Adaptive Gradient Langevin Algorithms for Stochastic Optimization and Bayesian Inference

Pierre Bras

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

LPSM
LABORATOIRE DE PROBABILITÉS STATISTIQUE & MODÉLISATION

THÈSE DE DOCTORAT

en vue de l'obtention du grade de

Docteur de SORBONNE UNIVERSITÉ

Discipline : **Mathématiques et Applications**

Laboratoire de Probabilités, Statistique et Modélisation – UMR 8001

École Doctorale de Sciences Mathématiques de Paris Centre – ED 386

Présentée par
Pierre BRAS

Adaptive Gradient Langevin Algorithms for Stochastic Optimization and Bayesian Inference

---

sous la direction de Gilles Pagès

Rapportée par: M. Éric Moulines *École Polytechnique*
M. Mike Giles *University of Oxford*

Soutenue le **11 Septembre 2023** devant le jury composé de:

M.	Francis Bach	*INRIA Paris*	Président du Jury
M.	Gilles Pagès	*Sorbonne Université*	Directeur
M.	Éric Moulines	*École Polytechnique*	Rapporteur
M.	Mike Giles	*University of Oxford*	Rapporteur
M.	Olivier Pironneau	*Sorbonne Université*	Examinateur
Mme.	Gersende Fort	*Institut de Mathématiques de Toulouse*	Examinatrice
M.	Nicolas Fournier	*Sorbonne Université*	Examinateur
M.	Josef Teichmann	*ETH Zürich*	Examinateur

Sorbonne Université
Laboratoire de Probabilités, Statistique
et Modélisation
UMR 8001, case 158
4 Place Jussieu
F-75252 Paris Cedex 5, France

École doctorale de Sciences
Mathématiques de Paris Centre
4 Place Jussieu
F-75252 Paris Cedex 5, France

Remerciements

Je remercie tout d'abord mon directeur de thèse, Gilles Pagès, pour m'avoir accepté en tant que doctorant et pour m'avoir guidé et accompagné pendant ces trois années. J'ai beaucoup appris sous sa direction et ses précieux conseils et remarques m'ont beaucoup aidé à prendre du recul et à organiser mes recherches. Tout au long de ma thèse, j'ai été impressionné par ses larges connaissances et son expertise dans de très nombreux aspects des probabilités numériques et appliquées, et il avait (presque) toujours réponse à mes questions. Je me souviens de tous les moments où je toquais à sa porte à l'improviste et malgré toutes ses responsabilités et occupations, mon professeur trouvait toujours un moment de libre pour discuter.

Je remercie ensuite les rapporteurs de cette thèse Éric Moulines et Mike Giles, pour avoir accepté cette charge. Leur rapport et leurs commentaires pertinents m'ont grandement aidé à améliorer la qualité de ce manuscrit. Je remercie également les autres membres du jury, Francis Bach, Olivier Pironneau, Gersende Fort, Nicolas Fournier et Josef Teichmann, pour l'honneur qu'ils me font en participant au jury de ma soutenance. Leur expertise et leur expérience dans leur domaine sont extrêmement précieuses pour évaluer mes travaux.

Durant ma thèse, j'ai pu bénéficier de l'accueil chaleureux et du soutien que m'ont témoigné les membres du Laboratoire de Probabilités, Statistique et Modélisation (LPSM) de Sorbonne Université. Je remercie particulièrement Vincent Lemaire et Idris Kharroubi pour nos discussions mathématiques et pour avoir répondu à mes questions sur leurs articles, et Daphné Giorgi pour organiser le séminaire Infomath et pour son aide précieuse en informatique. Je remercie Nicolas Fournier et encore une fois Idris Kharroubi pour leur participation à mon comité de thèse en début de troisième année. Je remercie Olivier Pironneau, professeur au laboratoire voisin Jacques-Louis Lion, pour nos échanges sur un sujet commun qui m'ont été très utiles pour la rédaction d'un article s'inspirant d'un de ses travaux.

Je remercie Fabien Panloup, professeur à l'Université d'Angers, pour avoir accepté de co-écrire un article avec Gilles Pagès et moi-même. Sa collaboration a été très précieuse pour établir un résultat important pour ma thèse.

Je remercie Arturo Kohatsu-Higa, professeur à Ritsumeikan University, pour m'avoir encadré pour un stage de recherche de six mois au Japon préalable à ma thèse, mais aussi pour s'être particulièrement occupé de moi et m'avoir accompagné pendant cette période. Grâce à lui, j'ai beaucoup appris sur la recherche mathématique et l'article que nous avons écrit ensemble fait également partie de ce manuscrit.

Je remercie Masaaki Fukasawa, professeur à l'université d'Osaka, pour m'avoir accueilli dans son laboratoire pour un stage de recherche d'été. Je suis heureux d'avoir collaboré avec lui pour la d'un rédaction d'un article, que j'ai aussi inclus dans cette thèse.

Toutes ces rencontres et échanges ont été très enrichissants pour moi et m'ont fait prendre

conscience que l'apprentissage machine et la simulation numérique sont des domaines très larges impliquant et mélangeant un large éventail de compétences et de spécialistes.

Pendant ces trois années au LPSM, j'ai eu la chance d'être entouré de collègues et amis doctorants sans qui cette aventure n'aurait pas été possible, et je souhaite les remercier tout particulièrement, pour nos conversations passionnantes à propos de probabilités et statistiques qui m'ont aidé à plusieurs reprises quand je bloquais sur un sujet, mais aussi pour tous les autres moments que nous avons partagés. Leur bonne humeur ont contribué à faire de ce laboratoire un endroit agréable au quotidien, malgré certaines périodes de confinements durant lesquelles nous n'avions pu nous voir qu'épisodiquement.

Je commencerai par les membres de mon bureau, je voudrais dire un grand merci à Loïc, que je voyais presque tout le temps au bureau, Christian, mon petit frère de thèse avec qui j'ai pu beaucoup échanger sur des sujets de recherche communs, Jean-David et ses nombreux récits d'expérience aux quatre coins du monde, Arthur, Lucas I. pour son aide et conseils sur le fonctionnement du laboratoire et séminaires en tant que "grand frère" et ancien du bureau. Je continue avec les doctorants des autres bureaux du couloir et du premier étage, en remerciant Robin et Yoan qui sont mes amis depuis bien avant la thèse et mes compagnons de l'ENS, mon frère de thèse Guillaume, Lucas B., Lucas D. (beaucoup trop de Lucas), Émilien, Jérôme, Sonia, David, Antonio, Alexandra, Bastien, Florian, Sergi, William et Isao. Enfin tous les doctorants que je n'ai pu connaître que plus brièvement lors de mon stage de pré-thèse au LPSM, en particulier mes grands frère et soeurs de thèse Yating, Thibault et Rancy. Certains d'entre vous ne sont plus au labo depuis un bout de temps mais vous faites (et ferez toujours) partie de l'histoire du LPSM et j'espère que vous lirez ces remerciements.

Ces trois dernières années ont aussi été marquées par des retours réguliers à Montpellier. Je remercie mes amis d'enfance et de lycée, en particulier Sylvain et Arthur que je connais depuis longtemps. Merci Arthur pour nos bavardages sur Messenger, nos sorties à Montpellier et à Sète, nos soirées sur Apex Legends et tout le reste.

Je remercie ma famille, Papa et Maman qui m'ont toujours encouragé dans mes études et apporté leur soutien inconditionnel dans tous mes choix, et surtout pendant cette thèse. Merci de vous occuper encore de moi, merci pour toutes les fois où vous êtes venus m'accompagner ou me chercher à la gare à cinq heures du matin comme à minuit. Merci à Mamie, toujours aussi bienveillante, je prends beaucoup de plaisir à te revoir chaque fois à Montpellier. Je remercie ma sœur Louise et Thibault, quand on se retrouve à Montpellier mais aussi pour chaque occasion où vous m'avez invité dans votre appartement parisien. Merci à ma tante Martine à Andrésy au bout du RER qui m'accueille de temps en temps depuis la prépa.

Et pour finir, merci à toi Yiming, tu es très importante dans ma vie et en rencontrant les mêmes difficultés de la thèse nous avons pu nous soutenir mutuellement, surtout dans les moments compliqués. Sans toi je n'aurais jamais pu aller aussi loin.

Résumé

Nous étudions les algorithmes adaptatifs de descente de gradient par dynamique de Langevin (SGLD) pour résoudre des problèmes d'optimisation et d'inférence. Les algorithmes SGLD consistent en une descente de gradient avec ajout de bruit exogène dans le but d'échapper aux minima locaux et aux points selle. Contrairement à l'équation différentielle stochastique (EDS) de Langevin classique, nous nous concentrons sur le cas où le bruit exogène est adaptatif i.e. non constant et dépend de la position de la procédure, donnant une convergence plus rapide que les algorithmes non adaptatifs. Bien que le cas constant ait été largement étudié, peu d'attention a été portée jusqu'à présent au cas général et la littérature manque d'un résultat théorique général de convergence.

Dans une première partie, nous prouvons la convergence de ces algorithmes pour la distance de Wasserstein L^1 et pour la distance de la variation totale, à la fois pour l'EDS continue et pour l'algorithme discret avec des mesures de gradient bruitées. Nous nous intéressons également aux algorithmes de Langevin-recuit simulé, où le bruit décroît lentement vers zéro au cours du temps à une vitesse appropriée. Nous investiguons aussi le cadre "dégénéré" i.e. où la matrice Hessienne en le minimum n'est pas définie positive, un aspect qui a été mis de côté par la littérature.

Dans une seconde partie nous appliquons les algorithmes SGLD à des problèmes d'optimisation et d'inférence apparaîssant en apprentissage machine et en probabilités numériques et nous comparons les performances de divers algorithmes de Langevin préconditionnés (adaptatifs) avec leurs équivalents respectifs non-Langevin. Nous observons que les algorithmes de Langevin améliorent la procédure d'entraînement pour des réseaux de neurones artificiels très profonds et que plus le réseau est profond, plus les gains apportés par les algorithmes de Langevin sont importants. Suivant cette heuristique nous introduisons une nouvelle variante des algorithmes de Langevin appelée "Langevin par couches", qui ajoute du bruit de Langevin sur seulement les couches les plus profondes du réseaux. Nous montrons les avantages des algorithmes de Langevin et de Langevin par couches pour l'entraînement d'architectures profondes en reconnaissance d'image (ResNet, DenseNet) et en contrôle stochastique (réseaux Markoviens).

Une dernière partie est consacrée à la simulation numérique de processus stochastiques. Nous démontrons des bornes pour la distance en variation totale entre une EDS et son schéma d'Euler-Maruyama en temps court, en utilisant une extrapolation de Richardson-Romberg pondérée. Ce résultat est crucial pour l'analyse de la convergence en variation totale des algorithmes de Langevin mentionnés ci-dessus. En utilisant l'analyse trajectorielle, nous étudions le taux d'erreur faible du schéma d'Euler-Maruyama pour les équations de Volterra stochastiques (EVSs), qui sont des équations différentielles stochastiques non Markoviennes avec un noyau de mémoire, tout en gardant à l'esprit le cas des modèles à volatilité rugueuse. Enfin, nous donnons des formules et des méthodes de simulation pour le mouvement Brownien réfléchi ou arrêté dans un cône en deux dimensions.

Mots clé– Optimisation Stochastique, Descente de Gradient, Equation de Langevin, Dynamique de Langevin, Recuit Simulé, Apprentissage Machine, Apprentissage Profond, Réseaux de Neurones, Schéma d'Euler-Maruyama, Mesures de Gibbs, Contrôle Stochastique, Méthodes de Monte Carlo, Méthodes MCMC.

Abstract

This thesis focuses on adaptive Stochastic Gradient Langevin Dynamics (SGLD) algorithms to solve optimization and Bayesian inference problems. SGLD algorithms consist in a stochastic gradient descent with exogenous noise added in order to escape local minima and saddle points. Contrary to the classic Langevin Stochastic Differential Equation (SDE), we study the case where the exogenous noise is adaptive i.e. not constant but depends on the position of the procedure, then yielding a convergence faster than non-adaptive algorithms. Although the constant case has been extensively studied, little attention has been paid so far to the general case and the literature lacks of a general theoretical convergence result.

In a first part we prove the convergence of SGLD algorithms for the L^1-Wasserstein distance and for the Total Variation distance for both the continuous stochastic differential equation and the discrete algorithm with noisy gradient measurements. We also study Langevin-simulated annealing algorithms, where the noise is slowly decreased to zero at an appropriate rate with time. We investigate degenerate settings as well i.e. when the Hessian matrix at the minimum is not definite, which is an aspect that had been put aside by the literature.

In a second part we apply SGLD algorithms to optimization and inference problems arising in Machine Learning and in Numerical Probability and we compare the performances of various preconditioned (adaptive) Langevin algorithms with their non-Langevin counterparts. We observe that Langevin algorithms improve the training procedure for very deep artificial neural networks and that the deeper a network is, the greater are the gains brought by Langevin algorithms. Following these heuristics, we introduce a new variant of Langevin algorithms called Layer Langevin, which adds Langevin noise only to the deepest layers of the network. We show the benefits of Langevin and Layer Langevin algorithms for the training of very deep architectures in image recognition (ResNet, DenseNet) and in stochastic optimal control (Markovian networks).

A last part is devoted to the numerical simulation of stochastic processes. We give bounds for the Total Variation distance between an SDE and its Euler-Maruyama scheme in small time, using a weighted multi-level Richardson-Romberg extrapolation. This result is crucial for the analysis of convergence in Total Variation of the above Langevin algorithms. Using path-wise analysis, we study the weak error rate for the Euler-Maruyama scheme for Stochastic Volterra Equations (SVEs) which are non-Markovian stochastic differential equations with memory kernel, while keeping in mind the case of rough volatility models. Lastly, we give density formulae and simulation methods for the reflected and stopped Brownian motion in a two-dimensional wedge.

Keywords– Stochastic Optimization, Gradient Descent, Langevin Equation, Langevin Dynamics, Simulated Annealing, Machine Learning, Deep Learning, Neural Networks, Euler-Maruyama Scheme, Gibbs Measures, Bayesian Inference, Stochastic Optimal Control, Monte Carlo Methods, MCMC Methods.

Contents

1 Introduction **3**
1.1 Stochastic algorithms . . . 3
1.1.1 Stochastic gradient descent and the Robbins-Siegmund lemma . . . 3
1.1.2 Classic examples of stochastic optimization problems . . . 4
1.1.3 Stochastic processes and the Euler-Maruyama scheme . . . 10
1.1.4 Bayesian inference and sampling from probability measures . . . 13
1.2 Langevin equation and Langevin algorithms . . . 15
1.2.1 The Langevin Equation . . . 15
1.2.2 Adaptive Langevin algorithms . . . 17
1.2.3 Langevin-Simulated anneling equation . . . 19
1.3 Contributions of the thesis . . . 20
1.3.1 Convergence of Langevin-Simulated annealing algorithms . . . 21
1.3.2 Adaptive Langevin algorithms for deep Neural Networks . . . 22
1.3.3 Simulation of stochastic processes and discretization schemes . . . 24

I Convergence of adaptive Langevin-Simulated Annealing algorithms **29**

2 Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise for the L^1-Wasserstein distance **31**
2.1 Introduction . . . 32
2.2 Assumptions and main results . . . 35
2.2.1 Assumptions . . . 35
2.2.2 Main results . . . 37
2.2.3 The degenerate case . . . 38
2.3 Application to optimization problems . . . 39
2.3.1 Potential function associated to a Neural Regression Problem . . . 39
2.3.2 Practitioner's corner: choices for σ . . . 40
2.4 Langevin equation with constant time coefficient . . . 41
2.4.1 Exponential contraction property . . . 41
2.4.2 Time schedule and Wasserstein distance between Gibbs measures . . . 44
2.5 Plateau case . . . 45
2.6 Continuously decreasing case . . . 47
2.6.1 Boundedness of the potential . . . 48
2.6.2 Strong and weak error bounds . . . 48
2.6.3 Proof of Theorem 2.2.1.(a) . . . 51
2.7 Continuously decreasing case : the Euler-Maruyama scheme . . . 53
2.7.1 Boundedness of the potential . . . 54
2.7.2 Strong and weak error bounds for the Euler-Maruyama scheme . . . 55
2.7.3 Proof of Theorem 2.2.1.(b) . . . 56

2.8 Convergence of the Euler-Maruyama scheme with plateau . . . 59
2.9 Experiments . . . 59
2.10 Conclusion and perspectives of future work . . . 61
2.11 Appendix . . . 62
2.12 Supplementary Material . . . 63
2.12.1 Proof of Proposition 2.4.4 . . . 63
2.12.2 Proof of Proposition 2.7.3 . . . 65
2.12.3 Proof of Theorem 2.2.4 . . . 67
2.12.4 Proof of Theorem 2.5.1 . . . 68

3 Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise II: Total Variation 69
3.1 Introduction . . . 69
3.2 Assumptions and main results . . . 71
3.2.1 Assumptions . . . 71
3.2.2 Main results . . . 73
3.2.3 Extensions and interpolations of the processes . . . 74
3.3 Bounds in total variation for small t . . . 75
3.3.1 Total variation bound in small time for the Euler-Maruyama scheme . . . 75
3.3.2 Total variation bound in small time for the continuous SDE . . . 77
3.4 Convergence of the plateau SDE X_t in total variation . . . 78
3.4.1 Exponential contraction in total variation . . . 78
3.4.2 Convergence of the plateau SDE . . . 80
3.5 Convergence of Y_t in total variation . . . 81
3.5.1 Preliminary lemmas . . . 81
3.5.2 Proof of Theorem 3.2.1(a) . . . 82
3.6 Convergence of the Euler-Maruyama scheme in total variation . . . 83
3.6.1 Preliminary lemmas . . . 84
3.6.2 Proof of Theorem 3.2.1(b) . . . 84
3.7 Experiments . . . 85
3.8 Appendix . . . 86
3.8.1 Proof of Proposition 3.4.2 . . . 86

4 Convergence rates of Gibbs measures with degenerate minimum 89
4.1 Introduction . . . 89
4.2 Definitions and notations . . . 92
4.3 Convergence of Gibbs measures . . . 94
4.3.1 Properties of Gibbs measures . . . 94
4.3.2 Statement of the problem . . . 95
4.3.3 Main results : rate of convergence of Gibbs measures . . . 96
4.4 Expansion of f at a local minimum with degenerate derivatives . . . 99
4.4.1 Expansion of f for any order p . . . 99
4.4.2 Review of the one dimensional case . . . 101
4.4.3 Proof of Theorem 4.4.1 for $p = 1$ . . . 101
4.4.4 Proof of Theorem 4.4.1 for $p = 2$ . . . 101
4.4.5 Difficulties beyond the 4th order and Hilbert's 17^{th} problem . . . 102
4.4.6 Proof of Theorem 4.4.1 for $p = 3$ . . . 103
4.4.7 Proof of Theorem 4.4.1 for $p = 4$ . . . 106
4.4.8 Counter-example and proof of Theorem 4.4.1 with $p \geq 5$ under the hypothesis (4.4.6) . . . 107

4.4.9 Proofs of the uniform convergence and of the non-constant property . . . 109
4.4.10 Non coercive case . . . 110
4.5 Proofs of Theorem 4.3.3 and Theorem 4.3.5 using Theorem 4.4.1 . . . 112
4.5.1 Single well case . . . 112
4.5.2 Multiple well case . . . 114
4.6 Infinitely flat minimum . . . 114
4.7 Simulations: computing high-order expansion of the loss with singular Hessian matrix . . . 115
4.8 Appendix: Properties of tensors . . . 116

II Adaptive Langevin algorithms for deep Neural Networks 119

5 Stochastic algorithms for artificial neural networks with simulations in TensorFlow 121
5.1 Artificial neural networks . . . 121
5.1.1 Calibration of artificial neural networks as a stochastic optimization problem 121
5.1.2 Neural networks architectures for tasks other than regression . . . 122
5.1.3 Neural networks, stochastic gradient and automatic differentiation . . . 123
5.2 Stochastic gradient algorithms . . . 125
5.2.1 Adaptive stochastic algorithms . . . 125
5.2.2 Gradient optimizers in TensorFlow . . . 127
5.3 Simulations . . . 129
5.3.1 Double well potential . . . 129
5.3.2 Bayesian Logistic Regression . . . 130
5.3.3 Image classification (1) . . . 130
5.3.4 Image classification (2) . . . 132
5.3.5 Time series analysis and prediction . . . 132
5.3.6 Optimal Quantization . . . 134

6 Langevin algorithms for very deep Neural Networks with application to image classification 137
6.1 Introduction . . . 137
6.2 Very deep neural networks . . . 138
6.3 Langevin algorithms for the training of deep neural networks . . . 139
6.3.1 Experimental setting . . . 139
6.3.2 Plain and convolutional networks . . . 139
6.3.3 Highway networks . . . 140
6.4 Layer Langevin algorithm . . . 143
6.5 Application to deep architectures for image recognition . . . 144

7 Langevin algorithms for Markovian Neural Networks and Deep Stochastic control 151
7.1 Introduction . . . 151
7.2 Stochastic control through gradient descent . . . 153
7.2.1 Stochastic optimal control . . . 153
7.2.2 Preconditioned stochastic gradient Langevin dynamics . . . 154
7.2.3 Experimental setting . . . 155
7.3 Fishing quotas . . . 156
7.4 Deep hedging . . . 159

7.5 Resource management . . . 160
7.6 Comments on the numerical experiments . . . 163

III Numerical simulation of stochastic processes 165

8 Total variation distance between two diffusions in small time with unbounded drift: application to the Euler-Maruyama scheme 167
8.1 Introduction . . . 167
8.2 Main results . . . 170
8.3 Proof of the Theorems . . . 173
8.3.1 Recalls on density estimates for SDEs with bounded drift . . . 173
8.3.2 Preliminary results . . . 174
8.3.3 Proof of Theorem 8.2.1 . . . 176
8.3.4 Proof of Theorem 8.2.2 using Theorem 8.2.7 . . . 178
8.3.5 Proof of Theorem 8.2.5 . . . 181
8.3.6 Proof of Theorem 8.2.8 . . . 181
8.4 Counterexample . . . 182
8.5 Appendix . . . 184

9 Weak error rates for numerical schemes of non-singular Stochastic Volterra equations with application to option pricing under path-dependent volatility 187
9.1 Introduction . . . 187
9.2 Setting and main results . . . 189
9.2.1 Setting . . . 189
9.2.2 Main results . . . 191
9.3 Preliminary results on infinite dimensional paths . . . 191
9.3.1 State space and path derivatives . . . 191
9.3.2 Expectation of the supremum of a random path process . . . 192
9.3.3 A general Itō formula for path-dependent functionals . . . 193
9.4 Proof of Theorem 9.2.2 . . . 195
9.4.1 Definition of the infinite dimensional semi-group and domino strategy . . . 195
9.4.2 Weak error in small time . . . 197
9.4.3 Proof that the derivatives of g are bounded . . . 202
9.4.4 Conclusion: proof of Theorem 9.2.2 . . . 203
9.4.5 Proof of weak error for the scheme with discretization of the kernels . . . 203
9.5 Simulations . . . 204
9.6 Appendix . . . 205

10 Simulation of Reflected Brownian motion on two dimensional wedges 207
10.1 Introduction . . . 207
10.2 Notations . . . 210
10.3 Setting of the problem . . . 211
10.4 Analytic formulas for the density of the reflected process . . . 213
10.5 Exact simulation algorithms . . . 214
10.5.1 Formulas for the simulation of (τ, W_τ) in the case $\alpha = \pi/m$ . . . 215
10.5.2 Algorithm for the simulation of the stopped Brownian motion: General case 217
10.5.3 Algorithm for the simulation of the reflected Brownian motion . . . 219
10.6 Folding number in a wedge and complexity . . . 220

10.6.1 Majoration of N for the simulation algorithm of the stopped Brownian motion . . . 221
10.6.2 Majoration of the number of folds of the Brownian motion in a wedge . . 221
10.6.3 Proposition of modification of Algorithm II . . . 223
10.7 Adaptation of the algorithms to general processes . . . 225
10.7.1 Stopped Brownian motion with constant drift . . . 225
10.7.2 Adaptation of the simulation algorithms for Itō processes . . . 226
10.8 Simulations . . . 229
10.9 Appendix: Auxiliary Lemmas . . . 230
10.9.1 Estimates on Bessel processes . . . 231
10.10 Appendix: Proofs in Section 10.4 . . . 233
10.11 Appendix: A hint for higher dimensions . . . 236
10.12 Supplementary material . . . 237
10.12.1 Proof of the convergence to the initial condition when $t \to 0$ . . . 237

Bibliography **245**

Notations

We give the notations that are commonly used throughout this thesis.

- We endow the space $\mathbb{R}^d$ with the canonical Euclidean norm denoted by $|\cdot|$ and we denote $\langle \cdot, \cdot \rangle$ the associated canonical inner product. For $x \in \mathbb{R}^d$ and for $R > 0$, we denote $\mathbf{B}(x, R) = \{y \in \mathbb{R}^d : \ |y - x| \leq R\}$.

- For x, $y \in \mathbb{R}^d$ we denote $(x, y) = \{ux + (1-u)y, \ u \in [0, 1]\}$ the geometric segment between x and y.

- For $a, b \in \mathbb{N}$ we denote $\mathcal{M}_{a,b}(\mathbb{R})$ the set of $a \times b$ matrices with real coefficients. If $a = b$ we sometimes use the notation $\mathcal{M}_a(\mathbb{R}) = \mathcal{M}_{a,a}(\mathbb{R})$.

- For $u, v \in \mathbb{R}^d$ we denote by $u * v$, or sometimes by $u \odot v$, the element-wise product (Schur product) i.e. $u * v = [u_i v_i]_{1 \leq i \leq d}$. Similarly we define $u \oslash v = [u_i / v_i]_{1 \leq i \leq d}$.

- For $u \in \mathbb{R}^{d_1}$ and $v \in \mathbb{R}^{d_2}$ we denote by $u \otimes v$ the tensor product with $a \otimes b = [a_i b_j]_{1 \leq i \leq d_1, 1 \leq j \leq d_2}$.

- For $d_1, \ldots, d_r \in \mathbb{N}$ we define $\mathbb{R}^{d_1} \otimes \ldots \otimes \mathbb{R}^{d_r}$ the set of tensors of order r with dimensions $d_1, \ldots, d_r$ i.e. tensors of the form $(x_{i_1, \ldots, i_r})_{1 \leq i_1 \leq d_1, \ldots, 1 \leq i_r \leq d_r}$.

- For $M \in (\mathbb{R}^d)^{\otimes k}$, we denote by $\|M\|$ its operator norm, i.e. $\|M\| = \sup_{u \in \mathbb{R}^{d \times k}, \ |u| = 1} M \cdot u$. If $M : \mathbb{R}^d \to (\mathbb{R}^d)^{\otimes k}$, we denote $\|M\|_\infty = \sup_{x \in \mathbb{R}^d} \|M(x)\|$.

- For $u \in \mathbb{R}^d$ we denote $\mathrm{diag}(u) \in \mathcal{M}_d(\mathbb{R})$ the diagonal matrix with diagonal u.

- For $(\mathcal{A}, d_\mathcal{A})$ and $(\mathcal{B}, d_\mathcal{B})$ two metric spaces, typically $\mathcal{A} = \mathbb{R}^{d_1}$ and $\mathcal{B} = \mathbb{R}^{d_2}$ endowed with the Euclidean distance, we denote $\mathcal{C}^k(\mathcal{A}, \mathcal{B})$ the set of functions $f : \mathcal{A} \to \mathcal{B}$ being k times differentiable with continuous k^{th} derivative. If there is no ambiguity, we simply write $f \in \mathcal{C}^k$.

- For $(\mathcal{A}, d_\mathcal{A})$ and $(\mathcal{B}, d_\mathcal{B})$ two metric spaces, we say that $f : \mathcal{A} \to \mathcal{B}$ is coercive if $d_\mathcal{B}(f(x), 0) \to \infty$ as $d_\mathcal{A}(x, 0) \to \infty$ uniformly. For normed spaces, we can rewrite $|f(x)| \to \infty$ as $|x| \to \infty$.

- We denote $\mathcal{C}^k_b$ the set of functions in $\mathcal{C}^k$ being bounded with bounded partial derivatives up to order k; we denote $\widetilde{\mathcal{C}}^k_b$ the set of functions in $\mathcal{C}^k$ with bounded partial derivatives up to order k but that are not necessarily bounded themselves.

- If $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is Lipschitz continuous, we denote by $[f]_{\text{Lip}}$ its Lipschitz constant.

- For $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ being $\mathcal{C}^k$, for $1 \leq j \leq k$ we denote by $\nabla^j f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2} \otimes (\mathbb{R}^{d_1})^{\otimes j}$ its derivative tensor of order j given by
$$\nabla^j f = \left(\frac{\partial^j f_{i_0}}{\partial x_{i_1} \dots \partial x_{i_j}} \right)_{1 \leq i_0 \leq d_2,\ 1 \leq i_1, \dots, i_j \leq d_1}.$$
- For $f : \mathbb{R}^d \to \mathbb{R}$ such that $\min_{\mathbb{R}^d} f$ exists, we denote
$$\operatorname{argmin}(f) = \left\{ x \in \mathbb{R}^d :\ f(x) = \min_{\mathbb{R}^d} f \right\}.$$
- For $f, g : \mathbb{R}^q \to \mathbb{R}$ and x_0 in $\bar{\mathbb{R}}$, we write $f(x) \sim g(x)$ as $x \to x_0$ meaning $f(x) = g(x) + o(g(x))$ as $x \to x_0$.
- If u_n and v_n are two real-valued sequences, we write $u_n \asymp v_n$ meaning that $u_n = O(v_n)$ and $v_n = O(u_n)$.
- For a random vector X, we denote by $[X]$ its law.
- We denote the total variation distance between two probability distributions π_1 and π_2 on $\mathbb{R}^d$:
$$\mathrm{d_{TV}}(\pi_1, \pi_2) = 2 \sup_{A \in \mathcal{B}or(\mathbb{R}^d)} |\pi_1(A) - \pi_2(A)|.$$
We have as well
$$\mathrm{d_{TV}}(\pi_1, \pi_2) = \sup \left\{ \int_{\mathbb{R}^d} f d\pi_1 - \int_{\mathbb{R}^d} f d\pi_1 :\ f : \mathbb{R}^d \to [-1, 1] \text{ measurable} \right\}.$$
Moreover, we recall that if π_1 and π_2 admit densities with respect to some measure reference λ, then
$$\mathrm{d_{TV}}(\pi_1, \pi_2) = \int_{\mathbb{R}^d} \left| \frac{d\pi_1}{d\lambda} - \frac{d\pi_2}{d\lambda} \right| d\lambda.$$
- We denote the L^p-Wasserstein distance between two distributions π_1 and π_2 on $\mathbb{R}^d$:
$$\mathcal{W}_p(\pi_1, \pi_2) = \inf \left\{ \left(\int_{\mathbb{R}^d} |x - y|^p \pi(dx, dy) \right)^{1/p} :\ \pi \in \mathcal{P}(\pi_1, \pi_2) \right\},$$
where $\mathcal{P}(\pi_1, \pi_2)$ stands for the set of probability distributions on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}or(\mathbb{R}^d)^{\otimes 2})$ with respective marginal laws π_1 and π_2. For $p = 1$, let us recall the Kantorovich-Rubinstein representation of the Wasserstein distance of order 1 [Vil09, Equation (6.3)]:
$$\mathcal{W}_1(\pi_1, \pi_2) = \sup \left\{ \int_{\mathbb{R}^d} f(x)(\pi_1 - \pi_2)(dx) :\ f : \mathbb{R}^d \to \mathbb{R},\ [f]_{\mathrm{Lip}} = 1 \right\}. \tag{0.0.1}$$
- For some distance $\mathcal{D}$ on the set of probability distributions on $\mathbb{R}^d$ and for X and Y two $\mathbb{R}^d$-valued random vectors, we denote without ambiguity $\mathcal{D}(X, Y) = \mathcal{D}([X], [Y])$.
- For $x \in \mathbb{R}^d$, we denote by δ_x the Dirac mass at x.
- We generally consider $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t \geq 0})$ a filtered probability space satisfying the usual conditions.
- We denote by ReLU or $(\cdot)_+$ the ReLU function: $\mathrm{ReLU}(x) = (x)_+ = \max(0, x)$.
- For u, $v \in \mathbb{R}$, we define $u \bmod(v) = u - v\lfloor u/v \rfloor$.
- We often use the notation C and c to denote positive constants, which may change from line to line.

Chapter 1

Introduction

Stochastic algorithms are a powerful tool to solve complex optimization and inference problems, where the state space is not explored deterministically but randomly. Stochastic algorithms have recently known a renewed interest, especially from the Machine Learning community; new research aims at reworking, modifying the existing stochastic algorithms in order to make them more efficient and to accelerate their convergence speed.

1.1 Stochastic algorithms

1.1.1 Stochastic gradient descent and the Robbins-Siegmund lemma

Let $V : \mathbb{R}^d \to \mathbb{R}$ and let us consider the following optimization problem with no constraint:

$$\underset{x\in\mathbb{R}^d}{\text{Minimize}} \quad V(x). \tag{1.1.1}$$

We assume that the dimension d is large so that this optimization problem is complex to solve. We generally make the following assumptions on the function V: V is $\mathcal{C}^1$, coercive i.e. $V(x) \to +\infty$ as $|x| \to +\infty$, which implies in particular that $\min(V)$ exists. The stochastic gradient descent (SGD) algorithm is a gradient descent with noise:

$$X_0 \in \mathbb{R}^d, \quad X_{n+1} = X_n - \gamma_{n+1} \left(\nabla V(X_n) + \zeta_{n+1} \right), \tag{1.1.2}$$

where (γ_n) is a positive sequence of steps which can be constant or decreasing to 0 and (ζ_n) is a sequence of noise corresponding to noisy observations of the true gradient $\nabla V(X_n)$. Illustrations of the gradient descent are given in Figures 1.1 and 1.2; we hope that for large enough n, X_n will be close to $\text{argmin}(V)$. A classical setting is where $\nabla V(X_n)$ is measured with no bias i.e. there exist a random variable Z with values in $\mathbb{R}^q$ and a function $v : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}$ such that

$$\forall x \in \mathbb{R}^d, \ \mathbb{E}[v(x, Z)] = V(x) \ \text{ and } \ \mathbb{E}[\partial_x v(x, Z)] = \nabla V(x). \tag{1.1.3}$$

The stochastic gradient algorithm then becomes:

$$X_{n+1} = X_n - \gamma_{n+1} \partial_x v(x_n, Z_{n+1}), \tag{1.1.4}$$

where (Z_n) is i.i.d. and where $Z_1 \sim Z$. The noise sequence ζ is then a sequence of increments of a martingale i.e. for all $n \in \mathbb{N}$, $\mathbb{E}[\zeta_{n+1} | X_0, \ldots, X_n] = 0$.

Stochastic algorithms were introduced by Robbins and Monro in [RM51] in 1951. The Robbins-Siegmund Lemma [RS71] guarantees the convergence of the stochastic gradient algorithm under some conditions.

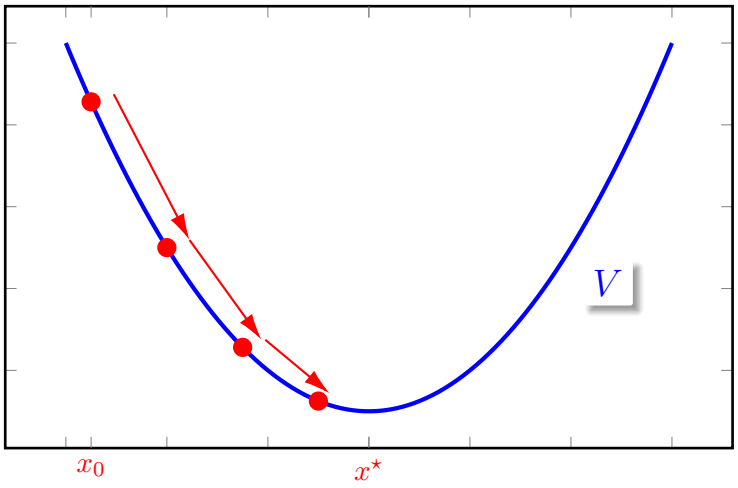


Figure 1.1: Gradient descent algorithm.

Lemma 1.1.1 (Robbins-Siegmund Lemma)**.** *Let $h : \mathbb{R}^d \to \mathbb{R}^d$ and let $H : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^d$ be such that*

$$\forall x \in \mathbb{R}^d,\ \mathbb{E}[H(x, Z)] = h(x).$$

Assume that there exists a $\mathcal{C}^1$ function $V : \mathbb{R}^d \to \mathbb{R}^+$ such that ∇V is Lipschitz-continuous and $|\nabla V|^2 \leq C(1 + V)$ for some constant $C > 0$. Assume furthermore that $\langle \nabla V, h \rangle \geq 0$ and that

$$\forall x \in \mathbb{R}^d,\ \|H(x, Z)\|_2^2 \leq C(1 + V(x)).$$

Let (γ_n) be a sequence of positive real numbers such that

$$\sum_{n=1}^{\infty} \gamma_n = +\infty, \quad \sum_{n=1}^{\infty} \gamma_n^2 < +\infty.$$

Then the sequence (X_n) defined by

$$X_{n+1} = X_n - \gamma_{n+1} H(X_n, Z_{n+1})$$

satisfies:

1. *$X_n - X_{n-1} \to 0$ almost surely and in L^2 as $n \to \infty$.*
2. *The sequence $\left(\mathbb{E}[V(X_n)]\right)_{n \in \mathbb{N}}$ is bounded.*
3. *$V(X_n)$ converges almost surely.*
4. *$\sum_{n \geq 1} \gamma_n \langle \nabla V, h \rangle (X_{n-1}) < +\infty$ almost surely.*
5. *The martingale sequence $(\sum_{k=1}^{n} \gamma_k (H(X_{k-1}, Z_k) - h(X_{k-1}))$ converges almost surely and in $L^2(\mathbb{P})$.*

1.1.2 Classic examples of stochastic optimization problems

We now give classic examples where such optimization problems arise.

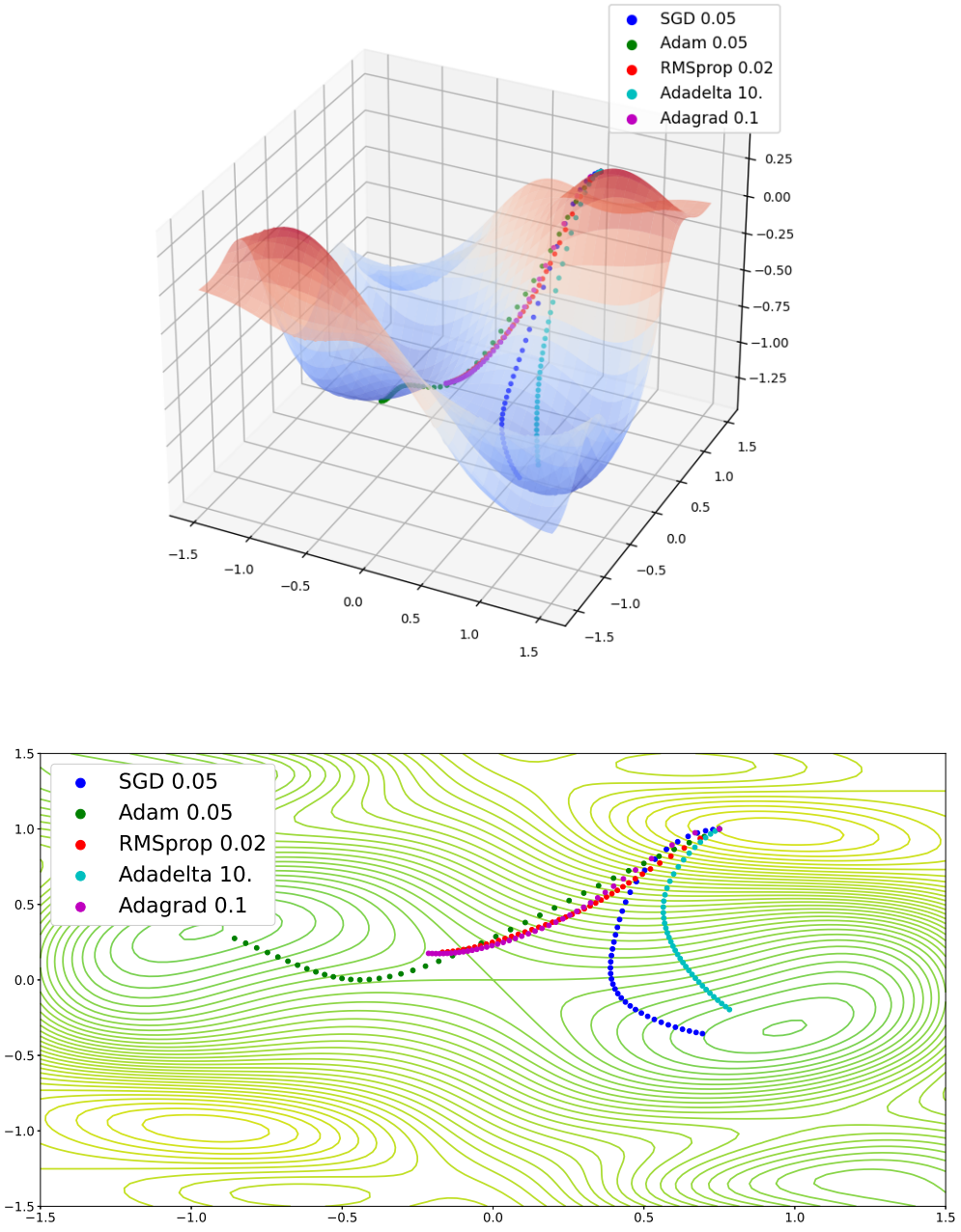


Figure 1.2: Example of gradient descent in $\mathbb{R}^2$ with different gradient descent algorithms for the function $V(x, y) = -\sin(x^2)\cos(3y^2)e^{-x^2y^2} - e^{-(x+y)^2}$ with two global minima next to $(0, 0)$. For each gradient descent method we indicate its learning rate in the legend.

1.1.2.1 Machine Learning and regression

Let us consider data samples $u_i \in \mathbb{R}^{d_{\text{in}}}$ and $y_i \in \mathbb{R}^{d_{\text{out}}}$ for $1 \leq i \leq M$ associated to a regression problem, where (u_i) are the inputs and (y_i) are the outputs. That is, we look for a function $\psi : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ which fits to the data i.e.

$$\forall\ 1 \leq i \leq M,\ \psi(u_i) \approx y_i.$$

In other words, the objective is to extract a model from the empirical data. We look for a function ψ in a family of functions parametrized by a finite-dimension parameter: $\{\psi_x,\ x \in \mathbb{R}^d\}$. For a loss function $L : \mathbb{R}^{d_{\text{out}}} \to \mathbb{R}^+$ which measures the error between the prediction $\psi_x(u_i)$ and the true data y_i, the regression problem then becomes the minimization of the average loss over the dataset and can be written as the following optimization problem:

$$\underset{x\in\mathbb{R}^d}{\text{Minimize}} \quad V(x) := \frac{1}{M}\sum_{i=1}^{M} L(\psi_x(u_i) - y_i). \tag{1.1.5}$$

Examples of parametrization:

- *Least-square regression:* Let us write $x = (\theta, \iota) \in \mathcal{M}_{d_{\text{out}},d_{\text{in}}}(\mathbb{R}) \times \mathbb{R}^{d_{\text{out}}}$ so that ψ_x is the affine function $\psi_{(\theta,\iota)}(u) = \theta \cdot u + \iota$ and L is the quadratic loss i.e. $L(w) = (1/2)|w|^2$.
- *Logistic regression:* Let us assume that $y_i \in \{\pm 1\}$ are labels associated to a binary classification problem and that $u_i \in \mathbb{R}^d$. Then the problems becomes

$$\underset{x\in\mathbb{R}^d}{\text{Minimize}} \quad V(x) := \frac{1}{M}\sum_{i=1}^{M} \log\left(1 + e^{-y_i\langle x, u_i\rangle}\right).$$

- *Fully connected artificial Neural Networks:* Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a non-linear function. The function φ is generally chosen to be a sigmoid-type or a ReLU-type function, see Figure 1.4. Let $K+1$, $K \in \mathbb{N}$, be the number of layers of the neural network and for $k = 0, \ldots, K$ let $d_k \in \mathbb{N}$ be the size of the k^{th} layer with $d_0 = d_{\text{in}}$ and $d_K = d_{\text{out}}$. For $k = 1, \ldots, K$, $u \in \mathbb{R}^{d_{k-1}}$ and for $\theta_k \in \mathcal{M}_{d_k,d_{k-1}}(\mathbb{R})$ and $\iota_k \in \mathbb{R}^{d_k}$ we define the vector in $\mathbb{R}^{d_k}$:

$$\varphi_{\theta_k,\iota_k}(u) := \left[\varphi([\theta_k \cdot u + \iota_k]_i)\right]_{1\leq i\leq d_k},$$

  i.e. the scalar function φ is applied to the vector $\theta_k \cdot u + \iota_k$ coordinate by coordinate. Let us write $x = (\theta_1, \iota_1, \ldots, \theta_K, \iota_K)$, then the output of the neural network is

$$\psi_x(u) = \theta_K \cdot \left(\varphi_{\theta_{K-1},\iota_{K-1}} \circ \cdots \circ \varphi_{\theta_1,\iota_1}(u)\right) + \iota_K.$$

  An illustration of a fully connected neural network is given in Figure 1.3. Neural networks are known for their ability to approximate a wide range of non-linear functions in high dimension and to fit to many non-linear regression tasks [Cyb89, LBH15].

The gradient descent algorithm for (1.1.5) then becomes

$$X_{n+1} = X_n - \frac{\gamma_{n+1}}{M}\sum_{i=1}^{M} \partial_x\psi_x(u_i) \cdot \nabla L(\psi_x(u_i) - y_i).$$

However, if the number of samples M is large then computing ∇V exactly at each iteration is costly. We instead replace this algorithm by a stochastic gradient algorithm: at each iteration n let $i_n \in \{1, \ldots, M\}$ be a random index chosen uniformly at random and the iteration becomes

$$X_{n+1} = X_n - \frac{\gamma_{n+1}}{M}\partial_x\psi_x(u_{i_{n+1}}) \cdot \nabla L(\psi_x(u_{i_{n+1}}) - y_{i_{n+1}}). \tag{1.1.6}$$

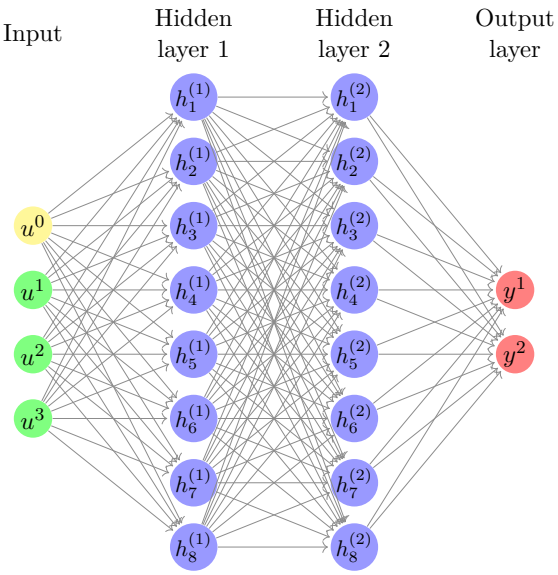


Figure 1.3: Scheme of a fully connected neural network with 3 layers (2 hidden layers) and with output in $\mathbb{R}^2$.

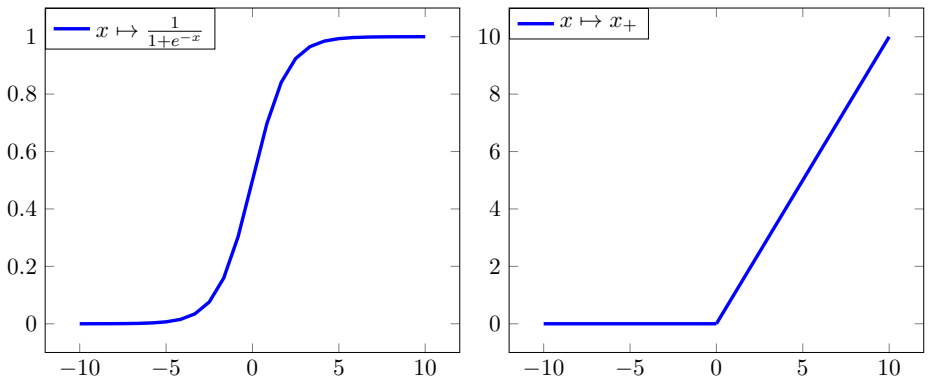


Figure 1.4: The Sigmoid and ReLU functions

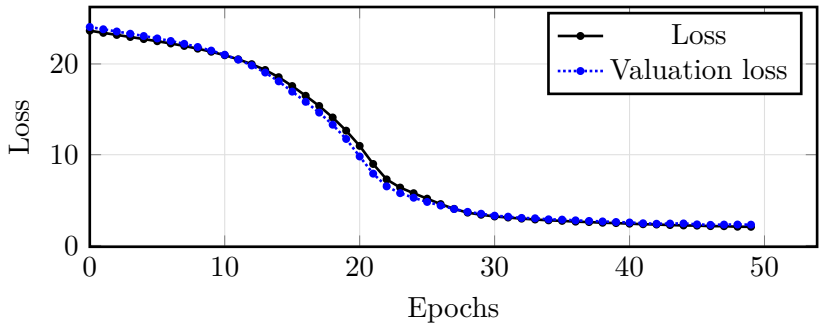


Figure 1.5: Loss and valuation loss during the training of a neural network with two hidden layers with 64 units each and ReLU activation for the prediction of fuel efficiency on the Auto MPG dataset [Qui93] using stochastic gradient descent.

As a more stable version of (1.1.6) and still computationally efficient, the gradient can be estimated by an average over a small batch of data:

$$X_{n+1} = X_n - \frac{\gamma_{n+1}}{N_{\text{batch}}} \sum_{i \in \mathcal{I}_{n+1}} \partial_x \psi_x(u_i) \cdot \nabla L(\psi_x(u_i) - y_i)$$

where $\mathcal{I}_{n+1}$ is a subset of $\{1, \dots, M\}$ of size N_{batch} taken uniformly at random with $N_{\text{batch}} \ll M$. An example of neural network training is given in Figure 1.5.

1.1.2.2 Computation of quantiles and Value at Risk

Let Z be a random variable taking its values in $\mathbb{R}$ and such that $\mathbb{E}|Z| < \infty$. For $\alpha \in [0, 1]$, let q_α be the quantile of order α, i.e.

$$q_\alpha = \inf\{u \in \mathbb{R} : \mathbb{P}(Z \leq u) \geq \alpha\}.$$

In the Mathematical Finance community, the quantile q_α is called Value at Risk (VaR) and is widely used with α close to 1 as a risk measure for risk management [Jor96]. Following [UR01] we have the characterization

$$q_\alpha = \underset{x \in \mathbb{R}}{\operatorname{argmin}}\ \mathbb{E}\left[x + \frac{1}{1-\alpha}(Z - x)_+\right] =: \underset{x \in \mathbb{R}}{\operatorname{argmin}} \mathbb{E}[v(x, Z)].$$

To compute q_α, one may apply a gradient descent algorithm however computing exactly the expectation may be impossible. Instead the stochastic gradient algorithm reads

$$X_{n+1} = X_n - \gamma_{n+1} \begin{cases} 1 & \text{if } X_n \geq Z_{n+1} \\ -\alpha/(1-\alpha) & \text{if } X_n < Z_{n+1}, \end{cases}$$

where (Z_n) is an i.i.d. sequence and where $Z_1 \sim Z$. The Conditional Value at Risk (CVaR) is then defined as

$$\mathbb{E}[Z | Z \geq q_\alpha] = \mathbb{E}[v(q_\alpha, Z)]$$

and can be directly estimated by the online computation [BFP09]:

$$\frac{1}{\Gamma_n} \sum_{k=1}^{n} \gamma_k v(X_k, Z_k) \quad \text{with } \Gamma_n := \gamma_1 + \ldots + \gamma_n.$$

1.1.2.3 Stochastic optimal control

Let us consider the following stochastic control problem associated to an SDE in continuous time:

$$\min_u J(u) := \mathbb{E}\left[\int_0^T G(t, Y_t^u)dt + F(Y_T^u)\right], \tag{1.1.7}$$

$$dY_t^u = b(Y_t^u, u_t)dt + \sigma(Y_t^u, u_t)dW_t, \ t \in [0, T] \tag{1.1.8}$$

where $b : \mathbb{R}^{d_1} \times \mathbb{R}^{d_3} \to \mathbb{R}^{d_1}$, $\sigma : \mathbb{R}^{d_1} \times \mathbb{R}^{d_3} \to \mathcal{M}_{d_1,d_2}(\mathbb{R})$, W is a $\mathbb{R}^{d_2}$-valued Brownian motion and u is a $\mathbb{R}^{d_3}$-valued continuous adapted process, $T > 0$, $G : \mathbb{R}^+ \times \mathbb{R}^{d_1} \to \mathbb{R}$ and $F : \mathbb{R}^{d_1} \to \mathbb{R}$.

