{W_k^i\}_i}\\
&\quad\times\mathbf{E}\Big[\mathrm{Q}^N[f](x_k^m,y_k^m,\{w_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{w_k^i\}_i)\Big]\Big|_{\{w_k^i\}_i=\{W_k^i\}_i}\\
&= 0.
\end{aligned}$$

where we have used **A3** at the two last equalities. We also have with the same arguments:

$$\begin{aligned}
&\mathbf{E}\Big[\frac{1}{|B_k|^2}\sum_{n=1}^{|B_k|}\big|\mathrm{Q}^N[f](x_k^n,y_k^n,\{W_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{W_k^i\}_i)\big|^2\Big|\mathcal{F}_k^N\Big]\\
&= \sum_{q\geq 1}\frac{1}{q^2}\sum_{n=1}^{q}\mathbf{E}[\mathbf{1}_{|B_k|=q}]\,\mathbf{E}\Big[\big|\mathrm{Q}^N[f](x_k^n,y_k^n,\{w_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{w_k^i\}_i)\big|^2\Big]\Big|_{\{w_k^i\}_i=\{W_k^i\}_i}\\
&= \sum_{q\geq 1}\frac{1}{q^2}\sum_{n=1}^{q}\mathbf{E}[\mathbf{1}_{|B_k|=q}]\mathbf{E}\Big[\big|\mathrm{Q}^N[f](x_1^1,y_1^1,\{w_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{w_k^i\}_i)\big|^2\Big]\Big|_{\{w_k^i\}_i=\{W_k^i\}_i}\\
&= \mathbf{E}\left[\frac{1}{|B_k|}\right]\mathrm{Var}_\pi\left(\mathrm{Q}^N[f](x,y,\{W_k^i\}_i)\right).
\end{aligned}$$

The notation Cov_π means that we consider the expectation only w.r.t. $(x,y)\sim\pi$ (see **A3**) . We

then have, for $k \geq 0$ (see (2.2.4)),

$$\begin{aligned}
\mathbf{E}[\langle f, M_k^N\rangle^2 | \mathcal{F}_k^N] &= \frac{\alpha^2}{N^2}\mathbf{E}\Big[\Big|\frac{1}{|B_k|}\sum_{(x,y)\in B_k}\big[\mathrm{Q}^N[f](x,y,\{W_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{W_k^i\}_i)\big]\Big|^2\Big|\mathcal{F}_k^N\Big] \\
&= \frac{\alpha^2}{N^2}\mathbf{E}\Big[\frac{1}{|B_k|^2}\sum_{n=1}^{|B_k|}\big|\mathrm{Q}^N[f](x_k^n,y_k^n,\{W_k^i\}_i) - \bar{\mathrm{Q}}^N[f](\{W_k^i\}_i)\big|^2\Big|\mathcal{F}_k^N\Big] \\
&= \frac{\alpha^2}{N^2}\mathbf{E}\left[\frac{1}{|B_k|}\right]\mathrm{Var}_\pi\left(\mathrm{Q}^N[f](x,y,\{W_k^i\}_i)\right).
\end{aligned}$$

Then, one has:

$$\begin{aligned}
N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}[\langle f, M_k^N\rangle^2|\mathcal{F}_k^N] &= \alpha^2\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}\mathbf{E}\left[\frac{1}{|B_{\lfloor Ns\rfloor}|}\right]\mathrm{Var}_\pi\left(\mathrm{Q}^N[f](x,y,\{W^i_{\lfloor Ns\rfloor}\}_i)\right)\mathrm{d}s \\
&= \alpha^2\int_0^t\mathbf{E}\left[\frac{1}{|B_{\lfloor Ns\rfloor}|}\right]\mathrm{Var}_\pi\left(\mathrm{Q}^N[f](x,y,\{W^i_{\lfloor Ns\rfloor}\}_i)\right)\mathrm{d}s \\
&\quad - \alpha^2\int_{\frac{\lfloor Nt\rfloor}{N}}^t\mathbf{E}\left[\frac{1}{|B_{\lfloor Ns\rfloor}|}\right]\mathrm{Var}_\pi\left(\mathrm{Q}^N[f](x,y,\{W^i_{\lfloor Ns\rfloor}\}_i)\right)\mathrm{d}s. \quad (2.3.43)
\end{aligned}$$

Using **A7**, a dominated convergence theorem, and the same arguments as those used in the proof of Lemma 2.20, we prove that if $m^N \to m$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$, we have for all $f \in \mathcal{C}^{2,\gamma}(\mathbf{R}^d)$ and $t \in \mathbf{R}_+$, as $N \to +\infty$,

$$\begin{aligned}
&\int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\mathbf{E}\left[\frac{1}{|B_{\lfloor Ns\rfloor}|}\right]\Big[(y - \langle\sigma_*(\cdot,x), m_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x), m_s^N\rangle \\
&\quad - \int_{\mathcal{X}\times\mathcal{Y}}(y' - \langle\sigma_*(\cdot,x'), m_s^N\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x'), m_s^N\rangle\pi(\mathrm{d}x',\mathrm{d}y')\Big]^2\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\to \int_0^t\int_{\mathcal{X}\times\mathcal{Y}}\mathbf{E}\left[\frac{1}{|B_\infty|}\right]\Big[(y - \langle\sigma_*(\cdot,x), m_s\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x), m_s\rangle \\
&\quad - \int_{\mathcal{X}\times\mathcal{Y}}(y' - \langle\sigma_*(\cdot,x'), m_s\rangle)\langle\nabla f\cdot\nabla\sigma_*(\cdot,x'), m_s\rangle\pi(\mathrm{d}x',\mathrm{d}y')\Big]^2\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s.
\end{aligned}$$

Recall that by Theorem 2.1, $\mu^N \xrightarrow{p} \bar{\mu}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$. Therefore, using the continuous mapping theorem, we have for all $t \in \mathbf{R}_+$ and $f \in \mathcal{C}^{2,\gamma}(\mathbf{R}^d)$:

$$\alpha^2\int_0^t\mathbf{E}\left[\frac{1}{|B_{\lfloor Ns\rfloor}|}\right]\mathrm{Var}_\pi(\mathrm{Q}^N[f](x,y,\{W^i_{\lfloor Ns\rfloor}\}_i))\mathrm{d}s \xrightarrow{p} \mathfrak{c}_t[f] := \alpha^2\int_0^t\mathbf{E}\left[\frac{1}{|B_\infty|}\right]\mathrm{Var}(\mathrm{Q}_s[f](x,y))\mathrm{d}s.$$

Note that $t \in \mathbf{R}_+ \mapsto \mathfrak{c}_t[f]$ is locally Lipschitz continuous since for all $s \in [0,t]$, $\mathrm{Var}(\mathrm{Q}_s[f](x,y)) \leq C\int_{\mathcal{X}\times Y}(|y|^2+1)\pi(\mathrm{d}x,\mathrm{d}y)\|f\|^2_{\mathcal{C}^{1,\gamma}}\sup_{s\in[0,t]}|\langle(1+|\cdot|^\gamma),\bar{\mu}_s\rangle|^2$. Let us now consider the second term in the right-hand side of (2.3.43). Using (2.2.20) and Lemma 2.11, $\mathbf{E}[|\mathrm{Q}^N[f](x,y,\{W_k^i\}_i)|^2] \leq C\|f\|^2_{\mathcal{C}^{2,\gamma}}$. Consequently, it holds:

$$\mathbf{E}\Big[\Big|\alpha^2\int_{\frac{\lfloor Nt\rfloor}{N}}^t\mathbf{E}\Big[\frac{1}{|B_{\lfloor Ns\rfloor}|}\Big]\mathrm{Var}_\pi\left(\mathrm{Q}^N[f](x,y,\{W^i_{\lfloor Ns\rfloor}\}_i)\right)\mathrm{d}s\Big|\Big] \xrightarrow[N\to\infty]{} 0.$$

We have thus shown that

$$N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}[\langle f, M_k^N\rangle^2|\mathcal{F}_k^N] \xrightarrow[N\to\infty]{p} \alpha^2\int_0^t\mathbf{E}\left[\frac{1}{|B_\infty|}\right]\mathrm{Var}(\mathrm{Q}_s[f](x,y))\mathrm{d}s.$$

At this point, the study of the first term in the right-hand side of (2.3.42) is complete. It remains to study the second term in the right-hand side of (2.3.42). Using (2.2.51), we obtain:

$$N^2\mathbf{E}\Big[\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,M_k^N\rangle^2-\mathbf{E}[\langle f,M_k^N\rangle^2|\mathcal{F}_k^N]\Big|^2\Big]=N^2\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}\Big[\Big|\langle f,M_k^N\rangle^2-\mathbf{E}[\langle f,M_k^N\rangle^2|\mathcal{F}_k^N]\Big|^2\Big]$$
$$\leq CN^2\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}[\langle f,M_k^N\rangle^4]\leq CN^2\|f\|^4_{\mathcal{C}^{2,\gamma_*}}/N^3\to 0.$$

In conclusion, we have proved that for every $t\in\mathbf{R}_+$,

$$\mathfrak{a}_t^N[f]\xrightarrow{p}\mathfrak{c}_t[f]\text{ as }N\to+\infty. \tag{2.3.44}$$

By [EK09, Theorem 7.1.4], the proof of Proposition 2.33 is complete.

□

Proposition 2.34. *Let $\beta>3/4$ and assume* **A1**-**A7**. *Consider a family $\mathscr{F}=\{f_a\}_{a\geq 1}$ of elements of $\mathcal{C}^{2,\gamma}(\mathbf{R}^d)$. Then, for $k\geq 1$, the sequence*

$$\{t\mapsto(\sqrt{N}\langle f_1,M_t^N\rangle,\dots,\sqrt{N}\langle f_k,M_t^N\rangle)^T,t\in\mathbf{R}_+\}_{N\geq 1}$$

converges in distribution in $\mathcal{D}(\mathbf{R}_+,\mathbf{R}^k)$ towards a process $Y_k^{\mathscr{F}}=\{t\mapsto(Y_t^1,\dots,Y_t^k)^T,t\in\mathbf{R}_+\}\in\mathcal{C}(\mathbf{R}_+,\mathbf{R}^k)$ with zero-mean and independent Gaussian increments (which is thus a martingale). In addition, for all $0\leq s\leq t$,

$$\mathrm{Cov}(Y_t^i,Y_s^j)=\alpha^2\mathbf{E}\left[\frac{1}{|B_\infty|}\right]\int_0^s\mathrm{Cov}(\mathrm{Q}_v[f_i](x,y),\mathrm{Q}_v[f_j](x,y))\mathrm{d}v,\quad 1\leq i,j\leq k. \tag{2.3.45}$$

Notice that (2.3.45) is exactly (2.1.15).

Proof. Set for ease of notation, $\mathscr{M}_t^N=(\sqrt{N}\langle f_1,M_t^N\rangle,\dots,\sqrt{N}\langle f_k,M_t^N\rangle)^T$, $t\in\mathbf{R}_+$. We have $\mathscr{M}_t^N=\sum_{q=0}^{\lfloor Nt\rfloor-1}\xi_q^N$, where $\xi_q^N=(\sqrt{N}\langle f_1,M_q^N\rangle,\dots,\sqrt{N}\langle f_k,M_q^N\rangle)^T$ (see indeed (2.2.6)). From [EK09, Remark 7.1.5], the covariation matrix of $\mathscr{M}_t^N$ is

$$\mathscr{A}_t^N[f_1,\dots,f_k]:=N\sum_{q=0}^{\lfloor Nt\rfloor-1}\xi_q^N(\xi_q^N)^T=N\sum_{q=0}^{\lfloor Nt\rfloor-1}(\langle f_i,M_q^N\rangle\langle f_j,M_q^N\rangle)_{i,j=1,\dots,k}.$$

If $t\geq s$, we have $\mathscr{A}_t^N[f_1,\dots,f_k]-\mathscr{A}_s^N[f_1,\dots,f_k]=N\sum_{q=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\xi_q^T\xi_q\geq 0$. By (2.3.41), Condition (a) in [EK09, Theorem 7.1.4] is satisfied for $\mathscr{M}^N$. Secondly, condition (1.19) in [EK09, Theorem 7.1.4] is satisfied, using the decomposition

$$\begin{aligned}\mathscr{A}_t^N[f_1,\dots,f_k]_{i,j}=N\sum_{q=0}^{\lfloor Nt\rfloor-1}\langle f_i,M_q^N\rangle\langle f_j,M_q^N\rangle&=N\sum_{q=0}^{\lfloor Nt\rfloor-1}\frac{1}{2}(\langle f_i+f_j,M_q^N\rangle^2-\langle f_i,M_q^N\rangle^2-\langle f_j,M_q^N\rangle^2)\\&=\frac{1}{2}\big(\mathfrak{a}_t^N[f_i+f_j]-\mathfrak{a}_t^N[f_i]-\mathfrak{a}_t^N[f_j]\big)\\&\xrightarrow{p}\frac{1}{2}\big(\mathfrak{c}_t[f_i+f_j]-\mathfrak{c}_t[f_i]-\mathfrak{c}_t[f_j]\big)\text{ (by (2.3.44))}\\&=\alpha^2\int_0^t\mathbf{E}\left[\frac{1}{|B_\infty|}\right]\mathrm{Cov}(\mathrm{Q}_v[f_i](x,y),\mathrm{Q}_v[f_j](x,y))\mathrm{d}v\\&:=(\mathfrak{C}_t)_{i,j}.\end{aligned}$$

It remains to check that $\mathfrak{C}$ satisfies the assumptions of [EK09, Theorem 7.1.1]. Clearly $\mathfrak{C}(0) = 0$ and $t \in \mathbf{R}_+ \mapsto \mathfrak{C}_t$ is continuous. In addition, if $0 \le s \le t$,

$$\begin{aligned}\mathfrak{C}_t - \mathfrak{C}_s &= \alpha^2 \Big(\int_s^t \mathbf{E}\left[\frac{1}{|B_\infty|}\right] \mathrm{Cov}(\mathrm{Q}_v[f_i](x,y), \mathrm{Q}_v[f_j](x,y))\mathrm{d}v \Big)_{i,j=1,\dots,k} \\ &= \alpha^2 \int_s^t \mathbf{E}\left[\frac{1}{|B_\infty|}\right] \mathbf{E}[\Xi_v(x,y)\Xi_v^T(x,y)]\mathrm{d}v,\end{aligned}$$

where $\Xi_v(x,y)_i = \mathrm{Q}_v[f_i](x,y) - \mathbf{E}[\mathrm{Q}_v[f_i](x,y)]$, $i \in \{1,\dots,k\}$. Thus, $\mathfrak{C}_t - \mathfrak{C}_s$ is symmetric and non negative definite. The proof of Proposition 2.34 is complete. □

Proposition 2.35. *Let $\beta > 3/4$ and assume that* **A1**-**A7** *hold. Then, the sequence $(\sqrt{N}M^N)_{N\ge1}$ converges in distribution in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$ to a* G*-process $\mathscr{G} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$ (see Definition 2.6).*

To prove Proposition 2.35, we will prove that there is a unique limit point of the sequence $(\sqrt{N}M^N)_{N\ge1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$ (recall that this sequence is relatively compact in this space, see Proposition 2.30), so that the whole sequence converges in distribution. Proposition 2.34 will then imply that this unique limit point is a G-process. Before, we need to introduce some definitions. For a family $\mathscr{F} = \{f_a\}_{a\ge1}$ of elements of $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$, we define, for $k \ge 1$, the projection

$$\pi_k^{\mathscr{F}} : \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d)) \to \mathcal{D}(\mathbf{R}_+, \mathbf{R})^k, \quad m \mapsto (\langle f_1, m\rangle, \dots, \langle f_k, m\rangle)^T.$$

The function $\pi_k^{\mathscr{F}}$ is continuous. In the following, $\mathscr{H} = \{h_a\}_{a\ge1}$ is an orthonormal basis of $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$. Let d_R be a metric for the Skorohod topology on $\mathcal{D}(\mathbf{R}_+, \mathbf{R})$. Introduce the space $\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty$ defined as the set of sequences taking values in $\mathcal{D}(\mathbf{R}_+, \mathbf{R})$. We endow $\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty$ with the metric $\rho(u,v) = \sum_{a\ge1} 2^{-a} \min(1, d_R(u_a, v_a))$. We consider on $\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty$ the topology associated with ρ. We have that $\rho(u^N, u) \to 0$ if and only if $d_R(u_a^N, u_a) \to 0$ for all $a \ge 1$. Notice with that with this metric ρ, $\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty$ is separable, since $\mathcal{D}(\mathbf{R}_+, \mathbf{R})$ is separable [EK09, Theorem 3.5.6]. We now define the map

$$\Pi : \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d)) \to \mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty, \quad m \mapsto (\langle h_a, m\rangle)_{a\ge1}^T.$$

This map is injective (because $\mathscr{H}$ is a basis of $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$) and continuous. The map Π depends on the orthonormal basis $\mathscr{H}$ but, for ease of notation, we have omitted to write it. Finally, we introduce the continuous function

$$p_k : \mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty \to \mathcal{D}(\mathbf{R}_+, \mathbf{R})^k, \quad (m_a)_{a\ge1}^T \mapsto (m_1, \dots, m_k)^T.$$

It holds

$$\pi_k^{\mathscr{H}} = p_k \circ \Pi.$$

We now introduce the set

$$\mathcal{C} := \big\{ p_k^{-1}(H), \quad H \in \mathcal{B}(\mathcal{D}(\mathbf{R}_+, \mathbf{R})^k),\ k \ge 1 \big\} \subset \mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty,$$

where $\mathcal{B}(\mathcal{D}(\mathbf{R}_+, \mathbf{R})^k)$ denotes the Borel $\boldsymbol{\sigma}$-algebra of $\mathcal{D}(\mathbf{R}_+, \mathbf{R})^k$. The continuity of p_k implies that $\mathcal{C} \subset \mathcal{B}(\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty)$. The following result shows that $\mathcal{C}$ is a separating class of $(\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty, \mathcal{B}(\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty))$, where we recall that this means by definition that two probability measures on $\mathcal{B}(\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty)$ which agree on $\mathcal{C}$ necessarily agree on $\mathcal{B}(\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty)$.

Lemma 2.36. *The set $\mathcal{C}$ is a separating class of $(\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty, \mathcal{B}(\mathcal{D}(\mathbf{R}_+, \mathbf{R})^\infty))$.*

Proof. We recall that any subset of $\mathcal{B}(\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty)$ which is a π-system (i.e. closed under finite intersection) and which generates the $\boldsymbol{\sigma}$-algebra $\mathcal{B}(\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty)$ is a separating class (see [Bil99, Page 9]). Let us first prove that $\mathcal{C}$ is a π-system. Notice that it holds $p_k^{-1}(H) = p_{k+1}^{-1}(H \times \mathcal{D}(\mathbf{R}_+,\mathbf{R}))$. Thus, if A and $A' \in \mathcal{C}$ (write $A = p_k^{-1}(H)$ and $A' = p_{k'}^{-1}(H')$, and assume that $k' \geq k$), then $A \cap A' = p_{k'}^{-1}((H \times \mathcal{D}(\mathbf{R}_+,\mathbf{R}) \ldots \times \mathcal{D}(\mathbf{R}_+,\mathbf{R})) \cap H') \in \mathcal{C}$. Consequently $\mathcal{C}$ is a π-system. It remains to show that the $\boldsymbol{\sigma}$-algebra generated by $\mathcal{C}$ is equal to $\mathcal{B}(\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty)$. To prove it, it is sufficient to prove that any open set of $\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty$ is a countable union of sets in $\mathcal{C}$. Introduce, for $x \in \mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty$, $k \geq 1$ and $\epsilon > 0$,

$$\mathcal{N}_{k,\epsilon}(x) := \{y \in \mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty;\ d_R(x_a, y_a) < \epsilon,\ 1 \leq a \leq k\} \subset \mathcal{C}. \tag{2.3.46}$$

By straightforward arguments $\mathcal{N}_{k,\epsilon}(x)$ is open in $(\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty, \rho)$. Remark that for all $y \in \mathcal{N}_{k,\epsilon}(x)$, it holds $\rho(x,y) < \epsilon + 2^{-k}$. Given $r > 0$, choose $\epsilon > 0$ and $k \geq 1$ such that $\epsilon + 2^{-k} < r$. Then, $\mathcal{N}_{k,\epsilon}(x) \subset B_\rho(x,r)$ where $B_\rho(x,r)$ is the open ball of center x and radius r for the metric ρ of $\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty$. The space $\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty$ is separable and we consider D a dense and countable subset of $(\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty, \rho)$. Let $\mathcal{O}$ be an open subset of $(\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty, \rho)$. We claim that

$$\mathcal{O} = \mathcal{N}_\mathcal{O} \text{ where } \mathcal{N}_\mathcal{O} := \bigcup_{\substack{x \in \mathrm{D} \cap \mathcal{O},\ k \geq 1 \\ \epsilon \in \mathbf{Q} \cap \mathbf{R}_+^*,\ \mathcal{N}_{k,\epsilon}(x) \subset \mathcal{O}}} \mathcal{N}_{k,\epsilon}(x).$$

We have $\mathcal{N}_\mathcal{O} \subset \mathcal{O}$. Let us now show that $\mathcal{O} \subset \mathcal{N}_\mathcal{O}$. To this end, pick $y \in \mathcal{O}$ and $r_0 > 0$ such that $B_\rho(y, r_0) \subset \mathcal{O}$. Choose $k_0 \geq 1$ such that

$$2^{-k_0} < \frac{r_0}{4}.$$

Consider, for $n \geq 1$, $x^n \in \mathrm{D}$ such that $\rho(y, x^n) < 1/n$. Choose $n \geq 1$ such that

$$\frac{1}{n} + \frac{r_0}{2} < r_0 \text{ and } \frac{2^{k_0}}{n} < \frac{r_0}{4}.$$

Since for all $a \geq 1$, $d_R(y_a, x_a^n) \to 0$ as $n \to +\infty$ (by definition of ρ), we choose if necessary $n \geq 1$ larger so that $\min(1, d_R(y_a, x_a^n)) = d_R(y_a, x_a^n)$ for all $a = 1, \ldots, k_0$. Finally, choose $\epsilon_{k_0,n} = 2^{k_0}/n \in \mathbf{Q}$. We have for all $a = 1, \ldots, k_0$, $d_R(x_n^a, y_a) \leq \rho(x^n, y) 2^a \leq \rho(x^n, y) 2^{k_0} < 2^{k_0}/n = \epsilon_{k_0,n}$, so that $y \in \mathcal{N}_{k_0,\epsilon_{k_0,n}}(x^n)$. It just remains to check that $\mathcal{N}_{k_0,\epsilon_{k_0,n}}(x^n) \subset \mathcal{O}$ to ensure that $\mathcal{N}_{k_0,\epsilon_{k_0,n}}(x^n) \subset \mathcal{N}_\mathcal{O}$. We have $\rho(y, x^n) < \epsilon_{k_0,n} + 2^{-k_0} < r_0/4 + r_0/4 = r_0/2$ and thus $\mathcal{N}_{k_0,\epsilon_{k_0,n}}(x^n) \subset B_\rho(x^n, r_0/2)$. Since $1/n + r_0/2 < r_0$, by triangular inequality, $B_\rho(x^n, r_0/2) \subset B_\rho(y, r_0) \subset \mathcal{O}$. Thus, $\mathcal{N}_{k_0,\epsilon_{k_0,n}}(x^n) \subset \mathcal{O}$, which proves that $\mathcal{N}_{k_0,\epsilon_{k_0,n}}(x^n) \subset \mathcal{N}_\mathcal{O}$. Consequently, we have proved that $y \in \mathcal{N}_{k_0,\epsilon_{k_0,n}}(x^n) \subset \mathcal{N}_\mathcal{O}$, and then that $\mathcal{O} \subset \mathcal{N}_\mathcal{O}$. Thus, $\mathcal{O} = \mathcal{N}_\mathcal{O}$. In conclusion, every open set $\mathcal{O}$ is a countable union of sets of the form (2.3.46). This implies that $\sigma(\mathcal{C}) = \mathcal{B}(\mathcal{D}(\mathbf{R}_+,\mathbf{R})^\infty)$ and therefore $\mathcal{C}$ is a separating class. □

Proposition 2.37. *Let P, Q be two probability measures on $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$ such that $\pi_k^\mathscr{H} P = \pi_k^\mathscr{H} Q$ for all $k \geq 1$. Then, $P = Q$.*

Proof. The equality $\pi_k^\mathscr{H} P = \pi_k^\mathscr{H} Q$ for all $k \geq 1$, writes $p_k \circ \Pi P = p_k \circ \Pi Q$. By Lemma 2.36, $\Pi P = \Pi Q$. Since Π is injective it admits a left inverse Π^{-1}, and therefore $P = Q$. The proof is complete. □

We are now in position to prove Proposition 2.35.

Proof of Proposition 2.35. By Proposition 2.30, $(\sqrt{N}M^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$. Let $\mathscr{M}^*$ be one of its limit point. Let us show that $\mathscr{M}^*$ is independent of the extracted subsequence, say of N'. Since $\pi_k^{\mathscr{H}}$ is continuous for all $k \geq 1$, the continuous mapping theorem implies that in $\mathcal{D}(\mathbf{R}_+, \mathbf{R})^k$,

$$\pi_k^{\mathscr{H}}(\sqrt{N'}M^{N'}) \to \pi_k^{\mathscr{H}}(\mathscr{M}^*) \text{ in distribution, as } N' \to +\infty.$$

For $k \geq 1$, introduce the following continuous and bijective mapping $\mathscr{Q}_k : \mathcal{D}(\mathbf{R}_+, \mathbf{R}^k) \to \mathcal{D}(\mathbf{R}_+, \mathbf{R})^k$ defined by: $m \mapsto (m \cdot e_1, \ldots, m \cdot e_k)^T$, where $\{e_1, \ldots, e_k\}$ denotes the canonical basis of $\mathbf{R}^k$. By Proposition 2.34 applied with $f_a = h_a$, $a \in \{1, \ldots, k\}$ (recall $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d) \subset \mathcal{C}^{2,\gamma}(\mathbf{R}^d)$), it holds in $\mathcal{D}(\mathbf{R}_+, \mathbf{R}^k)$,

$$\forall k \geq 1,\ \mathscr{Q}_k^{-1} \circ \pi_k^{\mathscr{H}}(\sqrt{N'}M^{N'}) \to Y_k^{\mathscr{H}} \text{ in distribution, as } N' \to +\infty.$$

Since $\mathscr{Q}_k$ is continuous, one then has in $\mathcal{D}(\mathbf{R}_+, \mathbf{R})^k$,

$$\forall k \geq 1,\ \pi_k^{\mathscr{H}}(\sqrt{N'}M^{N'}) \to \mathscr{Q}_k(Y_k^{\mathscr{H}}) \text{ in distribution, as } N' \to +\infty.$$

It follows that $\pi_k^{\mathscr{H}}(\mathscr{M}^*) = \mathscr{Q}_k(Y_k^{\mathscr{H}})$ in distribution. By Proposition 2.37, the distribution of $\mathscr{M}^*$ is fully determined by the collection of distributions of the processes $\pi_k^{\mathscr{H}}(\mathscr{M}^*)$, for $k \geq 1$. Thus, $\mathscr{M}^*$ is independent of the subsequence, and therefore the whole sequence $(\sqrt{N}M^N)_{N\geq 1}$ convergences to $\mathscr{M}^*$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$. By Lemma 2.32, $\mathscr{M}^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$. Let us now consider a family $\mathscr{F} = \{f_a\}_{a\geq 1}$ of elements of $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$. Since $\mathscr{D}_k$ and $\pi_k^{\mathscr{F}}$ are continuous, and by Proposition 2.34, one has that $\mathscr{Q}_k^{-1} \circ \pi_k^{\mathscr{F}}(\mathscr{M}^*) = Y_k^{\mathscr{F}} \in \mathcal{C}(\mathbf{R}_+, \mathbf{R}^k)$ in distribution. The proof of Proposition 2.35 is complete. □

2.3.5 Limit points of $(\eta^N)_{N\geq 1}$ and end of the proof of Theorem 2.8

On the limit points of the sequence $(\eta^N, \sqrt{N}M^N)_{N\geq 1}$

Lemma 2.38. *Assume* **A1-A7**. *Then, the sequence $(\eta_0^N)_{N\geq 1}$ converges in distribution in $\mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$ towards a variable ν_0 which is the unique (in distribution) $\mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$-valued random variable such that for all $k \geq 1$ and $f_1 \ldots, f_k \in \mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d)$, $(\langle f_1, \nu_0\rangle, \ldots, \langle f_k, \nu_0\rangle)^T \sim \mathcal{N}(0, \Gamma(f_1, \ldots, f_k))$, where $\Gamma(f_1, \ldots, f_k)$ is the covariance matrix of the vector $(f_1(W_0^1), \ldots, f_k(W_0^1))^T$.*

Proof. The sequence $(\eta_0^N)_{N\geq 1}$ is tight in $\mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$. Let $\mathscr{F} = \{f_a\}_{a\geq 1}$ be a family of elements of $\mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d)$. Define, for $k \geq 1$, the projection

$$\mathscr{P}_k^{\mathscr{F}} : \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d) \to \mathbf{R}^k, \quad m \mapsto (\langle f_1, m\rangle, \ldots, \langle f_k, m\rangle).$$

The map $\mathscr{P}_k^{\mathscr{F}}$ is continuous. By the standard vectorial central limit theorem, for $k \geq 1$, $\mathscr{P}_k^{\mathscr{F}}(\eta_0^N) \to \mathcal{N}(0, \Gamma(f_1, \ldots, f_k))$ in distribution. In addition, we show with the same arguments as those used to prove Lemma 2.36 and Proposition 2.37, that when $\mathscr{F}$ is an orthonormal basis of $\mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d)$, the distribution of a $\mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$-valued random variable ν is fully determined by the collection of the distributions $\{\mathscr{P}_k^{\mathscr{F}}(\nu), k \geq 1\}$. Hence, $(\eta_0^N)_{N\geq 1}$ has a unique limit point ν_0 in distribution which is the unique (in distribution) $\mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$-valued random variable such that for all $k \geq 1$, $\mathscr{P}_k^{\mathscr{F}}(\nu_0) \sim \mathcal{N}(0, \Gamma(f_1, \ldots, f_k))$. In particular, the whole sequence $(\eta_0^N)_{N\geq 1}$ converges in distribution towards ν_0. The proof of the lemma is complete. □

Set

$$\mathscr{E} := \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)) \times \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d)). \tag{2.3.47}$$

According to Propositions 2.27 and 2.30, $(\eta^N, \sqrt{N}\, M^N)_{N\geq 1}$ is tight in $\mathscr{E}$. Let $(\eta^*, \mathscr{G}^*)$ be one of its limit point in $\mathscr{E}$. Along some subsequence, it holds:

$$(\eta^{N'}, \sqrt{N'}\, M^{N'}) \to (\eta^*, \mathscr{G}^*), \text{ as } N' \to +\infty.$$

Considering the marginal distributions, and according to Lemmata 2.31 and 2.32, it holds a.s.

$$\eta^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)) \text{ and } \mathscr{G}^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d)). \tag{2.3.48}$$

By uniqueness of the limit in distribution, using Lemma 2.38 (together with the fact that the projection $m \in \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)) \mapsto m_0 \in \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$ is continuous) and Proposition 2.35, it also holds:

$$\eta_0^* = \nu_0 \text{ and } \mathscr{G}^* = \mathscr{G}, \text{ in distribution.} \tag{2.3.49}$$

Proposition 2.39. *Let $\beta > 3/4$ and assume* **A1**-**A7**. *Then, η^* is a weak solution of* (2.1.16) *(see Definition 2.7) with initial distribution ν_0 (see Lemma 2.38).*

Proof. Let us introduce, for $\Phi \in \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$, $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$, and $s \geq 0$:

$$\mathbf{U}_s[f](\Phi) := \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \bar{\mu}_s\rangle)\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \Phi\rangle \pi(\mathrm{d}x, \mathrm{d}y) \tag{2.3.50}$$

and

$$\mathbf{V}_s[f](\Phi) := \int_{\mathcal{X}\times\mathcal{Y}} \langle \sigma_*(\cdot, x), \Phi\rangle\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \bar{\mu}_s\rangle \pi(\mathrm{d}x, \mathrm{d}y). \tag{2.3.51}$$

Recall $\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t)$ (see (2.1.11)). Using (2.2.8) and Corollary 2.2, it holds:

$$\langle f, \eta_t^N\rangle - \langle f, \eta_0^N\rangle - \int_0^t (\mathbf{U}_s[f](\eta_s^N) - \mathbf{V}_s[f](\eta_s^N))\mathrm{d}s - \langle f, \sqrt{N} M_t^N\rangle = -\mathbf{e}_t^N[f], \tag{2.3.52}$$

where

$$\begin{aligned}\mathbf{e}_t^N[f] :=& \frac{1}{\sqrt{N}} \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\langle \sigma_*(\cdot, x), \eta_s^N\rangle\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \eta_s^N\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \sqrt{N}\langle f, V_t^N\rangle - \sqrt{N} \sum_{k=0}^{\lfloor Nt\rfloor - 1} \langle f, R_k^N\rangle - \frac{\sqrt{N}}{N^{1+\beta}} \sum_{k=0}^{\lfloor Nt\rfloor - 1} \sum_{i=1}^{N} \nabla f(W_k^i) \cdot \varepsilon_k^i.\end{aligned}$$

In what follows, $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $t \in \mathbf{R}_+$ are fixed.

Step 1. In this step we study the continuity of the mapping

$$\mathbf{B}_t[f] : m \in \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)) \mapsto \langle f, m_t\rangle - \int_0^t (\mathbf{U}_s[f](m_s) - \mathbf{V}_s[f](m_s))\mathrm{d}s \in \mathbf{R}.$$

Let $(m^N)_{N\geq 1}$ such that $m^N \to m$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$. Recall that $\sup_{x\in\mathcal{X}} \|\sigma_*(\cdot, x)\|_{\mathcal{H}^{J_0-1,j_0}} < +\infty$ (by **A2** and because $j_0 > d/2$). Then, for all $N \geq 1$ and $s \in [0, t]$, it holds:

$$\left|(y - \langle \sigma_*(\cdot, x), \bar{\mu}_s\rangle)\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), m_s^N\rangle\right| \leq C(|y| + 1)\|f\|_{\mathcal{H}^{J_0,j_0}} \sup_{N\geq 1} \sup_{s\in[0,t]} \|m_s^N\|_{\mathcal{H}^{-J_0+1,j_0}} < +\infty$$

and since $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{1,\gamma}(\mathbf{R}^d)$,

$$\begin{aligned}|\langle \sigma_*(\cdot, x), m_s^N\rangle\langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \bar{\mu}_s\rangle| \leq C \sup_{N\geq 1} \sup_{s\in[0,t]} &\|m_s^N\|_{\mathcal{H}^{-J_0+1,j_0}} \\ &\times \|f\|_{\mathcal{H}^{J_0,j_0}} \sup_{s\in[0,t]} |\langle (1 + |\cdot|^\gamma), \bar{\mu}_s\rangle| < +\infty,\end{aligned}$$

for some $C > 0$ independent of $N \geq 1$, $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$, $s \geq 0$, and $(x,y) \in \mathcal{X} \times \mathcal{Y}$. With these two upper bounds, and using the same arguments as those used in the proof of Lemma 2.20, one deduces that for all continuity points $t \in \mathbf{R}_+$ of $\{t \mapsto m_t, t \in \mathbf{R}_+\}$, we have $\mathbf{B}_t[f](m^N) \to \mathbf{B}_t[f](m)$ as $N \to +\infty$. Consequently, using (2.3.48) and the continuous mapping theorem [Bil99, Theorem 2.7], for all $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $t \in \mathbf{R}_+$, it holds in distribution and as $N' \to +\infty$:

$$\mathbf{B}_t[f](\eta^{N'}) - \langle f, \eta_0^{N'} \rangle - \langle f, \sqrt{N'} M_t^{N'} \rangle \to \mathbf{B}_t[f](\eta^*) - \langle f, \eta_0^* \rangle - \langle f, \mathscr{G}_t^* \rangle. \tag{2.3.53}$$

Step 2. In this step we prove that for any $t \in \mathbf{R}_+$ and $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$:

$$\mathbf{E}\big[\big|\mathbf{e}_t^N[f]\big|\big] \to 0 \text{ as } N \to +\infty. \tag{2.3.54}$$

By (2.3.24), we have since $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d) \hookrightarrow \mathcal{H}^{J_2+1,j_2}(\mathbf{R}^d)$ (because $J_0 \geq J_2 + 1$ and $j_2 \geq j_0$, see (2.1.12) and (2.1.13)),

$$\frac{1}{\sqrt{N}} \mathbf{E}\Big[\int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha |\langle \sigma_*(\cdot,x), \eta_s^N \rangle| |\langle \nabla f \cdot \nabla \sigma_*(\cdot,x), \eta_s^N \rangle| \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s\Big] \leq \frac{Ct \|\nabla f\|_{\mathcal{H}^{J_2,j_2}}}{\sqrt{N}} \mathbf{E}\Big[\sup_{t\in[0,T]} \|\eta_t^N\|_{\mathcal{H}^{-J_2,j_2}}^2\Big]$$
$$\leq \frac{Ct \|f\|_{\mathcal{H}^{J_2+1,j_2}}}{\sqrt{N}} \to 0 \text{ as } N \to +\infty.$$

Using Lemma 2.12, we also have since $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$,

$$\mathbf{E}\Big[\Big|\sqrt{N}\langle f, V_t^N \rangle - \sqrt{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \langle f, R_k^N \rangle - \frac{\sqrt{N}}{N^{1+\beta}} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \sum_{i=1}^{N} \nabla f(W_k^i) \cdot \varepsilon_k^i \Big|\Big] \leq C \|f\|_{\mathcal{C}^{2,\gamma_*}} \Big[\frac{\sqrt{N}}{N}$$
$$+ \sqrt{N} N \big[\frac{1}{N^2} + \frac{1}{N^{2\beta}}\big] + \frac{\sqrt{N}}{N^\beta}\Big].$$

The right-hand-side of the previous term goes to 0 as $N \to +\infty$, since $\beta > 3/4$. This proves (2.3.54).

Step 3. End of the proof of Proposition 2.39. By (2.3.53), (2.3.54), and (2.3.52), we deduce that for all $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $t \in \mathbf{R}^+$, it holds a.s. $\mathbf{B}_t[f](\eta^*) - \langle f, \eta_0^* \rangle - \langle f, \mathscr{G}_t^* \rangle = 0$. Since $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $\mathbf{R}_+$ are separable, we conclude by a standard continuity argument that a.s. for all $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $t \in \mathbf{R}^+$, $\mathbf{B}_t[f](\eta^*) - \langle f, \eta_0^* \rangle - \langle f, \mathscr{G}_t^* \rangle = 0$. Hence, η^* is a weak solution of (2.1.16) (see Definition 2.7) with initial distribution ν_0 (see (2.3.49)). This ends the proof of Proposition 2.39. □

Inspired by the proof of [DLR19a, Corollary 5.7] (see also [KX04]), to end the proof of Theorem 2.8, we will show that (2.1.16) has a unique strong solution. This is the purpose of the next section, where we also conclude the proof of Theorem 2.8.