Stochastic control problems are usually solved using specific strategies, such as Forward-Backward SDEs (FBSDEs) [PW99], or by solving Hamilton-Jacobi-Bellman (HJB) optimality conditions [Bel57] through Partial Differential Equation (PDE) methods or by stochastic dynamic programming [KD01]. Such problems can also be solved using Neural Networks calibrated

by SGD techniques [GM05, HE16, WLP+19, CL21, BHLP22]. More specifically, we proceed as follows. We approximate the control u as the output of a (fully connected) neural network

$$u_t = u_\theta(t, Y_t^u) \tag{1.1.9}$$

where u_θ is a neural function with finite-dimensional parameter $\theta \in \mathbb{R}^d$. Since (1.1.8) defines a Markovian process, we can assume that u_t depends only on t and on Y_t^u instead of t and the whole previous trajectory $(Y_s^u)_{s\in[0,t]}$. We consider a subdivision of $[0,T]$:

$$0 = t_0 < t_1 < \cdots < t_N = T \tag{1.1.10}$$

and approximate (1.1.7) by the time discretized model

$$\min_{\theta\in\mathbb{R}^d} \bar{J}(u_\theta) := \sum_{k=0}^{N-1} (t_{k+1} - t_k) G(t_{k+1}, \bar{Y}_{t_{k+1}}^{u_\theta}) + F(\bar{Y}_{t_N}^{u_\theta}), \tag{1.1.11}$$

$$\bar{Y}_{t_{k+1}}^{u_\theta} = \bar{Y}_{t_k}^{u_\theta} + (t_{k+1} - t_k) b(\bar{Y}_{t_k}^{u_\theta}, u_\theta(t_k, \bar{Y}_{t_k}^{u_\theta})) + \sqrt{t_{k+1} - t_k}\, \sigma(\bar{Y}_{t_k}^{u_\theta}, u_\theta(t_k, \bar{Y}_{t_k}^{u_\theta}))\xi_{k+1}, \tag{1.1.12}$$

$$\xi_k \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_2}). \tag{1.1.13}$$

For every $\theta \in \mathbb{R}^d$, $\nabla_\theta \bar{J}$ can be computed by automatic differentiation, where the gradient w.r.t. to θ is tracked all along the trajectories through the recursive dynamics of the Euler scheme (1.1.12) [GG05, Gil07]. Then the SGD algorithms reads

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla_\theta \bar{J}(u_{\theta_n}, (\xi_k^{n+1})_{1\le k\le N}) \tag{1.1.14}$$

where $(\xi_k^n)_{1\le k\le N, n\in\mathbb{N}}$ is an array of i.i.d. random vectors $\mathcal{N}(0, I_{d_2})$-distributed, $(\gamma_n)_{n\in\mathbb{N}}$ is a non-increasing positive step sequence and where we wrote the dependence of $\bar{J}$ in (ξ_k^n).

1.1.2.4 Variance reduction in Monte Carlo simulation

In Monte Carlo simulation, we aim to estimate $\mathbb{E}[\varphi(Z)]$ where Z is a random vector taking its values in $\mathbb{R}^d$, $\varphi : \mathbb{R}^d \to \mathbb{R}$ and such that $\mathbb{E}|\varphi(Z)|^2 < \infty$, using the estimator $\frac{1}{M}\sum_{i=1}^M \varphi(Z_i)$ where $M \in \mathbb{N}$ is large and Z_i are i.i.d. with $Z_1 \sim Z$. Variation reduction techniques help to estimate $\mathbb{E}[\varphi(Z)]$ with smaller variance, yielding tighter confidence intervals.

More specifically, let us consider variance reduction by unconstrained recursive importance sampling as introduced in [Aro04] and [LP10]. Let us assume that $Z \sim \mathcal{N}(0, I_d)$. For example, for a Call option on a basket of risky assets driven by a multi-dimensional Black-Scholes model we have

$$\varphi(Z) = \left(\sum_{i=1}^d a_i s_0^i \exp\left[\left(r - \sum_{j=1}^d \frac{\sigma_{ij}^2}{2} \right) T + \sqrt{T} \sum_{j=1}^d \sigma_{ij} Z^j \right] - K \right)_+,$$

where r, T, $K \in \mathbb{R}^+$, $\sigma \in \mathcal{M}_d(\mathbb{R})$ is definite positive definite symmetric and s_0, $a \in (0, s_0)^d$ and $Z = (Z^1, \ldots, Z^d)$.

For $\theta \in \mathbb{R}^d$ we have

$$\begin{aligned}\mathbb{E}[\varphi(Z)] &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \varphi(z) e^{-\frac{|z|^2}{2}} dz = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \varphi(z+\theta) e^{\frac{-|\theta|^2}{2} - \langle\theta, z\rangle} e^{-\frac{z^2}{2}} dz \\ &= \mathbb{E}\left[\varphi(Z+\theta) e^{\frac{-|\theta|^2}{2} - \langle\theta, Z\rangle} \right]\end{aligned} \tag{1.1.15}$$

and

$$\mathrm{Var}\left[\varphi(Z+\theta) e^{\frac{-|\theta|^2}{2} - \langle\theta, Z\rangle} \right] = \mathbb{E}\left[\varphi^2(Z+\theta) e^{-|\theta|^2 - 2\langle\theta, Z\rangle} \right] - \mathbb{E}\left[\varphi(Z+\theta) e^{\frac{-|\theta|^2}{2} - \langle\theta, Z\rangle} \right]^2$$

$$= \mathbb{E}\left[\varphi^2(Z+\theta)e^{-|\theta|^2-2\langle\theta,Z\rangle}\right] - \mathbb{E}[\varphi(Z)]^2.$$

The objective is to solve the minimization problem

$$\underset{\theta\in\mathbb{R}^d}{\text{Minimize}}\ V(\theta) := \mathbb{E}\left[\varphi^2(Z+\theta)e^{-|\theta|^2-2\langle\theta,Z\rangle}\right] = \mathbb{E}\left[\varphi^2(Z)e^{-\langle\theta,Z\rangle+|\theta|^2/2}\right]$$

so that to estimate $\mathbb{E}[\varphi(Z)]$ by Monte Carlo simulation using (1.1.15) with a smaller variance. The (naive) stochastic gradient algorithm writes

$$\theta_{n+1} = \theta_n - \gamma_{n+1}\varphi^2(Z_{n+1})e^{-\langle\theta_n,Z_{n+1}\rangle+|\theta_n|^2/2}(\theta_n - Z_{n+1}). \tag{1.1.16}$$

However (1.1.16) suffers from instability and explosion because of the factors of exponential growth. This algorithm is improved in [LP10] using the following new representation of the gradient:

$$\nabla V(\theta) = \mathbb{E}[\varphi^2(Z-\theta)(2\theta - Z)].$$

1.1.2.5 Optimal quantization

Vector quantization consists in mapping input values from a large (or continuous) set to output values in a (finite) smaller set. It and was originally introduced for data compression [Llo82] and is now more widely used in Machine Learning for unsupervised learning and clustering analysis, in numerical probability for (conditional) expectation computation and option pricing.

More specifically, for μ some probability distribution in $L^2(\mathbb{R}^q)$ and for fixed $K \in \mathbb{N}$, the optimal quantization problem reads

$$\min_{x=(x^1,\dots,x^K)\in(\mathbb{R}^q)^K} V(x) := \frac{1}{2}\int_{\mathbb{R}^q} \min_{1\le k\le K} |\xi - x^k|^2 \mu(d\xi). \tag{1.1.17}$$

Then defining the Voronoï partition of $\mathbb{R}^q$:

$$V_k(x^1,\dots,x^K) := \left\{\xi\in\mathbb{R}^q \middle|\ \forall\ 1\le j\le K,\ |\xi - x^k| \le |\xi - x^j|\right\},\ 1\le k\le K,$$

we have

$$\partial_{x^k} V(x) = \partial_{x^k}\frac{1}{2}\sum_{j=1}^K \int_{V_j(x)} |\xi - x^j|^2\mu(d\xi) = \int_{V_k(x)} (x^k - \xi)\mu(d\xi) = \mathbb{E}_{Y\sim\mu}\left[\mathbb{1}_{Y\in V_k(x)}(x^k - Y)\right]$$

and the corresponding SGD algorithm, also called Competitive Learning Vector Quantization (CLVQ) in this case, reads with $X_n = (X_n^1,\dots,X_n^K) \in (\mathbb{R}^q)^K$ and $Y_n \sim \mu$ and iid :

$$X_{n+1} = X_n - \gamma_{n+1}\left[\mathbb{1}_{Y_{n+1}\in V_k(X_n)}(X_n^k - Y_{n+1})\right]_{1\le k\le K}. \tag{1.1.18}$$

An example of optimal quantization using the CLVQ algorithm is given in Figure 1.6.

We refer to [Pag15, Section 3.2] for more details about the CLVQ algorithm.

1.1.3 Stochastic processes and the Euler-Maruyama scheme

1.1.3.1 Stochastic differential equations

Let us consider a general SDE having its values in $\mathbb{R}^d$:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 \perp\!\!\!\perp W \tag{1.1.19}$$

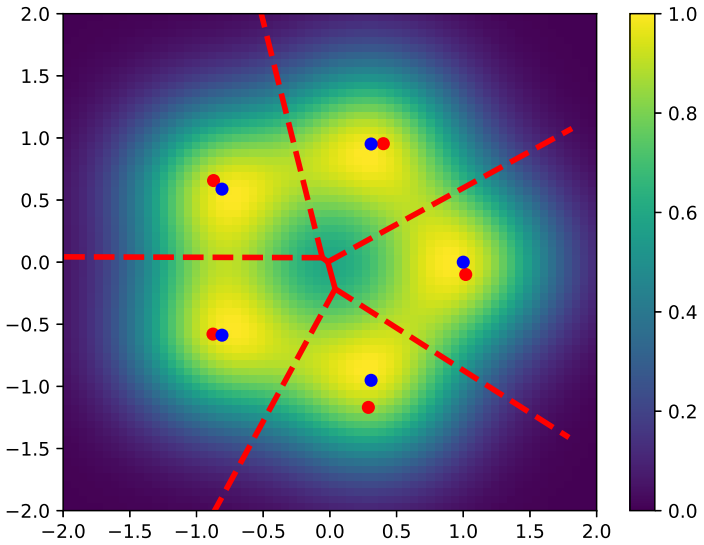


Figure 1.6: Example of optimal quantization in $\mathbb{R}^2$ where μ is the unweighted Gaussian mixture with component distributions $\mathcal{N}(y_k, (1/4)I_2)$, $1 \leq k \leq 5$, $y_k = (\cos(2k\pi/5), \sin(2k\pi/5))$ and where the number of quantizers is $K = 5$. We plot the density of μ, the points y_k in blue, the centroids x^k given by the CLVQ algorithm in red and the frontiers of the Voronoï partition $(V_k)_{1\leq k\leq 5}$ in red.

where W is a standard $\mathbb{R}^q$-valued Brownian motion and X_0 and W are both defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. A method of approximate numerical simulation of X_T for fixed time horizon $T > 0$ is the so-called Euler-Maruyama scheme which is the following time-discrete recursive algorithm:

$$\bar{X}_0 = X_0, \quad \bar{X}_{n+1} = \bar{X}_n + hb(\bar{X}_n) + h^{1/2}\sigma(\bar{X}_n)U_{n+1}, \quad U_n \sim \mathcal{N}(0, I_d) \text{ i.i.d.} \tag{1.1.20}$$

with $h = T/N$, $N \in \mathbb{N}$ and X_T is approximated by $\bar{X}_N$. This scheme is used in particular for Monte Carlo estimation of expectations where for some $f : \mathbb{R}^d \to \mathbb{R}$, $\mathbb{E}[f(X_T)]$ is approximated by the empirical mean $M^{-1}\sum_{i=1}^{M} f(\bar{X}_N^{(i)})$, where $\bar{X}_N^{(i)}$ are M independent simulations following (1.1.20). Conversely, some discrete stochastic algorithms can be seen as the discretization of some SDE, as seen in the following sections.

In order to evaluate the quality of the approximation $\bar{X}_N$ as $N \to \infty$, we use strong error and weak error bounds. The strong error is defined for $p \geq 1$ as the L^p norm

$$\left(\mathbb{E}|\bar{X}_N - X_T|^p\right)^{1/p} = \|\bar{X}_N - X_T\|_p.$$

Bounds on the strong error are obtained as follows. Assuming for example that b and σ are bounded and Lipschitz-continuous and writing $\bar{X}_N = Y_T$ with

$$dY_t = b(Y_{\underline{t}})dt + \sigma(Y_{\underline{t}})dW_t, \quad Y_0 = X_0$$

where $\underline{t} = h\lfloor t/h \rfloor$, for $p \geq 2$ and $t \in [0, T]$ we have

$$\|Y_t - X_t\|_p \leq \left\| \int_0^t (b(Y_{\underline{s}}) - b(Y_s))ds \right\|_p + \left\| \int_0^t (b(Y_s) - b(X_s))ds \right\|_p$$

$$
\begin{aligned}
&+ \Big\| \int_0^t (\sigma(Y_{\underline{s}}) - \sigma(Y_s)) dW_s \Big\|_p + \Big\| \int_0^t (\sigma(Y_s) - \sigma(X_s)) dW_s \Big\|_p \\
&\leq [b]_{\mathrm{Lip}} \Big(\int_0^t \|Y_s - X_s\|_p ds + \int_0^t \|Y_s - Y_{\underline{s}}\|_p ds \Big) \\
&\quad + C_p^{\mathrm{BDG}} [\sigma]_{\mathrm{Lip}} \Big(\Big| \int_0^t \|Y_s - X_s\|_p^2 ds \Big|^{1/2} + \Big| \int_0^t \|Y_s - Y_{\underline{s}}\|_p^2 ds \Big|^{1/2} \Big),
\end{aligned}
$$

where we used the Burkholder-Davis-Gundy and the regular and generalized Minkowski inequalities. But using the expression of the law of Y_s conditionally to $Y_{\underline{s}}$, we have $\|Y_s - Y_{\underline{s}}\|_p \leq C(s - \underline{s})^{1/2}$ so that with $\varphi_t := \sup_{s \in [0,t]} \|Y_s - X_s\|_p$ we get

$$
\varphi_t \leq [b]_{\mathrm{Lip}} \int_0^t \varphi_s ds + C_p^{\mathrm{BDG}} [\sigma]_{\mathrm{Lip}} (\int_0^t \varphi_s^2 ds)^{1/2} + C h^{1/2}
$$

but we have for every $\alpha > 0$:

$$
\Big(\int_0^t \varphi_s^2 ds \Big)^{1/2} \leq \varphi_t^{1/2} \Big(\int_0^t \varphi_s ds \Big)^{1/2} \leq \frac{\alpha}{2} \varphi_s + \frac{1}{2\alpha} \int_0^t \varphi_s ds
$$

and taking $\alpha = C_p^{\mathrm{BDG}} [\sigma]_{\mathrm{Lip}}$ yields

$$
\varphi_t \leq C \int_0^t \varphi_s ds + C h^{1/2},
$$

where C is a constant depending on p, T, $\|b\|_\infty$, $[b]_{\mathrm{Lip}}$, $\|\sigma\|_\infty$ and $[b]_{\mathrm{Lip}}$. By the Gronwall Lemma we obtain $\varphi_T \leq C h^{1/2}$ for $p \geq 2$; for $\in [1, 2)$ the result still holds remarking that $\| \cdot \|_p \leq \| \cdot \|_2$. Then the strong error $\|\bar{X}_N - X_T\| = \|Y_T - X_T\|_p$ is of order $N^{-1/2}$ as $N \to \infty$.

The weak error measures the error between the respective laws of X_T and $\bar{X}_N$, denoted $[X_T]$ and $[\bar{X}_N]$ respectively, and is generally defined as the total variation distance:

$$
\mathrm{d_{TV}}([\bar{X}_N], [X_T]) = 2 \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\mathbb{P}(\bar{X}_N \in A) - \mathbb{P}(X_T \in A)|,
$$

or as the L^p-Wasserstein distance:

$$
\mathcal{W}_p([\bar{X}_N], [X_T]) = \inf \left\{ \left(\int_{\mathbb{R}^d} |x - y|^p \pi(dx, dy) \right)^{1/p} : \ \pi \in \mathcal{P}([\bar{X}_N], [X_T]) \right\},
$$

where $\mathcal{P}([\bar{X}_N], [X_T])$ stands for the set of probability distributions on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)^{\otimes 2})$ with respective marginal distributions $[\bar{X}_N]$ and $[X_T]$, or as

$$
\sup \Big\{ \mathbb{E}[f(\bar{X}_N)] - \mathbb{E}[f(X_T)], \ f \in \mathcal{A} \Big\}
$$

where $\mathcal{A}$ is some class of Borel functions from $\mathbb{R}^d$ to $\mathbb{R}$; this last case corresponds to the total variation distance for $\mathcal{A}$ being the set of Borel measurable functions bounded by 1, and to the L^1-Wasserstein distance for $\mathcal{A}$ being the set of Lipschitz-continuous functions with Lipschitz constant no greater than 1, according to the Monge-Kantorovich duality for $\mathcal{W}_1$.

The weak error is generally much more difficult to analyse than the strong error. For the L^p-Wasserstein distance, the weak error can be directly bounded by the L^p-strong error, however better bounds can be obtained. Talay and Tubaro [TT90] and Bally and Talay [BT96] proved that the weak error is *in general* of order N^{-1}, against the order $N^{-1/2}$ for the strong error, thus showing that the weak error fundamentally differs from the strong error and generally converges faster.

1.1.3.2 Stochastic Volterra equations

Let us consider (continuous) stochastic Volterra processes in $\mathbb{R}^d$ i.e. which writes

$$X_t = X_0 + \int_0^t K_1(t,s)b(X_s)ds + \int_0^t K_2(t,s)\sigma(X_s)dW_s, \quad t \in [0,T], \tag{1.1.21}$$

where (W_t) is a standard Brownian motion in $\mathbb{R}^{q_3}$, where

$$b : \mathbb{R}^d \to \mathbb{R}^{q_1}, \quad K_1 : [0,T]^2 \to \mathcal{M}_{d,q_1}(\mathbb{R}), \quad \sigma : \mathbb{R}^d \to \mathcal{M}_{q_2,q_3}(\mathbb{R}), \quad K_2 : [0,T]^2 \to \mathcal{M}_{d,q_2}(\mathbb{R}),$$

and where $q_1, q_2, q_3 \in \mathbb{N}$. Stochastic Volterra equations (SVE) have been introduced for modelling in population dynamics, biology and physics [GLS90, Moh98], in order to generalize modelling to non-Markovian stochastic systems with some memory effect. They have been mathematically studied since [BM80] and [Pro85]. SVEs have recently attracted much attention in the mathematical finance community in the context of rough volatility modelling i.e. with $K_1(t,s) = K_2(t,s) = (t-s)^{H-1/2}$ for some $H \in (0,1/2)$, yielding less regular trajectories than for classic SDEs because of the singularity of the kernel for $s = t$, and which are more able to reproduce some features of asset prices for small H, typically $H \simeq 0.1$ [ALV07, GJR18, EEFR18, JR20, Fuk17, Fuk21]. We recall that a rough stochastic volatility model is a special case of a singular two-dimensional SVE, where the first process is an asset price satisfying

$$dS_t = S_t\sqrt{V_t}dB_t$$

for some Brownian motion B and where the second process (V_t) is the volatility satisfying some rough stochastic equation, then giving for the joint process (1.1.21) 2×2 matrix kernels K_1 and K_2 being diagonal and constant on their first coordinate.

An approximation of the solution of (1.1.21) is given by the Euler-Maruyama scheme for SVEs:

$$\bar{X}_{n+1} = X_0 + h\sum_{j=0}^{n} K_1((n+1)h, jh)b(\bar{X}_j) + h^{1/2}\sum_{j=0}^{n} K_2((n+1)h, jh)\sigma(\bar{X}_j)(W_{t_{j+1}} - W_{t_j})$$

with $h = T/N$, $N \in \mathbb{N}$ and X_T is approximated by $\bar{X}_N$.

The strong order of convergence of the Euler-Maruyama scheme for a rough SVE with parameter H is known to be N^{-H} in general, which is very slow for small $H \in (0,1/2)$. The weak order rate is still an open problem, see [BHT22, BFN22, Gas23, FSW22] for recent advances.

1.1.4 Bayesian inference and sampling from probability measures

Stochastic algorithms are also used for sampling from a probability measure.

In Bayesian inference, we consider a family of probability distributions on $\mathbb{R}^{d_0}$ parametrized by a finite-dimensional parameter $\{p(u|x)du : x \in \mathbb{R}^d\}$ and a prior distribution on the parameter $p_0(x)dx$. Assuming that the observations u_1, ..., u_M are i.i.d. and follow the distribution $p(u|x)du$ conditionally to the value of the parameter x, the posterior probability on $x \in \mathbb{R}^d$ knowing the observations has density proportional to $p_0(x)p(u_1|x)\dots p(u_M|x)$. Defining

$$\begin{aligned} V(x) &:= -\log(p_0(x)p(u_1|x)\dots p(u_M|x)) = -\log(p_0(x)) - \log(p(u_1|x)) - \dots - \log(p(u_M|x)) \\ &=: V_0(x) + V_1(x) + \dots + V_M(x), \end{aligned}$$

then the posterior parameter X has a law of density proportional to $e^{-V(x)}$. In this context, Bayesian inference requires sampling from a probability measure which is defined from a large amount of data if M is large.

Given a probability measure ν, Markov Chain Monte Carlo (MCMC) methods build an explicit Markov chain (X_n) admitting ν as its invariant measure, allowing to approximatively sample from ν. The Metropolis-Hastings algorithm was introduced by Metropolis in 1953 [MRR$^+$53] and extended by Hastings in 1970 [Has70]. Given a probability density ν and a parametrized Markov kernel chosen by the user $Q(x, dy)$ on $\mathbb{R}^d$, the Markov chain X on $\mathbb{R}^d$ is defined recursively by:

- Sample $\tilde{X}_{n+1}$ from $Q(X_n, dy)$.
- $X_{n+1} = \tilde{X}_{n+1}$ with acceptance probability

$$\min\left(1, \frac{\nu(\tilde{X}_{n+1})Q(\tilde{X}_{n+1}, X_n)}{\nu(X_n)Q(X_n, \tilde{X}_{n+1})}\right)$$

  and $X_{n+1} = X_n$ otherwise

so that the probability measure with density ν is the unique invariant distribution of X. One advantage of this algorithm is that it does not require to compute the normalization constant of the density ν nor the one of the kernel Q.

Lamberton and Pagès introduced in 2002 the method of sampling from a probability measure by solving SDEs [LP02]. That is, assuming that the SDE

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t \tag{1.1.22}$$

taking its values in $\mathbb{R}^d$ is ergodic with invariant measure ν, then considering the Euler-Maruyama scheme with decreasing steps

$$\bar{X}_{n+1} = \bar{X}_n + \gamma_{n+1}b(\bar{X}_n) + \sqrt{\gamma_{n+1}}\sigma(\bar{X}_n)U_{n+1}, \quad U_n \sim \mathcal{N}(0, I_d) \text{ i.i.d.}, \tag{1.1.23}$$

the measure ν is estimated by the weighted average of Dirac measures

$$\nu_n := \frac{1}{\Gamma_n}\sum_{k=1}^{n}\gamma_k\delta_{\bar{X}_k}, \quad \Gamma_n = \gamma_1 + \ldots + \gamma_n. \tag{1.1.24}$$

[LP02] shows that for a general mean-reverting diffusion (1.1.22), the averaged Euler scheme of an ergodic diffusion converges to the invariant probability measure of the diffusion. More precisely:

Theorem 1.1.2. *Assume that there exists a $\mathcal{C}^2$ and coercive Lyapunov function $V : \mathbb{R}^d \to [V^\star, \infty)$ such that $\nabla^2 V$ is bounded and*

$$|\nabla V|^2 + |b|^2 \leq CV, \quad \|\sigma(x)\sigma^\top(x)\| = o(V(x)) \quad \text{as } |x| \to \infty, \tag{1.1.25}$$

$$\exists \alpha > 0,\ \beta \in \mathbb{R},\ \langle \nabla V, b\rangle \leq -\alpha V + \beta, \tag{1.1.26}$$

and that (1.1.22) admits a unique invariant measure ν. Then for every $f : \mathbb{R}^d \to \mathbb{R}$ being continuous with $f(x) = o(V^k(x))$ for some $k \in \mathbb{N}$, then $\nu_n(f) \to \nu(f)$ as $n \to \infty$.

This method of simulation with respect to some probability measure fundamentally differs from using the Euler-Maruyama scheme for solving SDEs at fixed time horizon $T > 0$. It was then specifically analysed and extended in a series of papers, among them [LP03, Lem05, PP09, LP12, DM17, PR20].

More recently, [PP23] gives convergence rates for the law of $\bar{X}_n$ itself in the algorithm (1.1.23) for the L^1-Wasserstein distance and the total variation distance. Assuming furthermore that σ is bounded and elliptic and that b is Lipschitz-continuous and the following confluence assumption:

$$\forall x, y \in \mathbb{R}^d, \quad \mathcal{W}_1([X_t^x], [X_t^y]) \leq C|x-y|e^{-\rho t}, \tag{1.1.27}$$

which is fulfilled in particular if the uniform dissipative assumption outside some compact set is satisfied:

$$\exists \alpha_0 > 0, \quad \exists R > 0, \quad \forall x, y \in \mathbf{B}(0,R), \quad \langle b(x) - b(y), x - y\rangle \leq -\alpha_0 |x-y|^2, \tag{1.1.28}$$

the authors obtain a convergence speed at rate "almost" γ_n, which reads:

$$\mathcal{W}_1([\bar{X}_n^x], \nu) \leq C\gamma_n |\log(\gamma_n)| \vartheta(x), \tag{1.1.29}$$

$$\forall \varepsilon > 0, \ \exists C_\epsilon > 0, \ \mathrm{d_{TV}}([\bar{X}_n^x], \nu) \leq C_\varepsilon \gamma_n^{1-\varepsilon} \vartheta(x), \tag{1.1.30}$$

$$\vartheta(x) := (1 + |x|) \wedge V^2(x).$$

The strategy of proof relies on a *domino strategy* (telescopic sum) inspired by proofs of weak error expansion of discretization schemes of diffusion processes, see [TT90, BT96]. For a function $f : \mathbb{R}^d \to \mathbb{R}$ being either Lipschitz-continuous (for the L^1-Wasserstein distance) or measurable bounded (for the total variation distance), the *domino strategy* consists in a step-by-step decomposition of the weak error to produce an upper bound as follows:

$$\begin{aligned} |\mathbb{E}f(\bar{X}_n^x) - \mathbb{E}f(X_{\Gamma_n}^x)| &= |\bar{P}_{\gamma_1} \circ \cdots \circ \bar{P}_{\gamma_n} f(x) - P_{\Gamma_n} f(x)| \\ &= \left| \sum_{k=1}^n \bar{P}_{\gamma_1} \circ \cdots \circ \bar{P}_{\gamma_{k-1}} \circ (\bar{P}_{\gamma_k} - P_{\gamma_k}) \circ P_{\Gamma_n - \Gamma_k} f(x) \right| \\ &\leq \sum_{k=1}^n \left| \bar{P}_{\gamma_1} \circ \cdots \circ \bar{P}_{\gamma_{k-1}} \circ (\bar{P}_{\gamma_k} - P_{\gamma_k}) \circ P_{\Gamma_n - \Gamma_k} f(x) \right|, \end{aligned} \tag{1.1.31}$$

where P and $\bar{P}$ are the transition kernels associated to X and $\bar{X}$ respectively. Then three terms appear:

- *The time discretization error*, for large k, corresponding to the error between $\bar{X}$ and X with small time horizon and where the error is controlled by classic weak and strong bounds on the error of the Euler-Maruyama scheme.
- *The first ergodic error*, for small k, corresponding to the error between $\bar{X}$ and X with large time horizon. In this case, the ergodic properties of X also apply to $\bar{X}$.
- *The second ergodic error*, corresponding to the distance between X and its invariant measure ν.

The case of the total variation distance is more difficult to deal with than the case of the L^1-Wasserstein distance, since $\mathrm{d_{TV}}([\bar{X}_t^x], [X_t^x])$ is in general difficult to bound for small $t > 0$. Instead, [PP23] relies on Malliavin bounds using regularization properties of the semi-group.

1.2 Langevin equation and Langevin algorithms

1.2.1 The Langevin Equation

If the function V to be minimized in (1.1.1) is not convex, the gradient descent algorithm can be trapped into a local minimum which is not the global minimum, see Figure 1.7. Moreover,

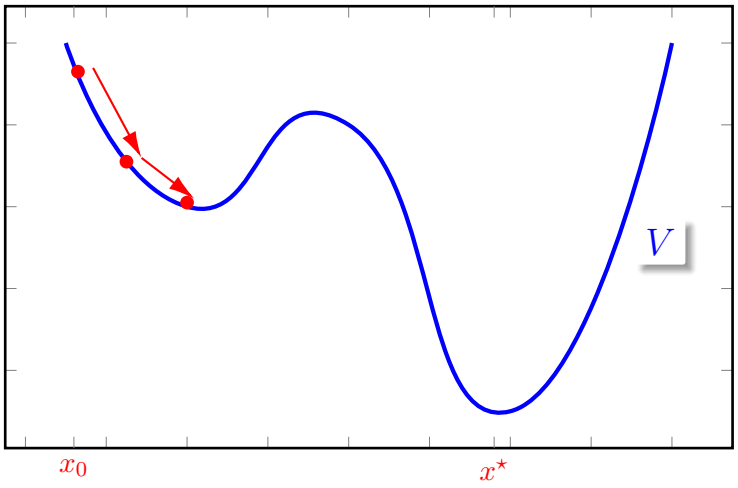


Figure 1.7: Example of trap for the gradient descent.

saddle points of V may slow down the convergence of the algorithm [DPG+14]. The stochastic gradient descent (1.1.2), adding a noise to the classic gradient descent, helps escaping from such traps [Laz92, BD96]. Another possibility is to add an exogenous noise to the gradient descent, which also adds regularization and stability to the algorithm. Considering (1.1.2) and adding a white noise yields:

$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1}\left(\nabla V(\bar{X}_n) + \zeta_{n+1}\right) + \sigma\sqrt{\gamma_{n+1}}U_{n+1}, \tag{1.2.1}$$

where (U_n) is i.i.d, $U_1 \sim \mathcal{N}(0, I_d)$ and $\sigma > 0$. Let us point out the fundamental difference between the noise $\gamma_{n+1}\zeta_{n+1}$ associated to the stochastic approximation of the gradient and the exogenous white noise $\sqrt{\gamma_{n+1}}U_{n+1}$, in particular since $\gamma_n \to 0$ the first one is of order γ_n and the second one of order $\sqrt{\gamma_n}$. This algorithm, called Stochastic Gradient Langevin Dynamics (SGLD) was introduced by Welling and Teh in [WT11]. Langevin algorithms as (1.2.1) are inspired by stochastic analysis and stochastic differential equations. Indeed, the continuous version of (1.2.1) is

$$dX_t = -\nabla V(X_t)dt + \sigma dW_t, \tag{1.2.2}$$

where W is a standard d-dimensional Brownian motion. Equation (1.2.2) was introduced by Langevin in 1908 [Lan08] to model the random movement of a particle in a fluid colliding with the other particles. Assuming that $\exp(-2V/\sigma^2) \in L^1(\mathbb{R}^d)$ and denoting $V^\star := \min_{\mathbb{R}^d} V$, the Langevin equation (1.2.2) admits an invariant measure which is the Gibbs measure given by the density

$$\nu_\sigma(dx) := \mathcal{Z}_\sigma e^{-2(V(x)-V^\star)/\sigma^2}dx \quad \text{with } \mathcal{Z}_\sigma := \left(\int_{\mathbb{R}^d} e^{-2(V(x)-V^\star)/\sigma^2}dx\right)^{-1}. \tag{1.2.3}$$

Indeed, for $g : \mathbb{R}^d \to \mathbb{R}$ being $\mathcal{C}^2$ with compact support, a quick calculation shows that

$$\begin{aligned}\frac{d}{dt}\mathbb{E}_{\nu_\sigma}[g(X_t)] &= \frac{d}{dt}\int_{\mathbb{R}^d}\mathbb{E}_x[g(X_t)]\nu_\sigma(dx) = \int_{\mathbb{R}^d}\mathcal{L}g(x)\nu_\sigma(dx)\\ &= \mathcal{Z}_\sigma\int_{\mathbb{R}^d}\big(-\nabla V(x)\cdot\nabla g(x) + \frac{\sigma^2}{2}\Delta g(x)\big)e^{-2(V(x)-V^\star)/\sigma^2}dx\\ &= \mathcal{Z}_\sigma\int_{\mathbb{R}^d}\nabla\cdot\big(\nabla g(x)e^{-2(V(x)-V^\star)/\sigma^2}\big)dx = 0,\end{aligned} \tag{1.2.4}$$

where we used an integration by parts formula to get the last line, where $\mathcal{L}$ is the infinitesimal generator associated to (1.2.2) and $\nabla\cdot$ denotes the divergence operator.

Moreover, for small σ, the measure ν_σ is concentrated around the set $\operatorname{argmin}(V)$. The following proposition makes this last statement precise.

Proposition 1.2.1. *Let $V : \mathbb{R}^d \to \mathbb{R}$ be a Borel function such that*

$$V^\star := \operatorname{essinf}(V) = \inf\{y : \ \lambda_d\{V \leq y\} > 0\} > -\infty.$$

Assume also that $e^{-V} \in L^1(\mathbb{R}^d)$ where λ_d denotes the Lebesgue measure on $\mathbb{R}^d$. Then

$$\forall \varepsilon > 0, \ \nu_\sigma(\{V \geq V^\star + \varepsilon\}) \to 0 \quad \text{as } \sigma \to 0.$$

Then if (X_t) defined in (1.2.2) is ergodic, solving (1.2.1) or (1.2.2) for small σ and large t guarantees that X_t is close to $\operatorname{argmin}(V)$ in some sense.

More precisely, convergence rates of Gibbs measures were studied by Hwang in 1980 [Hwa80]. Assuming that $\operatorname{argmin}(V) = \{x_1^\star, \dots, x_r^\star\}$ is finite and that each $\nabla^2 V(x_i^\star)$, $1 \leq i \leq r$, is positive definite then ν_σ weakly converges to a mixture of Dirac measures at the $x_i^\star$'s at rate σ.

[HRSS21] gives another heuristic for the Langevin equation and shows that some non-convex optimization problems can be seen as convex optimization problems when considering functionals defined on the infinite-dimensional space of probability measures on $\mathbb{R}^d$. The Brownian noise then appears as an entropic regularization term over the space of measures.

In the setting of Bayesian inference as presented in Section 1.1.4, [DM17, BDMS19, DM19, MMS20] study the method to sample from some target distribution ν defined on $\mathbb{R}^d$ from the Langevin equation with constant noise coefficient $\sigma \in \mathbb{R}$ (1.2.2) [WT11]. Indeed, assuming that ν has positive probability density $x \mapsto \nu(x)$, then ν is the invariant measure of (1.2.2) with

$$V(x) = -\frac{\sigma^2}{2}\log(\nu(x)), \quad \nabla V(x) = -\frac{\sigma^2}{2}\frac{\nabla\nu(x)}{\nu(x)}. \tag{1.2.5}$$

[DM17, DM19] give bounds for the total variation distance and the L^2-Wasserstein distance between the discrete algorithm $\bar{X}_n$ in (1.2.1) and the target distribution ν. [MMS20] focuses on the scheme with constant step size $\gamma_n \equiv \gamma > 0$ and analyses the resulting asymptotic bias, giving bounds for the L^1 and L^2-Wasserstein distances. As this is generally the case for MCMC methods, this method only requires the knowledge of the density of ν up to some normalization constant.

These results are valid only for constant coefficient σ. In the additive setting, the exact diffusion and its Euler scheme are linked through the Girsanov formula so that the error is bounded using the Pinsker inequality [PP23, Appendix B]. However, this strategy cannot be applied to the multiplicative noise case, which turns out to be more demanding. [PP23], which focuses on sampling from some distribution using (1.1.23), appears as an extension of [DM17] to the case where σ is not constant.

1.2.2 Adaptive Langevin algorithms

Choosing in (1.2.1) and (1.2.2) a constant noise coefficient $\sigma > 0$, yielding an homogeneous and isotropic white noise σdW_t may not be relevant in general. Indeed, the components of the gradient $\nabla V(X_t)$ may widely vary or may be correlated. To accelerate the convergence of X_t to its invariant distribution, we rework (1.2.2) by allowing σ to depend on the position X_t, leading to devise an adaptive Langevin algorithm. Allowing σ to be adaptive highly extends the range of applications of Langevin algorithms. The case where σ is not constant is called

the "multiplicative case" whereas the case where σ is constant is called the 'additive case". A general heuristic is the so-called Newton method, which consists in considering $\sigma = (\nabla^2 V)^{-1}$ and thus adding more noise in the directions where the gradient of V slowly varies in order to accelerate the exploration of the state space. However the exact computation of $\sigma = (\nabla^2 V)^{-1}$ is cumbersome because

- In high-dimensional optimization problems where the dimension d is large, the Hessian matrix has size $d \times d$ which is too large for both computation time and memory size.
- It becomes necessary to compute second-order derivatives whereas we only need to compute first-order derivatives for "classic" SGLD
- As the size $d \times d$ of the Hessian matrix is large, computing its inverse is also costly.

[DdVB15] adapted the well-known Newton method, which consists in considering $\sigma = (\nabla^2 V)^{-1}$, to SGLD. Since the size of the Hessian matrix may be too large in practice, because inverting it is computationally costly and because the Hessian matrix may not be positive in every point, it is suggested to consider instead $|\mathrm{diag}((\nabla^2 V))^2|^{-1}$. However, computing high-order derivatives may be cumbersome; [SBCR16] adapts the quasi-Newton method [NW06] to approximate the Hessian matrix to SGLD, yielding the Stochastic Quasi-Newton Langevin algorithm.

[DHS11] and [LCCC16] give algorithms where the choice for σ is $\sigma \simeq \mathrm{diag}\left((\lambda I_d + [|\partial_i V|]_{1\leq i\leq d})^{-1}\right)$, where $\lambda > 0$ guarantees numerical stability. The idea of using geometry has been explored in [PT13], where σ^{-2} defines the local curvature of a Riemannian manifold, giving the Stochastic Gradient Riemaniann Langevin Dynamics algorithm where σ is equal to $\mathcal{I}_x^{-1/2}$ where $\mathcal{I}_x$ is the Fisher information matrix, or to some other choices (see [PT13, Table 1]) as $\mathcal{I}_x$ may be intractable. [MCF15] extends the previous algorithm to Hamiltonian Monte Carlo methods, where a momentum variable is added in order to take into account the "inertia" of the trajectory [Nea96, DMS20], yielding the Stochastic Gradient Riemannian Hamiltonian Monte Carlo method.

Allowing the matrix σ to depend on the position yields a faster convergence; we refer to the previous references where the simulations prove that the new methods greatly improve classical stochastic gradients algorithms. In particular, we refer to the simulations [SBCR16, Figure 2], [PT13, Figure 2] and [MCF15, Figure 3] where the some of the above different methods are compared.

As presented in [MCF15] and [LCCC16, Equation (3)] and formally demonstrated in [PP23, Proposition 2.6], if σ is not constant then we need to add a correction term to the drift so that the Gibbs measure ν_σ remains the invariant measure.

Proposition 1.2.2. *Let $a > 0$ and let (X_t) denote the solution to the $\mathbb{R}^d$-valued SDE in $\mathbb{R}^d$*

$$dX_t = b(X_t)dt + a\sigma(X_t)dW_t, \quad X_0 \perp\!\!\!\perp W. \tag{1.2.6}$$

Assume that $\nabla^2 V$ is bounded, $e^{-2V/a^2} \in L^1(\mathbb{R}^d)$, σ is $\mathcal{C}^1$ and uniformly elliptic i.e.

$$\exists\ \underline{\sigma}_0 > 0,\ \forall x \in \mathbb{R}^d,\ \sigma(x)\sigma^\top(x) \geq \underline{\sigma}_0,$$

and bounded with bounded partial derivatives. If the drift b satisfies

$$b(x) = -(\sigma\sigma^\top \nabla V)(x) + a^2 \left[\sum_{j=1}^{d} \partial_j(\sigma\sigma^\top)(x)_{ij}\right]_{1\leq i\leq d}, \tag{1.2.7}$$

then the distribution ν_a defined in (1.2.3) *is the unique invariant distribution of the above Brownian diffusion.*

Defining the correction term $\Upsilon(x) := [\sum_{j=1}^d \partial_j(\sigma\sigma^\top)(x)_{ij}]_{1\le i\le d}$ which reduces to zero in the additive case, the discrete algorithm corresponding to the continuous SDE (1.2.6) reads

$$X_{n+1} = X_n - \gamma_{k+1}\sigma\sigma^\top(\nabla V(X_n) + \zeta_{n+1}) + a^2\gamma_{n+1}\Upsilon(X_n) + a\sqrt{\gamma_{n+1}}\sigma(X_n)U_{n+1}, \qquad (1.2.8)$$

where $(U_n)_{n\ge1}$ is i.i.d. and $\mathcal{N}(0, I_d)$-distributed.

Let us provide a simple example of how the above extension to non constant diffusion coefficients for Langevin simulation can be used, see [PP23, Section 2.4]. We consider the pseudo-Cauchy distribution on $\mathbb{R}^d$ with exponent $\kappa > 0$ defined by

$$\pi_\kappa(dx) = \frac{C_\kappa}{(1+x^2)^{1+\kappa}}dx = C_\kappa e^{-V(x)}dx \quad \text{with} \quad V(x) = (d+\kappa)\log(1+x^2) + 1.$$

Then using Proposition 1.2.2 with $\sigma = I_d$, the distribution π_κ is the invariant distribution of the one-dimensional Brownian diffusion

$$dY_t = -(d+\kappa)\frac{Y_t}{1+|Y|_t^2}dt + dW_t.$$

However, we have in this case $|b(x)| \to 0$ as $|x| \to \infty$, suggesting that the convergence to π_κ is slow. On the other hand, using Proposition 1.2.2 with $\sigma(x) = (1+|x|^2)^{1/2}I_d$, the measure π_κ is also the invariant distribution of the Brownian diffusion

$$dX_t = -(d+\kappa-1)X_t dt + \sqrt{1+|X|_t^2}\, dW_t$$

where the drift satisfies

$$\langle b(x) - b(y), x-y\rangle = -(\kappa - 1 + d/2)|x-y|^2$$

so that Assumption (1.1.28) is satisfied for $\kappa > 1 - d/2$, suggesting that the convergence to π_κ is fast. For more details in particular concerning mean-reverting diffusions we refer to [PP23, Section 2.4].

1.2.3 Langevin-Simulated anneling equation

Considering (1.2.2) with invariant measure ν_σ defined in (1.2.3) and since ν_σ is concentrated around $\operatorname{argmin}(V)$ for small σ (Proposition 1.2.1), a method for minimizing V is to solve (1.2.2) while making the drift coefficient decrease to zero along time, leading to the Langevin-simulated annealing equation:

$$dX_t = -\nabla V(X_t) + a(t)\sigma dW_t, \qquad (1.2.9)$$

where X_t is $\mathbb{R}^d$-valued, $\sigma > 0$ and $a : \mathbb{R}^+ \to \mathbb{R}^+$ is decreasing and $a(t) \to 0$ as $t \to \infty$. Equation (1.2.9) is a non-homogeneous Markov SDE; for every $t \in \mathbb{R}^+$ we can still define the "instantaneous" invariant measure as the invariant measure of the corresponding time-homogeneous SDE when freezing the time dependent coefficient $a(t)$. In our case it is the measure $\nu_{a(t)\sigma}$. The Langevin-simulated annealing equation shares indeed its heuristic with the original simulated annealing algorithm [KGV83, vLA87], which builds a Markov chain from the Gibbs measure ν_σ using the Metropolis-Hastings algorithm [MRR$^+$53] and where the parameter σ, interpreted as a temperature, slowly decreases to zero over the iterations. In Statistical Physics, the parameter sequence $(a(t)\sigma)$ is interpreted as the square root of the system temperature. At the beginning of the algorithm, the temperature is high, allowing to explore the state space more efficiently to the detriment of the optimization procedure; at the end the temperature is low allowing to focus on the minima of V.

In [CHS87, Roy89] is shown that choosing $a(t) = A\log^{-1/2}(t)$ for some $A > 0$ in (1.2.9) guarantees the convergence of X_t to $\nu^\star$ defined as the limit measure of (ν_a) as $a \to 0$, $\nu^\star$ being supported by $\operatorname{argmin}(V)$. [Mic92] proves again the convergence of the SDE using free energies inequalities. These studies deeply rely on some Poincaré and log-Sobolev inequalities and require the following assumptions on the potential function:

$$\lim_{|x|\to\infty} V(x) = \lim_{|x|\to\infty} |\nabla V(x)| = \infty \quad \text{and} \quad \forall x \in \mathbb{R}^d,\ \Delta V(x) \leq C + |\nabla V(x)|^2.$$

More specifically, taking $\sigma = 1$ in (1.2.9) and defining the Kullback-Liebler divergence (or free energy):

$$\mathcal{J}_t := \mathrm{d}_{\mathrm{KL}}\left(X_t \| \nu_{a(t)}\right) = \int_{\mathbb{R}^d} \log\left(\frac{p(t,x)}{\nu_{a(t)}(x)}\right) p(t,x)dx, \tag{1.2.10}$$

where $p(t,x)$ is the density of X_t at $x \in \mathbb{R}^d$ (see [Fri64, Chapter 9, Theorem 7]) and follows the Fokker-Planck equation:

$$\partial_t p(t,x) = \nabla \cdot (\nabla V(x) p(t,x)) + \frac{1}{2}a^2(t)\Delta p(t,x), \tag{1.2.11}$$

using the expression of $\partial_t p(t,x)$ in (1.2.11), the explicit expression of the density $\nu_{a(t)}(x)$ and the integration by parts formula, we obtain

$$\frac{d\mathcal{J}_t}{dt} = \frac{2}{a^4(t)}\frac{da^2(t)}{dt}\int_{\mathbb{R}^d} V(x)(\nu_{a(t)}(x) - p(t,x))dx - 2a^2(t)\int_{\mathbb{R}^d}\left|\nabla\sqrt{\frac{p(t,x)}{\nu_{a(t)}(x)}}\right|^2 \nu_{a(t)}(x)dx. \tag{1.2.12}$$

We then use a logarithmic Sobolev inequality:

$$\int_{\mathbb{R}^d} f^2\log(f^2)d\nu_{a(t)} \leq C\int_{\mathbb{R}^d}|\nabla f|^2 d\nu_{a(t)} + \left(\int_{\mathbb{R}^d} f^2 d\nu_{a(t)}\right)\log\left(\int_{\mathbb{R}^d} f^2 d\nu_{a(t)}\right) \tag{1.2.13}$$

with

$$f(x) = \sqrt{p(t,x)/\nu_{a(t)}(x)}$$

so that

$$-2\sigma^2(t)\int_{\mathbb{R}^d}\left|\nabla\sqrt{\frac{p(t,x)}{\nu_{a(t)}(x)}}\right|^2 \nu_{a(t)}(x)dx \leq -C\int_{\mathbb{R}^d}\frac{p(t,x)}{\nu_{a(t)}(x)}\log\left(\frac{p(t,x)}{\nu_{a(t)}(x)}\right)\nu_{a(t)}(x)dx = -C\mathcal{J}_t.$$

On the other hand,

$$\int_{\mathbb{R}^d} V(x)(\nu_{a(t)}(x) - p(t,x))dx = \nu_{a(t)}(V) - \mathbb{E}V(X_t)$$

can be bounded independently of t using classic bounds on the growth of the potential, see for example [PP23, Proposition A.1]. Then (1.2.12) gives a bound on $d\mathcal{J}_t/dt$ and solving the differential inequation with $a(t) = A\log^{-1/2}(t)$ yields the convergence of $\mathcal{J}_t$ to 0.

[Zit08] proves that the convergence still holds under weaker assumptions, in particular where the gradient of the potential is not coercive, using weak Poincaré inequalities. In [GM91] is proved the convergence of the associated stochastic gradient descent algorithm.

1.3 Contributions of the thesis

Our research results and articles are available on my personal research website at the address:

`https://perso.lpsm.paris/~pbras/`.

Our code for simulations along with Jupyter notebooks are available on my GitHub page at the address:

`https://github.com/Bras-P`.

1.3.1 Convergence of Langevin-Simulated annealing algorithms

1.3.1.1 Convergence of the adaptive Langevin equation and its Euler-Maruyama scheme for the L^1-Wasserstein distance and in total variation

In Chapters 2 and 3, we prove the convergence and give convergence rates for Langevin-simulated annealing algorithms with multiplicative noise i.e. for the SDE (1.2.6) and its associated Euler-Maruyama scheme (1.2.8) where the coefficient of the exogenous noise $(a(t))$ decreases to 0 as $t \to \infty$. The target measure denoted $\nu^\star$ is defined as the limit measure of ν_a as $a \to 0$ and the rates of convergence are given for the L^1-Wasserstein distance and the total variation distance. We make assumptions on V and σ similar to those in [PP23], in particular that $|\nabla V|^2 \leq CV$, ∇V is Lipschitz continuous and σ is bounded with bounded derivatives. We also assume that $\operatorname{argmin}(V)$ is finite and that for every $x^\star \in \operatorname{argmin}(V)$, either $\nabla^2 V(x^\star)$ is definite positive or $x^\star$ is a strict polynomial minimum.

We adopt a *domino* strategy (see (1.1.31) and the description following) and use methods described in Section 1.1.3.1 and in Section 1.1.4 however we need to adapt the strategy to non-homogeneous Markov diffusion processes with evanescent ellipticity, since the diffusion coefficient of the SDE depends on the time t and its ellipticity fades away as $t \to \infty$. We prove that choosing a coefficient $(a(t))$ of order $\log^{-1/2}(t)$ is a sufficient - and generally necessary - condition for convergence. We give a convergence rate for $\mathcal{W}_1([X_t], \nu^\star)$ which turns out to be somehow limited by $\mathcal{W}_1(\nu_{a(t)}, \nu^\star)$ which is of order $a(t) \simeq \log^{-1/2}(t)$ under the assumption that $\nabla^2 V$ is positive definite at every point of $\operatorname{argmin}(V)$. Still we establish sharper bounds for the convergence of $\mathcal{W}_1(X_t, \nu_{a(t)})$:

$$\mathcal{W}_1([X_t^{x_0}], \nu^\star) + \mathcal{W}_1([\bar{X}_{N(t)}^{x_0}], \nu^\star) \leq C \max(1 + |x_0|, V^2(x_0)) a(t),$$
$$\mathcal{W}_1([X_t^{x_0}], \nu_{a(t)}) + \mathcal{W}_1([\bar{X}_{N(t)}^{x_0}], \nu_{a(t)}) \leq C_\alpha \max(1 + |x_0|, V^2(x_0)) t^{-\alpha}$$

for every $\alpha \in (0, 1)$ with $N(t) = \min\{k \in \mathbb{N},\ \Gamma_{k+1} > t\}$, see Theorems 2.2.1 and 2.2.4.

For the total variation distance, we obtain for every $\alpha \in (0, 1)$:

$$\mathrm{d_{TV}}([X_t^{x_0}], \nu_{a(t)}) \leq C_\alpha e^{C \log^{1/2}(t)(1+|x_0|)} t^{-\alpha}, \tag{1.3.1}$$
$$\mathrm{d_{TV}}([\bar{X}_{N(t)}^{x_0}], \nu_{a(t)}) \leq C_\alpha \left(\log^{1/2}(t) \max(V^2(x_0), 1 + |x_0|) t^{-\alpha} + e^{C \log^{1/2}(t)(1+|x_0|^2)} t^{C/A^2} \gamma_{N(Ct)}^{1/2} \right) \tag{1.3.2}$$

with A defined by $a(t) = A \log^{-1/2}(e + t)$, see Theorem 3.2.1.

These results have been submitted as two separate articles as joints work with Gilles Pagès:

- The analysis of the convergence for the L^1-Wasserstein distance, entitled *Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise* and accepted for publication in *Mathematics of Computation*,
- The analysis of the convergence for the total variation distance, entitled *Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise II: Total Variation* and which has been published in *Monte Carlo Methods and Applications*.

1.3.1.2 Convergence rates of Gibbs measures

In order to analyse the rate of convergence of Langevin algorithms, for $\mathfrak{D}$ some distance on $\mathcal{P}(\mathbb{R}^d)$, let us write

$$\mathfrak{D}([X_t], \nu^\star) \leq \begin{cases} \mathfrak{D}([X_t], \nu_a) + \mathfrak{D}(\nu_a, \nu^\star) & \text{if } a > 0 \text{ is constant,} \\ \mathfrak{D}([X_t], \nu_{a(t)}) + \mathfrak{D}(\nu_{a(t)}, \nu^\star) & \text{if } a \text{ is decreasing to 0.} \end{cases} \tag{1.3.3}$$

The second term in this inequality $\mathfrak{D}(\nu_a, \nu^\star)$ corresponds to the bias term. It measures the distance between the invariant distribution $\nu^\star$ and its approximation using a Gibbs measure ν_a and converges to 0 as $a \to 0$. Consequently we also need to investigate the order of $\mathfrak{D}(\nu_a, \nu^\star)$ for small $a > 0$. It is known that if $\operatorname{argmin}(V)$ is finite and if for every $x^\star \in \operatorname{argmin}(V)$, $\nabla^2 V(x^\star)$ is definite positive then ν_a converges weakly to $\nu^\star$ which is a weighted sum of Dirac masses where each weight is proportional to $\det^{-1/2}(\nabla^2 V(x^\star))$ and the rate of convergence is of order a as $a \to 0$ [Hwa80]. However, there are few results available when $\nabla^2 V(x^\star)$ is not positive definite for some $x^\star \in \operatorname{argmin}(V)$. Assuming $\operatorname{argmin}(V) = \{x^\star\}$ for simplifying notations, [AH10] gives a convergence rate under the condition that there exists a function $g : \mathbb{R}^d \to \mathbb{R}$ with $e^{-g} \in L^1(\mathbb{R}^d)$ and $\alpha_1, \ \ldots, \ \alpha_d \in (0, +\infty)$ such that

$$\forall h \in \mathbb{R}^d, \quad \frac{1}{a}\left[V(x^\star + (a^{\alpha_1} h_1, \ldots, a^{\alpha_d} h_d)) - V(x^\star)\right] \underset{a \to 0}{\longrightarrow} g(h_1, \ldots, h_d). \tag{1.3.4}$$

In Chapter 4 we give conditions on V such that (1.3.4) is fulfilled and then elucidate the expression of g depending on V and its successive derivatives at $x^\star$. Instead of positive definiteness we assume $x^\star$ is a strictly polynomial minimum of V i.e. $V - V(x^\star)$ is bounded below in a neighbourhood of $x^\star$ by some non-negative polynomial function null only at $x^\star$. Under some assumptions, we give an algorithm to identify $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d$, an orthogonal transformation B and $g : \mathbb{R}^d \to \mathbb{R}$ a polynomial function which is not constant in any of its variables, depending on the successive derivatives of V such that

$$\forall h \in \mathbb{R}^d, \ \frac{1}{a}(V(x^\star + B \cdot (a^\alpha * h)) - V(x^\star)) \underset{a \to 0}{\longrightarrow} g(h)$$

where a^α is the vector $(a^{\alpha_1}, \ldots, a^{\alpha_d})$ and $*$ denotes the Schur product between two vectors i.e. for $u, v \in \mathbb{R}^d$, $u * v = (u_1 v_1, \ldots, u_d v_d)$. Then we obtain the following central limit theorem:

$$a^{-\alpha} * (B^{-1} \cdot (Y_a - x^\star)) \underset{a \to 0}{\longrightarrow} Y \quad \text{in law}$$

where $Y_a \sim \nu_{\sqrt{2a}}$ and $Y \sim e^{-g(y)} dy$. We refer to Theorems 4.3.3 and 4.3.5.

Although the case where $\nabla^2 V(x^\star)$ is not positive definite may seem singular and mostly of theoretical interest at first sight, it can actually occur in practice as pointed out by eminent Machine Learning researchers, among them L. Bottou and Y. LeCun [SBL16, SEU+17], for high-dimensional and over-parametrized optimization and inference problems, especially for the calibration of artificial neural networks.

These results have been published as the following article: Pierre Bras. Convergence rates of Gibbs measures with degenerate minimum. *Bernoulli*, 28(4):2431 – 2458, 2022.

1.3.2 Adaptive Langevin algorithms for deep Neural Networks

1.3.2.1 SGLD algorithms for deep neural networks, Layer Langevin algorithm, application to image classification

In Chapter 5 we implement and run simulations to evaluate the performances of SGLD algorithms described in Section 1.2.2 with different choice for the function σ on various problems coming from Machine Learning: regression, classification, image recognition, time series analysis etc. In particular we analyse the benefits of adding Gaussian noise during the training in comparison with the non-Langevin counterpart algorithms. More precisely, popular preconditioned stochastic gradient methods in Machine Learning such as RMSprop [TH12], Adam and Adamax [KB15] and Adadelta [Zei12] generally writes in a non-Langevin setting

$$X_{n+1} = X_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1} \tag{1.3.5}$$

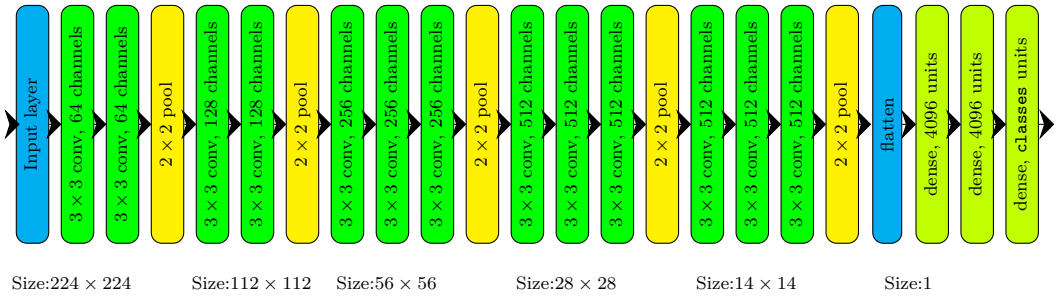


Figure 1.8: Architecture of the VGG-16 network for an input image of size 224×224.

where g_{n+1} is an estimation of the gradient $\nabla V(X_n)$ and P_{n+1} is some preconditioner rule, while similarly to (1.2.8), the corresponding Langevin counterpart method writes

$$X_{n+1} = X_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1} + a\gamma_{n+1}^{1/2} \mathcal{N}(0, P_{n+1}) + a^2 \gamma_{n+1} \big[\sum_{j=1}^{d} \partial_j P_{n+1}(X_n)_{ij} \big]_{1 \le i \le d}. \quad (1.3.6)$$

In Chapter 6 we proceed to a side by side comparison of (1.3.5) and (1.3.6) for different choices of stochastic gradient methods. As it was noted in [NVL+15, Ani19], adding gradient noise can in fact improve the learning for very deep neural networks. Indeed, the noise provides regularization and allows to escape from traps for the gradient descent such as local minima or saddle points [DPG+14]. Moreover, the deeper the neural network is, the more non-linear it is, thus increasing the number of such traps. Many advances in supervised learning were made possible using very deep neural networks, which are able to tackle much more difficult problems than shallow ones [KSH12, MPCB14, LBH15], in particular as it comes to image classification, involving very deep convolutional architectures [SLJ+15, SZ15, HZRS16, HLVDMW17]. An example of deep convolution architecture is given in Figure 1.8.

Still, as a price to be paid, deep neural networks are considerably more difficult to train [GB10, DPG+14]. To cope with this issue, highway networks [SGS15] and residual networks [HZRS16] were introduced. Their many successive layers behaves either as a dense layer or as the identity function, allowing the gradient information to propagate trough the successive layers.

We compare the benefits of preconditioned Langevin algorithms [LCCC16] for various architectures and for various depths of neural networks and we observe that the deeper the network is, the greater are the gains provided by Langevin algorithms. Based on this heuristic and since the most important non-linearities of the network are contained in the deepest layers, we introduce a new optimization method that we call Layer Langevin algorithm, which consists in training the network by adding Langevin noise only to the training of some weights and not to the other weights. That is, the stochastic gradient method described in (1.3.6) becomes

$$\begin{aligned} X_{n+1}^{(i)} = X_n^{(i)} - \gamma_{n+1} \big[P_{n+1} \cdot g_{n+1} \big]^{(i)} + \mathbb{1}_{i \in \mathcal{J}} a \gamma_{n+1}^{1/2} \big[\mathcal{N}(0, P_{n+1}) \big]^{(i)} \\ + \mathbb{1}_{i \in \mathcal{J}} a^2 \gamma_{n+1} \big[\sum_{j=1}^{d} \partial_j P_{n+1}(X_n)_{ij} \big]^{(i)}, \quad 1 \le i \le d \end{aligned}$$

for some $\mathcal{J}$ subset of $\{1, \dots, d\}$. In particular, we choose $\mathcal{J}$ to be the set of indexes of the weights of the k first (deepest) layers for some integer k. We refer to Section 6.4. We then highlight the possibilities of training acceleration using Langevin and Layer Langevin methods on deep residual [HZRS16] and dense convolutional networks [HLVDMW17] for image classification.

We give an implementation of Langevin and Layer Langevin algorithms in `TensorFlow` as instances of the base class `tf.keras.optimizers. Optimizers`. These optimizers are directly

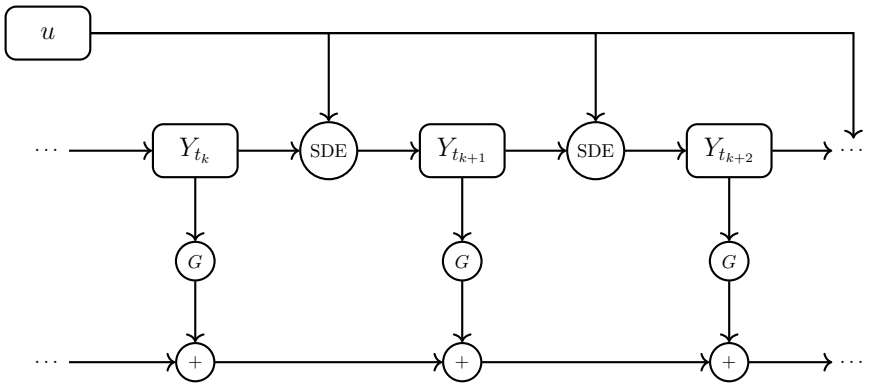


Figure 1.9: Depth of Markovian neural networks. The control u acts on Y_{t_k}, which itself acts on $Y_{t_{k+1}}$, $Y_{t_{k+2}}$, ..., Y_{t_N}, hence the depth of the network.

usable as the `Python` package `langevin_optimizers`, which is downloadable from the GitHub repository `https://github.com/Bras-P/langevin-for-stochastic-control` and with the command

```
$ pip install git+https://github.com/Bras-P/langevin-for-stochastic-control.git
```

These results have been presented at the *International Neural Network Society Workshop on Deep Learning Innovations and Applications* (INNS DLIA), part of the Machine Learning conference *International Joint Conference on Neural Networks* IJCNN 2023, and will be published in the first edition of the INNS workshop series in *Procedia Computer Science* as *Langevin algorithms for very deep Neural Networks with application to image classification*.

1.3.2.2 Langevin algorithms for deep stochastic control

In Chapter 7, we evaluate the benefits of Langevin algorithms on solving stochastic optimal control problems using neural networks as described in Section 1.1.2.3. Indeed, if the control is parametrized by a neural network and if it is applied at many discretization time steps, then the stochastic control problem reads as the training of a very deep neural network, see Figure 1.9.

As in Section 1.3.2.1, we compare preconditioned Langevin algorithms with their respective non-Langevin counterparts for solving various stochastic optimal control problems: fishing quotas [LPP23], deep hedging of financial options [BGTW19], oil drilling and resource management [GGKL21], and we show that Langevin and Layer Langevin optimizers can significantly improve the training procedure.

These results have been presented at the Machine Learning conference *International Joint Conference on Neural Networks* IJCNN 2023 and are published in the conference proceedings as a joint work with Gilles Pagès and under the title *Langevin algorithms for Markovian Neural Networks and Deep Stochastic Control*.

1.3.3 Simulation of stochastic processes and discretization schemes

1.3.3.1 Total variation convergence of the Euler-Maruyama scheme in small time

For the total variation bound in (1.3.2), while implementing the domino strategy described in Section 1.1.4 we need to establish a bound for the total variation between an SDE and its

one-step Euler-Maruyama scheme in small time i.e. a bound for $\mathrm{d_{TV}}([\bar{X}_t^x],[X_t^x])$ where

$$X_0^x = x, \quad dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad \bar{X}_t = x + tb(x) + \sigma(x)W_t$$

where b and σ are general Lipschitz continuous coefficients here. Indeed, the difficulty of the total variation distance in small time is the following: considering its representation formula and comparing it with the L^1-Wasserstein distance, if x and $y \in \mathbb{R}^d$ are close to each other and if $f : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous, then we can bound $|f(x) - f(y)|$ by $[f]_{\mathrm{Lip}}|x - y|$; whereas if f is simply measurable and bounded, then we cannot directly bound $|f(x) - f(y)|$ in terms of $|x - y|$. Previous results in the literature only focus on total variation bounds for fixed time horizon $T > 0$ [BT96, GL08].

In Chapter 8, we show more generally that for two SDEs in $\mathbb{R}^d$ starting at the same point:

$$\begin{aligned} X_0^x &= x \in \mathbb{R}^d, & dX_t^x &= b_1(X_t^x)dt + \sigma_1(X_t^x)dW_t, \\ Y_0^x &= x, & dY_t^x &= b_2(Y_t^x)dt + \sigma_2(Y_t^x)dW_t \end{aligned}$$

and assuming that for $i = 1, 2$, $\sigma_i \in \mathcal{C}_b^2$ and is elliptic and $b \in \mathcal{C}^1$ with bounded derivatives, we obtain

$$\mathrm{d_{TV}}([X_t^x],[Y_t^x]) \leq C(t^{1/2} + |\sigma_1 - \sigma_2|(x))^{2/3} + Ce^{c|x|^2}t^{1/2}, \tag{1.3.7}$$

see Theorem 8.2.1. This bound relies on the following result for $r = 1$ which links the total variation distance with the L^1-Wasserstein distance provided that the laws of X_t^x and Y_t^x are regular enough:

Theorem 1.3.1. *Let Z_1 and Z_2 be two random vectors in $L^1(\mathbb{R}^d)$ and admitting densities p_1 and p_2 respectively with respect to the Lebesgue measure. Assume furthermore that p_1 and p_2 are $\mathcal{C}^{2r}$ with $r \in \mathbb{N}$ and that $\nabla^k p_i \in L^1(\mathbb{R}^d)$ for $i = 1, 2$ and $k = 1, \ldots, 2r$. Then we have*

$$\mathrm{d_{TV}}(Z_1, Z_2) \leq C_{d,r}\mathcal{W}_1(Z_1, Z_2)^{2r/(2r+1)} \left(\int_{\mathbb{R}^d} \left(\|\nabla^{2r} p_1(\xi)\| + \|\nabla^{2r} p_2(\xi)\| \right) d\xi \right)^{1/(2r+1)} \tag{1.3.8}$$

where the constant $C_{d,r}$ depends only on d and on r.

In the case of SDEs, the densities of X_t^x and Y_t^x can be expressed as the solution a Fokker-Planck partial differential equation and we rely on Aronson bounds [Fri64] to control their regularity.

We furthermore extend (1.3.7): assuming that $\sigma \in \mathcal{C}_b^{2r}$ we prove that

$$\mathrm{d_{TV}}([X_t^x],[Y_t^x]) \leq C(t^{1/2} + |\sigma_1 - \sigma_2|(x))^{2r/(2r+1)} + Ce^{c|x|^2}t^{1/2}, \tag{1.3.9}$$

see Theorem 8.2.2. For $r \geq 2$, we use a multi-step Richardson-Romberg extrapolation [RG11, LP17] which is a method imported from numerical analysis and which builds a Taylor expansion with null coefficients up to some high order.

These results have been published as: Pierre Bras, Gilles Pagès, and Fabien Panloup. Total variation distance between two diffusions in small time with unbounded drift: application to the Euler-Maruyama scheme. *Electronic Journal of Probability*, 27:1–19, 2022.

1.3.3.2 Weak error rates for numerical schemes of Stochastic Volterra Equations with application to stochastic volatility models and option pricing under path-dependent volatility

As a further research on convergence of Euler-Maruyama schemes for SDEs, in Chapter 9 we study the rate of weak convergence of numerical schemes for Stochastic Volterra Equations as described in Section 1.1.3.2, that is for $f : \mathbb{R}^d \to \mathbb{R}$ being smooth we give bounds on

$$\mathbb{E}[f(\bar{X}_N)] - \mathbb{E}[f(X_T)]. \tag{1.3.10}$$

A first bound on the weak error can be obtained from bounds on the strong error, however this is sub-optimal in general. For example, for Stochastic Differential Equations (SDE) the strong error is of order $O(1/\sqrt{N})$ but the weak error is of order $O(1/N)$. Such bounds get even worse in the case of SVE with fractional kernel, giving a weak error bounded by the strong error which is order $O(N^{-H})$, where $H \in (0, 1/2)$ is the Hurst parameter of the fractional kernel and is small ($H \simeq 0.1$) in many financial applications. In [RTY21] are given bounds for the weak error for the multi-level Euler-Maruyama scheme, however the authors only assume that the weak error is bounded by the strong error (see [RTY21, Section 2.3]), which is largely suboptimal in general. In [Gas23] are given weak error rates from some rough volatility models and are proved to be of order $O(N^{-(3H+1/2)\wedge 1})$, yielding significantly better bounds in the case where H is close to 0. However the results are valid only for some special cases (semilinear or cubic test function), hinting that obtaining general results for fractional processes is difficult.

As a first step, we focus on the regular case i.e. where K_1 and K_2 are defined and continuous on $[0, T]^2$. Developing a *domino* strategy (1.1.31) that we adapt to the path-dependent case and using pathwise derivatives of path-dependent functionals [VZ19, Dup19], we prove that the convergence rate of (1.3.10) is of order $O(1/N)$ in the non-singular case and where N is the number of steps in the Euler-Maruyama scheme, see Theorem 9.2.2, thus giving a convergence rate similar to the Markovian SDE case.

These results are a joint work with Masaaki Fukasawa and have been submitted to *SIAM Journal on Financial Mathematics* (*SIFIN*) and are currently in revision for possible publication. The analysis of the singular rough case from adapting from the regular case is a work in progress.

1.3.3.3 Simulation of reflected Brownian motion on two dimensional wedges

In Chapter 10 we study the Brownian motion W in $\mathbb{R}^2$ which is reflected or stopped on a wedge, i.e. the subset of $\mathbb{R}^2$ defined as

$$\mathcal{D} = \{(r\cos(\theta), r\sin(\theta)),\ r \geq 0,\ \theta \in [0, \alpha]\}$$

for some $\alpha \in (0, 2\pi)$ [Iye85, Met10, Pil14].

We prove the following density formula for the reflected Brownian motion X_t^x starting at $x = (r_0\cos(\theta_0), r_0\sin(\theta_0))$:

$$\mathbb{P}^x(X_t \in dy) = \frac{2r}{t\alpha} e^{-\frac{r^2+r_0^2}{2t}} \left(\frac{1}{2} I_0\left(\frac{rr_0}{t}\right) + \sum_{n=1}^{\infty} I_{n\pi/\alpha}\left(\frac{rr_0}{t}\right) \cos\left(\frac{n\pi\theta}{\alpha}\right) \cos\left(\frac{n\pi\theta_0}{\alpha}\right)\right) dr d\theta \tag{1.3.11}$$

where I_a, $a \geq 0$, stands for the modified Bessel function of the first kind, see Theorem 10.4.2. Unfortunately, this formula involves oscillating infinite sums of Bessel functions that are hardly usable for simulation purposes. Instead of directly computing these sums, we propose an alternative simulation method which uses an extension of the reflection principle in two dimensions for a particular type of wedges with angle π/m, $m \in \mathbb{N}$ in Sections 10.5.2 and 10.5.3. As a first step, we obtain a simulation method for $\mathbb{E}[f(W_{T\wedge\tau})]$ where τ stands for the first time the process W touches the boundary of the wedge $\mathcal{D}$. Applying the methodology for the stopped process recursively one obtains an algorithm for the reflected process X. We then extend these methods to the simulation of the reflection on $\mathcal{D}$ of the process Z with $dZ_t = b(Z_t)dt + dW_t$ for some drift coefficient b.

The reflected Brownian motion and more generally reflected processes have applications in finance (for example, in stochastic models where the process, which can model an interest rates, is constrained to be non-negative [Ha09] or has other constraints like barriers such as in [IKP13]);

in queueing models [GNR86], etc. Considering the particular case of a wedge may open the way to new simulation algorithms for reflected processes adapted to non-smooth domains with corner points.

These results have been published as the following article: Pierre Bras and Arturo Kohatsu-Higa. Simulation of reflected Brownian motion on two dimensional wedges. *Stochastic Processes and their Applications*, 156:349–378, 2023.

Part I

Convergence of adaptive Langevin-Simulated Annealing algorithms

Chapter 2

Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise for the L^1-Wasserstein distance

The results presented in this chapter have been accepted for publication in *Mathematics of Computation* as a joint work with Gilles Pagès. An arXiv preprint is available [BP21].

Abstract

We study the convergence of Langevin-Simulated Annealing type algorithms with multiplicative noise, i.e. for $V : \mathbb{R}^d \to \mathbb{R}$ a potential function to minimize, we consider the stochastic differential equation $dY_t = -\sigma\sigma^\top \nabla V(Y_t)dt + a(t)\sigma(Y_t)dW_t + a(t)^2\Upsilon(Y_t)dt$, where (W_t) is a Brownian motion, where $\sigma : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ is an adaptive (multiplicative) noise, where $a : \mathbb{R}^+ \to \mathbb{R}^+$ is a function decreasing to 0 and where Υ is a correction term. This setting can be applied to optimization problems arising in Machine Learning; allowing σ to depend on the position brings faster convergence in comparison with the classical Langevin equation $dY_t = -\nabla V(Y_t)dt + \sigma dW_t$. The case where σ is a constant matrix has been extensively studied however little attention has been paid to the general case. We prove the convergence for the L^1-Wasserstein distance of Y_t and of the associated Euler scheme $\bar{Y}_t$ to some measure $\nu^\star$ which is supported by $\mathrm{argmin}(V)$ and give rates of convergence to the instantaneous Gibbs measure $\nu_{a(t)}$ of density $\propto \exp(-2V(x)/a(t)^2)$. To do so, we first consider the case where a is a piecewise constant function. We find again the classical schedule $a(t) = A\log^{-1/2}(t)$. We then prove the convergence for the general case by giving bounds for the Wasserstein distance to the stepwise constant case using ergodicity properties.

Keywords– Stochastic Optimization, Langevin Equation, Simulated Annealing, Neural Networks.

2.1 Introduction

Langevin-based algorithms are used to solve optimization problems in high dimension and have gained much interest in relation with Machine Learning [WT11, MCF15, LCCC16, DM17]. The Langevin equation is a Stochastic Differential Equation (SDE) which consists in a gradient descent with noise. More precisely, let $V : \mathbb{R}^d \to \mathbb{R}^+$ be a coercive potential function, then the associated Langevin equation reads

$$dX_t = -\nabla V(X_t)dt + \sigma dW_t,\ t \geq 0,$$

where (W_t) is a d-dimensional Brownian motion and where $\sigma > 0$. Under standard assumptions, the invariant measure of this SDE is the Gibbs measure of density proportional to $e^{-2V(x)/\sigma^2}$ and for small enough σ, this measure concentrates around $\operatorname{argmin}(V)$, see [Dal17] and Chapter 4. Adding a small noise to the gradient descent allows to explore the space and to escape from traps such as local minima or saddle points appearing in non-convex optimization problems [Laz92, DPG$^+$14, HRSS21]. This noise may also be interpreted as coming from the approximation of the gradient in stochastic gradient descent algorithms. Such methods have been recently brought up to light again with Stochastic Gradient Langevin Dynamics (SGLD) algorithms [WT11, LCCC16], especially for the deep learning and the calibration of large artificial neural networks, which is a high-dimensional non-convex optimization problem.

The Langevin-simulated annealing SDE is the Langevin equation where the noise parameter is slowly decreasing to 0, namely

$$dX_t = -\nabla V(X_t)dt + a(t)\sigma dW_t,\ t \geq 0, \tag{2.1.1}$$

where $a : \mathbb{R}^+ \to \mathbb{R}^+$ is non-increasing and converges to 0. The idea is that the "instantaneous" invariant measure $\nu_{a(t)\sigma}$ which is the Gibbs measure of density $\propto \exp(-2V(x)/(a(t)^2\sigma^2))$ converges itself to $\operatorname{argmin}(V)$. This method indeed shares similarities with the original simulated annealing algorithm [vLA87], which builds a Markov chain from the Gibbs measure using the Metropolis-Hastings algorithm and where the parameter σ, interpreted as a temperature, slowly decreases to zero over the iterations.

In [CHS87, Roy89] is shown that choosing $a(t) = A\log^{-1/2}(t)$ for some $A > 0$ in (2.1.1) guarantees the convergence of X_t to $\nu^\star$ defined as the limit measure of $(\nu_{a(t)})$ as $t \to \infty$ and which is supported by $\operatorname{argmin}(V)$. [Mic92] proves again the convergence of the SDE using free energies inequalities. These studies deeply rely on some Poincaré and log-Sobolev inequalities and require the following assumptions on the potential function:

$$\lim_{|x|\to\infty} V(x) = \lim_{|x|\to\infty} |\nabla V(x)| = \infty \quad \text{and} \quad \forall x \in \mathbb{R}^d,\ \Delta V(x) \leq C + |\nabla V(x)|^2.$$

[Zit08] proves that the convergence still holds under weaker assumptions, in particular where the gradient of the potential is not coercive, using weak Poincaré inequalities. In [GM91] is proved the convergence of the associated stochastic gradient descent algorithm.

All these results are established in the so-called additive case, i.e. they highly rely on the fact that σ is constant, whereas little attention has been paid to the multiplicative case, i.e. where $\sigma : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ is not constant and depends on X_t. Allowing σ to be adaptive and to depend on the position highly extends the range of applications of Langevin algorithms and such adaptive algorithms are already widely used by practitioners and prove to be faster than non-adaptive algorithms and competitive with standard non-Langevin algorithms or even faster. See Section 2.3.2 where various specifications for $\sigma(x)$ that can be found in the Stochastic Optimization literature are briefly presented, and Section 2.9 where we show results of simulations of the training of an artificial neural network for various choices of σ. However, to our knowledge,

a general result of convergence for Langevin algorithms with multiplicative noise is yet to be proved. [PP23, Proposition 2.6] gives a general formula on b and σ so that the associated Gibbs measure is still the invariant measure of the SDE $dX_t = b(X_t)dt + \sigma(X_t)dW_t$; a simple example of acceleration of convergence using non-constant σ is then given in [PP23, Section 2.4]. More generally, [MCF15] gives a characterization of any SGMCMC (Stochastic Gradient Markov Chain Monte Carlo) algorithm with multiplicative noise and with the corresponding Gibbs measure as a target. In practice, the matrix σ is often chosen so that $\sigma\sigma^\top \simeq (\nabla^2 V)^{-1}$ but approximations are needed because of the high dimensions of the matrix (e.g. only considering diagonal matrices). Still, our results hold also for non-diagonal σ, which opens the way to algorithms with such σ.

In this paper, we consider the following SDE:

$$dY_t = -(\sigma\sigma^\top \nabla V)(Y_t)dt + a(t)\sigma(Y_t)dW_t + \left(a^2(t) \left[\sum_{j=1}^{d} \partial_j(\sigma\sigma^\top)(Y_t)_{ij} \right]_{1\leq i\leq d} \right) dt \tag{2.1.2}$$

$$a(t) = \frac{A}{\sqrt{\log(t)}}, \tag{2.1.3}$$

where the expression of the drift comes from [PP23, Proposition 2.6] and where the second drift term is interpreted as a correction term so that $\nu_{a(t)}$ is still the "instantaneous" invariant measure and satisfying $-b_{a(t)}\nu + (1/2)\nabla \cdot (a(t)\sigma\nu) = 0$ for every t, where b_a denotes the associated drift. This last term boils down to 0 if σ is constant. The aim of this paper is to prove the convergence for the L^1-Wasserstein distance of the law of Y_t to $\nu^\star$ in the setting adopted in [PP23], assuming in particular the convex uniformity of the potential outside a compact set and the ellipticity and the boundedness of σ. We also prove the convergence of the corresponding Euler-Maruyama scheme with decreasing steps and with noisy measurements of the gradient arising from mini-batch sampling. In this paper we use Markov models along with stochastic calculus to study the convergence of adaptive Langevin algorithms as a first theoretical study, however in practice these algorithms often include history/momentum effects with time-exponential average, and do not solely depend on Y_t as assumed in (2.1.2). We refer to Section 5.2 for precise definitions of some popular gradient algorithms.

Considering the convex condition outside a compact set is in fact quite different from the convex setting and turns out to be more demanding. This setting often appears in optimization problems (see Section 2.3.1), where a characteristic set - the compact set - contains the interesting features of the model with traps such as local minima, and where outside of this set the loss function is coercive and convex. We give classic examples of neural networks where this setting applies.

We adopt a *domino strategy* like in [PP23], inspired by proofs of weak error expansion of discretization schemes of diffusion processes [TT90, BT96]. In [PP23] is proved the convergence of the Euler-Maruyama scheme $\bar{X}$ with decreasing steps (γ_n) of an ergodic and homogeneous SDE X with non constant σ, to the invariant measure of X; the additive case was tackled in [DM17, BDMS19, DM19] and in [MMS20] for the constant step case. It then appears that the multiplicative case is more demanding than the additive case. For a function $f : \mathbb{R}^d \to \mathbb{R}$, the *domino strategy* consists in a step-by-step decomposition of the weak error to produce an upper bound as follows:

$$\begin{aligned} |\mathbb{E}f(\bar{X}^x_{\Gamma_n}) - \mathbb{E}f(X^x_{\Gamma_n})| &= |\bar{P}_{\gamma_1} \circ \cdots \circ \bar{P}_{\gamma_n} f(x) - P_{\Gamma_n} f(x)| \\ &\leq \sum_{k=1}^{n} \left| \bar{P}_{\gamma_1} \circ \cdots \circ \bar{P}_{\gamma_{k-1}} \circ (\bar{P}_{\gamma_k} - P_{\gamma_k}) \circ P_{\Gamma_n - \Gamma_k} f(x) \right|, \end{aligned} \tag{2.1.4}$$

where P and $\bar{P}$ are the transition kernels associated to X and $\bar{X}$ respectively and where $\Gamma_n = \gamma_1 + \cdots + \gamma_n$. Then two terms appear: first the "error" term, for large k, where the error is controlled by classic weak and strong bounds on the error of an Euler-Maruyama scheme, and the "ergodic" term, for small k, where the ergodicity of X is used.

However, we cannot directly apply this strategy of proof to our problem since we consider a non homogeneous SDE Y, so we proceed as follows: we consider instead the SDE X where the coefficient $a(t)$ is non-increasing and piecewise constant and where the successive plateaux $[T_{n-1}, T_n)$ of a are increasingly larger time intervals. On each plateau we obtain a homogeneous and uniformly elliptic SDE with an invariant Gibbs distribution ν_{a_n} where a_n is the constant value of a on $[T_{n-1}, T_n)$, to which a *domino strategy* can be applied. This ellipticity fades with time since a_n goes to 0 and we need to carefully control its impact on the way the diffusion X gets close to its "instantaneous" invariant Gibbs distribution ν_{a_n}. To this end we have to refine several one step weak error results from [PP23] and ergodic bounds from [Wan20], which extends the ergodic contraction bounds from [Ebe16] to the multiplicative case. Doing so we derive by induction an upper-bound for the distance between X_t and $\nu^\star$ after each plateau and prove that a coefficient $(a(t))$ of order $\log^{-1/2}(t)$ is a sufficient and generally necessary condition for convergence. Using this result, we then prove the convergence of Y_t and its Euler-Maruyama scheme $\bar{Y}_t$ by bounding the distance between X_t and Y_t and X_t and $\bar{Y}_t$. We need to split the problem into the cases where $(a(t))$ is piecewise constant and where $(a(t))$ is continuously decaying and then use careful weak error comparisons to prove the convergence in the second case, as we cannot use ergodic properties for the non-homogeneous processes at first sight. We also consider the "Stochastic Gradient case" i.e. where the true gradient cannot be computed exactly and where a noise, which is a sequence of increments of a martingale, is added to the gradient. This case was treated in [GM91] in the additive setting. The process X is used as a tool for the proof of the convergence of Y_t, however the convergence of X_t to $\nu^\star$ also has its own interest since the "plateau" method is also used by practitioners.

We also establish a convergence rate which is somehow limited by $\mathcal{W}_1(\nu_{a(t)}, \nu^\star)$, which is of order $a(t)$ under the assumption that $\operatorname{argmin}(V)$ is finite and that $\nabla^2 V$ is positive definite at every element of $\operatorname{argmin}(V)$, thus giving the same convergence rate as in the additive case and with the same rate for the annealing schedule $(a(t))$, see Remarks 2.4.7 and 2.5.2. If $\operatorname{argmin}(V)$ is still finite but if $\nabla^2 V$ is not positive definite at every element of $\operatorname{argmin}(V)$, but if we assume instead that all the elements of $\operatorname{argmin}(V)$ are strictly polynomial minima, then the rate is of order $a(t)^\delta$ for some $\delta \in (0, 1)$, see Chapter 4. We pay particular attention to the non-definite case, since it was pointed out in [SBL16, SEU+17] that for some optimization problems arising in Machine Learning, the Hessian of the loss function at the end of the training tends to be extremely singular. Indeed, as the dimension of the parameter which is used to minimize the loss function is large and as the neural network can be over-parametrized, many eigenvalues of the Hessian matrix are close to zero. However, this subject is still new in the Stochastic Optimization literature and needs more theoretical background.

Still we give sharper bounds on the rate of convergence of the L^1-Wasserstein distance between X or Y and $\nu_{a(t)}$ as in practice the optimization procedure stops at some (large) t and the target distribution is actually $\nu_{a(t)}$ instead of $\nu^\star$.

In a next paper, we shall prove the convergence in total variation distance. In this last case, the domino strategy is more complex to implement and requires regularization lemmas, as in [PP23] which studies the convergence for both distances.

The article is organized as follows. In Section 2.2 we first give the setting and assumptions of the problem we consider. This setting is taken from [PP23]. We then state our main results of convergence as well as convergence rates. In Section 2.3 we show how this setting applies to some classic optimization problems arising in Machine Learning and present several general choices

for σ that are used in practice. In Section 2.4 we consider the case where the coefficient a is constant and give convergence rates to the invariant measure taking into account the ellipticity parameter. We also give preliminary lemmas for the rate of convergence of ν_{a_n} to $\nu^\star$. In Section 2.5 we prove the convergence of the solution of the SDE where a is piecewise constant, by "plateaux". Using the dependence in a of the rate of convergence to the invariant measure in the ergodic case we prove the convergence to $\operatorname{argmin}(V)$. In Section 2.6 we prove the convergence of the SDE in the case where a is not by plateau but is continuously decreasing. This is done by bounding the Wasserstein distance with the "plateau" case and revisiting the lemmas for strong and weak errors from [PP23]. In Section 2.7 and Section 2.8 we also prove the convergence for the corresponding Euler-Maruyama schemes. The proofs actually follow the same strategy as the previous one. In Section 2.9 we present experiments of training of neural networks using various specifications for σ; the algorithms with multiplicative σ prove to be faster than the algorithm with constant σ.

Notations

In addition to the notations given from page 1, we use the notation C to denote a positive real constant, which may change from line to line. The constant C depends on the parameters of the problem: the coefficients of the SDE, the choice of A in $a(t) = A\log^{-1/2}(t)$, the upper bound $\bar{\gamma}$ on the decreasing steps, but C does not depend on t nor x.

2.2 Assumptions and main results

2.2.1 Assumptions

Let $V : \mathbb{R}^d \to (0, +\infty)$ be a $\mathcal{C}^2$ potential function such that V is coercive and

$$(x \mapsto |x|^2 e^{-2V(x)/A^2}) \in L^1(\mathbb{R}^d) \text{ for some } A > 0. \tag{2.2.1}$$

Then V admits a minimum on $\mathbb{R}^d$. Moreover, let us assume that

$$V^\star := \min_{\mathbb{R}^d} V > 0, \quad \operatorname{argmin}(V) = \{x_1^\star, \dots, x_{m^\star}^\star\}, \quad \forall\, i = 1, \dots, m^\star,\ \nabla^2 V(x_i^\star) > 0, \tag{2.2.2, $\mathcal{H}_{V1}$}$$

i.e. $\min_{\mathbb{R}^d} V$ is attained at a finite number $m^\star$ of points and in each point the Hessian matrix is positive definite. We then define for $a \in (0, A]$ the Gibbs measure ν_a of density :

$$\nu_a(dx) = \mathcal{Z}_a e^{-2(V(x)-V^\star)/a^2} dx, \quad \mathcal{Z}_a = \left(\int_{\mathbb{R}^d} e^{-2(V(x)-V^\star)/a^2} dx \right)^{-1} \tag{2.2.3}$$

We use the previous normalization for ν_a and $\mathcal{Z}_a$ however we emphasize that computing the density $\nu_a(x)$ does not require the knowledge of $V^\star$ since we could write $\nu_a(dx) = \bar{\mathcal{Z}}_a \exp(-2V(x)/a^2)dx$ with the appropriate normalization constant $\bar{\mathcal{Z}}_a$. Following [Hwa80, Theorem 2.1], the measure ν_a converges weakly to $\nu^\star$ as $a \to 0$, where $\nu^\star$ is the weighted sum of Dirac measures:

$$\nu^\star = \left(\sum_{j=1}^{m^\star} \left(\det \nabla^2 V(x_j^\star) \right)^{-1/2} \right)^{-1} \sum_{i=1}^{m^\star} \left(\det \nabla^2 V(x_i^\star) \right)^{-1/2} \delta_{x_i^\star}. \tag{2.2.4}$$

We consider the following Langevin SDE in $\mathbb{R}^d$:

$$Y_0^{x_0} = x_0 \in \mathbb{R}^d, \tag{2.2.5}$$

$$dY_t^{x_0} = b_{a(t)}(Y_t^{x_0})dt + a(t)\sigma(Y_t^{x_0})dW_t,$$

where, for $a \geq 0$, the drift b_a is given by

$$b_a(x) = -(\sigma\sigma^\top \nabla V)(x) + a^2 \left[\sum_{j=1}^{d} \partial_j(\sigma\sigma^\top)_{ij}(x)\right]_{1\leq i\leq d} =: -(\sigma\sigma^\top \nabla V)(x) + a^2\Upsilon(x), \qquad (2.2.6)$$

where W is a standard $\mathbb{R}^d$-valued Brownian motion defined on some rich enough probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where $\sigma : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ is C^2 and

$$a(t) = \frac{A}{\sqrt{\log(t+e)}} \qquad (2.2.7)$$

where A is defined in (2.2.1) and with $\log(e) = 1$. This equation corresponds to a gradient descent on the potential V with preconditioning σ and multiplicative noise; the second term in the drift (2.2.6) is a correction term (see [PP23, Proposition 2.6]) which is zero for constant σ.

We make the following assumptions on the potential V:

$$\lim_{|x|\to+\infty} V(x) = +\infty, \quad |\nabla V|^2 \leq CV \quad \text{and} \quad \sup_{x\in\mathbb{R}^d} ||\nabla^2 V(x)|| < +\infty, \qquad (2.2.8,\ \mathcal{H}_{V2})$$

which implies in particular that V has at most a quadratic growth. Let us also assume that

$$\sigma \text{ is bounded and Lipschitz continuous, } \nabla^2\sigma \text{ is bounded, } \nabla(\sigma\sigma^\top)\nabla V \text{ is bounded, } \quad (2.2.9,\ \mathcal{H}_\sigma)$$

and that σ is uniformly elliptic, i.e.

$$\exists \underline{\sigma}_0 > 0,\ \forall x \in \mathbb{R}^d,\ (\sigma\sigma^\top)(x) \geq \underline{\sigma}_0^2 I_d. \qquad (2.2.10)$$

Assumptions (2.2.8, $\mathcal{H}_{V2}$) and (2.2.9, $\mathcal{H}_\sigma$) imply that Υ is also bounded and Lipschitz continuous and that b_a is Lipschitz continuous uniformly in $a \in [0, A]$. Let the minimal constant $[b]_{\text{Lip}}$ be such that:

$$\forall a \in [0, A],\ b_a \text{ is } [b]_{\text{Lip}}\text{-Lipschitz continuous.}$$

We make the non-uniform dissipative (or convexity) assumption outside of a compact set: there exists $\alpha_0 > 0$ and $R_0 > 0$ such that

$$\forall x, y \in \mathbf{B}(0, R_0)^c,\ \left\langle \left(\sigma\sigma^\top \nabla V\right)(x) - \left(\sigma\sigma^\top \nabla V\right)(y),\ x - y \right\rangle \geq \alpha_0 |x-y|^2. \qquad (2.2.11,\ \mathcal{H}_{cf})$$

Taking $y \in \mathbf{B}(0, R_0)^c$ fixed, letting $|x| \to \infty$ and using the boundedness of σ, (2.2.11, $\mathcal{H}_{cf}$) implies that $|\nabla V|$ is coercive. Using (2.2.8, $\mathcal{H}_{V2}$) and the boundedness of σ, there exists $C > 0$ (depending on A) such that:

$$\forall a \in [0, A],\ 1 + |b_a(x)| \leq CV^{1/2}(x).$$

Let $(\gamma_n)_{n\geq 1}$ be a non-increasing sequence of varying positive steps. We define $\Gamma_n := \gamma_1 + \cdots + \gamma_n$ and for $t \geq 0$:

$$N(t) := \min\{k \geq 0 :\ \Gamma_{k+1} > t\} = \max\{k \geq 0 :\ \Gamma_k \leq t\}.$$

We make the classical assumptions on the step sequence, namely

$$\gamma_n \downarrow 0, \quad \sum_{n\geq 1} \gamma_n = +\infty \quad \text{and} \quad \sum_{n\geq 1} \gamma_n^2 < +\infty \qquad (2.2.12,\ \mathcal{H}_{\gamma 1})$$

and we also assume that

$$\varpi := \limsup_{n\to\infty} \frac{\gamma_n - \gamma_{n+1}}{\gamma_{n+1}^2} < \infty. \qquad (2.2.13,\ \mathcal{H}_{\gamma 2})$$

For example, if $\gamma_n = \gamma_1/n^\alpha$ with $\alpha \in (1/2, 1)$ then $\varpi = 0$; if $\gamma_n = \gamma_1/n$ then $\varpi = \gamma_1$.

In Stochastic Gradient algorithms, the true gradient is estimated using mini-batches of randomly selected training data, which introduces noise in the gradient estimates. This noise is modelled by a zero-mean random vector ζ, which law only depends on the current position. That is, let us consider a family of random fields $(\zeta_n(x))_{x\in\mathbb{R}^d, n\in\mathbb{N}}$ such that for every $n \in \mathbb{N}$, $(\omega, x) \in \Omega \times \mathbb{R}^d \mapsto \zeta_n(x, \omega)$ is measurable and for all $x \in \mathbb{R}^d$, the law of $\zeta_n(x)$ only depends on x and $(\zeta_n(x))_{n\in\mathbb{N}}$ is an i.i.d. sequence independent of W. We make the following assumptions:

$$\forall x \in \mathbb{R}^d,\ \forall p \geq 1,\ \mathbb{E}[\zeta_1(x)] = 0 \quad \text{and} \quad \mathbb{E}[|\zeta_1(x)|^p] \leq C_p V^{p/2}(x). \qquad (2.2.14)$$

We then consider the Euler-Maruyama scheme with decreasing steps associated to (Y_t):

$$\begin{aligned} \bar{Y}_0^{x_0} = x_0, \quad \bar{Y}_{\Gamma_{n+1}}^{x_0} &= \bar{Y}_{\Gamma_n} + \gamma_{n+1}\left(b_{a(\Gamma_n)}(\bar{Y}_{\Gamma_n}^{x_0}) + \zeta_{n+1}(\bar{Y}_{\Gamma_n}^{x_0})\right) \\ &\quad + a(\Gamma_n)\sigma(\bar{Y}_{\Gamma_n}^{x_0})(W_{\Gamma_{n+1}} - W_{\Gamma_n}), \end{aligned} \qquad (2.2.15)$$

We extend $\bar{Y}_\cdot^{x_0}$ on $\mathbb{R}^+$ by considering its genuine continuous interpolation:

$$\begin{aligned} \forall t \in [\Gamma_n, \Gamma_{n+1}), \quad \bar{Y}_t^{x_0} &= \bar{Y}_{\Gamma_n}^{x_0} + (t - \Gamma_n)\left(b_{a(\Gamma_n)}(\bar{Y}_{\Gamma_n}^{x_0}) + \zeta_{n+1}(\bar{Y}_{\Gamma_n}^{x_0})\right) \\ &\quad + a(\Gamma_n)\sigma(\bar{Y}_{\Gamma_n}^{x_0})(W_t - W_{\Gamma_n}). \end{aligned} \qquad (2.2.16)$$

Assumption (2.2.11, $\mathcal{H}_{cf}$) along with the fact that b_a is Lipschitz-continuous guarantee the no-explosion of the above processes and numerical schemes [Tal90, LP02], see also Lemmas 2.6.1 and 2.7.1.

2.2.2 Main results

We now state our main results.

Theorem 2.2.1. *(a) Let Y be defined in* (2.2.5)*. Assume* (2.2.2, $\mathcal{H}_{V1}$)*,* (2.2.8, $\mathcal{H}_{V2}$)*,* (2.2.9, $\mathcal{H}_\sigma$)*,* (2.2.10) *and* (2.2.11, $\mathcal{H}_{cf}$)*. If A is large enough, then for every $x_0 \in \mathbb{R}^d$,*

$$\mathcal{W}_1([Y_t^{x_0}], \nu^\star) \underset{t\to\infty}{\longrightarrow} 0.$$

More precisely, for every $t > 0$:

$$\mathcal{W}_1([Y_t^{x_0}], \nu^\star) \leq C \max(1 + |x_0|, V(x_0))a(t)$$

and for every $\alpha \in (0, 1)$ we have

$$\mathcal{W}_1([Y_t^{x_0}], \nu_{a(t)}) \leq C \max(1 + |x_0|, V(x_0))t^{-\alpha}.$$

(b) Let $\bar{Y}$ be defined in (2.2.15)*. Assume* (2.2.2, $\mathcal{H}_{V1}$)*,* (2.2.8, $\mathcal{H}_{V2}$)*,* (2.2.9, $\mathcal{H}_\sigma$)*,* (2.2.10) *and* (2.2.11, $\mathcal{H}_{cf}$)*. Assume furthermore* (2.2.12, $\mathcal{H}_{\gamma 1}$) *and* (2.2.13, $\mathcal{H}_{\gamma 2}$)*, that V is $\mathcal{C}^3$ with $\|\nabla^3 V\| \leq CV^{1/2}$ and that σ is $\mathcal{C}^3$ with $\|\nabla^3(\sigma\sigma^\top)\| \leq CV^{1/2}$. If A is large enough then for every $x_0 \in \mathbb{R}^d$,*

$$\mathcal{W}_1([\bar{Y}_t^{x_0}], \nu^\star) \underset{t\to\infty}{\longrightarrow} 0.$$

More precisely, for every $t > 0$:

$$\mathcal{W}_1([\bar{Y}_t^{x_0}], \nu^\star) \leq C \max(1 + |x_0|, V^2(x_0))a(t),$$

and for every $\alpha \in (0,1)$ we have

$$\mathcal{W}_1([\bar{Y}_t^{x_0}], \nu_{a(t)}) \leq C \max(1 + |x_0|, V^2(x_0))t^{-\alpha}.$$

Remark 2.2.2. In particular, if $\operatorname{argmin} V = \{x^\star\}$ is reduced to a point, we can rewrite the conclusions of Theorem 2.2.1 as $\|Y_t^{x_0} - x^\star\|_1 \to 0$ and $\|\bar{Y}_t^{x_0} - x^\star\|_1 \to 0$ respectively and so on.

2.2.3 The degenerate case

In this subsection we consider the case where some of the $\nabla^2 V(x_i^\star)$'s may be not definite positive but where the $x_i^\star$'s are strictly polynomial minima, i.e. is $V(x) - V(x_i^\star)$ is bounded below in a neighbourhood of $x_i^\star$ by a non-negative polynomial function null only in $x_i^\star$. This case can be treated in a similar way using the change of variable given in Chapter 4.

First, let us restate the results from Theorem 4.3.3. To simplify, let us assume that $\operatorname{argmin}(V)$ is reduced to a point.

Theorem 2.2.3. *Assume that V is C^{2p} with $p \geq 2$, is coercive, that* $\operatorname{argmin}(V) = \{x^\star\}$*, that $e^{-AV} \in L^1(\mathbb{R}^d)$ for some $A > 0$ and that $x^\star$ is a strictly polynomial minimum of order $2p$ i.e. p is the smallest integer such that*

$$\exists r > 0,\ \forall h \in \boldsymbol{B}(0,r) \setminus \{0\},\ \sum_{k=2}^{2p} \frac{1}{k!} \nabla^k V(x^\star) \cdot h^{\otimes k} > 0.$$

Assume also the technical hypothesis (4.3.4) *if $p \geq 5$. Then there exist $B \in \mathcal{O}_d(\mathbb{R})$, $\alpha_1, \ldots, \alpha_d \in \{1/2, \ldots, 1/(2p)\}$ and a polynomial function $g : \mathbb{R}^d \to \mathbb{R}$ which is not constant in any of its variables such that*

$$\forall h \in \mathbb{R}^d,\ \frac{1}{s}[V(x^\star + B \cdot (s^{\alpha_1} h_1, \ldots, s^{\alpha_d} h_d)) - V(x^\star)] \underset{s \to 0}{\longrightarrow} g(h).$$

Moreover assume that g is coercive. Then if $Z_s \sim \nu_{\sqrt{2s}}$,

$$\left(\frac{(B^{-1} \cdot (Z_s - x^\star))_1}{s^{\alpha_1}}, \cdots, \frac{(B^{-1} \cdot (Z_s - x^\star))_d}{s^{\alpha_d}} \right) \xrightarrow{\mathscr{L}} Z \quad \text{as } s \to 0,$$

where Z has density proportional to $\exp(-g)$.

Theorem 2.2.4. *Let us make the same assumptions as in Theorem 2.2.3 and assume that $V^\star > 0$. Assume furthermore* (2.2.8, $\mathcal{H}_{V2}$)*,* (2.2.9, $\mathcal{H}_\sigma$)*,* (2.2.10) *and* (2.2.11, $\mathcal{H}_{cf}$)*. Assume furthermore* (2.2.12, $\mathcal{H}_{\gamma 1}$) *and* (2.2.13, $\mathcal{H}_{\gamma 2}$)*, that V is $\mathcal{C}^3$ with $\|\nabla^3 V\| \leq C V^{1/2}$ and that σ is $\mathcal{C}^3$ with $\|\nabla^3(\sigma\sigma^\top)\| \leq C V^{1/2}$. Let us denote $\alpha_{\min} := \min(\alpha_1, \ldots, \alpha_d)$. Then for every $\alpha \in (0,1)$ we have*

$$\begin{aligned}
&\mathcal{W}_1([Y_t^{x_0}], \nu_{a(t)}) \leq C \max(1 + |x_0|, V(x_0))t^{-\alpha},\\
&\mathcal{W}_1([\bar{Y}_t^{x_0}], \nu_{a(t)}) \leq C \max(1 + |x_0|, V^2(x_0))t^{-\alpha},\\
&\mathcal{W}_1([Y_t^{x_0}], \nu^\star) \leq C \max(1 + |x_0|, V(x_0))a(t)^{2\alpha_{\min}},\\
&\mathcal{W}_1([\bar{Y}_t^{x_0}], \nu^\star) \leq C \max(1 + |x_0|, V^2(x_0))a(t)^{2\alpha_{\min}}.
\end{aligned}$$

The proof is given in the Supplementary Material.

2.3 Application to optimization problems

2.3.1 Potential function associated to a Neural Regression Problem

The setting described in Section 2.2 can first be applied to convex optimization problems where the potential function V has a quadratic growth as $|x| \to \infty$. Classical examples are least-squares regression and logistic regression with quadratic regularization, that is:

$$\min_{x\in\mathbb{R}^d} \frac{1}{M}\sum_{i=1}^{M} \log(1+e^{-v_i\langle u_i,x\rangle}) + \frac{\lambda}{2}|x|^2,$$

where $v_i \in \{-1,+1\}$ and $u_i \in \mathbb{R}^d$ are the data samples associated with a binary classification problem and where $\lambda > 0$ is the regularization parameter.

We now consider a scalar regression problem with a fully connected neural network with quadratic regularization. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be the sigmoid function. To simplify the proofs, we may consider instead a smooth function approximating the sigmoid function such that φ' has compact support. Let $K \in \mathbb{N}$ be the number of layers and for $k = 1, \ldots, K$, let $d_k \in \mathbb{N}$ be the size of the k^{th} layer with $d_K = 1$. For $u \in \mathbb{R}^{d_{k-1}}$ and for $\theta \in \mathcal{M}_{d_k,d_{k-1}}(\mathbb{R})$, we define $\varphi_\theta(u) := [\varphi([\theta \cdot u]_i)]_{1\le i\le d_k}$. The output of the neural network is

$$\begin{aligned}&\psi : \mathbb{R}^{d_1,d_0} \times \cdots \times \mathbb{R}^{d_K,d_{K-1}} \times \mathbb{R}^{d_0} \to \mathbb{R}\\ &\psi(\theta_1,\ldots,\theta_K,u) = \psi(\theta,u) = \varphi_{\theta_K} \circ \ldots \circ \varphi_{\theta_1}(u).\end{aligned}$$

Let $u_i \in \mathbb{R}^{d_0}$ and $v_i \in \mathbb{R}$ be the data samples for $1 \le i \le M$. The objective is

$$\underset{\theta_1,\ldots,\theta_K}{\text{minimize}} \quad V(\theta) := \frac{1}{2M}\sum_{i=1}^{M}(\psi(\theta_1,\ldots,\theta_K,u_i) - v_i)^2 + \frac{\lambda}{2}|\theta|^2,$$

where $\theta = (\theta_1,\ldots,\theta_K)$ and where $\lambda > 0$.

Proposition 2.3.1. *Consider a neural network with a single layer : $\psi(\theta,u) = \varphi(\langle\theta,u\rangle)$. Assume that the data u and v are bounded, that u admits a continuous density and φ' has bounded support in $\mathbb{R}$. Then V satisfies* (2.2.8, $\mathcal{H}_{V2}$) *and for some $R_0,\ \alpha_0 > 0$,*

$$\forall x,\ y \in \boldsymbol{B}(0,R_0)^c,\ \langle\nabla V(x) - \nabla V(y), x-y\rangle \ge \alpha_0|x-y|^2 \tag{2.3.1}$$

Proof. Let us define the measure Q on $\mathbb{R}^{d_0}\times\mathbb{R}$ associated to the data, i.e. $Q = (1/M)\sum_{i=1}^{M}\delta_{u_i,v_i}$. Note that φ, φ' and φ'' are bounded. The function ψ is bounded so

$$2V(\theta) = \int(\varphi(\langle\theta,u\rangle) - v)^2 Q(du,dv) + \lambda|\theta|^2 \sim \lambda|\theta|^2 \quad \text{as } |\theta| \to \infty,$$

so V is coercive. Moreover, we have

$$\nabla V = \int u\varphi'(\langle\theta,u\rangle)(\varphi(\langle\theta,u\rangle) - v)Q(du,dv) + \lambda\theta$$

so $\nabla V(\theta) \sim \lambda\theta$ as $|\theta| \to \infty$ and $|\nabla V|^2 \le CV$. Then, let us assume that the support of φ' is included in $[-R,R]$, that u has its values in $\mathbf{B}(0,R)$ and v in $[-R,R]$. Then the set $\{u \in \mathbf{B}(0,R),\ |\langle\theta,u\rangle| < R\}$ has Lebesgue measure no larger than $C/|\theta|$ so

$$\left\|\nabla^2\int(\varphi(\langle\theta,u\rangle) - v)^2 Q(du,dv)\right\| \le C/|\theta|,$$

so outside the compact set $\{|\theta| \le 2C/\lambda\}$, we have $\|\nabla^2 V\| \ge \lambda/2$ which guarantees (2.3.1). □

This proposition can be extended to the case where φ is of sigmoid type, with fast decay of φ' as $\pm\infty$. However, we cannot directly extend this proposition to multi-layers neural networks. Nevertheless, if we consider that the training stops if a parameter becomes too large and if we replace $\psi(\theta, u)$ by $\psi(\phi(\theta), u)$ where $\phi : \mathbb{R} \to \mathbb{R}$ is a smooth approximation of $x \mapsto \min(x, R)\mathbb{1}_{x\geq 0} + \max(x, -R)\mathbb{1}_{x<0}$ where $R > 0$ is large and where ϕ is applied in order to avoid over-fitting coordinate by coordinate then the resulting potential V with quadratic regularization satisfies (2.2.8, $\mathcal{H}_{V2}$) and (2.3.1).

2.3.2 Practitioner's corner: choices for σ

In this section we briefly present general choices for the non-constant matrix σ that are often used in the Stochastic Optimization and Machine Learning literature.

[WT11] introduced the Stochastic Gradient Langevin Dynamics (SGLD) with constant preconditioner matrix σ. [DdVB15] adapted the well-known Newton method, which consists in considering $\sigma\sigma^\top = (\nabla^2 V)^{-1}$, to SGLD. Since the size of the Hessian matrix may be too large in practice, because inverting it is computationally costly and because the Hessian matrix may not be positive in every point, it is suggested to consider instead $|\mathrm{diag}((\nabla^2 V))^2|^{-1/2}$. However, computing high-order derivatives may be cumbersome; [SBCR16] adapts the quasi-Newton method [NW06] to approximate the Hessian matrix to SGLD, yielding the Stochastic Quasi-Newton Langevin algorithm.