Pathwise uniqueness

Proposition 2.40. *Let $\beta > 3/4$ and assume* **A1**-**A7**. *Then strong uniqueness holds for* (2.1.16), *namely, on a fixed probability space, given a $\mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d)$-valued random variable ν and a* G*-process $\mathscr{G} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R}^d))$, there exists at most one $\mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$-valued process η solution to* (2.1.16) *with $\eta_0 = \nu$ almost surely.*

Proof. By linearity of the involved operators in (2.1.16), it is enough to consider a $\mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$-valued process η solution to (2.1.16) when a.s. $\nu = 0$ and $\mathscr{G} = 0$, i.e., for every $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $t \in \mathbf{R}_+$:

$$\begin{cases} \langle f, \eta_t \rangle - \int_0^t (\mathbf{U}_s[f](\eta_s) - \mathbf{V}_s[f](\eta_s)) \mathrm{d}s = 0, \\ \langle f, \eta_0 \rangle = 0, \end{cases} \tag{2.3.55}$$

where $\mathbf{U}$ and $\mathbf{V}$ are defined respectively in (2.3.50) and (2.3.51). Pick $T > 0$. By (2.3.55), a.s. for all $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $t \in [0,T]$, we have $\langle f, \eta_t\rangle^2 = 2\int_0^t (\mathbf{U}_s[f](\eta_s) - \mathbf{V}_s[f](\eta_s))\langle f, \eta_s\rangle \mathrm{d}s$. Recall $\sup_{s\in[0,T]} |\langle (1+|\cdot|^\gamma), \bar{\mu}_s\rangle| < +\infty$ (because $\bar{\mu} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_\gamma(\mathbf{R}^d))$) and $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{1,\gamma}(\mathbf{R}^d)$. Then, by Cauchy-Schwarz inequality, a.s. for all $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$ and $t \in [0,T]$:

$$
\begin{aligned}
-\int_0^t \mathbf{V}_s[f](\eta_s)\langle f, \eta_s\rangle \mathrm{d}s &\leq \alpha \int_0^t \Big[\langle f, \eta_s\rangle^2 + \int_{\mathcal{X}\times\mathcal{Y}} \langle \sigma_*(\cdot, x), \eta_s\rangle^2 \langle \nabla f \cdot \nabla \sigma_*(\cdot, x), \bar{\mu}_s\rangle^2 \pi(\mathrm{d}x, \mathrm{d}y)\Big]\mathrm{d}s \\
&\leq C \int_0^t \Big[\langle f, \eta_s\rangle^2 + \|\eta_s\|^2_{\mathcal{H}^{-J_0,j_0}} \|f\|^2_{\mathcal{H}^{L,\gamma}}\Big]\mathrm{d}s. \qquad (2.3.56)
\end{aligned}
$$

Let $\{f_a\}_{a\geq 1}$ be an orthonormal basis of $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$. Using the operator $\mathsf{T}_x : f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d) \mapsto \nabla f \cdot \nabla \sigma(\cdot, x) \in \mathcal{H}^{J_0-1,j_0}(\mathbf{R}^d)$ defined for all $x \in \mathcal{X}$ and Lemma 2.43, we have a.s. for all $t \in [0,T]$:

$$
\begin{aligned}
\sum_{a\geq 1} \int_0^t \mathbf{U}_s[f_a](\eta_s)\langle f_a, \eta_s\rangle \mathrm{d}s &= \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} (y - \langle \sigma_*(\cdot, x), \bar{\mu}_s\rangle)\Big(\sum_{a\geq 1} \langle f_a, \eta_s\rangle \langle \mathsf{T}_x f_a, \eta_s\rangle\Big)\pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&= \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} (y - \langle \sigma_*(\cdot, x), \bar{\mu}_s\rangle)\langle \eta_s, \mathsf{T}_x^* \eta_s\rangle_{\mathcal{H}^{-J_0,j_0}} \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&\leq C \int_0^t \|\eta_s\|^2_{\mathcal{H}^{-J_0,j_0}} \mathrm{d}s. \qquad (2.3.57)
\end{aligned}
$$

Therefore, using the bounds (2.3.56) and (2.3.57), together with $\mathcal{H}^{J_0,j_0}(\mathbf{R}^d) \hookrightarrow_{\text{H.S.}} \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$, we have a.s. for all $t \in [0,T]$:

$$
\|\eta_t\|^2_{\mathcal{H}^{-J_0,j_0}} = \sum_{a\geq 1} \langle f_a, \eta_t\rangle^2 \leq C \int_0^t \|\eta_s\|^2_{\mathcal{H}^{-J_0,j_0}} \mathrm{d}s.
$$

By Gronwall's lemma, a.s. $\|\eta_t\|_{\mathcal{H}^{-J_0,j_0}} = 0$ for all $t \in [0,T]$. This concludes the proof of Proposition 2.40. □

End of the proof of Theorem 2.8

Proof of Theorem 2.8. Let $\ell \in \{1,2\}$ and N_ℓ be such that in distribution $\eta^{N_\ell} \to \eta^\ell$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$ (see Proposition 2.27). By Lemma 2.31, a.s. $\eta^\ell \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$. Consider now $(\eta^{\ell,*}, \mathscr{G}^{\ell,*})$ a limit point of $(\eta^{N_\ell}, \sqrt{N_\ell}\, M^{N_\ell})_{N_\ell \geq 1}$ in $\mathscr{E}$. Up to extracting a subsequence from N_ℓ, we assume that in distribution and as $N_\ell \to +\infty$,

$$
(\eta^{N_\ell}, \sqrt{N_\ell}\, M^{N_\ell}) \to (\eta^{\ell,*}, \mathscr{G}^{\ell,*}) \text{ in } \mathscr{E}.
$$

Considering the marginal distributions, we then have by uniqueness of the limit in distribution, for $\ell = 1, 2$ (see also Proposition 2.35):

$$
\eta^{\ell,*} = \eta^\ell, \text{ and } \mathscr{G}^{\ell,*} = \mathscr{G} \text{ in distribution.} \qquad (2.3.58)
$$

By Proposition 2.39, $\eta^{1,*}$ and $\eta^{2,*}$ are two weak solutions of (2.1.16) with initial distribution ν_0 (see also Lemma 2.38). Since strong uniqueness for (2.1.16) (see Proposition 2.40) implies weak uniqueness for (2.1.16), we deduce that $\eta^{1,*} = \eta^{2,*}$ in distribution. By (2.3.58), this implies that $\eta^1 = \eta^2$ in distribution and then, that the whole sequence $(\eta^N)_{N\geq 1}$ converges in distribution in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$. Denoting by η^* its limit, we have proved that η^* has the same distribution as the unique weak solution $\eta^\star$ of (2.1.16) with initial distribution ν_0. This concludes the proof of Theorem 2.8.

□

2.3.6 The case when $\beta = 3/4$

In this section, we assume that $d = 1$. Recall $\mathsf{f}_2 : x \in \mathbf{R} \mapsto |x|^2$, which belongs to $\mathcal{H}^{J_0,j_0}(\mathbf{R})$ because $j_0 - 2 = 3 - 2 > 1/2$ (see (2.1.12)).

Proof of Proposition 2.10. Assume $\beta = 3/4$. The proof of Proposition 2.10 is divided into two steps.

Step 1. Let $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R})$. When $\beta = 3/4$, it appears that for non affine test functions f, the term $\langle f, R_k^N\rangle$ is not negligible any more. In this step we simply rewrite (2.3.52) by decomposing the term $\langle f, R_k^N\rangle$ into two terms: $\langle f, R_k^N\rangle = \langle f, \mathscr{R}_k^N\rangle + \langle f, \mathscr{B}_k^N\rangle$ where $\langle f, \mathscr{R}_k^N\rangle$ will be negligible and $\langle f, \mathscr{B}_k^N\rangle$ will not be negligible. More precisely, by (2.2.2) and (2.1.2), it holds

$$\begin{aligned}\langle f, R_k^N\rangle &= \frac{1}{2N}\sum_{i=1}^N (W_{k+1}^i - W_k^i)^2 f(\widehat{W}_k^i)\\ &= \frac{1}{2N}\sum_{i=1}^N \Big|\frac{\alpha}{N|B_k|}\sum_{(x,y)\in B_k}(y - g_{W_k}^N(x))\nabla_W\sigma_*(W_k^i, x) + \frac{\varepsilon_k^i}{N^{3/4}}\Big|^2 f''(\widehat{W}_k^i)\\ &= \langle f, \mathscr{R}_k^N\rangle + \langle f, \mathscr{B}_k^N\rangle,\end{aligned}$$

where

$$\begin{aligned}\langle f, \mathscr{R}_k^N\rangle = \frac{1}{2N}\sum_{i=1}^N \Big|\frac{\alpha}{N|B_k|}\sum_{(x,y)\in B_k}(y - g_{W_k}^N(x))\nabla_W\sigma_*(W_k^i, x)\Big|^2 f''(\widehat{W}_k^i)\\ + \frac{1}{N}\sum_{i=1}^N \frac{\alpha}{N|B_k|}\sum_{(x,y)\in B_k}(y - g_{W_k}^N(x))\nabla_W\sigma_*(W_k^i, x)\frac{\varepsilon_k^i}{N^{3/4}} f''(\widehat{W}_k^i)\end{aligned}$$

and

$$\langle f, \mathscr{B}_k^N\rangle = \frac{1}{2N^{5/2}}\sum_{i=1}^N |\varepsilon_k^i|^2 f''(\widehat{W}_k^i).$$

From (2.3.52), one then has:

$$\langle f, \eta_t^N\rangle - \langle f, \eta_0^N\rangle - \int_0^t (\mathbf{U}_s[f](\eta_s^N) - \mathbf{V}_s[f](\eta_s^N))\mathrm{d}s - \langle f, \sqrt{N}M_t^N\rangle = -\tilde{\mathbf{e}}_t^N[f] + \sqrt{N}\sum_{k=0}^{\lfloor Nt\rfloor - 1}\langle f, \mathscr{B}_k^N\rangle, \tag{2.3.59}$$

where

$$\begin{aligned}\tilde{\mathbf{e}}_t^N[f] :&= \frac{1}{\sqrt{N}}\int_0^t\int_{\mathcal{X}\times\mathcal{Y}} \alpha\langle\sigma_*(\cdot, x), \eta_s^N\rangle\langle\nabla f\cdot\nabla\sigma_*(\cdot, x), \eta_s^N\rangle\pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s\\ &\quad - \sqrt{N}\langle f, V_t^N\rangle - \sqrt{N}\sum_{k=0}^{\lfloor Nt\rfloor - 1}\langle f, \mathscr{R}_k^N\rangle - \frac{\sqrt{N}}{N^{1+\beta}}\sum_{k=0}^{\lfloor Nt\rfloor - 1}\sum_{i=1}^N \nabla f(W_k^i)\,\varepsilon_k^i.\end{aligned}$$

Step 2. Let η be a limit point of $(\eta^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}))$. Let N' be such that in distribution $\eta^{N'} \to \eta$ as $N' \to +\infty$. In this step, we pass to the limit in (2.3.59) with the test function $\mathsf{f}_2 : x \in \mathbf{R} \mapsto |x|^2$. By Propositions 2.27 and 2.30, the sequence $(\eta^{N'}, \sqrt{N'}M^{N'})_{N'\geq 1}$ is tight in $\mathscr{E}$ (see (2.3.47)). Let $(\eta^*, \mathscr{G}^*)$ be one of its limit point in $\mathscr{E}$. Up to extracting a subsequence from N', it holds:

$$(\eta^{N'}, \sqrt{N'}\,M^{N'}) \to (\eta^*, \mathscr{G}^*), \text{ as } N' \to +\infty.$$

Considering the marginal distributions, it holds in distribution,

$$\eta^* = \eta \text{ and } \mathscr{G}^* = \mathscr{G} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_0,j_0}(\mathbf{R})).$$

Introduce $\mathcal{C}(\eta^*) \subset \mathbf{R}_+$, whose complementary in $\mathbf{R}_+$ is at most countable, such that for all $u \in \mathcal{C}(\eta^*)$, $s \in \mathbf{R}_+ \mapsto \eta^*_s \in \mathcal{H}^{-J_0+1,j_0}(\mathbf{R})$ is a.s. continuous at u. Then, with the same arguments as those used to derive (2.3.53) and using also the fact that $0 \in \mathcal{C}(\eta^*)$, one has for all $t \in \mathcal{C}(\eta^*)$ and in distribution,

$$\mathbf{B}_t[f](\eta^{N'}) - \langle f, \eta_0^{N'}\rangle - \langle f, \sqrt{N'} M_t^{N'}\rangle \to \mathbf{B}_t[f](\eta^*) - \langle f, \eta_0^*\rangle - \langle f, \mathscr{G}_t^*\rangle \text{ as } N' \to +\infty. \quad (2.3.60)$$

Let us now deal with the two terms in the right-hand side of (2.3.59). Using (3.9.3), (2.2.10) and **A3**,

$$\begin{aligned}\mathbf{E}\Big[\frac{1}{|B_k|}\Big|\sum_{(x,y)\in B_k}(y - g^N_{W_k}(x))\nabla_W\sigma_*(W_k^i, x)\varepsilon_k^i\Big|\Big] &\le 2\mathbf{E}\Big[\frac{1}{|B_k|}\sum_{(x,y)\in B_k}\big|(y - g^N_{W_k}(x))\nabla_W\sigma_*(W_k^i, x)\big|^2\Big] \\ &\quad + 2\mathbf{E}\Big[|\varepsilon_k^i|^2\Big] \\ &\le C\Big[\mathbf{E}\Big[\frac{1}{|B_k|}\sum_{(x,y)\in B_k}(|y|^2+1)\Big] + 1\Big] \le C.\end{aligned}$$

We now set $f = \mathsf{f}_2$. Then, we have $\mathbf{E}\big[|\langle \mathsf{f}_2, \mathscr{R}_k^N\rangle|\big] \le C(N^{-2} + N^{-7/4})$. Using also the lines below (2.3.54) and Lemma 2.12, it holds:

$$\mathbf{E}\big[|\tilde{\mathbf{e}}_t^N[\mathsf{f}_2]|\big] \le C\Big[\frac{1}{\sqrt{N}} + N^{3/2}\Big(\frac{1}{N^2} + \frac{1}{N^{7/4}}\Big) + \frac{1}{N^{1/4}}\Big]. \quad (2.3.61)$$

On the other hand, using **A5** and the law of large number, it holds a.s. as $N \to +\infty$,

$$\sqrt{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle \mathsf{f}_2, \mathscr{B}_k^N\rangle = \frac{1}{2N^2}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}|\varepsilon_k^i|^2\mathsf{f}_2''(\widehat{W}_k^i) = \frac{1}{N^2}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{i=1}^{N}|\varepsilon_k^i|^2 \to t\mathbf{E}[|\varepsilon_1^1|^2]. \quad (2.3.62)$$

Therefore, using (2.3.60),(2.3.61) and (2.3.62), it holds for all $t \in \mathcal{C}(\eta^*)$, a.s. $\mathbf{B}_t[\mathsf{f}_2](\eta^*) - \langle \mathsf{f}_2, \eta_0^*\rangle - \langle \mathsf{f}_2, \mathscr{G}_t^*\rangle = t\mathbf{E}[|\varepsilon_1^1|^2]$. The mapping[3] $s \in \mathbf{R}_+ \mapsto \mathbf{B}_s[\mathsf{f}_2](\eta^*)$ is right continuous and $s \mapsto \langle \mathsf{f}_2, \mathscr{G}_s^*\rangle$ is continuous. By a standard continuity argument (the same as the one used in Proposition 2.21), it holds a.s. for all $t \in \mathbf{R}_+$, $\mathbf{B}_t[\mathsf{f}_2](\eta^*) - \langle \mathsf{f}_2, \eta_0^*\rangle - \langle \mathsf{f}_2, \mathscr{G}_t^*\rangle = t\mathbf{E}[|\varepsilon_1^1|^2]$. The proof of Proposition 2.10 is complete. □

Acknowledgment. The authors are grateful to Benoît Bonnet for fruitful discussions about the proof of Proposition 2.22. A.D. is grateful for the support received from the Agence Nationale de la Recherche (ANR) of the French government through the program "Investissements d'Avenir" (16-IDEX-0001 CAP 20-25). A.G. is supported by the French ANR under the grant ANR-17-CE40-0030 (project *EFI*) and the Institut Universitaire de France. M.M. acknowledges the support of the the French ANR under the grant ANR-20-CE46-0007 (*SuSa* project). B.N. is supported by the grant IA20Nectoux from the Projet I-SITE Clermont CAP 20-25.

[3]For all $m \in \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_0+1,j_0}(\mathbf{R}^d))$ and $f \in \mathcal{H}^{J_0,j_0}(\mathbf{R}^d)$, $t \in \mathbf{R}_+ \mapsto \mathbf{B}_t[f](m)$ is right-continuous. This is clear since $t \mapsto \langle f, m_t\rangle$ is right-continuous, and because $s \mapsto \mathbf{U}_s[f](m_s) - \mathbf{V}_s[f](m_s) \in L^\infty_{loc}(\mathbf{R}_+)$.

2.4 A note on relative compactness

In this section, we prove, in our Hilbert setting, that the condition (4.21) in [Kur75, Theorem (4.20)] can be replaced by the slightly modified condition, namely the regularity condition of item 2 in Proposition 2.41 below.

In the following $\mathcal{H}_1$ and $\mathcal{H}_2$ are two Hilbert spaces (whose duals are respectively denoted by $\mathcal{H}_1^{-1}$ and $\mathcal{H}_2^{-1}$) such that $\mathcal{H}_1 \hookrightarrow_{\text{H.S.}} \mathcal{H}_2$.

Proposition 2.41. *Let $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}_2^{-1})$ be a sequence of processes satisfying the following two conditions:*

1. *Compact containment condition : for every $T > 0$ and $\eta > 0$, there exists $C > 0$ such that*

$$\sup_{N\geq 1} \mathbf{P}\big(\sup_{t\in[0,T]} \|\mu_t^N\|^2_{\mathcal{H}_2^{-1}} > C\big) \leq \eta.$$

2. *Regularity condition : for every $\delta > 0$, $N \geq 1$, $0 \leq t \leq T$ and $0 \leq u \leq (T-t) \wedge \delta$, there exists $F_N(\delta) < \infty$ such that*

$$\mathbf{E}\left[\|\mu_{t+u}^N - \mu_t^N\|^2_{\mathcal{H}_2^{-1}}\right] \leq F_N(\delta),$$

with $\lim\limits_{\delta\to 0} \limsup\limits_{N\geq 1} F_N(\delta) = 0$.

Then, the sequence $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}_1^{-1})$ is relatively compact.

Proof. To prove this result, we follow the proof of [Kur75, Theorem (4.20)]. More precisely, we show that the assumptions of [Kur75, Theorem 1] are satisfied (namely conditions (4.2) and (4.3) there), when $E = \mathcal{H}_1^{-1}$ there.

Step 1. The condition (4.2) in [Kur75, Theorem 1] is satisfied (when $E = \mathcal{H}_1^{-1}$ there).

We have that $\mathcal{H}_1$ is compactly embedded in $\mathcal{H}_2$ (since a Hilbert-Schmidt embedding is compact). By Schauder's theorem, $\mathcal{H}_2^{-1}$ is compactly embedded in $\mathcal{H}_1^{-1}$. Thus, for all $C > 0$, the set $\{\phi \in \mathcal{H}_1^{-1},\ \|\phi\|_{\mathcal{H}_2^{-1}} \leq C\}$ is compact. Therefore, the condition (4.2) in [Kur75] is satisfied.

Step 2. The condition (4.3) in [Kur75, Theorem 1] is satisfied (when $E = \mathcal{H}_1^{-1}$ there).

By [Kur75, Lemma 4.4], $\{t \mapsto \mu_t^N, t \in \mathbf{R}_+\}_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}_1^{-1})$ if for all $\varepsilon > 0$, there exists a tight sequence $\{t \mapsto \mu_t^{N,\varepsilon}, t \in \mathbf{R}_+\}_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}_1^{-1})$ which is ε-close to $\{t \mapsto \mu_t^N, t \in \mathbf{R}_+\}_{N\geq 1}$. Following [Kur75], we define, for $\varepsilon > 0$, the sequence $\{t \mapsto \mu_t^{N,\varepsilon}, t \in \mathbf{R}_+\}_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}_1^{-1})$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}_2^{-1})$ of pure jump processes as follows. Let us first introduce, for $N \geq 1$ and $\varepsilon > 0$, $\tau_0^{N,\varepsilon} := 0$ and, for $k > 0$:

$$\tau_k^{N,\varepsilon} := \inf\{t > \tau_{k-1}^{N,\varepsilon} : \|\mu_t^N - \mu^N_{\tau_{k-1}^{N,\varepsilon}}\|_{\mathcal{H}_2^{-1}} > \varepsilon\},$$
$$s_k^{N,\varepsilon} := \sup\{t < \tau_k^{N,\varepsilon} : \|\mu_t^N - \mu^N_{\tau_k^{N,\varepsilon}}\|_{\mathcal{H}_2^{-1}} \geq \varepsilon\}.$$

Then we define, for $\varepsilon > 0$,

$$\mu_t^{N,\varepsilon} := \begin{cases} \mu_0^N & \text{for} \quad t < \frac{1}{2}(s_1^{N,\varepsilon} + \tau_1^{N,\varepsilon}) \\ \mu^N_{\tau_k^{N,\varepsilon}} & \text{for } \frac{1}{2}(s_k^{N,\varepsilon} + \tau_k^{N,\varepsilon}) \leq t < \frac{1}{2}(s_{k+1}^{N,\varepsilon} + \tau_{k+1}^{N,\varepsilon}). \end{cases}$$

We claim that for any $\varepsilon > 0$, the sequence $\{t \mapsto \mu_t^{N,\varepsilon}, t \in \mathbf{R}_+\}_{N\geq 1}$ verifies condition (4.2) of [Kur75, Theorem 4.1] when $E = \mathcal{H}_1^{-1}$ there. Indeed, by the discussion in the first step above, this

follows from the compact containment condition verified by $\{t \mapsto \mu_t^N, t \in \mathbf{R}_+\}_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}_2^{-1})$ (see item 1 in Proposition 2.41) together with the fact that $\sup_{t\in\mathbf{R}_+} \|\mu_t^N - \mu_t^{N,\varepsilon}\|_{\mathcal{H}_2^{-1}} \leq C_0\varepsilon$, where the constant $C_0 > 0$ is independent of N and ε.

It remains to prove that for any $\varepsilon > 0$, $\{t \mapsto \mu_t^{N,\varepsilon}, t \in \mathbf{R}_+\}_{N\geq 1}$ satisfies the condition (4.3) in [Kur75, Theorem 4.1] when $E = \mathcal{H}_1^{-1}$ there (so that it will be tight for each $\varepsilon > 0$). Since $\mathcal{H}_2^{-1} \hookrightarrow \mathcal{H}_1^{-1}$, it is enough to show that for any $\varepsilon > 0$, $\{t \mapsto \mu_t^{N,\varepsilon}, t \in \mathbf{R}_+\}_{N\geq 1}$ satisfies the condition (4.3) in [Kur75, Theorem 4.1] when $E = \mathcal{H}_2^{-1}$ there. By [Kur75, Lemma 4.5] (and its note) and the construction of $\{t \mapsto \mu_t^{N,\varepsilon}, t \in \mathbf{R}_+\}_{N\geq 1}$, it is sufficient to bound $\delta \mapsto \mathbf{P}(\tau_1^{N,\varepsilon} \leq \delta)$ and $\delta \mapsto \mathbf{P}(\tau_{k+1}^{N,\varepsilon} - s_k^{N,\varepsilon} \leq \delta)$ by a function $\delta \mapsto G_N^\varepsilon(\delta)$ such that for every $\varepsilon > 0$,

$$\lim_{\delta\to 0} \limsup_N G_N^\varepsilon(\delta) = 0. \tag{2.4.1}$$

Let us prove that we can indeed bound these two terms by such $G_N^\varepsilon(\delta)$ satisfying (2.4.1). Recall that by item 2 in Proposition 2.41, for every $\delta > 0$, $N \geq 1$, $0 \leq t \leq T$ and $0 \leq u \leq (T-t) \wedge \delta$,

$$\mathbf{E}\left[\|\mu_{t+u}^N - \mu_t^N\|_{\mathcal{H}_2^{-1}}^2\right] \leq F_N(\delta), \text{ with } \lim_{\delta\to 0} \limsup_{N\geq 1} F_N(\delta) = 0.$$

Now, introduce as in [Kur75], the distance r on $\mathcal{H}_2^{-1}$ defined by $r(\varphi^1, \varphi^2) = \min(1, \|\varphi^1 - \varphi^2\|_{\mathcal{H}_2^{-1}})$. We thus have:

$$\mathbf{E}\left[r\left(\mu_{t+u}^N, \mu_t^N\right)^2\right] \leq F_N(\delta), \tag{2.4.2}$$

On the one hand, we have, for $0 < \varepsilon < 1$,

$$\mathbf{P}(\tau_1^{N,\varepsilon} \leq \delta) = \mathbf{P}\big(\|\mu^N_{\tau_1^{N,\varepsilon}\wedge\delta} - \mu_0^N\|_{\mathcal{H}_2^{-1}} > \varepsilon\big) = \mathbf{P}\big(r\big(\mu^N_{\tau_1^{N,\varepsilon}\wedge\delta}, \mu_0^N\big) > \varepsilon\big) \leq \frac{1}{\varepsilon^2}\mathbf{E}\big[r\big(\mu^N_{\tau_1^{N,\varepsilon}\wedge\delta}, \mu_0^N\big)^2\big]. \tag{2.4.3}$$

On the other hand, we have, for $k \geq 1$ and $0 < \varepsilon < 1$,

$$\begin{aligned}\mathbf{P}(\tau_{k+1}^{N,\varepsilon} - s_k^{N,\varepsilon} \leq \delta) \leq \mathbf{P}(\tau_{k+1}^{N,\varepsilon} - \tau_k^{N,\varepsilon} \leq \delta) &= \mathbf{P}\big(r(\mu^N_{\tau_{k+1}^{N,\varepsilon}\wedge(\tau_k^{N,\varepsilon}+\delta)}, \mu^N_{\tau_k^{N,\varepsilon}}) > \varepsilon\big) \\ &\leq \frac{1}{\varepsilon^2}\mathbf{E}\big[r\big(\mu^N_{\tau_{k+1}^{N,\varepsilon}\wedge(\tau_k^{N,\varepsilon}+\delta)}, \mu^N_{\tau_k^{N,\varepsilon}}\big)^2\big].\end{aligned} \tag{2.4.4}$$

From (2.4.2), and because the stopping times appearing in (2.4.3) and (2.4.4) can be approximated by sequences of decreasing discrete stopping times, we can indeed bound $\delta \mapsto \mathbf{P}(\tau_1^{N,\varepsilon} \leq \delta)$ and $\delta \mapsto \mathbf{P}(\tau_{k+1}^{N,\varepsilon} - s_k^{N,\varepsilon} \leq \delta)$ by $F_N(\delta)$ which satisfies (2.4.1). Consequently, for each $0 < \varepsilon < 1$, the condition (4.3) in [Kur75, Theorem 4.1] is satisfied for the sequence $\{t \mapsto \mu_t^{N,\varepsilon}, t \in \mathbf{R}_+\}_{N\geq 1}$ when $E = \mathcal{H}_2^{-1}$ there. We can thus apply [Kur75, Theorem 4.1] to $\{t \mapsto \mu_t^{N,\varepsilon}, t \in \mathbf{R}_+\}_{N\geq 1}$, which is therefore tight in $\mathcal{D}(\mathbf{R}_+, \mathcal{H}_1^{-1})$ for each $0 < \varepsilon < 1$. Using [Kur75, Lemma 4.4] (with $E = \mathcal{H}_1^{-1}$ there), this concludes the proof of the lemma. □

2.5 Technical lemmata

In this section we state and prove Lemma 2.42, Lemma 2.43 and Lemma 2.44.

Lemma 2.42. *Let $\beta \geq 1/2$ and assume* **A1**-**A7**. *Recall $\mathcal{H}^{J_1,j_1}(\mathbf{R}^d) \hookrightarrow \mathcal{H}^{L,\gamma}(\mathbf{R}^d)$ (see (2.1.14)). Then, for all $T > 0$, there exists $C > 0$ such that for all $f \in \mathcal{H}^{J_1,j_1}(\mathbf{R}^d)$, $N \geq 1$, and $t \in [0,T]$, it holds:*

(i) $\mathbf{E}\left[\sum_{k=0}^{\lfloor Nt\rfloor-1} 2\langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle\sqrt{N}\langle f, M^N_k\rangle + 4N\langle f, M^N_k\rangle^2\right] \leq C\|f\|^2_{\mathcal{H}^{L,\gamma}}$.

(ii) $\mathbf{E}\left[\sum_{k=0}^{\lfloor Nt\rfloor-1} \frac{2\sqrt{N}}{N^{1+\beta}}\langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle \sum_{i=1}^N \nabla f(W^i_k)\cdot\varepsilon^i_k + \frac{4}{N^{1+2\beta}}\Big|\sum_{i=1}^N \nabla f(W^i_k)\cdot\varepsilon^i_k\Big|^2\right] \leq \frac{C}{N^{2\beta-1}}\|f\|^2_{\mathcal{H}^{L,\gamma}}$.

(iii) $\mathbf{E}\left[\sum_{k=0}^{\lfloor Nt\rfloor-1} N|\langle f, R^N_k\rangle|^2\right] \leq C\|f\|^2_{\mathcal{H}^{L,\gamma}}\left[\frac{1}{N^2} + \frac{N^2}{N^{4\beta}}\right]$.

(iv) $\mathbf{E}\left[\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f, \Upsilon_{\frac{k+1}{N}-}\rangle\sqrt{N}\langle f, R^N_k\rangle\right] \leq C\|f\|^2_{\mathcal{H}^{L,\gamma}}\left[1 + \frac{1}{N} + \frac{N^3}{N^{4\beta}}\right] + \mathbf{E}\left[\int_0^t\langle f, \Upsilon^N_s\rangle^2 \mathrm{d}s\right]$.

(v) $\mathbf{E}\left[\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle \mathbf{a}^N_k[f] - \sqrt{N}\int_0^t\langle f, \Upsilon^N_s\rangle \mathbf{L}^N_s[f]\mathrm{d}s\Big|\right] \leq C\|f\|^2_{\mathcal{H}^{L,\gamma}}$.

(vi) $\mathbf{E}\left[\sum_{k=0}^{\lfloor Nt\rfloor-1}|\mathbf{a}^N_k[f]|^2\right] \leq C\|f\|^2_{\mathcal{H}^{L,\gamma}}$.

Proof. Let $T > 0$ and $f \in \mathcal{H}^{J_1,j_1}(\mathbf{R}^d)$. In what follows, $C > 0$ is a constant, independent of $N \geq 1$, $t \in [0,T]$, f, and $k \in \{0, \ldots, \lfloor NT\rfloor - 1\}$ which can change from one occurence to another. We recall that for $N \geq 1$ and $k \geq 1$, $\mathcal{F}^N_k$ is the $\boldsymbol{\sigma}$-algebra generated by $\{W^i_0\}_{i=1}^N$, B_j and $(\varepsilon^i_j)_{i=1}^N$ for $j = 0, \ldots, k-1$, and that $\mathcal{F}^N_0 := \boldsymbol{\sigma}\{\{W^i_0\}_{i=1}^N\}$, see (2.1.9). Recall also the definitions of M^N_k and R^N_k in (2.2.4) and (2.2.2) respectively, of $\mathbf{a}^N_s$ and $\mathbf{L}^N_s$ in (2.3.16), and that for $N \geq 1$ and $t \in \mathbf{R}_+$, $\Upsilon^N_t = \sqrt{N}(\mu^N_t - \bar{\mu}^N_t)$ (see also (2.3.3)). We start by proving item *(i)* in Lemma 2.42. For all $t \in [0,T]$, because for all $a \in \mathbf{N}$ and $b \in \{1, \ldots, N\}$, W^b_a is $\mathcal{F}^N_a$-measurable and $\varepsilon^b_a \perp\!\!\!\perp \mathcal{F}^N_a$ (see **A5**) together with the fact that $\bar{X}^b_s$ is $\mathcal{F}^N_0$-measurable (for all $s \geq 0$), one has using also (2.2.24):

$$\begin{aligned}\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle\sqrt{N}\langle f, M^N_k\rangle\Big] &= N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}\Big[\langle f, \mu^N_{\frac{k+1}{N}-} - \bar{\mu}^N_{\frac{k+1}{N}}\rangle\langle f, M^N_k\rangle\Big] \\ &= N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}\big[\langle f, \nu^N_k\rangle\mathbf{E}\left[\langle f, M^N_k\rangle\big|\,\mathcal{F}^N_k\right]\big] \\ &\quad - N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}\Big[\langle f, \bar{\mu}^N_{\frac{k+1}{N}}\rangle\mathbf{E}\Big[\langle f, M^N_k\rangle\big|\mathcal{F}^N_k\Big]\Big] = 0. \qquad (2.5.1)\end{aligned}$$

By (2.2.19) $\mathbf{E}\left[\langle f, M^N_k\rangle^2\right] \leq C\|f\|^2_{\mathcal{H}^{L,\gamma}}/N^2$. Together with (2.5.1), we deduce item *(i)*.

Let us now prove item *(ii)*. We have, using **A5**,

$$\begin{aligned}\mathbf{E}\big[\langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle\sum_{i=1}^N\nabla f(W^i_k)\cdot\varepsilon^i_k\big] &= \sqrt{N}\,\mathbf{E}\Big[(\langle f, \nu^N_k\rangle - \langle f, \bar{\mu}^N_{\frac{k+1}{N}}\rangle)\sum_{i=1}^N\nabla f(W^i_k)\cdot\varepsilon^i_k\Big] \\ &= \sqrt{N}\sum_{i=1}^N\mathbf{E}\left[(\langle f, \nu^N_k\rangle - \langle f, \bar{\mu}^N_{\frac{k+1}{N}}\rangle)\nabla f(W^i_k)\right]\cdot\mathbf{E}[\varepsilon^i_k] = 0.\end{aligned}$$

On the other hand, using (2.2.29), $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$ (see (2.1.6)), and the same arguments as those used in (2.2.30), it holds:

$$\mathbf{E}\big[\big|\sum_{i=1}^N\nabla f(W^i_k)\cdot\varepsilon^i_k\big|^2\big] = \sum_{i=1}^N\mathbf{E}\big[\big|\nabla f(W^i_k)\cdot\varepsilon^i_k\big|^2\big] \leq CN\|f\|^2_{\mathcal{C}^{2,\gamma_*}} \leq CN\|f\|^2_{\mathcal{H}^{L,\gamma}}.$$

This ends the proof of item *(ii)*. Item *(iii)* is a direct consequence of (2.2.34) and $\mathcal{H}^{L,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$.

Let us now prove item *(iv)*. We have that

$$\sum_{k=0}^{\lfloor Nt\rfloor-1} \langle f, \Upsilon_{\frac{k+1}{N}-}\rangle\sqrt{N}\langle f, R_k^N\rangle \le \sum_{k=0}^{\lfloor Nt\rfloor-1} \frac{1}{N}\langle f, \Upsilon_{\frac{k+1}{N}-}\rangle^2 + \sum_{k=0}^{\lfloor Nt\rfloor-1} N^2|\langle f, R_k^N\rangle|^2. \tag{2.5.2}$$

On the one hand, by item *(iii)*,

$$\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1} N^2|\langle f, R_k^N\rangle|^2\Big] \le C\|f\|^2_{\mathcal{H}^{L,\gamma}}\Big[\frac{1}{N} + \frac{N^3}{N^{4\beta}}\Big]. \tag{2.5.3}$$

On the other hand, we have

$$\begin{aligned}\Big|\int_0^{\frac{\lfloor Nt\rfloor}{N}} \langle f, \Upsilon_s^N\rangle^2 \mathrm{d}s - \sum_{k=0}^{\lfloor Nt\rfloor-1} \frac{1}{N}\langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle^2\Big| &= \Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}} \Big(\langle f, \Upsilon_s^N\rangle^2 - \langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle^2\Big)\mathrm{d}s\Big| \\ &\le \sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}} \Big|\langle f, \Upsilon_s^N\rangle^2 - \langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle^2\Big|\mathrm{d}s.\end{aligned}$$

Let $0 \le k < \lfloor Nt\rfloor$ and $s \in \big(\frac{k}{N}, \frac{k+1}{N}\big)$. We have

$$\begin{aligned}\langle f, \Upsilon_s^N\rangle^2 - \langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle^2 &= N\Big[\big(\langle f, \nu_k^N\rangle - \langle f, \bar\mu_s^N\rangle\big)^2 - \big(\langle f, \nu_k^N\rangle - \langle f, \bar\mu^N_{\frac{k+1}{N}}\rangle\big)^2\Big] \\ &= N\Big[2\langle f, \nu_k^N\rangle(\langle f, \bar\mu^N_{\frac{k+1}{N}}\rangle - \langle f, \bar\mu_s^N\rangle) + \langle f, \bar\mu_s^N\rangle^2 - \langle f, \bar\mu^N_{\frac{k+1}{N}}\rangle^2\Big] \\ &= N\Big[2\langle f, \nu_k^N\rangle\Big(\langle f, \bar\mu^N_{\frac{k+1}{N}}\rangle - \langle f, \bar\mu_s^N\rangle\Big) \\ &\quad + \big(\langle f, \bar\mu_s^N\rangle - \langle f, \bar\mu^N_{\frac{k+1}{N}}\rangle\big)\big(\langle f, \bar\mu_s^N\rangle + \langle f, \bar\mu^N_{\frac{k+1}{N}}\rangle\big)\Big].\end{aligned} \tag{2.5.4}$$

It holds using (2.3.5) and that $|\bar X_{\frac{k+1}{N}} - \bar X_s| \le C/N$ (by (2.3.4)):

$$\begin{aligned}&|\langle f, \bar\mu^N_{\frac{k+1}{N}}\rangle - \langle f, \bar\mu_s^N\rangle| = \Big|\frac{1}{N}\sum_{i=1}^N f(\bar X^i_{\frac{k+1}{N}}) - f(\bar X_s^i)\Big| \\ &\le \frac{1}{N}\Big[\sum_{i=1}^N |\bar X^i_{\frac{k+1}{N}} - \bar X_s^i||\nabla f(\bar X_s^i)| + C|\bar X^i_{\frac{k+1}{N}} - \bar X_s^i|^2 \sup_{t\in(0,1)}|\nabla^2 f|(t\bar X^i_{\frac{k+1}{N}} + (1-t)\bar X_s^i)\Big] \\ &\le \frac{C}{N}\|f\|_{\mathcal{C}^{2,\gamma_*}}\sum_{i=1}^N |\bar X^i_{\frac{k+1}{N}} - \bar X_s^i| + |\bar X^i_{\frac{k+1}{N}} - \bar X_s^i|^2 \le \frac{C}{N}\|f\|_{\mathcal{H}^{L,\gamma}}.\end{aligned} \tag{2.5.5}$$

Going back to (2.5.4), and using also (2.3.5), we have:

$$\begin{aligned}\Big|\langle f, \Upsilon_s^N\rangle^2 - \langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle^2\Big| &\le C\|f\|_{\mathcal{H}^{L,\gamma}}\Big(|\langle f, \nu_k^N\rangle| + \Big|\langle f, \bar\mu_s^N\rangle + \langle f, \bar\mu^N_{\frac{k+1}{N}}\rangle\Big|\Big) \\ &\le C\|f\|_{\mathcal{H}^{L,\gamma}}\Big(\frac{\|f\|_{\mathcal{C}^{2,\gamma_*}}}{N}\sum_{i=1}^N(1 + |W_k^i|^{\gamma_*}) + C\|f\|_{\mathcal{C}^{2,\gamma_*}}\Big).\end{aligned}$$

Therefore, using Lemma 2.11, we have shown that

$$\begin{aligned}\mathbf{E}\Big[\Big|\int_0^{\frac{\lfloor Nt\rfloor}{N}} \langle f, \Upsilon_s^N\rangle^2 \mathrm{d}s - \sum_{k=0}^{\lfloor Nt\rfloor-1} \frac{1}{N}\langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle^2\Big|\Big] &\le \sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}} \mathbf{E}\Big[\Big|\langle f, \Upsilon_s^N\rangle^2 - \langle f, \Upsilon^N_{\frac{k+1}{N}-}\rangle^2\Big|\Big]\mathrm{d}s \\ &\le C\|f\|^2_{\mathcal{H}^{L,\gamma}},\end{aligned}$$

so that

$$\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\frac{1}{N}\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle^2\Big]\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}+\mathbf{E}\Big[\int_0^{\frac{\lfloor Nt\rfloor}{N}}\langle f,\Upsilon^N_s\rangle^2\mathrm{d}s\Big].$$

On the other hand, using (2.2.31), (2.3.5), and $\mathcal{H}^{L,\gamma}(\mathbf{R}^d)\hookrightarrow\mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$,

$$\begin{aligned}\mathbf{E}[\langle f,\Upsilon^N_s\rangle^2]\le CN\Big[\mathbf{E}\big[\langle f,\mu^N_s\rangle^2\big]+\mathbf{E}\big[\langle f,\bar\mu^N_s\rangle^2\big]\Big]&\le CN\Big[\|f\|^2_{\mathcal{C}^{2,\gamma_*}}+\frac{1}{N^2}\mathbf{E}\big[\big|\sum_{i=1}^N f(\bar X^i_t)\big|^2\big]\Big]\\&\le CN\Big[\|f\|^2_{\mathcal{C}^{2,\gamma_*}}+\frac{1}{N}\mathbf{E}\big[\sum_{i=1}^N|f(\bar X^i_t)|^2\big]\Big]\\&\le CN\|f\|^2_{\mathcal{H}^{L,\gamma}}.\end{aligned}$$

Therefore, it holds: $\mathbf{E}[\int_{\frac{\lfloor Nt\rfloor}{N}}^t\langle f,\Upsilon^N_s\rangle^2\mathrm{d}s]\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}$. Hence,

$$\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\frac{1}{N}\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle^2\Big]\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}+\mathbf{E}\Big[\int_0^t\langle f,\Upsilon^N_s\rangle^2\mathrm{d}s\Big]\tag{2.5.6}$$

Item *(iv)* is then a consequence of (2.5.2), (2.5.3), and (2.5.6).