[DHS11] and [LCCC16] give algorithms where the choice for σ is $\sigma \simeq \mathrm{diag}((\lambda + |\nabla V|)^{-1})$, where $\lambda > 0$ guarantees numerical stability. The idea of using geometry has been explored in [PT13], where σ^{-2} defines the local curvature of a Riemannian manifold, giving the Stochastic Gradient Riemaniann Langevin Dynamics algorithm where σ is equal to $\mathcal{I}_x^{-1/2}$ where $\mathcal{I}_x$ is the Fisher information matrix, or to some other choices (see [PT13, Table 1]) as $\mathcal{I}_x$ may be intractable. [MCF15] extends the previous algorithm to Hamiltonian Monte Carlo methods, where a momentum variable is added in order to take into account the "inertia" of the trajectory [Nea96, DMS20], yielding the Stochastic Gradient Riemannian Hamiltonian Monte Carlo method.

Allowing the matrix σ to depend on the position yields a faster convergence; we refer to the previous references where various simulations confirm that these new methods greatly improve classical stochastic gradients algorithms. In particular, we refer to the simulations [SBCR16, Figure 2], [PT13, Figure 2] and [MCF15, Figure 3] where the different methods based on multiplicative noise are compared.

In popular choices for σ like RMSprop [LCCC16] and Adam [KB15], the order of growth of σ is roughly

$$\sigma(x) \simeq \mathrm{diag}\left(\eta_1(\eta_2 + U(x))^{-1/4}\right),$$

with $\eta_1, \eta_2 > 0$ and $U(x) = \nabla V(x)^{*2}$ where $*$ denotes the component-wise product. Then σ is bounded and we have $\nabla U = 2\nabla V * \nabla^2 V$ and

$$\nabla\sigma = -(\eta_1/2)(\eta_2 + U)^{-5/4} * \nabla V * \nabla^2 V,$$

where the shapes of the tensors are broadcast with $*$ if the orders are different. We obtain then that $\nabla\sigma$ is bounded. Moreover

$$\nabla(\sigma\sigma)^\top \cdot \nabla V = -\eta_1((\eta_2 + U)^{-3/2} * \nabla V * \nabla^2 V) \cdot \nabla V$$

is also bounded. Moreover

$$\nabla^2\sigma = ((5/4)\eta_1(\eta_2 + U)^{-9/4} * \nabla V * \nabla^2 V) \otimes_2 (\nabla V * \nabla^2 V)$$

$$- (1/2)\eta_1(\eta_2 + U)^{-5/4}(\nabla^2 V^{\otimes_2 2} + \nabla V * \nabla^3 V)$$

with the rule $a \otimes_2 b = (a_{ij}b_{ik})_{ijk}$, so that $\nabla^2\sigma$ is also bounded. Thus the assumptions on the growth of σ and on its derivatives (2.2.9, $\mathcal{H}_\sigma$) are satisfied.

In practice, the annealing schedule $(a(t))$ converges slowly to 0 and practitioners often use adaptive Langevin algorithms with a being constant or decreasing on a few plateau values instead. At the end of the optimization procedure, the coefficient $a(t)$ may thus not be zero, yielding a convergence to the Gibbs measure $\nu_{a(t)}$, which still gives a good approximation of $\operatorname{argmin}(V)$ if $a(t)$ is small enough and the theoretical bounds we give for the "plateau" case (see Theorems 2.5.1 and 2.8.1) can still be applied for fixed $t \in \mathbb{R}$. Moreover, adding small Langevin noise can improve the learning procedure in comparison with non-Langevin gradient descent algorithms, even if a is taken constant [LCCC16].

2.4 Langevin equation with constant time coefficient

In this section, we consider the following $\mathbb{R}^d$-valued homogeneous SDE:

$$X_0^x = x \in \mathbb{R}^d, \quad dX_t^x = b_a(X_t^x)dt + a\sigma(X_t^x)dW_t, \tag{2.4.1}$$

with $a \in (0, A]$ and where b_a is defined in (2.2.6). The drift is specified in such a way that the Gibbs measure ν_a defined in (2.2.3) is the unique invariant distribution of (X_t^x) (see [PP23, Proposition 2.6]).

2.4.1 Exponential contraction property

We now prove contraction properties of the SDE (2.4.1) under the uniform convex setting on the whole $\mathbb{R}^d$ or outside a compact set (2.2.11, $\mathcal{H}_{cf}$). If the uniform dissipative assumption holds on $\mathbb{R}^d$ then we have the following contraction property.

Proposition 2.4.1. *Let Z be the solution of*

$$Z_0^x = x \in \mathbb{R}^d, \quad dZ_t^x = b^Z(Z_t^x)dt + \sigma^Z(Z_t^x)dW_t,$$

where the coefficients b^Z and σ^Z are (globally) Lipschitz continuous. Assume the uniform convexity i.e. there exists $\alpha > 0$ such that

$$\forall x, y \in \mathbb{R}^d, \langle b^Z(x) - b^Z(y),\ x - y\rangle + \frac{1}{2}||\sigma^Z(x) - \sigma^Z(y)||^2 \leq -\alpha|x - y|^2. \tag{2.4.2}$$

Then:

$$\forall x, y \in \mathbb{R}^d,\ \mathcal{W}_1\left([Z_t^x], [Z_t^y]\right) \leq C|x - y|e^{-\alpha t}.$$

Proof. By the Itō lemma, $t \mapsto e^{2\alpha t}\left|Z_t^x - Z_t^y\right|^2$ is a super-martingale, so

$$\mathbb{E}\left|Z_t^x - Z_t^y\right|^2 \leq e^{-2\alpha t}|x - y|^2,$$

which yields the desired result. □

This proposition can be applied to X under the assumption

$$\forall x, y \in \mathbb{R}^d, \langle b_a(x) - b_a(y),\ x - y\rangle + \frac{a^2}{2}||\sigma(x) - \sigma(y)||^2 \leq -\alpha|x - y|^2,$$

which may be hard to check because of the dependence in a. In [PP23, Corollary 2.5] is proved that this contraction property is still true under uniform convexity outside a compact set (2.2.11, $\mathcal{H}_{cf}$) using bounds from [Wan20]. [Wan20, Theorem 2.6] relies on explicit couplings of $[X_t^x]$ and $[X_t^y]$ by reflection. Such reflection couplings were introduced in [Ebe16] for the additive case $\sigma = I_d$ and were extended to the multiplicative case in [Wan20]. In the following theorem we make explicit the dependence in a.

Theorem 2.4.2. *Under the assumption* (2.2.11, $\mathcal{H}_{cf}$),

(a) For every x, $y \in \mathbb{R}^d$,

$$\mathcal{W}_1\left([X_t^x], [X_t^y]\right) \leq Ce^{C_1/a^2}|x-y|e^{-\rho_a t}, \quad \rho_a := e^{-C_2/a^2} \tag{2.4.3, $\mathcal{P}_{cf}$}$$

where the constants C, C_1, C_2 do not depend on a.

(b) For every $x \in \mathbb{R}^d$,

$$\mathcal{W}_1\left([X_t^x], \nu_a\right) \leq Ce^{C_1/a^2}e^{-\rho_a t}\mathcal{W}_1(\delta_x, \nu_a). \tag{2.4.4}$$

Proof. (a) We refine the proof of [Wan20, Theorem 2.6] to enhance the dependence of the constants in the parameter a. First we remark as in [PP23, Section 4.5] in the proof of Corollary 2.5, that Assumption (2.17) of [Wan20], stating that there exist constants K_1, K_2 and $r_0 > 0$ such that, with $\underline{\sigma} := \sqrt{\sigma\sigma^\top - \underline{\sigma}_0^2 I_d}$,

$$\begin{aligned}\frac{a^2}{2}\|\underline{\sigma}(x) - \underline{\sigma}(y)\|^2 - \frac{|a^2(\sigma(x)-\sigma(y))^\top(x-y)|^2}{2|x-y|^2} + \langle b_a(x) - b_a(y), x-y\rangle \\ \leq \left((K_1+K_2)\mathbb{1}_{|x-y|\leq r_0} - K_2\right)|x-y|^2,\ x,\ y \in \mathbb{R}^d,\end{aligned} \tag{2.4.5}$$

is true, since $a \in (0, A]$ and $\sigma\sigma^\top$ bounded, as soon as there exist positive constants $\widetilde{K_1}$, $\widetilde{K_2}$ and R_1 such that

$$\forall x, y \in \mathbb{R}^d,\ \langle b_a(x) - b_a(y), x-y\rangle \leq \widetilde{K_1}\mathbb{1}_{|x-y|\leq R_1} - \widetilde{K_2}|x-y|^2,$$

which is, up to changing the positive constants, equivalent to

$$\forall x, y \in \mathbb{R}^d,\ \langle b_0(x) - b_0(y), x-y\rangle \leq \widetilde{K_1}\mathbb{1}_{|x-y|\leq R_1} - \widetilde{K_2}|x-y|^2,$$

which is in turn equivalent to (2.2.11, $\mathcal{H}_{cf}$). Then we repeat the argument leading to (4.3) in [Wan20]. We reformulate the assumption of ellipticity (2.2.10) as:

$$dX_t = b_a(X_t)dt + a(\underline{\sigma}(X_t)dW_t^1 + \underline{\sigma}_0 dW_t^2),$$

where $\underline{\sigma} \geq 0$ and where (W_t^1) and (W_t^2) are two independent Brownian motions in $\mathbb{R}^d$ (which can be expressed in terms of W). For $x \neq y$, let X^x be the solution of this SDE with $X_0 = x$ and let Y^y solve the following coupled SDE for $Y_0^y = y$:

$$dY_t^y = b_a(Y_t^y)dt + a\underline{\sigma}(Y_t^y)dW_t^1 + a\underline{\sigma}_0\left(dW_t^2 - 2\frac{\langle X_t^x - Y_t^y, dW_t^2\rangle(X_t^x - Y_t^y)}{|X_t^x - Y_t^y|^2}\right).$$

The process Y^y is in fact defined by orthogonally symmetrizing the component of the noise in W^2 w.r.t. $X_t^x - Y_t^y$ at every instant t. This SDE has a unique solution up to the coupling time

$$T_{x,y} := \inf\{t \geq 0 :\ X_t^x = Y_t^y\},$$

and for $t \geq T_{x,y}$ we set $Y_t^y = X_t^x$. Then Y^y has the same distribution as X^y i.e. is a weak solution of (2.4.1) with starting value y and it follows from (2.4.5) and from the Itō formula applied to $|X^x - Y^y|$ that for every $0 \leq u \leq t \leq T_{x,y}$,

$$|X_t^x - Y_t^y| - |X_u^x - Y_u^y| \leq M_t - M_u + \int_u^t \Big((K_1+K_2)\mathbb{1}_{|X_s^x - Y_s^y| \leq r_0} - K_2\Big) |X_s^x - Y_s^y| ds,$$

where

$$M_t = \int_0^t \frac{a\langle 2\underline{\sigma}_0 dW_s^2 + (\underline{\sigma}(X_s) - \underline{\sigma}(Y_s^y))dW_t^1, X_s^x - Y_s^y\rangle}{|X_s^x - Y_s^y|}$$

is a true Brownian martingale with bracket process satisfying

$$\langle M \rangle_t \geq 4a^2 \underline{\sigma}_0^2 t. \tag{2.4.6}$$

We now set, still like in the proof of (4.3) in [Wan20],

$$p_t := |X_t^x - Y_t^y| \quad \text{and} \quad \bar{p}_t := \varepsilon p_t + 1 - e^{-Np_t},$$

where

$$N := \frac{r_0}{a^2 \underline{\sigma}_0^2}(K_1 + K_2) \quad \text{and} \quad \varepsilon := Ne^{-Nr_0}.$$

Then we have :

$$\varepsilon p_t \leq \bar{p}_t \leq (N+\varepsilon)p_t, \quad \text{and} \quad \forall r \in [0, r_0),\ \frac{2N^2}{r(\varepsilon e^{Nr} + N)} \geq \frac{K_1 + K_2}{a^2 \underline{\sigma}_0^2}.$$

Then using (2.4.6) we derive for all $0 \leq u \leq t \leq T_{x,y}$:

$$\begin{aligned}
\bar{p}_t - \bar{p}_u &\leq \int_u^t (\varepsilon + Ne^{-Np_s}) dM_s \\
&\quad + \int_u^t (\varepsilon + Ne^{-Np_s}) \left((K_1+K_2)\mathbb{1}_{p_s \leq r_0} - K_2 - \frac{2N^2 a^2 \underline{\sigma}_0^2}{p_s(\varepsilon e^{Np_s} + N)} \right) p_s ds \\
&\leq \tilde{M}_t - \tilde{M}_u - K_2 \int_u^t (\varepsilon + Ne^{-Np_s}) p_s ds \leq \tilde{M}_t - \tilde{M}_u - \varepsilon K_2 \int_u^t p_s ds \\
&\leq \tilde{M}_t - \tilde{M}_u - \frac{\varepsilon K_2}{N+\varepsilon} \int_u^t \bar{p}_s ds.
\end{aligned}$$

So that we have

$$\mathbb{E}[\bar{p}_t - \bar{p}_u] = \mathbb{E}[(\bar{p}_t - \bar{p}_u)\mathbb{1}_{t \leq T_{x,y}}] \leq -\frac{\varepsilon K_2}{N+\varepsilon} \int_u^t \mathbb{E}\bar{p}_s ds,$$

so that

$$\frac{d}{dt}\mathbb{E}[\bar{p}_t] \leq -\frac{\varepsilon K_2}{N+\varepsilon}\mathbb{E}[\bar{p}_t]$$

and then

$$\mathbb{E}\bar{p}_t \leq \bar{p}_0 e^{-\frac{\varepsilon K_2}{N+\varepsilon}t}.$$

Noting that $\bar{p}_0 \leq (N+\varepsilon)|x-y|$, we have

$$\mathbb{E}p_t \leq \frac{N+\varepsilon}{\varepsilon}|x-y|e^{-\frac{\varepsilon K_2}{N+\varepsilon}t}, .$$

so that

$$\mathcal{W}_1\left([X_t^x], [X_t^y]\right) \leq \frac{N+\varepsilon}{\varepsilon}|x-y|e^{-\frac{\varepsilon K_2}{N+\varepsilon}t} \leq Ce^{C_1/a^2}|x-y|e^{-e^{-C_2/a^2}t}.$$

(b) As ν_a is the invariant distribution of the diffusion (2.4.1), using (2.4.3, $\mathcal{P}_{cf}$) we have

$$\begin{aligned}\mathcal{W}_1\left([X_t^x], \nu_a\right) &= \int_{\mathbb{R}^d} \mathcal{W}_1\left([X_t^x], [X_t^y]\right) \nu_a(dy) \leq C e^{C_1/a^2} e^{-\rho_a t} \int_{\mathbb{R}^d} |x-y| \nu_a(dy) \\ &\leq C e^{C_1/a^2} e^{-\rho_a t} \mathcal{W}_1(\delta_x, \nu_a).\end{aligned}$$

□

2.4.2 Time schedule and Wasserstein distance between Gibbs measures

For $C_{(T)} > 0$ and for $\beta > 0$, let us define the time schedule that will be used for the plateau SDE in the next section:

$$T_n := C_{(T)} n^{1+\beta}, \tag{2.4.7}$$

and by a slight abuse of notation we define

$$a_n := a(T_n) = \frac{A}{\sqrt{\log(T_n + e)}} \quad \text{and} \quad \rho_n := \rho_{a_n} = e^{-C_2/a_n^2}. \tag{2.4.8}$$

Lemma 2.4.3. *The sequence $a_n = A \log^{-1/2}(T_n + e)$ satisfies*

$$0 \leq a_n - a_{n+1} \asymp (n \log^{3/2}(n))^{-1}. \tag{2.4.9}$$

Proof. One straightforwardly checks that

$$\begin{aligned}a_n - a_{n+1} &\sim -\frac{d}{dn}\left(\frac{A}{\sqrt{\log(C_{(T)} n^{1+\beta} + e)}}\right) \\ &= \frac{A(1+\beta)}{2\log^{3/2}(C_{(T)} n^{1+\beta} + e)\left(n + e/(C_{(T)} n^{\beta})\right)} \asymp \frac{1}{n \log^{3/2}(n)}.\end{aligned}$$

□

We prove the following result that will be useful to study the convergence of the plateau SDE.

Proposition 2.4.4. *Let ν_a, $a \in (0, A]$ be the Gibbs measure defined in (2.2.3). Assume that V is coercive, that $(x \mapsto |x|^2 e^{-2V(x)/A^2}) \in L^1(\mathbb{R}^d)$ and (2.2.2, $\mathcal{H}_{V1}$). Then for $n \in \mathbb{N}$,*

$$\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) \leq \frac{C}{n \log^{3/2}(n)}.$$

Moreover, for every s, $t \in [a_{n+1}, a_n]$, we have

$$\mathcal{W}_1(\nu_s, \nu_t) \leq \frac{C}{n \log^{3/2}(n)}.$$

The proof of this proposition is given in the Supplementary Material. It relies on the following lemma.

Lemma 2.4.5. *Let μ and ν be two probability distributions on $\mathbb{R}^d$ with densities f and g respectively with finite moments of order p. Assume that there exists $M \geq 1$ such that $f \leq Mg$. Then*

$$\mathcal{W}_p(\mu, \nu)^p \leq \mathbb{E}|X - Y|^p - \frac{1}{M}\mathbb{E}|X - \tilde{X}|^p,$$

where X and $\tilde{X} \sim \mu$, $Y \sim \nu$ and X, $\tilde{X}$ and Y are mutually independent.

Proof. We define a coupling on μ and ν inspired from the acceptance rejection sampling as follows. Let $X \sim \mu$, $Y \sim \nu$, $U \sim \mathcal{U}([0,1])$ and X, Y, U are independent, and let

$$X' = Y\mathbb{1}\{U \leq f(Y)/(Mg(Y))\} + X\mathbb{1}\{U > f(Y)/(Mg(Y))\}.$$

Then adapting the proof of the acceptance rejection method, $X' \sim \mu$ and we have:

$$\begin{aligned}
\mathbb{E}|X' - Y|^p &= \mathbb{E}|Y - X|^p \mathbb{1}\{U > f(Y)/(Mg(Y))\} \\
&= \int_{(\mathbb{R}^d)^2} |y-x|^p \left(\int_0^1 \mathbb{1}\{u > f(y)/(Mg(y))\} du \right) f(x)g(y)dxdy \\
&= \int_{(\mathbb{R}^d)^2} |y-x|^p f(x)g(y)dxdy - \frac{1}{M} \int_{(\mathbb{R}^d)^2} |y-x|^p f(x)f(y)dxdy \\
&= \mathbb{E}|X-Y|^p - \frac{1}{M}\mathbb{E}|X - \tilde{X}|^p.
\end{aligned}$$

□

Lemma 2.4.6. *We have*

$$\mathcal{W}_1(\nu_{a_n}, \nu^\star) \leq Ca_n. \tag{2.4.10}$$

Proof. First let us prove that $\mathcal{W}_1(\nu_a, \nu^\star) \to 0$ as $a \to 0$. By Proposition 2.4.4 and using that

$$\textstyle\sum_{n\geq 2} (n \log^{3/2}(n))^{-1} < \infty,$$

(ν_{a_n}) is a Cauchy sequence in $(L^1(\mathbb{R}^d), \mathcal{W}_1)$ so converges to some limit measure $\widetilde{\nu}$. But (ν_{a_n}) also weakly converges to $\widetilde{\nu}$, so $\widetilde{\nu} = \nu^\star$. Moreover, $\mathcal{W}_1(\nu_{a_n}, \nu^\star)$ is bounded by the tail of the above series, which is of order $\log^{-1/2}(n)$. □

Remark 2.4.7. Considering $\nu_a(dx) = \mathcal{Z}_a \exp(-2|x|^2/a^2)dx$ with $V(x) = |x|^2$ gives

$$\mathcal{W}_1(\nu_a, \nu^\star) = \mathbb{E}|Z_a - 0| = (a/2)\mathbb{E}|\mathcal{N}(0, I_d)|$$

with $Z_a \sim \nu_a$, showing that we cannot get better convergence rates in Lemma 2.4.6 in general.

2.5 Plateau case

We define (X_t) as the solution the following SDE where the coefficients piecewisely depend on the time; X is then said to be "by plateaux":

$$X_0^{x_0} = x_0, \quad dX_t^{x_0} = b_{a_{k+1}}(X_t^{x_0})dt + a_{k+1}\sigma(X_t^{x_0})dW_t, \quad t \in [T_k, T_{k+1}], \tag{2.5.1}$$

where b_a is defined in (2.2.6), (T_n) is defined in (2.4.7) and (a_n) is defined in (2.4.8). We first prove the convergence of (X_t) using ergodic properties of the corresponding SDE on each plateau $[T_n, T_{n+1}]$. We note that although the coefficients are not continuous, the process $(X_t^{x_0})$ is well defined as it is the continuous concatenation of the solutions of the equations on the intervals $[T_k, T_{k+1}]$. More generally, we define $(X_t^{x,n})$ as the solution of

$$X_0^{x,n} = x, \ dX_t^{x,n} = b_{a_{k+1}}(X_t^{x,n})dt + a_{k+1}\sigma(X_t^{x,n})dW_t, \ t \in [T_k - T_n, T_{k+1} - T_n], \ k \geq n,$$

i.e. $(X_t^{x,n})$ has the law of $(X_{T_n+t})_{t\geq 0}$ conditionally to $X_{T_n} = x$. We have $X_t^x = X_t^{x,0}$.

Theorem 2.5.1. *Let X be defined in* (2.5.1)*. If*

$$A > \max\left(\sqrt{(1+\beta^{-1})C_2}, \sqrt{(1+\beta)C_1}\right), \tag{2.5.2}$$

where C_1 and C_2 are given in (2.4.3, $\mathcal{P}_{cf}$)*, then for every $x_0 \in \mathbb{R}^d$:*

$$\mathcal{W}_1([X_t^{x_0}], \nu^\star) \underset{t\to\infty}{\longrightarrow} 0.$$

More precisely, for $t \geq 0$ we have:

(i) $\mathcal{W}_1([X_t^{x_0}], \nu^\star) \leq Ca(t)(1+|x_0|)$,

(ii) for all $C' < C_{(T)}$, for all large enough $n \geq n(C'_{(T)})$,

$$\mathcal{W}_1([X_{T_n}^{x_0}], \nu_{a_n}) \leq Cn^{-1+(\beta+1)C_1/A^2} e^{-(C')^{1-C_2/A^2}(\beta+1)n^{\beta-(\beta+1)C_2/A^2}}(1+|x_0|),$$

(iii) $\mathcal{W}_1([X_t^{x_0}], \nu_{a(t)}) \leq \dfrac{C(1+|x_0|)}{t^{(\beta+1)^{-1}-C_1/A^2}\log^{3/2}(t)}$.

Proof. For fixed $x \in \mathbb{R}^d$ and using Theorem 2.4.2 we have:

$$\mathcal{W}_1([X^{x,n}_{T_{n+1}-T_n}], \nu_{a_{n+1}}) \leq Ce^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \mathcal{W}_1(\delta_x, \nu_{a_{n+1}}).$$

So integrating x with respect to the law of $X_{T_n}^{x_0}$ (and using the existence of the optimal coupling, see for example [Wan12, Proposition 1.3]) yields:

$$\mathcal{W}_1([X_{T_{n+1}}^{x_0}], \nu_{a_{n+1}}) \leq Ce^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \left(\mathcal{W}_1([X_{T_n}^{x_0}], \nu_{a_n}) + \mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}})\right). \tag{2.5.3}$$

Iterating this relation yields

$$\begin{aligned}\mathcal{W}_1([X_{T_{n+1}}^{x_0}], \nu_{a_{n+1}}) &\leq \mu_{n+1}\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) + \mu_{n+1}\mu_n \mathcal{W}_1(\nu_{a_{n-1}}, \nu_{a_n}) + \cdots \\ &\quad + \mu_{n+1}\cdots\mu_1 \mathcal{W}_1(\nu_{a_0}, \nu_{a_1}) + \mu_{n+1}\cdots\mu_1 \mathcal{W}_1(\delta_{x_0}, \nu_{a_0}).\end{aligned} \tag{2.5.4}$$

where

$$\begin{aligned}\mu_n &:= Ce^{C_1/a_n^2} e^{-\rho_{a_n}(T_n - T_{n-1})} = C(T_n+e)^{C_1/A^2} e^{-(T_n+e)^{-C_2/A^2}(T_n - T_{n-1})} \\ &\leq C(C_{(T)}n^{\beta+1}+e)^{C_1/A^2} e^{-(C_{(T)}n^{\beta+1}+e)^{-C_2/A^2} C_{(T)}(\beta+1)(n-1)^\beta} \\ &\leq Cn^{(\beta+1)C_1/A^2} e^{-(C')^{1-C_2/A^2}(\beta+1)n^{\beta-(\beta+1)C_2/A^2}},\end{aligned} \tag{2.5.5}$$

where we have used (2.4.7) and where the last inequality holds for large enough n. Note that μ_n is bounded by a sequence in the form of $n^\delta \exp(-Ln^\eta) = o(n^{-\ell})$ for every $\ell \geq 0$. Owing to (2.5.2), we have $\beta - (\beta+1)C_2/A^2 > 0$.

On the other hand, if $Z \sim \nu_{a_0}$ then $\mathcal{W}_1(\delta_{x_0}, \nu_{a_0}) = \mathbb{E}|x_0 - Z| \leq |x_0| + \mathbb{E}|Z|$. Plugging this into (2.5.4) and using that $\mu_n \to 0$ so is bounded and smaller than 1 for n large enough and then $(\mu_{n-1}\cdots\mu_k)_{1\leq k\leq n-1}$ is bounded ; using Proposition 2.4.4 and that $\sum_n (n\log^{3/2}(n))^{-1} < \infty$ yields

$$\begin{aligned}\mathcal{W}_1([X_{T_{n+1}}^{x_0}], \nu_{a_{n+1}}) &\leq \mu_{n+1}\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) + C\mu_{n+1}\mu_n\big(\mathcal{W}_1(\nu_{a_{n-1}}, \nu_{a_n}) + \cdots \\ &\quad + \mathcal{W}_1(\nu_{a_0}, \nu_{a_1})\big) + C\mu_{n+1}\mu_n \mathcal{W}_1(\delta_{x_0}, \nu_{a_0}) \\ &\leq \mu_{n+1}\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) + C\mu_{n+1}\mu_n + C\mu_{n+1}\mu_n(1+|x_0|)\end{aligned}$$

$$\leq C\frac{\mu_{n+1}}{n\log^{3/2}(n)}(1+|x_0|) \leq C\mu_{n+1}a_{n+1}(1+|x_0|),$$

where we used that $\mu_n = o(\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}))$. Then using Lemma 2.4.6 we have

$$\mathcal{W}_1([X^{x_0}_{T_{n+1}}], \nu^\star) \leq \mathcal{W}_1([X^{x_0}_{T_{n+1}}], \nu_{a_{n+1}}) + \mathcal{W}_1(\nu_{a_{n+1}}, \nu^\star) \leq Ca_n(1+|x_0|),$$

where we used once again $\mu_n \to 0$.

Now, let us prove that $\mathcal{W}_1([X^{x_0}_t], \nu^\star) \to 0$ as $t \to \infty$. For $t \in [0, T_{n+1} - T_n)$ we integrate (2.4.4) with respect to the law of $X^{x_0}_{T_n}$, giving

$$\begin{aligned}
\mathcal{W}_1([X^{x_0}_{T_n+t}], \nu_{a_{n+1}}) &\leq Ce^{C_1 a_{n+1}^{-2}} e^{-\rho_{a_{n+1}} t}\mathcal{W}_1([X^{x_0}_{T_n}], \nu_{a_{n+1}}) \\
&\leq Ce^{C_1 a_{n+1}^{-2}} \left(\mathcal{W}_1([X^{x_0}_{T_n}], \nu_{a_n}) + \mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}})\right) \\
&\leq Ce^{C_1 a_{n+1}^{-2}} \mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}})(1+|x_0|) \\
&\leq \frac{C(1+|x_0|)}{n^{1-(\beta+1)C_1/A^2}\log^{3/2}(n)}. \qquad (2.5.6)
\end{aligned}$$

Now, for $t \geq 0$, let n be such that $t \in [T_n, T_{n+1})$. Then $(n+1) \geq t^{1/(\beta+1)}$ and

$$\begin{aligned}
\mathcal{W}_1([X^{x_0}_t], \nu_{a(t)}) &\leq \mathcal{W}_1([X^{x_0}_t], \nu_{a_{n+1}}) + \mathcal{W}_1(\nu_{a_{n+1}}, \nu_{a(t)}) \\
&\leq \frac{C(1+|x_0|)}{t^{(\beta+1)^{-1}-C_1/A^2}\log^{3/2}(t)}, \qquad (2.5.7)
\end{aligned}$$

where we used the second claim of Proposition 2.4.4.

Furthermore owing to (2.5.2) we have $(\beta+1)C_1/A^2 < 1$, so that

$$\mathcal{W}_1([X^{x_0}_{T_n+t}], \nu^\star) \leq \mathcal{W}_1([X^{x_0}_{T_n+t}], \nu_{a_{n+1}}) + \mathcal{W}_1(\nu_{a_{n+1}}, \nu^\star) \leq Ca_n(1+|x_0|).$$

□

Remark 2.5.2. We find again the classic schedule $a(t)$ of order $\log^{-1/2}(t)$. If for example we choose instead $a_n = \log(T_n)^{-(1+\varepsilon)/2}$ for some $\varepsilon > 0$, then we obtain

$$\log(\mu_1 \cdots \mu_n) = n\log(C) + \frac{C_1}{A^2}\sum_{k=1}^{n}\log^{1+\varepsilon}(T_k) - \sum_{k=1}^{n}\frac{T_k - T_{k-1}}{T_k^{\log^\varepsilon(T_k)C_2/A^2}}.$$

Hence, as $T_n - T_{n-1} = o(T_n^{\log^\varepsilon(T_n)C_2/A^2})$, $\mu_1 \cdots \mu_n$ does not converge to 0 whatever the value of $A > 0$ is.

2.6 Continuously decreasing case

We now consider (Y_t) solution to (2.2.5) i.e. the Langevin equation where the time coefficient $a(t)$ before σ is continuously decreasing. More generally, since Y is solution to a non-homogeneous SDE, we define for every $x \in \mathbb{R}^d$ and for every fixed $u \geq 0$:

$$Y^x_{0,u} = x, \quad dY^x_{t,u} = b_{a(t+u)}(Y^x_{t,u})dt + a(t+u)\sigma(Y^x_{t,u})dW_t, \qquad (2.6.1)$$

so that $Y^x = Y^x_{\cdot,0}$. We define the kernel associated to Y between the times t and $t+u$ as $P^Y_{t,u}$ such that for all $f : \mathbb{R}^d \to \mathbb{R}^+$ measurable, $P^Y_{t,u}f(x) = \mathbb{E}[f(Y^x_{t,u})]$. We also consider X as defined in (2.5.1) and its associated kernel denoted as $P^{X,n}_t$ such that for every $f : \mathbb{R}^d \to \mathbb{R}^+$ measurable, $P^{X,n}_t f(x) = \mathbb{E}[f(X^{x,n}_t)]$. In this section we prove the convergence of (Y_t) by giving bounds on $\mathcal{W}_1(Y^x_t, X^x_t)$ and using Theorem 2.5.1.

2.6.1 Boundedness of the potential

Lemma 2.6.1. *Let $p > 0$. Then there exists $C > 0$ such that for every $n \geq 0$, for every $u \geq 0$ and for every $x \in \mathbb{R}^d$:*

$$\sup_{t\geq 0} \mathbb{E}V^p(X_t^{x,n}) \leq CV^p(x) \quad \textit{and} \quad \sup_{t\geq 0} \mathbb{E}V^p(Y_{t,u}^x) \leq CV^p(x).$$

Proof. By the Itō Lemma, we have for $k \geq n$ and for $t \in [T_k - T_n, T_{k+1} - T_n)$:

$$\begin{aligned} dV^p(X_t^{x,n}) &= p\nabla V(X_t^{x,n})^\top \cdot V^{p-1}(X_t^{x,n}) a_{k+1}\sigma(X_t^{x,n})dW_t \\ &+ p\nabla V(X_t^{x,n})^\top \cdot V^{p-1}(X_t^{x,n}) \left(-\sigma\sigma^\top(X_t^{x,n})\nabla V(X_t^{x,n}) + a_{k+1}^2 \Upsilon(X_t^{x,n})\right) dt \\ &+ \frac{p}{2}\left(\nabla^2 V(X_t^{x,n})V^{p-1}(X_t^{x,n}) + (p-1)|\nabla V(X_t^{x,n})|^2 \cdot V^{p-2}(X_t^{x,n})\right) a_{k+1}^2 \sigma\sigma^\top(X_t^{x,n})dt. \end{aligned}$$

Using the facts that (a_k), Υ, σ, $\nabla^2 V$ are bounded, that $|\nabla V| \leq CV^{1/2}$ and that V, $|\nabla V|$ are coercive and $\sigma\sigma^\top \geq \underline{\sigma}_0^2 I_d$, there exists $R > 0$ such that if $|X_t^{x,n}| \geq R$ then the coefficient of dt in the last equation is bounded above by

$$pV^{p-1}\nabla V(X_t^{x,n})^T \cdot \left(-\underline{\sigma}_0^2(X_t^{x,n})\nabla V(X_t^{x,n}) + C\right) + CV^{p-1}(X_t^{x,n})||\sigma||_\infty^2 \leq 0,$$

so that

$$\mathbb{E}[V^p(X_t^{x,n})] \leq \max\left(\sup_{|z|\leq R} V^p(z), V^p(x)\right).$$

The proof is the same for Y, replacing a_{k+1} by $a(t)$. □

2.6.2 Strong and weak error bounds

In this subsection we adapt the proofs to bound weak and strong errors from [PP23] while paying attention to the dependence in a_n.

Lemma 2.6.2. *Let $p \geq 1$ and let $\bar{\gamma} > 0$. There exists $C > 0$ such that for all $n \geq 0$, u, $t \geq 0$ such that $u \in [T_n, T_{n+1}]$, $u + t \in [T_n, T_{n+1}]$ and $t \leq \bar{\gamma}$,*

$$||X_t^{x,n} - Y_{t,u}^x||_p \leq C\sqrt{t}(a_n - a_{n+1}).$$

Proof. We first consider the case $p \geq 2$. Noting that $a_{n+1} \leq a(u+s) \leq a_n$ for all $s \in [0,t]$ and using Lemma 2.11.1 in the Appendix, with in mind that $b_a = b_0 + a^2\Upsilon$, we have

$$\begin{aligned} \|X_t^{x,n} - Y_{t,u}^x\|_p &\leq \left\| \int_0^t (b_{a_{n+1}}(X_s^{x,n}) - b_{a(u+s)}(Y_{s,u}^x))ds \right\|_p \\ &\quad + \left\| \int_0^t (a_{n+1}\sigma(X_s^{x,n}) - a(u+s)\sigma(Y_{s,u}^x))dW_s \right\|_p \\ &\leq [b]_{\mathrm{Lip}} \int_0^t ||X_s^{x,n} - Y_{s,u}^x||_p ds + \int_0^t ||a_{n+1}^2\Upsilon(X_s^{x,n}) - a(u+s)^2\Upsilon(Y_{s,u}^x)||_p ds \\ &+ C_p^{BDG} a_{n+1}[\sigma]_{\mathrm{Lip}} \left(\int_0^t ||X_s^{x,n} - Y_{s,u}^x||_p^2 ds\right)^{1/2} + \left\| \int_0^t \sigma(Y_{s,u}^x)(a_{n+1} - a(u+s))dW_s \right\|_p \\ &\leq [b]_{\mathrm{Lip}} \int_0^t ||X_s^{x,n} - Y_{s,u}^x||_p ds + ||\Upsilon||_\infty (a_n^2 - a_{n+1}^2)t + a_{n+1}^2 [\Upsilon]_{\mathrm{Lip}} \int_0^t ||X_s^{x,n} - Y_{s,u}^x||_p ds \\ &\quad + C_p^{BDG} a_{n+1}[\sigma]_{\mathrm{Lip}} \left(\int_0^t ||X_s^{x,n} - Y_{s,u}^x||_p^2 ds\right)^{1/2} + ||\sigma||_\infty ||W_1||_p \sqrt{t}(a_n - a_{n+1}), \end{aligned}$$

where we used the generalized Minkowski inequality. Set $\varphi(t) := \sup_{0\le s\le t} ||X_s^{x,n} - Y_{s,u}^x||_p$ and $\psi(t) := ||\Upsilon||_\infty(a_n^2 - a_{n+1}^2)t + ||\sigma||_\infty||W_1||_p\sqrt{t}(a_n - a_{n+1})$. Both functions are non-decreasing and

$$\varphi(t) \le \psi(t) + ([b]_{\text{Lip}} + a_{n+1}^2[\Upsilon]_{\text{Lip}}) \int_0^t ||X_s^{x,n} - Y_{s,u}^x||_p ds + C_p^{BDG} a_{n+1}[\sigma]_{\text{Lip}} \left(\int_0^t ||X_s^{x,n} - Y_{s,u}^x||_p^2 ds \right)^{1/2}.$$

Moreover, for every $\eta > 0$:

$$\left(\int_0^t \varphi(s)^2 ds \right)^{1/2} \le \sqrt{\varphi(t)} \sqrt{\int_0^t \varphi(s) ds} \le \frac{\eta}{2}\varphi(t) + \frac{1}{2\eta} \int_0^t \varphi(s) ds.$$

Taking $\eta = \left(C_p^{BDG} a_{n+1}[\sigma]_{\text{Lip}} \right)^{-1}$ yields:

$$\varphi(t) \le 2\psi(t) + \left(2[b]_{\text{Lip}} + 2a_{n+1}^2[\Upsilon]_{\text{Lip}} + (C^{BDG} a_{n+1}[\sigma]_{\text{Lip}})^2 \right) \int_0^t \varphi(s) ds.$$

So the Gronwall Lemma yields for every $t \in [0, \bar{\gamma}]$

$$\varphi(t) \le 2e^{(2[b]_{\text{Lip}} + 2a_{n+1}^2[\Upsilon]_{\text{Lip}} + (C_p^{BDG} a_{n+1}[\sigma]_{\text{Lip}})^2)\bar{\gamma}} \psi(t),$$

which completes the proof for $p \ge 2$, noting that $a_n^2 - a_{n+1}^2 \le 2a_n(a_n - a_{n+1}) = o(a_n - a_{n+1})$. If $p \in [1, 2)$, the inequality is still true remarking that $\|\cdot\|_p \le \|\cdot\|_2$. □

Lemma 2.6.3. *Let $p \ge 1$ and let $\bar{\gamma} > 0$. There exists a real constant $C \ge 0$ such that for all $n \ge 0$,*

$$\forall t \in [0, \bar{\gamma}],\ ||X_t^{x,n} - x||_p \le C V^{1/2}(x) \sqrt{t}.$$

Proof. We perform a proof similar to the proof of Lemma 2.6.2. For $p \ge 2$ we have

$$\begin{aligned} ||X_t^{x,n} - x||_p &\le \left\| \int_0^t b_{a_{n+1}}(X_s^{x,n}) ds \right\|_p + \left\| \int_0^t a_{n+1}\sigma(X_s^{x,n}) ds \right\|_p \\ &\le t|b_{a_{n+1}}(x)| + A\|\sigma\|_\infty \|W\|_1 \sqrt{t} + [b]_{\text{Lip}} \int_0^t ||X_s^{x,n} - x||_p ds \\ &\quad + A[\sigma]_{\text{Lip}} C_p^{BDG} \left(\int_0^t ||X_s^{x,n} - x||_p^2 ds \right)^{1/2}. \end{aligned}$$

From here we use the Gronwall Lemma as in the proof of Lemma 2.6.2. For $p \in [1, 2)$, we have $\|\cdot\|_p \le \|\cdot\|_2$. □

Proposition 2.6.4. *Let $\bar{\gamma} > 0$. There exists $C > 0$ such that for every $g : \mathbb{R}^d \to \mathbb{R}$ being $\mathcal{C}^2$, for every $\gamma \in (0, \bar{\gamma}]$, every $n \ge 0$ and every $u \ge 0$ such that $u \in [T_n, T_{n+1}]$ and $u + \gamma \in [T_n, T_{n+1}]$:*

$$|\mathbb{E}\left[g(Y_{\gamma,u}^x) \right] - \mathbb{E}\left[g(X_\gamma^{x,n}) \right]| \le C\gamma(a_n - a_{n+1})\Phi_g(x)$$

$$\text{with}\quad \Phi_g(x) = \max\left(|\nabla g(x)|, \left\| \sup_{\xi \in (X_\gamma^{x,n}, Y_{\gamma,u}^x)} ||\nabla^2 g(\xi)|| \right\|_2, V^{1/2}(x) \left\| \sup_{\xi \in (x, X_\gamma^{x,n})} ||\nabla^2 g(\xi)|| \right\|_2 \right).$$

Proof. By the second order Taylor formula, for every $y,\ z \in \mathbb{R}^d$:

$$g(z) - g(y) = \langle \nabla g(y) | z - y \rangle + \int_0^1 (1 - s)\nabla^2 g(sz + (1 - s)y) ds (z - y)^{\otimes 2}.$$

Applying this expansion with $y = X_\gamma^{x,n}$ and $z = Y_{\gamma,u}^x$ yields:

$$\mathbb{E}[g(Y_{\gamma,u}^x) - g(X_\gamma^{x,n})] = \langle \nabla g(x) | \mathbb{E}[Y_{\gamma,u}^x - X_\gamma^{x,n}] \rangle + \mathbb{E}[\langle \nabla g(X_\gamma^{x,n}) - \nabla g(x), Y_{\gamma,u}^x - X_\gamma^{x,n} \rangle] + \int_0^1 (1-s)\mathbb{E}\left[\nabla^2 g(sY_{\gamma,u}^x + (1-s)X_\gamma^{x,n})(Y_{\gamma,u}^x - X_\gamma^{x,n})^{\otimes 2}\right] ds. \quad (2.6.2)$$

The first term is bounded by $|\nabla g(x)| \cdot |\mathbb{E}[Y_{\gamma,u}^x - X_\gamma^{x,n}]|$, with

$$\begin{aligned} |\mathbb{E}[Y_{\gamma,u}^x - X_\gamma^{x,n}]| &= \Big| \mathbb{E}\left[\int_0^\gamma (b_{a(s+u)}(Y_{s,u}^x) - b_{a(s+u)}(X_s^{x,n}))ds\right] \\ &\quad + \mathbb{E}\left[\int_0^\gamma (b_{a(u+s)}(X_s^{x,n}) - b_{a_{n+1}}(X_s^{x,n}))ds\right] \Big| \\ &\leq C[b]_{\mathrm{Lip}}(a_n - a_{n+1}) \int_0^\gamma \sqrt{s} ds + ||\Upsilon||_\infty \gamma(a_n^2 - a_{n+1}^2) \leq C\gamma(a_n - a_{n+1}), \end{aligned}$$

where we used Lemma 2.6.2. Using Lemma 2.6.3 and Lemma 2.6.2 again, the second term in the right hand side of (2.6.2) is bounded by

$$C \left|\left| \sup_{\xi \in (x, X_\gamma^{x,n})} ||\nabla^2 g(\xi)|| \right|\right|_2 \sqrt{\gamma} V^{1/2}(x) \sqrt{\gamma}(a_n - a_{n+1}).$$

Using Lemma 2.6.2, the third term is bounded by

$$\frac{1}{2} C\gamma(a_n - a_{n+1})^2 \left|\left| \sup_{\xi \in (X_\gamma^{x,n}, Y_{\gamma,u}^x)} ||\nabla^2 g(\xi)|| \right|\right|_2.$$

□

Proposition 2.6.5. *Let T, $\bar{\gamma} > 0$. There exists $C > 0$ such that for every Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$ and every $t \in (0, T]$, for all $n \geq 0$, for all $\gamma < \bar{\gamma}$ and every $u \in [T_n, T_{n+1}]$ such that $u + t + \gamma \in [T_n, T_{n+1}]$,*

$$\left| \mathbb{E}\left[P_t^{X,n} f(Y_{\gamma,u}^x)\right] - \mathbb{E}\left[P_t^{X,n} f(X_\gamma^{x,n})\right] \right| \leq C a_{n+1}^{-2}(a_n - a_{n+1})[f]_{\mathrm{Lip}} \gamma t^{-1/2} V(x).$$

Proof. We apply Proposition 2.6.4 to $g_t := P_t^{X,n} f$ with $t > 0$. Following [PP23, Proposition 3.2(b)] while paying attention to the dependence in the ellipticity parameter a, we have

$$\Phi_{g_t}(x) \leq C[f]_{\mathrm{Lip}} a_{n+1}^{-2} t^{-1/2} \max\left(V^{1/2}(x), \left|\left| \sup_{\xi \in (X_\gamma^{x,n}, Y_{\gamma,u}^x)} V^{1/2}(\xi) \right|\right|_2, V^{1/2}(x) \left|\left| \sup_{\xi \in (x, X_\gamma^{x,n})} V^{1/2}(\xi) \right|\right|_2 \right).$$

But following (2.2.8, $\mathcal{H}_{V2}$), $\nabla V / V^{1/2}$ is bounded so $V^{1/2}$ is Lipschitz continuous and then

$$\begin{aligned} \left|\left| \sup_{\xi \in (x, X_\gamma^{x,n})} V^{1/2}(\xi) \right|\right|_2 &\leq \left|\left| V^{1/2}(x) + [V^{1/2}]_{\mathrm{Lip}} |X_\gamma^{x,n} - x| \right|\right|_2 \leq C V^{1/2}(x) \\ \left|\left| \sup_{\xi \in (X_\gamma^{x,n}, Y_{\gamma,u}^x)} V^{1/2}(\xi) \right|\right|_2 &\leq \left|\left| V^{1/2}(x) + [V^{1/2}]_{\mathrm{Lip}} \max(|X_\gamma^{x,n} - x|, |Y_{\gamma,u}^x - x|) \right|\right|_2 \leq C V^{1/2}(x), \end{aligned}$$

where we used Lemmas 2.6.3 and 2.6.2. We thus obtain the desired result. □

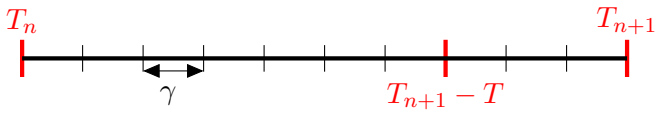


Figure 2.1: *Intervals for the domino strategy.*

2.6.3 Proof of Theorem 2.2.1.(a)

More precisely, we prove that for all $\beta > 0$, if

$$A > \max\left(\sqrt{(\beta+1)(2C_1+C_2)}, \sqrt{(1+\beta^{-1})C_2}\right), \tag{2.6.3}$$

then

$$\mathcal{W}_1([Y_t^{x_0}], \nu_{a(t)}) \leq \frac{C\max(1+|x_0|, V(x_0))}{\log^{3/2}(t)t^{(1+\beta)^{-1}-(2C_1+C_2)/A^2}}.$$

Proof. We apply the *domino strategy* (2.1.4). Let us fix $T \in (0, T_1)$ and $\gamma \in (0, T_1 - T)$. Here γ is not linked to any Euler-Maruyama scheme but is an auxiliary tool for the proof. Let $n \geq 0$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be Lipschitz continuous. We divide the two intervals $[T_n, T_{n+1} - T]$ and $[T_{n+1} - T, T_{n+1}]$ into smaller intervals of size γ (see Figure 2.1) and for $x \in \mathbb{R}^d$ using the semi-group property of $P^{X,n}$ on $[T_n, T_{n+1})$ we write:

$$\begin{aligned}
&\left|\mathbb{E}f(X^{x,n}_{T_{n+1}-T_n}) - \mathbb{E}f(Y^x_{T_{n+1}-T_n,T_n})\right| \\
&\leq \sum_{k=1}^{\lfloor (T_{n+1}-T_n-T)/\gamma \rfloor} \left|P^Y_{(k-1)\gamma,T_n} \circ (P^Y_{\gamma,T_n+(k-1)\gamma} - P^{X,n}_\gamma) \circ P^{X,n}_{T_{n+1}-T_n-k\gamma} f(x)\right| \\
&+ \sum_{k=\lfloor (T_{n+1}-T_n-T)/\gamma \rfloor+1}^{\lfloor (T_{n+1}-T_n)/\gamma \rfloor-1} \left|P^Y_{(k-1)\gamma,T_n} \circ (P^Y_{\gamma,T_n+(k-1)\gamma} - P^{X,n}_\gamma) \circ P^{X,n}_{T_{n+1}-T_n-k\gamma} f(x)\right| \\
&+ \left|P^Y_{\gamma(\lfloor (T_{n+1}-T_n)/\gamma \rfloor-1),T_n} \circ (P^Y_{\gamma+(T_{n+1}-T_n)\mathrm{mod}\gamma,T_n+\gamma(\lfloor (T_{n+1}-T_n)/\gamma \rfloor-1)} - P^{X,n}_{\gamma+(T_{n+1}-T_n)\mathrm{mod}(\gamma)}) f(x)\right| \\
&=: (a) + (b) + (c).
\end{aligned}$$

The term (a) is the "ergodic term", for which the exponential contraction from Theorem 2.4.2 can be exploited. The terms (b) and (c) are the "error terms" where we bound the error on intervals of length no larger than T. The term (c) is a remainder term due to the fact that $T_{n+1} - T_n$ is generally not a multiple of γ.

- **Term** (a) **:** It follows from Theorem 2.4.2 and Lemma 2.6.2 that

$$\begin{aligned}
&|(P^Y_{\gamma,T_n+(k-1)\gamma} - P^{X,n}_\gamma) \circ P^{X,n}_{T_{n+1}-T_n-k\gamma} f(x)| \\
&\quad = |\mathbb{E}P^{X,n}_{T_{n+1}-T_n-k\gamma} f(X^{x,n}_\gamma) - \mathbb{E}P^X_{T_{n+1}-T_n-k\gamma,n} f(Y^x_{\gamma,T_n+(k-1)\gamma})| \\
&\quad \leq Ce^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n-k\gamma)} [f]_{\mathrm{Lip}} \mathbb{E}|X^{x,n}_\gamma - Y^x_{\gamma,T_n+(k-1)\gamma}| \\
&\quad \leq Ce^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n-k\gamma)} [f]_{\mathrm{Lip}} \sqrt{\gamma}(a_n - a_{n+1})
\end{aligned}$$

Integrating with respect to $P^Y_{(k-1)\gamma,T_n}$ and summing up yields

$$\begin{aligned}
(a) &\leq Ce^{C_1 a_{n+1}^{-2}} [f]_{\mathrm{Lip}} \sqrt{\gamma}(a_n - a_{n+1}) \frac{e^{-\rho_{n+1}T} - e^{-\rho_{n+1}(T_{n+1}-T_n)}}{e^{\gamma\rho_{n+1}} - 1} \\
&\leq Ce^{C_1 a_{n+1}^{-2}} [f]_{\mathrm{Lip}} \sqrt{\gamma}(a_n - a_{n+1})(\gamma\rho_{n+1})^{-1}.
\end{aligned}$$

- **Term** (b)**:** Applying Proposition 2.6.5 yields:

$$|(P^Y_{\gamma,T_n+(k-1)\gamma} - P^{X,n}_{\gamma}) \circ P^{X,n}_{T_{n+1}-T_n-k\gamma} f(x)| \leq C a_{n+1}^{-2}(a_n - a_{n+1})[f]_{\mathrm{Lip}} \frac{\gamma}{\sqrt{T_{n+1}-T_n-k\gamma}} V(x).$$

Integrating with respect to $P^Y_{(k-1)\gamma,T_n}$ and using Lemma 2.6.1 which guarantees that $P^Y_{(k-1)\gamma,T_n} V(x) \leq CV(x)$ and summing with respect to k implies

$$(b) \leq C a_n^{-2}(a_n - a_{n+1})[f]_{\mathrm{Lip}} \gamma V(x) \sum_{k=1}^{\lceil T/\gamma \rceil} (k\gamma)^{-1/2} \leq C a_n^{-2}(a_n - a_{n+1})[f]_{\mathrm{Lip}} T^{1/2} V(x).$$

- **Term** (c)**:** Noting that $\gamma + (T_{n+1} - T_n) \bmod(\gamma) \leq 2\gamma$, Lemma 2.6.2 yields

$$(c) \leq C[f]_{\mathrm{Lip}} \sqrt{\gamma}(a_n - a_{n+1}).$$

Now we sum up the terms (a), (b) and (c). Since γ is constant we have:

$$\left| \mathbb{E} f(X^{x,n}_{T_{n+1}-T_n}) - \mathbb{E} f(Y^x_{T_{n+1}-T_n,T_n}) \right| \leq C e^{C_1 a_{n+1}^{-2}} (a_n - a_{n+1}) \rho_{n+1}^{-1} [f]_{\mathrm{Lip}} V(x),$$

so that for all $x \in \mathbb{R}^d$,

$$\mathcal{W}_1([X^{x,n}_{T_{n+1}-T_n}], [Y^x_{T_{n+1}-T_n,T_n}]) \leq C e^{C_1 a_{n+1}^{-2}} (a_n - a_{n+1}) \rho_{n+1}^{-1} V(x). \tag{2.6.4}$$

Temporarily setting $x_n := X^{x_0}_{T_n}$ and $y_n := Y^{x_0}_{T_n}$, we derive

$$\begin{aligned}
\mathcal{W}_1([X^{x_0}_{T_{n+1}}], [Y^{x_0}_{T_{n+1}}]) &= \mathcal{W}_1([X^{x_n,n}_{T_{n+1}-T_n}], [Y^{y_n}_{T_{n+1}-T_n,T_n}]) \\
&\leq \mathcal{W}_1([X^{x_n,n}_{T_{n+1}-T_n}], [X^{y_n,n}_{T_{n+1}-T_n}]) + \mathcal{W}_1([X^{y_n,n}_{T_{n+1}-T_n}], [Y^{y_n}_{T_{n+1}-T_n,T_n}]) \\
&\leq C e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n)} \mathcal{W}_1([X^{x_0}_{T_n}], [Y^{x_0}_{T_n}]) + C e^{C_1 a_{n+1}^{-2}} (a_n - a_{n+1}) \rho_{n+1}^{-1} \mathbb{E} V(Y^{x_0}_{T_n}),
\end{aligned}$$

where we used Theorem 2.4.2 and (2.6.4). We then apply Lemma 2.6.1 which guarantees that $(\mathbb{E}V(Y^{x_0}_{T_n}))_n$ is bounded by $CV(x_0)$. Let us denote

$$\lambda_n := C e^{C_1 a_n^{-2}} (a_{n-1} - a_n) \rho_n^{-1} = C e^{(C_1+C_2) a_n^{-2}} (a_{n-1} - a_n).$$

Owing to (2.6.3) we have $\lambda_n \to 0$. Iterating this relation and using (μ_n) defined in (2.5.5) yields like in the proof of Theorem 2.5.1:

$$\begin{aligned}
\mathcal{W}_1([X^{x_0}_{T_{n+1}}], [Y^{x_0}_{T_{n+1}}]) &\leq CV(x_0) \left(\lambda_{n+1} + \mu_{n+1}\lambda_n + \mu_{n+1}\mu_n \lambda_{n-1} + \cdots + \mu_{n+1} \cdots \mu_2 \lambda_1 \right) \\
&\leq CV(x_0) \left(\lambda_{n+1} + \mu_{n+1} \left(\lambda_n + \cdots + \lambda_1 \right) \right) \\
&\leq CV(x_0) \left(\lambda_{n+1} + n\mu_{n+1} \right).
\end{aligned}$$

But following (2.5.5) one checks that $n\mu_{n+1} = o(\lambda_{n+1})$ so that

$$\mathcal{W}_1([X^{x_0}_{T_{n+1}}], [Y^{x_0}_{T_{n+1}}]) \leq CV(x_0)\lambda_{n+1} \leq \frac{CV(x_0)}{\log^{3/2}(n+1)(n+1)^{1-(\beta+1)(C_1+C_2)/A^2}}.$$

Moreover, owing to (2.6.3) and combining with Theorem 2.5.1 we get

$$\mathcal{W}_1([Y^{x_0}_{T_n}], \nu_{a_n}) \leq \mathcal{W}_1([Y^{x_0}_{T_n}], [X^{x_0}_{T_n}]) + \mathcal{W}_1([X^{x_0}_{T_n}], \nu_{a_n}) \leq \frac{C \max(1 + |x_0|, V(x_0))}{\log^{3/2}(n) n^{1-(\beta+1)(C_1+C_2)/A^2}}$$

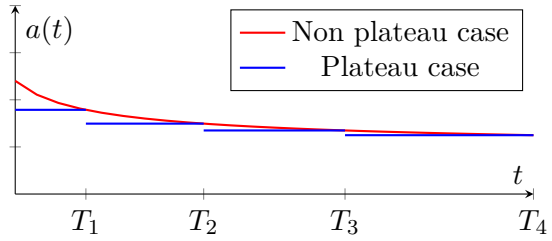


Figure 2.2: *Decreasing of the noise coefficient a for the plateau and non plateau cases.*

and as the right hand side of these inequalities is in $o(a_n)$, we derive

$$\mathcal{W}_1([Y^{x_0}_{T_n}],\nu^\star) \le \mathcal{W}_1([Y^{x_0}_{T_n}],[X^{x_0}_{T_n}]) + \mathcal{W}_1([X^{x_0}_{T_n}],\nu^\star) \le Ca_n \max(1+|x_0|, V(x_0)).$$

• **Convergence for $t \to \infty$:** Now let us prove that $\mathcal{W}_1([Y^{x_0}_t],\nu^\star) \to 0$ as $t \to \infty$. As before, let $T > 0$. For $t \ge T$, then we perform the same domino strategy where we replace T_{n+1} by $T_n + t$ and we consider the intervals $[T_n, T_n + t - T]$ and $[T_n + t - T, T_n + t]$. For $t < T$ then we only consider the terms (b) and (c) and we replace T by t in (b). Doing so we obtain

$$\mathcal{W}_1([X^{x,n}_t],[Y^x_{t,T_n}]) \le Ce^{C_1 a_{n+1}^{-2}}(a_n - a_{n+1})\rho_{n+1}^{-1}V(x).$$

So that, as before:

$$\begin{aligned}\mathcal{W}_1([X^{x_0}_{T_n+t}],[Y^{x_0}_{T_n+t}]) &\le Ce^{C_1 a_{n+1}^{-2}}\mathcal{W}_1([X^{x_0}_{T_n}],[Y^{x_0}_{T_n}]) + Ce^{C_1 a_{n+1}^{-2}}(a_n - a_{n+1})\rho_{n+1}^{-1}V(x_0)\\ &\le \frac{CV(x_0)}{\log^{3/2}(n)n^{1-(\beta+1)(2C_1+C_2)/A^2}}.\end{aligned}$$

Owing to (2.6.3) we have $1-(\beta+1)(2C_1+C_2)/A^2 > 1$, so that, using (2.5.6),

$$\mathcal{W}_1([Y^{x_0}_{T_n+t}],\nu_{a_{n+1}}) \le \mathcal{W}_1([Y^{x_0}_{T_n+t}],[X^{x_0}_{T_n+t}]) + \mathcal{W}_1([X^{x_0}_{T_n+t}],\nu_{a_{n+1}}) \le \frac{C\max(1+|x_0|,V(x_0))}{\log^{3/2}(n)n^{1-(\beta+1)(2C_1+C_2)/A^2}}.$$

We then prove the bound for $\mathcal{W}_1([Y^{x_0}_t],\nu_{a(t)})$ the same way as for (2.5.7), using the second claim of Proposition 2.4.4.

□

2.7 Continuously decreasing case : the Euler-Maruyama scheme

We now consider $(\bar{Y}_n)$ to be the Euler-Maruyama scheme of (Y_t) with steps (γ_n) defined in (2.2.15) and we also consider its genuine interpolation defined in (2.2.16). In this section we prove the convergence of $(\bar{Y}_t)$ by giving bounds on $\mathcal{W}_1(Y^x_t, X^x_t)$ and using Theorem 2.5.1. As with (2.6.1), we define more generally for every $n \ge 0$, $(\bar{Y}^x_{t,\Gamma_n})_{t\ge 0}$, first at times $\Gamma_k - \Gamma_n$, $k \ge n$, by

$$\begin{aligned}\bar{Y}^x_{0,\Gamma_n} = x,\ \bar{Y}^x_{\Gamma_{k+1}-\Gamma_n,\Gamma_n} &= \bar{Y}^x_{\Gamma_k-\Gamma_n,\Gamma_n} + \gamma_{k+1}\left(b_{a(\Gamma_k)}(\bar{Y}^x_{\Gamma_k-\Gamma_n,\Gamma_n}) + \zeta_{k+1}(\bar{Y}^x_{\Gamma_k-\Gamma_n,\Gamma_n})\right)\\ &\quad + a(\Gamma_k)\sigma(\bar{Y}^x_{\Gamma_k-\Gamma_n,\Gamma_n})(W_{\Gamma_{k+1}} - W_{\Gamma_k}),\end{aligned}$$

then at every time t by the genuine interpolation on the intervals $([\Gamma_k - \Gamma_n, \Gamma_{k+1} - \Gamma_n))_{k\ge n}$ as before. In particular $\bar{Y}^x = \bar{Y}^x_{\cdot,0}$. Still more generally, we define $\bar{Y}^x_{t,u}$ where $u \in (\Gamma_n, \Gamma_{n+1})$ as

$$\bar{Y}^x_{0,u} = x,\ \bar{Y}^x_{t,u} = \begin{cases} x + t(b_a(x) + \zeta_{n+1}(x)) + a^2(u)\sigma(x)(W_t - W_{\Gamma_u}) & \text{if } t \in [u, \Gamma_{n+1}]\\ \bar{Y}^{\bar{Y}^x_{\Gamma_{n+1}-u,u}}_{t-(\Gamma_{n+1}-u),\Gamma_{n+1}} & \text{if } t > \Gamma_{n+1}.\end{cases}$$

For n, $k \geq 0$, for $u \in [\Gamma_k, \Gamma_{k+1})$ and $\gamma \in [0, \Gamma_{k+1} - u]$, let $P^{\bar{Y}}_{\gamma,u}$ be the transition kernel associated to $\bar{Y}_{\cdot,u}$ between the times 0 and γ i.e. for all $f : \mathbb{R}^d \to \mathbb{R}^+$ measurable, $P^{\bar{Y}}_{\gamma,u} f(x) = \mathbb{E}[f(\bar{Y}^x_{\gamma,u})]$.

2.7.1 Boundedness of the potential

Lemma 2.7.1. *Let $p \geq 1/2$. There exists a constant $C > 0$ such that for every $k \geq 0$, for every $u \in [\Gamma_k, \Gamma_{k+1})$ and for every $x \in \mathbb{R}^d$:*

$$\sup_{n \geq k+1} \mathbb{E} V^p(\bar{Y}^x_{\Gamma_n - u, u}) \leq C V^p(x).$$

Proof. We rework the proof of Lemma 2(b) in [LP02]. Let us assume directly that $u = \Gamma_k$. To simplify the notations, we define $\widetilde{y}_n := \bar{Y}^x_{\Gamma_n - \Gamma_k, \Gamma_k}$ for $n \geq k$ and $\Delta\widetilde{y}_{n+1} := \widetilde{y}_{n+1} - \widetilde{y}_n$. The Taylor formula applied to V^p between $\widetilde{y}_n$ and $\widetilde{y}_{n+1}$ yields for some $\xi_{n+1} \in (\widetilde{y}_n, \widetilde{y}_{n+1})$ and with $\nabla^2(V^p) = p(V^{p-1}\nabla^2 V + (p-1)V^{p-2}\nabla V \nabla V^T)$:

$$\begin{aligned}
V^p(\widetilde{y}_{n+1}) &= V^p(\widetilde{y}_n) + pV^{p-1}\langle \nabla V(\widetilde{y}_n), \Delta\widetilde{y}_{n+1}\rangle + \frac{1}{2}\nabla^2(V^p)(\xi_{n+1}) \cdot (\Delta\widetilde{y}_{n+1})^{\otimes 2} \\
&= V^p(\widetilde{y}_n) + pV^{p-1}\nabla V(\widetilde{y}_n)^T \cdot \big(- \gamma_{n+1}\sigma\sigma^\top(\widetilde{y}_n)\nabla V(\widetilde{y}_n) + \gamma_{n+1} a^2(\Gamma_n)\Upsilon(\widetilde{y}_n) \\
&\quad + \gamma_{n+1}\zeta_{n+1}(\widetilde{y}_n) + \sqrt{\gamma_{n+1}} a(\Gamma_n)\sigma(\widetilde{y}_n) U_{n+1}\big) + \frac{1}{2}\nabla^2(V^p)(\xi_{n+1}) \cdot (\Delta\widetilde{y}_{n+1})^{\otimes 2},
\end{aligned}$$

where $U_{n+1} \sim \mathcal{N}(0, I_d)$. Moreover using (2.2.8, $\mathcal{H}_{V2}$), $\sqrt{V}$ is Lipschitz continuous so

$$\begin{aligned}
&\mathbb{E}\big[\sup_{z \in (\widetilde{y}_n, \widetilde{y}_{n+1})} V^{1/2}(z) | \widetilde{y}_1, \ldots, \widetilde{y}_n\big] \\
&\quad \leq V^{1/2}(\widetilde{y}_n) + [\sqrt{V}]_{\mathrm{Lip}} \mathbb{E}[|\widetilde{y}_{n+1} - \widetilde{y}_n| | \widetilde{y}_1, \ldots, \widetilde{y}_n] \leq C V^{1/2}(\widetilde{y}_n),
\end{aligned} \tag{2.7.1}$$

and in particular

$$\mathbb{E}[\|\nabla^2(V^p)(\xi_{n+1})\| | \widetilde{y}_1, \ldots, \widetilde{y}_n] \leq C\|\nabla^2(V^p)(\widetilde{y}_n)\|.$$

Moreover using that $\nabla^2 V$ is bounded and that $|\nabla V| \leq C V^{1/2}$ we have

$$\|\nabla^2(V^p)(\widetilde{y}_n)\| \leq C\|(V^{p-1}\nabla^2 V + V^{p-2}\nabla V \nabla V^T)(\widetilde{y}_n)\| \leq C V^{p-1}(\widetilde{y}_n).$$

Then using the facts that a, Υ, σ, $\nabla^2 V$ are bounded and that $\gamma_n^2 = o(\gamma_n)$, that V, ∇V are coercive and $\sigma\sigma^\top \geq \underline{\sigma}_0^2 I_d$ and (2.2.14), there exists $R > 0$ and $N \in \mathbb{N}$ such that if $|\widetilde{y}_n| \geq R$ and $n \geq N$ then

$$\begin{aligned}
&\mathbb{E}[V^p(\widetilde{y}_{n+1}) - V^p(\widetilde{y}_n) | \widetilde{y}_1, \ldots, \widetilde{y}_n] \\
&\quad \leq pV^{p-1}\nabla V(\widetilde{y}_n)^T \cdot \Big(- \gamma_{n+1}\underline{\sigma}_0^2(\widetilde{y}_n)\nabla V(\widetilde{y}_n) + C\gamma_{n+1}\Big) \\
&\qquad + C\|\nabla^2(V^p)(\widetilde{y}_n)\| \cdot \Big(\gamma_{n+1}^2 \|\sigma\|_\infty^4 |\nabla V(\widetilde{y}_n)|^2 + C\gamma_{n+1}^2 + C\gamma_{n+1}^2 V(x) + C\gamma_{n+1}\mathbb{E}|\mathcal{N}(0, I_d)|^2\Big) \\
&\quad \leq C\gamma_{n+1} V^{p-1}(\widetilde{y}_n)\Big[|\nabla V(\widetilde{y}_n)|\big(- |\nabla V(\widetilde{y}_n)| + 1\big) + \gamma_{n+1}(|\nabla V(\widetilde{y}_n)|^2 + 1) + 1\Big] \leq 0.
\end{aligned}$$

On the other side, if $|\widetilde{y}_n| \leq R$ then

$$\mathbb{E}[|V^p(\widetilde{y}_{n+1}) - V^p(\widetilde{y}_n)| | \widetilde{y}_1, \ldots, \widetilde{y}_n] \leq C\gamma_{n+1} \sup_{|x| \leq R} V^p(x).$$

Moreover for $n \in \{k, \ldots, N\}$ using (2.7.1) we have

$$\mathbb{E}[|V^p(\widetilde{y}_{n+1}) - V^p(\widetilde{y}_n)| | \widetilde{y}_1, \ldots, \widetilde{y}_n] \leq C V^p(\widetilde{y}_n)$$

so that

$$\sup_{k \leq n \leq N+1} \mathbb{E}[V^p(\widetilde{y}_n)] \leq C^{N-k} V^p(x).$$

Finally we obtain

$$\sup_{n \geq k} \mathbb{E}[V^p(\widetilde{y}_n)] \leq C V^p(x).$$

□

2.7.2 Strong and weak error bounds for the Euler-Maruyama scheme

Lemma 2.7.2. *Let $p \geq 1$. There exists $C > 0$ such that for every n, $k \geq 0$, for every $u \in [\Gamma_k, \Gamma_{k+1})$ and every $t > 0$ such that $u \in [T_n, T_{n+1}]$, $t \leq \Gamma_{k+1} - u$ and $u + t \in [T_n, T_{n+1}]$,*

$$||X_t^{x,n} - \bar{Y}_{t,u}^x||_p \leq C\left(V^{1/2}(x)t + \sqrt{t}(a_n - a_{n+1})\right).$$

Proof. As in the proof of Lemma 2.6.2, if $p \geq 2$ we have

$$\begin{aligned}||X_t^{x,n} - \bar{Y}_{t,u}^x||_p &\leq [b]_{\text{Lip}} \int_0^t ||X_s^{x,n} - x||_p ds + ||\Upsilon||_\infty (a_n^2 - a_{n+1}^2)t \\ &+ a_{n+1}^2 [\Upsilon]_{\text{Lip}} \int_0^t ||X_s^{x,n} - x||_p ds + \|\zeta_1(x)\|_p t \\ &+ C^{BDG} a_{n+1} [\sigma]_{\text{Lip}} \left(\int_0^t ||X_s^{x,n} - x||_p^2 ds\right)^{1/2} + ||\sigma||_\infty ||W_1||_p \sqrt{t}(a_n - a_{n+1}).\end{aligned}$$

Plugging Lemma 2.6.3 and (2.2.14) into this inequality yields:

$$||X_t^{x,n} - \bar{Y}_{t,u}^x||_p \leq CV^{1/2}(x)t^{3/2} + ||\Upsilon||_\infty (a_n^2 - a_{n+1}^2)t + CV^{1/2}(x)t + C\sqrt{t}(a_n - a_{n+1}),$$

which completes the proof for $p \geq 2$. If $p \in [1, 2)$, we remark that $\|\cdot\|_p \leq \|\cdot\|_2$. □

Proposition 2.7.3. *For every $g : \mathbb{R}^d \to \mathbb{R}$ being $\mathcal{C}^3$, for every n, $k \geq 0$ and every $u \in [\Gamma_k, \Gamma_{k+1})$ such that $u \in [T_n, T_{n+1}]$, $\gamma \leq \Gamma_{k+1} - u$ and $u + \gamma \in [T_n, T_{n+1}]$:*

$$\begin{aligned}|\mathbb{E}\left[g(\bar{Y}_{\gamma,u}^x)\right] - \mathbb{E}\left[g(X_\gamma^{x,n})\right]| &\leq CV^{1/2}(x)\left(V^{1/2}(x)\gamma^2 + \gamma(a_n - a_{n+1})\right)\bar{\Phi}_{g,1}(x) \\ &+ CV(x)\left(V^{1/2}(x)\gamma^2 + \gamma^{3/2}(a_n - a_{n+1})\right)\bar{\Phi}_{g,2}(x),\end{aligned}$$

with

$$\begin{aligned}\bar{\Phi}_{g,1}(x) &= \max\left(|\nabla g(x)|, ||\nabla^2 g(x)||, \left|\left|\sup_{\xi \in (X_\gamma^{x,n}, \bar{Y}_{\gamma,u}^x)} ||\nabla^2 g(\xi)||\right|\right|_2\right), \\ \bar{\Phi}_{g,2}(x) &= \left|\left|\sup_{\xi \in (x, X_\gamma^{x,n})} ||\nabla^3 g(\xi)||\right|\right|_4.\end{aligned}$$

The proof is given in the Supplementary Material.

Proposition 2.7.4. *Let $T > 0$. There exists $C > 0$ such that for every Lipschitz continuous function f and every $t \in (0, T]$, for all n, $k \geq 0$, for all $u \in [\Gamma_k, \Gamma_{k+1})$, for all γ such that $\Gamma_k \in [T_n, T_{n+1}]$, $\gamma \leq \Gamma_{k+1} - u$ and $u + t + \gamma \in [T_n, T_{n+1}]$,*

$$\begin{aligned}&\left|\mathbb{E}\left[P_t^{X,n} f(\bar{Y}_{\gamma,u}^x)\right] - \mathbb{E}\left[P_t^{X,n} f(X_\gamma^{x,n})\right]\right| \\ &\quad \leq C[f]_{\text{Lip}} V^2(x) \cdot \left(a_{n+1}^{-2} t^{-1/2}\left(\gamma^2 + (a_n - a_{n+1})\gamma\right) + a_{n+1}^{-3} t^{-1}\left(\gamma^2 + \gamma^{3/2}(a_n - a_{n+1})\right)\right).\end{aligned}$$

Proof. The proof is the same as for Proposition 2.6.5. □

2.7.3 Proof of Theorem 2.2.1.(b)

More precisely, we prove that for all $\beta > 0$, if

$$A > \max\left(\sqrt{(\beta+1)(2C_1+C_2)}, \sqrt{(1+\beta^{-1})C_2}\right), \tag{2.7.2}$$

then

$$\mathcal{W}_1([\bar{Y}_t^{x_0}], \nu_{a(t)}) \leq \frac{C \max\left(1+|x_0|, V^2(x_0)\right)}{t^{(1+\beta)^{-1}-(2C_1+C_2)/A^2}}.$$

Proof. We apply the same *domino strategy* as in Section 2.6.3. Let $n \geq 0$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be Lipschitz continuous. Let us denote

$$\gamma^{\text{init}} := \Gamma_{N(T_n)+1} - T_n \leq \gamma_{N(T_n)+1} \quad \text{and} \quad \gamma^{\text{end}} := T_{n+1} - \Gamma_{N(T_{n+1})} \leq \gamma_{N(T_{n+1})+1}.$$

For $x \in \mathbb{R}^d$ we write:

$$\begin{aligned}
&\left|\mathbb{E}f(X^{x,n}_{T_{n+1}-T_n}) - \mathbb{E}f(\bar{Y}^x_{T_{n+1}-T_n,T_n})\right| \leq \left|(P^{\bar{Y}}_{\gamma^{\text{init}},T_n} - P^{X,n}_{\gamma^{\text{init}}}) \circ P^{X,n}_{T_{n+1}-\Gamma_{N(T_n)+1}} f(x)\right| \\
&+ \sum_{k=N(T_n)+2}^{N(T_{n+1}-T)} \left|P^{\bar{Y}}_{\gamma^{\text{init}},T_n} \circ P^{\bar{Y}}_{\gamma_{N(T_n)+2},\Gamma_{N(T_n)+1}} \circ \cdots \circ P^{\bar{Y}}_{\gamma_{k-1},\Gamma_{k-2}} \circ (P^{\bar{Y}}_{\gamma_k,\Gamma_{k-1}} - P^{X,n}_{\gamma_k}) \circ P^{X,n}_{T_{n+1}-\Gamma_k} f(x)\right| \\
&+ \sum_{k=N(T_{n+1}-T)+1}^{N(T_{n+1})-1} \left|P^{\bar{Y}}_{\gamma^{\text{init}},T_n} \circ P^{\bar{Y}}_{\gamma_{N(T_n)+2},\Gamma_{N(T_n)+1}} \circ \cdots \circ P^{\bar{Y}}_{\gamma_{k-1},\Gamma_{k-2}} \circ (P^{\bar{Y}}_{\gamma_k,\Gamma_{k-1}} - P^{X,n}_{\gamma_k}) \circ P^{X,n}_{T_{n+1}-\Gamma_k} f(x)\right| \\
&+ \Big|P^{\bar{Y}}_{\gamma^{\text{init}},T_n} \circ P^{\bar{Y}}_{\gamma_{N(T_n)+2},\Gamma_{N(T_n)+1}} \circ \cdots \circ P^{\bar{Y}}_{\gamma_{N(T_{n+1})-1},\Gamma_{N(T_{n+1})-2}} \\
&\quad \circ (P^{\bar{Y}}_{\gamma^{\text{end}}+\gamma_{N(T_{n+1})},\Gamma_{N(T_{n+1})-1}} - P^{X,n}_{\gamma^{\text{end}}+\gamma_{N(T_{n+1})}}) f(x)\Big| \\
&=: (c^{\text{init}}) + (a) + (b) + (c^{\text{end}}).
\end{aligned}$$

- **Term** (a)**:** we have

$$\begin{aligned}
&|(P^{\bar{Y}}_{\gamma_k,\Gamma_{k-1}} - P^{X,n}_{\gamma_k}) \circ P^{X,n}_{T_{n+1}-\Gamma_k} f(x)| \\
&= |P^{\bar{Y}}_{\gamma_k,\Gamma_{k-1}} \circ P^{X,n}_{T/2} \circ P^{X,n}_{T_{n+1}-\Gamma_k-T/2} f(x) - P^{X,n}_{\gamma_k,n} \circ P^{X,n}_{T/2} \circ P^{X,n}_{T_{n+1}-\Gamma_k-T/2} f(x)| \\
&\leq |\mathbb{E}P^{X,n}_{T_{n+1}-\Gamma_k-T/2}(\Xi^x_k) - \mathbb{E}P^{X,n}_{T_{n+1}-\Gamma_k-T/2}(\bar{\Xi}^x_k)| \\
&\leq Ce^{Ca^{-2}_{n+1}} e^{-\rho_{n+1}(T_{n+1}-\Gamma_k-T/2)} [f]_{\text{Lip}} \mathbb{E}|\Xi^x_k - \bar{\Xi}^x_k|,
\end{aligned}$$

where Ξ^x_k and $\bar{\Xi}^x_k$ are any random vectors with laws $\left[X^{X^{x,n}_{\gamma_k},n}_{T/2}\right]$ and $\left[X^{\bar{Y}^x_{\gamma_k,\Gamma_{k-1}},n}_{T/2}\right]$ respectively and where we used Theorem 2.4.2 to get the last inequality. Thus, it follows from the definition of the Wasserstein distance that

$$\begin{aligned}
&|(P^{\bar{Y}}_{\gamma_k,\Gamma_{k-1}} - P^{X,n}_{\gamma_k}) \circ P^{X,n}_{T_{n+1}-\Gamma_k} f(x)| \\
&\qquad \leq Ce^{Ca^{-2}_{n+1}} e^{-\rho_{n+1}(T_{n+1}-\Gamma_k)} [f]_{\text{Lip}} \mathcal{W}_1\left(X^{X^{x,n}_{\gamma_k},n}_{T/2}, X^{\bar{Y}^x_{\gamma_k,\Gamma_{k-1}},n}_{T/2}\right).
\end{aligned}$$

On the other hand, the Kantorovich-Rubinstein representation of the L^1-Wasserstein distance (see [Vil09, Equation (6.3)]) reads

$$\mathcal{W}_1\left(X^{X^{x,n}_{\gamma_k},n}_{T/2}, X^{\bar{Y}^x_{\gamma_k,\Gamma_{k-1}},n}_{T/2}\right) = \sup_{[g]_{\text{Lip}}=1} \mathbb{E}\left[g\left(X^{X^{x,n}_{\gamma_k},n}_{T/2}\right) - g\left(X^{\bar{Y}^x_{\gamma_k,\Gamma_{k-1}},n}_{T/2}\right)\right]$$

$$= \sup_{[g]_{\mathrm{Lip}}=1} \mathbb{E}\left[P^{X,n}_{T/2} g(X^{x,n}_{\gamma_k}) - P^{X,n}_{T/2} g(\bar{Y}^{x}_{\gamma_k,\Gamma_{k-1}})\right].$$

It follows from Proposition 2.7.4 and using $[g]_{\mathrm{Lip}} = 1$ that

$$\mathbb{E}\left[P^{X,n}_{T/2} g(X^{x,n}_{\gamma_k}) - P^{X,n}_{T/2} g(\bar{Y}^{x}_{\gamma_k,\Gamma_{k-1}})\right] \leq C a_{n+1}^{-3}\left(\gamma_k^2 + (a_n - a_{n+1})\gamma_k\right) V^2(x),$$

so that

$$\begin{aligned}
&|(P^{\bar{Y}}_{\gamma_k,\Gamma_{k-1}} - P^{X,n}_{\gamma_k}) \circ P^{X,n}_{T_{n+1}-\Gamma_k} f(x)| \\
&\quad \leq C e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-\Gamma_k)} [f]_{\mathrm{Lip}} a_{n+1}^{-3}\left(\gamma_k^2 + (a_n - a_{n+1})\gamma_k\right) V^2(x).
\end{aligned}$$

Finally, integrating with respect to $P^{\bar{Y}}_{\gamma^{\mathrm{init}},T_n} \circ P^{\bar{Y}}_{\gamma_{N(T_n)+2},\Gamma_{N(T_n)+1}} \circ \cdots \circ P^{\bar{Y}}_{\gamma_{k-1},\Gamma_{k-2}}$ yields:

$$\begin{aligned}
&\left|P^{\bar{Y}}_{\gamma^{\mathrm{init}},T_n} \circ P^{\bar{Y}}_{\gamma_{N(T_n)+2},\Gamma_{N(T_n)+1}} \circ \cdots \circ P^{\bar{Y}}_{\gamma_{k-1},\Gamma_{k-2}} \circ (P^{\bar{Y}}_{\gamma_k,\Gamma_{k-1}} - P^{X,n}_{\gamma_k}) \circ P^{X,n}_{T_{n+1}-\Gamma_k} f(x)\right| \\
&\quad \leq C e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-\Gamma_k)} [f]_{\mathrm{Lip}} a_{n+1}^{-3}\left(\gamma_k^2 + (a_n - a_{n+1})\gamma_k\right) \\
&\qquad \cdot \left(\sup_{\ell \geq N(T_n)+1} \mathbb{E} V^2(\bar{Y}^{x}_{\gamma^{\mathrm{init}}+\Gamma_\ell - \Gamma_{N(T_n)+1},T_n})\right) \\
&\quad \leq C e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-\Gamma_k)} [f]_{\mathrm{Lip}} a_{n+1}^{-3}\left(\gamma_k^2 + (a_n - a_{n+1})\gamma_k\right) V^2(x),
\end{aligned}$$

where we used Lemma 2.7.1. Now, summing up over k yields:

$$\begin{aligned}
(a) &\leq C a_{n+1}^{-3} e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1} T_{n+1}} [f]_{\mathrm{Lip}} V^2(x) \sum_{k=N(T_n)+2}^{N(T_{n+1}-T)} ((a_n - a_{n-1}) + \gamma_k)\gamma_k e^{\rho_{n+1}\Gamma_k} \\
&\leq C a_{n+1}^{-3} e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1} T_{n+1}} [f]_{\mathrm{Lip}} ((a_n - a_{n-1}) + \gamma_{N(T_n)}) V^2(x) \sum_{k=N(T_n)+2}^{N(T_{n+1}-T)} \gamma_k e^{\rho_{n+1}\Gamma_{k-1}} \\
&\leq C a_{n+1}^{-3} e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1} T_{n+1}} [f]_{\mathrm{Lip}} ((a_n - a_{n-1}) + \gamma_{N(T_n)}) V^2(x) \int_{T_n}^{T_{n+1}-T} e^{\rho_{n+1} u} du \\
&\leq C a_{n+1}^{-3} e^{C_1 a_{n+1}^{-2}} [f]_{\mathrm{Lip}} ((a_n - a_{n-1}) + \gamma_{N(T_n)}) V^2(x) \rho_{n+1}^{-1} \\
&\leq C a_{n+1}^{-3} e^{C_1 a_{n+1}^{-2}} [f]_{\mathrm{Lip}} (a_n - a_{n-1}) V^2(x) \rho_{n+1}^{-1},
\end{aligned}$$

where we used that $(e^{\rho_{n+1}\gamma_k})_{n,k\geq 0}$ is bounded and Lemma 2.11.3 in the last inequality. We obtain likewise

$$(c^{\mathrm{init}}) \leq C e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n)} [f]_{\mathrm{Lip}} a_{n+1}^{-3} (a_n - a_{n+1}) \gamma_{N(T_n)+1} V^2(x).$$

- **Term** (b)**:** Applying Proposition 2.7.4 yields:

$$\begin{aligned}
(b) &\leq C a_{n+1}^{-3}\left(\gamma_{N(T_{n+1}-T)} + \sqrt{\gamma_{N(T_{n+1}-T)}}(a_n - a_{n+1})\right) \\
&\quad \cdot [f]_{\mathrm{Lip}} V^2(x) \sum_{k=N(T_{n+1}-T)+1}^{N(T_{n+1})-1} \frac{\gamma_k}{T_{n+1} - \Gamma_k} \\
&\quad + C a_{n+1}^{-2}\left(\gamma_{N(T_{n+1}-T)} + (a_n - a_{n+1})\right) [f]_{\mathrm{Lip}} V^2(x) \sum_{k=N(T_{n+1}-T)+1}^{N(T_{n+1})-1} \frac{\gamma_k}{\sqrt{T_{n+1} - \Gamma_k}}
\end{aligned}$$

$$
\begin{aligned}
&\leq C a_{n+1}^{-3} \left(\gamma_{N(T_{n+1}-T)} + \sqrt{\gamma_{N(T_{n+1}-T)}}(a_n - a_{n+1}) \right) \\
&\quad \cdot [f]_{\mathrm{Lip}} V^2(x) \int_{T_{n+1}-T}^{T_{n+1}-\gamma_{N(T_{n+1})}} \frac{1}{T_{n+1}-u} du \\
&\quad + C a_{n+1}^{-2} \left(\gamma_{N(T_{n+1}-T)} + (a_n - a_{n+1}) \right) [f]_{\mathrm{Lip}} V^2(x) \int_{T_{n+1}-T}^{T_{n+1}-\gamma_{N(T_{n+1})}} \frac{1}{\sqrt{T_{n+1}-u}} du \\
&\leq C a_{n+1}^{-3} \left(\gamma_{N(T_{n+1}-T)} + \sqrt{\gamma_{N(T_{n+1}-T)}}(a_n - a_{n+1}) \right) [f]_{\mathrm{Lip}} V^2(x) \log(1/\gamma_{N(T_{n+1})}) \\
&\quad + C a_{n+1}^{-2} (a_n - a_{n+1}) [f]_{\mathrm{Lip}} V^2(x).
\end{aligned}
$$

Using Lemma 2.11.4 in Appendix,

$$
\sqrt{\gamma_{N(T_{n+1}-T)}} \log(1/\gamma_{N(T_{n+1})}) \leq C \sqrt{\gamma_{N(T_{n+1})}} \log(1/\gamma_{N(T_{n+1})}) \to 0
$$

and using Lemma 2.11.3 we also have

$$
\gamma_{N(T_{n+1}-T)} \log(1/\gamma_{N(T_{n+1})}) \leq C \gamma_{N(T_{n+1})}^{1-\varepsilon} = o\left(n^{-1-\beta'} \right) = o(a_n - a_{n+1})
$$

where $\beta' > 0$ for small enough ε. So that

$$
(b) \leq C a_{n+1}^{-3} (a_n - a_{n+1}) [f]_{\mathrm{Lip}} V^2(x).
$$

- **Term** $(c^{\mathbf{end}})$**:** Using Lemma 2.7.2 and $\gamma^{\mathrm{end}} \leq \gamma_{N(T_{n+1})+1} \leq \gamma_{N(T_n)}$ yields:

$$
\begin{aligned}
&|(P^{\bar{Y}}_{\gamma^{\mathrm{end}}+\gamma_{N(T_{n+1})},\Gamma_{N(T_{n+1})-1}} - P^{X,n}_{\gamma^{\mathrm{end}}+\gamma_{N(T_{n+1})}}) f(x)| \\
&\qquad \leq C[f]_{\mathrm{Lip}} \left(\sqrt{\gamma_{N(T_n)}}(a_n - a_{n+1}) + \gamma_{N(T_n)} \right) V^{1/2}(x).
\end{aligned}
$$

Then we integrate with respect to $P^{\bar{Y}}_{\gamma^{\mathrm{init}},T_n} \circ P^{\bar{Y}}_{\gamma_{N(T_n)+2},\Gamma_{N(T_n)+1}} \circ \cdots \circ P^{\bar{Y}}_{\gamma_{k-1},\Gamma_{k-2}}$ and apply Lemma 2.7.1.

- So we have finally that $|\mathbb{E} f(X^{x,n}_{T_{n+1}-T_n}) - \mathbb{E} f(\bar{Y}^x_{T_{n+1}-T_n,T_n})|$ is bounded by

$$
C a_{n+1}^{-3} [f]_{\mathrm{Lip}} (a_n - a_{n+1}) e^{C_1 a_{n+1}^{-2}} \rho_{n+1}^{-1} V^2(x),
$$

which implies that, for every $x \in \mathbb{R}^d$,

$$
\mathcal{W}_1([X^{x,n}_{T_{n+1}-T_n}], [\bar{Y}^x_{T_{n+1}-T_n,T_n}]) \leq C a_{n+1}^{-3} (a_n - a_{n+1}) e^{C_1 a_{n+1}^{-2}} \rho_{n+1}^{-1} V^2(x).
$$

We integrate this inequality with respect to the laws of $X^{x_0}_{T_n}$ and $\bar{Y}^{x_0}_{T_n}$ and obtain, temporarily setting $x_n := X^{x_0}_{T_n}$ and $\bar{y}_n := \bar{Y}^{x_0}_{T_n}$,

$$
\begin{aligned}
&\mathcal{W}_1([X^{x_0}_{T_{n+1}}], [\bar{Y}^{x_0}_{T_{n+1}}]) = \mathcal{W}_1([X^{x_n,n}_{T_{n+1}-T_n}], [\bar{Y}^{\bar{y}_n}_{T_{n+1}-T_n,T_n}]) \\
&\leq \mathcal{W}_1([X^{x_n,n}_{T_{n+1}-T_n}], [X^{\bar{y}_n,n}_{T_{n+1}-T_n}]) + \mathcal{W}_1([X^{\bar{y}_n,n}_{T_{n+1}-T_n}], [\bar{Y}^{\bar{y}_n}_{T_{n+1}-T_n,T_n}]) \\
&\leq C e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n)} \mathcal{W}_1([X^{x_0}_{T_n}], [\bar{Y}^{x_0}_{T_n}]) + C a_{n+1}^{-3} (a_n - a_{n+1}) e^{C_1 a_{n+1}^{-2}} \rho_{n+1}^{-1} \mathbb{E} V^2(\bar{Y}^{x_0}_{T_n}) \\
&\leq C e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n)} \mathcal{W}_1([X^{x_0}_{T_n}], [\bar{Y}^{x_0}_{T_n}]) + C a_{n+1}^{-3} (a_n - a_{n+1}) e^{C_1 a_{n+1}^{-2}} \rho_{n+1}^{-1} V^2(x_0) \\
&=: \mu_{n+1} \mathcal{W}_1([X^{x_0}_{T_n}], [\bar{Y}^{x_0}_{T_n}]) + v_{n+1} V^2(x_0),
\end{aligned}
$$

where μ_n is defined in (2.5.5) and where we used again Lemma 2.7.1. We use Lemma 2.4.3 to bound $(a_n - a_{n+1})$ and owing to (2.7.2) we have $v_n \to 0$, so is bounded. We iterate this inequality and obtain

$$
\mathcal{W}_1([X^{x_0}_{T_{n+1}}], [\bar{Y}^{x_0}_{T_{n+1}}]) \leq C V^2(x_0) \left(v_{n+1} + \mu_{n+1} v_n + \mu_{n+1} \mu_n v_{n-1} + \cdots + \mu_{n+1} \cdots \mu_2 v_1 \right)
$$

$$\leq CV^2(x_0)\left(v_{n+1} + Cn\mu_{n+1}\right).$$

But following (2.5.5) we have $n\mu_n = O(v_n)$ so that

$$\mathcal{W}_1([X^{x_0}_{T_{n+1}}], [\bar{Y}^{x_0}_{T_{n+1}}]) \leq CV^2(x_0)v_{n+1} \leq \frac{CV^2(x_0)}{(n+1)^{1-(\beta+1)(C_1+C_2)/A^2}}.$$

Moreover, owing to (2.7.2) and combining with Theorem 2.5.1 we get

$$\mathcal{W}_1([\bar{Y}^{x_0}_{T_n}], \nu_{a_n}) \leq \mathcal{W}_1([\bar{Y}^{x_0}_{T_n}], [X^{x_0}_{T_n}]) + \mathcal{W}_1([X^{x_0}_{T_n}], \nu_{a_n}) \leq \frac{C\max(1+|x_0|, V^2(x_0))}{n^{1-(\beta+1)(C_1+C_2)/A^2}}$$

and

$$\mathcal{W}_1([\bar{Y}^{x_0}_{T_n}], \nu^\star) \leq \mathcal{W}_1([\bar{Y}^{x_0}_{T_n}], [X^{x_0}_{T_n}]) + \mathcal{W}_1([X^{x_0}_{T_n}], \nu^\star) \leq Ca_n \max(1+|x_0|, V^2(x_0)).$$

Finally, to prove that $\mathcal{W}_1([\bar{Y}^{x_0}_t], \nu^\star) \to 0$ as $t \to \infty$, we conclude as in the end of Section 2.6.3.

□

2.8 Convergence of the Euler-Maruyama scheme with plateau

In this section, we consider the Euler-Maruyama scheme for (X_t), that is

$$\bar{X}^{x_0}_0 = x_0, \quad \bar{X}^{x_0}_{\Gamma_{k+1}} = \bar{X}_{\Gamma_k} + \gamma_{k+1} b_{a_{n+1}}(\bar{X}_{\Gamma_k}) + a_{n+1}\sigma(\bar{X}_{\Gamma_k})(W_{\Gamma_{k+1}} - W_{\Gamma_k})$$

for $k \in \{N(T_n), \dots, N(T_{n+1}) - 1\}$. We also define as in Section 2.7 the genuine time-continuous scheme and the Euler-Maruyama scheme for $(X^{x,n}_t)_t$ so that $\bar{X}^{x_0,0} = \bar{X}^x$.

Although we already proved the convergence of the Euler-Maruyama scheme for (Y_t), we shall also prove the convergence of the present scheme, since this algorithm is also used by practitioners within the framework of batch methods.

Theorem 2.8.1. *Assume* (2.2.2, $\mathcal{H}_{V1}$), (2.2.8, $\mathcal{H}_{V2}$), (2.2.9, $\mathcal{H}_\sigma$), (2.2.10) *and* (2.2.11, $\mathcal{H}_{cf}$). *Assume furthermore* (2.2.12, $\mathcal{H}_{\gamma 1}$) *and* (2.2.13, $\mathcal{H}_{\gamma 2}$), *that* V *is* $\mathcal{C}^3$ *with* $\|\nabla^3 V\| \leq CV^{1/2}$ *and that* σ *is* $\mathcal{C}^3$ *with* $\|\nabla^3(\sigma\sigma^\top)\| \leq CV^{1/2}$. *Then for large enough* $A > 0$ *and for every* $x_0 \in \mathbb{R}^d$,

$$\mathcal{W}_1(\bar{X}^{x_0}_t, \nu^\star) \underset{t\to\infty}{\longrightarrow} 0.$$

The proof of this theorem is given in the Supplementary Material.

2.9 Experiments

In this section, we compare the performances of adaptive Langevin-Simulated Annealing algorithms versus vanilla SGLD, that is the Langevin algorithm with constant (additive) σ[1]. We train an artificial neural network on the MNIST dataset [LBBH98], which is composed of grayscale images of size 28×28 of handwritten digits (from 0 to 9). The goal is to recognize the handwritten digit and to classify the images. 60000 images are used for training and 10000 images are used for test.

As done by practitioners in general, we do not include the correction term Υ in the Langevin optimizers as computing second-order derivatives highly increases the training time. Moreover

[1]Our code is available at `https://github.com/Bras-P/langevin-simulated-annealing`.

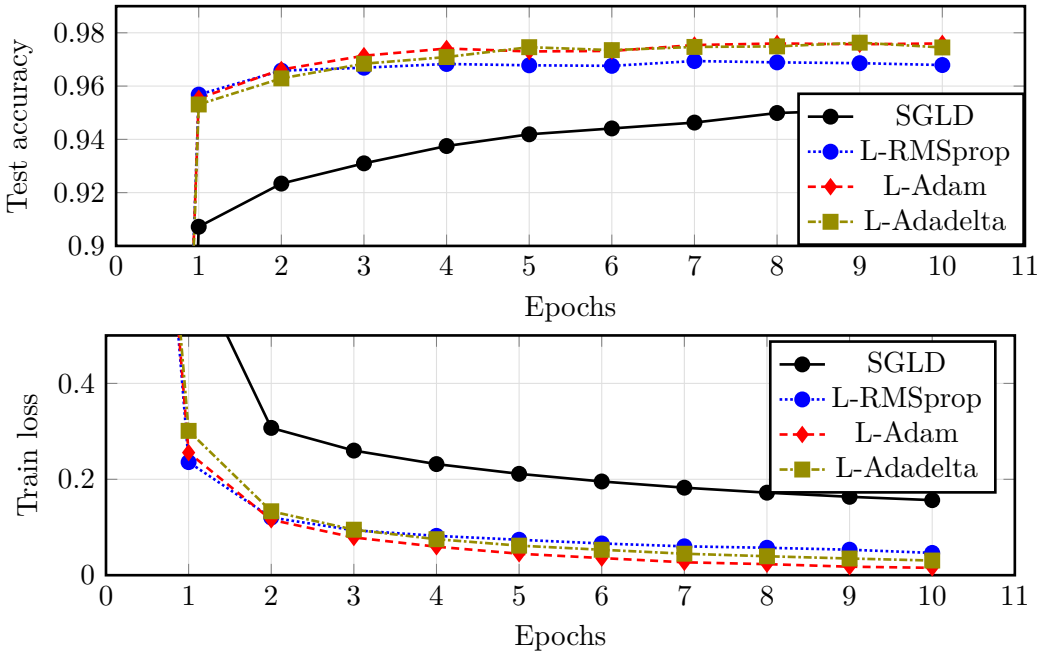


Figure 2.3: *Performance of preconditioned Langevin algorithms compared with vanilla SGLD on the MNIST dataset. The values of the hyperparameters are* $a(n) = A\log^{-1/2}(c_1 n + e)$ *with* $A = 2.10^{-3}$ *and where* $c_1 n = 1$ *after* 5 *epochs;* $\gamma_n = \gamma_1/(1 + c_2 n)$ *where* $c_2 n = 1$ *after* 5 *epochs and where for SGLD,* $\gamma_1 = 0.001$ *for L-RMSprop and L-Adam and* $\gamma_1 = 0.1$ *for L-Adadelta.*

Preconditioner	SGLD	L-RMSprop	L-Adam	L-Adadelta
Best accuracy	95,24 %	96,94 %	97,60 %	97,63 %

Table 2.1: *Best accuracy performance on the MNIST test set after 10 epochs.*

the time step coefficient of Υ is $a_n^2\gamma_n$ with $a_n \ll 1$ and $\gamma_n \ll 1$, in comparison with γ_n for the gradient and $a_n\gamma_n^{1/2}$ for the Gaussian noise. We refer to Chapter 5 for more details.

We consider a feedforward neural network with two hidden dense layers with 128 units each and with ReLU activation. For the adaptive Langevin algorithms, we choose the function σ as a diagonal matrix which is the square root of the preconditioner in RMSprop [LCCC16], in Adam [KB15] and in Adadelta [Zei12] respectively (see also Section 2.3.2), giving L-RMSprop, L-Adam and L-Adadelta respectively. The results are given in Figure 2.3 and in Table 2.1.

As pointed out in the literature (see the references Section 2.3.2), the preconditioned Langevin algorithms show significant improvement compared with the vanilla SGLD algorithm. The convergence is faster and they achieve a lower error on the test set. We also display the value of the loss function on the train set during the training to show that the better performances of the preconditioned algorithms are not due to some overfitting effect.

We also compare preconditioned Langevin algorithms with their respective non-Langevin counterpart. For shallow neural networks, adding an exogenous noise does not seem to improve significantly the performances of the optimization algorithm. However, for deep neural networks, which are highly non-linear and which loss function has many local minima, the Langevin version is competitive with the currently widely used non-Langevin algorithms and can even lead to improvement. The results are given in Figure 2.4 where we used a deep neural network with 20 hidden layers with 32 units each and with ReLU activation.

In order to understand how sensitive are these methods to poor initialization, we run an

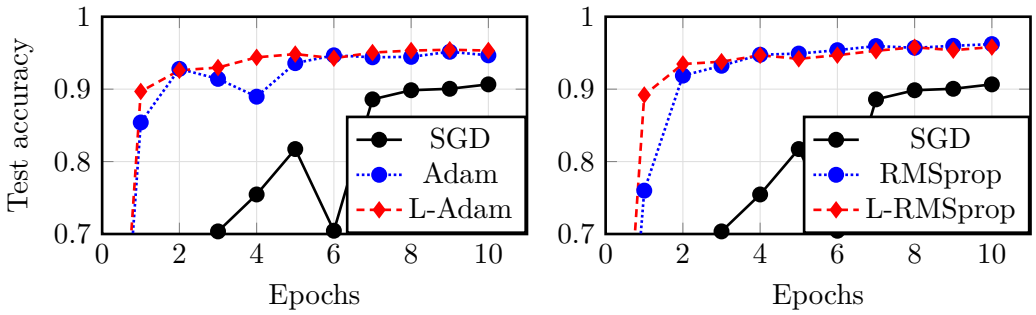


Figure 2.4: *Side-by-side comparison of optimization algorithms with their respective Langevin counterparts for the training of a deep neural network on the MNIST dataset. We display the performance of SGD for reference. The values of the hyperparameters are* $a(n) = A \log^{-1/2}(c_1 n + e)$ *with* $A = 1.10^{-3}$ *for L-Adam and* $A = 5.10^{-4}$ *for L-RMSprop and where* $c_1 n = 1$ *after* 5 *epochs;* $\gamma_n = \gamma_1/(1 + c_2 n)$ *where* $c_2 n = 1$ *after* 5 *epochs and where* $\gamma_1 = 0.01$ *for SGLD and* $\gamma_1 = 0.001$ *for the others.*

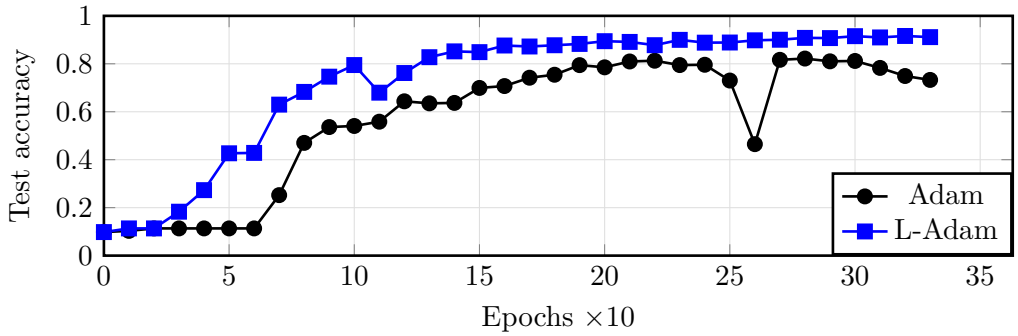


Figure 2.5: *Performance of the Adam optimizer compared with its Langevin version at the beginning of the training of a deep neural network on the MNIST dataset with poor initialization. We record the accuracy on the test set 10 times per epoch.*

experiment on the previous deep neural network where all the weights are initialized to zero, as in [NVL+15, Section 4.1]. We plot the accuracy on the test set in Figure 2.5. We observe that the non-Langevin optimizer needs some time before escaping from the neighbourhood of the initial point whereas in its Langevin version, the Gaussian noise is effective to rapidly escape from highly degenerated saddle points of the loss.

2.10 Conclusion and perspectives of future work

We proved the convergence of adaptive Langevin-Simulated annealing algorithms using both ergodic properties and weak error analysis, which fills a gap between the theory, which focused on the additive case, and the practice, where adaptive algorithm are commonly used to obtain faster convergence rates. Dealing with non constant diffusion coefficient σ makes the study more difficult as we need to rely on fine stochastic analysis. We finally obtained the same annealing schedule in $\log^{-1/2}(t)$ as in the additive case.

Future work would include relaxation of the assumptions on the coefficients of the SDE, in particular on the growth the potential V at infinity, as it is done for the additive case in [MFT21]. Another perspective would include weak error analysis of the corresponding projected Langevin-Simulated Annealing algorithm for optimization under constraint. We refer to [BEL15, Lam21]

where $\mathcal{W}_1$-contraction properties from [Ebe16, Wan20] are also used in an additive noise setting, but the multiplicative case is yet to be addressed by the literature.

2.11 Appendix

Lemma 2.11.1. *Let Z and $\widetilde{Z}$ be two continuous diffusion processes. Then for all $t \geq 0$ and for all $p \geq 2$:*

$$\left\| \int_0^t (\sigma(Z_s) - \sigma(\widetilde{Z}_s)) dW_s \right\|_p \leq C_p^{BDG} [\sigma]_{\mathrm{Lip}} \left(\int_0^t \|Z_s - \widetilde{Z}_s\|_p^2 ds \right)^{1/2},$$

where C_p^{BDG} is a constant which only depends on p.

Proof. It follows from the generalized Minkowski and the Burkholder-Davis-Gundy inequalities that

$$\begin{aligned} \left\| \int_0^t (\sigma(Z_s) - \sigma(\widetilde{Z}_s)) dW_s \right\|_p &\leq C_p^{BDG} [\sigma]_{\mathrm{Lip}} \left\| \int_0^t |Z_s - \widetilde{Z}_s|^2 ds \right\|_{p/2}^{1/2} \\ &\leq C_p^{BDG} [\sigma]_{\mathrm{Lip}} \left(\int_0^t \|Z_s - \widetilde{Z}_s\|_p^2 ds \right)^{1/2} \end{aligned}$$

□

We now give some results on the step sequence (γ_n) associated to the Euler-Maruyama scheme. Let us recall that the sequence (T_n) is defined in (2.4.7).