Let us now prove item *(v)*. We have (see (2.3.16))

$$\begin{aligned}&\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle\mathbf{a}^N_k[f]-\sqrt{N}\int_0^{\frac{\lfloor Nt\rfloor}{N}}\langle f,\Upsilon^N_s\rangle\mathbf{L}^N_s[f]\mathrm{d}s\Big]\\&\qquad=\sqrt{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}\mathbf{E}\Big[\Big(\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle-\langle f,\Upsilon^N_s\rangle\Big)\mathbf{L}^N_s[f]\Big]\mathrm{d}s.\end{aligned}$$

Using (2.5.5), for $s\in\big(\frac{k}{N},\frac{k+1}{N}\big)$, it holds:

$$\big|\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle-\langle f,\Upsilon^N_s\rangle\big|=\sqrt{N}\,\big|\langle f,\bar\mu^N_s\rangle-\langle f,\bar\mu^N_{\frac{k+1}{N}}\rangle\big|\le C\frac{\|f\|_{\mathcal{H}^{L,\gamma}}}{\sqrt{N}},$$

and using (2.2.17), Lemma 2.11, and $\mathcal{H}^{L,\gamma}(\mathbf{R}^d)\hookrightarrow\mathcal{C}^{2,\gamma_*}(\mathbf{R}^d)$, one deduces that:

$$\mathbf{E}\big[\,\big|\mathbf{L}^N_s[f]\big|^2\,\big]\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}.$$

Thus,

$$\mathbf{E}\Big[\Big|\sum_{k=0}^{\lfloor Nt\rfloor-1}\langle f,\Upsilon^N_{\frac{k+1}{N}-}\rangle\mathbf{a}^N_k[f]-\sqrt{N}\int_0^{\frac{\lfloor Nt\rfloor}{N}}\langle f,\Upsilon^N_s\rangle\mathbf{L}^N_s[f]\mathrm{d}s\Big|\Big]\le\sqrt{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}C\frac{\|f\|^2_{\mathcal{H}^{L,\gamma}}}{\sqrt{N}}\mathrm{d}s\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}.$$

In addition we have:

$$\begin{aligned}\mathbf{E}\Big[\sqrt{N}\Big|\int_{\frac{\lfloor Nt\rfloor}{N}}^t\langle f,\Upsilon^N_s\rangle\mathbf{L}^N_s[f]\mathrm{d}s\Big|\Big]&\le\sqrt{N}\int_{\frac{\lfloor Nt\rfloor}{N}}^t\sqrt{\mathbf{E}[|\langle f,\Upsilon^N_s\rangle|^2]}\sqrt{\mathbf{E}[|\mathbf{L}^N_s[f]|^2]}\mathrm{d}s\\&\le C\|f\|^2_{\mathcal{H}^{L,\gamma}}.\end{aligned}$$

We have thus proved item *(v)*.

Finally,

$$\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}|\mathbf{a}^N_k[f]|^2\Big]=N\mathbf{E}\Big[\sum_{k=0}^{\lfloor Nt\rfloor-1}\Big|\int_{\frac{k}{N}}^{\frac{k+1}{N}}\mathbf{L}^N_s[f]\mathrm{d}s\Big|^2\Big]\le\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}\mathbf{E}[|\mathbf{L}^N_s[f]|^2]\mathrm{d}s\le C\|f\|^2_{\mathcal{H}^{L,\gamma}},$$

which proves item *(vi)*. This ends the proof of the lemma. □

Lemma 2.43. *Let $J \geq 1$ and $\gamma \geq 0$. For $x \in \mathcal{X}$, recall the definition of T_x (see (2.3.21)), $\mathsf{T}_x : f \in \mathcal{H}^{J,\gamma}(\mathbf{R}^d) \mapsto \nabla f \cdot \nabla \sigma_*(\cdot, x) \in \mathcal{H}^{J-1,\gamma}(\mathbf{R}^d)$. Then, there exists $C > 0$ such that for any $\Upsilon \in \mathcal{H}^{-J+1,\gamma}(\mathbf{R}^d)$ and $x \in \mathcal{X}$,*

$$|\langle \Upsilon, \mathsf{T}_x^* \Upsilon \rangle_{\mathcal{H}^{-J,\gamma}}| \leq C \|\Upsilon\|^2_{\mathcal{H}^{-J,\gamma}}. \tag{2.5.7}$$

This result is stronger than what one obtains with the Cauchy-Schwarz inequality. Indeed, the Cauchy-Schwarz inequality only implies

$$|\langle \Upsilon, \mathsf{T}_x^* \Upsilon \rangle_{\mathcal{H}^{-J,\gamma}}| \leq \|\Upsilon\|_{\mathcal{H}^{-J,\gamma}} \|\mathsf{T}_x^* \Upsilon\|_{\mathcal{H}^{-J,\gamma}} \leq C \|\Upsilon\|_{\mathcal{H}^{-J,\gamma}} \|\Upsilon\|_{\mathcal{H}^{-J+1,\gamma}}.$$

Let us mention that Lemma 2.43 extends [SS20a, Lemma B.1] to the non compact and weighted case.

Proof. Let $x \in \mathcal{X}$ and $\Upsilon \in \mathcal{H}^{-J+1,\gamma}(\mathbf{R}^d) \hookrightarrow \mathcal{H}^{-J,\gamma}(\mathbf{R}^d)$. By the Riesz representation theorem, there exists a unique $\Psi \in \mathcal{H}^{J,\gamma}(\mathbf{R}^d)$ such that,

$$\langle f, \Upsilon \rangle = \langle f, \Psi \rangle_{\mathcal{H}^{J,\gamma}}, \text{ for } f \in \mathcal{H}^{J,\gamma}(\mathbf{R}^d).$$

We set $F(\Upsilon) = \Psi$. The density of $\mathcal{C}_c^\infty(\mathbf{R}^d)$ in $\mathcal{H}^{J,\gamma}(\mathbf{R}^d)$ implies that $\{\Upsilon \in \mathcal{H}^{-J,\gamma}(\mathbf{R}^d) : F(\Upsilon) \in \mathcal{C}_c^\infty(\mathbf{R}^d)\}$ is dense in $\mathcal{H}^{-J,\gamma}(\mathbf{R}^d)$. It is thus sufficient to show (2.5.7) for Υ such that $\Psi = F(\Upsilon) \in \mathcal{C}_c^\infty(\mathbf{R}^d)$. We have

$$\langle \Upsilon, \mathsf{T}_x^* \Upsilon \rangle_{\mathcal{H}^{-J,\gamma}} = \langle \Psi, \mathsf{T}_x^* \Upsilon \rangle = \langle \mathsf{T}_x \Psi, \Upsilon \rangle = \langle \mathsf{T}_x \Psi, \Psi \rangle_{\mathcal{H}^{J,\gamma}}. \tag{2.5.8}$$

Let us prove that $|\langle \mathsf{T}_x \Psi, \Psi \rangle_{J,\gamma}| \leq C \|\Psi\|^2_{\mathcal{H}^{J,\gamma}}$ for $\Psi \in \mathcal{C}_c^\infty(\mathbf{R}^d)$. By definition, we have

$$\langle \mathsf{T}_x \Psi, \Psi \rangle_{J,\gamma} = \sum_{|k| \leq J} \int_{\mathbf{R}^d} \Big[D^k (\nabla \Psi(w) \cdot \nabla \sigma_*(w, x)) D^k \Psi(w) \Big] \times \frac{1}{1 + |w|^{2\gamma}} \mathrm{d}w.$$

In the previous sum, the only terms involving derivatives of Ψ of order greater than J are the terms for which $|k| = J$. Therefore, it is sufficient to only deal with such k. Pick a multi-index k such that $|k| = J$. For all $x \in \mathcal{X}$, we have

$$\begin{aligned}
\int_{\mathbf{R}^d} \big[D^k (\nabla \Psi(w) \cdot \nabla \sigma_*(w, x)) D^k \Psi(w) \big] \times \frac{\mathrm{d}w}{1 + |w|^{2\gamma}} &= \int_{\mathbf{R}^d} D^k \Big(\sum_{i=1}^d \partial_i \Psi(w) \partial_i \sigma_*(w, x) \Big) \times \frac{D^k \Psi(w)}{1 + |w|^{2\gamma}} \mathrm{d}w \\
&= \sum_{i=1}^d \int_{\mathbf{R}^d} D^k \left(\partial_i \Psi(w) \partial_i \sigma_*(w, x) \right) \times \frac{D^k \Psi(w)}{1 + |w|^{2\gamma}} \mathrm{d}w.
\end{aligned}$$

Let us consider the case when $i = 1$ and $k = (J, 0 \ldots, 0)$. The other cases can be treated similarly. For all $x \in \mathcal{X}$,

$$\begin{aligned}
\int_{\mathbf{R}^d} D^k \left(\partial_1 \Psi(w) \partial_1 \sigma_*(w, x) \right) \times \frac{D^k \Psi(w)}{1 + |w|^{2\gamma}} \mathrm{d}w &= \int_{\mathbf{R}^d} \partial_1^J \left(\partial_1 \Psi(w) \partial_1 \sigma_*(w, x) \right) \times \frac{\partial_1^J \Psi(w)}{1 + |w|^{2\gamma}} \mathrm{d}w \\
&= \int_{\mathbf{R}^d} \partial_1^{J+1} \Psi(w) \partial_1 \sigma_*(w, x) \times \frac{\partial_1^J \Psi(w)}{1 + |w|^{2\gamma}} \mathrm{d}w \\
&\quad + \sum_{j=0}^{J-1} \binom{J}{j} \int_{\mathbf{R}^d} \partial_1^{j+1} \Psi(w) \partial_1^{J-j+1} \sigma_*(w, x) \times \frac{\partial_1^J \Psi(w)}{1 + |w|^{2\gamma}} \mathrm{d}w.
\end{aligned} \tag{2.5.9}$$

Since σ_* and all its derivatives are bounded, one has:

$$\sum_{j=0}^{J-1} \binom{J}{j} \int_{\mathbf{R}^d} \left| \partial_1^{j+1} \Psi(w) \partial_1^{J-j+1} \sigma_*(w, x) \times \frac{\partial_1^J \Psi(w)}{1 + |w|^{2\gamma}} \right| \mathrm{d}w \leq C \|\Psi\|_{\mathcal{H}^{J,\gamma}}.$$

Let us now deal with the first term in the right-hand side of (2.5.9). By Fubini's theorem :

$$\int_{\mathbf{R}^d} \partial_1^{J+1}\Psi(w)\partial_1\sigma_*(w,x)\times\frac{\partial_1^J\Psi(w)}{1+|w|^{2\gamma}}\mathrm{d}w = \int_{\mathbf{R}^{d-1}}\int_{\mathbf{R}} \partial_1^{J+1}\Psi(z,w')\partial_1\sigma_*(z,w',x)\times\frac{\partial_1^J\Psi(z,w')}{1+|(z,w')|^{2\gamma}}\mathrm{d}z\mathrm{d}w'.$$

An integration by parts yields, for all $w'\in\mathbf{R}^{d-1}$, using that Ψ is compactly supported,

$$\begin{aligned}
&2\int_{\mathbf{R}} \partial_1^{J+1}\Psi(z,w')\partial_1\sigma_*(z,w',x)\times\frac{\partial_1^J\Psi(z,w')}{1+|(z,w')|^{2\gamma}}\mathrm{d}z\\
&=\left[|\partial_1^J\Psi(z,w')|^2\frac{\partial_1\sigma_*(z,w',x)}{1+|(z,w')|^{2\gamma}}\right]_{-\infty}^{+\infty}-\int_{\mathbf{R}}|\partial_1^J\Psi(z,w')|^2\partial_1\left(\frac{\partial_1\sigma_*(z,w',x)}{1+|(z,w')|^{2\gamma}}\right)\mathrm{d}z\\
&=\int_{\mathbf{R}}|\partial_1^J\Psi(z,w')|^2\times\frac{2\gamma z|(z,w')|^{2\gamma-2}\partial_1\sigma_*(z,w',x)-\partial_1^2\sigma_*(z,w',x)(1+|(z,w')|^{2\gamma})}{(1+|(z,w')|^{2\gamma})^2}\mathrm{d}z.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\int_{\mathbf{R}^d} \partial_1^{J+1}\Psi(w)\partial_1\sigma_*(w,x)\times\frac{\partial_1^J\Psi(w)}{1+|w|^{2\gamma}}\mathrm{d}w\right|\\
&=\frac{1}{2}\left|\int_{\mathbf{R}^{d-1}}\int_{\mathbf{R}}|\partial_1^J\Psi(z,w')|^2\times\frac{2\gamma w|(z,w')|^{2\gamma-2}\partial_1\sigma_*(z,w',x)-\partial_1^2\sigma_*(z,w',x)(1+|(z,w')|^{2\gamma})}{(1+|(z,w')|^{2\gamma})^2}\mathrm{d}z\mathrm{d}w'\right|\\
&\le C\int_{\mathbf{R}^{d-1}}\int_{\mathbf{R}}\frac{|\partial_1^J\Psi(z,w')|^2}{1+|(z,w')|^{2\gamma}}\times\mathrm{d}z\mathrm{d}w'\le C\|\Psi\|_{\mathcal{H}^{J,\gamma}}^2.
\end{aligned}$$

To summarize, we have shown the existence of $C<\infty$ (independent of x) such that for any $\Psi\in\mathcal{C}_c^\infty(\mathbf{R}^d)$, $|\langle\mathsf{T}_x\Psi,\Psi\rangle_{J,\gamma}|\le C\|\Psi\|_{\mathcal{H}^{J,\gamma}}^2$. Consequently, by (2.5.8), $|\langle\Upsilon,\mathsf{T}_x^*\Upsilon\rangle_{-J,\gamma}|\le C\|\Upsilon\|_{\mathcal{H}^{J,\gamma}}^2$. This completes the proof of the lemma. □

Lemma 2.44. *Let $N\ge1$ and $f:\mathbf{R}_+\to\mathbf{R}$ be a piecewise continuous function whose jumps occur only at the times k/N, $k\ge1$. Introduce the function $g:\mathbf{R}_+\to\mathbf{R}$ defined by, for all $t\ge0$, $g(t)=\sum_{k=0}^{\lfloor Nt\rfloor-1}\alpha_k$ where for all $k\ge0$, $\alpha_k\in\mathbf{R}$ (recall the convention $\sum_{k=0}^{-1}=0$). Set for $t\ge0$,*

$$F(t)=\int_0^t f(s)\mathrm{d}s \text{ and } \psi(t)=F(t)+g(t).$$

Then, for all $t\ge0$,

$$\psi(t)^2=2\int_0^t f(s)\psi(s)\mathrm{d}s+\sum_{k=0}^{\lfloor Nt\rfloor-1}|\alpha_k|^2+2\sum_{k=0}^{\lfloor Nt\rfloor-1}\psi\Big(\frac{k+1}{N}^-\Big)\Big(\underbrace{g\Big(\frac{k+1}{N}\Big)-g\Big(\frac{k}{N}\Big)}_{=\alpha_k}\Big).$$

Proof. For all $k\ge0$ and $t\in[\frac{k}{N},\frac{k+1}{N})$, it holds $\psi(t)^2-\psi(\frac{k}{N})^2=2\int_{\frac{k}{N}}^t\psi'(s)\psi(s)\mathrm{d}s$. Letting $t\to\frac{k+1}{N}^-$, we obtain

$$\psi\Big(\frac{k+1}{N}^-\Big)^2-\psi\Big(\frac{k}{N}\Big)^2=2\int_{\frac{k}{N}}^{\frac{k+1}{N}}\psi'(s)\psi(s)\mathrm{d}s. \tag{2.5.10}$$

Since F is continuous and by definition of g, it holds $\psi(\frac{k+1}{N}^-)^2=(F(\frac{k+1}{N})+g(\frac{k}{N}))^2$. Hence,

$$\psi\Big(\frac{k+1}{N}^-\Big)^2-\psi\Big(\frac{k}{N}\Big)^2=F\Big(\frac{k+1}{N}\Big)^2-F\Big(\frac{k}{N}\Big)^2+2g\Big(\frac{k}{N}\Big)\Big(F\Big(\frac{k+1}{N}\Big)-F\Big(\frac{k}{N}\Big)\Big).$$

Therefore, (2.5.10) reads (using also that $g'(s) = 0$ for all $s \in (\frac{k}{N}, \frac{k+1}{N})$)

$$F\Big(\frac{k+1}{N}\Big)^2 - F\Big(\frac{k}{N}\Big)^2 + 2g\Big(\frac{k}{N}\Big)\Big(F\Big(\frac{k+1}{N}\Big) - F\Big(\frac{k}{N}\Big)\Big) = 2\int_{\frac{k}{N}}^{\frac{k+1}{N}} f(s)\psi(s)\mathrm{d}s.$$

Now, for all $t \geq 0$, denoting $k = \lfloor Nt \rfloor$,

$$\begin{aligned}
2\int_0^t f(s)\psi(s)\mathrm{d}s &= \sum_{j=0}^{k-1} 2\int_{\frac{j}{N}}^{\frac{j+1}{N}} f(s)\psi(s)\mathrm{d}s + 2\int_{\frac{k}{N}}^{t} f(s)\psi(s)\mathrm{d}s \\
&= \sum_{j=0}^{k-1} F\Big(\frac{j+1}{N}\Big)^2 - F\Big(\frac{j}{N}\Big)^2 + 2g\Big(\frac{j}{N}\Big)\Big(F\Big(\frac{j+1}{N}\Big) - F\Big(\frac{j}{N}\Big)\Big) + \psi(t)^2 - \psi\Big(\frac{k}{N}\Big)^2 \\
&= F\Big(\frac{k}{N}\Big)^2 + \sum_{j=0}^{k-1} 2g\Big(\frac{j}{N}\Big)\Big(F\Big(\frac{j+1}{N}\Big) - F\Big(\frac{j}{N}\Big)\Big) + \psi(t)^2 - \psi\Big(\frac{k}{N}\Big)^2 \\
&= F\Big(\frac{k}{N}\Big)^2 + \sum_{j=0}^{k-2} 2F\Big(\frac{j+1}{N}\Big)\Big(g\Big(\frac{j}{N}\Big) - g\Big(\frac{j+1}{N}\Big)\Big) + 2g\Big(\frac{k-1}{N}\Big)F\Big(\frac{k}{N}\Big) + \psi(t)^2 - \psi\Big(\frac{k}{N}\Big)^2 \\
&= -g\Big(\frac{k}{N}\Big)^2 + \sum_{j=0}^{k-1} 2F\Big(\frac{j+1}{N}\Big)\Big(g\Big(\frac{j}{N}\Big) - g\Big(\frac{j+1}{N}\Big)\Big) + \psi(t)^2.
\end{aligned}$$

Hence,

$$\psi(t)^2 = 2\int_0^t f(s)\psi(s)\mathrm{d}s + g\Big(\frac{k}{N}\Big)^2 + 2\sum_{j=0}^{k-1} F\Big(\frac{j+1}{N}\Big)\Big(g\Big(\frac{j+1}{N}\Big) - g\Big(\frac{j}{N}\Big)\Big).$$

Using that $g(0) = 0$, one can write $g(\frac{k}{N})^2 = \sum_{j=0}^{k-1} |\alpha_k|^2 + 2\sum_{j=0}^{k-1}(g(\frac{j+1}{N}) - g(\frac{j}{N}))g(\frac{j}{N})$. This yields,

$$\begin{aligned}
\psi(t)^2 &= 2\int_0^t f(s)\psi(s)\mathrm{d}s + \sum_{j=0}^{k-1} |\alpha_k|^2 + 2\sum_{j=0}^{k-1} \Big(F\Big(\frac{j+1}{N}\Big) + g\Big(\frac{j}{N}\Big)\Big)\Big(g\Big(\frac{j+1}{N}\Big) - g\Big(\frac{j}{N}\Big)\Big) \\
&= 2\int_0^t f(s)\psi(s)\mathrm{d}s + \sum_{j=0}^{k-1} |\alpha_k|^2 + 2\sum_{j=0}^{k-1} \psi\Big(\frac{j+1^-}{N}\Big)\Big(g\Big(\frac{j+1}{N}\Big) - g\Big(\frac{j}{N}\Big)\Big),
\end{aligned}$$

which is the desired formula. □

Remark 2.45. *Notice by Lemma 2.44 and* (3.9.3) *(with $m = 4$ there), if $\alpha_k = \alpha_k^1 + \alpha_k^2 + \alpha_k^3 + \alpha_k^4$, it holds*

$$\psi(t)^2 \leq 2\int_0^t f(s)\psi(s)\mathrm{d}s + 4\sum_{\ell=1}^{4}\sum_{k=0}^{\lfloor Nt \rfloor - 1} |\alpha_k^\ell|^2 + 2\sum_{\ell=1}^{4}\sum_{k=0}^{\lfloor Nt \rfloor - 1} \alpha_k^\ell\, \psi\Big(\frac{k+1^-}{N}\Big).$$

Chapter 3

Law of Large Numbers for Bayesian two-layer Neural Network trained with Variational Inference

This chapter corresponds to the article [DHG+23], published in the *36th Conference on Learning Theory* (COLT 2023). It is a joint work with Tom Huix, Arnaud Guillin, Manon Michel, Éric Moulines and Boris Nectoux.

Contents

Abstract

We provide a rigorous analysis of training by variational inference (VI) of Bayesian neural networks in the two-layer and infinite-width case. We consider a regression problem with a regularized evidence lower bound (ELBO) which is decomposed into the expected log-likelihood of the data and the Kullback-Leibler (KL) divergence between the a priori distribution and the variational posterior. With an appropriate weighting of the KL, we prove a law of large numbers for three different training schemes: (i) the idealized case with exact estimation of a multiple

Gaussian integral from the reparametrization trick, (ii) a minibatch scheme using Monte Carlo sampling, commonly known as *Bayes by Backprop*, and (iii) a new and computationally cheaper algorithm which we introduce as *Minimal VI*. An important result is that all methods converge to the same mean-field limit. Finally, we illustrate our results numerically and discuss the need for the derivation of a central limit theorem.

3.1 Introduction

Deep Learning has led to a revolution in machine learning with impressive successes. However, some limitations of DL have been identified and, despite, many attempts, our understanding of DL is still limited. A long-standing problem is the assessment of predictive uncertainty: DL tends to be overconfident in its predictions [APH$^+$21], which is a problem in applications such as autonomous driving [MGK$^+$17, MWL$^+$20], medical diagnosis [KG17, FFG$^+$19], or finance; cf [KA13, Gha15]. Therefore, on the one hand, analytical efforts are being made to thoroughly investigate the performance of DL; and on the other hand, many approaches have been proposed to alleviate its shortcomings. The Bayesian paradigm is an attractive way to tackle predictive uncertainty, as it provides a framework for training uncertainty-aware neural networks (NNs) (e.g. [Gha15, BCKW15, GG16]).

Thanks to a fully probabilistic approach, Bayesian Neural Networks (BNN) combine the impressive neural-network expressivity with the decision-theoretic approach of Bayesian inference, making them capable of providing predictive uncertainty; see [BCKW15, MWL$^+$20, MGK$^+$17, FFG$^+$19]. However, Bayesian inference requires deriving the posterior distribution of the NN weights. This posterior distribution is typically not tractable. A classical approach is to sample the posterior distribution using Markov chain Monte Carlo methods (such as Hamilton-Monte-Carlo methods). There are however long-standing difficulties, such as the proper choice of the prior and fine-tuning of the sampler. Such difficulties often become prohibitive in large-dimensional cases,[CJ21]. An alternative is to use variational inference, which has a long history [HC93, Mac95, M$^+$95]. Simpler methods that do not require exact computation of integrals over the variational posterior were then developed, e.g. first by [Gra11] thanks to some approximation and then by [BCKW15] with the *Bayes by Backprop* approach. In the latter, the posterior distribution is approximated by a parametric distribution and a generalisation of the reparametrization trick used by [KW14] leads to an unbiased estimator of the gradient of the ELBO; see also [GG16, LW17, KNT$^+$18]. Despite the successful application of this approach, little is known about the overparameterized limit and appropriate weighting that must be assumed to obtain a nontrivial Bayesian posterior, see [IVHW21]. Recently, [HMD$^+$22] outlined the importance of balancing in ELBO the integrated log-likelihood term and the KL regularizer, to avoid both overfitting and dominance of the prior. However, a suitable limiting theory has yet to be established, as well as guarantees for the practical implementation of the stochastic gradient descent (SGD) used to estimate the parameters of the variational distribution.

Motivated by the need to provide a solid theoretical framework, asymptotic analysis of NN has gained much interest recently. The main focus has been on the gradient descent algorithm and its variants [RVE18a, CB18b, MMN18, SS20b, DGMN22]. In much of these works, a mean-field analysis is performed to characterize the limiting nonlinear evolution of the weights of a two-layer NN, allowing the derivation of a law of large numbers and a central limit theorem for the empirical distribution of neuron weights. A long-term goal of these works is to demonstrate convergence toward a global minimum of these limits for the mean field. Despite some progress in this direction, this is still an open and highly challenging problem; cf [CB18b, Chi22, CCFRF22]. Nevertheless, this asymptotic analysis is also of interest in its own right, as we show here in the case of variational inference for Bayesian neural networks. Indeed, based on this asymptotic analysis, we develop an efficient and new variant of the stochastic gradient descent (SGD) algorithm for variational inference in BNN that computes only the information necessary to recover the

limit behavior.

Our goal, then, is to work at the intersection of analytical efforts to gain theoretical guarantees and insights and of practical methods for a workable variational inference procedure. By adapting the framework developed by [DGMN22], we produce a rigorous asymptotic analysis of BNN trained in a variational setting for a regression task. From the limit equation analysis, we first find that a proper regularisation of the Kullback-Leibler divergence term in relation with the integrated loss leads to their right asymptotic balance. Second, we prove the asymptotic equivalence of the idealized and Bayes-by-Backprop SGD schemes, as both preserve the same core contributions to the limit. Finally, we introduce a computationally more favourable scheme, directly stemming from the effective asymptotic contributions. This scheme is the true mean-field algorithmic approach, as only deriving from non-interacting terms.

More specifically, our contributions are the following:

- We first focus on the idealized SGD algorithm, where the variational expectations of the derivative of the loss from the reparametrization trick of [BCKW15] are computed exactly. More precisely, we prove that with the number of neurons $N \to +\infty$, the sequence of trajectories of the scaled empirical distributions of the parameters satisfies a law of large numbers. This is the purpose of Theorem 3.1. The proof is completely new: it establishes directly the limit in the topology inherited by the Wasserstein distance bypassing the highly technical Sobolev space arguments used in [DGMN22].

The idealized SGD requires the computation of some integrals, which in practice prevents a direct application of this algorithm. However, we can prove its convergence to an explicit nonlinear process. These integrals are usually obtained by a Monte Carlo approximation, leading to the *Bayes-by-Backprop* SGD, see [BCKW15].

- We show for the *Bayes-by-Backprop* SGD (see Theorem 3.2) that the sequence of trajectories of the scaled empirical distributions of the parameters satisfies the same law of large numbers as that in Theorem 3.1, which justifies such an approximation procedure. Note that each step of the algorithm involves the simulation of $O(N)$ Gaussian random variables, which can make the associated gradient evaluation prohibitively expensive.

- A careful analysis of the structure of the limit equation (3.3.4) allows us to develop a new algorithm, called *Minimal-VI* SGD, which at each step generates only two Gaussian random variables and for which we prove the same limiting behavior. The key idea here is to keep only those contributions which affect the asymptotic behavior and which can be understood as the mean-field approximation from the uncorrelated degrees of freedom. This is all the more interesting since we observe numerically that the number weights N required to reach this asymptotic limit is quite small which makes this variant of immediate practical interest.

- We numerically investigate the convergence of the three methods to the common limit behavior on a toy example. We observe that the mean-field method is effective for a small number of neurons ($N = 300$). The differences between the methods are reflected in the variances.

The paper is organized as follows: Section 3.2 introduces the variational inference in BNN, as well as the SGD schemes commonly considered, namely the idealized and *Bayes-by-backprop* variants. Then, in Section 3.3 we establish our initial result, the LLN for the idealized SGD. In Section 3.4 we prove the LLN for the *Bayes-by-backprop* SGD and its variants. We show that both SGD schemes have the same limit behavior. Based on an analysis of the obtained limit equation, we present in Section 3.5 the new *minimal- VI*. Finally, in Section 3.6 we illustrate our findings using numerical experiments. The proofs of the mean-field limits, which are original and quite technically demanding, are gathered in the supplementary paper.

Related works. Law of Large Numbers (LLN) for mean-field interacting particle systems, have attracted a lot of attentions; see for example [HM86, Szn91, FM97a, JM98a, DLR19a, DMG99, KX04] and references therein. The use of mean-field particle systems to analyse two-layer neural networks with random initialization have been considered in [MMN18, MMM19], which establish a LLN on the empirical measure of the weights at fixed times - we consider in this paper the trajectory convergence, i.e. the whole empirical measure process (time indexed) converges uniformly w.r.t. Skorohod topology. It enables not only to use the limiting PDE, for example to study the convergence of the weights towards the infimum of the loss function (see [CB18b] for preliminary results), but is is also crucial to establish the central limit theorem, see for example [DGMN22]. [RVE18a] give conditions for global convergence of GD for exact mean-square loss and online stochastic gradient descent (SGD) with mini-batches increasing in size with the number of weights N. A LLN for the entire trajectory of the empirical measure is also given in [SS20b] for a standard SGD. [DBDFS20] establish the propagation of chaos for SGD with different step size schemes. Compared to the existing literature dealing with the SGD empirical risk minimization in two-layer neural networks, [DGMN22] provide the first rigorous proof of the existence of the limit PDE, and in particular its uniqueness, in the LLN.

We are interested here in deriving a LLN but for Variational Inference (VI) of two-layer Bayesian Neural Networks (BNN), where we consider a regularized version of the Evidence Lower Bound (ELBO).

3.2 Variational inference in BNN: Notations and common SGD schemes

3.2.1 Variational inference and Evidence Lower Bound

Setting. Let X and Y be subsets of $\mathbf{R}^n$ ($n \geq 1$) and $\mathbf{R}$ respectively. For $N \geq 1$ and $\boldsymbol{w} = (w^1, \dots, w^N) \in (\mathbf{R}^d)^N$, let $f^N_{\boldsymbol{w}} : \mathsf{X} \to \mathbf{R}$ be the following two-layer neural network: for $x \in \mathsf{X}$,

$$f^N_{\boldsymbol{w}}(x) := \frac{1}{N} \sum_{i=1}^{N} s(w^i, x) \in \mathbf{R},$$

where $s : \mathbf{R}^d \times \mathsf{X} \to \mathbf{R}$ is the activation function. We work in a Bayesian setting, in which we seek a distribution of the latent variable $\boldsymbol{w}$ which represents the weights of the neural network. The standard problem in Bayesian inference over complex models is that the posterior distribution is hard to sample. To tackle this problem, we consider Variational Inference, in which we consider a family of distribution $\mathcal{Q}^N = \{q^N_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Xi^N\}$ (where Ξ is some parameter space) easy to sample. The objective is to find the best $q^N_{\boldsymbol{\theta}} \in \mathcal{Q}^N$, the one closest in KL divergence (denoted $\mathscr{D}_{\mathrm{KL}}$) to the exact posterior. Because we cannot compute the KL, we optimize the evidence lower bound (ELBO), which is equivalent to the KL up to an additive constant.

Denoting by $\mathfrak{L} : \mathbf{R} \times \mathbf{R} \to \mathbf{R}_+$ the negative log-likelihood (by an abuse of language, we call this quantity the *loss*), the ELBO (see [BKM17]) is defined, for $\boldsymbol{\theta} \in \Xi^N$, $(x, y) \in \mathsf{X} \times \mathsf{Y}$, by

$$\mathrm{E}_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y) := -\int_{(\mathbf{R}^d)^N} \mathfrak{L}(y, f^N_{\boldsymbol{w}}(x)) q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} - \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P^N_0),$$

where P^N_0 is some prior on the weights of the NN. The ELBO is decomposed into two terms: one corresponding to the Kullback-Leibler (KL) divergence between the variational density and the prior and the other to a marginal likelihood term. It was empirically found that the maximization of the ELBO function is prone to yield very poor inferences [CPDV21]. It is argued in [CPDV21] and [HMD$^+$22] that optimizing the ELBO leads as $N \to \infty$ to the collapse of the variational posterior to the prior. [HMD$^+$22] proposed to consider a regularized version of the ELBO, which consists in multiplying the KL term by a parameter which is scaled by the inverse of the number

of neurons:

$$\mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y) := -\int_{(\mathbf{R}^d)^N} \mathfrak{L}(y, f^N_{\boldsymbol{w}}(x)) q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} - \frac{1}{N} \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P^N_0), \tag{3.2.1}$$

A first objective of this paper is to show that the proposed regularization leads to a stable asymptotic behavior and the effect of both the integrated loss and Kullback-Leibler terms on the limiting behavior are balanced in the limit $N \to \infty$. The maximization of $\mathrm{E}^N_{\mathrm{lbo}}$ is carried out using SGD.

The variational family $\mathcal{Q}^N$ we consider is a Gaussian family of distributions. More precisely, we assume that for any $\boldsymbol{\theta} = (\theta^1, \dots, \theta^N) \in \Xi^N$, the variational distribution $q^N_{\boldsymbol{\theta}}$ factorizes over the neurons: for all $\boldsymbol{w} = (w^1, \dots, w^N) \in (\mathbf{R}^d)^N$, $q^N_{\boldsymbol{\theta}}(\boldsymbol{w}) = \prod_{i=1}^N q^1_{\theta^i}(w^i)$, where $\theta = (m, \rho) \in \Xi := \mathbf{R}^d \times \mathbf{R}$ and q^1_θ is the probability density function (pdf) of $\mathcal{N}(m, g(\rho)^2 I_d)$, with $g(\rho) = \log(1 + e^\rho)$, $\rho \in \mathbf{R}$.

In the following, we simply write $\mathbf{R}^{d+1}$ for $\mathbf{R}^d \times \mathbf{R}$. In addition, following the reparameterisation trick of [BCKW15], $q^1_\theta(w)\mathrm{d}w$ is the pushforward of a reference probability measure with density γ by Ψ_θ (see more precisely Assumption **A1**). In practice, γ is the pdf of $\mathcal{N}(0, I_d)$ and $\Psi_\theta(z) = m + g(\rho)z$. With these notations, (3.2.1) writes

$$\mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y) = -\int_{(\mathbf{R}^d)^N} \mathfrak{L}\Big(y, \frac{1}{N}\sum_{i=1}^N s(\Psi_{\theta^i}(z^i), x)\Big) \gamma(z^1) \dots \gamma(z^N) \mathrm{d}z_1 \dots \mathrm{d}z_N - \frac{1}{N} \mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P^N_0).$$

Loss function and prior distribution. In this work, we focus on the regression problem, i.e. $\mathfrak{L}$ is the Mean Square Loss: for $y_1, y_2 \in \mathbf{R}$, $\mathfrak{L}(y_1, y_2) = \frac{1}{2}|y_1 - y_2|^2$. We also introduce the function $\phi : (\theta, z, x) \in \mathbf{R}^{d+1} \times \mathbf{R}^d \times \mathsf{X} \mapsto s(\Psi_\theta(z), x)$. On the other hand, we assume that the prior distribution P^N_0 write, for all $\boldsymbol{w} \in (\mathbf{R}^d)^N$, $P^N_0(\boldsymbol{w}) = \prod_{i=1}^N P^1_0(w^i)$, where $P^1_0 : \mathbf{R}^d \to \mathbf{R}_+$ is the pdf of $\mathcal{N}(m_0, \sigma_0^2 I_d)$, and $\sigma_0 > 0$. Therefore $\mathscr{D}_{\mathrm{KL}}(q^N_{\boldsymbol{\theta}} | P^N_0) = \sum_{i=1}^N \mathscr{D}_{\mathrm{KL}}(q_{\theta^i} | P^1_0)$ and, for $\theta = (m, \rho) \in \mathbf{R}^{d+1}$,

$$\mathscr{D}_{\mathrm{KL}}(q^1_\theta | P^1_0) = \int_{\mathbf{R}^d} q^1_\theta(x) \log(q^1_\theta(x)/P^1_0(x)) \mathrm{d}x = \frac{\|m - m_0\|_2^2}{2\sigma_0^2} + \frac{d}{2}\Big(\frac{g(\rho)^2}{\sigma_0^2} - 1\Big) + \frac{d}{2}\log\Big(\frac{\sigma_0^2}{g(\rho)^2}\Big).$$

Note that $\mathscr{D}_{\mathrm{KL}}$ has at most a quadratic growth in m and ρ.

Note that we assume here a Gaussian prior to get an explicit expression of the Kullback-Leibler divergence. Most arguments extend to sufficiently regular densities and are essentially the same for exponential families, using conjugate families for the variational approximation.

3.2.2 Common SGD schemes in backpropagation in a variational setting

Idealized SGD. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Consider a data set $\{(x_k, y_k)\}_{k \geq 0}$ i.i.d. w.r.t. $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$, the space of probability measures over $\mathsf{X} \times \mathsf{Y}$. For $N \geq 1$ and given a learning rate $\eta > 0$, the maximization of $\theta \in \mathbf{R}^{d+1} \mapsto \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y)$ with a SGD algorithm writes as follows: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases} \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \nabla_{\boldsymbol{\theta}} \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}_k, x_k, y_k) \\ \boldsymbol{\theta}_0 \sim \mu_0^{\otimes N}, \end{cases} \tag{3.2.2}$$

where $\mu_0 \in \mathcal{P}(\mathbf{R}^{d+1})$ and $\boldsymbol{\theta}_k = (\theta^1_k, \dots, \theta^N_k)$. We now compute $\nabla_{\boldsymbol{\theta}} \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y)$.