Lemma 2.11.2. *Let (u_n) be a positive and non-increasing sequence such that $\sum_n u_n < \infty$. Then $u_n = o(n^{-1})$.*

Proof. We have $N u_{2N} \leq \sum_{n=N}^{2N} u_n \to 0$ as $N \to \infty$. □

Lemma 2.11.3. *We have*

$$\gamma_{N(T_n)} = o\left(n^{-(1+\beta)}\right). \tag{2.11.1}$$

Proof. Using the previous lemma, $\gamma_n = o(n^{-1/2})$ so that $\Gamma_n = o(n^{1/2})$ and then $x^2 = o(N(x))$ as $x \to \infty$ and then $\gamma_{N(T_n)} = o\left(N(T_n)^{-1/2}\right) = o\left(n^{-(1+\beta)}\right)$. □

Lemma 2.11.4. *The sequence $(\gamma_{N(T_{n+1}-T)}/\gamma_{N(T_{n+1})})$ is bounded.*

Proof. Using (2.2.13, $\mathcal{H}_{\gamma 2}$), we have for $\varpi' > \varpi$ and for large enough k, $(\gamma_k - \gamma_{k+1})/\gamma_{k+1}^2 \leq \varpi'$ so that $\gamma_k/\gamma_{k+1} \leq 1 + \varpi' \gamma_{k+1}$ and then

$$\begin{aligned} \log\left(\frac{\gamma_{N(T_{n+1}-T))}}{\gamma_{N(T_{n+1})}} \right) &= \sum_{k=N(T_{n+1}-T)}^{N(T_{n+1})-1} \log\left(\frac{\gamma_k}{\gamma_{k+1}} \right) \leq C \sum_{k=N(T_n)}^{N(T_{n+1})-1} \gamma_k \\ &= C\left(\Gamma_{N(T_{n+1})} - \Gamma_{N(T_{n+1}-T)} \right) \leq C(T_{n+1} - (T_{n+1} - T)). \end{aligned}$$

□

2.12 Supplementary Material

2.12.1 Proof of Proposition 2.4.4

Proof. We have

$$\frac{\nu_{a_{n+1}}(x)}{\nu_{a_n}(x)} = \frac{\mathcal{Z}_{a_{n+1}}}{\mathcal{Z}_{a_n}} e^{-2(V(x)-V^\star)(a_{n+1}^{-2}-a_n^{-2})} \leq \frac{\mathcal{Z}_{a_{n+1}}}{\mathcal{Z}_{a_n}} =: M_n.$$

We now consider $(P_i)_{1\leq i\leq m^\star}$ a partition of $\mathbb{R}^d$ such that for all i, $x_i^\star \in \mathring{P}_i$. Let us prove that for all $1 \leq i \leq m^\star$,

$$\mathcal{Z}_{a,i}^{-1} := \int_{\mathbb{R}^d} e^{-2(V(x)-V^\star)/a^2} \mathbb{1}_{x\in P_i} dx \underset{a\to 0}{\sim} a^d \int_{\mathbb{R}^d} e^{-x^\top \nabla^2 V(x_i^\star)x} dx. \tag{2.12.1}$$

Let $r > 0$; let us consider $\widetilde{V}_i$ defined as

$$\widetilde{V}_i(x) = \begin{cases} V(x) & \text{if } x \in \mathbf{B}(x_i^\star, r) \\ |x - x_i^\star|^2 + V^\star & \text{otherwise.} \end{cases}$$

We also define $\widetilde{\mathcal{Z}}_{a,i}^{-1} := \int_{\mathbb{R}^d} e^{-2(\widetilde{V}_i(x)-V^\star)/a^2} \mathbb{1}_{x\in P_i} dx$. Then, owing to $V^\star > 0$ and (2.2.8, $\mathcal{H}_{V2}$),

$$\forall x \in \mathbb{R}^d,\ C|x - x_i^\star|^2 \leq \widetilde{V}_i(x) - V^\star \leq C'|x - x_i^\star|^2 \tag{2.12.2}$$

and then

$$\widetilde{\mathcal{Z}}_{a,i}^{-1} = a^d \int_{\mathbb{R}^d} e^{-2(\widetilde{V}_i(ax+x_i^\star)-V^\star)/a^2} \mathbb{1}_{x\in a^{-1}(P_i-x_i^\star)} dx \underset{a\to 0}{\sim} a^d \int_{\mathbb{R}^d} e^{-x^\top \nabla^2 V(x_i^\star)x} dx,$$

where we get the equivalence by dominated convergence ; the domination comes from (2.12.2). Then

$$\mathcal{Z}_{a,i}^{-1} - \widetilde{\mathcal{Z}}_{a,i}^{-1} = \int_{\mathbf{B}(x_i^\star,r)^c} e^{-2(V(x)-V^\star)/a^2} \mathbb{1}_{x\in P_i} dx - \int_{\mathbf{B}(x_i^\star,r)^c} e^{-2(\widetilde{V}_i(x)-V^\star)/a^2} \mathbb{1}_{x\in P_i} dx =: I_1 - I_2,$$

$$I_2 = a^d \int_{\mathbf{B}(0,r/a)^c} e^{-2|x|^2} \mathbb{1}_{x\in a^{-1}(P_i-x_i^\star)} dx \leq a^d \int_{\mathbf{B}(0,r/a)^c} e^{-2|x|^2} dx = o(a^d) = o\left(\widetilde{\mathcal{Z}}_{a,i}^{-1}\right).$$

Moreover using Proposition 4.3.1 we have $\mathcal{Z}_{a,i} I_1 \to 0$ as $a \to 0$, so that

$$\mathcal{Z}_{a,i}^{-1} = \widetilde{\mathcal{Z}}_{a,i}^{-1} + o\left(\mathcal{Z}_{a,i}^{-1}\right) + o\left(\widetilde{\mathcal{Z}}_{a,i}^{-1}\right) \sim \widetilde{\mathcal{Z}}_{a,i}^{-1},$$

which proves (2.12.1) and then

$$\mathcal{Z}_a^{-1} \underset{a\to 0}{\sim} a^d \sum_{i=1}^{m^\star} \int_{\mathbb{R}^d} e^{-x^\top \nabla^2 V(x_i^\star)x} dx. \tag{2.12.3}$$

We now prove that

$$\mathcal{Z}_{a_n}^{-1} - \mathcal{Z}_{a_{n+1}}^{-1} \leq C a_{n+1}^{d-1}(a_n - a_{n+1}). \tag{2.12.4}$$

Indeed, by convexity we have for all $z \in \mathbb{R}$

$$\left|e^{-2z/a_n^2} - e^{-2z/a_{n+1}^2}\right| \leq 2e^{-2z/a_n^2} z \left|\frac{1}{a_n^2} - \frac{1}{a_{n+1}^2}\right| \leq 4e^{-2z/a_n^2} \frac{z}{a_{n+1}^2} \frac{(a_n - a_{n+1})}{a_n}. \tag{2.12.5}$$

and then

$$\begin{aligned}\mathcal{Z}_{a_n,i}^{-1} - \mathcal{Z}_{a_{n+1},i}^{-1} &= a_{n+1}^d \int_{\mathbb{R}^d} \Big(e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_n^2} \mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)} \\ &\quad - e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_{n+1}^2} \mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)} \Big) dx \\ &\le 4a_{n+1}^{d-1}(a_n - a_{n+1}) \underbrace{\int_{\mathbb{R}^d} e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_n^2} \frac{V(a_{n+1}x+x_i^\star)-V^\star}{a_{n+1}^2} \mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)} dx}_{:=I_3}.\end{aligned}$$

Let us also define

$$\widetilde{I}_3 := \int_{\mathbb{R}^d} e^{-2(\widetilde{V}_i(a_{n+1}x+x_i^\star)-V^\star)/a_n^2} \frac{\widetilde{V}_i(a_{n+1}x+x_i^\star)-V^\star}{a_{n+1}^2} \mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)} dx.$$

Then $\widetilde{I}_3$ converges by dominated convergence and $|I_3 - \widetilde{I}_3|$ is bounded by

$$\begin{aligned}&\Big| \int_{\mathbb{R}^d} \Big(e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_n^2} \frac{V(a_{n+1}x+x_i^\star)-V^\star}{a_{n+1}^2} \\ &\qquad - e^{-2(\widetilde{V}_i(a_{n+1}x+x_i^\star)-V^\star)/a_n^2} \frac{\widetilde{V}_i(a_{n+1}x+x_i^\star)-V^\star}{a_{n+1}^2} \Big) \mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)} dx \Big| \\ &\le a_{n+1}^{-d-2} \int_{\mathbf{B}(x_i^\star,r)^c} e^{-2(V(x)-V^\star)/a_n^2} (V(x)-V^\star) \mathbb{1}_{x\in P_i} dx \\ &\quad + \int_{\mathbf{B}(0,r/a_{n+1})^c} e^{-2(\widetilde{V}_i(a_{n+1}x+x_i^\star)-V^\star)/a_n^2} \frac{\widetilde{V}_i(a_{n+1}x+x_i^\star)-V^\star}{a_{n+1}^2} \mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)} dx.\end{aligned}$$

The second integral converges to 0 by dominated convergence by similar arguments as for I_2. Moreover we have for every $x \in \mathbf{B}(x_i^\star, r)^c \cap P_i$, $V(x) - V^\star \ge \varepsilon$ for some $\varepsilon > 0$ and then for n such that $a_n \le A/\sqrt{2}$:

$$\begin{aligned}&a_{n+1}^{-d-2} \int_{\mathbf{B}(x_i^\star,r)^c} e^{-2(V(x)-V^\star)/a_n^2} (V(x)-V^\star) \mathbb{1}_{x\in P_i} dx \\ &\quad \le C a_{n+1}^{-d-2} \int_{\mathbf{B}(x_i^\star,r)^c} e^{-2(V(x)-V^\star)/a_n^2} |x-x_i^\star|^2 \mathbb{1}_{x\in P_i} dx \\ &\quad \le C a_{n+1}^{-d-2} e^{-\varepsilon/a_n^2} \int_{\mathbf{B}(x_i^\star,r)^c} e^{-(V(x)-V^\star)/a_n^2} |x-x_i^\star|^2 \mathbb{1}_{x\in P_i} dx \\ &\quad \le C a_{n+1}^{-d-2} e^{-\varepsilon/a_n^2} \int_{\mathbb{R}^d} e^{-2(V(x)-V^\star)/A^2} |x-x_i^\star|^2 dx \underset{n\to\infty}{\longrightarrow} 0,\end{aligned}$$

where we used that $(x \mapsto |x|^2 e^{-2(V(x)-V^\star)/A^2}) \in L^1(\mathbb{R}^d)$. Then we obtain that I_3 converges to $\widetilde{I}_3$, which proves (2.12.4). Then we have

$$1 - M_n^{-1} = \frac{\mathcal{Z}_{a_n}^{-1} - \mathcal{Z}_{a_{n+1}}^{-1}}{\mathcal{Z}_{a_n}^{-1}} \le C \frac{a_n - a_{n+1}}{a_n} \le \frac{C}{n \log(n)}.$$

On the other hand, if $X \sim \nu_{a_{n+1}}$, $\tilde{X} \sim \nu_{a_{n+1}}$, $Y \sim \nu_{a_n}$ and X, $\tilde{X}$ and Y are mutually independent then

$$\begin{aligned}&\Big| \mathbb{E}|X-Y| - \mathbb{E}|X-\tilde{X}| \Big| \\ &= \Big| a_{n+1}^d \mathcal{Z}_{a_n} a_{n+1}^d \mathcal{Z}_{a_{n+1}} \sum_{i,j=1}^{m^\star} \int\int a_{n+1} |x-y| e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_{n+1}^2}\end{aligned}$$

$$
\begin{aligned}
&\cdot e^{-2(V(a_{n+1}y+x_i^\star)-V^\star)/a_n^2}\cdot \mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)}\mathbb{1}_{y\in a_{n+1}^{-1}(P_j-x_i^\star)}dxdy\\
&-(a_{n+1}^d\mathcal{Z}_{a_{n+1}})^2\sum_{i,j=1}^{m^\star}\int\int a_{n+1}|x-y|e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_{n+1}^2}\\
&\cdot e^{-2(V(a_{n+1}y+x_i^\star)-V^\star)/a_{n+1}^2}\mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)}\mathbb{1}_{y\in a_{n+1}^{-1}(P_j-x_i^\star)}dxdy\Big|\\
=\;& a_{n+1}^{2d+1}\mathcal{Z}_{a_{n+1}}\sum_{i,j=1}^{m^\star}\int\int|x-y|e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_{n+1}^2}\\
&\cdot\Big|\mathcal{Z}_{a_n}e^{-2(V(a_{n+1}y+x_i^\star)-V^\star)/a_n^2}-\mathcal{Z}_{a_{n+1}}e^{-2(V(a_{n+1}y+x_i^\star)-V^\star)/a_{n+1}^2}\Big|\\
&\cdot\mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)}\mathbb{1}_{y\in a_{n+1}^{-1}(P_j-x_i^\star)}dxdy\\
\le\;& a_{n+1}\left(a_{n+1}^{2d}\mathcal{Z}_{a_{n+1}}^2\right)\sum_{i,j=1}^{m^\star}\int\int|x-y|e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_{n+1}^2}\\
&\cdot\Big|e^{-2(V(a_{n+1}y+x_i^\star)-V^\star)/a_n^2}-e^{-2(V(a_{n+1}y+x_i^\star)-V^\star)/a_{n+1}^2}\Big|\,\mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)}\mathbb{1}_{y\in a_{n+1}^{-1}(P_j-x_i^\star)}dxdy\\
&+a_{n+1}\left(a_{n+1}^{2d}\mathcal{Z}_{a_{n+1}}^2\right)\sum_{i,j=1}^{m^\star}\int\int|x-y|e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_{n+1}^2}e^{-2(V(a_{n+1}y+x_i^\star)-V^\star)/a_n^2}\\
&\quad\cdot\left|1-\frac{\mathcal{Z}_{a_n}}{\mathcal{Z}_{a_{n+1}}}\right|\mathbb{1}_{x\in a_{n+1}^{-1}(P_i-x_i^\star)}\mathbb{1}_{y\in a_{n+1}^{-1}(P_j-x_i^\star)}dxdy.
\end{aligned}
$$

So using (2.12.5), dominated convergence as for the proof of (2.12.3), (2.12.3) itself with (2.4.9) and the bound for $1-\mathcal{Z}_{a_n}/\mathcal{Z}_{a_{n+1}}=1-M_n^{-1}$ we have

$$
\begin{aligned}
&\limsup_{n\to\infty}\left[n\log^{3/2}(n)\left|\mathbb{E}|X-Y|-\mathbb{E}|X-\tilde{X}|\right|\right]\\
&\quad\le C\sum_{i=1}^{m^\star}\int\int|x-y|e^{-x^\top\nabla^2V(x_i^\star)x}e^{-y^\top\nabla^2V(x_i^\star)y}\left(1+y^\top\nabla^2V(x_i^\star)y\right)dxdy.
\end{aligned}
$$

So that using Lemma 2.4.5 and the fact that $\mathbb{E}|X-\tilde{X}|$ is of order a_n we have

$$
\begin{aligned}
\mathcal{W}_1(\nu_{a_n},\nu_{a_{n+1}})&\le\mathbb{E}|X-Y|-\frac{1}{M_n}\mathbb{E}|X-\tilde{X}|\\
&\le\mathbb{E}|X-Y|-\mathbb{E}|X-\tilde{X}|+\frac{C}{n\log(n)}\mathbb{E}|X-\tilde{X}|\le\frac{C}{n\log^{3/2}(n)}.
\end{aligned}
$$

The proof for the second claim is similar. □

2.12.2 Proof of Proposition 2.7.3

Proof. As in the proof of [PP23, Proposition 3.5], we split $|\mathbb{E}[g(\bar{Y}_{\gamma,u}^x)]-\mathbb{E}\left[g(X_\gamma^{x,n})\right]|$ into four terms A_1, A_2, A_3 and A_4, that is, by the Taylor formula, for every y, $z\in\mathbb{R}^d$,

$$
g(z)-g(y)=\langle\nabla g(y)|z-y\rangle+\int_0^1(1-u)\nabla^2g\left(uz+(1-u)y\right)du(z-y)^{\otimes2}.
$$

For a given $x\in\mathbb{R}^d$, it follows that

$$
g(z)-g(y)=\langle\nabla g(x)|z-y\rangle+\langle\nabla g(y)-\nabla g(x)|z-y\rangle
$$

$$
\begin{aligned}
&+ \int_0^1 (1-u)\nabla^2 g\left(uz + (1-u)y\right)(z-y)^{\otimes 2} du \\
&= \langle \nabla g(x) | z-y \rangle + \langle \nabla^2 g(x)(y-x) | z-y \rangle \\
&+ \int_0^1 (1-u)\nabla^3 g(uy + (1-u)x)(y-x)^{\otimes 2}(z-y) du \\
&+ \int_0^1 (1-u)\nabla^2 g\left(uz + (1-u)y\right) du (z-y)^{\otimes 2}.
\end{aligned}
$$

Applying this expansion with $y = X_\gamma^{x,n}$ and $z = \bar{Y}_{\gamma,u}^x$, this yields:

$$
\begin{aligned}
\mathbb{E}[g(\bar{Y}_{\gamma,u}^x) - g(X_\gamma^{x,n})] &= \underbrace{\langle \nabla g(x) | \mathbb{E}[\bar{Y}_{\gamma,u}^x - X_\gamma^{x,n}] \rangle}_{=:A_1} + \underbrace{\mathbb{E}\left[\langle \nabla^2 g(x)(X_\gamma^{x,n} - x) | \bar{Y}_{\gamma,u}^x - X_\gamma^{x,n} \rangle\right]}_{=:A_2} \\
&+ \underbrace{\mathbb{E}\left[\int_0^1 (1-u)\nabla^3 g(uX_\gamma^{x,n} + (1-u)x)(X_\gamma^{x,n} - x)^{\otimes 2}(\bar{Y}_{\gamma,u}^x - X_\gamma^{x,n}) du\right]}_{=:A_3} \\
&+ \underbrace{\int_0^1 (1-u)\mathbb{E}\left[\nabla^2 g\left(u\bar{Y}_{\gamma,u}^x + (1-u)X_\gamma^{x,n}\right)(\bar{Y}_{\gamma,u}^x - X_\gamma^{x,n})^{\otimes 2}\right] du}_{=:A_4}.
\end{aligned}
$$

- **Term** A_1: The term A_1 is bounded by $|\nabla g(x)| \cdot |\mathbb{E}[\bar{Y}_{\gamma,u}^x - X_\gamma^{x,n}]|$, with

$$
\begin{aligned}
\mathbb{E}[\bar{Y}_{\gamma,u}^x - X_\gamma^{x,n}] &= \mathbb{E}\left[\int_0^\gamma b_{a(u)}(x) - b_{a(u)}(X_s^{x,n}))ds\right] + \mathbb{E}\left[\int_0^\gamma (b_{a(u)}(X_s^{x,n}) - b_{a_{n+1}}(X_s^{x,n}))ds\right] \\
&=: A_{11} + A_{12}.
\end{aligned}
$$

We have $|A_{12}| \leq \gamma ||\Upsilon||_\infty (a_n^2 - a_{n+1}^2)$ and

$$
\begin{aligned}
|A_{11}| &= \left| \int_0^\gamma \int_0^s \mathbb{E}\left[\nabla b_{a(u)}(X_v^{x,n}) b_{a(u)}(X_v^{x,n}) + \frac{1}{2}\nabla^2 b_{a(u)}(X_v^{x,n}) a_{n+1}^2 \sigma\sigma^\top(X_v^{x,n})\right] dv \right| \\
&\leq C\gamma^2 \sup_{v \in [0,\gamma]} \mathbb{E}[V^{1/2}(X_v^{x,n})] \leq C\gamma^2 V^{1/2}(x),
\end{aligned}
$$

where we used that $|\nabla b_a| \leq C$ and $\|\nabla^2 b_a\| \leq CV^{1/2}$ because we assumed $\|\nabla^3 V\| \leq CV^{1/2}$ and $\|\nabla^3(\sigma\sigma^\top)\| \leq CV^{1/2}$.

- **Term** A_2: We have:

$$
|A_2| \leq \sum_{1 \leq i,j \leq d} |\partial_{ij} g(x)| |\mathbb{E}[(X_\gamma^{x,n} - x)_i (X_\gamma^{x,n} - \bar{Y}_{\gamma,u}^x)_j]|
$$

and we have

$$
\mathbb{E}[(X_\gamma^{x,n} - x)_i (X_\gamma^{x,n} - \bar{Y}_{\gamma,u}^x)_j] = \mathbb{E}[(X_\gamma^{x,n} - \bar{Y}_{\gamma,u}^x)_i (X_\gamma^{x,n} - \bar{Y}_{\gamma,u}^x)_j] + \mathbb{E}[(\bar{Y}_{\gamma,u}^x - x)_i (X_\gamma^{x,n} - \bar{Y}_{\gamma,u}^x)_j].
$$

Using Lemma 2.7.2, the first term of the right-hand side is bounded by $C(V^{1/2}(x)\gamma + \sqrt{\gamma}(a_n - a_{n+1}))^2$ and in the second term we write $(\bar{Y}_{\gamma,u}^x - x)_i = \left(\gamma b_{a(u)}(x) + \gamma\zeta_{k+1}(x) + a(u)\sigma(x)W_\gamma\right)_i$ and we have

$$
|\mathbb{E}[(\gamma b_{a(u)}(x) + \gamma\zeta_{k+1}(x))_i (X_\gamma^{x,n} - \bar{Y}_{\gamma,u}^x)_j]| \leq \gamma V^{1/2}(x)(V^{1/2}(x)\gamma + \sqrt{\gamma}(a_n - a_{n+1}))
$$

and using that the increments of ζ and W are independent,

$$
|\mathbb{E}[(a(u)\sigma(x)W_\gamma)_i (\gamma\zeta_{k+1}(x))_j]| = 0
$$

and using the Itō isometry:

$$\begin{aligned}
&\left|\mathbb{E}\left[(a(u)\sigma(x)W_\gamma)_i\left(\int_0^\gamma (b_{a_{n+1}}(X_s^{x,n}) - b_{a_{n+1}}(x) + b_{a_{n+1}}(x) - b_{a(u)}(x))_j ds\right.\right.\right.\\
&\quad\left.\left.\left.+\int_0^\gamma ((a_{n+1}\sigma(X_s^{x,n}) - a_{n+1}\sigma(x) + a_{n+1}\sigma(x) - a(u)\sigma(x))dW_s)_j\right)\right]\right|\\
&\leq C[b]_{\mathrm{Lip}}\int_0^\gamma ||W_\gamma||_2||X_s^{x,n} - x||_2 ds + C(a_n^2 - a_{n+1}^2)||W_\gamma||_1\gamma||\Upsilon||_\infty\\
&\quad + C\left|\sum_{k=1}^d \int_0^\gamma \mathbb{E}[\sigma_{ik}(x)(\sigma_{jk}(X_s^{x,n}) - \sigma_{jk}(x)]ds\right| + C(a_n - a_{n+1})\mathbb{E}[W_\gamma^2]\\
&\leq CV^{1/2}(x)\gamma^2 + C(a_n - a_{n+1})\gamma^{3/2} + CV^{1/2}(x)\gamma^2 + C(a_n - a_{n+1})\gamma,
\end{aligned}$$

where we used an argument similar to A_{11} to bound the third term, using that $\nabla\sigma$ and $\nabla^2\sigma$ are bounded.

• **Term** A_3: Using the three fold Cauchy-Schwarz inequality, Lemma 2.6.3 and Lemma 2.7.2, A_3 is bounded by

$$C\left|\left|\sup_{\xi\in(x,X_\gamma^{x,n})}||\nabla^3 g(\xi)||\right|\right|_4 V(x)\gamma\left(V^{1/2}(x)\gamma + \sqrt{\gamma}(a_n - a_{n+1})\right).$$

• **Term** A_4: Using Lemma 2.7.2, A_4 is bounded by

$$C\left(V^{1/2}(x)\gamma + \sqrt{\gamma}(a_n - a_{n+1})\right)^2\left|\left|\sup_{\xi\in(X_\gamma^{x,n},\bar{Y}_{\gamma,u}^x)}||\nabla^2 g(\xi)||\right|\right|_2.$$

□

2.12.3 Proof of Theorem 2.2.4

Proof. We remark that according to Proposition 4.3.1, for all $\kappa > 0$ we have $e^{-\kappa g} \in L^1(\mathbb{R}^d)$. We first prove that

$$\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) \leq \frac{C}{n\log^{1+\alpha_{\min}}(n)}, \tag{2.12.6}$$

so that $(\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}))$ is still a converging Bertrand series. To do so, we directly adapt the proof of Proposition 2.4.4, replacing the change of variables in the integrals in ax by the change of variables in $B\cdot(a^{2\alpha_1}x_1, \dots, a^{2\alpha_d}x_d)$. Still using (2.12.5), we successively obtain

$$\begin{aligned}
&\mathcal{Z}_a^{-1} \underset{a\to 0}{\sim} a^{2\alpha_1+\dots+2\alpha_d}\int_{\mathbb{R}^d} e^{-2g(x)}dx\\
&\mathcal{Z}_{a_n}^{-1} - \mathcal{Z}_{a_{n+1}}^{-1} \leq 4a_{n+1}^{2\alpha_1+\dots+2\alpha_d-1}(a_n - a_{n+1})\int_{\mathbb{R}^d} e^{-2g(x)}g(x)dx\\
&1 - M_n^{-1} \leq \frac{C}{n\log(n)}\\
&\left|\mathbb{E}|X - Y| - \mathbb{E}|X - \tilde{X}|\right| \leq \frac{Ca_{n+1}^{2\alpha_{\min}}}{n\log(n)}.
\end{aligned}$$

Then, using (2.12.6) we prove that $\mathcal{W}_1(\nu_n, \nu^\star) \leq Ca_n^{2\alpha_{\min}}$ the same way as in Lemma 2.4.6.

The next parts of the proof are the same as for the definite positive case. □

2.12.4 Proof of Theorem 2.5.1

To prove Theorem 2.8.1, we proceed as for the proof of Theorem 2.2.1.

In the following, for $\gamma > 0$ we denote by $(\bar{X}_t^{x,n,\gamma})_{t\in[0,\gamma]}$ the Euler-Maruyama scheme over one step with coefficient a_{n+1}. We first recall [PP23, Lemma 3.4(b), Proposition 3.5(a)] giving bounds for the weak and strong errors for the one-step Euler-Maruyama scheme, which do not depend on the ellipticity parameter a_n.

Lemma 2.12.1. *Let $p \geq 1$ and let $\bar{\gamma} > 0$. There exists $C > 0$ such that for every $n \geq 0$, for every $\gamma \in (0, \bar{\gamma}]$ and every $t \in [0, \gamma]$:*

$$||X_t^{x,n} - \bar{X}_t^{x,n,\gamma}||_p \leq CV^{1/2}(x)t.$$

Proposition 2.12.2. *Let $\bar{\gamma} > 0$. Then for every $g : \mathbb{R}^d \to \mathbb{R}$ being $\mathcal{C}^3$ and for every $0 \leq \gamma \leq \gamma' \leq \bar{\gamma}$:*

$$\left|\mathbb{E}\left[g(\bar{X}_\gamma^{x,n,\gamma'})\right] - \mathbb{E}\left[g(X_\gamma^{x,n})\right]\right| \leq CV^{3/2}(x)\gamma^2\Phi_g(x),$$

where

$$\Phi_g(x) = \max\left(|\nabla g(x)|, ||\nabla^2 g(x)||, \left|\left|\sup_{\xi\in(X_\gamma^{x,n},\bar{X}_\gamma^{x,n,\gamma'})} ||\nabla^2 g(\xi)||\right|\right|_2, \left|\left|\sup_{\xi\in(x,X_\gamma^{x,n})} ||\nabla^3 g(\xi)||\right|\right|_4\right).$$

Proposition 2.12.3. *Let T, $\bar{\gamma} > 0$. Then for every Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, for every $n \geq 0$ and every $t \in (0, T]$ and every $0 \leq \gamma \leq \gamma' \leq \bar{\gamma}$:*

$$\left|\mathbb{E}\left[P_t f(\bar{X}_\gamma^{x,n,\gamma'})\right] - \mathbb{E}\left[P_t f(X_\gamma^{x,n})\right]\right| \leq Ca_n^{-3}[f]_{\mathrm{Lip}}\gamma^2 t^{-1}V^2(x).$$

Proof. The proof is the same as in [PP23, Proposition 3.6]. When applying [PP23, Proposition 3.2(b)], we remark that the lowest exponent of ϱ_0 is -3. □

Moreover, by the same proof as in Lemma 2.7.1 we get

$$\sup_{m\geq k+1} \mathbb{E}V^p(\bar{X}_{\Gamma_m-\Gamma_k}^{x,n}) \leq CV^p(x).$$

We now prove Theorem 2.8.1.

Proof. Let us write:

$$\mathcal{W}_1([\bar{X}_{T_n}^{x_0}], \nu^\star) \leq \mathcal{W}_1([\bar{X}_{T_n}^{x_0}], [X_{T_n}^{x_0}]) + \mathcal{W}_1([X_{T_n}^{x_0}], \nu^\star).$$

Temporarily setting $\bar{x}_n := \bar{X}_{T_n}^{x_0}$ and $x_n := X_{T_n}^{x_0}$, we have

$$\begin{aligned}\mathcal{W}_1([\bar{X}_{T_{n+1}}^{x_0}], [X_{T_{n+1}}^{x_0}]) &= \mathcal{W}_1([\bar{X}_{T_{n+1}-T_n}^{\bar{x}_n,n}], [X_{T_{n+1}-T_n}^{x_n,n}]) \\ &\leq \mathcal{W}_1([\bar{X}_{T_{n+1}-T_n}^{\bar{x}_n,n}], [X_{T_{n+1}-T_n}^{\bar{x}_n,n}]) + \mathcal{W}_1([X_{T_{n+1}-T_n}^{\bar{x}_n,n}], [X_{T_{n+1}-T_n}^{x_n,n}]),\end{aligned}$$

and we find a bound on the first term using the same proof as in [PP23, Section 4.2]. For $x \in \mathbb{R}^d$, we split $|\mathbb{E}f(\bar{X}_{T_{n+1}-T_n}^{x,n}) - \mathbb{E}f(X_{T_{n+1}-T_n}^{x,n})|$ into three terms (a), (b) and (c). We however pay attention to the dependence in a_n when applying Lemma 2.12.1, Proposition 2.12.3 and Theorem 2.4.2. We then have:

$$\begin{aligned}(c) &\leq C[f]_{\mathrm{Lip}}\gamma_{N(T_n)}V^{1/2}(x), \\ (b) &\leq Ca_{n+1}^{-3}\gamma_{N(T_n)}\log\left(\frac{T + ||\gamma||_\infty}{\gamma_{N(T_n)}}\right), \\ (a) &\leq Ca_{n+1}^{-3}e^{C_1 a_{n+1}^{-2}}V(x)\gamma_{N(T_n)}\rho_{n+1}^{-1}.\end{aligned}$$

Then we establish a recursive relation and prove the convergence as in the proof of Theorem 2.2.1.

□

Chapter 3

Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise II: Total Variation

The results presented in this chapter have been published in *Monte Carlo Methods and Applications* as a joint work with Gilles Pagès [BP23a].

Abstract

We study the convergence of Langevin-Simulated Annealing type algorithms with multiplicative noise, i.e. for $V : \mathbb{R}^d \to \mathbb{R}$ a potential function to minimize, we consider the stochastic differential equation $dY_t = -\sigma\sigma^\top \nabla V(Y_t)dt + a(t)\sigma(Y_t)dW_t + a(t)^2\Upsilon(Y_t)dt$, where (W_t) is a Brownian motion, where $\sigma : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ is an adaptive (multiplicative) noise, where $a : \mathbb{R}^+ \to \mathbb{R}^+$ is a function decreasing to 0 and where Υ is a correction term. Allowing σ to depend on the position brings faster convergence in comparison with the classical Langevin equation $dY_t = -\nabla V(Y_t)dt + \sigma dW_t$. In a previous paper we established the convergence in L^1-Wasserstein distance of Y_t and of its associated Euler scheme $\bar{Y}_t$ to $\operatorname{argmin}(V)$ with the classical schedule $a(t) = A\log^{-1/2}(t)$. In the present paper we prove the convergence in total variation distance. The total variation case appears more demanding to deal with and requires regularization lemmas.

Keywords– Stochastic Optimization, Langevin Equation, Simulated Annealing, Neural Networks.

3.1 Introduction

Langevin-based algorithms are used to solve optimization problems in high dimension and have gained much interest in relation with Machine Learning. The Langevin equation is a Stochastic

Differential Equation (SDE) which consists in a gradient descent with noise. More precisely, let $V : \mathbb{R}^d \to \mathbb{R}^+$ be a coercive potential function, then the associated Langevin equation reads

$$dX_t = -\nabla V(X_t)dt + \sigma dW_t,\ t \geq 0,$$

where (W_t) is a d-dimensional Brownian motion and where $\sigma > 0$. Under standard assumptions, the invariant measure of this SDE is the Gibbs measure ν_{σ^2} of density proportional to $e^{-2V(x)/\sigma^2}$ and for small enough σ, this measure concentrates around $\operatorname{argmin}(V)$ see [Dal17] and Chapter 4. Adding a small noise to the gradient descent allows to explore the space and to escape from traps such as local minima or saddle points appearing in non-convex optimization problems [Laz92, DPG$^+$14]. Such methods have been recently brought up to light again with Stochastic Gradient Langevin Dynamics (SGLD) algorithms [WT11, LCCC16], especially for the deep learning and the calibration of large artificial neural networks.

The Langevin-simulated annealing SDE is the Langevin equation where the noise parameter is slowly decreasing to 0, namely

$$dX_t = -\nabla V(X_t)dt + a(t)\sigma dW_t,\ t \geq 0, \tag{3.1.1}$$

where $a : \mathbb{R}^+ \to \mathbb{R}^+$ is non-increasing and converges to 0. The idea is that the "instantaneous" invariant measure $\nu_{a(t)\sigma}$ which is the Gibbs measure of density $\propto \exp(-2V(x)/(a(t)^2\sigma^2))$ converges itself to $\operatorname{argmin}(V)$. Although the additive case i.e. where σ is constant has been extensively studied (see for example [DM17, DM19]), little attention has been paid to the multiplicative case i.e. where $\sigma : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ depends on X_t.

The objective of the present paper is to study the convergence in total variation of the Langevin-Simulated annealing SDE, i.e. (3.1.1) with non-constant σ. Following [PP23, Proposition 2.6], we need to add a correction term in the drift, giving

$$dY_t = -(\sigma\sigma^\top \nabla V)(Y_t)dt + a(t)\sigma(Y_t)dW_t + \left(a^2(t)\left[\sum_{j=1}^{d} \partial_j(\sigma\sigma^\top)(Y_t)_{ij}\right]_{1\leq i\leq d}\right)dt, \tag{3.1.2}$$

$$a(t) = \frac{A}{\sqrt{\log(t+e)}}, \tag{3.1.3}$$

so that $\nu_{a(t)}$ is still the the "instantaneous" invariant measure. We also study the convergence of its Euler-Maruyama scheme $\bar{Y}_t$ with decreasing steps and with noisy gradient estimates coming from stochastic gradient algorithms. We assume in particular the convex uniformity of the potential V outside a compact set (but we do not assume that the potential is convex) and the ellipticity and the boundedness of σ.

We studied this SDE and proved the convergence in L^1-Wasserstein distance of Y and $\bar{Y}$ to $\nu^\star$ which is the limit measure of ν_a as $a \to 0$, in a previous chapter, see Chapter 2. More precisely, we proved that $\mathcal{W}_1(Y_t, \nu^\star)$ is of order $a(t)$ as $t \to \infty$ and that $\mathcal{W}_1(Y_t, \nu_{(a(t))})$ is of order $t^{-\alpha}$ for every $\alpha \in (0, 1)$. For more details, we refer to the introduction of Chapter 2. In particular, for applications to optimization problems arising in Stochastic Optimization and in Machine Learning and for choices of $\sigma : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ used by practitioners, we refer to Section 2.3.

The proof for the total variation distance case relies on the same strategy developed in Chapter 2. We first introduce the process X where the coefficient $(a(t))$ is "by plateaux" i.e. non-increasing and piecewise constant on time intervals $[T_n, T_{n+1}]$. Then we give bounds on $\mathrm{d_{TV}}(X_t, Y_t)$ using a *domino strategy* (2.1.4). However the main difference with the L^1-Wasserstein distance concerns the total variation distance between X and Y in small time as in

general, it is more difficult to give bounds in small time for the total variation distance between two processes with close coefficients. Indeed, considering the functional characterization and comparing it with the L^1-Wasserstein distance, if x and $y \in \mathbb{R}^d$ are close to each other and if $f : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz-continuous, then we can bound $|f(x) - f(y)|$ by $[f]_{\mathrm{Lip}}|x - y|$; however if f is measurable bounded, then we cannot directly bound $|f(x) - f(y)|$ in terms of $|x - y|$. Instead, the common strategy of proof in the literature is to use Malliavin calculus in order to perform an integration by parts and to use bounds on the derivatives of the density. In this context, [PP23] relies on a highly technical Malliavin approach inducing a "regularization from the past" (see [PP23, Theorem 3.7 and Appendix B]).

We give bounds in small time relying on Chapter 8 and we adapt some of the proofs to the non-homogeneous Markovian setting. These bounds rely on estimates of the density of the solutions to SDE's and their derivatives [Fri64]. The strategy of proof is the following: we first reduce to the null drift case using a Girsanov change of measure. Then we introduce an artificial regularization in order to perform a Malliavin-type integration by parts and we use Aronson's bounds on the density and its derivatives; we need to pay attention to the dependency in the parameter a, controlling the ellipticity of the SDE and which converges to 0, of the constants that appear in the Aronson bounds. Moreover, we rely on [DMR18] to give bounds on the total variation between two Gaussian laws.

Contrary to the L^1-Wasserstein distance, we do not prove the convergence as $t \to \infty$ of Y_t and $\bar{Y}_t$ to $\nu^\star$ since in most of the cases, $\nu^\star$ is supported by a finite number of points and then if Y_t has a density then $\mathrm{d_{TV}}(Y_t, \nu^\star) = 2$. Instead, we prove the convergence in total variation of Y_t and $\bar{Y}_t$ to their "instantaneous invariant measure" $\nu_{a(t)}$ which itself converges to $\nu^\star$ (in law, for the L^1-Wasserstein distance etc, see for example [Hwa80, Theorem 2.1] and Lemma 2.4.6 and we give bounds on $\mathrm{d_{TV}}(Y_t, \nu_{a(t)})$ and on $\mathrm{d_{TV}}(\bar{Y}_t, \nu_{a(t)})$ as $t \to \infty$.

The paper is organized as follows. In Section 3.2 we give the setting and assumptions of the problem we consider and state our main results of convergence with convergence rates. This setting is the same as in Chapter 2. In Section 3.3 we establish bounds in small time for $\mathrm{d_{TV}}(X_t, Y_t)$ and for $\mathrm{d_{TV}}(X_t, \bar{Y}_t)$, inspired from Chapter 8. In Section 3.4, we prove the convergence of the plateaux SDE X using exponential contraction properties. Using this convergence, the convergences of $\mathrm{d_{TV}}(Y_t, \nu_{a(t)})$ and $\mathrm{d_{TV}}(\bar{Y}_t, \nu_{a(t)})$ are proved in Section 3.5 and 3.6 respectively. In section 3.7 we compare additive and multiplicative Langevin algorithms on a numerical optimization problem and we give numerical evidence that multiplicative Langevin algorithms improve the optimization procedure.

We use the notations defined from page 1.

3.2 Assumptions and main results

3.2.1 Assumptions

Let us briefly recall the setting adopted in Chapter 2. Let $V : \mathbb{R}^d \to (0, +\infty)$ be a $\mathcal{C}^2$ potential function such that V is coercive and

$$(x \mapsto |x|^2 e^{-2V(x)/A^2}) \in L^1(\mathbb{R}^d) \text{ for some } A > 0. \tag{3.2.1}$$

Then V admits a minimum on $\mathbb{R}^d$. Moreover, let us assume that

$$V^\star := \min_{\mathbb{R}^d} V > 0, \quad \mathrm{argmin}(V) = \{x_1^\star, \dots, x_{m^\star}^\star\}, \quad \forall\, i = 1, \dots, m^\star,\ \nabla^2 V(x_i^\star) > 0, \tag{3.2.2, $\mathcal{H}_{V1}$}$$

i.e. $\min_{\mathbb{R}^d} V$ is attained at a finite number $m^\star$ of points and at each point the Hessian matrix is positive definite. We then define for $a \in (0, A]$ the Gibbs measure ν_a of density :

$$\nu_a(dx) = \mathcal{Z}_a e^{-2(V(x)-V^\star)/a^2} dx, \quad \mathcal{Z}_a = \left(\int_{\mathbb{R}^d} e^{-2(V(x)-V^\star)/a^2} dx \right)^{-1} \tag{3.2.3}$$

Following [Hwa80, Theorem 2.1], the measure ν_a converges weakly to $\nu^\star$ as $a \to 0$, where $\nu^\star$ is the weighted sum of Dirac measures:

$$\nu^\star = \left(\sum_{j=1}^{m^\star} \left(\det \nabla^2 V(x_j^\star) \right)^{-1/2} \right)^{-1} \sum_{i=1}^{m^\star} \left(\det \nabla^2 V(x_i^\star) \right)^{-1/2} \delta_{x_i^\star}. \tag{3.2.4}$$

Following Lemma 2.4.6, ν_a also converges to $\nu^\star$ as $a \to 0$ for the L^1-Wasserstein distance.

We consider the following Langevin SDE in $\mathbb{R}^d$:

$$Y_0^{x_0} = x_0 \in \mathbb{R}^d, \quad dY_t^{x_0} = b_{a(t)}(Y_t^{x_0})dt + a(t)\sigma(Y_t^{x_0})dW_t, \tag{3.2.5}$$

where, for $a \geq 0$, the drift b_a is given by

$$b_a(x) = -(\sigma\sigma^\top \nabla V)(x) + a^2 \left[\sum_{j=1}^{d} \partial_j (\sigma\sigma^\top)_{ij}(x) \right]_{1 \leq i \leq d} =: -(\sigma\sigma^\top \nabla V)(x) + a^2 \Upsilon(x), \tag{3.2.6}$$

where W is a standard $\mathbb{R}^d$-valued Brownian motion defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where $\sigma : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ is $\mathcal{C}^2$ and

$$a(t) = \frac{A}{\sqrt{\log(t+e)}} \tag{3.2.7}$$

where A is defined in (3.2.1) and with $\log(e) = 1$. This equation corresponds to a gradient descent on the potential V with preconditioning σ and multiplicative noise ; the second term in the drift (3.2.6) is a correction term (see [PP23, Proposition 2.6]) which is zero for constant σ.

We make the following assumptions on the potential V:

$$|\nabla V|^2 \leq CV \quad \text{and} \quad \sup_{x \in \mathbb{R}^d} ||\nabla^2 V(x)|| < +\infty, \tag{3.2.8, $\mathcal{H}_{V2}$}$$

which implies in particular that V has at most a quadratic growth. Let us also assume that

$$\sigma \text{ is bounded and Lipschitz-continuous, } \nabla^2\sigma \text{ is bounded, } \nabla(\sigma\sigma^\top)\nabla V \text{ is bounded,} \tag{3.2.9, $\mathcal{H}_\sigma$}$$

and that σ is uniformly elliptic, i.e.

$$\exists \underline{\sigma}_0 > 0, \ \forall x \in \mathbb{R}^d, \ (\sigma\sigma^\top)(x) \geq \underline{\sigma}_0^2 I_d. \tag{3.2.10}$$

Assumptions (3.2.8, $\mathcal{H}_{V2}$) and (3.2.9, $\mathcal{H}_\sigma$) imply that Υ is also bounded and Lipschitz-continuous and that b_a is Lipschitz-continuous uniformly in $a \in [0, A]$. Let the minimal constant $[b]_{\text{Lip}}$ be such that:

$$\forall a \in [0, A], \ b_a \text{ is } [b]_{\text{Lip}}\text{-Lipschitz continuous.} \tag{3.2.11}$$

We make the non-uniform dissipative (or convexity) assumption outside of a compact set: there exists $\alpha_0 > 0$ and $R_0 > 0$ such that

$$\forall x, y \in \mathbf{B}(0, R_0)^c, \ \left\langle \left(\sigma\sigma^\top \nabla V \right)(x) - \left(\sigma\sigma^\top \nabla V \right)(y), \ x - y \right\rangle \geq \alpha_0 |x - y|^2. \tag{3.2.12, $\mathcal{H}_{cf}$}$$

Taking $y \in \mathbf{B}(0,R_0)^c$ fixed, letting $|x| \to \infty$ and using the boundedness of σ, (3.2.12, $\mathcal{H}_{cf}$) implies that $|\nabla V|$ is coercive. Using (3.2.8, $\mathcal{H}_{V2}$) and the boundedness of σ, there exists $C>0$ (depending on A) such that:

$$\forall a \in [0,A],\ 1+|b_a(x)| \le CV^{1/2}(x).$$

Let $(\gamma_n)_{n\ge1}$ be a non-increasing sequence of varying positive steps. We define $\Gamma_n := \gamma_1 + \cdots + \gamma_n$ and for $t \ge 0$:

$$N(t) := \min\{k \ge 0 :\ \Gamma_{k+1} > t\} = \max\{k \ge 0 :\ \Gamma_k \le t\}. \tag{3.2.13}$$

We make the classical assumptions on the step sequence, namely

$$\gamma_n \downarrow 0, \quad \sum_{n\ge1}\gamma_n = +\infty \quad \text{and} \quad \sum_{n\ge1}\gamma_n^2 < +\infty \tag{3.2.14, $\mathcal{H}_{\gamma1}$}$$

and we also assume that

$$\varpi := \limsup_{n\to\infty} \frac{\gamma_n - \gamma_{n+1}}{\gamma_{n+1}^2} < \infty. \tag{3.2.15, $\mathcal{H}_{\gamma2}$}$$

For example, if $\gamma_n = \gamma_1/n^\eta$ with $\eta \in (1/2,1)$ then $\varpi = 0$; if $\gamma_n = \gamma_1/n$ then $\varpi = \gamma_1$.

In stochastic gradient algorithms, the true gradient is measured with a zero-mean noise ζ, which law only depends on the current position. That is, let us consider a family of random fields $(\zeta_n(x))_{x\in\mathbb{R}^d, n\in\mathbb{N}}$ such that for every $n \in \mathbb{N}$, $(\omega,x) \in \Omega\times\mathbb{R}^d \mapsto \zeta_n(x,\omega)$ is measurable and for all $x \in \mathbb{R}^d$, the law of $\zeta_n(x)$ only depends on x and $(\zeta_n(x))_{n\in\mathbb{N}}$ is an i.i.d. sequence independent of W. We make the following assumptions:

$$\forall x \in \mathbb{R}^d,\ \forall p \ge 1,\ \mathbb{E}[\zeta_1(x)] = 0 \quad \text{and} \quad \mathbb{E}[|\zeta_1(x)|^p] \le C_p V^{p/2}(x). \tag{3.2.16}$$

We then consider the Euler-Maruyama scheme with decreasing steps associated to (Y_t):

$$\bar{Y}_0^{x_0} = x_0, \quad \bar{Y}_{\Gamma_{n+1}}^{x_0} = \bar{Y}_{\Gamma_n} + \gamma_{n+1}\left(b_{a(\Gamma_n)}(\bar{Y}_{\Gamma_n}^{x_0}) + \zeta_{n+1}(\bar{Y}_{\Gamma_n}^{x_0})\right) + a(\Gamma_n)\sigma(\bar{Y}_{\Gamma_n}^{x_0})(W_{\Gamma_{n+1}} - W_{\Gamma_n}), \tag{3.2.17}$$

We extend $\bar{Y}_\cdot^{x_0}$ on $\mathbb{R}^+$ by considering its genuine continuous interpolation:

$$\forall t \in [\Gamma_n, \Gamma_{n+1}),\ \bar{Y}_t^{x_0} = \bar{Y}_{\Gamma_n}^{x_0} + (t-\Gamma_n)\left(b_{a(\Gamma_n)}(\bar{Y}_{\Gamma_n}^{x_0}) + \zeta_{n+1}(\bar{Y}_{\Gamma_n}^{x_0})\right) + a(\Gamma_n)\sigma(\bar{Y}_{\Gamma_n}^{x_0})(W_t - W_{\Gamma_n}). \tag{3.2.18}$$

3.2.2 Main results

Theorem 3.2.1. *(a) Let Y be defined in* (3.2.5)*. Assume* (3.2.2, $\mathcal{H}_{V1}$)*,* (3.2.8, $\mathcal{H}_{V2}$)*,* (3.2.9, $\mathcal{H}_\sigma$)*,* (3.2.10) *and* (3.2.12, $\mathcal{H}_{cf}$)*. Then, for every $\alpha \in (0,1)$, if A is large enough, then for every $x_0 \in \mathbb{R}^d$ and for every $t > 0$:*

$$\mathrm{d_{TV}}\left(Y_t^{x_0}, \nu_{a(t)}\right) \le Ce^{C\sqrt{\log(t)}(1+|x_0|^2)}t^{-\alpha}. \tag{3.2.19}$$

(b) Let $\bar{Y}$ be defined in (3.2.17)*. Assume* (3.2.2, $\mathcal{H}_{V1}$)*,* (3.2.8, $\mathcal{H}_{V2}$)*,* (3.2.9, $\mathcal{H}_\sigma$)*,* (3.2.10) *and* (3.2.12, $\mathcal{H}_{cf}$)*. Assume furthermore that $\sigma \in \mathcal{C}_b^{2r}$. Assume furthermore* (3.2.14, $\mathcal{H}_{\gamma1}$) *and* (3.2.15, $\mathcal{H}_{\gamma2}$)*, that V is $\mathcal{C}^3$ with $\|\nabla^3 V\| \le CV^{1/2}$ and that σ is $\mathcal{C}^3$ with $\|\nabla^3(\sigma\sigma^\top)\| \le CV^{1/2}$. Then, for every $\alpha \in (0,1)$, if A is large enough, then for every $x_0 \in \mathbb{R}^d$ and for every $t > 0$:*

$$\mathrm{d_{TV}}\left(\bar{Y}_t^{x_0}, \nu_{a(t)}\right) \le C\left(\log^{1/2}(t)\max\left[V^2(x_0), 1+|x_0|\right]t^{-\alpha} + e^{C\sqrt{\log(t)}(1+|x_0|^2)}t^{C/A^2}\gamma_{N(Ct)}^{r/(2r+1)}\right). \tag{3.2.20}$$

Remark 3.2.2. Depending on the step sequence (γ_n), we can compare the two terms arising in the right-hand side of (3.2.20). For example, if $\gamma_n = \gamma_1 n^{-\eta}$ for some $\eta \in (1/2, 1]$, then

- If $\eta = 1$, then $\gamma_{N(Ct)} \asymp e^{-Ct}$ and the first term is the dominating term.
- If $\eta \in (1/2, 1)$ then $\gamma_{N(Ct)} \asymp (Ct)^{-\eta/(1-\eta)}$.

3.2.3 Extensions and interpolations of the processes

Let us define the following processes that will be used as auxiliary tools in the proofs.

• We define (X_t) as the solution the following SDE where the coefficients piecewisely depend on the time; X is then said to be "by plateaux":

$$X_0^{x_0} = x_0, \quad dX_t^{x_0} = b_{a_{k+1}}(X_t^{x_0})dt + a_{k+1}\sigma(X_t^{x_0})dW_t, \quad t \in [T_k, T_{k+1}], \tag{3.2.21}$$

where b_a is defined in (3.2.6) and the time schedule (T_n) is defined by

$$T_n := C_{(T)} n^{1+\beta}, \tag{3.2.22}$$

where $C_{(T)} > 0$, $\beta > 0$ and $a_n := a(T_n)$. More generally, we define $(X_t^{x,n})$ as the solution of

$$X_0^{x,n} = x, \quad dX_t^{x,n} = b_{a_{k+1}}(X_t^{x,n})dt + a_{k+1}\sigma(X_t^{x,n})dW_t, \quad t \in [T_k - T_n, T_{k+1} - T_n], \ k \geq n, \tag{3.2.23}$$

i.e. $(X_t^{x,n})$ has the conditional law of $(X_{T_n+t})_{t\geq 0}$ given $X_{T_n} = x$. We have $X_t^x = X_t^{x,0}$. The Markov transition kernel associated to $X^{\cdot,n}$ denoted $P_t^{X,n}$ reads on Borel functions $f : \mathbb{R}^d \to \mathbb{R}^+$, $P_t^{X,n} f(x) = \mathbb{E}[f(X_t^{x,n})]$.

• Considering now the original SDE (3.2.5), we also define for every $x \in \mathbb{R}^d$ and every fixed $u \geq 0$:

$$Y_{0,u}^x = x, \quad dY_{t,u}^x = b_{a(t+u)}(Y_{t,u}^x)dt + a(t+u)\sigma(Y_{t,u}^x)dW_t, \tag{3.2.24}$$

so that $Y^x = Y_{\cdot,0}^x$. We define the Markov transition kernel associated to Y between the times t and $t+u$ by $P_{t,u}^Y$ such that for all Borel functions $f : \mathbb{R}^d \to \mathbb{R}^+$, $P_{t,u}^Y f(x) = \mathbb{E}[f(Y_{t,u}^x)]$.

• Considering finally (3.2.17) and (3.2.18), we define for every $n \geq 0$, $(\bar{Y}_{t,\Gamma_n}^x)_{t\geq 0}$, first at times $\Gamma_k - \Gamma_n$, $k \geq n$, by

$$\begin{aligned} \bar{Y}_{0,\Gamma_n}^x = x, \quad \bar{Y}_{\Gamma_{k+1}-\Gamma_n,\Gamma_n}^x &= \bar{Y}_{\Gamma_k-\Gamma_n,\Gamma_n}^x + \gamma_{k+1}\left(b_{a(\Gamma_k)}(\bar{Y}_{\Gamma_k-\Gamma_n,\Gamma_n}^x) + \zeta_{k+1}(\bar{Y}_{\Gamma_k-\Gamma_n,\Gamma_n}^x)\right) \\ &\quad + a(\Gamma_k)\sigma(\bar{Y}_{\Gamma_k-\Gamma_n,\Gamma_n}^x)(W_{\Gamma_{k+1}} - W_{\Gamma_k}), \end{aligned} \tag{3.2.25}$$

then at every time t by the genuine interpolation on the intervals $([\Gamma_k - \Gamma_n, \Gamma_{k+1} - \Gamma_n))_{k\geq n}$ as before. In particular $\bar{Y}^x = \bar{Y}_{\cdot,0}^x$. Still more generally, we define $\bar{Y}_{t,u}^x$ where $u \in (\Gamma_n, \Gamma_{n+1})$ as

$$\bar{Y}_{0,u}^x = x, \quad \bar{Y}_{t,u}^x = \begin{cases} x + t(b_a(x) + \zeta_{n+1}(x)) + a^2(u)\sigma(x)(W_t - W_{\Gamma_u}) & \text{if } t \in [u, \Gamma_{n+1}] \\ = \bar{Y}_{t-(\Gamma_{n+1}-u),\Gamma_{n+1}}^{\bar{Y}_{\Gamma_{n+1}-u,u}^x} & \text{if } t > \Gamma_{n+1}. \end{cases}$$

For n, $k \geq 0$, for $u \in [\Gamma_k, \Gamma_{k+1})$ and $\gamma \in [0, \Gamma_{k+1} - u]$, let $P_{\gamma,u}^{\bar{Y}}$ be the Markov transition kernel associated to $\bar{Y}_{\cdot,u}$ between the times 0 and γ i.e. for all Borel functions $f : \mathbb{R}^d \to \mathbb{R}^+$, $P_{\gamma,u}^{\bar{Y}} f(x) = \mathbb{E}[f(\bar{Y}_{\gamma,u}^x)]$.

3.3 Bounds in total variation for small t

In this section we give bounds for the total variation distance between the processes X, Y and $\bar{Y}$. Although such bounds are straightforward for L^p-distances, they are more difficult to establish for $\mathrm{d_{TV}}$. To this end we adopt a strategy similar to Chapter 8.