First, under regularity assumptions on the function ϕ (which will be formulated later, see **A1**

and **A3** below) and by assumption on $\mathfrak{L}$, we have for all $i \in \{1, \dots, N\}$ and all $(x, y) \in \mathsf{X} \times \mathsf{Y}$,

$$\begin{aligned}
&\int_{(\mathbf{R}^d)^N} \nabla_{\theta^i} \mathfrak{L}\Big(y, \frac{1}{N}\sum_{j=1}^N \phi(\theta^j, z^j, x)\Big)\gamma(z^1)\dots\gamma(z^N)\mathrm{d}z^1\dots\mathrm{d}z^N \\
&= -\frac{1}{N^2}\sum_{j=1}^N \int_{(\mathbf{R}^d)^N} (y - \phi(\theta^j, z^j, x))\nabla_\theta\phi(\theta^i, z^i, x)\gamma(z^1)\dots\gamma(z^N)\mathrm{d}z^1\dots\mathrm{d}z^N \qquad (3.2.3)\\
&= -\frac{1}{N^2}\Big[\sum_{j=1, j\neq i}^N (y - \langle\phi(\theta^j, \cdot, x), \gamma\rangle)\langle\nabla_\theta\phi(\theta^i, \cdot, x), \gamma\rangle + \langle(y - \phi(\theta^i, \cdot, x))\nabla_\theta\phi(\theta^i, \cdot, x), \gamma\rangle\Big],
\end{aligned}$$

where we have used the notation $\langle U, \nu\rangle = \int_{\mathbf{R}^q} U(z)\nu(\mathrm{d}z)$ for any integrable function $U : \mathbf{R}^q \to \mathbf{R}$ w.r.t. a measure ν (with a slight abuse of notation, we denote by γ the measure $\gamma(z)\mathrm{d}z$). Second, for $\theta \in \mathbf{R}^{d+1}$, we have

$$\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) = \begin{pmatrix} \nabla_m \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) \\ \partial_\rho \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_0^2}(m - m_0) \\ \frac{d}{\sigma_0^2} g'(\rho)g(\rho) - d\frac{g'(\rho)}{g(\rho)} \end{pmatrix}. \qquad (3.2.4)$$

In conclusion, the SGD (3.2.2) writes: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases}
\theta_{k+1}^i = \theta_k^i - \dfrac{\eta}{N^2}\displaystyle\sum_{j=1, j\neq i}^N \Big(\langle\phi(\theta_k^j, \cdot, x_k), \gamma\rangle - y_k\Big)\langle\nabla_\theta\phi(\theta_k^i, \cdot, x_k), \gamma\rangle \\
\qquad\qquad -\dfrac{\eta}{N^2}\Big\langle(\phi(\theta_k^i, \cdot, x_k) - y_k)\nabla_\theta\phi(\theta_k^i, \cdot, x_k), \gamma\Big\rangle - \dfrac{\eta}{N}\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{\theta_k^i}^1 | P_0^1) \\
\theta_0^i \sim \mu_0.
\end{cases} \qquad (3.2.5)$$

We shall call this algorithm *idealised* SGD because it contains an intractable term given by the integral w.r.t. γ. This has motivated the development of methods where this integral is replaced by an unbiased Monte Carlo estimator (see [BCKW15]) as detailed below.

***Bayes-by-Backprop* SGD**. The second SGD algorithm we study is based on an approximation, for $i \in \{1, \dots, N\}$, of $\int_{(\mathbf{R}^d)^N}(y - \phi(\theta^j, z^j, x))\nabla_\theta\phi(\theta^i, z^i, x)\gamma(z^1)\dots\gamma(z^N)\mathrm{d}z^1\dots\mathrm{d}z^N$ (see (3.2.3)) by

$$\frac{1}{B}\sum_{\ell=1}^B \big(y - \phi(\theta^j, \mathsf{Z}^{j,\ell}, x)\big)\nabla_\theta\phi(\theta^i, \mathsf{Z}^{i,\ell}, x) \qquad (3.2.6)$$

where $B \in \mathbf{N}^*$ is a fixed integer and $(\mathsf{Z}^{q,\ell}, q \in \{i, j\}, 1 \leq \ell \leq B)$ is a i.i.d finite sequence of random variables distributed according to $\gamma(z)\mathrm{d}z$. In this case, for $N \geq 1$, given a dataset $(x_k, y_k)_{k\geq 0}$, the maximization of $\theta \in \mathbf{R}^{d+1} \mapsto \mathrm{E}_{\mathrm{lbo}}^N(\boldsymbol{\theta}, x, y)$ with a SGD algorithm is the following: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases}
\theta_{k+1}^i = \theta_k^i - \dfrac{\eta}{N^2 B}\displaystyle\sum_{j=1}^N\sum_{\ell=1}^B \big(\phi(\theta_k^j, \mathsf{Z}_k^{j,\ell}, x_k) - y_k\big)\nabla_\theta\phi(\theta_k^i, \mathsf{Z}_k^{i,\ell}, x_k) - \dfrac{\eta}{N}\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{\theta_k^i}^1 | P_0^1) \\
\theta_0^i = (m_0^i, \rho_0^i) \sim \mu_0,
\end{cases} \qquad (3.2.7)$$

where $\eta > 0$ and $(\mathsf{Z}_k^{j,\ell}, 1 \leq j \leq N, 1 \leq \ell \leq B, k \geq 0)$ is a i.i.d sequence of random variables distributed according to γ.

3.3 Law of large numbers for the idealized SGD

Assumptions and notations. When E is a metric space and $\mathscr{I} = \mathbf{R}_+$ or $\mathscr{I} = [0, T]$ $(T \geq 0)$, we denote by $\mathcal{D}(\mathscr{I}, E)$ the Skorohod space of càdlàg functions on $\mathscr{I}$ taking values in E and $\mathcal{C}(\mathscr{I}, E)$

the space of continuous functions on $\mathscr{I}$ taking values in E. The evolution of the parameters $(\{\theta_k^i, i=1,\dots,N\})_{k\geq 1}$ defined by (3.2.5) is tracked through their empirical distribution ν_k^N (for $k\geq 0$) and its scaled version μ_t^N (for $t\in\mathbf{R}_+$), which are defined as follows:

$$\nu_k^N := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_k^i} \quad \text{and} \quad \mu_t^N := \nu_{\lfloor Nt\rfloor}^N, \text{ where the } \theta_k^i\text{'s are defined (3.2.5).} \tag{3.3.1}$$

Fix $T>0$. For all $N\geq 1$, $\mu^N := \{\mu_t^N, t\in[0,T]\}$ is a random element of $\mathcal{D}([0,T],\mathcal{P}(\mathbf{R}^{d+1}))$, where $\mathcal{P}(\mathbf{R}^{d+1})$ is endowed with the weak convergence topology. For $N\geq 1$ and $k\geq 1$, we introduce the following $\boldsymbol{\sigma}$-algebras:

$$\mathcal{F}_0^N = \boldsymbol{\sigma}(\theta_0^i, 1\leq i\leq N) \quad \text{and} \quad \mathcal{F}_k^N = \boldsymbol{\sigma}(\theta_0^i, (x_q,y_q), 1\leq i\leq N, 0\leq q\leq k-1). \tag{3.3.2}$$

Recall $q_\theta^1 : \mathbf{R}^d\to\mathbf{R}_+$ be the pdf of $\mathcal{N}(m, g(\rho)^2 I_d)$ ($\theta=(m,\rho)\in\mathbf{R}^{d+1}$). In this work, we assume the following.

A1. There exists a pdf $\gamma : \mathbf{R}^d\to\mathbf{R}_+$ such that for all $\theta\in\mathbf{R}^{d+1}$, $q_\theta^1\mathrm{d}x = \Psi_\theta\#\gamma\mathrm{d}x$, where $\{\Psi_\theta, \theta\in\mathbf{R}^{d+1}\}$ is a family of $\mathcal{C}^1$-diffeomorphisms over $\mathbf{R}^d$ such that for all $z\in\mathbf{R}^d$, $\theta\in\mathbf{R}^{d+1}\mapsto\Psi_\theta(z)$ is of class $\mathcal{C}^\infty$. Finally, there exists $\mathfrak{b}:\mathbf{R}^d\to\mathbf{R}_+$ such that for all multi-index $\alpha\in\mathbf{N}^{d+1}$ with $|\alpha|\geq 1$, there exists $C_\alpha>0$, for all $z\in\mathbf{R}^d$ and $\theta=(\theta_1,\dots,\theta_{d+1})\in\mathbf{R}^{d+1}$,

$$\left|\partial_\alpha\Psi_\theta(z)\right| \leq C_\alpha\mathfrak{b}(z) \quad \text{with for all } q\geq 1,\ \langle\mathfrak{b}^q,\gamma\rangle<+\infty, \tag{3.3.3}$$

where $\partial_\alpha = \partial_{\theta_1}^{\alpha_1}\dots\partial_{\theta_{d+1}}^{\alpha_{d+1}}$ and $\partial_{\theta_j}^{\alpha_j}$ is the partial derivatives of order α_j w.r.t. to θ_j.

A2. The sequence $\{(x_k,y_k)\}_{k\geq 0}$ is i.i.d. w.r.t. $\pi\in\mathcal{P}(\mathsf{X}\times\mathsf{Y})$. The set $\mathsf{X}\times\mathsf{Y}\subset\mathbf{R}^d\times\mathbf{R}$ is compact. For all $k\geq 0$, $(x_k,y_k)\perp\!\!\!\perp\mathcal{F}_k^N$, where $\mathcal{F}_k^N$ is defined in (3.3.2).

A3. The activation function $s:\mathbf{R}^d\times\mathsf{X}\to\mathbf{R}$ belongs to $\mathcal{C}_b^\infty(\mathbf{R}^d\times\mathsf{X})$ (the space of smooth functions over $\mathbf{R}^d\times\mathsf{X}$ whose derivatives of all order are bounded).

A4. The initial parameters $(\theta_0^i)_{i=1}^N$ are i.i.d. w.r.t. $\mu_0\in\mathcal{P}(\mathbf{R}^{d+1})$ which has compact support.

Note that **A1** is satisfied when γ is the pdf of $\mathcal{N}(0,I_d)$ and $\Psi_\theta(z)=m+g(\rho)z$, with $\mathfrak{b}(z)=1+|z|$. With these assumptions, for every fixed $T>0$, the sequence $(\{\theta_k^i, i=1,\dots,N\})_{k=0,\dots,\lfloor NT\rfloor}$ defined by (3.2.5) is a.s. bounded:

Lemme 3.1 (Uniform bound on the parameters)**.** *Assume* **A1**→**A4**. *Then, there exists $C>0$ such that a.s. for all $T>0$, $N\geq 1$, $i\in\{1,\dots,N\}$, and $0\leq k\leq\lfloor NT\rfloor$, $|\theta_k^i|\leq Ce^{[C(2+T)]T}$.*

Lemma 3.1 implies that a.s. for all $T>0$ and $N\geq 1$, $\mu^N\in\mathcal{D}([0,T],\mathcal{P}(\Theta_T))$, where

$$\Theta_T = \{\theta\in\mathbf{R}^{d+1}, |\theta|\leq Ce^{[C(2+T)]T}\}.$$

Law of large numbers for $(\mu^N)_{N\geq 1}$ defined in (3.3.1). The first main result of this work is the following.

Theorem 3.1. *Assume* **A1**→**A4**. *Let $T>0$. Then, the sequence $(\mu^N)_{N\geq 1}\subset\mathcal{D}([0,T],\mathcal{P}(\Theta_T))$ defined in* (3.3.1) *converges in probability to the unique deterministic solution $\bar\mu\in\mathcal{C}([0,T],\mathcal{P}(\Theta_T))$ to the following measure-valued evolution equation: $\forall f\in\mathcal{C}^\infty(\Theta_T)$ and $\forall t\in[0,T]$,*

$$\begin{aligned}\langle f,\bar\mu_t\rangle - \langle f,\mu_0\rangle = &-\eta\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\big\langle\phi(\cdot,\cdot,x)-y,\bar\mu_s\otimes\gamma\big\rangle\big\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar\mu_s\otimes\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\ &-\eta\int_0^t\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\bar\mu_s\big\rangle\mathrm{d}s.\end{aligned} \tag{3.3.4}$$

The proof of Theorem 3.1 is given in Section 3.8. We stress here the most important steps and used techniques. In a first step, we derive an identity satisfied by $(\mu^N)_{N\geq 1}$, namely the pre-limit equation (3.8.9); see Sec. 3.8.1. Then we show in Sec. 3.8.2 that $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}([0,T],\mathcal{P}(\Theta_T))$. To do so, we check that the sequence $(\mu^N)_{N\geq 1}$ satisfies all the required assumptions of [Jak86, Theorem 3.1] when $E=\mathcal{P}(\Theta_T)$ there. In Sec. 3.8.2 we prove that every limit point of $(\mu^N)_{N\geq 1}$ satisfies the limit equation (3.3.4). Then, in Section 3.8.2, we prove that there is a unique solution of the measure-valued equation (3.3.4). To prove the uniqueness of the solution of (3.3.4), we use techniques developed in [PRT15] which are based on a representation formula for solution to measure-valued equations [Vil03, Theorem 5.34] together with estimates in Wasserstein distances between two solutions of (3.3.4) derived in [PR16]. In Section 3.8.2, we also conclude the proof of Theorem 3.1. Compared to [DGMN22, Theorem 1], the fact that $(\{\theta_k^i, i=1,\dots,N\})_{k=0,\dots,\lfloor NT\rfloor}$ defined by (3.2.5) are a.s. bounded allows to use different and more straightforward arguments to prove (i) the relative compactness in $\mathcal{D}([0,T],\mathcal{P}(\Theta_T))$ of $(\mu^N)_{N\geq 1}$ (defined in (3.3.1)) (ii) the continuity property of the operator $\mathsf{m}\mapsto \mathbf{\Lambda}_t[f](\mathsf{m})$ defined in (3.8.16) w.r.t. the topology of $\mathcal{D}([0,T],\mathcal{P}(\Theta_T))$ and (iii) $(\mu^N)_{N\geq 1}$ has limit points in $\mathcal{C}([0,T],\mathcal{P}(\Theta_T))$. Step (ii) is necessary in order to pass to the limit $N\to+\infty$ in the pre-limit equation and Step (iii) is crucial since we prove that there is at most one solution of (3.3.4) in $\mathcal{C}([0,T],\mathcal{P}(\Theta_T))$. It is worthwhile to emphasize that, as $N\to\infty$, the effects of the integrated loss and of the KL terms are balanced, as conjectured in [HMD+22].

To avoid further technicalities, we have chosen what may seem restrictive assumptions on the data or the activation function. Note however that it readily extends to unbounded set X, and also unbounded Y assuming that π as polynomial moments of sufficiently high order. Also, RELU (or more easily leaky RELU) may be considered by using weak derivatives (to consider the singularity at 0), and a priori moment bounds on the weights.

3.4 LLN for the *Bayes-by-Backprop* SGD

The sequence $\{\theta_k^i, i\in\{1,\dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ defined recursively by the algorithm (3.2.7) is in general not bounded, since $\nabla_\theta\phi(\theta,\mathsf{Z},x)$ is not necessarily bounded if $\mathsf{Z}\sim\gamma(s)\mathrm{d}z$. Therefore, we cannot expect Lemma 3.1 to hold for $\{\theta_k^i, i\in\{1,\dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ set by (3.2.7). Thus, the sequence $\{\theta_k^i, i\in\{1,\dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ is considered on the whole space $\mathbf{R}^{d+1}$.

Wasserstein spaces and results. For $N\geq 1$, and $k\geq 1$, we set

$$\mathcal{F}_k^N=\boldsymbol{\sigma}\Big(\theta_0^i,\mathsf{Z}_q^{j,\ell},(x_q,y_q),1\leq i,j\leq N,1\leq\ell\leq B,0\leq q\leq k-1\Big\}\Big). \tag{3.4.1}$$

In addition to **A1**→**A4** (where in **A2**, when $k\geq 1$, $\mathcal{F}_k^N$ is now the one defined in (3.4.1)), we assume:

A5. The sequences $(\mathsf{Z}_k^{j,\ell},1\leq j\leq N,1\leq\ell\leq B,k\geq 0)$ and $((x_k,y_k),k\geq 0)$ are independent. In addition, for $k\geq 0$, $\big((x_k,y_k),\mathsf{Z}_k^{j,\ell},1\leq j\leq N,1\leq\ell\leq B\big)\perp\!\!\!\perp\mathcal{F}_k^N$.

Note that the last statement of **A5** implies the last statement of **A2**. We introduce the scaled empirical distribution of the parameters of the algorithm (3.2.7), i.e. for $k\geq 0$ and $t\geq 0$:

$$\nu_k^N:=\frac{1}{N}\sum_{i=1}^N\delta_{\theta_k^i}\quad\text{and}\quad\mu_t^N:=\nu_{\lfloor Nt\rfloor}^N,\ \text{where the }\theta_k^i\text{'s are defined (3.2.7).} \tag{3.4.2}$$

One can no longer rely on the existence of a compact subset $\Theta_T\subset\mathbf{R}^{d+1}$ such that a.s. $(\mu^N)_{N\geq 1}\subset\mathcal{D}([0,T],\mathcal{P}(\Theta_T))$, where $\mu^N=\{t\geq 0\mapsto\mu_t^N\}$ is defined in (3.4.2). For this reason, we will work in Wasserstein spaces $\mathcal{P}_q(\mathbf{R}^{d+1})$, $q\geq 0$, which, we recall, are defined by

$$\mathcal{P}_q(\mathbf{R}^{d+1})=\Big\{\nu\in\mathcal{P}(\mathbf{R}^{d+1}),\int_{\mathbf{R}^{d+1}}|\theta|^q\nu(\mathrm{d}\theta)<+\infty\Big\}. \tag{3.4.3}$$

These spaces are endowed with the Wasserstein metric W_q, see e.g. [San15, Chapter 5] for more materials on Wasserstein spaces. For all $q \geq 0$, $(\mu^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{P}_q(\mathbf{R}^{d+1}))$. The second main results of this work is a LLN for $(\mu^N)_{N\geq 1}$ defined in (3.4.2).

Theorem 3.2. *Assume* **A1**→**A5**. *Let* $\gamma_0 > 1 + \frac{d+1}{2}$. *Then, the sequence* $(\mu^N)_{N\geq 1}$ *defined in* (3.4.2) *converges in probability in* $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ *to a deterministic element* $\bar{\mu} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$, *where* $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ *is the unique solution in* $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ *to the following measure-valued evolution equation:*$\forall f \in \mathcal{C}_b^\infty(\mathbf{R}^{d+1})$ *and* $\forall t \in \mathbf{R}_+$,

$$\begin{aligned}\langle f, \bar{\mu}_t\rangle - \langle f, \mu_0\rangle = -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ - \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_s\big\rangle \mathrm{d}s. \qquad (3.4.4)\end{aligned}$$

Theorem 3.2 is proved in Section 3.9. Since $\{\theta_k^i, i \in \{1,\dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ defined by (3.2.7) is not bounded in general, we work in the space $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. The proof of Theorem 3.2 is more involved than that of Theorem 3.1, and generalizes the latter to the case where the parameters of the SGD algorithm are unbounded. We prove that $(\mu^N)_{N\geq 1}$ (defined in (3.4.2)) is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. To this end we now use [Jak86, Theorem 4.6]. The compact containment, which is the purpose of Lemma 3.19, is not straightforward since $\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1})$ is not compact contrary to Theorem 3.1 where we used the compactness of $\mathcal{P}(\Theta_T)$. More precisely, the compact containment of $(\mu^N)_{N\geq 1}$ relies on a characterization of the compact subsets of $\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1})$ (see Proposition 3.17) and moment estimates on $\{\theta_k^i, i \in \{1,\dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ (see Lemma 3.16). We also mention that contrary to what is done in the proof of Theorem 3.1, we do not show that every limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ is continuous in time but we still manage to prove that they all satisfy (3.4.4). Then, using the duality formula for the W_1-distance together with rough estimates on the jumps of $t \mapsto \langle f, \mu_t^N\rangle$ (for f uniformly Lipschitz over $\mathbf{R}^{d+1}$), we then show that every limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ belongs a.s. to $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$. Again this is important since we have uniqueness of (3.4.4) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$.

We conclude this section with the following important uniqueness result.

Proposition 3.3. *Under the assumptions of Theorems 3.1 and 3.2, the solution to* (3.3.4) *is independent of* T *and is equal to the solution to* (3.4.4).

This uniqueness result states that both idealized and *Bayes-by-backprop* SGD have the same limiting behavior. It is also noteworthy that the mini-batch B is held fixed B. The effect of batch size can be seen at the level of the central limit theorem, which we leave for future work.

3.5 The *Minimal-VI* SGD algorithm

The idea behing the *Bayes-by-Backprop* SGD stems from the fact that there are integrals wrt γ in the loss function that cannot be computed in practice and it is quite natural up to a reparameterization trick, to replace these integrals by a Monte Carlo approximation (with i.i.d. gaussian random variables). To devise a new cheaper algorithm based on the only terms impacting the asymptotic limit, we directly analyse the limit equation (3.3.4) and remark that it can be rewritten as, $\forall f \in \mathcal{C}^\infty(\Theta_T)$ and $\forall t \in [0,T]$,

$$\begin{aligned}&\langle f, \bar{\mu}_t\rangle - \langle f, \mu_0\rangle \\ &= -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}\times(\mathbf{R}^d)^2} \big\langle \phi(\cdot, z_1, x) - y, \bar{\mu}_s\big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z_2, x), \bar{\mu}_s\big\rangle \gamma^{\otimes 2}(\mathrm{d}z_1\mathrm{d}z_2)\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &\quad - \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_s\big\rangle \mathrm{d}s.\end{aligned}$$

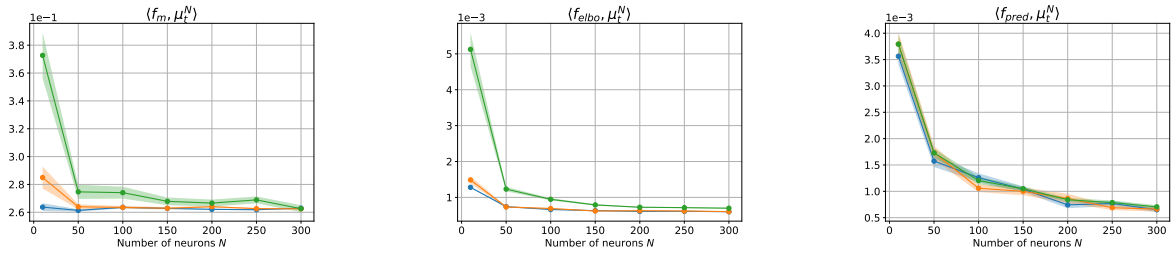


Figure 3.1: Convergence of $\langle f, \mu_T^N \rangle$ to $\langle f, \bar{\mu}_T \rangle$, for the idealized (blue), *Bayes-by-Backprop* (orange) and *Minimal-VI* (green) SGD algorithms over 50 realizations.

Thus, the integration over $\gamma^{\otimes 2}$ can be considered as that over π, i.e., we can consider them as two more data variables that only need to be sampled at each new step. In this case, the SGD (3.2.7) becomes: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases} \theta_{k+1}^i = \theta_k^i - \dfrac{\eta}{N^2} \displaystyle\sum_{j=1}^{N} \big(\phi(\theta_k^j, \mathsf{Z}_k^1, x_k) - y_k\big) \nabla_\theta \phi(\theta_k^i, \mathsf{Z}_k^2, x_k) - \dfrac{\eta}{N} \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{\theta_k^i}^1 | P_0^1) \\ \theta_0^i = (m_0^i, \rho_0^i) \sim \mu_0, \end{cases} \tag{3.5.1}$$

where $\eta > 0$ and $(\mathsf{Z}_k^p, p \in \{1,2\}, k \geq 0)$ is a i.i.d sequence of random variables distributed according to $\gamma^{\otimes 2}$. We call this backpropagation scheme *minimal- VI SGD* which is much cheaper in terms of computational complexity, with the same limiting behavior as we now discuss.

We introduce the $\boldsymbol{\sigma}$-algebra for $N, k \geq 1$:

$$\mathcal{F}_k^N = \boldsymbol{\sigma}\Big(\theta_0^i, \mathsf{Z}_q^p, (x_q, y_q), 1 \leq i \leq N, p \in \{1,2\}, 0 \leq q \leq k-1\Big\}\Big). \tag{3.5.2}$$

In addition to **A1**→**A4** (where in **A2**, $\mathcal{F}_k^N$ is now the one defined above in (3.5.2) when $k \geq 1$), the following assumption

A6. The sequences $(\mathsf{Z}_k^p, p \in \{1,2\}, k \geq 0)$ and $((x_k, y_k), k \geq 0)$ are independent. In addition, for $k \geq 0$, $\big((x_k, y_k), \mathsf{Z}_k^p, p \in \{1,2\}\big) \perp\!\!\!\perp \mathcal{F}_k^N$, where $\mathcal{F}_k^N$ is defined in (3.5.2).

Set for $k \geq 0$ and $t \geq 0$, $\nu_k^N := \frac{1}{N} \sum_{i=1}^N \delta_{\theta_k^i}$ and $\mu_t^N := \nu_{\lfloor Nt \rfloor}^N$, where the θ_k^i's are defined in (3.5.1). The last main result of this work states that the sequence $(\mu^N)_{N \geq 1}$ satisfies the same law of large numbers when $N \to +\infty$ as the one satisfied by (3.4.2), whose proof will be omitted as it is the same as the one made for Theorem 3.2.

Theorem 3.4. *Assume* **A1**→**A4** *and* **A6**. *Then, the sequence of* $(\mu^N)_{N \geq 1}$ *satisfies all the statements of Theorem 3.2.*

3.6 Numerical experiments

In this section we illustrate the theorems 3.1, 3.2, and 3.4 using the following toy model. We set $d = 5$. Given $\theta^* \in \mathbf{R}^d$ (drawn from a normal distribution and scaled to the unit norm), we draw i.i.d observations as follows: Given $x \sim \mathcal{U}([-1,1]^d)$, we draw $y = \tanh(x^\top \theta^*) + \epsilon$, where ϵ is zero mean with variance 10^{-4}. The initial distribution of parameters is centered around the prior: $\boldsymbol{\theta}_0 \sim (\mathcal{N}(m_0, 0.01 I_d) \times \mathcal{N}(g^{-1}(\sigma_0), 0.01))^{\otimes N}$, with $m_0 = 0$ and $\sigma_0 = 0.2$. Since the idealized algorithm cannot be implemented exactly, a mini-batch of size 100 is used as a proxy for the following comparisons of the different algorithms. For the algorithm (3.2.7) SGD we set $B = 1$.

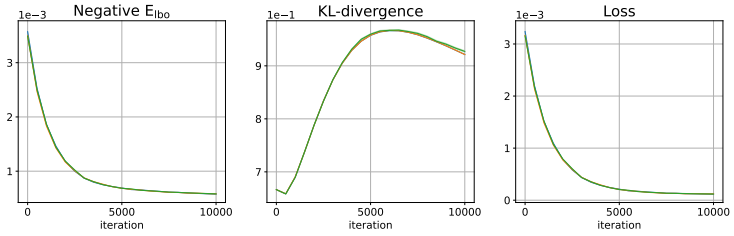


Figure 3.2: Decay of the negative ELBO (left) and its two components (KL (middle), loss (right)) during the training process done by the idealized (blue), *Bayes-by-Backprop* (orange) and *Minimal-VI* (green) SGD algorithms, for $N = 10000$.

Evolution and limit of the distribution Fig. 1.24 displays the histograms of $\{F(\theta^i_{\lfloor Nt \rfloor}), i = 1, \dots, N\}$ ($F(\theta) = \|m\|_2, g(\rho)$ or m, where $\theta = (m, \rho) \in \mathbf{R}^d \times \mathbf{R}$), for $N = 10000$, at initialization, halfway through training, and at the end of training. The empirical distributions illustrated by these histograms are very similar over the course of training. It can be seen that for $N = 10000$ the limit of the mean field is reached.

Convergence with respect to the numbers of neurons. We investigate here the speed of convergence of μ^N_t to $\bar{\mu}_t$ (as $N \to +\infty$), when tested against test functions f. More precisely, we fix a time T (end of training) and Figure 3.1 represents the empirical mean of $\langle f, \mu^N_T \rangle$ over 50 realizations. The test functions used for this experiment are $f_m(\theta) = \|m\|_2$, $f_{\mathrm{Elbo}(\theta)} = -\hat{\mathrm{E}}_{\mathrm{lbo}}(\theta)^N$ where $\hat{\mathrm{E}}_{\mathrm{lbo}}$ is the empirical $\mathrm{E}^N_{\mathrm{lbo}}$ (see (3.2.1)) computed with 100 samples of (x, y) and $(z^1, \dots, z^N)$. Finally, $f_{pred}(\theta) = \hat{\mathbb{E}}_x \left[\hat{\mathbb{V}}_{\boldsymbol{w} \sim q^N_{\boldsymbol{\theta}}} [f^N_{\boldsymbol{w}}(x)]^{1/2} \right]$ where $\hat{\mathbb{E}}$ and $\hat{\mathbb{V}}$ denote respectively the empirical mean and the empirical variance over 100 samples. All algorithms are converging to the same limit and are performing similarly even with a limited number of neurons ($N = 300$ in this example).

Convergence with respect to time. This section illustrates the training process of a BNN with a given number of neurons $N = 10000$. In Figure 3.2, we plot the negative ELBO on a test set and its two components, the loss and the KL-divergence terms. Figure 3.2 shows that the BNN is able to learn on this specific task and all algorithms exhibit a similar performance. It illustrates the trajectorial convergence of $\{\mu^N_t, t \in [0, T]\}_{N \geq 1}$ to $\{\bar{\mu}_t, t \in [0, T]\}$ as $N \to +\infty$.

Behavior around the limit $\bar{\mu}$. On Figure 1.5, we plot the boxplots of $\langle f, \mu^N_t \rangle$ for 50 realizations and $N = 10000$, at different times of the training. *Minimal-VI* scheme (which is computationally cheaper as explained in 3.5) exhibit a larger variance than the other algorithms.

3.7 Conclusion

By establishing the limit behavior of the idealized SGD for the variational inference of BNN with the weighting suggested by [HMD+22], we have rigorously shown that the most-commonly used in practice *Bayes-by-Backprop* scheme indeed exhibits the same limit behavior. Furthermore, the analysis of the limit equation led us to validate the correct scaling of the KL divergence term in with respect to the loss. Notably, the mean-field limit dynamics has also helped us to devise a far less costly new SGD algorithm, the *Minimal-VI*. This scheme shares the same limit behavior, but only stems from the non-vanishing asymptotic contributions, hence the reduction of the computational cost. Aside from confirming the analytical results, the first simulations presented here show that the three algorithms, while having the same limit, may differ in terms of variance. Thus, deriving a CLT result and discussing the right trade-off between computational complexity and variance will be done in future work. Also, on a more general level regarding uncertainty quantification, an interesting question is to analyse the impact of the correct scaling of the KL divergence term on the error calibration and how to apply the same analysis in the context of deep ensembles.

3.8 Proof of Theorem 3.1

For simplicity, we prove the theorem 3.1 when $T = 1$, and we denote Θ_1 simply by Θ. In this section we assume **A1**–**A4**.

3.8.1 Pre-limit equation (3.8.9) and error terms in (3.8.9)

Derivation of the pre-limit equation

The aim of this section is to establish the so-called pre-limit equation (3.8.9), which will be our starting point to derive Equation (3.3.4). Let $N \geq 1$, $k \in \{0, \dots, N\}$, and $f \in \mathcal{C}^\infty(\Theta)$. Recall that by Lemma 3.1 and since $0 \leq k \leq N$, a.s. $\theta_k^i \in \Theta$, and thus a.s. $f(\theta_k^i)$ is well-defined. The Taylor-Lagrange formula yields

$$\begin{aligned}\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle &= \frac{1}{N}\sum_{i=1}^N f(\theta_{k+1}^i) - f(\theta_k^i) \\ &= \frac{1}{N}\sum_{i=1}^N \nabla_\theta f(\theta_k^i)\cdot(\theta_{k+1}^i - \theta_k^i) + \frac{1}{2N}\sum_{i=1}^N (\theta_{k+1}^i - \theta_k^i)^T \nabla^2 f(\widehat{\theta}_k^i)(\theta_{k+1}^i - \theta_k^i),\end{aligned}$$

where, for all $i \in \{1, \dots, N\}$, $\widehat{\theta}_k^i \in (\theta_k^i, \theta_{k+1}^i) \subset \Theta$. Using (3.2.5), we then obtain

$$\begin{aligned}\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle = &-\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1, j\neq i}^N \big(\langle\phi(\theta_k^j, \cdot, x_k), \gamma\rangle - y_k\big)\big\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i, \cdot, x_k), \gamma\big\rangle \\ &- \frac{\eta}{N^2}\big\langle\big(\phi(\cdot, \cdot, x_k) - y_k\big)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot, \cdot, x_k), \nu_k^N\otimes\gamma\big\rangle \\ &- \frac{\eta}{N}\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \nu_k^N\big\rangle + \mathbf{R}_k^N[f], \end{aligned} \tag{3.8.1}$$

where

$$\mathbf{R}_k^N[f] := \frac{1}{2N}\sum_{i=1}^N (\theta_{k+1}^i - \theta_k^i)^T\nabla^2 f(\widehat{\theta}_k^i)(\theta_{k+1}^i - \theta_k^i).$$

Let us define

$$\begin{aligned}\mathbf{D}_k^N[f] := &\mathbf{E}\Big[-\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1, j\neq i}^N \big(\langle\phi(\theta_k^j, \cdot, x_k), \gamma\rangle - y_k\big)\big\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i, \cdot, x_k), \gamma\big\rangle\Big|\mathcal{F}_k^N\Big] \\ &- \mathbf{E}\Big[\frac{\eta}{N^2}\big\langle(\phi(\cdot, \cdot, x_k) - y_k)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot, \cdot, x_k), \nu_k^N\otimes\gamma\big\rangle\Big|\mathcal{F}_k^N\Big]. \end{aligned} \tag{3.8.2}$$

Note that using (3.8.26) and (3.8.28) together with the fact that $|\nabla_\theta f(\theta_k^i)| \leq \sup_{\theta\in\Theta}|\nabla_\theta f(\theta)|$, the integrant in (3.8.2) is integrable and thus $\mathbf{D}_k^N[f]$ is well defined. Using the fact that $(x_k, y_k) \perp\!\!\!\perp \mathcal{F}_k^N$ by **A2** and that $\{\theta_k^i, i = 1, \dots, N\}$ is $\mathcal{F}_k^N$-measurable by (3.2.5), we have:

$$\begin{aligned}\mathbf{D}_k^N[f] = &-\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1, j\neq i}^N\int_{\mathsf{X}\times\mathsf{Y}} \big(\langle\phi(\theta_k^j, \cdot, x), \gamma\rangle - y\big)\big\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i, \cdot, x), \gamma\big\rangle\pi(\mathrm{d}x, \mathrm{d}y) \\ &- \frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\big\langle(\phi(\cdot, \cdot, x) - y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot, \cdot, x), \nu_k^N\otimes\gamma\big\rangle\pi(\mathrm{d}x, \mathrm{d}y). \end{aligned} \tag{3.8.3}$$

Introduce also

$$\begin{aligned}\mathbf{M}_k^N[f] := &-\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1, j\neq i}^N (\langle\phi(\theta_k^j, \cdot, x_k), \gamma\rangle - y_k)\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i, \cdot, x_k), \gamma\rangle \\ &- \frac{\eta}{N^2}\langle(\phi(\cdot, \cdot, x_k) - y_k)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot, \cdot, x_k), \nu_k^N\otimes\gamma\rangle - \mathbf{D}_k^N[f].\end{aligned}$$

Note that $\mathbf{E}\big[\mathbf{M}_k^N[f]|\mathcal{F}_k^N\big] = 0$. Equation (3.8.1) then writes

$$\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle = \mathbf{D}_k^N[f] + \mathbf{M}_k^N[f] - \frac{\eta}{N}\big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \nu_k^N\big\rangle + \mathbf{R}_k^N[f].$$

Notice also that

$$\begin{aligned}
\mathbf{D}_k^N[f] &= -\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1}^N \int_{\mathsf{X}\times\mathsf{Y}} (\langle\phi(\theta_k^j,\cdot,x),\gamma\rangle - y)\langle \nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&\quad + \frac{\eta}{N^3}\sum_{i=1}^N \int_{\mathsf{X}\times\mathsf{Y}} (\langle\phi(\theta_k^i,\cdot,x),\gamma\rangle - y)\langle \nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&\quad - \frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}} \langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \nu_k^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&= -\frac{\eta}{N}\int_{\mathsf{X}\times\mathsf{Y}} \langle\phi(\cdot,\cdot,x)-y,\nu_k^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&\quad + \frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}} \Big\langle(\langle\phi(\cdot,\cdot,x),\gamma\rangle - y)\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle, \nu_k^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&\quad - \frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}} \langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \nu_k^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y). \qquad (3.8.4)
\end{aligned}$$

Now, we define for $t\in[0,1]$:

$$\mathbf{D}_t^N[f] := \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{D}_k^N[f],\quad \mathbf{R}_t^N[f] := \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{R}_k^N[f],\quad \text{and}\quad \mathbf{M}_t^N[f] := \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{M}_k^N[f]. \qquad (3.8.5)$$

We can rewrite $\mathbf{D}_t^N[f]$ has follows:

$$\mathbf{D}_t^N[f] = \sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}} N\mathbf{D}_{\lfloor Ns\rfloor}^N[f]\mathrm{d}s = N\int_0^t \mathbf{D}_{\lfloor Ns\rfloor}^N[f]\mathrm{d}s - N\int_{\frac{\lfloor Nt\rfloor}{N}}^t \mathbf{D}_{\lfloor Ns\rfloor}^N[f]\mathrm{d}s.$$

Since $\nu_{\lfloor Ns\rfloor}^N = \mu_s^N$ (by definition, see (3.3.1)), we have, using also (3.8.4) with $k=\lfloor Ns\rfloor$,

$$\begin{aligned}
\mathbf{D}_t^N[f] &= -\eta\int_0^t\int_{\mathsf{X}\times\mathsf{Y}} \langle\phi(\cdot,\cdot,x)-y,\mu_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad + \frac{\eta}{N}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}} \Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad - \frac{\eta}{N}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}} \Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s - \mathbf{V}_t^N[f], \qquad (3.8.6)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{V}_t^N[f] &:= -\eta\int_{\frac{\lfloor Nt\rfloor}{N}}^t\int_{\mathsf{X}\times\mathsf{Y}} \langle\phi(\cdot,\cdot,x)-y,\mu_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad + \frac{\eta}{N}\int_{\frac{\lfloor Nt\rfloor}{N}}^t\int_{\mathsf{X}\times\mathsf{Y}} \Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad - \frac{\eta}{N}\int_{\frac{\lfloor Nt\rfloor}{N}}^t\int_{\mathsf{X}\times\mathsf{Y}} \Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s.
\end{aligned}$$

On the other hand, we also have for $t\in[0,1]$,

$$\sum_{k=0}^{\lfloor Nt\rfloor-1} -\frac{\eta}{N}\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\nu_k^N\big\rangle = -\eta\int_0^{\frac{\lfloor Nt\rfloor}{N}}\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\mu_s^N\big\rangle\mathrm{d}s. \qquad (3.8.7)$$

We finally set:

$$\mathbf{W}_t^N[f] := -\mathbf{V}_t^N[f] + \eta \int_{\frac{\lfloor Nt \rfloor}{N}}^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle \mathrm{d}s. \tag{3.8.8}$$

Since $\langle f, \mu_t^N \rangle - \langle f, \mu_0^N \rangle = \sum_{k=0}^{\lfloor Nt \rfloor - 1} \langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle$, we deduce from (3.8.1), (3.8.5), (3.8.6), (3.8.7) and (3.8.8), the so-called pre-limit equation satisfied by μ^N: for $N \geq 1$, $t \in [0,1]$, and $f \in \mathcal{C}^\infty(\Theta)$,

$$\begin{aligned}
\langle f, \mu_t^N \rangle - \langle f, \mu_0^N \rangle &= -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&\quad - \eta \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \rangle \mathrm{d}s \\
&\quad + \frac{\eta}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&\quad - \frac{\eta}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\
&\quad + \mathbf{M}_t^N[f] + \mathbf{W}_t^N[f] + \mathbf{R}_t^N[f].
\end{aligned} \tag{3.8.9}$$

The last five terms in (3.8.9) **are error terms**

The purpose of this section is to show that the last five terms appearing in the r.h.s. of (3.8.9) are error terms when $N \to +\infty$. For $J \in \mathbf{N}^*$ and $f \in \mathcal{C}^J(\Theta)$, set $\|f\|_{\mathcal{C}^J(\Theta)} := \sum_{|k| \leq J} \|\partial_k f\|_{\infty,\Theta}$, where $\|g\|_{\infty,\Theta} = \sup_{\theta\in\Theta} |g(\theta)|$ for $g : \Theta \to \mathbf{R}^m$.