For $x \in \mathbb{R}^d$ and for $a \in \mathbb{R}^+$ we define the "cut" drift $\tilde{b}_a^x : \mathbb{R}^d \to \mathbb{R}^d$ which is the drift b_a which is null outside a compact set centred on x. More precisely, we choose $R > 0$ and we consider a $\mathcal{C}^\infty$ decreasing function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\psi = 1$ on $[0, R^2]$ and $\psi = 0$ on $[(R+1)^2, \infty)$ and we define $\tilde{b}_a^x(y) := b_a(y)\psi(|y-x|^2)$, so that $|\tilde{b}_a^x|$ is bounded by $C(1+|x|)$ since b_a is Lipschitz-continuous.

For $\sigma : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$, we denote the martingale:

$$M(\sigma)_0^x = x, \quad dM(\sigma)_t^x = \sigma(M(\sigma)_t^x)dW_t \tag{3.3.1}$$

with its associated one-step Euler-Maruyama scheme:

$$\bar{M}(\sigma)_t^x = x + \sigma(x)W_t. \tag{3.3.2}$$

Lemma 3.3.1. *Let Z be solution of the following SDE:*

$$dZ_t^x = u(t)\sigma_{_Z}(Z_t^x)dW_t,$$

where $u : \mathbb{R}^+ \to (0, \infty)$ is $\mathcal{C}^1$ and bounded. Then $(Z_t) \sim (M(\sigma)_{F^{(-1)}(t)})$, where $F : \mathbb{R}^+ \to \mathbb{R}^+$ is solution of the differential equation

$$F(0) = 0, \quad F'(t) = \frac{1}{u^2(F(t))}$$

and where $F^{(-1)}$ denotes the (continuous) inverse function of F.

Proof. First, F is well defined and is strictly increasing with $F(t) \to \infty$ as $t \to \infty$ since u is bounded, so that $F^{(-1)} : \mathbb{R}^+ \to \mathbb{R}^+$ is well defined as well. We have

$$d\left(Z_{F(t)}^x\right) = u(F(t))\sigma_{_Z}(Z_{F(t)}^x)d\left(W_{F(t)}\right) = F'(t)^{1/2}u(F(t))\sigma_{_Z}(Z_{F(t)}^x)d\widetilde{W}_t = \sigma_{_Z}(Z_{F(t)}^x)d\widetilde{W}_t.$$

where $\widetilde{W}$ is the Brownian motion defined by $\widetilde{W}_t = \int_0^t (F'(s))^{-1/2} dW_{F(s)}$. □

3.3.1 Total variation bound in small time for the Euler-Maruyama scheme

Proposition 3.3.2. *Assume that $\sigma \in \mathcal{C}_b^{2r}$. There exists $C > 0$ such that for every n, $k \geq 0$, for every $u \in [\Gamma_k, \Gamma_{k+1})$ and every $t > 0$ such that $u \in [T_n, T_{n+1}]$, $t \leq \Gamma_{k+1} - u$ and $u+t \in [T_n, T_{n+1}]$,*

$$\mathrm{d_{TV}}(X_t^{x,n}, \bar{Y}_{t,u}^x) \leq Ce^{Ca_{n+1}^{-1}(1+|x|^2)}t^{r/(2r+1)} + Ca_n^{-2}(a_n - a_{n+1}). \tag{3.3.3}$$

Proof. We apply a strategy of proof similar to Theorem 8.2.2. However we need to pay attention to the dependency of the constants in the bounds in (a_n). Let us write

$$\begin{aligned}\mathrm{d_{TV}}(X_t^{x,n}, \bar{Y}_{t,u}^x) \leq\ & \mathrm{d_{TV}}(X_t^{x,n}, \widetilde{X}_t^{x,n}) + \mathrm{d_{TV}}(\widetilde{X}_t^{x,n}, Z_t^{x,n}) + \mathrm{d_{TV}}(Z_t^{x,n}, \bar{Z}_t^{x,n}) \\ & + \mathrm{d_{TV}}(\bar{Z}_t^{x,n}, \bar{X}_t^{x,n}) + \mathrm{d_{TV}}(\bar{X}_t^{x,n}, \bar{Y}_{t,u}^x),\end{aligned} \tag{3.3.4}$$

where

$$\widetilde{X}_0^{x,n} = x, \quad d\widetilde{X}_t^{x,n} = \tilde{b}_{a_{n+1}}^x(\widetilde{X}_t^{x,n})dt + a_{n+1}\sigma(\widetilde{X}_t^{x,n})dW_t,$$

$$Z_0^{x,n} = x, \quad dZ_t^{x,n} = a_{n+1}\sigma(Z_t^{x,n})dW_t,$$
$$\bar{Z}_0^{x,n} = x, \quad \bar{Z}_t^{x,n} = x + a_{n+1}\sigma(x)W_t.$$

• Using Lemma 8.3.3, we have

$$\mathrm{d_{TV}}(X_t^{x,n}, \widetilde{X}_t^{x,n}) \leq C(1+|x|^2)t,$$

where the constant C does not depend on n.

• We use [QZ04, Theorem 2.4] and we rework the bound from Lemma 8.3.6 to make explicit the dependency in a_n. Reworking Lemma 8.3.5, we have for $q \geq 1$:

$$\mathbb{E}\left[\sup_{s\in[0,t]} |U_s^{x,n}|^{2q}\right] \leq Ce^{Cqa_{n+1}^{-1}(1+|x|^2)}, \tag{3.3.5}$$
$$U_0^{x,n} = 1, \quad dU_s^{x,n} = a_{n+1}^{-1}U_s^{x,n}\left\langle \sigma^{-1}(Z_s^{x,n})\tilde{b}_{a_{n+1}}^x(Z_s^{x,n}), dW_s\right\rangle.$$

Moreover, following Lemma 3.3.1 we have $(Z_t^{x,n}) \sim (M(\sigma)^x_{F^{(-1)}(t)})$ where the process $(M(\sigma)_t)$ does not depend on n and where $F^{(-1)}(t) = a_{n+1}^2 t$. Thus following [Fri64, Chapter 9, Theorem 7] (also see Theorem 8.3.1 for the application to SDE's) and since $\sigma \in \mathcal{C}_b^2$ we have

$$|\nabla_x p_{M(\sigma)}(t,x,y)| \leq \frac{C}{t^{(d+1)/2}} e^{-c|y-x|^2/t} \tag{3.3.6}$$

and then

$$|\nabla_x p_Z(t,x,y)| = |\nabla_x p_{M(\sigma)}(a_{n+1}^2 t, x, y)| \leq \frac{Ca_{n+1}^{-(d+1)}}{t^{(d+1)/2}} e^{-ca_{n+1}^{-2}|y-x|^2/t} \leq \frac{Ca_{n+1}^{-(d+1)}}{t^{(d+1)/2}} e^{-c|y-x|^2/t}.$$

Then using Lemma 8.3.6 with the adapted bound on $U_s^{x,n}$ (3.3.5) along with [QZ04, Theorem 2.4], we obtain

$$\mathrm{d_{TV}}(\widetilde{X}_t^{x,n}, Z_t^{x,n}) \leq Ce^{Ca_{n+1}^{-1}(1+|x|^2)}t^{1/2}. \tag{3.3.7}$$

The same way, we obtain

$$\mathrm{d_{TV}}(\bar{Z}_t^{x,n}, \bar{X}_t^{x,n}) \leq Ce^{Ca_{n+1}^{-1}(1+|x|^2)}t^{1/2}.$$

• Following Lemma 3.3.1, we have $(Z_t^{x,n}) \sim (M(\sigma)^x_{a_{n+1}^2 t})$ and $(\bar{Z}_t^{x,n}) \sim (\bar{M}(\sigma)^x_{a_{n+1}^2 t})$, where both processes $M(\sigma)$ and $\bar{M}(\sigma)$ do not depend on n. We then use Theorem 8.2.2 to get

$$\mathrm{d_{TV}}(M(\sigma)_t^x, \bar{M}(\sigma)_t^x) \leq Ce^{C|x|^2}t^{r/(2r+1)}$$

which implies

$$\mathrm{d_{TV}}(Z_t^{x,n}, \bar{Z}_t^{x,n}) \leq Ce^{C|x|^2}a_{n+1}^{2r/(2r+1)}t^{r/(2r+1)} \leq Ce^{C|x|^2}t^{r/(2r+1)}.$$

• Let us now investigate $\mathrm{d_{TV}}(\bar{X}_t^{x,n}, \bar{Y}_{t,u}^x)$. Conditionally to $\zeta(x)$, both random vectors are Gaussian vectors with

$$\bar{X}_t^{x,n} \sim \mathcal{N}\left(x + tb_{a_{n+1}}(x), a_{n+1}^2 t\sigma\sigma^\top(x)\right), \quad \bar{Y}_{t,u}^x \sim \mathcal{N}\left(x + tb_{a(u)}(x) + t\zeta(x), a^2(u)t\sigma\sigma^\top(x)\right).$$

Then, conditionally to $\zeta(x)$ we have

$\mathrm{d_{TV}}(\bar{X}_t^{x,n}, \bar{Y}_{t,u}^x)$

$$\leq \mathrm{d_{TV}}\left(\mathcal{N}\left(x+tb_{a_{n+1}}(x), a_{n+1}^2 t\sigma\sigma^\top(x)\right), \mathcal{N}\left(x+tb_{a(u)}(x)+t\zeta(x), a_{n+1}^2 t\sigma\sigma^\top(x)\right)\right)$$
$$+ \mathrm{d_{TV}}\left(\mathcal{N}\left(x+tb_{a(u)}(x)+t\zeta(x), a_{n+1}^2 t\sigma\sigma^\top(x)\right), \mathcal{N}\left(x+tb_{a(u)}(x)+t\zeta(x), a^2(u) t\sigma\sigma^\top(x)\right)\right)$$
$$=: D_1 + D_2.$$

We then refer to [DMR18] which gives bounds on the total variation between two Gaussian laws, first in the case $d > 1$. Using [DMR18, Theorem 1.1] with $\lambda_1 = \cdots = \lambda_d = (a(u)^2 - a_{n+1}^2)/a_{n+1}^2$, we have

$$D_2 \leq C\left(\frac{a^2(u) - a_{n+1}^2}{a_{n+1}^2}\right) \leq C a_n^{-1}(a_n - a_{n+1}).$$

Using [DMR18, Theorem 1.2], since the ρ_i's are bounded independently of n and since for every $y \in \mathbb{R}^d$, $y^\top \sigma\sigma^\top(x) y \geq \varrho_0^2 |y|^2$, we have

$$D_1 \leq C\sqrt{t} a_{n+1}^{-1}(1 + |\zeta(x)|^{1/2}).$$

Now, integrating over the law of $\zeta(x)$ and using that $\mathbb{E}|\zeta(x)| \leq CV(x)$, we obtain

$$\mathrm{d_{TV}}(\bar{X}_t^{x,n}, \bar{Y}_{t,u}^x) \leq C a_n^{-1}(a_n - a_{n+1}) + C\sqrt{t}(1 + V^{1/2}(x)).$$

In the case $d = 1$, we use [DMR18, Theorem 1.3] and obtain the same bounds.

- **Conclusion:** Considering (3.3.4), we get

$$\begin{aligned}\mathrm{d_{TV}}(X_t^{x,n}, \bar{Y}_{t,u}^x) &\leq C(1+|x|^2)t + Ce^{Ca_{n+1}^{-1}(1+|x|^2)} t^{1/2} + Ce^{C|x|^2} t^{r/(2r+1)} \\ &\quad + C a_n^{-1}(a_n - a_{n+1}) + C\sqrt{t}(1 + V^{1/2}(x)) \\ &\leq Ce^{Ca_{n+1}^{-1}(1+|x|^2)} t^{r/(2r+1)} + C a_n^{-1}(a_n - a_{n+1}).\end{aligned}$$

□

3.3.2 Total variation bound in small time for the continuous SDE

Proposition 3.3.3. *Assume that $\sigma \in \mathcal{C}_b^{2r}$ and let $\bar{\gamma} > 0$. There exists $C > 0$ such that for all $\varepsilon > 0$, $n \geq 0$, u, $t \geq 0$ such that $u \in [T_n, T_{n+1}]$, $u + t \in [T_n, T_{n+1}]$ and $t \leq \bar{\gamma}$,*

$$\mathrm{d_{TV}}(X_t^{x,n}, Y_{t,u}^x) \leq Ce^{Ca_{n+1}^{-1}(1+|x|^2)} t^{1/2} + C a_{n+1}^{-(d+r)}(a(u) - a_{n+1})^{2r/(2r+1)}. \tag{3.3.8}$$

Proof. We have

$$\begin{aligned}\mathrm{d_{TV}}(X_t^{x,n}, Y_{t,u}^x) &\leq \mathrm{d_{TV}}(X_t^{x,n}, \widetilde{X}_t^{x,n}) + \mathrm{d_{TV}}(\widetilde{X}_t^{x,n}, Z_t^{x,n}) + \mathrm{d_{TV}}(Z_t^{x,n}, \widetilde{Z}_{t,u}^x) \\ &\quad + \mathrm{d_{TV}}(\widetilde{Z}_{t,u}^x, \widetilde{Y}_{t,u}^x) + \mathrm{d_{TV}}(\widetilde{Y}_{t,u}^x, Y_{t,u}^x)\end{aligned} \tag{3.3.9}$$

where

$$\begin{aligned}d\widetilde{X}_t^{x,n} &= \tilde{b}_{a_{n+1}}^x(\widetilde{X}_t^{x,n})dt + a_{n+1}\sigma(\widetilde{X}_t^{x,n})dW_t, \\ dZ_t^{x,n} &= a_{n+1}\sigma(Z_t^{x,n})dW_t, \\ d\widetilde{Z}_{t,u}^x &= a(u+t)\sigma(\widetilde{Z}_{t,u}^x)dW_t, \\ d\widetilde{Y}_{t,u}^x &= \tilde{b}_{a(u+t)}^x(\widetilde{Y}_{t,u}^x)dt + a(u+t)\sigma(\widetilde{Y}_{t,u}^x)dW_t.\end{aligned}$$

Using Lemma 8.3.3, we have

$$\mathrm{d_{TV}}(X_t^{x,n}, \widetilde{X}_t^{x,n}) + \mathrm{d_{TV}}(\widetilde{Y}_{t,u}^x, Y_{t,u}^x) \leq C(1+|x|^2)t.$$

Using (3.3.7) again, we have

$$\mathrm{d_{TV}}(\widetilde{X}_t^{x,n}, Z_t^{x,n}) \leq Ce^{Ca_{n+1}^{-1}(1+|x|^2)}t^{1/2}.$$

Moreover, using [QZ04, Theorem 2.4] (with an immediate adaptation to the non-homogeneous case) and establishing the same bounds as in Lemma 8.3.6, we also have

$$\mathrm{d_{TV}}(\widetilde{Z}_{t,u}^{x}, \widetilde{Y}_{t,u}^{x}) \leq Ce^{Ca_{n+1}^{-1}(1+|x|^2)}t^{1/2}. \tag{3.3.10}$$

We now turn to $\mathrm{d_{TV}}(Z_t^{x,n}, \widetilde{Z}_{t,u}^{x})$. Using Lemma 3.3.1 as in (3.3.6) we have

$$|\nabla_y^{2r} p_Z(t,x,y)| = |\nabla_y^{2r} p_{M(\sigma)}(a_{n+1}^2 t, x, y)| \leq C\frac{a_{n+1}^{-(d+r)}}{t^{(d+r)/2}}e^{-c|x-y|^2/t}.$$

To bound $p_{\widetilde{Z}}$ we use the change of time F satisfying $F'(t) = a^{-2}(u + F(t))$ so that

$$a_n^{-2}t \leq F(t) \leq a_{n+1}^{-2}t \quad \text{and} \quad a_{n+1}^2 t \leq F^{(-1)}(t) \leq a_n^2 t$$

and then

$$|\nabla_y^{2r} p_{\widetilde{Z}}(t,x,y)| = |\nabla_y^{2r} p_{M(\sigma)}(F^{(-1)}(t), x, y)| \leq C\frac{a_{n+1}^{-(d+r)}}{t^{(d+r)/2}}e^{-c|x-y|^2/t}.$$

We prove as in Lemma 2.6.2 that

$$\|Z_t^{x,n} - \widetilde{Z}_{t,u}^{x}\|_1 \leq C(a(u) - a_{n+1})t^{1/2}$$

and then using Theorem 8.2.7 we get

$$\mathrm{d_{TV}}(Z_t^{x,n}, \widetilde{Z}_{t,u}^{x}) \leq Ca_{n+1}^{-(d+r)}(a(u) - a_{n+1})^{2r/(2r+1)}.$$

- **Conclusion:** considering (3.3.9), we get

$$\mathrm{d_{TV}}(X_t^{x,n}, Y_{t,u}^{x}) \leq C(1+|x|^2)t + Ce^{Ca_{n+1}^{-1}(1+|x|^2)}t^{1/2} + Ca_{n+1}^{-(d+r)}(a(u) - a_{n+1})^{2r/(2r+1)}.$$

□

Remark 3.3.4. As in Theorem 8.2.5, we could improve the dependency in $|x|$ in (3.3.3) and (3.3.8), at the expanse of further assumptions on V. However it would require to track the dependency in the ellipticity (in a_n) in the bounds proved in [MPZ21], which rely on Malliavin calculus. We believe that it would considerably increase the length and the technicality of the present article, while bringing no significant improvement to our final results.

3.4 Convergence of the plateau SDE X_t in total variation

In this section, we prove the convergence of the plateau SDE (X_t) defined in (3.2.21).

3.4.1 Exponential contraction in total variation

We first show that the property of exponential contraction that holds for the L^1-Wasserstein distance under the setting described in Section 3.2.1 (see Theorem 2.4.2) also holds for the total variation distance.

Theorem 3.4.1. *Let X be the solution to*

$$X_0^x = x, \quad dX_t^x = b_a(X_t^x)dt + a\sigma(X_t^x)dW_t, \tag{3.4.1}$$

with $a \in (0, A]$ and where b_a is defined in (3.2.6)*, so that ν_a defined in* (3.2.3) *is the unique invariant distribution of X ([PP23, Proposition 2.6]). Let $t_0 \in (0, 1]$. Under the assumption* (3.2.12, $\mathcal{H}_{cf}$)*,*

(a) For every x, $y \in \mathbb{R}^d$ and for every $t \geq t_0$ we have

$$\mathrm{d_{TV}}(X_t^x, X_t^y) \leq Ca^{-1}e^{C_1/a^2}e^{-\rho_a t}|x - y|, \quad \rho_a := e^{-C_2/a^2}. \tag{3.4.2}$$

(b) For every $x \in \mathbb{R}^d$ and for every $t \geq t_0$ we have

$$\mathrm{d_{TV}}(X_t^x, \nu_a) \leq Ca^{-1}e^{C_1/a^2}e^{-\rho_a t}\nu_a(|x - \cdot|). \tag{3.4.3}$$

Proof. (a) Following Theorem 2.4.2, we have

$$\forall x, y \in \mathbb{R}^d, \ \mathcal{W}_1\left(X_t^x, X_t^y\right) \leq Ce^{C_1/a^2}|x - y|e^{-\rho_a t}.$$

Let $t \geq t_0$ and let $f : \mathbb{R}^d \to \mathbb{R}$ a Borel bounded function. Then

$$\mathbb{E}[f(X_t^x)] - \mathbb{E}[f(X_t^y)] = \mathbb{E}[P_{t_0}^X f(X_{t-t_0}^x)] - \mathbb{E}[P_{t_0}^X f(X_{t-t_0}^x)],$$

where P^X denotes the kernel associated to X. But using [PP23, Proposition 3.1] we have for every z_1 and $z_2 \in \mathbb{R}^d$,

$$\begin{aligned} P_{t_0}^X f(z_2) - P_{t_0}^X f(z_1) &= \langle \nabla P_{t_0}^X f(\xi), z_2 - z_1 \rangle \\ &= \frac{1}{t_0}\mathbb{E}\left[f(X_t^\xi) \left\langle \int_0^{t_0} (a^{-1}\sigma^{-1}(X_s^\xi)Y_s^\xi)^\top dW_s, z_2 - z_1 \right\rangle \right], \end{aligned}$$

where $\xi \in (z_1, z_2)$ and $(Y_s^\xi)_{s\geq 0}$ denotes the tangent process of (X_s^ξ), i.e.

$$Y_0^\xi = I_d, \quad dY_s^\xi = \nabla b_a(X_s^\xi)Y_s^\xi ds + a\nabla\sigma(X_s^\xi)Y_s^\xi \otimes dW_s. \tag{3.4.4}$$

Since $\nabla\sigma$ and ∇b_a are bounded (uniformly in a), we have

$$\sup_{\xi\in\mathbb{R}^d, s\in[0,t_0]} \mathbb{E}\|Y_s^\xi\|_2^2 < +\infty,$$

where the bound does not depend on a. So that

$$P_{t_0}^X f(z_2) - P_{t_0}^X f(z_1) \leq C||f||_\infty |z_2 - z_2|a^{-1} \sup_{\xi\in\mathbb{R}^d, s\in[0,t_0]} \mathbb{E}\|Y_s^\xi\|_2,$$

and then $[P_{t_0}^X f]_{\mathrm{Lip}} \leq Ca^{-1}\|f\|_\infty$. Then we obtain

$$\mathrm{d_{TV}}(X_t^x, X_t^y) \leq Ca^{-1}\mathcal{W}_1(X_{t-t_0}^x, X_{t-t_0}^y) \leq Ca^{-1}e^{C_1/a^2}e^{-\rho_a t}|x - y|.$$

(b) As ν_a is the invariant distribution of the diffusion (3.4.1) we have

$$\begin{aligned} \mathrm{d_{TV}}\left(X_t^x, \nu_a\right) &\leq \int_{\mathbb{R}^d} \mathrm{d_{TV}}\left(X_t^x, X_t^y\right)\nu_a(dy) \leq Ce^{C_1/a^2}e^{-\rho_a t}\int_{\mathbb{R}^d} |x - y|\nu_a(dy) \\ &\leq Ce^{C_1/a^2}e^{-\rho_a t}\nu_a(|x - \cdot|). \end{aligned}$$

□

3.4.2 Convergence of the plateau SDE

Let (T_n) be the time schedule defined in (3.2.22) and by a slight abuse of notation we define

$$a_n := a(T_n) = \frac{A}{\sqrt{\log(T_n + e)}} \quad \text{and} \quad \rho_n := \rho_{a_n} = e^{-C_2/a_n^2}. \tag{3.4.5}$$

We recal Lemma 2.4.3:

$$0 \leq a_n - a_{n+1} \asymp (n \log^{3/2}(n))^{-1}. \tag{3.4.6}$$

Proposition 3.4.2. *Let ν_a, $a \in (0, A]$, be the Gibbs measure defined in* (3.2.3)*. Assume that V is coercive, that $(x \mapsto |x|^2 e^{-2V(x)/A^2}) \in L^1(\mathbb{R}^d)$ and* (3.2.2, $\mathcal{H}_{V1}$)*. Then for $n \geq 2$,*

$$\mathrm{d_{TV}}(\nu_{a_n}, \nu_{a_{n+1}}) \leq \frac{C}{n \log(n)}. \tag{3.4.7}$$

Moreover, for every s, $t \in [a_{n+1}, a_n]$, we have

$$\mathrm{d_{TV}}(\nu_s, \nu_t) \leq \frac{C}{n \log(n)}. \tag{3.4.8}$$

The proof is given in the Appendix 3.8.1.

We now prove the convergence of the SDE "by plateaux" for the total variation distance.

Theorem 3.4.3. *Let X be the process defined in* (3.2.21) *and* (3.2.23)*. Let t_0 be defined as in Theorem 3.4.1. If $A > \max(\sqrt{(1+\beta^{-1})C_2}, \sqrt{(1+\beta)C_1})$ where C_1 and C_2 are defined in* (3.4.2)*, then for all $x_0 \in \mathbb{R}^d$ and for all $C'_{(T)} < C_{(T)}$, for all large enough $n \geq n(C'_{(T)})$, on the time schedule (T_n) we have*

$$\mathrm{d_{TV}}(X^{x_0}_{T_n}, \nu_{a_n}) \leq C a_n^{-1} n^{-1+(\beta+1)C_1/A^2} \exp\left(-(C'_{(T)})^{1-C_2/A^2}(\beta+1)n^{\beta-(\beta+1)C_2/A^2}\right)(1+|x_0|) \tag{3.4.9}$$

and for every $t \in \mathbb{R}^+ \setminus (\bigcup_{n\geq 1}[T_n, T_n + t_0])$ we have

$$\mathrm{d_{TV}}(X^{x_0}_t, \nu_{a(t)}) \leq \frac{C(1+|x_0|)}{t^{(1+\beta)^{-1}-C_1/A^2}\log(t+e)}. \tag{3.4.10}$$

Proof. For fixed $x \in \mathbb{R}^d$ and using Theorem 3.4.1, we have for every bounded Borel function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}[f(X^{x,n}_{T_{n+1}-T_n})] - \mathbb{E}[f(Z_{a_{n+1}})] \leq C a_{n+1}^{-1} e^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \|f\|_\infty \mathbb{E}|x - Z_{a_{n+1}}|,$$

where $Z_{a_{n+1}} \sim \nu_{a_{n+1}}$. Now integrating x with respect to the law of $X^{x_0}_{T_n}$ yields

$$\begin{aligned}
\mathrm{d_{TV}}(X^{x_0}_{T_{n+1}}, \nu_{a_{n+1}}) &\leq C a_{n+1}^{-1} e^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \left(\mathcal{W}_1(X^{x_0}_{T_n}, \nu_{a_n}) + \mathcal{W}_1(\nu_{a_{n+1}}, \nu_{a_n})\right) \\
&\leq C \frac{a_{n+1}^{-1}\mu_{n+1}}{n \log^{3/2}(n)}(1+|x_0|), \\
\mu_n &:= e^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)}
\end{aligned} \tag{3.4.11}$$

where we used Theorem 2.5.1 and Proposition 2.4.4. We use the bound on μ_n given by (2.5.5). Then to bound $\mathrm{d_{TV}}(X^{x_0}_t, \nu_{a_{n+1}})$ for any $t \in (T_n + t_0, T_{n+1})$, we apply Theorem (3.4.1) on the time interval $[T_n, t]$ which length is not smaller than t_0 and we conclude as in the proof of Theorem 2.5.1. □

Remark 3.4.4. The condition that t does not belong in any interval $[T_n, T_n + t_0]$ is a technical condition which is specific to our strategy of proof. However this condition is not a problem for the convergence of Y_t and $\bar{Y}_t$ since for these two processes, the time schedule (T_n) is only a tool for the proof.

3.5 Convergence of Y_t in total variation

We now consider (Y_t) as defined in (3.2.5) with extended definition (3.2.24).

3.5.1 Preliminary lemmas

Lemma 3.5.1. *Let $\lambda \in \mathbb{R}^+$. There exists $C > 0$ such that for every $n \geq 0$, $u \geq 0$ and every $x \in \mathbb{R}^d$:*

$$\sup_{t\geq 0} \mathbb{E}\left[e^{\lambda|X_t^{x,n}|^2}\right] \leq Ce^{\lambda|x|^2} \quad \text{and} \quad \sup_{t\geq 0} \mathbb{E}\left[e^{\lambda|Y_{t,u}^{x}|^2}\right] \leq Ce^{\lambda|x|^2}. \tag{3.5.1}$$

Sketch of proof. By Itō's Lemma, we have for $k \geq n$ and for $t \in [T_k - T_n, T_{k+1} - T_n)$:

$$\begin{aligned} d\left(e^{\lambda|X_t^{x,n}|^2}\right) &= \lambda e^{\lambda|X_t^{x,n}|^2}\left(2\langle X_t^{x,n}, dX_t^{x,n}\rangle + d\langle X^{x,n}\rangle_t\right) + 2\lambda^2|X_t^{x,n}|^2 e^{\lambda|X_t^{x,n}|^2} d\langle X^{x,n}\rangle_t \\ &= \lambda e^{\lambda|X_t^{x,n}|^2}\Big(- 2\langle \sigma\sigma^\top \nabla V(X_t^{x,n}), X_t^{x,n}\rangle dt + 2a_{n+1}^2\langle X_t^{x,n}, \Upsilon(X_t^{x,n})\rangle dt \\ &\quad + 2a_{n+1}\langle X_t^{x,n}, \sigma(X_t^{x,n})dW_t\rangle + a_{n+1}^2 \operatorname{Tr}(\sigma\sigma^\top(X_t))dt\Big) \\ &\quad + 2\lambda^2 e^{\lambda|X_t^{x,n}|^2} a_{n+1}^2 (X_t^{x,n})^\top \sigma\sigma^\top(X_t^{x,n}) X_t^{x,n} dt \end{aligned}$$

the "dominating" term is $-\langle \sigma\sigma^\top \nabla V(X_t^{x,n}), X_t^{x,n}\rangle dt$ which makes $\mathbb{E}[e^{\lambda|X_t^{x,n}|^2}]$ decrease. Using assumption (3.2.12, $\mathcal{H}_{cf}$), we have for $|X_t^{x,n}|$ large enough,

$$-\langle \sigma\sigma^\top \nabla V(X_t^{x,n}), X_t^{x,n}\rangle \leq -C\underline{\sigma}_0^2\alpha_0|X_t^{x,n}|^2.$$

Moreover, using the facts that Υ and σ are bounded, that $a_n \to 0$, that $|\nabla V| \leq CV^{1/2}$ and that $\sigma\sigma^\top \geq \underline{\sigma}_0^2 I_d$, for large enough $|X_t^{x,n}|$ and large enough n, the coefficient in dt in the last equation is negative. We deal with the cases where $|X_t^{x,n}|$ is not large enough or where n is not large enough the same way as in the proof of Lemma 2.6.1 and Lemma 2.7.1, where more details can be found.

The proof is the same for Y, replacing a_{k+1} by $a(u+t)$. □

Proposition 3.5.2. *Let T, $\bar{\gamma} > 0$. There exists $C > 0$ such that for every Borel bounded function $f : \mathbb{R}^d \to \mathbb{R}$ and every $t \in (0, T]$, for all $n \geq 0$, for all $\gamma < \bar{\gamma}$ such that $u \in [T_n, T_{n+1}]$ and $u + t + \gamma \in [T_n, T_{n+1}]$,*

$$\left|\mathbb{E}\left[P_t^{X,n} f(Y_{\gamma,u}^x)\right] - \mathbb{E}\left[P_t^{X,n} f(X_\gamma^{x,n})\right]\right| \leq Ca_{n+1}^{-2}(a_n - a_{n+1})\|f\|_\infty \gamma t^{-1} V(x). \tag{3.5.2}$$

Proof. We apply Proposition 2.6.4 to $g_t := P_t^{X,n} f$ with $t > 0$. Following [PP23, Proposition 3.2(b)], we have

$$\Phi_{g_t}(x) \leq C\|f\|_\infty a_{n+1}^{-2} t^{-1} \max\left(V^{1/2}(x), \left\|\sup_{\xi\in(X_\gamma^{x,n}, Y_{\gamma,u}^x)} V^{1/2}(\xi)\right\|_2, V^{1/2}(x)\left\|\sup_{\xi\in(x, X_\gamma^{x,n})} V^{1/2}(\xi)\right\|_2\right).$$

We conclude as in the proof of Proposition 2.6.5.

□

3.5.2 Proof of Theorem 3.2.1(a)

More precisely, we prove that for all $\beta > 0$, if

$$A > \max\left(\sqrt{(\beta+1)(2C_1+C_2)}, \sqrt{(1+\beta^{-1})C_2}\right), \tag{3.5.3}$$

then

$$\mathrm{d_{TV}}\left(Y_t^{x_0}, \nu_{a(t)}\right) \leq \frac{Ce^{C\sqrt{\log(t)}(1+|x_0|^2)}}{t^{(1+\beta)^{-1}-(2C_1+C_2)/A^2}}. \tag{3.5.4}$$

Proof. We follow the proof of Theorem 2.2.1(b) in Section 2.7.3 based on a domino strategy with respect to some decreasing step sequence (γ_n), even though Y is not an Euler-Maruyama scheme. In this case, the step sequence (γ_n) is only a tool for the proof. This way we can choose freely the sequence (γ_n) in this section. We use Theorem 3.4.1 in place of Theorem 2.4.2 and Proposition 3.5.2 in place of Proposition 2.7.4. For $f : \mathbb{R}^d \to \mathbb{R}$ bounded measurable and for $x \in \mathbb{R}^d$ we write

$$\begin{aligned}
&\left|\mathbb{E}f(X^{x,n}_{T_{n+1}-T_n}) - \mathbb{E}f(Y^x_{T_{n+1}-T_n,T_n})\right| \leq \left|(P^Y_{\gamma^{\mathrm{init}},T_n} - P^{X,n}_{\gamma^{\mathrm{init}}}) \circ P^{X,n}_{T_{n+1}-\Gamma_{N(T_n)+1}} f(x)\right| \\
&+ \sum_{k=N(T_n)+2}^{N(T_{n+1}-T)} \left|P^Y_{\gamma^{\mathrm{init}},T_n} \circ P^Y_{\gamma_{N(T_n)+2},\Gamma_{N(T_n)+1}} \circ \cdots \circ P^Y_{\gamma_{k-1},\Gamma_{k-2}} \circ (P^Y_{\gamma_k,\Gamma_{k-1}} - P^{X,n}_{\gamma_k}) \circ P^{X,n}_{T_{n+1}-\Gamma_k} f(x)\right| \\
&+ \sum_{k=N(T_{n+1}-T)+1}^{N(T_{n+1})-1} \left|P^Y_{\gamma^{\mathrm{init}},T_n} \circ P^Y_{\gamma_{N(T_n)+2},\Gamma_{N(T_n)+1}} \circ \cdots \circ P^Y_{\gamma_{k-1},\Gamma_{k-2}} \circ (P^Y_{\gamma_k,\Gamma_{k-1}} - P^{X,n}_{\gamma_k}) \circ P^{X,n}_{T_{n+1}-\Gamma_k} f(x)\right| \\
&+ \Big|P^Y_{\gamma^{\mathrm{init}},T_n} \circ P^Y_{\gamma_{N(T_n)+2},\Gamma_{N(T_n)+1}} \circ \cdots \circ P^Y_{\gamma_{N(T_{n+1})-1},\Gamma_{N(T_{n+1})-2}} \\
&\qquad \circ (P^Y_{\gamma^{\mathrm{end}}+\gamma_{N(T_{n+1})},\Gamma_{N(T_{n+1})-1}} - P^{X,n}_{\gamma^{\mathrm{end}}+\gamma_{N(T_{n+1})}}) f(x)\Big| \\
&=: (c^{\mathrm{init}}) + (a) + (b) + (c^{\mathrm{end}}),
\end{aligned}$$

where

$$\gamma^{\mathrm{init}} := \Gamma_{N(T_n)+1} - T_n \leq \gamma_{N(T_n)+1} \quad \text{and} \quad \gamma^{\mathrm{end}} := T_{n+1} - \Gamma_{N(T_{n+1})} \leq \gamma_{N(T_{n+1})+1}.$$

Then we have

$$\begin{aligned}
(a) &\leq Ca_{n+1}^{-3} e^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}T_{n+1}} \|f\|_\infty V(x)(a_n - a_{n+1}) \sum_{k=N(T_n)+2}^{N(T_{n+1}-T)} \gamma_k e^{\rho_{n+1}\Gamma_k} \\
&\leq Ca_{n+1}^{-3} e^{C_1 a_{n+1}^{-2}} \|f\|_\infty (a_n - a_{n+1}) V(x) \rho_{n+1}^{-1}.
\end{aligned}$$

We obtain likewise

$$(c^{\mathrm{init}}) \leq Ca_{n+1}^{-3} e^{-\rho_{n+1}(T_{n+1}-T_n)} \|f\|_\infty (a_n - a_{n+1}) \gamma_{N(T_n)} V(x).$$

Applying Proposition 3.5.2 yields

$$\begin{aligned}
(b) &\leq Ca_{n+1}^{-2}(a_n - a_{n+1}) \|f\|_\infty V(x) \sum_{k=N(T_{n+1}-T)+1}^{N(T_{n+1})-1} \frac{\gamma_k}{T_{n+1} - \Gamma_k} \\
&\leq Ca_{n+1}^{-2}(a_n - a_{n+1}) \|f\|_\infty V(x) \log(1/\gamma_{N(T_{n+1})}).
\end{aligned}$$

Applying Proposition 3.3.3 with $r = 1$ along with Lemma 3.5.1 yields

$$(c^{\mathrm{end}}) \le C\|f\|_\infty \left(e^{Ca_{n+1}^{-1}(1+|x|^2)}\gamma_{N(T_n)}^{1/2} + a_{n+1}^{-(d+1)}(a(T_{n+1}-\gamma_{N(T_{n+1})}) - a_{n+1})^{2/3}\right).$$

But we have

$$\begin{aligned} a(T_{n+1}-\gamma_{N(T_{n+1})}) - a_{n+1} &= a(T_{n+1}-\gamma_{N(T_{n+1})}) - a(T_{n+1}) \le C\frac{da}{dt}(T_{n+1})\cdot\gamma_{N(T_{n+1})} \\ &\le \frac{C\gamma_{N(T_{n+1})}}{T_{n+1}}. \end{aligned}$$

We now choose $\gamma_n = \gamma_1 n^{-2/3}$ so that $\gamma_{N(T_n)} \asymp n^{-2}$ and then

$$(c^{\mathrm{end}}) \le Ce^{Ca_{n+1}^{-1}(1+|x|^2)}n^{-1}.$$

This way we obtain for every $x \in \mathbb{R}^d$:

$$|\mathbb{E}f(X^{x,n}_{T_{n+1}-T_n}) - \mathbb{E}f(Y^x_{T_{n+1}-T_n,T_n})| \le C\|f\|_\infty \underbrace{a_{n+1}^{-3}e^{C_1a_{n+1}^{-2}}(a_n - a_{n+1})V(x)\rho_{n+1}^{-1}}_{=:v_{n+1}} e^{Ca_{n+1}^{-1}(1+|x|^2)}. \tag{3.5.5}$$

We integrate this inequality with respect to the laws of $X^{x_0}_{T_n}$ and $\bar{Y}^{x_0}_{T_n}$ and obtain, temporarily setting $x_n := X^{x_0}_{T_n}$ and $y_n := Y^{x_0}_{T_n}$ and using Lemma 2.6.1 and Lemma 3.5.1,

$$\begin{aligned} \mathrm{d_{TV}}(X^{x_0}_{T_{n+1}}, Y^{x_0}_{T_{n+1}}) &\le \mathrm{d_{TV}}(X^{x_n,n}_{T_{n+1}-T_n}, X^{y_n,n}_{T_{n+1}-T_n}) + \mathrm{d_{TV}}(X^{y_n,n}_{T_{n+1}-T_n}, Y^{\bar{y}_n}_{T_{n+1}-T_n,T_n}) \\ &\le \underbrace{Ca_{n+1}^{-1}e^{C_1a_{n+1}^{-2}}e^{-\rho_{n+1}(T_{n+1}-T_n)}}_{:=\mu'_{n+1}=a_{n+1}^{-1}\mu_{n+1}} \mathrm{d_{TV}}(X^{x_0}_{T_n}, Y^{x_0}_{T_n}) + \underbrace{Cv_{n+1}e^{Ca_{n+1}^{-1}(1+|x_0|^2)}}_{:=w_{n+1}}, \end{aligned}$$

where μ_n is defined in (3.4.11). Iterating this inequality yields

$$\mathrm{d_{TV}}(X^{x_0}_{T_{n+1}}, Y^{x_0}_{T_{n+1}}) \le C(w_{n+1} + \mu'_{n+1}w_n + \cdots + \mu'_{n+1}\cdots\mu'_2w_1) \le Cw_{n+1},$$

where we used, since A satisfies (3.5.3), that $\mu'_n = O(e^{-Cn^\eta})$ for some $\eta > 0$ (see (2.5.5)) and that w_n is bounded as it converges to 0. Moreover using Theorem 3.4.3 we have

$$\mathrm{d_{TV}}(Y^{x_0}_{T_n}, \nu_{a_n}) \le \mathrm{d_{TV}}(X^{x_0}_{T_n}, Y^{x_0}_{T_n}) + \mathrm{d_{TV}}(X^{x_0}_{T_n}, \nu_{a_n}) \le \frac{Ce^{C\sqrt{\log(n)}(1+|x_0|^2)}}{n^{1-(\beta+1)(C_1+C_2)/A^2}}. \tag{3.5.6}$$

Finally, let us bound $\mathrm{d_{TV}}(X^{x_0}_t, Y^{x_0}_t)$ for any $t \in [T_n, T_{n+1}]$. If $t \in [T_n + t_0, T_{n+1}]$ then we can apply Theorem 3.4.1 and we proceed as in the end of Section 2.6.3. If $t \in [T_n, T_n + t_0]$, then we consider another shifted time schedule $\bar{T}_n := C_{(T)}n^{1+\beta} + 2t_0$ such that

$$\bigcup_{i=0}^{\infty}[T_n, T_n + t_0] \cap \bigcup_{i=0}^{\infty}[\bar{T}_n, \bar{T}_n + t_0] = \varnothing.$$

Making use of the new time schedule we obtain as before a bound on $\mathrm{d_{TV}}(Y^{x_0}_t, \nu_{a(t)})$ for every $t \notin \bigcup_{i=0}^{\infty}[\bar{T}_n, \bar{T}_n + t_0]$. Since the time schedules (T_n) and $(\bar{T}_n)$ are only tools for the proof of convergence of Y_t, we then obtain a bound on $\mathrm{d_{TV}}(Y_t, \nu_{a(t)})$ for every $t \in \mathbb{R}^+$. □

3.6 Convergence of the Euler-Maruyama scheme in total variation

We now consider $(\bar{Y}_n)$ as in (3.2.17) with extended definition (3.2.25).

3.6.1 Preliminary lemmas

Lemma 3.6.1. *Let $\lambda \in \mathbb{R}^+$. There exists a constant $C > 0$ such that for every $k \geq 0$, for every $u \in [\Gamma_k, \Gamma_{k+1})$ and for every $x \in \mathbb{R}^d$:*

$$\sup_{n \geq k+1} \mathbb{E}\left[e^{\lambda |Y^x_{\Gamma_n - u, u}|^2}\right] \leq C e^{\lambda |x|^2}. \tag{3.6.1}$$

Proof. The prove is the same as for Lemma 3.5.1. For the adaptation to discrete time, we refer to the proof of Lemma 2.7.1. □

Proposition 3.6.2. *Let $T > 0$. There exists $C > 0$ such that for every Lipschitz continuous function f and every $t \in (0, T]$, for all $n \geq 0$, for all γ such that $\Gamma_k \in [T_n, T_{n+1}]$, $\gamma \leq \gamma_{k+1}$ and $\Gamma_k + t + \gamma \in [T_n, T_{n+1}]$,*

$$\begin{aligned}
&\left|\mathbb{E}\left[P_t f(\bar{Y}^x_{\gamma, \Gamma_k})\right] - \mathbb{E}\left[P_t f(X^{x,n}_\gamma)\right]\right| \\
&\quad \leq C \|f\|_\infty V^2(x) \left(a_{n+1}^{-2} t^{-1} \left(\gamma^2 + (a(\Gamma_k) - a_{n+1})\gamma\right) + a_{n+1}^{-3} t^{-3/2} \left(\gamma^2 + \gamma^{3/2}(a(\Gamma_k) - a_{n+1})\right)\right).
\end{aligned} \tag{3.6.2}$$

Proof. The proof is the same as the proof of Proposition 3.5.2, using Proposition 2.7.3. We also remark that we can directly improve the bound in $(a_n - a_{n+1})$ into $(a(\Gamma_k) - a_{n+1})$. □

3.6.2 Proof of Theorem 3.2.1(b)

More precisely, we prove that for all $\beta > 0$, if $\sigma \in \mathcal{C}_b^{2r}$ and if

$$A > \max\left(\sqrt{(\beta+1)(2C_1 + C_2)}, \sqrt{(1+\beta^{-1})C_2}\right) \tag{3.6.3}$$

and if A is large enough so that

$$n^{(\beta+1)C_1/A^2} \gamma_{N(T_n)}^{r/(2r+1)} \underset{n \to \infty}{\longrightarrow} 0, \tag{3.6.4}$$

then

$$\mathrm{d_{TV}}(\bar{Y}^{x_0}_t, \nu_{a(t)}) \leq C \left(\frac{\log^{1/2}(t) \max\left[V^2(x_0), 1 + |x_0|\right]}{t^{(\beta+1)^{-1} - (2C_1 + C_2)/A^2}} + e^{C\sqrt{\log(t)}(1+|x_0|^2)} t^{C_1/A^2} \gamma_{Ct}^{r/(2r+1)}\right). \tag{3.6.5}$$

Proof. We still follow the proof of Theorem 2.2.1 in Section 2.7.3 based on a domino strategy, using Theorem 3.4.1 in place of Theorem 2.4.2 and Proposition 3.6.2 in place of Proposition 2.7.4. Let $n \geq 0$, for $f : \mathbb{R}^d \to \mathbb{R}$ bounded measurable, we split $|\mathbb{E}f(X^{x,n}_{T_{n+1} - T_n}) - \mathbb{E}f(\bar{Y}^x_{T_{n+1} - T_n, T_n})|$ into four terms (c^{init}), (a), (b), (c^{end}).

Using Theorem 3.4.1, Lemma 2.7.1 and Proposition 3.6.2 we get as in Section 2.7.3:

$$\begin{aligned}
(a) &\leq C a_{n+1}^{-4} e^{C_1 a_{n+1}^{-2}} \|f\|_\infty (a_n - a_{n+1}) V^2(x) \rho_{n+1}^{-1}. \\
(c^{\mathrm{init}}) &\leq C a_{n+1}^{-4} e^{C_1 a_{n+1}^{-2}} e^{-\rho_n (T_{n+1} - T_n)} \|f\|_\infty (a_n - a_{n+1}) \gamma_{N(T_n)+1} V^2(x).
\end{aligned}$$

Using Proposition 3.6.2 and Lemma 2.7.1, we obtain

$$(b) \leq C a_{n+1}^{-3} \left(\gamma_{N(T_{n+1} - T)} + \sqrt{\gamma_{N(T_{n+1} - T)}}(a_n - a_{n+1})\right) \|f\|_\infty V^2(x) \sum_{k = N(T_{n+1} - T) + 1}^{N(T_{n+1}) - 1} \frac{\gamma_k}{(T_{n+1} - \Gamma_k)^{3/2}}$$

$$+ Ca_{n+1}^{-2} \left(\sum_{k=N(T_{n+1}-T)+1}^{N(T_{n+1})-1} \frac{\gamma_{N(T_{n+1}-T)}\gamma_k}{T_{n+1}-\Gamma_k} + \sum_{k=N(T_{n+1}-T)+1}^{N(T_{n+1})-1} \frac{\gamma_k(a(\Gamma_k)-a_{n+1})}{T_{n+1}-\Gamma_k} \right) \|f\|_\infty V^2(x).$$

But we remark that

$$a(\Gamma_k) - a_{n+1} = a(\Gamma_k) - a(T_{n+1}) \leq C\frac{da}{dt}(T_{n+1}) \cdot (\Gamma_k - T_{n+1}) \leq \frac{C(\Gamma_k - T_{n+1})}{T_{n+1}\log^{3/2}(T_{n+1})}$$

and then

$$\begin{aligned}
(b) &\leq Ca_{n+1}^{-3} \left(\gamma_{N(T_{n+1}-T)} + \sqrt{\gamma_{N(T_{n+1}-T)}}(a_n - a_{n+1}) \right) \|f\|_\infty V^2(x) \int_{T_{n+1}-T}^{T_{n+1}-\gamma_{N(T_{n+1})}} \frac{du}{(T_{n+1}-u)^{3/2}} \\
&\quad + Ca_{n+1}^{-2} \left(\gamma_{N(T_{n+1}-T)} \int_{T_{n+1}-T}^{T_{n+1}-\gamma_{N(T_{n+1})}} \frac{du}{T_{n+1}-u} + \frac{1}{T_{n+1}} \int_{T_{n+1}-T}^{T_{n+1}-\gamma_{N(T_{n+1})}} du \right) \|f\|_\infty V^2(x) \\
&\leq Ca_{n+1}^{-3} \left(\gamma_{N(T_{n+1}-T)} + \sqrt{\gamma_{N(T_{n+1}-T)}}(a_n - a_{n+1}) \right) \|f\|_\infty V^2(x) \gamma_{N(T_{n+1})}^{-1/2} \\
&\quad + Ca_{n+1}^{-2} \left(\gamma_{N(T_{n+1})} |\log(\gamma_{N(T_{n+1})})| + T_{n+1}^{-1} \right) \|f\|_\infty V^2(x) \\
&\leq Ca_{n+1}^{-3} \left(\gamma_{N(T_{n+1})}^{1/2} + (a_n - a_{n+1}) \right) \|f\|_\infty V^2(x).
\end{aligned}$$

Applying Proposition 3.3.2 along with Lemma 3.6.1 yields

$$(c^{\mathrm{end}}) \leq C\|f\|_\infty \left(e^{Ca_{n+1}^{-1}(1+|x|^2)} \gamma_{N(T_{n+1})}^{r/(2r+1)} + a_n^{-2}(a_n - a_{n+1}) \right).$$

We finally obtain for every $x \in \mathbb{R}^d$:

$$\begin{aligned}
&|\mathbb{E}f(X_{T_{n+1}-T_n}^{x,n}) - \mathbb{E}f(\bar{Y}_{T_{n+1}-T_n,T_n}^{x})| \\
&\quad \leq C\|f\|_\infty (a_{n+1}^{-4} e^{C_1 a_{n+1}^{-2}} (a_n - a_{n+1}) V^2(x) \rho_{n+1}^{-1} + e^{Ca_{n+1}^{-1}(1+|x|^2)} \gamma_{N(T_{n+1})}^{r/(2r+1)}).
\end{aligned}$$

The same way as in Section 3.5.2 we get

$$\begin{aligned}
&\mathrm{d_{TV}}(\bar{Y}_{T_{n+1}}^{x_0}, \nu_{a_{n+1}}) \\
&\quad \leq C \left(a_{n+1}^{-4} e^{C_1 a_{n+1}^{-2}} (a_n - a_{n+1}) \max \left[V^2(x_0), 1 + |x_0| \right] \rho_{n+1}^{-1} + e^{Ca_{n+1}^{-1}(1+|x_0|^2)} \gamma_{N(T_{n+1})}^{r/(2r+1)} \right)
\end{aligned}$$

and, for $t \in [T_n, T_{n+1}]$,

$$\begin{aligned}
&\mathrm{d_{TV}}(\bar{Y}_t^{x_0}, \nu_{a(t)}) \\
&\quad \leq Ce^{C_1 a_{n+1}^{-2}} \left(a_{n+1}^{-4} e^{C_1 a_{n+1}^{-2}} (a_n - a_{n+1}) \max \left[V^2(x_0), 1 + |x_0| \right] \rho_{n+1}^{-1} + e^{Ca_{n+1}^{-1}(1+|x_0|^2)} \gamma_{N(T_{n+1})}^{r/(2r+1)} \right).
\end{aligned}$$

□

3.7 Experiments

In this section, we compare the performances of adaptive Langevin-Simulated Annealing algorithms versus vanilla SGLD, that is the Langevin algorithm with constant (additive) σ[1]. We train an artificial neural network on the CIFAR-10 dataset [KH09], which is composed of RGB images of size 32×32 belonging to ten different classes: airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. 50000 images are used for training and 10000 images are used

[1] An implementation of Langevin optimizers in `TensorFlow` is available at `https://github.com/Bras-P/deep-layer-langevin`.

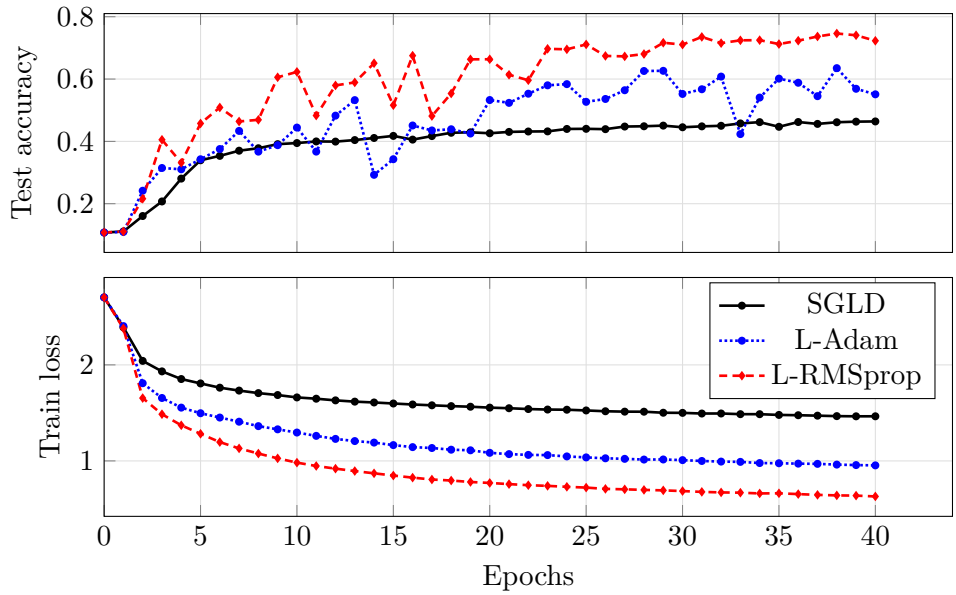


Figure 3.1: Comparison of Langevin optimizers for the training of ResNet20 on the CIFAR-10 dataset. Batch size is 512. The schedules are $\gamma_n = \gamma_1/(1 + c_0 n)$ where $\gamma_1 = 5\,\mathrm{e}{-3}$ and $c_0 n = 1$ after $n = 20$ epochs, and $a_n = A \log^{-1/2}(c_0 n + e)$ where $A = 5\,\mathrm{e}{-3}$.

for test. We use the architecture ResNet20 [HZRS16] with initial width parameter 8. We consider the Adam [KB15] and the RMSprop [TH12] preconditioners for the adaptive Langevin algorithms, giving L-Adam and L-RMSprop respectively. The results are given in Figure 3.1.

As pointed out in the literature [LCCC16], the preconditioned Langevin algorithms show significant improvement compared with the vanilla SGLD algorithm. The convergence is faster and they achieve a lower error on the test set. We also display the value of the loss function on the train set during the training to show that the better performances of the preconditioned algorithms are not due to some overfitting effect.