Lemma 3.5 (Error terms)**.** *Assume* **A1**→**A4**. *Then, there exists $C > 0$ such that a.s. for all $f \in \mathcal{C}^\infty(\Theta)$ and $N \geq 1$,*

1. $\frac{\eta}{N} \int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \Big| \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \leq C\|f\|_{\mathcal{C}^1(\Theta)}/N.$
2. $\frac{\eta}{N} \int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \Big| \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \leq C\|f\|_{\mathcal{C}^1(\Theta)}/N.$
3. $\sup_{t\in[0,1]} |\mathbf{W}_t^N[f]| + \sup_{t\in[0,1]} |\mathbf{R}_t^N[f]| \leq C\|f\|_{\mathcal{C}^2(\Theta)}/N.$

Finally, $\sup_{t\in[0,1]} \mathbf{E}\big[|\mathbf{M}_t^N[f]|\big] \leq C\|f\|_{\mathcal{C}^1(\Theta)}/\sqrt{N}$.

Proof. All along the proof, $C > 0$ denotes a positive constant independent of $N \geq 1, k \in \{0, \ldots, N-1\}, (s,t) \in [0,1]^2, (x,y) \in \mathsf{X} \times \mathsf{Y}, \theta \in \Theta, z \in \mathbf{R}^d$, and $f \in \mathcal{C}^\infty(\Theta)$ which can change from one occurrence to another. Using (3.8.28), the Cauchy-Schwarz inequality, and the fact that $\nabla_\theta f$ is bounded over Θ imply:

$$|\langle \nabla_\theta f(\theta) \cdot \nabla_\theta \phi(\theta,\cdot,x), \gamma \rangle| \leq \langle |\nabla_\theta f(\theta) \cdot \nabla_\theta \phi(\theta,\cdot,x)|, \gamma \rangle \leq C\|f\|_{\mathcal{C}^1(\Theta)}. \tag{3.8.10}$$

Combining (3.8.26) and (3.8.10), we obtain:

$$\int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma \rangle, \mu_s^N \Big\rangle \Big| \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \leq C\|f\|_{\mathcal{C}^1(\Theta)}$$

and

$$\int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_m f \cdot \nabla_m \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \Big| \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \leq C\|f\|_{\mathcal{C}^1(\Theta)},$$

which proves Items 1 and 2.

Let us now prove Item 3. By (3.8.26) and (3.8.10), $\sup_{t\in[0,1]}|\mathbf{V}_t^N[f]| \le C\|f\|_{\mathcal{C}^1(\Theta)}/N$. On the other hand, because $f\in\mathcal{C}^\infty(\Theta)$ and $\theta\mapsto\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)$ is continuous (see (3.2.4)) over Θ which is compact, it holds, $\|\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)\|_{\infty,\Theta}<+\infty$. Hence, it holds:

$$\sup_{t\in[0,1]}\Big|\int_{\frac{\lfloor Nt\rfloor}{N}}^t\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\mu_s^N\big\rangle\mathrm{d}s\Big|\le C\|f\|_{\mathcal{C}^1(\Theta)}/N.$$

Using (3.8.8), it then holds $\sup_{t\in[0,1]}|\mathbf{W}_t^N[f]|\le C\|f\|_{\mathcal{C}^1(\Theta)}/N$. Since $f\in\mathcal{C}^\infty(\Theta)$, we have, for $N\ge1$ and $0\le k\le N-1$, $|\mathbf{R}_k^N[f]|\le\|f\|_{\mathcal{C}^2(\Theta)}\frac{C}{N}\sum_{i=1}^N|\theta_{k+1}^i-\theta_k^i|^2$. By (3.8.29) and Lemma 3.1, $|\theta_{k+1}^i-\theta_k^i|^2\le C/N^2$ and consequently, one has:

$$|\mathbf{R}_k^N[f]|\le C\|f\|_{\mathcal{C}^2(\Theta)}/N^2. \tag{3.8.11}$$

Hence, for all $t\in[0,1]$, $|\mathbf{R}_t^N[f]|\le C\|f\|_{\mathcal{C}^2(\Theta)}/N$. This proves Item 3.

Let us now prove the last item in Lemma 3.5. Let $t\in[0,1]$. We have, by (3.8.5),

$$|\mathbf{M}_t^N[f]|^2=\sum_{k=0}^{\lfloor Nt\rfloor-1}\big|\mathbf{M}_k^N[f]\big|^2+2\sum_{k<j}\mathbf{M}_k^N[f]\,\mathbf{M}_j^N[f].$$

For all $0\le k<j<\lfloor Nt\rfloor$, $\mathbf{M}_k^N[f]$ is $\mathcal{F}_j^N$-measurable (see (3.3.2)), and since $\mathbf{E}\big[\mathbf{M}_j^N[f]|\mathcal{F}_j^N\big]=0$, one deduces that $\mathbf{E}\big[\,\mathbf{M}_k^N[f]\,\mathbf{M}_j^N[f]\,\big]=\mathbf{E}\big[\mathbf{M}_k^N[f]\,\mathbf{E}\big[\mathbf{M}_j^N[f]|\mathcal{F}_j^N\big]\,\big]=0$. Hence, $\mathbf{E}[|\mathbf{M}_t^N[f]|^2]=\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}[|\mathbf{M}_k^N[f]|^2]$. By (3.8.26) and (3.8.10), one has a.s. for all $0\le k\le N-1$,

$$|\mathbf{M}_k^N[f]|\le C\|f\|_{\mathcal{C}^1(\Theta)}/N. \tag{3.8.12}$$

Hence, $\mathbf{E}[|\mathbf{M}_t^N[f]|^2]\le C\|f\|_{\mathcal{C}^1(\Theta)}/N$, which proves the last inequality in Lemma 3.5. □

3.8.2 Convergence to the limit equation as $N\to+\infty$

In this section we prove the relative compactness of $(\mu^N)_{N\ge1}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$. We then show that any of its limit points satisfies the limit equation (3.3.4).

Wasserstein spaces and duality formula

In this section we recall some basic results which will be used throughout this work on the space $\mathcal{P}(\mathcal{S})$ when $(\mathcal{S},\mathsf{d})$ is a Polish space. First when endowed with the weak convergence topology, $\mathcal{P}(\mathcal{S})$ is a Polish space [Bil99, Theorem 6.8]. In addition, $\mathcal{P}_q(\mathcal{S})=\{\nu\in\mathcal{P}(\mathcal{S}),\int_{\mathcal{S}}\mathsf{d}(w_0,w)^q\nu(\mathrm{d}w)<+\infty\}$, where $w_0\in\mathcal{S}$ is arbitrary (note that this space was defined previously in (3.4.3) when $\mathcal{S}=\mathbf{R}^{d+1}$) when endowed with the W_q metric is also a Polish space [Vil09, Theorem 6.18]. Recall also the duality formula for the W_1-distance on $\mathcal{P}_1(\mathcal{S})$ (see e.g [Vil09, Remark 6.5]):

$$\mathsf{W}_1(\mu,\nu)=\sup\Big\{\Big|\int_{\mathcal{S}}f(w)\mathrm{d}\mu(w)-\int_{\mathcal{S}}f(w)\nu(\mathrm{d}w)\Big|,\ \|f\|_{\mathrm{Lip}}\le1\Big\}. \tag{3.8.13}$$

Finally, when $\mathcal{K}\subset\mathbf{R}^{d+1}$ is compact, the convergence in W_q-distance is equivalent to the usual weak convergence on $\mathcal{P}(\mathcal{K})$ (see e.g. [Vil09, Corollary 6.13]).

Relative compactness

The main result of this section is to prove that $(\mu^N)_{N\ge1}$ is relatively compact in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$, which is the purpose of Proposition 3.7 below. To this end, we need to prove that for all $f\in\mathcal{C}^\infty(\Theta)$, every sequence $(\langle f,\mu_t^N\rangle)_{N\ge1}$ satisfies some regularity conditions, which is the purpose of the next result.

Lemma 3.6 (Regularity condition)**.** *Assume* **A1**→**A4**. *Then there exists* $C > 0$ *such that a.s. for all* $f \in \mathcal{C}^\infty(\Theta)$, $0 \le r < t \le 1$, *and* $N \ge 1$*:*

$$|\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle| \le C\|f\|_{\mathcal{C}^2(\Theta)}\Big[|t-r| + \frac{|t-r|}{N} + \frac{1}{N}\Big]. \tag{3.8.14}$$

Proof. Let $f \in \mathcal{C}^\infty(\Theta)$ and let $N \ge 1$ and $0 \le r < t \le 1$. In the following $C > 0$ is a positive constant independent of $f \in \mathcal{C}^\infty(\Theta)$, $N \ge 1$, and $0 \le r < t \le 1$, which can change from one occurrence to another. From (3.8.9), we have

$$\begin{aligned}\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle &= \mathbf{A}_{r,t}^N[f] - \eta\int_r^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \mu_s^N\rangle \mathrm{d}s \\ &\quad + \mathbf{M}_t^N[f] - \mathbf{M}_r^N[f] + \mathbf{W}_t^N[f] - \mathbf{W}_r^N[f] + \mathbf{R}_t^N[f] - \mathbf{R}_r^N[f],\end{aligned} \tag{3.8.15}$$

where

$$\begin{aligned}\mathbf{A}_{r,t}^N[f] &= -\eta\int_r^t\int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma\rangle\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y) \\ &\quad + \frac{\eta}{N}\int_r^t\int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y) \\ &\quad - \frac{\eta}{N}\int_r^t\int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y)\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y).\end{aligned}$$

By (3.8.26) and (3.8.10), $|\mathbf{A}_{r,t}^N[f]| \le C\|f\|_{\mathcal{C}^1(\Theta)}\big[|t-r| + \frac{|t-r|}{N}\big]$. In addition, since $\theta \mapsto \mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)$ is bounded over Θ (since it is smooth and Θ is compact),

$$\Big|\int_r^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \mu_s^N\rangle \mathrm{d}s\Big| \le C\|f\|_{\mathcal{C}^1(\Theta)}|t-r|.$$

Furthermore, using (3.8.12),

$$|\mathbf{M}_t^N[f] - \mathbf{M}_r^N[f]| = \Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor - 1} \mathbf{M}_k^N[f]\Big| \le (\lfloor Nt\rfloor - \lfloor Nr\rfloor)C\|f\|_{\mathcal{C}^1(\Theta)}/N.$$

Next, we have, by Item 3 in Lemma 3.5, $|\mathbf{W}_t^N[f] - \mathbf{W}_r^N[f]| \le |\mathbf{W}_t^N[f]| + |\mathbf{W}_r^N[f]| \le C\|f\|_{\mathcal{C}^2(\Theta)}/N$. Finally, by (3.8.11),

$$|\mathbf{R}_t^N[f] - \mathbf{R}_r^N[f]| = \Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor - 1} \mathbf{R}_k^N[f]\Big| \le (\lfloor Nt\rfloor - \lfloor Nr\rfloor)C\|f\|_{\mathcal{C}^2(\Theta)}/N^2.$$

The proof of Proposition 3.6 is complete plugging all the previous estimates in (3.8.15). □

Proposition 3.7 (Relative compactness)**.** *Assume* **A1**→**A4**. *Then, the sequence* $(\mu^N)_{N\ge1}$ *is relatively compact in* $\mathcal{D}([0,1], \mathcal{P}(\Theta))$.

Proof. The proof consists in applying [Jak86, Theorem 3.1] with $E = \mathcal{P}(\Theta)$ endowed with the weak convergence topology. Set $\mathbb{F} = \{\mathfrak{L}_f, f \in \mathcal{C}^\infty(\Theta)\}$ where

$$\mathsf{L}_f : \nu \in \mathcal{P}(\Theta) \mapsto \langle f, \nu\rangle.$$

The class of continuous functions $\mathbb{F}$ on $\mathcal{P}(\Theta)$ satisfies Conditions [Jak86, (3.1) and (3.2) in Theorem 3.1].

On the other hand, the condition [Jak86, (3.3) in Theorem 3.1] is satisfied since $\mathcal{P}(\Theta)$ is compact because Θ is compact (see e.g. [PZ20, Corollary 2.2.5] together with [Vil09, Corollary 6.13]).

It remains to verify Condition (3.4) of [Jak86, Theorem 3.1], i.e. that for all $f \in \mathcal{C}^\infty(\Theta)$, $(\langle f, \mu^N\rangle)_{N\geq 1}$ is relatively compact in $\mathcal{D}([0,1],\mathbf{R})$. To this end, we apply [Bil99, Theorem 13.2]. Condition (i) in [Bil99, Theorem 13.2] is satisfied because $|\langle f, \mu_t^N\rangle| \leq \|f\|_{\infty,\Theta}$ for all $t \in [0,1]$ and $N \geq 1$. Let us now show that Condition (ii) in [Bil99, Theorem 13.2] holds. For this purpose, we use Lemma 3.6. For $\delta, \beta > 0$ sufficiently small, it is possible to construct a subdivision $\{t_i\}_{i=0}^v$ of $[0,1]$ such that $t_0 = 0$, $t_v = 1$, $t_{i+1} - t_i = \delta + \beta$ for $i \in \{0, \dots, v-2\}$ and $\delta + \beta \leq t_v - t_{v-1} \leq 2(\delta+\beta)$. According to the terminology introduced in [Bil99, Section 12], $\{t_i\}_{i=0}^v$ is δ-sparse. Then, by Lemma 3.6, there exists $C > 0$ such that a.s. for all $\delta, \beta > 0$, all such subdivision $\{t_i\}_{i=0}^v$, $i \in \{0, \dots, v-1\}$, and $N \geq 1$,

$$\sup_{t,r\in[t_i,t_{i+1}]} |\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle| \leq C\Big(|t_{i+1} - t_i| + \frac{|t_{i+1} - t_i|}{N} + \frac{1}{N}\Big) \leq C\Big(2(\delta+\beta) + \frac{2(\delta+\beta)}{N} + \frac{1}{N}\Big).$$

Thus, one has:

$$\inf_{\beta>0} \max_i \sup_{t,r\in[t_i,t_{i+1}]} |\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle| \leq C\Big(2\delta + \frac{2\delta}{N} + \frac{1}{N}\Big).$$

Consequently, there exists $C > 0$ such that a.s. for all $\delta > 0$ small enough and $N \geq 1$,

$$w'_{\langle f,\mu^N\rangle}(\delta) := \inf_{\substack{\{t_i\}\\ \delta\text{-sparse}}} \max_i \sup_{t,r\in[t_i,t_{i+1}]} |\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle| \leq C\Big(2\delta + \frac{2\delta}{N} + \frac{1}{N}\Big).$$

This implies $\lim_{\delta\to 0} \limsup_{N\to+\infty} \mathbf{E}[w'_{\langle f,\mu^N\rangle}(\delta)] = 0$. By Markov's inequality, this proves Condition (ii) of [Bil99, Theorem 13.2]. Therefore, for all $f \in \mathcal{C}^\infty(\Theta)$, using also Prokhorov theorem, the sequence $(\langle f, \mu^N\rangle)_{N\geq 1} \subset \mathcal{D}([0,1],\mathbf{R})$ is relatively compact. In conclusion, according to [Jak86, Theorem 3.1], $(\mu^N)_{N\geq 1} \subset \mathcal{D}([0,1],\mathcal{P}(\Theta))$ is tight. □

Limit points satisfy the limit equation (3.3.4)

In this section we prove that every limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$ satisfies (3.3.4).

Lemma 3.8. *Let $\mathsf{m}, (\mathsf{m}^N)_{N\geq 1} \subset \mathcal{D}([0,1],\mathcal{P}(\Theta))$ be such that $\mathsf{m}^N \to \mathsf{m}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$. Then, for all Lipschitz continuous function $f : \Theta \to \mathbf{R}$, we have $\langle f, \mathsf{m}^N\rangle \to \langle f, \mathsf{m}\rangle$ in $\mathcal{D}([0,1],\mathbf{R})$.*

Proof. Let f be such a function. By [Bil99, p.124], $\mathsf{m}^N \to \mathsf{m}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$ iff there exist functions $\lambda_N : [0,1] \to [0,1]$ continuous, increasing onto itself such that $\sup_{t\in[0,1]} |\lambda_N(t) - t| \to_{N\to\infty} 0$ and $\sup_{t\in[0,1]} \mathsf{W}_1(\mathsf{m}^N_{\lambda_N(t)}, \mathsf{m}_t) \to_{N\to\infty} 0$. Then $\langle f, \mathsf{m}^N\rangle \to \langle f, \mathsf{m}\rangle$ in $\mathcal{D}([0,1],\mathbf{R})$ since by (3.8.13), $\sup_{t\in[0,1]} |\langle f, \mathsf{m}^N_{\lambda_N(t)}\rangle - \langle f, \mathsf{m}_t\rangle| \leq \|f\|_{\mathrm{Lip}} \sup_{t\in[0,1]} \mathsf{W}_1(\mathsf{m}^N_{\lambda_N(t)}, \mathsf{m}_t) \to_{N\to\infty} 0$. □

Proposition 3.9 (Continuity of the limit points of $\langle f, \mu^N\rangle$)**.** *Let $f \in \mathcal{C}^\infty(\Theta)$. Then, any limit point of $(\langle f, \mu^N\rangle)_{N\geq 1} \subset \mathcal{D}([0,1],\mathbf{R})$ belong a.s. to $\mathcal{C}([0,1],\mathbf{R})$.*

Proof. Fix $t \in (0,1]$. Letting $r \to t$ in (3.8.14), we obtain $|\langle f, \mu_t^N\rangle - \langle f, \mu_{t^-}^N\rangle| \leq C/N$. Therefore $\sup_{t\in(0,1]} |\langle f, \mu_t^N\rangle - \langle f, \mu_{t^-}^N\rangle| \xrightarrow{\mathscr{D}} 0$ as $N \to +\infty$. The result follows from [Bil99, Theorem 13.4]. □

Proposition 3.10 (Continuity of the limit points of μ^N)**.** *Let $\mu^* \in \mathcal{D}([0,1],\mathcal{P}(\Theta))$ be a limit point of $(\mu^N)_{N\geq 1} \subset \mathcal{D}([0,1],\mathcal{P}(\Theta))$. Then, a.s. $\mu^* \in \mathcal{C}([0,1],\mathcal{P}(\Theta))$.*

Proof. Up to extracting a subsequence, we assume that $\mu^N \xrightarrow{\mathscr{D}} \mu^*$. By Skorohod representation theorem, there exists another probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbf{P}})$ on which are defined random elements $(\hat{\mu}^N)_{N\geq 1}$ and $\hat{\mu}^*$, where,

$$\hat{\mu}^* \overset{\mathscr{D}}{=} \mu^*, \quad \text{and for all } N \geq 1,\ \hat{\mu}^N \overset{\mathscr{D}}{=} \mu^N,$$

and such that $\hat{\mathbf{P}}$-a.s., $\hat{\mu}^N \to \hat{\mu}^*$ in $\mathcal{D}([0,1], \mathcal{P}(\Theta))$ as $N \to +\infty$. Fix $f \in \mathcal{C}^\infty(\Theta)$. We have, by Lemma 3.8,

$$\hat{\mathbf{P}}-\text{a.s.}, \quad \langle f, \hat{\mu}^N\rangle \to_{N\to+\infty} \langle f, \hat{\mu}^*\rangle \ \text{ in } \ \mathcal{D}([0,1], \mathbf{R}).$$

In particular, $\langle f, \hat{\mu}^N\rangle \to_{N\to+\infty} \langle f, \hat{\mu}^*\rangle$ in distribution. By Proposition 3.9, there exists $\hat{\Omega}_f \subset \hat{\Omega}$ of $\hat{\mathbf{P}}$-mass 1 such that for all $\omega \in \hat{\Omega}_f$, $\langle f, \hat{\mu}^*(\omega)\rangle \in \mathcal{C}([0,1], \mathbf{R})$. Denote by $\mathscr{F}$ the class polynomial functions with rational coefficients. Since this class is countable, the set $\hat{\Omega}_{\mathscr{F}} := \cap_{f\in\mathscr{F}} \hat{\Omega}_f$ is of $\hat{\mathbf{P}}$-mass 1. Consider now an arbitrary $f \in \mathcal{C}(\Theta)$ and let us show that for all $\omega \in \hat{\Omega}_{\mathscr{F}}$, $\langle f, \hat{\mu}^*(\omega)\rangle \in \mathcal{C}([0,1], \mathbf{R})$. By the Stone-Weierstrass theorem, there exist $(f_n)_{n\geq 1} \subset \mathscr{F}$ such that $\|f_n - f\|_{\infty,\Theta} \to_{n\to+\infty} 0$. On $\hat{\Omega}_{\mathscr{F}}$, for all n, $t \in [0,1] \mapsto \langle f_n, \hat{\mu}^*_t\rangle$ is continuous and converges uniformly to $t \in [0,1] \mapsto \langle f, \hat{\mu}^*_t\rangle$. Hence, for all $\omega \in \hat{\Omega}_{\mathscr{F}}$ and $f \in \mathcal{C}(\Theta)$, $\langle f, \hat{\mu}^*(\omega)\rangle \in \mathcal{C}([0,1], \mathbf{R})$, i.e. for all $\omega \in \hat{\Omega}_{\mathscr{F}}$, $\hat{\mu}^*(\omega) \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$. This concludes the proof. □

Now, we introduce, for $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, the function $\mathbf{\Lambda}_t[f] : \mathcal{D}([0,1], \mathcal{P}(\Theta)) \to \mathbf{R}_+$ defined by:

$$\begin{aligned}\mathbf{\Lambda}_t[f] : \mathsf{m} \mapsto \Big| &\langle f, \mathsf{m}_t\rangle - \langle f, \mu_0\rangle \\ &+ \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &+ \eta \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P^1_0), \mathsf{m}_s\rangle \mathrm{d}s \Big|. \end{aligned} \tag{3.8.16}$$

We now study the continuity of $\mathbf{\Lambda}_t[f]$.

Lemma 3.11. *Let $(\mathsf{m}^N)_{N\geq 1} \subset \mathcal{D}([0,1], \mathcal{P}(\Theta))$ converge to $\mathsf{m} \in \mathcal{D}([0,1], \mathcal{P}(\Theta))$. Then, for all continuity point $t \in [0,1]$ of m and all $f \in \mathcal{C}^\infty(\Theta)$, we have $\mathbf{\Lambda}_t[f](\mathsf{m}^N) \to \mathbf{\Lambda}_t[f](\mathsf{m})$.*

Proof. Let $f \in \mathcal{C}^\infty(\Theta)$ and denote by $\mathcal{C}(\mathsf{m}) \subset [0,1]$ the set of continuity points of m. Let $t \in \mathcal{C}(\mathsf{m})$. From [Bil99, p. 124], we have, for all $s \in \mathcal{C}(\mathsf{m})$,

$$\mathsf{m}^N_s \to \mathsf{m}_s \ \text{ in } \ \mathcal{P}(\Theta). \tag{3.8.17}$$

Thus, $\langle f, \mathsf{m}^N_t\rangle \to_{N\to\infty} \langle f, \mathsf{m}_t\rangle$. For all $z \in \mathbf{R}^d$ and $(x,y) \in \mathsf{X}\times\mathsf{Y}$, **A1** and **A3** ensure that the functions $\theta \in \Theta \mapsto \phi(\theta, z, x) - y$ and $\theta \in \Theta \mapsto \nabla_\theta f(\theta) \cdot \nabla_\theta \phi(\theta, z, x)$ are continuous and also bounded because Θ is compact. Hence, for all $s \in [0,t] \cap \mathcal{C}(\mathsf{m})$, using (3.8.17),

$$\langle \phi(\cdot, z, x) - y, \mathsf{m}^N_s\rangle \to \langle \phi(\cdot, z, x) - y, \mathsf{m}_s\rangle \ \text{ and } \ \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z, x), \mathsf{m}^N_s\rangle \to \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z, x), \mathsf{m}_s\rangle$$

Since $[0,1]\backslash\mathcal{C}(\mathsf{m})$ is at most countable (see [Bil99, p. 124]) we have that for a.e. $(s, z', z, x, y) \in [0,t] \times \mathbf{R}^d \times \mathbf{R}^d \times \mathsf{X} \times \mathsf{Y}$,

$$\langle \phi(\cdot, z', x) - y, \mathsf{m}^N_s\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z, x), \mathsf{m}^N_s\rangle \to \langle \phi(\cdot, z', x) - y, \mathsf{m}_s\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z, x), \mathsf{m}_s\rangle.$$

Since $\phi(\theta, z', x) - y$ is bounded and by (3.8.27), there exists $C > 0$ such that for all $(s, z', z, x, y) \in [0,t] \times \mathbf{R}^d \times \mathbf{R}^d \times \mathsf{X} \times \mathsf{Y}$, $\langle |\phi(\cdot, z', x) - y|, \mathsf{m}^N_s\rangle \langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z, x)|, \mathsf{m}^N_s\rangle \leq C \|\nabla_\theta f\|_{\infty,\Theta} \mathfrak{b}(z)$. By the dominated convergence theorem, we then have:

$$\begin{aligned} &\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}^N_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}^N_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ \xrightarrow[N\to+\infty]{} &\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s. \end{aligned}$$

With the same arguments as above, one shows that $\int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P_0^1), \mathsf{m}_s^N \rangle \mathrm{d}s \to \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P_0^1), \mathsf{m}_s \rangle \mathrm{d}s$. The proof of the lemma is complete. □

Proposition 3.12 (Convergence to the limit equation)**.** *Let $\mu^* \in \mathcal{D}([0,1], \mathcal{P}(\Theta))$ be a limit point of $(\mu^N)_{N \geq 1} \subset \mathcal{D}([0,1], \mathcal{P}(\Theta))$. Then, a.s. μ^* satisfies* (3.3.4).

Proof. Up to extracting a subsequence, we can assume that $\mu^N \xrightarrow{\mathscr{D}} \mu^*$ as $N \to +\infty$. Let $f \in \mathcal{C}^\infty(\Theta)$. The pre-limit equation (3.8.9) and Lemma 3.5 imply that a.s. for all $N \geq 1$ and $t \in [0,1]$, $\mathbf{\Lambda}_t[f](\mu^N) \leq C/N + \mathbf{M}_t^N[f]$. Hence, using the last statement in Lemma 3.5, it holds for all $t \in [0,1]$,

$$\lim_{N \to \infty} \mathbf{E}[\mathbf{\Lambda}_t[f](\mu^N)] = 0.$$

In particular, $\mathbf{\Lambda}_t[f](\mu^N) \xrightarrow{\mathscr{D}} 0$. Let us now show that $\mathbf{\Lambda}_t[f](\mu^N) \xrightarrow{\mathscr{D}} \mathbf{\Lambda}_t[f](\mu^*)$. Denoting by $\mathsf{D}(\mathbf{\Lambda}_t[f])$ the set of discontinuity points of $\mathbf{\Lambda}_t[f]$, we have, from Proposition 3.10 and Lemma 3.11, for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$,

$$\mathbf{P}(\mu^* \in \mathsf{D}(\mathbf{\Lambda}_t[f])) = 0.$$

By the continuous mapping theorem, $\mathbf{\Lambda}_t[f](\mu^N) \xrightarrow{\mathscr{D}} \mathbf{\Lambda}_t[f](\mu^*)$. By uniqueness of the limit in distribution, we have that for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, a.s. $\mathbf{\Lambda}_t[f](\mu^*) = 0$. Let us now prove that a.s. for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, $\Lambda_t[f](\mu^*) = 0$.

On the one hand, for all $f \in \mathcal{C}^\infty(\Theta)$ and $\mathsf{m} \in \mathcal{D}([0,1], \mathcal{P}(\Theta))$, the function $t \mapsto \Lambda_t[f](\mathsf{m})$ is right-continuous. Since $[0,1]$ is separable, we have that for all $f \in \mathcal{C}^\infty(\Theta)$, a.s. for all $t \in [0,1]$, $\mathbf{\Lambda}_t[f](\mu^*) = 0$.

One the other hand $\mathcal{C}^\infty(\Theta)$ is separable (when endowed with the norm $\|f\|_{\mathcal{C}^\infty(\Theta)} = \sum_{k \geq 0} 2^{-k} \min(1, \sum_{|j|=k} \|\partial_j f\|_{\infty,\Theta})$) and the function $f \in \mathcal{C}^\infty(\Theta) \mapsto \mathbf{\Lambda}_t[f](\mathsf{m})$ is continuous (for fixed $t \in [0,1]$ and $\mathsf{m} \in \mathcal{D}([0,1], \mathcal{P}(\Theta))$) relatively to the topology induced by $\|f\|_{\mathcal{C}^\infty(\Theta)}$.

Hence, we obtain that a.s. for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, $\mathbf{\Lambda}_t[f](\mu^*) = 0$. The proof of the proposition is thus complete. □

Uniqueness and end of the proof of Theorem 3.1

Proposition 3.13. *There exists a unique solution to* (3.3.4) *in $\mathcal{C}([0,1], \mathcal{P}(\Theta))$.*

Proof. First of all, the fact that there is a solution to (3.3.4) is provided by Propositions 3.7, 3.10 and 3.12. The proof of the fact that there is a unique solution to (3.3.4) relies on the same arguments as those used in the proof of [DGMN22, Proposition 2.14].

For $\mu \in \mathcal{P}(\mathbf{R}^{d+1})$, we introduce $\boldsymbol{v}[\mu] : \mathbf{R}^{d+1} \to \mathbf{R}^{d+1}$ defined, for $\theta = (m, \rho) \in \mathbf{R}^{d+1}$, by

$$\boldsymbol{v}[\mu](\theta) = -\eta \int_{\mathsf{X} \times \mathsf{Y}} \langle \phi(\cdot, \cdot, x) - y, \mu \otimes \gamma \rangle \langle \nabla_\theta \phi(\theta, \cdot, x), \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y) - \eta \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1). \tag{3.8.18}$$

In addition, if $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ is solution to (3.3.4), it satisfies also (3.3.4) with test functions $f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$. Then, adopting the terminology of [San15, Section 4.1.2], any solution $\bar{\mu}$ to (3.3.4) is a *weak solution*[1] on $[0,T]$ of the measure-valued equation

$$\begin{cases} \partial_t \bar{\mu}_t = \mathrm{div}(\boldsymbol{v}[\bar{\mu}_t] \bar{\mu}_t) \\ \bar{\mu}_0 = \mu_0. \end{cases} \tag{3.8.19}$$

Let us now prove that:

[1] We mention that according to [San15, Proposition 4.2], the two notions of solutions of (3.8.19) (namely the weak solution and the *distributional* solution) are equivalent.

1. There exists $C > 0$ such that for all $\mu \in \mathcal{P}(\mathbf{R}^{d+1})$ and $\theta \in \mathbf{R}^{d+1}$,
$$|\mathrm{J}_\theta \boldsymbol{v}[\mu](\theta)| \leq C.$$
2. There exists $C > 0$ such that for all $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ solution to (3.3.4), $0 \leq s, t \leq 1$, and $\theta \in \mathbf{R}^{d+1}$,
$$|\boldsymbol{v}[\bar{\mu}_t](\theta) - \boldsymbol{v}[\bar{\mu}_s](\theta)| \leq C|t-s|.$$
3. There exists $L' > 0$ such that for all $\mu, \nu \in \mathcal{P}_1(\mathbf{R}^{d+1})$,
$$\sup_{\theta \in \mathbf{R}^d} |\boldsymbol{v}[\mu](\theta) - \boldsymbol{v}[\nu](\theta)| \leq L' \mathsf{W}_1(\mu, \nu).$$

Before proving the three items above, we quickly conclude the proof of the proposition. Items 1 and 2 above imply that $v(t,\theta) = \boldsymbol{v}[\bar{\mu}_t](\theta)$ is globally Lipschitz continuous over $[0,1] \times \mathbf{R}^{d+1}$ when $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ is a solution to (3.3.4). Since $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta)) \subset \mathcal{C}([0,1], \mathcal{P}(\mathbf{R}^{d+1}))$, this allows to use the representation theorem [Vil03, Theorem 5.34] for the solution of (3.8.19) in $\mathcal{C}([0,1], \mathcal{P}(\mathbf{R}^{d+1}))$, i.e. it holds:
$$\forall t \in [0,1],\ \bar{\mu}_t = \phi_t \# \mu_0, \tag{3.8.20}$$
where ϕ_t is the flow generated by the vector field $\boldsymbol{v}[\bar{\mu}_t](\theta)$ over $\mathbf{R}^{d+1}$. Equation (3.8.20) and the fact that $\mathcal{C}([0,1], \mathcal{P}(\Theta)) \subset \mathcal{C}([0,1], \mathcal{P}_1(\mathbf{R}^{d+1}))$ together with Item 3 above and the same arguments as those used in the proof of [DGMN22, Proposition 2.14] (which we recall is based estimates in Wasserstein distances between two solutions of (3.3.4) derived in [PR16]), one deduces that there is a unique solution to (3.3.4).

Let us prove Item 1. Recall $g(\rho) = \ln(1 + e^\rho)$. The functions
$$\rho \mapsto g''(\rho) g(\rho),\ \rho \mapsto g'(\rho),\ \rho \mapsto \frac{g'(\rho)}{g(\rho)},\ \text{ and } \rho \mapsto \frac{g''(\rho)}{g(\rho)}$$
are bounded on $\mathbf{R}$. Thus, in view of (3.2.4), $\|\mathrm{Hess}_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\theta | P^1_0)\|_{\infty, \mathbf{R}^{d+1}} < +\infty$. On the other hand, by **A1** and **A3**, for $x \in \mathsf{X}$, $z \in \mathbf{R}^d$, $\theta \in \Theta \mapsto \phi(\theta, z, x)$ is smooth and there exists $C > 0$, for all $x \in \mathsf{X}$, $\theta \in \mathbf{R}^{d+1}$, $z \in \mathbf{R}^d$:
$$|\mathrm{Hess}_\theta \phi(\theta, z, x)| \leq C(\mathfrak{b}(z)^2 + \mathfrak{b}(z)).$$
This bound allows us to differentiate under the integral signs in (3.8.18) and proves that $|\mathrm{J}_\theta \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu \otimes \gamma \rangle \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y)| \leq C$, where $C > 0$ is independent of $\mu \in \mathcal{P}(\Theta)$ and $\theta \in \Theta$. The proof of Item 1 is complete.

Let us prove Item 2. Let $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ be a solution to (3.3.4), $0 \leq s \leq t \leq 1$, and $\theta \in \mathbf{R}^{d+1}$. We have
$$\boldsymbol{v}[\bar{\mu}_t](\theta) - \boldsymbol{v}[\bar{\mu}_s](\theta) = -\eta \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), (\bar{\mu}_t - \bar{\mu}_s) \otimes \gamma \rangle \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y). \tag{3.8.21}$$
Let $z \in \mathbf{R}^d$ and $x \in \mathsf{X}$. By **A1** and **A3**, $\phi(\cdot, z, x) \in \mathcal{C}^\infty(\Theta)$. Therefore, by (3.3.4),
$$\begin{aligned}\langle \phi(\cdot,z,x), \bar{\mu}_t - \bar{\mu}_s \rangle = &-\eta \int_s^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x') - y, \bar{\mu}_r \otimes \gamma \rangle \langle \nabla_\theta \phi(\cdot,z,x) \cdot \nabla_\theta \phi(\cdot,\cdot,x'), \bar{\mu}_r \otimes \gamma \rangle \pi(\mathrm{d}x', \mathrm{d}y) \mathrm{d}r \\ &- \eta \int_s^t \langle \nabla_\theta \phi(\cdot,z,x) \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P^1_0), \bar{\mu}_r \rangle \mathrm{d}r\end{aligned}$$
We have $\|\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\theta | P^1_0)\|_{\infty,\Theta} < +\infty$. Using also (3.8.27) and the fact that $\mathsf{X} \times \mathsf{Y}$ is a compact (see **A2**), it holds:
$$|\langle \phi(\cdot,z,x), \bar{\mu}_t - \bar{\mu}_s \rangle| \leq C\mathfrak{b}(z)|t-s|.$$

Hence, for all $x' \in \mathsf{X}$,

$$|\langle \phi(\cdot,\cdot,x'), (\bar\mu_t - \bar\mu_s)\otimes\gamma\rangle| \le \langle |\langle \phi(\cdot,\cdot,x'), \bar\mu_t - \bar\mu_s\rangle|, \gamma\rangle \le C|t-s|.$$

Thus, by (3.8.21) and (3.8.28), $|\boldsymbol{v}[\bar\mu_t](\theta) - \boldsymbol{v}[\bar\mu_s](\theta)| \le C|t-s|$. This ends the proof of Item 2.

Let us now prove Item 3. Fix $\mu,\nu \in \mathcal{P}_1(\mathbf{R}^{d+1})$ and $\theta \in \mathbf{R}^{d+1}$. We have

$$\boldsymbol{v}[\mu](\theta) - \boldsymbol{v}[\nu](\theta) = -\eta \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), (\mu-\nu)\otimes\gamma\rangle \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y) \tag{3.8.22}$$

For all $x \in \mathsf{X}$, using (3.8.13) and (3.8.27), it holds:

$$\begin{aligned}|\langle \phi(\cdot,\cdot,x), (\mu-\nu)\otimes\gamma\rangle| &\le \int_{\mathbf{R}^d} |\langle \phi(\cdot,z,x),\mu\rangle - \langle \phi(\cdot,z,x),\nu\rangle|\gamma(z)\mathrm{d}z \\ &\le C\int_{\mathbf{R}^d} \mathsf{W}_1(\mu,\nu)\mathfrak{b}(z)\gamma(z)\mathrm{d}z \le C\mathsf{W}_1(\mu,\nu).\end{aligned}$$

Finally, using in addition (3.8.28) and (3.8.22), we deduce Item 3. This ends the proof of the proposition. □

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Recall Lemma 3.1 ensures that a.s. $(\mu^N)_{N\ge1} \subset \mathcal{D}([0,1],\mathcal{P}(\Theta))$. By Proposition 3.7, this sequence is relatively compact. Let $\mu^* \in \mathcal{D}([0,1],\mathcal{P}(\Theta))$ be a limit point. Along some subsequence N', it holds:

$$\mu^{N'} \xrightarrow{\mathscr{D}} \mu^*.$$

In addition, a.s. $\mu^* \in \mathcal{C}([0,1],\mathcal{P}(\Theta))$ (by Proposition 3.10) and μ^* satisfies (3.3.4) (by Proposition 3.12). By Proposition 3.13, (3.3.4) admits a unique solution $\bar\mu \in \mathcal{C}([0,1],\mathcal{P}(\Theta))$. Hence, a.s. $\mu^* = \bar\mu$. Therefore,

$$\mu^{N'} \xrightarrow{\mathscr{D}} \bar\mu.$$

Since the sequence $(\mu^N)_{N\ge1}$ admits a unique limit point, the whole sequence converges in distribution to $\bar\mu$. The convergence also holds in probability since $\bar\mu$ is deterministic. The proof of Theorem 3.1 is complete. □

3.8.3 Proof of Lemma 3.1

In this section we prove Lemma 3.1. We start with the following simple result.

Lemma 3.14. *Let $T>0$, $N\ge1$, and $c_1>0$. Consider a sequence $(u_k)_{0\le k\le \lfloor NT\rfloor} \subset \mathbf{R}_+$ for which there exists v_0 such that $u_0 \le v_0$ and for all $1\le k\le \lfloor NT\rfloor$, $u_k \le c_1(1+\frac{1}{N}\sum_{\ell=0}^{k-1} u_\ell)$. Then, for all $0\le k\le \lfloor NT\rfloor$, $u_k \le v_0 e^{c_1 T}$.*

Proof. Define $v_k = c_1(1+\frac{1}{N}\sum_{\ell=0}^{k-1} v_\ell)$. For all $0\le k\le \lfloor NT\rfloor$, $u_k \le v_k$ and $v_k = v_{k-1}(1+c_1/N)$. Hence $v_k = v_0\big(1+c_1/N\big)^k \le v_0\big(1+c_1/N\big)^{\lfloor NT\rfloor} \le v^0 e^{c_1 T}$. This ends the proof of the Lemma. □

Proof of Lemma 3.1. Since $\rho\mapsto g'(\rho)$ and $\rho\mapsto g'(\rho)/g(\rho)$ are bounded continuous functions over $\mathbf{R}$, and since $|g(\rho)| \le C(1+|\rho|)$, according to (3.2.4), there exists $c>0$, for all $\theta\in\mathbf{R}^{d+1}$,

$$|\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)| \le c(1+|\theta|). \tag{3.8.23}$$

All along the proof, $C > 0$ is a constant independent of $N \geq 1$, $T > 0$, $i \in \{1, \dots, N\}$, $1 \leq k \leq \lfloor NT \rfloor$, $(x, y) \in \mathsf{X} \times \mathsf{Y}$, $\theta \in \mathbf{R}^{d+1}$, and $z \in \mathbf{R}^d$, which can change from one occurence to another. It holds:

$$|\theta_k^i| \leq |\theta_0^i| + \sum_{\ell=0}^{k-1} |\theta_{\ell+1}^i - \theta_\ell^i|. \tag{3.8.24}$$

Using (3.2.5), we have, for $0 \leq \ell \leq k-1$,

$$\begin{aligned} |\theta_{\ell+1}^i - \theta_\ell^i| \leq \frac{\eta}{N^2} \sum_{j=1, j\neq i}^{N} \Big| (\langle \phi(\theta_\ell^j, \cdot, x_\ell), \gamma \rangle - y_\ell) \langle \nabla_\theta \phi(\theta_\ell^i, \cdot, x_\ell), \gamma \rangle \Big| \\ + \frac{\eta}{N^2} \Big| \Big\langle (\phi(\theta_\ell^i, \cdot, x_\ell) - y_\ell) \nabla_\theta \phi(\theta_\ell^i, \cdot, x_\ell), \gamma \Big\rangle \Big| + \frac{\eta}{N} |\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta_\ell^i} | P_0^1)|. \end{aligned} \tag{3.8.25}$$

For all $\theta \in \mathbf{R}^{d+1}$, $z \in \mathbf{R}^d$, $(x, y) \in \mathsf{X} \times \mathsf{Y}$, we have, by **A2** and **A3**, since $\phi(\theta, z, x) = s(\Psi_\theta(z), x)$,

$$|\phi(\theta, z, x) - y| \leq C. \tag{3.8.26}$$

Moreover, we have $\nabla_\theta \phi(\theta, z, x) = \nabla_1 s(\Psi_\theta(z), x) \mathrm{J}_\theta \Psi_\theta(z)$ (here $\nabla_1 s$ refers to the gradient of s w.r.t. its first variable). By **A3**, $|\nabla_1 s(\Psi_\theta(z), x)| \leq C$ and, hence, denoting by J_θ the Jacobian w.r.t. θ, using (3.3.3),

$$|\nabla_\theta \phi(\theta, z, x)| \leq C |\mathrm{J}_\theta \Psi_\theta(z)| \leq C \mathfrak{b}(z). \tag{3.8.27}$$

Therefore, by (3.3.3),

$$\langle |\nabla_\theta \phi(\theta, \cdot, x)|, \gamma \rangle \leq C. \tag{3.8.28}$$

Hence, we obtain, using (3.8.25) and (3.8.23),

$$|\theta_{\ell+1}^i - \theta_\ell^i| \leq \frac{\eta}{N^2} \sum_{j=1, j\neq i}^{N} C + \frac{\eta}{N^2} C + \frac{c\eta}{N} (1 + |\theta_\ell^i|) \leq \frac{C}{N} (1 + |\theta_\ell^i|). \tag{3.8.29}$$

Using **A4**, there exists $K_0 > 0$ such that a.s. for all i, $|\theta_0^i| \leq K_0$. Then, from (3.8.24) and (3.8.29), for $1 \leq k \leq \lfloor NT \rfloor$, it holds:

$$|\theta_k^i| \leq K_0 + \frac{C}{N} \sum_{\ell=0}^{k-1} (1 + |\theta_\ell^i|) \leq K_0 + CT + \frac{C}{N} \sum_{\ell=0}^{k-1} |\theta_\ell^i| \leq C_{0,T} (1 + \frac{1}{N} \sum_{\ell=0}^{k-1} |\theta_\ell^i|),$$

with $C_{0,T} = \max(K_0 + CT, C) \leq K_0 + C(1+T)$. Then, by Lemma 3.14 and **A4**, we have that for all $N \geq 1$, $i \in \{1, \dots, N\}$ and $0 \leq k \leq \lfloor NT \rfloor$, $|\theta_k^i| \leq K_0 e^{[K_0 + C(1+T)]T}$. The proof of Lemma 3.1 is thus complete. □

3.9 Proof of Theorem 3.2

In this section, we assume **A1**→**A5** (where in **A2**, when $k \geq 1$, $\mathcal{F}_k^N$ is now the one defined in (3.4.1)) and the θ_k^i's (resp. μ^N) are those defined by (3.2.7) for $i \in \{1, \dots, N\}$ and $k \geq 0$ (resp. by (3.4.2) for $N \geq 1$).

3.9.1 Preliminary analysis and pre-limit equation

Notation and weighted Sobolev embeddings

For $J \in \mathbf{N}$ and $\beta \geq 0$, let $\mathcal{H}^{J,\beta}(\mathbf{R}^{d+1})$ be the closure of the set $\mathcal{C}_c^\infty(\mathbf{R}^{d+1})$ for the norm

$$\|f\|_{\mathcal{H}^{J,\beta}} := \Big(\sum_{|k| \leq J} \int_{\mathbf{R}^{d+1}} \frac{|\partial_k f(\theta)|^2}{1 + |\theta|^{2\beta}} \mathrm{d}\theta \Big)^{1/2}.$$

The space $\mathcal{H}^{J,\beta}(\mathbf{R}^{d+1})$ is a separable Hilbert space and we denote its dual space by $\mathcal{H}^{-J,\beta}(\mathbf{R}^{d+1})$ (see e.g. [FM97a, JM98a]). The associated scalar product on $\mathcal{H}^{J,\beta}(\mathbf{R}^{d+1})$ will be denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}^{J,\beta}}$. For $\Phi \in \mathcal{H}^{-J,\beta}(\mathbf{R}^{d+1})$, we use the notation

$$\langle f, \Phi \rangle_{J,\beta} = \Phi[f], \ f \in \mathcal{H}^{J,\beta}(\mathbf{R}^{d+1}).$$

For ease of notation, and if no confusion is possible, we simply denote $\langle f, \Phi \rangle_{J,\beta}$ by $\langle f, \Phi \rangle$. The set $\mathcal{C}_0^{J,\beta}(\mathbf{R}^{d+1})$ (resp. $\mathcal{C}^{J,\beta}(\mathbf{R}^{d+1})$) is defined as the space of functions $f : \mathbf{R}^{d+1} \to \mathbf{R}$ with continuous partial derivatives up to order $J \in \mathbf{N}$ such that

$$\text{for all } |k| \leq J, \ \lim_{|\theta| \to \infty} \frac{|\partial_k f(\theta)|}{1 + |\theta|^\beta} = 0 \ \ (\text{resp. } \sum_{|k| \leq J} \sup_{\theta \in \mathbf{R}^{d+1}} \frac{|\partial_k f(\theta)|}{1 + |\theta|^\beta} < +\infty).$$

The spaces $\mathcal{C}^{J,\beta}(\mathbf{R}^{d+1})$ and $\mathcal{C}_0^{J,\beta}(\mathbf{R}^{d+1})$ is endowed with the norm

$$\|f\|_{\mathcal{C}^{J,\beta}} := \sum_{|k| \leq J} \sup_{\theta \in \mathbf{R}^{d+1}} \frac{|\partial_k f(\theta)|}{1 + |\theta|^\beta}.$$

We note that

$$\theta \in \mathbf{R}^{d+1} \mapsto (1 - \chi(\theta))|\theta|^\alpha \in \mathcal{H}^{J,\beta}(\mathbf{R}^{d+1}) \text{ if } \beta - \alpha > (d+1)/2, \tag{3.9.1}$$

where $\chi \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$ equals 1 near 0. We recall that from [FM97a, Section 2], for $m' > (d+1)/2$ and $\alpha, j \geq 0$, $\mathcal{H}^{m'+j,\alpha}(\mathbf{R}^{d+1}) \hookrightarrow \mathcal{C}_0^{j,\alpha}(\mathbf{R}^{d+1})$. In the following, we consider $\gamma_0, \gamma_1 \in \mathbf{R}$ and $L_0 \in \mathbf{N}$ such that

$$\gamma_1 > \gamma_0 > \frac{d+1}{2} + 1 \text{ and } L_0 > \frac{d+1}{2} + 1.$$

We finally recall the following standard result.

Proposition 3.15. *Let $q > p \geq 1$ and $C > 0$. The set $\mathscr{K}_C^q := \{\mu \in \mathcal{P}_p(\mathbf{R}^{d+1}), \int_{\mathbf{R}^{d+1}} |x|^q \mu(\mathrm{d}x) \leq C\}$ is compact.*

Bound on the moments of the θ_k^i's

We have the following uniform bound in $N \geq 1$ on the moments of the sequence $\{\theta_k^i, i \in \{1, \ldots, N\}\}_{k=0,\ldots,\lfloor NT \rfloor}$ defined by (3.2.7).

Lemma 3.16. *Assume* **A1**$\to$ **A5**. *For all $T > 0$ and $p \geq 1$, there exists $C > 0$ such that for all $N \geq 1$, $i \in \{1, \ldots, N\}$ and $0 \leq k \leq \lfloor NT \rfloor$,*

$$\mathbf{E}[|\theta_k^i|^p] \leq C.$$

Proof. Let $p \geq 1$. By **A4**, $\mathbf{E}[|\theta_0^i|^p] \leq C_p$ for all $i \in \{1, \ldots, N\}$. Let $T > 0$. In the following $C > 0$ is a constant independent of $N \geq 1$, $i \in \{1, \ldots, N\}$, and $1 \leq k \leq \lfloor NT \rfloor$. Using (3.2.7), the fact that ϕ is bounded, Y is bounded, and (3.8.27), we have, for $0 \leq n \leq k-1$,

$$\begin{aligned} |\theta_{n+1}^i - \theta_n^i| &\leq \frac{C}{N^2 B} \sum_{j=1}^N \sum_{\ell=1}^B \mathfrak{b}(\mathsf{Z}_n^{i,\ell}) + \frac{C}{N} |\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta_n^i} | P_0^1)| \\ &\leq \frac{C}{NB} \sum_{\ell=1}^B (1 + \mathfrak{b}(\mathsf{Z}_n^{i,\ell})) + \frac{C}{N}(1 + |\theta_n^i|), \end{aligned} \tag{3.9.2}$$

where we have also used (3.8.23) for the last inequality. Let us recall the following convexity inequality: for $m, p \geq 1$ and $x_1, \ldots, x_p \in \mathbf{R}_+$,

$$\Big(\sum_{n=1}^{m} x_n\Big)^p \leq m^{p-1} \sum_{n=1}^{m} x_n^p. \tag{3.9.3}$$

Using (3.8.24), **A1** with $q = p$, and the fact that $1 \leq k \leq \lfloor NT \rfloor$, one has setting $u_k = \mathbf{E}[|\theta_k^i|^p]$, $u_k \leq C(1 + \frac{1}{N}\sum_{n=0}^{k-1} u_n)$. The result then follows from Lemma 3.14. □

Pre-limit equation

In this section, we derive the pre-limit equation for μ^N defined by (3.4.2). For simplicity we will keep the same notations as those introduced in Section 3.8.1, though these objects will now be defined with θ_k^i set by (3.2.7), and on $\mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$, for all integer $k \geq 0$, and all time $t \geq 0$. Let $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$. Then, set for $k \geq 0$,

$$\begin{aligned}\mathbf{D}_k^N[f] = &-\frac{\eta}{N^3}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\int_{\mathsf{X}\times\mathsf{Y}} \big(\langle\phi(\theta_k^j,\cdot,x),\gamma\rangle - y\big)\big\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\ &-\frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y).\end{aligned}$$

Note that $\mathbf{D}_k^N$ above is the one defined in (3.8.3) but now on $\mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$ and with θ_k^i defined by (3.2.7). For $k \geq 0$, we set

$$\mathbf{M}_k^N[f] = -\frac{\eta}{N^3 B}\sum_{i,j=1}^{N}\sum_{\ell=1}^{B}(\phi(\theta_k^j,\mathsf{Z}_k^{j,\ell},x_k)-y_k)\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\mathsf{Z}_k^{i,\ell},x_k) - \mathbf{D}_k^N[f]. \tag{3.9.4}$$

By Lemma 3.16 together with (3.8.26) and (3.8.27), $\mathbf{M}_k^N[f]$ is integrable. Also, using **A5** and the fact that θ_k^j is $\mathcal{F}_k^N$-measurable (see (3.4.1)),

$$\mathbf{E}[\mathbf{M}_k^N[f]|\mathcal{F}_k^N] = 0.$$

Set $\mathbf{M}_t^N[f] = \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{M}_k^N[f]$, $t \geq 0$. We now extend the definition of $\mathbf{W}_t^N[f]$ and $\mathbf{R}_k^N[f]$ to any time $t \geq 0$, $k \geq 0$, and $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$, and with θ_k^i set by (3.2.7). We then set

$$\mathbf{R}_t^N[f] = \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{R}_k^N[f],\ t \geq 0.$$

With the same algebraic computations as those made in Section 3.8.1, one obtains the following pre-limit equation: for $N \geq 1$, $t \geq 0$, and $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$,

$$\begin{aligned}\langle f,\mu_t^N\rangle - \langle f,\mu_0^N\rangle = &-\eta\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\mu_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\ &-\eta\int_0^t\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\mu_s^N\big\rangle\mathrm{d}s\\ &+\frac{\eta}{N}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\ &-\frac{\eta}{N}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\ &+\mathbf{M}_t^N[f]+\mathbf{W}_t^N[f]+\mathbf{R}_t^N[f].\end{aligned} \tag{3.9.5}$$

We will now show that the sequence $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+,\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$.

3.9.2 Relative compactness and convergence to the limit equation

Relative compactness in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$

In this section we prove the following result.

Proposition 3.17. *Assume* **A1**→**A5**. *Recall* $\gamma_0 > \frac{d+1}{2} + 1$. *Then, the sequence* $(\mu^N)_{N\geq 1}$ *is relatively compact in* $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$.

We start with the following lemma.

Lemma 3.18. *Assume* **A1**→ **A5**. *Then,* $\forall T > 0$ *and* $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$,

$$\sup_{N\geq 1} \mathbf{E}\Big[\sup_{t\in[0,T]} \langle f, \mu_t^N\rangle^2 \Big] < +\infty.$$

Proof. Let $T > 0$. In what follows, $C > 0$ is a constant independent of $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$, $(s,t) \in [0,T]^2$, and $z \in \mathbf{R}^d$ which can change from one occurence to another. We have by **A4**, $\mathbf{E}[\langle f, \mu_0^N\rangle^2] \leq C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}$. By (3.9.5) and (3.8.26), it holds:

$$\begin{aligned}
\sup_{t\in[0,T]} \langle f, \mu_t^N\rangle^2 \leq C\Big[\|f\|^2_{\mathcal{C}^{2,\gamma_1}} &+ \int_0^T \int_{\mathsf{X}\times\mathsf{Y}} \Big|\Big\langle \langle \big|\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)\big|, \gamma\rangle, \mu_s^N \Big\rangle\Big|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\int_0^T \big|\big\langle \big|\nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P_0^1)\big|, \mu_s^N\big\rangle\big|^2 \mathrm{d}s \\
&+ \frac{1}{N^2}\int_0^T \int_{\mathsf{X}\times\mathsf{Y}} \Big|\Big\langle \langle \big|\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)\big|, \gamma\rangle, \mu_s^N \Big\rangle\Big|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&+ \sup_{t\in[0,T]} |\mathbf{M}_t^N[f]|^2 + \sup_{t\in[0,T]} |\mathbf{W}_t^N[f]|^2 + \sup_{t\in[0,T]} |\mathbf{R}_t^N[f]|^2 \Big].
\end{aligned} \tag{3.9.6}$$

We have using (3.8.27), for $s \in [0,T]$ and $z \in \mathbf{R}^d$,

$$|\nabla_\theta f(\theta^i_{\lfloor Ns\rfloor}) \cdot \nabla_\theta \phi(\theta^i_{\lfloor Ns\rfloor}, z, x)| \leq C\|f\|_{\mathcal{C}^{1,\gamma_1}} \mathfrak{b}(z)(1 + |\theta^i_{\lfloor Ns\rfloor}|^{\gamma_1}). \tag{3.9.7}$$

Thus, using Lemma 3.16,

$$\mathbf{E}\big[\big\langle \langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)|, \gamma\rangle, \mu_s^N\big\rangle^2\big] \leq C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}. \tag{3.9.8}$$

Using (3.8.23), for $s \in [0,T]$, it holds:

$$\big|\nabla_\theta f(\theta^i_{\lfloor Ns\rfloor}) \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_{\lfloor Ns\rfloor}} | P_0^1)\big| \leq C\|f\|_{\mathcal{C}^{1,\gamma_1}}(1 + |\theta^i_{\lfloor Ns\rfloor}|^{\gamma_1+1}). \tag{3.9.9}$$

Thus, using Lemma 3.16,

$$\mathbf{E}\big[\big|\big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P_0^1), \mu_s^N\big\rangle\big|^2\big] \leq C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}. \tag{3.9.10}$$

On the other hand, we have using (3.9.3):

$$\sup_{t\in[0,T]} |\mathbf{M}_t^N[f]|^2 \leq \lfloor NT\rfloor \sum_{k=0}^{\lfloor NT\rfloor - 1} |\mathbf{M}_k^N[f]|^2. \tag{3.9.11}$$

Recall (3.9.4). By (3.8.3), (3.9.3), **A1**, and (3.9.7), it holds:

$$|\mathbf{D}_k^N[f]|^2 \leq C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}\Big[\frac{1}{N^4}\sum_{i\neq j=1}^N (1 + |\theta_k^i|^{2\gamma_1}) + \frac{1}{N^4}(1 + \langle |\cdot|^{2\gamma_1}, \nu_k^N\rangle)\Big] \leq \frac{C}{N^2}\|f\|^2_{\mathcal{C}^{1,\gamma_1}}(1 + |\theta_k^i|^{2\gamma_1})$$

and

$$|\mathbf{M}_k^N[f]|^2 \le \frac{C}{N^4 B}\sum_{i,j=1}^{N}\sum_{\ell=1}^{B}\|f\|^2_{\mathcal{C}^{1,\gamma_1}}|\mathfrak{b}(\mathsf{Z}_k^{i,\ell})|^2(1+|\theta^i_{\lfloor Ns\rfloor}|^{2\gamma_1})+|\mathbf{D}_k^N[f]|^2.$$

By Lemma 3.16 and **A1**, one deduces that

$$\mathbf{E}[|\mathbf{M}_k^N[f]|^2] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N^2. \tag{3.9.12}$$

Going back to (3.9.11), we then have $\mathbf{E}[\sup_{t\in[0,T]}|\mathbf{M}_t^N[f]|^2] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}$. Using the same arguments as those used so far, one also deduces that for $t\in[0,T]$

$$\begin{aligned}\sup_{t\in[0,T]}|\mathbf{W}_t^N[f]|^2 &\le \frac{C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}}{N^2}\sup_{t\in[0,T]}(1+\langle|\cdot|^{\gamma_1+1},\nu^N_{\lfloor Nt\rfloor}\rangle)^2\\ &= \frac{C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}}{N^2}\max_{0\le k\le\lfloor NT\rfloor}(1+\langle|\cdot|^{\gamma_1+1},\nu^N_k\rangle)^2\\ &\le \frac{C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}}{N^2}\sum_{k=0}^{\lfloor NT\rfloor}(1+\langle|\cdot|^{\gamma_1+1},\nu^N_k\rangle)^2.\end{aligned}$$

and thus

$$\mathbf{E}\Big[\sup_{t\in[0,T]}|\mathbf{W}_t^N[f]|^2\Big] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N. \tag{3.9.13}$$

Let us finally deal with the term involving $\mathbf{R}_t^N[f]$. One has using (3.9.3):

$$\sup_{t\in[0,T]}|\mathbf{R}_t^N[f]|^2 \le \lfloor NT\rfloor\sum_{k=0}^{\lfloor NT\rfloor-1}|\mathbf{R}_k[f]|^2.$$

For $0\le k\le\lfloor NT\rfloor-1$, we have

$$\begin{aligned}|\mathbf{R}_k^N[f]|^2 &\le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}}{N}\sum_{i=1}^{N}|\theta^i_{k+1}-\theta^i_k|^4(1+|\hat\theta^i_k|^{\gamma_1})^2\\ &\le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}}{N}\sum_{i=1}^{N}|\theta^i_{k+1}-\theta^i_k|^4(1+|\theta^i_{k+1}|^{2\gamma_1}+|\theta^i_k|^{2\gamma_1}).\end{aligned}$$

Using (3.9.2),

$$|\theta^i_{k+1}-\theta^i_k|^4 \le C\Big[\frac{1}{N^4}+\frac{|\theta^i_k|^4}{N^4}+\frac{1}{N^4B}\sum_{\ell=1}^{B}|\mathfrak{b}(\mathsf{Z}_k^{i,\ell})|^4\Big].$$

By Lemma 3.16 and **A1**, it then holds $\mathbf{E}[|\theta^i_{k+1}-\theta^i_k|^4(1+|\theta^i_{k+1}|^{2\gamma_1}+|\theta^i_k|^{2\gamma_1})] \le C/N^4$. Hence, one deduces that

$$\mathbf{E}[\sup_{t\in[0,T]}|\mathbf{R}_t^N[f]|^2] \le C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}/N^2. \tag{3.9.14}$$

This ends the proof of Lemma 3.18. □

Lemma 3.19 (Compact containment for $(\mu^N)_{N\ge1}$)**.** *Assume* **A1**→**A5**. *Let* $0<\epsilon<\gamma_1-\gamma_0$. *For every* $T>0$,

$$\sup_{N\ge1}\mathbf{E}\Big[\sup_{t\in[0,T]}\int_{\mathbf{R}^{d+1}}|x|^{\gamma_0+\epsilon}\mu_t^N(\mathrm{d}x)\Big] < +\infty. \tag{3.9.15}$$

Proof. Apply Lemma 3.18 with $f : \theta \mapsto (1-\chi)|\theta|^{\gamma_0+\epsilon} \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$. □

Lemma 3.20. *Assume* **A1**→**A5**. *Let $T > 0$ and $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$. Then, there exists $C > 0$ such that for all $\delta > 0$ and $0 \le r < t \le T$ such that $t - r \le \delta$, one has for all $N \ge 1$,*

$$\mathbf{E}\big[|\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle|^2\big] \le C(\delta^2 + \delta/N + 1/N).$$

Proof. Using (3.9.5), Jensen's inequality, (3.8.26), (3.9.8), and (3.9.10), one has for $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$,

$$\begin{aligned}\mathbf{E}\big[|\langle f, \mu_t^N\rangle - \langle f, \mu_r^N\rangle|^2\big] \le C\Big[(t-r)^2(1+1/N^2)\|f\|^2_{\mathcal{C}^{1,\gamma_1}} + \mathbf{E}\big[\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{M}_k^N[f]\Big|^2\big] \\ + \mathbf{E}\big[\,\big|\mathbf{W}_t^N[f] - \mathbf{W}_r^N[f]\big|^2\,\big] + \mathbf{E}\big[\,\big|\mathbf{R}_t^N[f] - \mathbf{R}_r^N[f]\big|^2\,\big]. \qquad (3.9.16)\end{aligned}$$

We also have with the same arguments as those used just before (3.8.12)

$$\mathbf{E}\big[\Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{M}_k^N[f]\Big|^2\big] = \sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{E}[|\mathbf{M}_k^N[f]|^2].$$

Using in addition (3.9.12), one has $\mathbf{E}\big[\big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{M}_k^N[f]\big|^2\big] \le C(N\delta+1)\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N^2$. Note that with this argument, we also deduce that

$$\mathbf{E}[|\mathbf{M}_t^N[f]|^2] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N. \qquad (3.9.17)$$

On the other hand, by (3.9.13) and (3.9.14), one has

$$\mathbf{E}\big[\,\big|\mathbf{W}_t^N[f] - \mathbf{W}_r^N[f]\big|^2\,\big] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N \text{ and } \mathbf{E}\big[\,\big|\mathbf{R}_t^N[f] - \mathbf{R}_r^N[f]\big|^2\,\big] \le C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}/N^2.$$

One then plugs all the previous estimates in (3.9.16) to deduce the result of Lemma 3.20. □

We are now in position to prove Proposition 3.17.

Proof of Proposition 3.17. The proof consists in applying [Jak86, Theorem 4.6] with $E = \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1})$ and $\mathbb{F} = \{\mathsf{H}_f, f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})\}$ where

$$\mathsf{H}_f : \nu \in \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}) \mapsto \langle f, \nu\rangle.$$

The set $\mathbb{F}$ on $\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1})$ satisfies Conditions [Jak86, (3.1) and (3.2) in Theorem 3.1]. Condition (4.8) there follows from Proposition 3.15, Lemma 3.19, and Markov's inequality. Let us now show [Jak86, Condition (4.9)] is verified, i.e. that for all $f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$, the family $(\langle f, \mu^N\rangle)_{N\ge1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathbf{R})$. To do this, it suffices to use Lemma 3.20 and [DGMN22, Proposition A.1] (with $\mathcal{H}_1 = \mathcal{H}_2 = \mathbf{R}$ there). In conclusion, according to [Jak86, Theorem 4.6], the sequence $(\mu^N)_{N\ge1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ is relatively compact. □

Limit points satisfy the limit equation (3.4.4)

For $f \in \mathcal{C}^{1,\gamma_0-1}(\mathbf{R}^{d+1})$ and $t \ge 0$, we introduce for $\mathsf{m} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$,

$$\begin{aligned}\mathbf{\Phi}_t[f] : \mathsf{m} \mapsto \Big|\langle f, \mathsf{m}_t\rangle - \langle f, \mu_0\rangle \\ + \eta\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s \otimes \gamma\rangle\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s\otimes\gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ + \eta\int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot|P_0^1), \mathsf{m}_s\big\rangle \mathrm{d}s\Big|. \qquad (3.9.18)\end{aligned}$$

Note that $\mathbf{\Phi}_t[f]$ is the function $\mathbf{\Lambda}_t[f]$ previously defined in (3.8.16) for test functions $f \in \mathcal{C}^{1,\gamma_0-1}(\mathbf{R}^{d+1})$ and for $\mathsf{m} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$.

Lemma 3.21. *Assume* **A1**→**A5**. *Let* $f \in \mathcal{C}^{1,\gamma_0-1}(\mathbf{R}^{d+1})$. *Then* $\mathbf{\Phi}_t[f]$ *is well defined. In addition, if a sequence* $(\mathsf{m}^N)_{N\geq 1}$ *converges to* m *in* $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$, *then, for all continuity point* $t \geq 0$ *of* m, *we have* $\mathbf{\Phi}_t[f](\mathsf{m}^N) \to \mathbf{\Phi}_t[f](\mathsf{m})$.

Proof. Using **A1**, and because Y is bounded and the function ϕ is bounded, $\mathscr{G}_1^{x,y} : \theta \mapsto \langle \phi(\theta, \cdot, x) - y, \gamma\rangle \in \mathcal{C}_b^\infty(\mathbf{R}^{d+1})$. In addition, for all multi-index $\alpha \in \mathbf{N}^{d+1}$, there exists $C > 0$, for all $x, y \in \mathsf{X} \times \mathsf{Y}$ and all $\theta \in \mathbf{R}^{d+1}$, $|\partial_\alpha \mathscr{G}_1^{x,y}(\theta)| \leq C$. The same holds for the function $\mathscr{G}_2^x : \theta \in \mathbf{R}^{d+1} \mapsto \langle \nabla_\theta \phi(\theta, \cdot, x), \gamma\rangle$. Consequently, $\theta \mapsto \nabla_\theta f(\theta) \cdot \mathscr{G}_2^x(\theta) \in \mathcal{C}^{0,\gamma_0-1}(\mathbf{R}^{d+1}) \hookrightarrow \mathcal{C}^{0,\gamma_0}(\mathbf{R}^{d+1})$. Then, there exists $C > 0$ independent of $(x, y) \in \mathsf{X} \times \mathsf{Y}$ and $s \in [0, t]$ such that

$$|\langle \mathscr{G}_1^{x,y}, \mathsf{m}_s\rangle| \leq C,$$

and

$$|\langle \nabla_\theta f \cdot \mathscr{G}_2^x, \mathsf{m}_s\rangle| \leq C\|f\|_{\mathcal{C}^{1,\gamma_0-1}} \langle 1 + |.|^{\gamma_0}, \mathsf{m}_s\rangle.$$

Finally, the function $\theta \mapsto \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)$ is smooth (see (3.2.4)) and (3.8.23) extends to all its derivatives, i.e. for all multi-index $\alpha \in \mathbf{N}^{d+1}$, there exists $c > 0$, for all $\theta \in \mathbf{R}^{d+1}$,

$$|\partial_\alpha \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)| \leq c(1 + |\theta|).$$

Thus, $\nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) \in \mathcal{C}^{0,\gamma_0}(\mathbf{R}^{d+1})$ and for some $C > 0$ independent of $s \in [0, t]$

$$|\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mathsf{m}_s\rangle| \leq C\|f\|_{\mathcal{C}^{1,\gamma_0-1}} \langle 1 + |.|^{\gamma_0}, \mathsf{m}_s\rangle.$$

Since in addition $\sup_{s\in[0,t]} \langle 1 + |.|^{\gamma_0}, \mathsf{m}_s\rangle < +\infty$ (since $s \mapsto \langle 1 + |.|^{\gamma_0}, \mathsf{m}_s\rangle \in \mathcal{D}(\mathbf{R}_+, \mathbf{R})$), $\mathbf{\Phi}_t[f]$ is well defined. To prove the continuity property of $\mathbf{\Phi}_t[f]$ it then suffices to use the previous upper bounds together similar arguments as those used in the proof of Lemma 3.11 (see also [DGMN22]). □

Proposition 3.22. *Assume* **A1**→**A5**. *Let* μ^* *be a limit point of* $(\mu^N)_{N\geq 1}$ *in* $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. *Then,* μ^* *satisfies a.s. Equation* (3.4.4).

Proof. Let us consider $f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$ and μ^* be a limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. Recall that by [EK09, lemma 7.7 in Chapter 3], the complementary of the set

$$\mathcal{C}(\mu^*) = \{t \geq 0, \mathbf{P}(\mu_{t^-}^* = \mu_t^*) = 1\}$$

is at most countable. Let $t_* \in \mathcal{C}(\mu^*)$. Then, by Lemma 3.21, one has that $\mathbf{P}(\mu^* \in \mathsf{D}(\mathbf{\Phi}_{t_*}[f])) = 0$. Thus, by the continuous mapping theorem, it holds

$$\mathbf{\Phi}_{t_*}[f](\mu^N) \xrightarrow{\mathscr{D}} \mathbf{\Phi}_{t_*}[f](\mu^*).$$

On the other hand, using (3.9.5) and the estimates (3.9.14), (3.9.13), (3.9.17), (3.9.8), and (3.9.10), it holds

$$\lim_{N\to\infty} \mathbf{E}[\mathbf{\Phi}_{t_*}[f](\mu^N)] = 0.$$

Consequently, for all $f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$ and $t_* \in \mathcal{C}(\mu^*)$, it holds a.s. $\mathbf{\Phi}_{t_*}[f](\mu^*) = 0$. On the other hand, for all $\psi \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$, $\mathsf{m} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$, and $s \geq 0$, the mappings

$$t \geq 0 \mapsto \mathbf{\Phi}_t[\psi](\mathsf{m})$$

is right continuous, and

$$f \in \mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1}) \mapsto \mathbf{\Phi}_s[f](\mathsf{m})$$

is continuous (because $\mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1}) \hookrightarrow \mathcal{C}_0^{1,\gamma_0-1}(\mathbf{R}^{d+1})$). In addition, $\mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1})$ admits a dense and countable subset of elements in $\mathcal{C}_c^\infty(\mathbf{R}^{d+1})$. Moreover, there exists a countable subset

$\mathcal{T}_{\mu^*}$ of $\mathcal{C}(\mu^*)$ such that for all $t \geq 0$ and $\epsilon > 0$, there exists $s \in \mathcal{T}_{\mu^*}$, $s \in [t, t+\epsilon]$. We prove this claim. Since $\mathbb{R}_+$ is a metric space, $\mathcal{C}(\mu^*)$ is separable and thus admits a dense subset $\mathcal{O}_{\mu^*}$. Since $[t+\epsilon/4, t+3\epsilon/4] \cap \mathcal{C}(\mu^*) \neq \emptyset$, there exists $u \in [t+\epsilon/4, t+3\epsilon/4] \cap \mathcal{C}(\mu^*)$. Consider now $s \in \mathcal{O}_{\mu^*}$ such that $|s-u| \leq \epsilon/4$. It then holds $t \leq s \leq t+\epsilon$, proving the claim with $\mathcal{T}_{\mu^*} = \mathcal{O}_{\mu^*}$.

Hence, we have with a classical argument that a.s. for all $f \in \mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1})$ and $t \geq 0$, $\mathbf{\Lambda}_t[f](\mu^*) = 0$. Note also that $\mathcal{C}_b^\infty(\mathbf{R}^{d+1}) \subset \mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1})$ since $2\gamma_0 > d+1$. This ends the proof of the proposition. □

3.9.3 Uniqueness of the limit equation and end of the proof of Theorem 3.2

In this section, we prove that there is a unique solution to (3.4.4) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$. To this end, we first need to prove that every limit points of $(\mu^N)_{N\geq 1}$ a.s. belongs to $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$.

Limit points belong to $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$

Proposition 3.23. *Assume* **A1→A5**. *Let $\mu^* \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ be a limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. Then, a.s. $\mu^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$.*

Proof. Note that since $\mathsf{W}_1 \leq \mathsf{W}_{\gamma_0}$, $\mu^{N'} \xrightarrow{\mathscr{D}} \mu^*$ also in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$, along some subsequence N'. According to [JS87, Proposition 3.26 in Chapter VI], $\mu^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ a.s. if for all $T > 0$, $\lim_{N\to+\infty} \mathbf{E}\big[\sup_{t\in[0,T]} \mathsf{W}_1(\mu^N_{t-}, \mu^N_t)\big] = 0$. Using (3.8.13), this is equivalent to prove that

$$\lim_{N\to+\infty} \mathbf{E}\Big[\sup_{t\in[0,T]} \sup_{\|f\|_{\mathrm{Lip}}\leq 1} |\langle f, \mu^N_{t-}\rangle - \langle f, \mu^N_t\rangle|\Big] = 0. \tag{3.9.19}$$

Let us consider $T > 0$ and a Lipschitz function $f : \mathbf{R}^{d+1} \to \mathbf{R}$ such that $\|f\|_{\mathrm{Lip}} \leq 1$. We have $\langle f, \mu^N_t\rangle = \langle f, \mu^N_0\rangle + \sum_{k=0}^{\lfloor Nt\rfloor -1} \langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle$ (with usual convention $\sum_0^{-1} = 0$). Thus the discontinuity points of $t \in [0,T] \mapsto \langle f, \mu^N_t\rangle$ lies exactly at $\{1/N, 2/N, \ldots, \lfloor NT\rfloor/N\}$ and

$$|\langle f, \mu^N_{t-}\rangle - \langle f, \mu^N_t\rangle| \leq \max_{k=0,\ldots,\lfloor NT\rfloor-1} |\langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle|, \quad \forall t \in [0,T],\ f \text{ Lipschitz}. \tag{3.9.20}$$

Pick $k = 0, \ldots, \lfloor NT\rfloor - 1$. We have by (3.9.2),

$$|\langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle| \leq \frac{1}{N}\sum_{i=1}^N |\theta^i_{k+1} - \theta^i_k| \leq \frac{C}{N}\sum_{i=1}^N \Big[\frac{1}{NB}\sum_{\ell=1}^B (1 + \mathfrak{b}(\mathsf{Z}^{i,\ell}_k)) + \frac{1}{N}(1 + |\theta^i_k|)\Big] =: d^N_k \tag{3.9.21}$$

Hence, it holds:

$$|d^N_k|^2 \leq \frac{C}{N}\sum_{i=1}^N \Big[\frac{1}{N^2B}\sum_{\ell=1}^B (1 + \mathfrak{b}^2(\mathsf{Z}^{i,\ell}_k)) + \frac{1}{N^2}(1 + |\theta^i_k|^2)\Big],$$

where thanks to Lemma 3.16 and **A1**, for all $k = 0, \ldots, \lfloor NT\rfloor - 1$, $\mathbf{E}[|d^N_k|^2] \leq C/N^2$ for some $C > 0$ independent of $N \geq 1$ and $k = 0, \ldots, \lfloor NT\rfloor - 1$. Thus, using (3.9.20) and (3.9.21),

$$\begin{aligned}
\mathbf{E}\Big[\sup_{t\in[0,T]} \sup_{\|f\|_{\mathrm{Lip}}\leq 1} |\langle f, \mu^N_{t-}\rangle - \langle f, \mu^N_t\rangle|\Big] &\leq \mathbf{E}\Big[\sup_{\|f\|_{\mathrm{Lip}}\leq 1} \max_{k=0,\ldots,\lfloor NT\rfloor-1} |\langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle|\Big] \\
&\leq \mathbf{E}\Big[\max_{k=0,\ldots,\lfloor NT\rfloor-1} d^N_k\Big] \\
&\leq \mathbf{E}\Big[\sqrt{\sum_{k=0}^{\lfloor NT\rfloor-1} |d^N_k|^2}\Big] \\
&\leq \sqrt{\mathbf{E}\Big[\sum_{k=0}^{\lfloor NT\rfloor-1} |d^N_k|^2\Big]} \leq \frac{C}{\sqrt{N}}.
\end{aligned}$$

This concludes the proof of Proposition 3.23. □

Uniqueness of the solution to (3.4.4)

Proposition 3.24. *There is a unique solution* $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ *to* (3.4.4).

Proof. First of all, the existence of a solution is provided by Propositions 3.17, 3.23 and 3.22. Let us now prove that there is a unique solution to (3.4.4) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$.

Recall the definition of $\boldsymbol{v}[\mu]$ in (3.8.18). We claim that for all $T > 0$ and all solution $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ of (3.4.4), there exists $C > 0$ such that

$$|\boldsymbol{v}[\bar{\mu}_t](\theta) - \boldsymbol{v}[\bar{\mu}_s](\theta)| \le C|t-s|, \quad \text{for all } 0 \le s \le t \le T \text{ and } \theta \in \mathbf{R}^{d+1}. \tag{3.9.22}$$

The proof of item (3.9.22) is the same as the one made for Item 2 in Proposition 3.13 since it holds using (3.8.23) and (3.8.27), for all $0 \le s \le t \le T$ and $z \in \mathbf{R}^d$,

$$\begin{aligned}\Big| \int_s^t \langle \nabla_\theta \phi(\cdot, z, x) \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot | P^1_0), \bar{\mu}_r \rangle \mathrm{d}r \Big| &\le C\mathfrak{b}(z) \int_s^t \langle (1 + |\cdot|), \bar{\mu}_r \rangle \mathrm{d}r \\ &\le C\mathfrak{b}(z) \max_{r \in [0,T]} \langle (1 + |\cdot|), \bar{\mu}_r \rangle |t-s|.\end{aligned}$$

We now conclude the proof of Proposition 3.24. Item 1 in the proof of Proposition 3.13 and (3.9.22) imply that $v(t, \theta) = \boldsymbol{v}[\bar{\mu}_t](\theta)$ is globally Lipschitz on $[0, T] \times \mathbf{R}^{d+1}$, for all $T > 0$, when $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ is a solution of (3.4.4). Since in addition a solution $\bar{\mu}$ to (3.4.4) is a weak solution on $\mathbf{R}_+$ to (3.8.19) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^{d+1}))$, it holds by [Vil03, Theorem 5.34]:

$$\forall t \ge 0, \ \bar{\mu}_t = \phi_t \# \mu_0, \tag{3.9.23}$$

where ϕ_t is the flow generated by the vector field $\boldsymbol{v}[\bar{\mu}_t](\theta)$ over $\mathbf{R}^{d+1}$. Together with Item 3 in the proof of Proposition 3.13 and using the same arguments as those used in Step 3 of the proof of [DGMN22, Proposition 2.14], two solutions agrees on each $[0, T]$ for all $T > 0$. One then deduces the uniqueness of the solution to (3.3.4). The proof of Proposition 3.24 is complete. □

We are now in position to end the proof of Theorem 3.2.

Proof of Theorem 3.2. By Proposition 3.17, $(\mu^N)_{N \ge 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. Let $\mu^1, \mu^2 \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ be two limit points of this sequence. By Proposition 3.23, a.s. $\bar{\mu}^1, \bar{\mu}^2 \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$. In addition, according to Proposition 3.22, μ^1 and μ^2 are a.s. solutions of (3.4.4). Denoting by $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ the unique solution to (3.4.4) (see Proposition 3.24), we have a.s.

$$\bar{\mu}^1 = \bar{\mu} \text{ and } \bar{\mu}^2 = \bar{\mu} \text{ in } \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1})).$$

In particular $\bar{\mu} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ and $\bar{\mu}^j = \bar{\mu}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$, $j \in \{1, 2\}$. As a consequence, $\bar{\mu}$ is the unique limit point of $(\mu^N)_{N \ge 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ and the whole sequence $(\mu^N)_{N \ge 1}$ converges to $\bar{\mu}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. Since $\bar{\mu}$ is deterministic, the convergence also holds in probability. The proof of Theorem 3.2 is complete. □

Let us now prove Proposition 3.3.

Proof of Proposition 3.3. Any solution to (3.3.4) in $\mathcal{C}([0, T], \mathcal{P}(\Theta_T))$ is a solution to (3.4.4) in $\mathcal{C}([0, T], \mathcal{P}_1(\mathbf{R}^{d+1}))$. The result follows from Proposition 3.24. □

Chapter 4

Central Limit Theorem for Bayesian Neural Networks trained with Variational Inference

This chapter is devoted to our work in preparation, presented in Section 1.3.3. More precisely, we discuss about the proof of Theorem 1.27. The proof follows the same lines as the proof of Theorem 1.17 (see more precisely Theorem 2.8), thus we focus on the convergence the martingale term to the G-process, which leads to different covariance structures, as well as a novel proof in the case of the *Bayes-by-backprop* algorithm (see Section 4.2). Each algorithm is treated separetely in its proper section. Let us recall the fluctuation process (1.3.11):

$$\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t),\ t \geq 0, N \geq 1.$$

Contents

4.1 The case of the *Idealized* algorithm

Let us first recall Theorem 1.27.

Theorem 4.1. *Assume* **A1**→**A4** *and let* $J_3 = 4\lceil \frac{d+1}{2} \rceil + 8$, $j_3 = \lceil \frac{d+1}{2} \rceil + 1$. *Consider* μ^N *defined by* (Alg-Id) *and* $\bar{\mu}$ *defined by* (1.3.8). *Then, the sequence* $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_3+1,j_3}(\mathbf{R}^{d+1}))$ *converges in distribution to* $\bar{\eta} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_3+1,j_3}(\mathbf{R}^{d+1}))$, *the unique weak solution (see Definitions 1.15 and 1.16) of:* $\forall t \geq 0$, $\forall f \in \mathcal{H}_0^{J_3,j_3-1}(\mathbf{R}^{d+1})$,

$$\begin{aligned}\langle f, \bar{\eta}_t\rangle - \langle f, \bar{\eta}_0\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\eta}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \bar{\eta}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\eta}_s \big\rangle \mathrm{d}s + \mathscr{G}_t[f],\end{aligned}$$

where $\bar{\eta}_0$ *is defined by*

$$(\langle f_1, \bar{\eta}_0\rangle, \ldots, \langle f_k, \bar{\eta}_0\rangle)^T \sim \mathcal{N}(0, \Gamma(f_1, \ldots, f_k)), \;\; f_1, \ldots, f_k \in \mathcal{H}_0^{J_3-1,j_3}(\mathbf{R}^{d+1}), \;\; k \geq 1,$$

with $\Gamma(f_1, \ldots, f_k)$ *the covariance matrix of* $(f_1(\theta_0^1), \ldots, f_k(\theta_0^1))^T$ *and where* $\mathscr{G}$ *is a* G*-process (see Definition 1.14) with covariance structure:* $\forall f, g \in \mathcal{H}_0^{J_3,j_3}(\mathbf{R}^{d+1}), \forall 0 \leq s \leq t$,

$$\mathrm{Cov}(\mathscr{G}_t[f], \mathscr{G}_s[g]) = \eta^2 \int_0^s \mathrm{Cov}(\mathrm{Q}[f](x,y,\bar{\mu}_v), \mathrm{Q}[g](x,y,\bar{\mu}_v))\mathrm{d}v, \tag{4.1.1}$$

where $\mathrm{Q}[f](x,y,\bar{\mu}_v) = \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_v \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_v \otimes \gamma\rangle$.

As already mentioned in the introduction, the proof of this theorem is similar to the proof of Theorem 1.17 (see more precisely Theorem 2.8). In the following, we enounce the pre-limit equation for η^N and discuss about the convergence of the martingale term $\sqrt{N}\mathbf{M}^N$ towards a G-process.

Pre-limit equation. Recall μ^N satisfies the pre-limit equation (3.8.9) and $\bar{\mu}$ satisfies (3.3.4). Hence, we obtain the pre-limit equation satisfied by $\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t)$:

$$\begin{aligned}\langle f, \eta_t^N\rangle - \langle f, \eta_0^N\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \frac{\eta}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \eta_s^N \rangle \mathrm{d}s \\ &+ \frac{\eta}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \frac{\eta}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &+ \sqrt{N}\mathbf{M}_t^N[f] + \sqrt{N}\mathbf{W}_t^N[f] + \sqrt{N}\mathbf{R}_t^N[f].\end{aligned} \tag{4.1.2}$$

where $\mathbf{M}_t^N[f]$, $\mathbf{W}_t^N[f]$ and $\mathbf{R}_t^N[f]$ and given by (3.8.5) and (3.8.8).

Convergence of $\sqrt{N}\mathbf{M}^N$ toward a G-process. To show that $\sqrt{N}\mathbf{M}^N$ converges to G-process with covariance structure given by (4.1.1), we proceed similarly as in Section 2.3.4, i.e., we first apply the central limit theorem for martingales for $\sqrt{N}\mathbf{M}^N[f]$, where f is a test function (see Proposition 2.33). It is then straightforward to obtain the same result in the multidimensional case (see Proposition 2.34). Combined with the relative compactness of $\sqrt{N}\mathbf{M}^N$ (see Proposition 2.30), we obtain that $\sqrt{N}\mathbf{M}^N$ converges to a G-process (see Proposition 2.35). Since the proofs of these results are similar the ones in Section 2.3.4, we confine ourselves to the application of the central limit theorem for martingales for $\sqrt{N}\mathbf{M}^N[f]$ (similarly as Proposition 2.33), since these calculations give the covariance structure of the G-process. According to [EK09, Theorem 1.4 in Chapter 7], the covariance structure of the G-process is given by the limit in probability (when $t \geq 0$ is fixed) of the covariation of $\mathbf{M}_t^N[f]$, that is, the limit in probability of

$$\mathfrak{a}_t^N[f] = N \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbf{M}_k^N[f]^2, \text{ as } N \to \infty. \tag{4.1.3}$$

In the following, we identify this limit. By (3.8.3), for $k \geq 0$,

$$\begin{aligned}
\mathbf{D}_k^N[f] &= -\frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1, j\neq i}^N \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\theta_k^j, \cdot, x) - y, \gamma \big\rangle \big\langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x), \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\
&\quad - \frac{\eta}{N^2} \int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot, \cdot, x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x), \nu_k^N \otimes \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\
&= -\frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1}^N \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\theta_k^j, \cdot, x) - y, \gamma \big\rangle \big\langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x), \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\
&\quad + \frac{\eta}{N^3} \sum_{i=1}^N \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\theta_k^i, \cdot, x) - y, \gamma \big\rangle \big\langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x), \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\
&\quad - \frac{\eta}{N^2} \int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot, \cdot, x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x), \nu_k^N \otimes \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y).
\end{aligned}$$

Let us introduce, for any $\nu \in \mathcal{H}^{J_0, j_0}(\mathbf{R}^{d+1})$ (where $J_0 = \lceil \frac{d+1}{2} \rceil + 3$ and $j_0 = 4\lceil \frac{d+1}{2} \rceil + 5$),

$$\bar{\mathrm{Q}}[f](\nu) = \int_{\mathsf{X}\times\mathsf{Y}} \big(\big\langle \phi(\cdot, \cdot, x), \nu \otimes \gamma \big\rangle - y \big) \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x), \nu \otimes \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y).$$

Setting, for $k \geq 0$ and $N \geq 1$,

$$\begin{aligned}
\mathfrak{R}_k^N[f] &:= \frac{\eta}{N^3} \sum_{i=1}^N (\langle \phi(\theta_k^i, \cdot, x_k), \gamma \rangle - y_k) \langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x_k), \gamma \rangle \\
&\quad - \frac{\eta}{N^2} \langle (\phi(\cdot, \cdot, x_k) - y_k) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x_k), \nu_k^N \otimes \gamma \rangle \\
&\quad - \frac{\eta}{N^3} \sum_{i=1}^N \int_{\mathsf{X}\times\mathsf{Y}} \big(\big\langle \phi(\theta_k^i, \cdot, x), \gamma \big\rangle - y \big) \big\langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x), \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\
&\quad + \frac{\eta}{N^2} \int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot, \cdot, x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x), \nu_k^N \otimes \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y),
\end{aligned}$$

we obtain

$$\mathbf{M}_k^N[f] = -\frac{\eta}{N} \mathrm{Q}[f](x_k, y_k, \nu_k^N) + \frac{\eta}{N} \bar{\mathrm{Q}}[f](\nu_k^N) + \mathfrak{R}_k^N[f]. \tag{4.1.4}$$

Hence, by (4.1.3) and (4.1.4), for all $t \in \mathbf{R}_+$,

$$\mathfrak{a}_t^N[f] = \frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} (\mathrm{Q}[f](x_k, y_k, \nu_k^N) - \bar{\mathrm{Q}}[f](\nu_k^N))^2 + 2\eta \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathfrak{R}_k^N[f](\bar{\mathrm{Q}}[f](\nu_k^N) - \mathrm{Q}[f](x_k, y_k, \nu_k^N)) + N \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathfrak{R}_k^N[f]^2. \tag{4.1.5}$$

Fix $t \geq 0$. Our aim is to determine the limit in probability of the sequence $(\mathfrak{a}_t^N[f])_{N \geq 1} \subset \mathbf{R}$. Observe that the two last terms of (4.1.5) converge to zero in probability, using the bounds $\mathbf{E}[|\mathfrak{R}_k^N[f]|^2] \leq C\|f\|^2_{\mathcal{C}^{1,j_0}}/N^4$ and $\mathbf{E}[|\bar{\mathrm{Q}}[f](\nu_k^N)|^2 + |\mathrm{Q}[f](x_k, y_k \nu_k^N)|^2] \leq C$ (these two bounds come from **A1**-**A2**-**A3** and Lemma 3.16). Therefore, one just needs to determine the limit in probability of

$$\begin{aligned}\frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} (\mathrm{Q}[f](x_k, y_k, \nu_k^N) - \bar{\mathrm{Q}}[f](\nu_k^N))^2 = \frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \nu_k^N)) \\ + \frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} (\mathrm{Q}[f](x_k, y_k, \nu_k^N) - \bar{\mathrm{Q}}[f](\nu_k^N))^2 \\ - \frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \nu_k^N)).\end{aligned}$$

On the one hand,

$$\begin{aligned}\frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \nu_k^N)) &= \eta^2 \sum_{k=0}^{\lfloor Nt \rfloor - 1} \int_{\frac{k}{N}}^{\frac{k+1}{N}} \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \mu_s^N)) \mathrm{d}s \\ &= \eta^2 \int_0^t \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \mu_s^N)) \mathrm{d}s - \eta^2 \int_{\frac{\lfloor Nt \rfloor}{N}}^t \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \mu_s^N)) \mathrm{d}s \\ &\xrightarrow[N \to +\infty]{p} \eta^2 \int_0^t \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \bar{\mu}_s)) \mathrm{d}s,\end{aligned}$$

using Theorem 3.1 together with the continuous mapping theorem and the dominated convergence theorem. On the other hand, we have, for some $C > 0$ independent of N,

$$\begin{aligned}&\mathbf{E}\Big[\Big|\frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} (\mathrm{Q}[f](x_k, y_k, \nu_k^N) - \bar{\mathrm{Q}}[f](\nu_k^N))^2 - \frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \nu_k^N))\Big|^2\Big] \\ &= \frac{\eta^4}{N^2} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbf{E}\Big[\Big|(\mathrm{Q}[f](x_k, y_k, \nu_k^N) - \bar{\mathrm{Q}}[f](\nu_k^N))^2 - \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \nu_k^N))\Big|^2\Big] \\ &\leq \frac{C}{N^2} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \mathbf{E}[|\mathrm{Q}[f](x_k, y_k, \nu_k^N)|^4] \leq \frac{C}{N} \to_{N \to +\infty} 0.\end{aligned}$$

We have thus shown that

$$\mathfrak{a}_t^N[f] \xrightarrow[N \to +\infty]{p} \eta^2 \int_0^t \mathrm{Var}_\pi(\mathrm{Q}[f](x, y, \bar{\mu}_s)) \mathrm{d}s.$$

In the following section, we do the same analysis in the case of algorithm (Alg-Bbb).

4.2 The case of the *Bayes-by-backprop* algorithm

Let us first recall Theorem 1.27.

Theorem 4.2. *Assume* **A1**→**A5** *and let* $J_3 = 4\lceil \frac{d+1}{2} \rceil + 8$, $j_3 = \lceil \frac{d+1}{2} \rceil + 1$. *Consider* μ^N *defined by* (Alg-Bbb) *and* $\bar{\mu}$ *defined by* (1.3.8). *Then, the sequence* $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_3+1,j_3}(\mathbf{R}^{d+1}))$ *converges in distribution to* $\bar{\eta} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_3+1,j_3}(\mathbf{R}^{d+1}))$, *the unique weak solution (see Definitions 1.15 and 1.16) of:* $\forall t \geq 0$, $\forall f \in \mathcal{H}_0^{J_3,j_3-1}(\mathbf{R}^{d+1})$,

$$\begin{aligned}\langle f, \bar{\eta}_t\rangle - \langle f, \bar{\eta}_0\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\eta}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \bar{\eta}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\eta}_s \big\rangle \mathrm{d}s + \mathscr{G}_t[f],\end{aligned}$$

where $\bar{\eta}_0$ *is defined by*

$$(\langle f_1, \bar{\eta}_0\rangle, \ldots, \langle f_k, \bar{\eta}_0\rangle)^T \sim \mathcal{N}(0, \Gamma(f_1, \ldots, f_k)),\ f_1, \ldots, f_k \in \mathcal{H}_0^{J_3-1,j_3}(\mathbf{R}^{d+1}),\ k \geq 1,$$

with $\Gamma(f_1, \ldots, f_k)$ *the covariance matrix of* $(f_1(\theta_0^1), \ldots, f_k(\theta_0^1))^T$ *and where* $\mathscr{G}$ *is a* G*-process (see Definition 1.14) with covariance structure:* $\forall f, g \in \mathcal{H}_0^{J_3,j_3}(\mathbf{R}^{d+1}), \forall 0 \leq s \leq t$,

$$\mathrm{Cov}(\mathscr{G}_t[f], \mathscr{G}_s[g]) = \eta^2 \int_0^s \mathrm{Cov}(\mathrm{Q}[f](x,y,\bar{\mu}_v), \mathrm{Q}[g](x,y,\bar{\mu}_v))\mathrm{d}v,$$

where $\mathrm{Q}[f](x,y,\bar{\mu}_v) = \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_v \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_v \otimes \gamma\rangle$.

Similarly as in Section 4.1, we first enounce the pre-limit equation satisfied by η^N, and then discuss about the convergence of $\sqrt{\mathbf{M}}^N$ towards a G-process.

Pre-limit equation. Using (3.9.5) and (3.4.4), we obtain the following pre-limit equation for $\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t)$:

$$\begin{aligned}\langle f, \eta_t^N\rangle - \langle f, \eta_0^N\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \frac{\eta}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \eta_s^N \rangle \mathrm{d}s \\ &+ \frac{\eta}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \frac{\eta}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &+ \sqrt{N}\mathbf{M}_t^N[f] + \sqrt{N}\mathbf{W}_t^N[f] + \sqrt{N}\mathbf{R}_t^N[f].\end{aligned}$$

Note that the terms $\mathbf{W}_t^N[f]$ and $\mathbf{R}_t^N[f]$ have the same expression as those which appear in (4.1.2). We also have

$$\begin{aligned}\mathbf{M}_k^N[f] &= -\frac{\eta}{N^3B}\sum_{i,j=1}^N\sum_{\ell=1}^B(\phi(\theta_k^j,\mathsf{Z}_k^{j,\ell},x_k)-y_k)\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\mathsf{Z}_k^{i,\ell},x_k)-\mathbf{D}_k^N[f]\\ &= -\frac{\eta}{N^3B}\sum_{i,j=1,j\neq i}^N\sum_{\ell=1}^B(\phi(\theta_k^j,\mathsf{Z}_k^{j,\ell},x_k)-y_k)\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\mathsf{Z}_k^{i,\ell},x_k)\\ &\quad+\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1,j\neq i}^N\int_{\mathsf{X}\times\mathsf{Y}}\big(\langle\phi(\theta_k^j,\cdot,x),\gamma\rangle-y\big)\big\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y)+\mathscr{R}_k^N[f],\end{aligned}$$

where

$$\begin{aligned}\mathbf{D}_k^N[f] &= -\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1,j\neq i}^N\int_{\mathsf{X}\times\mathsf{Y}}\big(\langle\phi(\theta_k^j,\cdot,x),\gamma\rangle-y\big)\big\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\ &\quad-\frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\end{aligned}$$

and

$$\begin{aligned}\mathscr{R}_k^N[f] &= -\frac{\eta}{N^3B}\sum_{i=1}^N\sum_{\ell=1}^B(\phi(\theta_k^i,\mathsf{Z}_k^{i,\ell},x_k)-y_k)\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\mathsf{Z}_k^{i,\ell},x_k)\\ &\quad+\frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y).\end{aligned}$$

Convergence of $\sqrt{N}\mathbf{M}^N$ toward a G-process. We refer to our discussion in 4.1 about the proof of convergence convergence of $\sqrt{N}\mathbf{M}^N$ toward a G-process, and confine ourselves here only to the convergence of $\sqrt{N}\mathbf{M}_t^N[f]$, for fixed $t\geq 0$ (see more precisely Proposition 2.33 and [EK09, Theorem 1.4 in Chapter 7]). Since our analysis requires $B=1$, we set $B=1$ throughout this paragraph. Again, we define, as in (4.1.3),

$$\mathfrak{a}_t^N[f] = N\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{M}_k^N[f]^2.$$

Let us introduce the random probability measures over $\mathbf{R}^{d+1}\times\mathbf{R}^d$

$$r_k^N=\frac{1}{N}\sum_{i=1}^N\delta_{(\theta_k^i,\mathsf{Z}_k^i)}\text{ and }\rho_t^N=r_{\lfloor Nt\rfloor}^N,\ k\geq 0,\ t\geq 0. \tag{4.2.1}$$

We also set, for $(x,y)\in\mathsf{X}\times\mathsf{Y}$ and $\rho\in\mathcal{P}(\mathbf{R}^{d+1}\times\mathbf{R}^d)$,

$$Q[f](x,y,\rho)=\langle\phi(\cdot,\cdot,x)-y,\rho\rangle\langle\nabla_\theta f(e_1(\cdot))\cdot\nabla_\theta\phi(\cdot,\cdot,x),\rho\rangle,$$

where e_1 is defined by $e_1(\theta^1,\ldots,\theta^{d+1})=(\theta^1,\ldots,\theta^d)$, for $(\theta^1,\ldots,\theta^{d+1})\in\mathbf{R}^{d+1}$. We have

$$\begin{aligned}\mathbf{M}_k^N[f] &= -\frac{\eta}{N^3}\sum_{i,j=1}^N(\phi(\theta_k^j,\mathsf{Z}_k^j,x_k)-y_k)\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\mathsf{Z}_k^i,x_k)-\mathbf{D}_k^N[f]\\ &= -\frac{\eta}{N}\langle\phi(\cdot,\cdot,x_k)-y_k,r_k^N\rangle\langle\nabla_\theta f(e_1(\cdot))\cdot\nabla_\theta\phi(\cdot,\cdot,x_k),r_k^N\rangle-\mathbf{D}_k^N[f]\\ &= -\frac{\eta}{N}Q(x_k,y_k,r_k^N)-\mathbf{D}_k^N[f]=\mathrm{F}^N(x_k,y_k,r_k^N)-\mathbf{D}_k^N[f]\end{aligned}$$

where

$$\mathrm{F}^N(x_k, y_k, r_k^N) = -\frac{\eta}{N} Q(x_k, y_k, r_k^N).$$

Fix $t \geq 0$. Our aim is to determine the limit in probability of the sequence $(\mathfrak{a}_t^N[f])_{N\geq 1} \subset \mathbf{R}$. We define a new filtration, different from $\mathcal{F}_k^N$ (see (3.4.1)), in which the $\mathbf{Z}^i$'s at iteration k are considered:

$$\widetilde{\mathcal{F}}_k^N = \boldsymbol{\sigma}\Big(\theta_0^i, \mathsf{Z}_{q'}^j, (x_q, y_q), 1 \leq i, j \leq N, 0 \leq q \leq k-1, 0 \leq q' \leq k\Big),\ k \geq 1.$$

We have

$$\mathfrak{a}_t^N[f] = N \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathbf{M}_k^N[f]^2 = N \sum_{k=0}^{\lfloor Nt\rfloor -1} \Big(\mathbf{E}\big[\mathbf{M}_k^N[f]^2 \big| \widetilde{\mathcal{F}}_k^N\big] + \mathbf{M}_k^N[f]^2 - \mathbf{E}\big[\mathbf{M}_k^N[f]^2 \big| \widetilde{\mathcal{F}}_k^N\big]\Big). \quad (4.2.2)$$

Since $\mathbf{E}[|\mathbf{M}_k^N[f]|^4] \leq C\|f\|^4_{\mathcal{C}^{1,j_0}}/N^4$,

$$\begin{aligned}\mathbf{E}\Big[\Big(N \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathbf{M}_k^N[f]^2 - \mathbf{E}[\mathbf{M}_k^N[f]^2|\widetilde{\mathcal{F}}_k^N]\Big)^2\Big] &= N^2 \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathbf{E}\Big[\Big(\mathbf{M}_k^N[f]^2 - \mathbf{E}[\mathbf{M}_k^N[f]^2|\widetilde{\mathcal{F}}_k^N]\Big)^2\Big] \\ &\leq CN^2 \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathbf{E}[\mathbf{M}_k^N[f]^4] \leq CN^2\|f\|^4_{\mathcal{C}^{1,j_0}}/N^3 \to 0.\end{aligned}$$

Hence, the two last terms of (4.2.2) converge to zero in probability:

$$N \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathbf{M}_k^N[f]^2 - \mathbf{E}\big[\mathbf{M}_k^N[f]^2 \big| \widetilde{\mathcal{F}}_k^N\big] \xrightarrow[N\to\infty]{p} 0. \quad (4.2.3)$$

Therefore, the limit in probability $\mathfrak{c}_t[f]$ of $\mathfrak{a}_t^N[f]$ is given by the limit in probability of

$$N \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathbf{E}\big[\mathbf{M}_k^N[f]^2 \big| \widetilde{\mathcal{F}}_k^N\big] = N \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathrm{Var}_\pi(\mathrm{F}^N(x, y, r_k^N)).$$

It holds

$$\begin{aligned} N \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathrm{Var}_\pi(\mathrm{F}^N(x, y, r_k^N)) &= \frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt\rfloor -1} \mathrm{Var}_\pi(Q(x, y, r_k^N) \\ &= \eta^2 \sum_{k=0}^{\lfloor Nt\rfloor -1} \int_{\frac{k}{N}}^{\frac{k+1}{N}} \mathrm{Var}_\pi(Q(x, y, \rho_s^N))\mathrm{d}s \\ &= \eta^2 \int_0^t \mathrm{Var}_\pi(Q(x, y, \rho_s^N))\mathrm{d}s - \eta^2 \int_{\frac{\lfloor Nt\rfloor}{N}}^t \mathrm{Var}_\pi(Q(x, y, \rho_s^N))\mathrm{d}s \end{aligned} \quad (4.2.4)$$

Our aim is now to pass to the limit in probability in (4.2.4) as $N \to +\infty$, for fixed $t \geq 0$. To this end, we introduce (N') a subsequence of (N). We seek a subsequence $(N'') \subset (N')$ such that $N'' \sum_{k=0}^{\lfloor N''t\rfloor -1} \mathrm{Var}_\pi(\mathrm{F}^{N''}(x, y, r_k^{N''}))$ converges almost surely to $\eta^2 \int_0^t \mathrm{Var}_\pi(Q(x, y, \bar{\mu}_s \otimes \gamma))\mathrm{d}s$. We know that $\mu^N \xrightarrow{p} \bar{\mu}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{j_0}(\mathbf{R}^{d+1}))$. Hence, there exists a subsequence (N'') of (N') such that $\mu^{N''} \xrightarrow{a.s.} \bar{\mu}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{j_0}(\mathbf{R}^{d+1}))$. By Lemma 4.3 below, a.s. for all $0 \leq s \leq t$, $\rho_s^{N''} \xrightarrow[N\to+\infty]{} \bar{\mu}_s \otimes \gamma$. Using **A1-A2-A3** and the fact that for almost every $\omega \in \Omega$, there exists $C(\omega) > 0$ such that for all N'' and $s \in [0, t]$, $\langle 1+|\cdot|^{j_0}, \mu_s^{N''}(\omega)\rangle \leq C(\omega)$, we have, using the dominated convergence

theorem, that almost surely for all $0 \le s \le t$, $\mathrm{Var}_\pi(Q(x,y,\rho_s^{N''})) \to \mathrm{Var}_\pi(Q(x,y,\bar\mu_s \otimes \gamma))$. Using the dominated convergence theorem again,

$$\eta^2 \int_0^t \mathrm{Var}_\pi(Q(x,y,\rho_s^{N''}))\mathrm{d}s \xrightarrow[N''\to\infty]{a.s.} \eta^2 \int_0^t \mathrm{Var}_\pi(Q(x,y,\bar\mu_s\otimes\gamma))\mathrm{d}s.$$

We now show that the last term of (4.2.4) tends to zero almost surely, up to a subsequence. Since $\mathbf{E}[|Q(x,y,\rho_s^N)|^2] \le C\|f\|^2_{\mathcal{C}^{1,j_0}}$, we have

$$\mathbf{E}\Big[\Big|\int_{\frac{\lfloor Nt\rfloor}{N}}^t \mathrm{Var}_\pi(Q(x,y,\rho_s^N))\mathrm{d}s\Big|\Big] \le \frac{C\|f\|^2_{\mathcal{C}^{1,j_0}}}{N} \xrightarrow[N\to\infty]{} 0.$$

Therefore, up to a subsequence, $\int_{\frac{\lfloor Nt\rfloor}{N}}^t \mathrm{Var}_\pi(Q(x,y,\rho_s^N))\mathrm{d}s \xrightarrow[N\to\infty]{p.s.} 0$. Since we have found a subsequence (N''') such that almost sure convergence holds, we have shown that

$$N\sum_{k=0}^{\lfloor Nt\rfloor-1} \mathrm{Var}_\pi(\mathrm{F}^N(x,y,r_k^N)) \xrightarrow[N\to\infty]{p} \eta^2\int_0^t \mathrm{Var}_\pi(Q(x,y,\bar\mu_s\otimes\gamma))\mathrm{d}s,$$

which is the desired result since $Q(x,y,\bar\mu_s\otimes\gamma) = \mathrm{Q}(x,y,\bar\mu_s)$.

Lemma 4.3. *Consider algorithm* (Alg-Bbb). *Assume* $\mu^{N'} \xrightarrow{a.s.} \bar\mu$ *in* $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{j_0}(\mathbf{R}^{d+1}))$ *for some subsequence. Then, almost surely, for all* $0 \le t \le T$,

$$\rho_t^{N'} \xrightarrow[N\to+\infty]{} \bar\mu_t \otimes \gamma \text{ in } \mathcal{P}(\mathbf{R}^{d+1}\times\mathbf{R}^d).$$

Proof. For simplicity, we denote N instead of N'. Let $0 \le t \le T$ and $g \in \mathcal{C}_b(\mathbf{R}^{d+1}\times\mathbf{R}^d)$. We have

$$\begin{aligned}\langle g, \rho_t^N\rangle - \langle g, \bar\mu_t\otimes\gamma\rangle = \frac{1}{N}\sum_{i=1}^N \Big(g(\theta^i_{\lfloor Nt\rfloor}, \mathsf{Z}^i_{\lfloor Nt\rfloor}) - \int_{\mathbf{R}^d} g(\theta^i_{\lfloor Nt\rfloor}, z)\gamma(z)\mathrm{d}z\Big) \\ + \frac{1}{N}\sum_{i=1}^N \int_{\mathbf{R}^d} g(\theta^i_{\lfloor Nt\rfloor}, z)\gamma(z)\mathrm{d}z - \langle g, \bar\mu_t\otimes\gamma\rangle. \qquad (4.2.5)\end{aligned}$$

On the one hand,

$$
\begin{aligned}
&\mathbf{E}\Big[\sup_{t\in[0,T]}\Big(\frac{1}{N}\sum_{i=1}^{N} g(\theta^i_{\lfloor Nt\rfloor},\mathsf{Z}^i_{\lfloor Nt\rfloor}) - \int_{\mathbf{R}^d} g(\theta^i_{\lfloor Nt\rfloor},z)\gamma(z)\mathrm{d}z\Big)^6\Big]\\
&\le \sum_{k=0}^{\lfloor NT\rfloor}\mathbf{E}\Big[\Big(\frac{1}{N}\sum_{i=1}^{N} g(\theta^i_k,\mathsf{Z}^i_k) - \int_{\mathbf{R}^d} g(\theta^i_k,z)\gamma(z)\mathrm{d}z\Big)^6\Big]\\
&= \frac{1}{N^6}\sum_{k=0}^{\lfloor NT\rfloor}\sum_{i=1}^{N}\mathbf{E}\Big[\Big(g(\theta^i_k,\mathsf{Z}^i_k) - \int_{\mathbf{R}^d} g(\theta^i_k,z)\gamma(z)\mathrm{d}z\Big)^6\Big]\\
&\quad + \frac{1}{N^6}\sum_{k=0}^{\lfloor NT\rfloor}\sum_{i\neq j}\mathbf{E}\Big[\Big(g(\theta^i_k,\mathsf{Z}^i_k) - \int_{\mathbf{R}^d} g(\theta^i_k,z)\gamma(z)\mathrm{d}z\Big)^3\Big(g(\theta^j_k,\mathsf{Z}^j_k) - \int_{\mathbf{R}^d} g(\theta^j_k,z)\gamma(z)\mathrm{d}z\Big)^3\Big]\\
&\quad + \frac{1}{N^6}\sum_{k=0}^{\lfloor NT\rfloor}\sum_{i\neq j}\mathbf{E}\Big[\Big(g(\theta^i_k,\mathsf{Z}^i_k) - \int_{\mathbf{R}^d} g(\theta^i_k,z)\gamma(z)\mathrm{d}z\Big)^4\Big(g(\theta^j_k,\mathsf{Z}^j_k) - \int_{\mathbf{R}^d} g(\theta^j_k,z)\gamma(z)\mathrm{d}z\Big)^2\Big]\\
&\quad + \frac{1}{N^6}\sum_{k=0}^{\lfloor NT\rfloor}\sum_{i\neq j\neq \ell}\mathbf{E}\Big[\Big(g(\theta^i_k,\mathsf{Z}^i_k) - \int_{\mathbf{R}^d} g(\theta^i_k,z)\gamma(z)\mathrm{d}z\Big)^2\Big(g(\theta^j_k,\mathsf{Z}^j_k) - \int_{\mathbf{R}^d} g(\theta^j_k,z)\gamma(z)\mathrm{d}z\Big)^2\\
&\qquad \times\Big(g(\theta^\ell_k,\mathsf{Z}^\ell_k) - \int_{\mathbf{R}^d} g(\theta^\ell_k,z)\gamma(z)\mathrm{d}z\Big)^2\Big]\\
&\le C_g/N^2.
\end{aligned}
$$

Hence,

$$
\sum_{N\ge 1}\mathbf{E}\Big[\sup_{t\in[0,T]}\Big(\frac{1}{N}\sum_{i=1}^{N} g(\theta^i_{\lfloor Nt\rfloor},\mathsf{Z}^i_{\lfloor Nt\rfloor}) - \int_{\mathbf{R}^d} g(\theta^i_{\lfloor Nt\rfloor},z)\gamma(z)\mathrm{d}z\Big)^6\Big] < +\infty.
$$

By Borel-Cantelli lemma,

$$
\sup_{t\in[0,T]}\Big|\frac{1}{N}\sum_{i=1}^{N} g(\theta^i_{\lfloor Nt\rfloor},\mathsf{Z}^i_{\lfloor Nt\rfloor}) - \int_{\mathbf{R}^d} g(\theta^i_{\lfloor Nt\rfloor},z)\gamma(z)\mathrm{d}z\Big| \xrightarrow{a.s.} 0. \tag{4.2.6}
$$

On the other hand, let us show that a.s., for all $t\in\mathbf{R}_+$,

$$
\frac{1}{N}\sum_{i=1}^{N}\int_{\mathbf{R}^d} g(\theta^i_{\lfloor Nt\rfloor},z)\gamma(z)\mathrm{d}z - \langle g,\bar\mu_t\otimes\gamma\rangle \to_{N\to+\infty} 0. \tag{4.2.7}
$$

Since $\mathsf{W}_1\le\mathsf{W}_{j_0}$, we have that

$$
\mu^N \xrightarrow{a.s.} \bar\mu \text{ in } \mathcal{D}(\mathbf{R}_+,\mathcal{P}_1(\mathbf{R}^{d+1})).
$$

As $\bar\mu\in\mathcal{C}(\mathbf{R}_+,\mathcal{P}_1(\mathbf{R}^{d+1}))$, it holds a.s. for all $t\in\mathbf{R}_+$, $\mu^N_t\to_{N\to\infty}\bar\mu_t$ in $\mathcal{P}_1(\mathbf{R}^{d+1})$. Considering $\mathsf{g}:\theta\in\mathbf{R}^{d+1}\mapsto\int_{\mathbf{R}^d} g(\theta,z)\gamma(z)\mathrm{d}z$, which is bounded continuous, we have a.s. for all $t\in\mathbf{R}_+$, $\langle\mathsf{g},\mu^N_t\rangle\to_{N\to\infty}\langle\mathsf{g},\bar\mu_t\rangle$, which is exactly (4.2.7). At this point, considering (4.2.5) together with (4.2.6) and (4.2.7), we have shown that for all $g\in\mathcal{C}_b(\mathbf{R}^{d+1}\times\mathbf{R}^d)$, a.s. for all $t\in[0,T]$,

$$
\langle g,\rho^N_t\rangle\to_{N\to\infty}\langle g,\bar\mu_t\otimes\gamma\rangle. \tag{4.2.8}
$$

To prove that a.s., for all $t\in[0,T]$, $\rho^N_t\to\bar\mu_t\otimes\gamma$, it is sufficient to show that a.s. for all $t\in[0,T]$ and $g\in\mathcal{C}^\infty_c(\mathbf{R}^{d+1}\times\mathbf{R}^d)$, $\langle g,\rho^N_t\rangle\to_{N\to\infty}\langle g,\bar\mu_t\otimes\gamma\rangle$ (see [AGS05, Remark 5.1.6]). Since the space $\mathcal{C}^\infty_c(\mathbf{R}^{d+1}\times\mathbf{R}^d)$ is separable, this last statement follows from (4.2.8) and a continuity argument. Hence, we have proved that a.s., for all $t\in[0,T]$, $\rho^N_t\to\bar\mu_t\otimes\gamma$, as desired. □

4.3 The case of the *Minimal-VI* algorithm

Let us first recall Theorem 1.27.

Theorem 4.4. *Assume* **A1**→**A4** *and* **A6** *and let* $J_3 = 4\lceil\frac{d+1}{2}\rceil + 8$, $j_3 = \lceil\frac{d+1}{2}\rceil + 1$. *Consider* μ^N *defined by* (Alg-Id) *and* $\bar{\mu}$ *defined by* (1.3.8). *Then, the sequence* $(\eta^N)_{N\geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{H}^{-J_3+1,j_3}(\mathbf{R}^{d+1}))$ *converges in distribution to* $\bar{\eta} \in \mathcal{C}(\mathbf{R}_+, \mathcal{H}^{-J_3+1,j_3}(\mathbf{R}^{d+1}))$, *the unique weak solution (see Definitions 1.15 and 1.16) of:* $\forall t \geq 0$, $\forall f \in \mathcal{H}_0^{J_3,j_3-1}(\mathbf{R}^{d+1})$,

$$\begin{aligned}\langle f, \bar{\eta}_t\rangle - \langle f, \bar{\eta}_0\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\eta}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \bar{\eta}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\eta}_s \big\rangle \mathrm{d}s + \mathscr{G}_t[f],\end{aligned}$$

where $\bar{\eta}_0$ *is defined by*

$$(\langle f_1, \bar{\eta}_0\rangle, \ldots, \langle f_k, \bar{\eta}_0\rangle)^T \sim \mathcal{N}(0, \Gamma(f_1, \ldots, f_k)),\ f_1, \ldots, f_k \in \mathcal{H}_0^{J_3-1,j_3}(\mathbf{R}^{d+1}),\ k \geq 1,$$

with $\Gamma(f_1, \ldots, f_k)$ *the covariance matrix of* $(f_1(\theta_0^1), \ldots, f_k(\theta_0^1))^T$ *and where* $\mathscr{G}$ *is a* G*-process (see Definition 1.14) with covariance structure:* $\forall f, g \in \mathcal{H}_0^{J_3,j_3}(\mathbf{R}^{d+1}), \forall 0 \leq s \leq t$,

$$\mathrm{Cov}(\mathscr{G}_t[f], \mathscr{G}_s[g]) = \eta^2 \int_0^s \mathrm{Cov}(\mathrm{Q}[f](x,y,z^1,z^2,\bar{\mu}_v), \mathrm{Q}[g](x,y,z^1,z^2,\bar{\mu}_v))\mathrm{d}v,$$

where $\mathrm{Q}[f](x,y,z^1,z^2,\bar{\mu}_v) = \langle \phi(\cdot,z^1,x) - y, \bar{\mu}_v\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,z^2,x), \bar{\mu}_v\rangle$.

Pre-limit equation. Using (3.9.5) and (3.4.4), we obtain the following pre-limit equation for $\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t)$:

$$\begin{aligned}\langle f, \eta_t^N\rangle - \langle f, \eta_0^N\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &- \frac{\eta}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \eta_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \eta_s^N\rangle \mathrm{d}s \\ &+ \frac{\eta}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &- \frac{\eta}{\sqrt{N}} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &+ \sqrt{N}\mathbf{M}_t^N[f] + \sqrt{N}\mathbf{W}_t^N[f] + \sqrt{N}\mathbf{R}_t^N[f].\end{aligned}$$

where

$$\mathbf{M}_k^N[f] = -\frac{\eta}{N} \langle \phi(\cdot, \mathsf{Z}_k^1, x_k) - y_k, \nu_k^N\rangle \langle \nabla_\theta f \cdot \nabla \phi(\cdot, \mathsf{Z}_k^2, x_k), \nu_k^N\rangle - \mathbf{D}_k^N[f].$$

Note that the terms $\mathbf{W}_t^N[f]$ and $\mathbf{R}_t^N[f]$ have the same expression as those which appear in (4.1.2).

Convergence of $\sqrt{N}\mathbf{M}^N$ toward a G-process. We refer to our discussion in 4.1 about the proof of convergence convergence of $\sqrt{N}\mathbf{M}^N$ toward a G-process, and confine ourselves here only to the convergence of $\sqrt{N}\mathbf{M}_t^N[f]$, for fixed $t \geq 0$ (see more precisely Proposition 2.33 and [EK09, Theorem 1.4 in Chapter 7]). The computations of this paragraph are similar to the ones of Section 4.1. We use again the decomposition

$$\begin{aligned}\mathfrak{a}_t^N[f] &= N \sum_{k=0}^{\lfloor Nt\rfloor-1} \mathbf{M}_k^N[f]^2 \\ &= N \sum_{k=0}^{\lfloor Nt\rfloor-1} \mathbf{E}[\mathbf{M}_k^N[f]^2|\mathcal{F}_k^N] + N \sum_{k=0}^{\lfloor Nt\rfloor-1} \mathbf{M}_k^N[f]^2 - \mathbf{E}[\mathbf{M}_k^N[f]^2|\mathcal{F}_k^N]. \end{aligned} \tag{4.3.1}$$

Introduce

$$\mathrm{Q}[f](x,y,z^1,z^2,\mu) = \langle \phi(\cdot,z^1,x) - y, \mu\rangle \langle \nabla_\theta f \cdot \nabla\phi(\cdot,z^2,x),\mu\rangle.$$

On the one hand, considering the fist term in the right hand side of (4.3.1), we have, considering Z^1 and Z^2 as two additionnal data,

$$\begin{aligned} N \sum_{k=0}^{\lfloor Nt\rfloor-1} \mathbf{E}[\mathbf{M}_k^N[f]^2|\mathcal{F}_k^N] &= \frac{\eta^2}{N} \sum_{k=0}^{\lfloor Nt\rfloor-1} \mathrm{Var}_{\pi\otimes\gamma^{\otimes 2}}(\mathrm{Q}[f](x,y,z^1,z^2,\nu_k^N)) \\ &= \eta^2 \sum_{k=0}^{\lfloor Nt\rfloor-1} \int_{\frac{k}{N}}^{\frac{k+1}{N}} \mathrm{Var}_{\pi\otimes\gamma^{\otimes 2}}(\mathrm{Q}[f](x,y,z^1,z^2,\mu_s^N))\mathrm{d}s \\ &= \eta^2 \int_0^t \mathrm{Var}_{\pi\otimes\gamma^{\otimes 2}}(\mathrm{Q}[f](x,y,z^1,z^2,\mu_s^N))\mathrm{d}s \\ &\quad - \int_{\frac{\lfloor Nt\rfloor}{N}}^t \mathrm{Var}_{\pi\otimes\gamma^{\otimes 2}}(\mathrm{Q}[f](x,y,z^1,z^2,\mu_s^N))\mathrm{d}s \\ &\xrightarrow[N\to+\infty]{p} \eta^2 \int_0^t \mathrm{Var}_{\pi\otimes\gamma^{\otimes 2}}(\mathrm{Q}[f](x,y,z^1,z^2,\bar{\mu}_s))\mathrm{d}s. \end{aligned}$$

On the other hand, the L^2-norm of the second term in the right hand side of (4.3.1) vanishes as $N \to +\infty$. Therefore, we have shown that

$$\mathfrak{a}_t^N[f] \xrightarrow[N\to+\infty]{p} \eta^2 \int_0^t \mathrm{Var}_{\pi\otimes\gamma^{\otimes 2}}(\mathrm{Q}[f](x,y,z^1,z^2,\bar{\mu}_s))\mathrm{d}s,$$

as desired.

Appendix A

Mathematical tools

Contents

A.1 Skorohod spaces

Our main references for this section are [Bil99] and [EK09]. Throughout this section (E, r) is a metric space, assumed separable and complete.

Definition A.1. *Let $0 < T < \infty$ and $I = [0, T]$ or $[0, \infty)$. A function $x : I \to E$ is said to belong to the Skorohod space $\mathcal{D}(I, E)$ if it is càdlàg, i.e., if it satisfies, for all $t \in I$,*

$$\lim_{\substack{s \to t \\ s > t}} x(s) = x(t) \text{ and } \lim_{\substack{s \to t \\ s < t}} x(s) \text{ exists.}$$

To define a metric on this space, contrary to space of continuous functions, we allow a uniformly small deformation of the time scale. Physically, this amounts to the recognition that we cannot measure time with perfect accuracy any more than we can position. The following topology, devised by Skorohod, embodies this idea. We start by focusing on the case when $I = [0, T]$. Let Λ denote the space of continuous increasing functions $\lambda : [0, T] \to [0, T]$ with $\lambda(0) = 0$ and $\lambda(T) = T$. The following metric defines the so called Skorohod $J1$-topology: for $x, y \in \mathcal{D}([0, T], E)$,

$$d_T(x, y) = \inf_{\lambda \in \Lambda} \Big(\sup_{t \in [0,T]} r(x(\lambda(t)), y(t)) + \sup_{t \in [0,T]} |\lambda(t) - t| \Big).$$

We have the immediate following property.

Proposition A.1. *Let $x, x_1, x_2, ... \in \mathcal{D}([0, T], E)$. Then, $d_T(x_n, x) \to 0$ if and only if there exists a sequence $(\lambda_n)_{n \geq 1} \subset \Lambda$ such that $x_n(\lambda_n(t)) \to x(t)$ uniformly in t and $\lambda_n(t) \to t$ uniformly in t.*

Endowed with this metric, the Skorohod space $\mathcal{D}([0, T], E)$ is not complete. Therefore, we define the following metric d^o, equivalent to d, for which $(\mathcal{D}([0, T], E), d^o)$ is complete:

$$d^o_T(x, y) = \inf_{\lambda \in \Lambda} \Big(\sup_{t \in [0,T]} r(x(\lambda(t)), y(t)) + \sup_{0 \leq s < t \leq T} \Big| \log \frac{\lambda(t) - \lambda(s)}{t - s} \Big| \Big).$$

Note that instead of requirering λ to be uniformly close to the identity, we require the slope of each chord to be near 1 or, what is equivalent, that its logarithm be close to 0.

We now extend this ideas to case $I = [0, \infty)$. A natural way to extend this topology to the case $I = [0, \infty)$ would be to require that $x_n \to x$ in $\mathcal{D}([0, \infty), E)$ if $d_T(x_n, x) \to 0$ for all $T > 0$. But consider the case $x_n = \mathbb{1}_{[0,1+1/n)}$ and $x = \mathbb{1}_{[0,1)}$. We have $d_2(x_n, x) \to 0$ but $d_1(x_n, x) = 1$. In fact, we have the following result:

Lemma A.2. *Let $t > s$. If $d_t(x_n, x) \to 0$ and if x is continuous at s, then $d_s(x_n, x) \to 0$.*

To fix with this problem, in case E is a vector space, one can transform functions in such a way that they become continuous at some point. More precisely, given $m > 0$, one can multiply a function $x \in \mathcal{D}([0, \infty), E)$ by the following function g_m in order to obtain a function that is continuous at m:

$$g_m(t) = \begin{cases} 1 \text{ if } t \leq m - 1 \\ m - t \text{ if } m - 1 \leq t \leq m \\ 0 \text{ if } t > m \end{cases}$$

This is the approach of [Bil99, Section 16], and one can define, on $\mathcal{D}([0, \infty), E)$,

$$d(x, y) = \sum_{m=1}^{\infty} 2^{-m} \max(d_m^o(xg_m, yg_m), 1), \ E \text{ being a vector space.} \tag{A.1.1}$$

We have the following caracterization of convergence.

Proposition A.3 ([Bil99, Theorem 16.2])**.** *Let $x, x_1, x_2, ... \in \mathcal{D}([0, T], E)$. Then, $d(x_n, x) \to 0$ if and only if $d_t(x_n, x) \to 0$ (or equivalently $d_t^o(x_n, x) \to 0$) for each continuity point t of x.*

In case (E, r) is only a metric space, we first introduce Λ_∞, the set of continuous increasing functions $\lambda : [0, \infty) \to [0, \infty)$ with $\lambda(0) = 0$ and $\lim \lambda(t) = \infty$ as $t \to \infty$. For $x, y \in \mathcal{D}([0, \infty), E)$, we define

$$d(x, y) = \inf_{\lambda \in \Lambda_\infty} \Big(\int_0^\infty e^{-u} d(x, y, \lambda, u) \mathrm{d}u + \sup_{0 \leq s < t \leq T} \Big| \log \frac{\lambda(t) - \lambda(s)}{t - s} \Big| \Big), \tag{A.1.2}$$

where, denoting by q the metric $r \wedge 1$,

$$d(x, y, \lambda, u) = \sup_{t \geq 0} q(x(t \wedge u), y(\lambda(t) \wedge u))$$

This last distance enjoys the following property, similar to Proposition A.1.

Proposition A.2 ([EK09, Proposition 5.3 in Chapter 3])**.** *Let $x, x_1, x_2, ... \in \mathcal{D}([0, \infty), E)$. Then, $d(x_n, x) \to 0$ if and only if for each $T > 0$, there exists a sequence $(\lambda_n^T)_{n \geq 1} \subset \Lambda_\infty$ such that*

$$\sup_{t \in [0,T]} |\lambda_n^T(t) - t| \to_{n \to \infty} 0$$

and

$$\sup_{t \in [0,T]} r(x_n(t), x(\lambda_n^T(t))) \to_{n \to \infty} 0.$$

Endowed with d defined either in (A.1.1) or (A.1.2), the space $\mathcal{D}([0, \infty), E)$ is separable and complete.

Random elements in $\mathcal{D}(I, E)$. Given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, a function $X : \Omega \to \mathcal{D}(I, E)$ is called a stochastic process with sample paths in $\mathcal{D}(I, E)$ if it is measurable and, for each $\omega \in \Omega$, $X(\omega) \in \mathcal{D}(I, E)$. The following property shows that for almost every time $t \geq 0$, such a random process is almost surely continuous a t. This remarkable result is crucial in our proofs (see for instance Proposition 2.21 in Chapter 2).

Proposition A.3 ([EK09, Lemma 3.7.7]). *Let X be a process with sample paths in $\mathcal{D}(I,E)$ with $I=[0,T]$ or $[0,+\infty)$. Then, the complement of the set*

$$\mathsf{C}(X)=\{t\in I,\ \mathbf{P}(X_t=X_{t^-})=1\}$$

is at most countable.

Note that the set $\mathsf{C}(X)$ depends only on the law of X. We now give an example of relative compactness criteria, in the case $I=[0,T]$ and E being a vector space. This kind of results is at the core of our proofs. Let us introduce the following modulus of continuity, designed for Skorohod spaces: for $0<\delta<1$,

$$w'_x(\delta)=\inf_{\{t_i\}}\max_{1\le i\le v}\sup_{s,t\in[t_{i-1},t_i)}|x(s)-x(t)|,$$

where the infimum extends over all δ-sparse sets $\{t_i\}$, i.e., which satisfy $0=t_0<t_1<\cdots<t_v=T$ and $t_i-t_{i-1}>\delta$ for all $i=1,\dots,v$.

Theorem A.4 ([Bil99, Theorem 13.2]). *Let $I=[0,T]$, $(P^N)_{N\ge1}\subset\mathcal{P}(\mathcal{D}(I,E))$ a sequence of probability measures on the Skorohod space $\mathcal{D}(I,E)$, E being a vector space. The sequence $(P^N)_{N\ge1}$ is relatively compact if and only if these two conditions hold:*

- *Compact containment condition:*

$$\lim_{a\to\infty}\limsup_N P^N\Big(x\in\mathcal{D}(I,E);\ \sup_{t\in[0,T]}|x(t)|\ge a\Big)=0.$$

- *Regularity condition: For each $\varepsilon>0$,*

$$\lim_{\delta\to0}\limsup_N P^N(x\in\mathcal{D}(I,E);w'_x(\delta)\ge\varepsilon)=0.$$

This result could be generalized to arbitrary metric spaces E and to $I=[0,\infty)$, see for instance [EK09, Theorem 8.6 in Chapter 3]. These Arzela-Ascoli type theorems are always composed by so called compact containment and regularity conditions. In particular, we proved a criterion in the case of Hilbert spaces having Hilbert-Schmidt embeddings (see Proposition 2.41 for the result and Appendix A.3 for a definition of Hilbert-Schmidt embeddings).

A.2 Wasserstein spaces and optimal transport

Our main references for this section are the famous books of Cédric Villani [Vil03] and [Vil09]. Given a metric space (X,d), separable and complete, we denote by $\mathcal{P}_p(X)$ the set of probability measures with finite moment of order $p\ge1$, i.e., probability measures $\mu\in\mathcal{P}(X)$ such that there exists $x_0\in X$ with

$$\int_X d(x,x_0)^p\mu(\mathrm{d}x)<\infty.$$

When d is bounded, one has, of course, $\mathcal{P}_p(X)=\mathcal{P}(X)$.
Given $\mu,\nu\in\mathcal{P}(X)$, let $\Pi(\mu,\nu)$ denote the set of probability measure over $X\times X$ whose first (resp. second) marginal is μ (resp. ν), i.e.,

$$\Pi(\mu,\nu)=\{\pi\in\mathcal{P}(X\times X),\ \pi(A\times X)=\mu(A),\ \pi(X\times A)=\nu(A),\forall\text{Borelian } A\subset X\}.$$

The Wasserstein distance of order $p\ge1$, is then defined as

$$\mathsf{W}_p(\mu,\nu)=\Big(\inf_{\pi\in\Pi(\mu,\nu)}\int_{X\times X}d(x,y)^p\pi(\mathrm{d}x,\mathrm{d}y)\Big)^{1/p},\ \mu,\nu\in\mathcal{P}_p(X).\tag{A.2.1}$$

This could be written in terms of random variables: considering the infimum over all random elements (U, V) (i.e., over all probability spaces $(\Omega, \mathcal{F}, \mathbf{P})$ and all measurable functions $(U, V) : \Omega \to X \times X$),

$$\mathsf{W}_p(\mu, \nu) = \inf\{\mathbf{E}[d(U, V)^p],\ U \sim \mu,\ V \sim \nu\}^{1/p}.$$

It is a non trivial fact that W_p defines a distance (see [Vil03, Theorem 7.3]). In case $p = 1$, we mention the so called Kantorovich-Rubinstein duality, which we use in our proofs (see for instance Proposition 3.23) :

$$\mathsf{W}_1(\mu, \nu) = \sup_{\|\phi\|_{\mathrm{Lip}} \leq 1} \Big[\int_X \phi(x)\mu(\mathrm{d}x) - \int_X \phi(x)\nu(\mathrm{d}x) \Big].$$

The minimization of (A.2.1) is a special case of Kantorovich's optimal transportation problem [Kan42, Kan48] (in which d^p is replaced by a nonnegative function c, called the *cost function*), which consists in minimizing the quantity

$$K[\pi] = \int_{X \times X} c(x, y)\pi(\mathrm{d}x, \mathrm{d}y) \tag{A.2.2}$$

over $\Pi(\mu, \nu)$, μ and ν being given. It is itself a generalized version of Monge's problem [Mon81], formulated in 1781, which remained unsolved at the time Kantorovich introduced his relaxed version of the problem. Monge's problem consists in minimizing

$$M[T] = \int_X c(x, T(x))\mu(\mathrm{d}x) \tag{A.2.3}$$

over all functions $T : X \to X$ such that $T_{\#}\mu = \nu$. It corresponds to minimizing $K[\pi]$ over

$$\Pi_M(\mu, \nu) = \{(Id \times T)_{\#}\mu,\ T_{\#}\mu = \nu\} \subset \Pi(\mu, \nu),$$

where $(Id \times T) : x \in X \mapsto (x, T(x)) \in X \times X$. We refer to [Vil09, Vil03] for an overview of problems (A.2.2) and (A.2.3).

A.3 Sobolev spaces

In this section, we introduce Sobolev spaces, that are extensively used in this thesis, to prove the central limit theorems. Our references for this section are [Eva22] and [AF03]. Let U be an open subset of $\mathbf{R}^n$, of class C^1. Sobolev spaces are subspaces of Lebesgue spaces $L^p(U)$, designed to weaken the classical notion of derivative.
Let us first introduce some notations. For $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbf{N}^n$ a multiindex we denote

$$D^\alpha u = \partial_1^{\alpha_1} \ldots \partial_n^{\alpha_n} u.$$

For $1 \leq p \leq \infty$ and any positive integer m, we define the following Sobolev norms:

$$\begin{cases} \|u\|_{W^{m,p}} = \Big(\sum_{0 \leq |\alpha| \leq m} \|D^\alpha u\|_{L^p}^p \Big)^{1/p}, \text{ if } 1 \leq p < \infty, \\ \|u\|_{W^{m,\infty}} = \max_{0 \leq |\alpha| \leq m} \|D^\alpha u\|_{L^\infty}, \end{cases} \tag{A.3.1}$$

for any function $u : U \to \mathbf{R}$ for which the right side makes sense. In the following definition, we leverage the classical integration by parts formula to introduce the notion of weak partial derivative.

Definition A.5. *Let $u, v \in L^1_{loc}(U)$ and α be a multiindex. We say that v is the α^{th}- weak partial derivative of u, written*

$$D^\alpha u = v$$

provided

$$\int_U u(x) D^\alpha \phi(x) \mathrm{d}x = (-1)^{|\alpha|} \int_U v(x)\phi(x)\mathrm{d}x, \tag{A.3.2}$$

for all test functions $\phi \in \mathcal{C}^\infty_c(U)$.

Note that there is no boundary term in the formula (A.3.2) since U is open and ϕ has compact support. We now define two function spaces, equiped with the norm (A.3.1):

Definition A.6. *For $1 \leq p \leq \infty$ and any positive integer m, we define the Sobolev space*

$$W^{m,p}(U)$$

as the set of all $u \in L^1_{loc}(U)$ such that for each multiindex α with $|\alpha| \leq m$, $D^\alpha u$ exists (in the sense of Definition A.5) and belongs to $L^p(U)$. We denote by

$$W^{m,p}_0(U)$$

the closure of $\mathcal{C}^\infty_c(U)$ in $W^{m,p}(U)$.

The spaces $W^{m,p}(U)$ and $W^{m,p}_0(U)$ are Banach spaces for each $1 \leq p \leq \infty$ and $m \geq 1$. When $p = 2$, it is usual to write $\mathcal{H}^m(U)$ (resp. $\mathcal{H}^m_0(U)$) in place of $W^{m,2}(U)$ (resp. $W^{m,2}_0(U)$).
In the following, we enounce conditions on p, m and the domain U to ensure regularity of elements of a Sobolev space. We start by recalling what we mean by "having a continuous representative".

Definition A.7. *We say $u^* : U \to \mathbf{R}$ is a version of a given function $u \in L^1_{loc}(U)$ provided*

$$u = u^* \ a.e.,$$

i.e., for any representative $\tilde{u} : U \to \mathbf{R}$ of u, there exists a measurable set $\mathscr{N} \subset U$ of mass 0 (with respect to the Lebesgue measure) such that

$$\tilde{u}(x) = u^*(x), \ \forall x \in \mathscr{N}^c.$$

We now introduce the Hölder spaces.

Definition A.8. *For any $0 < \gamma \leq 1$ and any nonnegative integer m, we define the Hölder space*

$$\mathcal{C}^{m,\gamma}_{\text{Holder}}(U)$$

as the space of all function $u \in \mathcal{C}^m(U)$ for which the norm

$$\|u\|_{\mathcal{C}^{m,\gamma}_{\text{Holder}}} = \sum_{|\alpha| \leq m} \sup_{x \in U} |D^\alpha u(x)| + \sup_{x,y \in U, x \neq y} \frac{|D^\alpha u(x) - D^\alpha u(y)|}{|x-y|^\gamma}$$

is finite.

We are now ready to state the continuous embedding theorem.

Theorem A.9 ([Eva22, Theorem 6 in Section 5.6.3] and [AF03, Theorem 4.12]). *Let $m \geq 1$ and $1 \leq p \leq \infty$ be such that $m > n/p$, and set*

$$\gamma = \begin{cases} \lfloor \frac{n}{p} \rfloor - \frac{n}{p} + 1, & \text{if } \frac{n}{p} \text{ is not an integer} \\ \text{any positive number} < 1, & \text{if } \frac{n}{p} \text{ is an integer.} \end{cases}$$

Then, we have the following.

- *If $U \subset \mathbf{R}^n$ is open and bounded with C^1-boundary, then any $u \in W^{m,p}(U)$ admits a version $u^* \in \mathcal{C}^{m-\lfloor \frac{n}{p} \rfloor - 1, \gamma}_{\text{Holder}}(U)$. Moreover, there exists a constant $C > 0$, independent of u, such that*

$$\|u^*\|_{\mathcal{C}^{m-\lfloor \frac{n}{p} \rfloor - 1, \gamma}_{\text{Holder}}} \leq C \|u\|_{W^{m,p}}.$$

  In other words, the embedding $W^{m,p}(U) \hookrightarrow \mathcal{C}^{m-\lfloor \frac{n}{p} \rfloor - 1, \gamma}_{\text{Holder}}(U)$ is continuous.

- *If $U \subset \mathbf{R}^n$ is arbitrary, then any $u \in W^{m,p}_0(U)$ admits a version $u^* \in \mathcal{C}^{m-\lfloor \frac{n}{p} \rfloor - 1, \gamma}_{\text{Holder}}(U)$. Moreover, there exists a constant $C > 0$, independent of u, such that*

$$\|u^*\|_{\mathcal{C}^{m-\lfloor \frac{n}{p} \rfloor - 1, \gamma}_{\text{Holder}}} \leq C \|u\|_{W^{m,p}}.$$

  In other words, the embedding $W^{m,p}_0(U) \hookrightarrow \mathcal{C}^{m-\lfloor \frac{n}{p} \rfloor - 1, \gamma}_{\text{Holder}}(U)$ is continuous.

In the following, we enounce some Hilbert-Schmidt embeddings between Sobolev space, which are a subclass of compact embeddings. We first define the notion of Hilbert-Schmidt operators. Let X and Y be two separable Hilbert spaces and let $\{e_i\}_{i=1}^{\infty}$ be an orthonormal basis of X. Let $A : X \to Y$ be a bounded linear operator. If

$$\sum_{i=1}^{\infty} \|Ae_i\|^2$$

is finite, A is called a Hilbert-Schmidt operator. We end this appendix by enouncing the following theorem, due to Maurin, extensively used in our proofs of the central limit theorems, to obtain moment estimates.

Theorem A.10 ([AF03, Theorem 6.61]). *Let m, j be nonnegative integers with $j > n/2$. Then, for any bounded domain $U \subset \mathbf{R}^n$, the embedding*

$$\mathcal{H}^{m+j}_0(U) \to \mathcal{H}^m_0(U)$$

is a Hilbert-Schmidt operator, which we write $\mathcal{H}^{m+j}_0(U) \hookrightarrow_{\text{H.S.}} \mathcal{H}^m_0(U)$. If U is a bounded domain of class $\mathcal{C}^2$, it also holds $\mathcal{H}^{m+j}(U) \hookrightarrow_{\text{H.S.}} \mathcal{H}^m(U)$.

Bibliography

[ABF+19] Giacomo Albi, Nicola Bellomo, Luisa Fermo, S-Y Ha, J Kim, Lorenzo Pareschi, David Poyato, and Juan Soler. Vehicular traffic, crowds, and swarms: From kinetic theory and multiscale methods to applications and research perspectives. *Mathematical Models and Methods in Applied Sciences*, 29(10):1901–2005, 2019.

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[AF03] R. Adams and J. Fournier. *Sobolev Spaces*. Elsevier, 2003.

[AG08] Luigi Ambrosio and Wilfrid Gangbo. Hamiltonian odes in the wasserstein space of probability measures. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(1):18–53, 2008.

[AGS05] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zurich. Birkhäuser Verlag, Basel, 2005.

[ALMV20] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020.

[AOY19] Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*, 2019.

[APH+21] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[Bar93] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

[BCKW15] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[BCN18] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[BFOZ18] Paolo Buttà, Franco Flandoli, Michela Ottobre, and Boguslaw Zegarlinski. A non-linear kinetic model of self-propelled particles with multiple equilibria. *arXiv preprint arXiv:1804.01247*, 2018.

[BGSR16] Thierry Bodineau, Isabelle Gallagher, and Laure Saint-Raymond. The brownian motion as the limit of a deterministic system of hard-spheres. *Inventiones mathematicae*, 203:493–553, 2016.

[Bil99] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2nd edition, 1999.

[BKM17] D.M. Blei, A. Kucukelbir, and J.D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[Car10] Pierre Cardaliaguet. Notes on mean field games. Technical report, Technical report, 2010.

[CB18a] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[CB18b] Lenaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

[CB20] Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1305–1338. PMLR, 09–12 Jul 2020.

[CCFRF22] Lénaïc Chizat, Maria Colombo, Xavier Fernandez-Real, and Alessio Figalli. Infinite-width limit of deep linear neural networks, 2022.

[CD+18] René Carmona, François Delarue, et al. *Probabilistic theory of mean field games with applications I-II*. Springer, 2018.

[CD22] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. i. models and methods. *arXiv preprint arXiv:2203.00446*, 2022.

[CDLL19] P. Cardaliaguet, F. Delarue, J.M. Lasry, and P.L. Lions. *The Master Equation and The Convergence Problem in Mean Field Games*. Number 201. Annals of Mathematical Studies. Princeton University Press, 2019.

[Chi22] Lénaïc Chizat. Mean-field langevin dynamics: Exponential convergence and annealing, 2022.

[CJ21] Adam D Cobb and Brian Jalaian. Scaling hamiltonian monte carlo inference for bayesian neural networks with symmetric splitting. In *Uncertainty in Artificial Intelligence*, pages 675–685. PMLR, 2021.

[CJLZ21] José A Carrillo, Shi Jin, Lei Li, and Yuhua Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 27:S5, 2021.

[CPDV21] Beau Coker, Weiwei Pan, and Finale Doshi-Velez. Wide mean-field variational bayesian neural networks ignore the data. *arXiv preprint arXiv:2106.07052*, 2021.

[CPT11] Emiliano Cristiani, Benedetto Piccoli, and Andrea Tosin. Multiscale modeling of granular flows with application to crowd dynamics. *Multiscale Modeling & Simulation*, 9(1):155–182, 2011.

[CRBVE20] Z. Chen, G.M. Rotskoff, J. Bruna, and E. Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22217–22230. Curran Associates, Inc., 2020.

[CVEB22] Zhengdao Chen, Eric Vanden-Eijnden, and Joan Bruna. A functional-space mean-field theory of partially-trained three-layer neural networks. *arXiv preprint arXiv:2210.16286*, 2022.

[DBDFS20] V. De Bortoli, A. Durmus, X. Fontaine, and U. Simsekli. Quantitative propagation of chaos for SGD in wide neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 278–288. Curran Associates, Inc., 2020.

[DGMN22] A. Descours, A. Guillin, M. Michel, and B. Nectoux. Law of large numbers and central limit theorem for wide two-layer neural networks: the mini-batch and noisy case. *arXiv preprint arXiv:2207.12734*, 2022.

[DGW04] H. Djellout, A. Guillin, and L. Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *The Annals of Probability*, 32(3B):2702 – 2732, 2004.

[DHG+23] A. Descours, T. Huix, A. Guillin, M. Michel, É. Moulines, and B. Nectoux. Law of large numbers for bayesian two-layer neural network trained with variational inference. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4657–4695. PMLR, 12–15 Jul 2023.

[DL18] Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1329–1338. PMLR, 10–15 Jul 2018.

[DLR19a] F. Delarue, D. Lacker, and K. Ramanan. From the master equation to mean field game limit theory: a central limit theorem. *Electronic Journal of Probability*, 24:1–54, 2019.

[DLR19b] François Delarue, Daniel Lacker, and Kavita Ramanan. From the master equation to mean field game limit theory: a central limit theorem. 2019.

[DLR20] François Delarue, Daniel Lacker, and Kavita Ramanan. From the master equation to mean field game limit theory: Large deviations and concentration of measure. *The Annals of Probability*, 48(1):211 – 263, 2020.

[DMG99] P. Del Moral and A. Guionnet. Central limit theorem for nonlinear filtering and interacting particle systems. *The Annals of Applied Probability*, 9(2):275–297, 1999.

[dRBK20] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.

[E20] W. E. Machine learning and computational mathematics. *Communications in Computational Physics*, 28(5):1639–1670, 2020.

[EK09] S. Ethier and T. Kurtz. *Markov Processes: Characterization and Convergence*, volume 282. John Wiley & Sons, 2009.

[Eva22] Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

[FFG+19] Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

[FG15] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

[FM97a] Begona Fernandez and Sylvie Méléard. A hilbertian approach for fluctuations on the mckean-vlasov model. *Stochastic processes and their applications*, 71(1):33–53, 1997.

[FM97b] Begoña Fernandez and Sylvie Méléard. A hilbertian approach for fluctuations on the mckean-vlasov model. *Stochastic Processes and their Applications*, 71(1):33–53, 1997.

[FRF22] Xavier Fernández-Real and Alessio Figalli. The continuous formulation of shallow neural networks as wasserstein-type gradient flows. In *Analysis at Large: Dedicated to the Life and Work of Jean Bourgain*, pages 29–57. Springer, 2022.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT press, 2016.

[GG16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[Gha15] Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

[GKM+96] Carl Graham, Thomas G Kurtz, Sylvie Méléard, Philip E Protter, Mario Pulvirenti, Denis Talay, and Sylvie Méléard. Asymptotic behaviour of some interacting particle systems; mckean-vlasov and boltzmann models. *Probabilistic Models for Nonlinear Partial Differential Equations: Lectures given at the 1st Session of the Centro Internazionale Matematico Estivo (CIME) held in Montecatini Terme, Italy, May 22–30, 1995*, pages 42–95, 1996.

[GKNV22] Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks. *Constructive approximation*, 55(1):259–367, 2022.

[GLQ+19] R.M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General analysis and improved rates. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 09–15 Jun 2019.

[Gol16] François Golse. On the dynamics of large particle systems in the mean field limit. *Macroscopic and large scale phenomena: coarse graining, mean field limits and ergodicity*, pages 1–144, 2016.

[Got98] Alexander David Gottlieb. *Markov transitions and the propagation of chaos*. University of California, Berkeley, 1998.

[GP21] Sara Grassi and Lorenzo Pareschi. From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 31(08):1625–1657, 2021.

[Gra11] A. Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

[GSRT13] Isabelle Gallagher, Laure Saint-Raymond, and Benjamin Texier. *From Newton to Boltzmann: hard spheres and short-range potentials*. European Mathematical Society Zürich, Switzerland, 2013.

[HC93] Geoffrey Hinton and Drew Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, pages 5–13. ACM Press, 1993.

[HJ14] Maxime Hauray and Pierre-Emmanuel Jabin. Particles approximations of vlasov equations with singular forces: Propagation of chaos. *preprint, arXiv1107*, 382, 2014.

[HM86] M. Hitsuda and I. Mitoma. Tightness problem and stochastic evolution equation arising from fluctuation phenomena for interacting diffusions. *Journal of Multivariate Analysis*, 19(2):311–328, 1986.

[HMD+22] Tom Huix, Szymon Majewski, Alain Durmus, Eric Moulines, and Anna Korba. Variational inference of overparameterized bayesian neural networks: a theoretical and empirical study, 2022.

[HRSS21] K. Hu, Z. Ren, D. Siska, and L. Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(4):2043 – 2065, 2021.

[IVHW21] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are bayesian neural network posteriors really like? *International Conference on Machine Learning*, 2021.

[Jak86] A. Jakubowski. On the skorokhod topology. In *Annales de l'IHP Probabilités et statistiques*, volume 22, pages 263–285, 1986.

[JLB+22] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[JM98a] B. Jourdain and S. Méléard. Propagation of chaos and fluctuations for a moderate model with smooth initial data. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 34(6):727–766, 1998.

[JM98b] Benjamin Jourdain and Sylvie Méléard. Propagation of chaos and fluctuations for a moderate model with smooth initial data. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 34, pages 727–766. Elsevier, 1998.

[JMM20] Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *The Annals of Statistics*, 48(6):3619 – 3642, 2020.

[JNG+21] C. Jin, P. Netrapalli, R. Ge, S. Kakade, and M. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

[JS87] J. Jacod and A. Shiryaev. *Skorokhod Topology and Convergence of Processes.* Springer, 1987.

[JT21] Benjamin Jourdain and Alvin Tse. Central limit theorem over non-linear functionals of empirical measures with applications to the mean-field fluctuation of interacting diffusions. *Electronic Journal of Probability*, 26:1–34, 2021.

[KA13] M. Krzywinski and N. Altman. Importance of being uncertain. *Nature methods*, 10(9):809–811, 2013.

[Kac56] Mark Kac. Foundations of kinetic theory. In *Proceedings of The third Berkeley symposium on mathematical statistics and probability*, volume 3, pages 171–197, 1956.

[Kan42] Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[Kan48] Leonid V Kantorovich. On a problem of monge. In *CR (Doklady) Acad. Sci. URSS (NS)*, volume 3, pages 225–226, 1948.

[KG17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[KMN+17] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[KNT+18] Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. In *International Conference on Machine Learning*, pages 2611–2620. PMLR, 2018.

[Kur75] T. Kurtz. Semigroups of conditioned shifts and approximation of Markov processes. *The Annals of Probability*, pages 618–642, 1975.

[KW14] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

[KX04] T. Kurtz and J. Xiong. A stochastic evolution equation arising from the fluctuations of a class of interacting particle systems. *Communications in Mathematical Sciences*, 2(3):325–358, 2004.

[LP17] Dustin Lazarovici and Peter Pickl. A mean field limit for the vlasov–poisson system. *Archive for Rational Mechanics and Analysis*, 225:1201–1231, 2017.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[LW17] Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *International Conference on Machine Learning*, pages 2218–2227. PMLR, 2017.

[M+95] David JC MacKay et al. Ensemble learning and evidence maximization. In *Proc. Nips*, volume 10, page 4083. Citeseer, 1995.

[Mac95] David JC MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469, 1995.

[MGK+17] Rowan McAllister, Yarin Gal, Alex Kendall, Mark van der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. In *IJCAI*, 2017.

[MJ69] HP McKean Jr. Propagation of chaos for a class of nonlinear parabolic equations, inlecture series in differential equations, vol. 2, 1969.

[MM13] Stéphane Mischler and Clément Mouhot. Kac's program in kinetic theory. *Inventiones mathematicae*, 193:1–147, 2013.

[MMM19] S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

[MMN18] S. Mei, A. Montanari, and P-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

[MT14] Adrian Muntean and Federico Toschi. *Collective dynamics from bacteria to crowds: an excursion through modeling, analysis and simulation*, volume 553. Springer Science & Business Media, 2014.

[MWL+20] Rhiannon Michelmore, Matthew Wicker, Luca Laurenti, Luca Cardelli, Yarin Gal, and Marta Kwiatkowska. Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7344–7350. IEEE, 2020.

[MWW+20] C. Ma, S. Wojtowytsch, L. Wu, et al. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *CSIAM Transactions on Applied Mathematics*, 1(4):561–615, 2020.

[Ngu19] Phan-Minh Nguyen. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv preprint arXiv:1902.02880*, 2019.

[NP20] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv preprint arXiv:2001.11443*, 2020.

[NPT10] Giovanni Naldi, Lorenzo Pareschi, and Giuseppe Toscani. *Mathematical modeling of collective behavior in socio-economic and life sciences*. Springer Science & Business Media, 2010.

[OSK+19] Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with bayesian principles. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[PR13] Benedetto Piccoli and Francesco Rossi. Transport equation with nonlocal velocity in wasserstein spaces: convergence of numerical schemes. *Acta applicandae mathematicae*, 124:73–105, 2013.

[PR16] B. Piccoli and F. Rossi. On properties of the generalized Wasserstein distance. *Archive for Rational Mechanics and Analysis*, 222(3):1339–1365, 2016.

[PRT15] B. Piccoli, F. Rossi, and E. Trélat. Control to flocking of the kinetic Cucker–Smale model. *SIAM Journal on Mathematical Analysis*, 47(6):4685–4719, 2015.

[PTTM17] René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 27(01):183–204, 2017.

[PZ20] V.M. Panaretos and Y. Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer Nature, 2020.

[RCC99] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[RVE18a] G.M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *Preprint arXiv:1805.00915, to appear in Comm. Pure App. Math.*, 2018.

[RVE18b] Grant M Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.

[RVE22] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

[San15] F. Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 55. Springer, 2015.

[SGN+19] U. Simsekli, M. Gurbuzbalaban, T. Nguyen, G. Richard, and L. Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks, 2019.

[SJL19] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

[SL18] S. L. Smith and Q. V. Le. A bayesian perspective on generalization and stochastic gradient descent. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[Spo12] Herbert Spohn. *Large scale dynamics of interacting particles*. Springer Science & Business Media, 2012.

[SS20a] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

[SS20b] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[SS22] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.

[Szn91] A-S. Sznitman. Topics in propagation of chaos. In *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251. Springer, 1991.

[Vil03] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[Vil09] C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[VRF22] Loucas Pillaud Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2127–2159. PMLR, 02–05 Jul 2022.

[WHX+20] J. Wu, W. Hu, H. Xiong, J. Huan, V. Braverman, and Z. Zhu. On the noisy gradient descent that generalizes as SGD. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10367–10376. PMLR, 13–18 Jul 2020.

[Woj20] Stephan Wojtowytsch. On the convergence of gradient descent training for two-layer relu-networks in the mean field regime. *arXiv preprint arXiv:2005.13530*, 2020.

[WRV+20] Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10248–10259. PMLR, 13–18 Jul 2020.

[YW71] Toshio Yamada and Shinzo Watanabe. On the uniqueness of solutions of stochastic differential equations. *Journal of Mathematics of Kyoto University*, 11(1):155–167, 1971.

[ZSDG18] Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5852–5861. PMLR, 10–15 Jul 2018.