3.8 Appendix

3.8.1 Proof of Proposition 3.4.2

Proof. We use the characterization of the total variation distance as the L^1-distance between the densities, which reads

$$\begin{aligned}
\mathrm{d_{TV}}(\nu_{a_n}, \nu_{a_{n+1}}) &= \int_{\mathbb{R}^d} \left| \mathcal{Z}_{a_n} e^{-2(V(x)-V^\star)/a_n^2} - \mathcal{Z}_{a_{n+1}} e^{-2(V(x)-V^\star)/a_{n+1}^2} \right| dx \\
&\leq \mathcal{Z}_{a_{n+1}} \int_{\mathbb{R}^d} \left| e^{-2(V(x)-V^\star)/a_n^2} - e^{-2(V(x)-V^\star)/a_{n+1}^2} \right| dx + |\mathcal{Z}_{a_n} - \mathcal{Z}_{a_{n+1}}| \int_{\mathbb{R}^d} e^{-2(V(x)-V^\star)/a_n^2} dx \\
&= \mathcal{Z}_{a_{n+1}} a_{n+1}^d \int_{\mathbb{R}^d} \left| e^{-2(V(a_{n+1}x)-V^\star)/a_n^2} - e^{-2(V(a_{n+1}x)-V^\star)/a_{n+1}^2} \right| dx \\
&\quad + \left| 1 - \frac{\mathcal{Z}_{a_n}}{\mathcal{Z}_{a_{n+1}}} \right| \mathcal{Z}_{a_{n+1}} a_n^d \int_{\mathbb{R}^d} e^{-2(V(a_n x)-V^\star)/a_n^2} dx.
\end{aligned}$$

Using (2.12.3) and (2.12.5), the first term is bounded by

$$C \frac{a_n - a_{n+1}}{a_n} \int_{\mathbb{R}^d} e^{-2(V(a_{n+1}y)-V^\star)/a_n^2} \frac{V(a_{n+1}y) - V^\star}{a_n^2} dx \leq C \frac{a_n - a_{n+1}}{a_n},$$

because the integral converges by dominated convergence as for the proof of (2.12.3). Using (2.12.3) and (2.12.4), the second term is bounded by $C(n\log(n))^{-1}$. □

Chapter 4

Convergence rates of Gibbs measures with degenerate minimum

This chapter corresponds to the article [Bra22] published in *Bernoulli*.

Abstract

We study convergence rates of Gibbs measures, with density proportional to $e^{-f(x)/t}$, as $t \to 0$ where $f : \mathbb{R}^d \to \mathbb{R}$ admits a unique global minimum at $x^\star$. We focus on the case where the Hessian is not definite at $x^\star$. We assume instead that the minimum is strictly polynomial and give a higher order nested expansion of f at $x^\star$, which depends on every coordinate. We give an algorithm yielding such an expansion if the polynomial order of $x^\star$ is no more than 8, in connection with Hilbert's 17^{th} problem. However, we prove that the case where the order is 10 or higher is fundamentally different and that further assumptions are needed. We then give the rate of convergence of Gibbs measures using this expansion. Finally we adapt our results to the multiple well case.

4.1 Introduction

Gibbs measures and their convergence properties are often used in stochastic optimization to minimize a function defined on $\mathbb{R}^d$. That is, let $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable function and let $x^\star \in \mathbb{R}^d$ be such that f admits a global minimum at $x^\star$. It is well known [Hwa80] that under standard assumptions, the associated Gibbs measure with density proportional to $e^{-f(x)/t}$ for $t > 0$, converges weakly to the Dirac mass at $x^\star$, $\delta_{x^\star}$, when $t \to 0$. The Langevin equation $dX_s = -\nabla f(X_s)ds + \sigma dW_s$ consists in a gradient descent with Gaussian noise. For $\sigma = \sqrt{2t}$, its invariant measure π_t has a density proportional to $e^{-f(x)/t}$ (see for example [Kha12], Lemma 4.16), so for small t we can expect it to converge to $\operatorname{argmin}(f)$ [Dal17] [BV22]. The simulated annealing algorithm [vLA87] builds a Markov chain from the Gibbs measure where the parameter t converges to zero over the iterations. This idea is also used in [GM91], giving a stochastic gradient descent algorithm where the noise is gradually decreased to zero. Adding a small noise to the gradient descent allows to explore the space and to escape from traps such as local minima and saddle points which appear in non-convex optimization problems [Laz92] [DPG$^+$14].

Such methods have been recently brought up to light again with SGLD (Stochastic Gradient Langevin Dynamics) algorithms [WT11] [LCCC16], especially for Machine Learning and training of artificial neural networks, which is a high-dimensional non-convex optimization problem.

The rates of convergence of Gibbs measures have been studied in [Hwa80], [Hwa81] and [AH10] under differentiability assumptions on f. It turns out to be of order $t^{1/2}$ as soon as the Hessian matrix $\nabla^2 f(x^\star)$ is positive definite. Furthermore, in the multiple well case i.e. if the minimum of f is attained at finitely many points $x_1^\star$, ..., $x_m^\star$, [Hwa80] proves that the limit distribution is a sum of Dirac masses $\delta_{x_i^\star}$ with coefficients proportional to $\det(\nabla^2 f(x_i^\star))^{-1/2}$ as soon as all the Hessian matrices are positive definite. If such is not the case, we can conjecture that the limit distribution is concentrated around the $x_i^\star$ where the degeneracy is of the highest order.

The aim of this paper is to provide a rate of convergence in this degenerate setting, i.e. when $x^\star$ is still a strict global minimum but $\nabla^2 f(x^\star)$ is no longer definite, which extends the range of applications of Gibbs measure-based algorithms, especially SGLD algorithms, where positive definiteness is generally assumed in the literature. In particular, our results are useful for establishing bounds for the bias made when $\delta_{x^\star}$ is approximated by π_t in Langevin algorithms.

As pointed out by eminent Machine Learning researchers, among them L. Bottou and Y. LeCun, in [SBL16] and [SEU+17], for some classification problems using neural networks, the Hessian of the loss function at the end of the training tends out to be extremely singular. Indeed, as the dimension of the parameter which is used to minimize the loss function is large and as the neural network can be over-parametrized, many eigenvalues of the Hessian matrix are close to zero. "Therefore, a lot of methodology that assumes non-singular Hessian cannot be applied without an appropriate modification" [SBL16]. However, this subject is still new in the Stochastic Optimization literature and needs more theoretical funding. Thus, studying the case of a degenerate minimum helps establishing theoretically the convergence of Langevin algorithms in these cases, as in Chapter 2, where is established the convergence of Langevin-simulated annealing algorithms in the multiplicative case i.e. the diffusion has a noise coefficient σ depending on the position and slowly decreases with time ; the case where the minimum is degenerate is then dealt with using the results of the present article. Furthermore following Section 2.2.3, it is not needed to know beforehand the degree of degeneracy of the minimum in order to establish the convergence to the minimum, however knowing the degree of degeneracy additionally gives the precise rate of convergence of such Langevin algorithms.

A general framework is given in [AH10], which provides rates of convergence based on dominated convergence. However a strong and rather technical assumption on f is needed and checking it seems, to some extent, more demanding than proving the result. To be more precise, the assumption reads as follows: there exists a function $g : \mathbb{R}^d \to \mathbb{R}$ with $e^{-g} \in L^1(\mathbb{R}^d)$ and $\alpha_1, \ldots, \alpha_d \in (0, +\infty)$ such that

$$\forall h \in \mathbb{R}^d, \quad \frac{1}{t}\left[f(x^\star + (t^{\alpha_1} h_1, \ldots, t^{\alpha_d} h_d)) - f(x^\star)\right] \underset{t\to 0}{\longrightarrow} g(h_1, \ldots, h_d). \tag{4.1.1}$$

Our objective is to give conditions on f such that (4.1.1) is fulfilled and then to elucidate the expression of g depending on f and its derivatives by studying the behaviour of f at $x^\star$ in every direction. In particular, we detail the different cases that can occur. Doing so we can apply the results from [AH10] yielding the convergence rate of the corresponding Gibbs measures. The orders α_1, ..., α_d must be chosen carefully and not too large, as the function g needs to depend on every of its variables h_1, ..., h_d, which is a necessary condition for e^{-g} to be integrable. We also extend our results to the multiple well case.

We generally assume f to be coercive, i.e. $f(x) \to +\infty$ as $||x|| \to +\infty$, $\mathcal{C}^{2p}$ in a neighbourhood of $x^\star$ for some $p \in \mathbb{N}$ and we assume that the minimum is polynomial strict, i.e.

the function f is bounded below in a neighbourhood of $x^\star$ by some non-negative polynomial function, null only at $x^\star$. Thus we can apply a multi-dimensional Taylor expansion to f at $x^\star$, where the successive derivatives of $f : \mathbb{R}^d \to \mathbb{R}$ are seen as symmetric tensors of $\mathbb{R}^d$. The idea is then to consider the successive subspaces where the derivatives of f are null up to some order ; using that the Taylor expansion of $f(x^\star+h)-f(x^\star)$ is non-negative, some cross derivative terms are null. However a difficulty arises at orders 6 and higher, as the set where the derivatives of f are null up to some order is no longer a vector subspace in general. This difficulty is linked with Hilbert's 17^{th} problem [Hil88], stating that a non-negative multivariate polynomial cannot be written as the sum of squares of polynomials in general. We thus need to change the definition of the subspaces we consider. Following this, we give a recursive algorithm yielding an adapted decomposition of $\mathbb{R}^d$ into vector subspaces and a function g satisfying (4.1.1) up to a change of basis, giving a canonical higher order nested decomposition of f at $x^\star$ in degenerate cases. An interesting fact is that the case where the polynomial order of $x^\star$ is 10 or higher fundamentally differs from those of orders 2, 4, 6 and 8, owing to the presence of even cross terms which may be not null. The algorithm we provide works at the orders 10 or higher only under the assumption that all such even cross terms are null. In general, it is more difficult to get a general expression of g for the orders 10 and higher. We then apply our results to [AH10], where we give conditions such that the hypotheses of [AH10], especially (4.1.1), are satisfied so as to infer rates of convergence of Gibbs measures in the degenerate case where $\nabla^2 f(x^\star)$ is not necessarily positive definite. The function g given by our algorithm is a non-negative polynomial function and non-constant in any of its variables, however it needs to be assumed to be coercive to be applied to [AH10]. We study the case where g is not coercive and give a method to deal with simple generic non-coercive cases, where our algorithm seems to be a first step to a more general procedure. However, we do not give a general method in this case.

Our results are applied to Gibbs measures but they can also be applied to more general contexts, as we give a canonical higher order nested expansion of f at a minimum, in the case where some derivatives are degenerate.

For general properties of symmetric tensors we refer to [CGLM08]. In the framework of stochastic approximation, [FP99] Section 3.1 introduced the notion of strict polynomial local extremum and investigated their properties as higher order "noisy traps".

The paper is organized as follows. In Section 4.3, we recall convergence properties of Gibbs measures and revisit the main theorem from [AH10]. This theorem requires, as an hypothesis, to find an expansion of f at its global minimum ; we properly state this problem in Section 4.3.2 under the assumption of strict polynomial minimum. In Section 4.3.3, we state our main result for both single well and multiple well cases, as well as our algorithm. In Section 4.4, we detail the expansion of f at its minimum for each order and provide the proof. We give the general expression of the canonical higher order nested expansion at any order in Section 4.4.1, where we distinguish the orders 10 and higher from the lower ones. We then provide the proof for each order 2, 4, 6 and 8 in Sections 4.4.3, 4.4.4, 4.4.6 and 4.4.7 respectively. We need to prove that, with the exponents $\alpha_1, \ldots, \alpha_d$ we specify, the convergence in (4.1.1) holds ; we do so by proving that, using the non-negativity of the Taylor expansion, some cross derivative terms are zero. Because of Hilbert's 17^{th} problem, we need to distinguish the orders 6 and 8 from the orders 2 and 4, as emphasized in Section 4.4.5. For orders 10 and higher, such terms are not necessarily zero and must then be assumed to be zero. In Section 4.4.8, we give a counter-example if this assumption is not satisfied before proving the result. In Section 4.4.9, we prove that for every order the resulting function g is constant in none of its variables and that the convergence in (4.1.1) is uniform on every compact set. In Section 4.4.10, we study the case where the function g is not coercive and give a method to deal with the simple generic case. In Section 4.5, we prove our main theorems stated in Section 4.3.3 using the expansion of f established in Section 4.4.

In Section 4.6, we deal with a "flat" example where all the derivatives in the local minimum are zero and where we cannot apply our main theorems. In Section 4.7 we give a practical example of a function f arising from an optimization problem with non definite Hessian matrix at the minimum and we numerically compute high order derivatives.

4.2 Definitions and notations

We give a brief list of notations that are used throughout the paper.

We endow $\mathbb{R}^d$ with its canonical basis $(e_1, \dots, e_d)$ and the Euclidean norm denoted by $||\cdot||$. For $x \in \mathbb{R}^d$ and $r > 0$ we denote by $\mathcal{B}(x, r)$ the Euclidean ball of $\mathbb{R}^d$ of center x and radius r. For E a vector subspace of $\mathbb{R}^d$, we denote by $p_E : \mathbb{R}^d \to E$ the orthogonal projection on E. For a decomposition of $\mathbb{R}^d$ into orthogonal subspaces, $\mathbb{R}^d = E_1 \oplus \cdots \oplus E_p$, we say that an orthogonal transformation $B \in \mathcal{O}_d(\mathbb{R})$ is adapted to this decomposition if for all $j \in \{1, \dots, p\}$,

$$\forall i \in \{\dim(E_1) + \cdots + \dim(E_{j-1}) + 1, \dots, \dim(E_1) + \cdots + \dim(E_j)\},\ B \cdot e_i \in E_j.$$

For $a,\ b \in \mathbb{R}^d$, we denote by $a * b$ the element-wise product, i.e.

$$\forall i \in \{1, \dots, d\},\ (a * b)_i = a_i b_i.$$

For v^1, …, v^k vectors in $\mathbb{R}^d$ and T a tensor of order k of $\mathbb{R}^d$, we denote the tensor product

$$T \cdot (v^1 \otimes \cdots \otimes v^k) = \sum_{i_1, \dots, i_k \in \{1, \dots, d\}} T_{i_1 \cdots i_k} v^1_{i_1} \dots v^k_{i_k}.$$

More generally, if $j \leq k$ and $v^1,\ \dots,\ v^j$ are j vectors in $\mathbb{R}^d$, then $T \cdot (v^1 \otimes \cdots \otimes v^j)$ is a tensor of order $k - j$ such that:

$$T \cdot (v^1 \otimes \cdots \otimes v^j)_{i_{j+1} \dots i_k} = \sum_{i_1, \dots, i_j \in \{1, \dots, d\}} T_{i_1 \dots i_k} v^1_{i_1} \dots v^j_{i_j}.$$

For $h \in \mathbb{R}^d$, $h^{\otimes k}$ denotes the tensor of order k such that

$$h^{\otimes k} = (h_{i_1} \dots h_{i_k})_{i_1, \dots, i_k \in \{1, \dots, d\}}.$$

For a function $f \in \mathcal{C}^p\left(\mathbb{R}^d, \mathbb{R}\right)$, we denote $\nabla^k f(x)$ the differential of order $k \leq p$ of f at x, as $\nabla^k f(x)$ is the tensor of order k defined by:

$$\nabla^k f(x) = \left(\frac{\partial^k f(x)}{\partial x_{i_1} \cdots \partial x_{i_k}} \right)_{i_1, i_2, \dots, i_k \in \{1, \dots, d\}}.$$

By Schwarz's theorem, this tensor is symmetric, i.e. for all permutation $\sigma \in \mathfrak{S}_k$,

$$\frac{\partial^k f(x)}{\partial x_{i_{\sigma(1)}} \cdots \partial x_{i_{\sigma(k)}}} = \frac{\partial^k f(x)}{\partial x_{i_1} \cdots \partial x_{i_k}}.$$

We recall the Taylor-Young formula in any dimension, and the Newton multinomial formula.

Theorem 4.2.1 (Taylor-Young formula)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^p$ and let $x \in \mathbb{R}^d$. Then:*

$$f(x + h) \underset{h \to 0}{=} \sum_{k=0}^{p} \frac{1}{k!} \nabla^k f(x) \cdot h^{\otimes k} + ||h||^p o(1).$$

We denote by $\binom{k}{i_1,\dots,i_p}$ the p-nomial coefficient, defined as:

$$\binom{k}{i_1,\dots,i_p} = \frac{k!}{i_1!\dots i_p!}.$$

Theorem 4.2.2 (Newton multinomial formula)**.** *Let $h_1, \dots, h_p \in \mathbb{R}^d$, then*

$$(h_1 + h_2 + \cdots + h_p)^{\otimes k} = \sum_{\substack{i_1,\dots,i_p \in \{0,\dots,k\} \\ i_1+\cdots+i_p=k}} \binom{k}{i_1,\dots,i_p} h_1^{\otimes i_1} \otimes \cdots \otimes h_p^{\otimes i_p}. \tag{4.2.1}$$

For T a tensor of order k, we say that T is non-negative (resp. positive) if

$$\forall h \in \mathbb{R}^d,\ T \cdot h^{\otimes k} \geq 0 \text{ (resp. } T \cdot h^{\otimes k} > 0). \tag{4.2.2}$$

We denote $L^1(\mathbb{R}^d)$ the set of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ that are integrable with respect to the Lebesgue measure on $\mathbb{R}^d$. We denote by λ_d the Lebesgue measure on $\mathbb{R}^d$. For $f : \mathbb{R}^d \to \mathbb{R}$ such that $e^{-f} \in L^1(\mathbb{R}^d)$, we define for $t > 0$, $C_t := \left(\int_{\mathbb{R}^d} e^{-f/t}\right)^{-1}$ and π_t the Gibbs measure

$$\pi_t(x)dx := C_t e^{-f(x)/t} dx.$$

For a family of random variables $(Y_t)_{t\in(0,1]}$ and Y a random variable, we write $Y_t \xrightarrow[t\to 0]{\mathscr{L}} Y$ meaning that (Y_t) weakly converges to Y.

We give the following definition of a strict polynomial local minimum of f:

Definition 4.2.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^{2p}$ for $p \in \mathbb{N}$ and let $x^\star$ be a local minimum of f. We say that f has a strict polynomial local minimum at $x^\star$ of order $2p$ if p is the smallest integer such that:*

$$\exists r > 0,\ \forall h \in \mathcal{B}(0,r) \setminus \{0\},\ \sum_{k=2}^{2p} \frac{1}{k!} \nabla^k f(x^\star) \cdot h^{\otimes k} > 0. \tag{4.2.3}$$

Remarks :

1. A local minimum $x^\star$ of f is not necessarily strictly polynomial, for example, $f : x \mapsto e^{-||x||^{-2}}$ and $x^\star = 0$.

2. If $x^\star$ is polynomial strict, then the order is necessarily even, because if $x^\star$ is not polynomial strict of order $2l$ for some $l \in \mathbb{N}$, then we have $h_n \to 0$ such that the Taylor expansion in h_n up to order $2l$ is zero ; by the minimum condition, the Taylor expansion in h_n up to order $2l+1$ must be non-negative, so we also have $\nabla^{2l+1} f(x^\star) \cdot h_n^{\otimes 2l+1} = 0$.

For $f : \mathbb{R}^d \to \mathbb{R}$ such that $\min_{\mathbb{R}^d}(f)$ exists, we denote by $\operatorname{argmin}(f)$ the arguments of the minima of f, i.e.

$$\operatorname{argmin}(f) = \left\{ x \in \mathbb{R}^d :\ f(x) = \min_{\mathbb{R}^d}(f) \right\}.$$

Without ambiguity, we write "minimum" or "local minimum" to designate $f(x^\star)$ as well as $x^\star$. Finally, we define, for $x^\star \in \mathbb{R}^d$ and $p \in \mathbb{N}$:

$$\mathscr{A}_p(x^\star) := \left\{ f \in \mathcal{C}^{2p}(\mathbb{R}^d, \mathbb{R}) :\ f \text{ admits a local minimum at } x^\star \right\}.$$
$$\mathscr{A}_p^\star(x^\star) := \left\{ f \in \mathcal{C}^{2p}(\mathbb{R}^d, \mathbb{R}) :\ f \text{ admits a strict polynomial local minimum at } x^\star \text{ of order } 2p \right\}.$$

4.3 Convergence of Gibbs measures

4.3.1 Properties of Gibbs measures

Let us consider a Borel function $f : \mathbb{R}^d \to \mathbb{R}$ with $e^{-f} \in L^1(\mathbb{R}^d)$. We study the asymptotic behaviour of the probability measures of density for $t \in (0, \infty)$:

$$\pi_t(x)dx = C_t e^{-\frac{f(x)}{t}} dx$$

when $t \to 0$. When t is small, the measure π_t tends to the set $\operatorname{argmin}(f)$. The following proposition makes this statement precise.

Proposition 4.3.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Borel function such that*

$$f^\star := \operatorname{essinf}(f) = \inf\{y : \ \lambda_d\{f \leq y\} > 0\} > -\infty,$$

and $e^{-f} \in L^1(\mathbb{R}^d)$. Then

$$\forall \varepsilon > 0, \ \pi_t(\{f \geq f^\star + \varepsilon\}) \underset{t\to 0}{\longrightarrow} 0.$$

Proof. As $f^\star > -\infty$, we may assume without loss of generality that $f^\star = 0$ by replacing f by $f - f^\star$. Let $\varepsilon > 0$. It follows from the assumptions that $f \geq 0$ λ_d-*a.e.* and $\lambda_d\{f \leq \varepsilon\} > 0$ for every $\varepsilon > 0$. As $e^- f \in L^1(\mathbb{R}^d)$, we have

$$\lambda_d\{f \leq \varepsilon/3\} \leq e^{\varepsilon/3} \int_{\mathbb{R}^d} e^{-f} d\lambda_d < +\infty.$$

Moreover by dominated convergence, it is clear that

$$C_t^{-1} \downarrow \lambda_d\{f = 0\} < +\infty.$$

We have

$$C_t \leq \left(\int_{f \leq \varepsilon/3} e^{-\frac{f(x)}{t}} dx \right)^{-1} \leq \left(e^{-\frac{\varepsilon}{3t}} \underbrace{\lambda_d\{f \leq \frac{\varepsilon}{3}\}}_{>0} \right)^{-1}.$$

Then

$$\pi_t\{f \geq \varepsilon\} = C_t \int_{f \geq \varepsilon} e^{-\frac{f(x)}{t}} dx \leq \frac{e^{\varepsilon/3t} \int_{f\geq\varepsilon} e^{-f(x)/t} dx}{\lambda_d\{f \leq \frac{\varepsilon}{3}\}} \leq \frac{e^{-\varepsilon/3t} C_{3t}^{-1}}{\lambda_d\{f \leq \frac{\varepsilon}{3}\}} \underset{t\to 0}{\longrightarrow} 0,$$

because if $f(x) \geq \varepsilon$, then $e^{-\frac{f(x)}{t}} \leq e^{-\frac{2\varepsilon}{3t}} e^{-\frac{f(x)}{3t}}$, and where we used that $C_{3t}^{-1} \leq C_1^{-1}$ if $t \leq 1/3$ □

Now, let us assume that $f : \mathbb{R}^d \to \mathbb{R}$ is continuous, $e^{-f} \in L^1(\mathbb{R}^d)$ and f admits a unique global minimum at $x^\star$ so that $\operatorname{argmin}(f) = \{x^\star\}$. In [AH10] is proved the weak convergence of π_t to $\delta_{x^\star}$ and a rate of convergence depending on the behaviour of $f(x^\star + h) - f(x^\star)$ for small enough h. Let us recall this result in detail.

Theorem 4.3.2 (Athreya-Hwang, 2010)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Borel function and let $x^\star \in \mathbb{R}^d$ such that :*

1. *$e^{-f} \in L^1(\mathbb{R}^d)$.*
2. *For all $\delta > 0$, $\inf\{f(x) - f(x^\star), \ ||x - x^\star|| > \delta\} > 0$.*

3. There exist $\alpha_1, \ldots, \alpha_d > 0$ *such that for all* $(h_1, \ldots, h_d) \in \mathbb{R}^d$,

$$\frac{1}{t}[f(x^\star + (t^{\alpha_1}h_1, \ldots, t^{\alpha_d}h_d)) - f^\star] \xrightarrow[t\to 0]{} g(h_1, \ldots, h_d) \in \mathbb{R}.$$

4. $\displaystyle\int_{\mathbb{R}^d} \sup_{0<t<1} e^{-\frac{f(x^\star+(t^{\alpha_1}h_1,\ldots,t^{\alpha_d}h_d))-f(x^\star)}{t}} dh_1 \ldots dh_d < \infty.$

For $0 < t < 1$, *let* X_t *be a random vector with distribution* π_t. *Then* $e^{-g} \in L^1(\mathbb{R}^d)$ *and*

$$\left(\frac{(X_t - x^\star)_1}{t^{\alpha_1}}, \ldots, \frac{(X_t - x^\star)_d}{t^{\alpha_d}}\right) \xrightarrow{\mathscr{L}} X \quad \text{as } t \to 0 \tag{4.3.1}$$

where the distribution of X *has a density proportional to* $e^{-g(x_1,\ldots,x_d)}$.

Remark: Hypothesis 2. is verified as soon as f is continuous, coercive (i.e. $f(x) \longrightarrow +\infty$ when $||x|| \to +\infty$) and that $\operatorname{argmin}(f) = \{0\}$.

To study the rate of convergence of the measure π_t when $t \to 0$ using Theorem 4.3.2, we need to identify $\alpha_1, \ldots, \alpha_d$ and g such that the condition (4.3.1) holds, up to a possible change of basis. Since $x^\star$ is a local minimum, the Hessian $\nabla^2 f(x^\star)$ is positive semi-definite. Moreover, if $\nabla^2 f(x^\star)$ is positive definite, then choosing $\alpha_1 = \cdots = \alpha_d = \frac{1}{2}$, we have:

$$\frac{1}{t}[f(x^\star + t^{1/2}h) - f(x^\star)] \xrightarrow[t\to 0]{} \frac{1}{2}h^T \cdot \nabla^2 f(x^\star) \cdot h := g(x).$$

And using an orthogonal change of variable:

$$\int_{\mathbb{R}^d} e^{-g(x)}dx = \int_{\mathbb{R}^d} e^{-\frac{1}{2}\sum_{i=1}^d \beta_i y_i^2} dy_1 \ldots dy_d < \infty,$$

where the eigenvalues β_i are positive. However, if $\nabla^2 f(x^\star)$ is not positive definite, then some of the β_i are zero and the integral does not converge.

4.3.2 Statement of the problem

We still consider the function $f : \mathbb{R}^d \to \mathbb{R}$ and assume that $f \in \mathscr{A}_p^\star(x^\star)$ for some $x^\star \in \mathbb{R}^d$ and some integer $p \geq 1$. Then our objective is to find $\alpha_1 \geq \cdots \geq \alpha_d \in (0, +\infty)$ and an orthogonal transformation $B \in \mathcal{O}_d(\mathbb{R})$ such that:

$$\forall h \in \mathbb{R}^d, \quad \frac{1}{t}\left[f(x^\star + B \cdot (t^\alpha * h)) - f(x^\star)\right] \xrightarrow[t\to 0]{} g(h_1, \ldots, h_d), \tag{4.3.2}$$

where t^α denotes the vector $(t^{\alpha_1}, \ldots, t^{\alpha_d})$ and where $g : \mathbb{R}^d \to \mathbb{R}$ is a measurable function which is not constant in any $h_1, \ldots, h_d$, i.e. for all $i \in \{1, \ldots, d\}$, there exist $h_1, \ldots, h_{i-1}, h_{i+1}, \ldots, h_d \in \mathbb{R}^d$ such that

$$h_i \mapsto g(h_1, \ldots, h_d) \text{ is not constant.} \tag{4.3.3}$$

Then we say that $\alpha_1, \ldots, \alpha_d$, B and g are a solution of the problem (4.3.2). The hypothesis that g is not constant in any of its variables is important ; otherwise, we could simply take $\alpha_1 = \cdots = \alpha_d = 1$ and obtain, by the first order condition:

$$\frac{1}{t}\left[f(x^\star + t(h_1, \ldots, h_d)) - f(x^\star)\right] \xrightarrow[t\to 0]{} 0.$$

4.3.3 Main results : rate of convergence of Gibbs measures

4.3.3.1 Single well case

Theorem 4.3.3 (Single well case)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^{2p}$ with $p \in \mathbb{N}$ and let $x^\star \in \mathbb{R}^d$ such that:*

1. *f is coercive, i.e. $f(x) \to +\infty$ when $||x|| \to +\infty$.*
2. *$\operatorname{argmin}(f) = x^\star$.*
3. *$f \in \mathscr{A}_p^\star(x^\star)$.*
4. *$e^{-f} \in L^1(\mathbb{R}^d)$.*

Let $(E_k)_k$, $(\alpha_i)_i$, B and g to be defined as in Algorithm 1 stated right after, so that for all $h \in \mathbb{R}^d$,

$$\frac{1}{t}\left[f\left(x^\star + B \cdot (t^\alpha * h)\right) - f(x^\star)\right] \underset{t\to 0}{\longrightarrow} g(h),$$

and where g is not constant in any of its variables. Moreover, assume that g is coercive and the following technical hypothesis if $p \geq 5$:

$$\begin{aligned} &\forall h \in \mathbb{R}^d,\ \forall (i_1, \dots, i_p) \in \{0, 2, \cdots, 2p\}^p, \\ &\frac{i_1}{2} + \cdots + \frac{i_p}{2p} < 1 \implies \nabla^{i_1+\cdots+i_p} f(x^\star) \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} = 0. \end{aligned} \tag{4.3.4}$$

Then the conclusion of Theorem 4.3.2 holds, with:

$$\left(\frac{1}{t^{\alpha_1}}, \dots, \frac{1}{t^{\alpha_d}}\right) * (B^{-1} \cdot (X_t - x^\star)) \xrightarrow{\mathscr{L}} X \quad \text{as } t \to 0,$$

where X has a density proportional to $e^{-g(x)}$.

Algorithm 1. Let $f \in \mathscr{A}_p^\star(x^\star)$ for $p \in \mathbb{N}$.

1. Define $(F_k)_{0 \leq k \leq p-1}$ recursively as:

$$\begin{cases} F_0 = \mathbb{R}^d \\ F_k = \{h \in F_{k-1} :\ \forall h' \in F_{k-1},\ \nabla^{2k} f(x^\star) \cdot h \otimes h'^{\otimes 2k-1} = 0\}. \end{cases}$$

2. For $1 \leq k \leq p-1$, define the subspace E_k as the orthogonal complement of F_k in F_{k-1}. By abuse of notation, define $E_p := F_{p-1}$.

3. Define $B \in \mathcal{O}_d(\mathbb{R})$ as an orthogonal transformation adapted to the decomposition

$$\mathbb{R}^d = E_1 \oplus \cdots \oplus E_p.$$

4. Define for $1 \leq i \leq d$,

$$\alpha_i := \frac{1}{2j} \quad \text{for } i \in \{\dim(E_1) + \cdots + \dim(E_{j-1}) + 1, \dots, \dim(E_1) + \cdots + \dim(E_j)\}. \tag{4.3.5}$$

5. Define $g : \mathbb{R}^d \to \mathbb{R}$ as

$$g(h) = \sum_{k=2}^{2p} \frac{1}{k!} \sum_{\substack{i_1, \dots, i_p \in \{0, \dots, k\} \\ i_1 + \cdots + i_p = k \\ \frac{i_1}{2} + \cdots + \frac{i_p}{2p} = 1}} \binom{k}{i_1, \dots, i_p} \nabla^k f(x^\star) \cdot p_{E_1}(B \cdot h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(B \cdot h)^{\otimes i_p}. \tag{4.3.6}$$

Remarks :

1. The function g is not unique, as we can choose any base B adapted to the decomposition $\mathbb{R}^d = E_1 \oplus \cdots \oplus E_p$.
2. The case $p \geq 5$ is fundamentally different from the case $p \leq 4$, since Algorithm 1 may fail to provide such $(E_k)_k$, $(\alpha_i)_i$, B and g if the technical hypothesis (4.3.4) is not fulfilled, as explained in Section 4.4.8. This yields fewer results for the case $p \geq 5$.
3. For $p \in \{1, 2, 3, 4\}$, the detail the expression of g in (4.4.7), (4.4.8), (4.4.10) and (4.4.13) respectively.
4. The function g has the following general properties : g is a non-negative polynomial of order $2p$; $g(0) = 0$ and $\nabla g(0) = 0$.
5. The condition on g to be coercive may seem not natural. We give more details about the case where g is not coercive in Section 4.4.10 and give a way to deal with the simple generic case of non-coercivity. However dealing with the general case where g is not coercive goes beyond the scope of our work.
6. The hypothesis that g is coercive is a necessary condition for $e^{-g} \in L^1(\mathbb{R}^d)$. We actually prove in Proposition 4.4.6 that it is a sufficient condition.

Practical aspect: How to check the technical hypothesis (4.3.4)**:** Let us define $T_k := \nabla^k f(x^\star)$ for $k \leq 2p$ and $\mathcal{I}_p := \{(i_1, \ldots, i_p) \in \{0, 2, \ldots, 2p\}^{2p} : \ i_1/2 + \cdots + i_p/(2p) < 1\}$.

We first apply Algorithm 1. We perform the change of basis given by B; then the new derivative tensor after change of coordinates is given by $(B^\top)^{\otimes k} \cdot T_k$. To simplify the notations, let us assume that $B = I_d$. For every $(i_1, \ldots, i_p) \in \mathcal{I}_p$ we check the condition (4.3.4) as follows. Let $k := i_1 + \cdots + i_p$. The function $h \mapsto T_k \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p}$ is a polynomial function, so we need to check that all of its coefficient are null. Let us define the sets of indexes:

$$I_\ell := \{i_1 + \cdots + i_{\ell-1} + 1, \ldots, i_1 + \cdots + i_\ell\}, \qquad 1 \leq \ell \leq p,$$
$$L_\ell := \{\dim(E_1) + \cdots + \dim(E_{\ell-1}) + 1, \ldots, \dim(E_1) + \cdots + \dim(E_\ell)\}, \qquad 1 \leq \ell \leq p.$$

Having in mind that $B = I_d$ so that for all $h \in \mathbb{R}^d$, $1 \leq m \leq p$ and $1 \leq j \leq d$, we have $[p_{E_m}(h)]_j = h_j$ if $j \in L_m$ and 0 otherwise. Then we have

$$\begin{aligned} T_k \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} &= \sum_{j_1, \ldots, j_k \in \{1, \ldots, d\}} [T_k]_{j_1, \ldots, j_k} \prod_{m=1}^{p} \prod_{\ell \in I_m} [p_{E_m}(h)]_{j_l} \\ &= \sum_{(j_1, \ldots, j_k) \in L_1^{i_1} \times \cdots \times L_p^{i_p}} [T_k]_{j_1, \ldots, j_k} \prod_{\ell=1}^{k} h_{j_\ell}. \end{aligned}$$

Moreover, having in mind that the tensor T_k is a symmetric tensor, we deduce that the condition (4.3.4) is satisfied if and only if

$$\forall (i_1, \ldots, i_p) \in \mathcal{I}_p, \ \forall (j_1, \ldots, j_{i_1 + \cdots + i_p}) \in L_1^{i_1} \times \cdots \times L_p^{i_p}, \ [T_{i_1 + \cdots + i_p}]_{j_1, \ldots, j_{i_1 + \cdots + i_p}} = 0.$$

Practical aspect: If the function f is analytically known, then the derivative tensors can be computed analytically, in particular using automatic differentiation. If the function f is not analytically known, in particular if f is known only at some points, then the present article gives possible convergence rates of Gibbs-based minimization algorithms and proves that the convergence rate is still polynomial under the assumption that the minimum is strictly polynomial. Degenerate cases yield a different behaviour that can be empirically detected; the order of degeneracy can be estimated as well.

4.3.3.2 Multiple well case

Still following [AH10], we study the multiple well case, i.e. the global minimum is attained in a finite number of points in $\mathbb{R}^d$, say $\{x_1^\star, \dots, x_m^\star\}$ for some $m \in \mathbb{N}$. In this case, the limiting measure of π_t will have its support in $\{x_1^\star, \dots, x_m^\star\}$, with different weights.

Theorem 4.3.4 (Athreya-Hwang, 2010)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Borel function, let $f^\star \in \mathbb{R}$ and let $x_1^\star, \dots, x_m^\star \in \mathbb{R}^d$ such that :*

1. *$e^{-f} \in L^1(\mathbb{R}^d)$.*
2. *For all $\delta > 0$, $\inf\{f(x) - f^\star,\ ||x - x_i^\star|| > \delta,\ 1 \le i \le m\} > 0$.*
3. *There exist $(\alpha_{ij})_{\substack{1\le i\le m\\ 1\le j\le d}}$ such that for all i, j, $\alpha_{ij} \ge 0$ and for all i:*

$$\frac{1}{t}[f(x_i^\star + (t^{\alpha_{i1}}h_1, \dots, t^{\alpha_{id}}h_d)) - f^\star] \underset{t\to 0}{\longrightarrow} g_i(h_1, \dots, h_d) \in [0, \infty).$$

4. *For all $i \in \{1, \dots, m\}$,*

$$\int_{\mathbb{R}^d} \sup_{0<t<1} e^{-\frac{f(x_i^\star + (t^{\alpha_{i1}}h_1, \dots, t^{\alpha_{id}}h_d)) - f^\star}{t}} dh_1 \dots dh_d < \infty.$$

Then, let $\alpha := \min_{1\le i\le m}\left\{\sum_{j=1}^d \alpha_{ij}\right\}$ and let $J := \left\{i \in \{1, \dots, m\} :\ \sum_{j=1}^d \alpha_{ij} = \alpha\right\}$. For $0 < t < 1$, let X_t be a random vector with distribution π_t. Then:

$$X_t \underset{t\to 0}{\overset{\mathscr{L}}{\longrightarrow}} \frac{1}{\sum_{j\in J}\int_{\mathbb{R}^d} e^{-g_j(x)}dx} \sum_{i\in J} \int_{\mathbb{R}^d} e^{-g_i(x)}dx \cdot \delta_{x_i^\star}.$$

Theorem 4.3.5 (Multiple well case)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^{2p}$ for $p \in \mathbb{N}$, let $f^\star \in \mathbb{R}$ and let $x_1^\star, \dots, x_m^\star \in \mathbb{R}^d$ such that:*

1. *f is coercive i.e. $f(x) \to +\infty$ when $||x|| \to +\infty$.*
2. *$\operatorname{argmin}(f) = \{x_1^\star, \dots, x_m^\star\}$ and for all i, $f(x_i^\star) = f^\star$.*
3. *For all $i \in \{1, \dots, m\}$, $f \in \mathscr{A}_{p_i}^\star(x_i^\star)$ for some $p_i \le p$.*
4. *$e^{-f} \in L^1(\mathbb{R}^d)$.*

Then, for every $i \in \{1, \dots, m\}$, we consider $(E_{ik})_k$, $(\alpha_{ij})_j$, B_i and g_i as defined in Algorithm 1, where we consider f to be in $\mathscr{A}_{p_i}^\star(x_i^\star)$, so that for every $h \in \mathbb{R}^d$:

$$\frac{1}{t}[f(x_i^\star + B_i \cdot (t^{\alpha_i} * h)) - f^\star] \underset{t\to 0}{\longrightarrow} g_i(h_1, \dots, h_d) \in [0, \infty),$$

where t^{α_i} is the vector $(t^{\alpha_{i1}}, \dots, t^{\alpha_{id}})$ and where g_i is not constant in any of its variables. Furthermore, we assume that for all i, g_i is coercive and the following technical hypothesis for every i such that $p_i \ge 5$:

$$\forall h \in \mathbb{R}^d,\ \forall (i_1, \dots, i_{p_i}) \in \{0, 2, \dots, 2p_i\}^{p_i},$$
$$\frac{i_1}{2} + \dots + \frac{i_{p_i}}{2p} < 1 \implies \nabla^{i_1+\dots+i_{p_i}} f(x_i^\star) \cdot p_{E_{i1}}(h)^{\otimes i_1} \otimes \dots \otimes p_{E_{ip_i}}(h)^{\otimes i_{p_i}} = 0.$$

Let $\alpha := \min_{1\leq i\leq m}\left\{\sum_{j=1}^d \alpha_{ij}\right\}$ *and let* $J := \left\{i \in \{1,\dots,m\} :\ \sum_{j=1}^d \alpha_{ij} = \alpha\right\}$. *Then:*

$$X_t \underset{t\to 0}{\longrightarrow} \frac{1}{\sum_{j\in J}\int_{\mathbb{R}^d} e^{-g_j(x)}dx}\sum_{i\in J}\int_{\mathbb{R}^d} e^{-g_i(x)}dx\cdot\delta_{x_i^\star}.$$

Moreover, let $\delta > 0$ *be small enough so that the balls* $\mathcal{B}(x_i^\star,\delta)$ *are disjoint, and define the random vector* X_{it} *to have the law of* X_t *conditionally to the event* $||X_t - x_i^\star|| < \delta$. *Then:*

$$\left(\frac{1}{t^{\alpha_{i1}}},\dots,\frac{1}{t^{\alpha_{id}}}\right) * (B_i^{-1}\cdot(X_{it} - x_i^\star)) \xrightarrow{\mathscr{L}} X_i \quad \text{as } t\to 0,$$

where X_i *has a density proportional to* $e^{-g_i(x)}$.

4.4 Expansion of f at a local minimum with degenerate derivatives

In this section, we aim at answering to the problem stated in (4.3.2) in order to devise conditions to apply Theorem 4.3.2. This problem can also be considered in a more general setting, independently of the study of the convergence of Gibbs measures. It provides a non degenerate higher order nested expansion of f at a local minimum when some of the derivatives of f are degenerate. Note here that we only need $x^\star$ to be a local minimum instead of a global minimum, since we only give local properties.

For $k \leq 2p$, we define the tensor of order k, $T_k := \nabla^k f(x^\star)$.

4.4.1 Expansion of f for any order p

In this section, we state our result in a synthetic form. The proofs of the cases $p = 1,2,3,4$ are individually detailled in Sections 4.4.3, 4.4.4, 4.4.6 and 4.4.7 respectively.

Theorem 4.4.1. *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be* $\mathcal{C}^{2p}$ *for some* $p \in \mathbb{N}$ *and assume that* $f \in \mathscr{A}_p^\star(x^\star)$ *for some* $x^\star \in \mathbb{R}^d$.

1. *If* $p \in \{1,2,3,4\}$, *then there exists orthogonal subspaces of* $\mathbb{R}^d$, $E_1,\ \dots,\ E_p$ *such that*

$$\mathbb{R}^d = E_1 \oplus \cdots \oplus E_p,$$

and satisfying for every $h \in \mathbb{R}^d$:

$$\frac{1}{t}\left[f\left(x^\star + t^{1/2}p_{E_1}(h) + \cdots + t^{1/(2p)}p_{E_p}(h)\right) - f(x^\star)\right] \tag{4.4.1}$$

$$\underset{t\to 0}{\longrightarrow} \sum_{k=2}^{2p}\frac{1}{k!}\sum_{\substack{i_1,\dots,i_p\in\{0,\cdots,k\}\\ i_1+\cdots+i_p=k\\ \frac{i_1}{2}+\cdots+\frac{i_p}{2p}=1}}\binom{k}{i_1,\dots,i_p}T_k\cdot p_{E_1}(h)^{\otimes i_1}\otimes\cdots\otimes p_{E_p}(h)^{\otimes i_p}. \tag{4.4.2}$$

The convergence is uniform with respect to h *on every compact set. Moreover, let* $B \in \mathcal{O}_d(\mathbb{R})$ *be an orthogonal transformation adapted to the decomposition* $E_1 \oplus \cdots \oplus E_p$, *then*

$$\frac{1}{t}\left[f\left(x^\star + B\cdot(t^\alpha * h)\right) - f(x^\star)\right] \underset{t\to 0}{\longrightarrow} g(h), \tag{4.4.3}$$

where

$$g(h) = \sum_{k=2}^{2p} \frac{1}{k!} \sum_{\substack{i_1,\dots,i_p \in \{0,\dots,k\} \\ i_1+\cdots+i_p=k \\ \frac{i_1}{2}+\cdots+\frac{i_p}{2p}=1}} \binom{k}{i_1,\dots,i_p} T_k \cdot p_{E_1}(B\cdot h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(B\cdot h)^{\otimes i_p} \tag{4.4.4}$$

is not constant in any of its variables $h_1, \dots, h_d$ *and*

$$\alpha_i := \frac{1}{2j} \quad \text{for } i \in \{\dim(E_1)+\cdots+\dim(E_{j-1})+1, \dots, \dim(E_1)+\cdots+\dim(E_j)\}. \tag{4.4.5}$$

2. *If* $p \geq 5$ *and if there exist orthogonal subspaces of* $\mathbb{R}^d$, $E_1, \dots, E_p$ *such that*

$$\mathbb{R}^d = E_1 \oplus \cdots \oplus E_p$$

and satisfying the following additional assumption

$$\begin{aligned} &\forall h \in \mathbb{R}^d, \ \forall (i_1,\dots,i_p) \in \{0,2,\dots,2p\}^p, \\ &\frac{i_1}{2}+\cdots+\frac{i_p}{2p} < 1 \implies T_{i_1+\cdots+i_p} \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} = 0, \end{aligned} \tag{4.4.6}$$

then (4.4.2) *stills holds true, as well as the uniform convergence on every compact set. Moreover, if* $B \in \mathcal{O}_d(\mathbb{R})$ *is an orthogonal transformation adapted to the previous decomposition, then* (4.4.3) *still hold true. However, depending on the function* f*, such subspaces do not necessarily exist.*

Remarks:

1. The limit (4.4.2) can be rewritten as:

$$\sum_{k=2}^{2p} \sum_{\substack{i_1,\dots,i_p \in \{0,\cdots,k\} \\ i_1+\cdots+i_p=k \\ \frac{i_1}{2}+\cdots+\frac{i_p}{2p}=1}} T_k \cdot \frac{p_{E_1}(h)^{\otimes i_1}}{i_1!} \otimes \cdots \otimes \frac{p_{E_p}(h)^{\otimes i_p}}{i_p!}.$$

2. For $p \in \{1,2,3,4\}$, we explicitly give the expression of the sum (4.4.2) and the p-tuples $(i_1,\dots,i_p)$ such that $\frac{i_1}{2}+\cdots+\frac{i_p}{2p}=1$, in (4.4.7), (4.4.8), (4.4.10) and (4.4.13) respectively.

3. For $p \in \{1,2,3,4\}$, we give in Algorithm 1 an explicit construction of the orthogonal subspaces $E_1, \dots, E_p$ as complementaries of annulation sets of some derivatives of f.

4. The case $p \geq 5$ is fundamentally different from the case $p \in \{1,2,3,4\}$. The strategy of proof developed for $p \in \{1,2,3,4\}$ fails if the assumption (4.4.6) is not satisfied. In 4.4.8 a counter-example is detailed. The case $p \geq 5$ yields fewer results than for $p \leq 4$, as the assumption (4.4.6) is strong.

5. For $p \geq 5$, such subspaces $E_1, \dots, E_p$ may also be obtained from Algorithm 1, however (4.4.6) is not necessarily true in this case.

The proof of Theorem 4.4.1 is given first individually for each $p \in \{1,2,3,4\}$, in Sections 4.4.3, 4.4.4, 4.4.6, 4.4.7 respectively. The proof for $p \geq 5$ is given in Section 4.4.8. The proof of the uniform convergence and of the fact that g is not constant is given in Section 4.4.9.

4.4.2 Review of the one dimensional case

We review the case $d = 1$, as it guides us for the proof in the case $d \geq 2$. The strategy is to find the first derivative $f^{(m)}(x^\star)$ which is non zero and then to choose $\alpha_1 = 1/m$.

Proposition 4.4.2. *Let $f : \mathbb{R} \to \mathbb{R}$ be $\mathcal{C}^p$ for some $p \in \mathbb{N}$ and let $x^\star$ be a strict polynomial local minimum of f. Then :*

1. *The order of the local minimum m is an even number and $f^{(m)}(x^\star) > 0$.*
2. $f(x^\star + h) \underset{h\to 0}{=} f(x^\star) + \frac{f^{(m)}(x^\star)}{m!}h^p + o(h^m)$

Then $\alpha_1 := 1/m$ is the solution of (4.3.2) and

$$\frac{1}{t}(f(x^\star + t^{1/m}h) - f(x^\star)) \underset{t\to 0}{\longrightarrow} \frac{f^{(m)}(x^\star)}{m!}h^m$$

which is a non-constant function of h, since $f^{(m)}(x^\star) \neq 0$. The direct proof using the Taylor formula is left to the reader.

4.4.3 Proof of Theorem 4.4.1 for $p = 1$

Let $f \in \mathscr{A}_1^\star(x^\star)$. The assumption that $x^\star$ is a strict polynomial local minimum at order 2 implies that $\nabla^2 f(x^\star)$ is positive definite. Let us denote $(\beta_i)_{1\leq i\leq d}$ its positive eigenvalues. By the spectral theorem, let us write $\nabla^2 f(x^\star) = B\mathrm{Diag}(\beta_{1:d})B^T$ for some $B \in \mathcal{O}_d(\mathbb{R})$. Then:

$$\frac{1}{t}(f(x^\star + t^{1/2}B \cdot h) - f(x^\star)) \underset{t\to 0}{\longrightarrow} \frac{1}{2}\sum_{i=1}^{d} \beta_i h_i^2. \tag{4.4.7}$$

Thus, a solution of (4.3.2) is $\alpha_1 = \cdots = \alpha_d = \frac{1}{2}$, B, and $g(h_1, \ldots, h_d) = \frac{1}{2}\sum_{i=1}^d \beta_i h_i^2$, which is a non-constant function of every $h_1, \ldots, h_d$, since for all i, β_i is positive.

In the following, our objective is to establish a similar result when $\nabla^2 f(x^\star)$ is not necessarily positive definite.

4.4.4 Proof of Theorem 4.4.1 for $p = 2$

Theorem 4.4.3. *Let $f \in \mathscr{A}_2(x^\star)$. Then there exist orthogonal subspaces E and F such that $\mathbb{R}^d = E \oplus F$, and that for all $h \in \mathbb{R}^d$:*

$$\begin{aligned}&\frac{1}{t}\left[f(x^\star + t^{1/2}p_{_E}(h) + t^{1/4}p_{_F}(h)) - f(x^\star)\right]\\ \underset{t\to 0}{\longrightarrow}\ &\frac{1}{2}\nabla^2 f(x^\star)\cdot p_{_E}(h)^{\otimes 2} + \frac{1}{2}\nabla^3 f(x^\star)\cdot p_{_E}(h)\otimes p_{_F}(h)^{\otimes 2} + \frac{1}{4!}\nabla^4 f(x^\star)\cdot p_{_F}(h)^{\otimes 4}.\end{aligned} \tag{4.4.8}$$

Moreover, if $f \in \mathscr{A}_2^\star(x^\star)$, then this is a solution to the problem (4.3.2)*, with $E_1 = E$, $E_2 = F$, α defined in* (4.4.5)*, B adapted to the previous decomposition and g defined in* (4.4.4)*.*

Remark: The set of 2-tuples (i_1, i_2) such that $\frac{i_1}{2} + \frac{i_2}{4} = 1$, are $(2, 0)$, $(1, 2)$ and $(0, 4)$, which gives the terms appearing in the sum in (4.4.2).

Proof. Let $F := \{h \in \mathbb{R}^d : \ \nabla^2 f(x^\star)\cdot h^{\otimes 2} = 0\}$. By the spectral theorem and since $\nabla^2 f(x^\star)$ is positive semi-definite, $F = \{h \in \mathbb{R}^d : \ \nabla^2 f(x^\star)\cdot h = 0^{\otimes 1}\}$ is a vector subspace of $\mathbb{R}^d$. Let E be the orthogonal complement of F in $\mathbb{R}^d$.

For $h \in \mathbb{R}^d$ we expand the left term of (4.4.8) using the Taylor formula up to order 4 and the multinomial formula (4.2.1), giving

$$\sum_{k=2}^{4} \frac{1}{k!} \sum_{\substack{i_1,i_2 \in \{0,\dots,k\} \\ i_1+i_2=k}} \binom{k}{i_1, i_2} t^{\frac{i_1}{2}+\frac{i_2}{4}-1} T_k \cdot p_{_E}(h)^{\otimes i_1} \otimes p_{_F}(h)^{\otimes i_2} + o(1).$$

The terms with coefficient t^a, $a > 0$, are $o(1)$ as $t \to 0$. By definition of F we have $\nabla^2 f(x^\star) \cdot p_{_F}(h) = 0^{\otimes 1}$, so we also have

$$\nabla^3 f(x^\star) \cdot p_{_F}(h)^{\otimes 3} = 0$$

by the local minimum condition. This yields the convergence stated in (4.4.8).

Moreover, if $x^\star$ is a local minimum of polynomial order 4, then by the local minimum condition, $\nabla^4 f(x^\star) > 0$ on F in the sense of (4.2.2). Moreover, since $\nabla^2 f(x^\star) > 0$ on E, then the limit is not constant in any $h_1, \dots, h_d$. □

Remark: The cross odd term is not necessarily null. For example, consider

$$\begin{array}{rccc} f: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & x^2 + y^4 + xy^2. \end{array}$$

Then f admits a global minimum at $x^\star = 0$ since $|xy^2| \leq \frac{1}{2}(x^2 + y^4)$. We have $E_1 = \mathbb{R}(1,0)$, $E_2 = \mathbb{R}(0,1)$ and for all $(x,y) \in \mathbb{R}^2$, $T_3 \cdot (xe_1) \otimes (ye_2)^{\otimes 2} = 2xy^2$ is not identically null.

4.4.5 Difficulties beyond the 4th order and Hilbert's 17th problem

If we do not assume as in the previous section that $\nabla^4 f(x^\star)$ is not positive on F, then we carry on the development of $f(x^\star + h)$ up to higher orders. A first idea is to consider $F_2 := \{h \in F : \nabla^4 f(x^\star) \cdot h^{\otimes 4} = 0\} \subseteq F$ and E_2 a complement subspace of F_2 in F, and to continue this process by induction as in Section 4.4.4. However, F_2 is not necessarily a subspace of F.

Indeed, let T be a symmetric tensor defined on $\mathbb{R}^{d'}$ of order $2k$ with $k \in \mathbb{N}$. As T is symmetric, there exist vectors $v^1, \dots, v^q \in \mathbb{R}^{d'}$, and scalars $\lambda_1, \dots, \lambda_q \in \mathbb{R}$ such that $T = \sum_i \lambda_i (v^i)^{\otimes 2k}$ (see [CGLM08], Lemma 4.2.), so

$$\forall h \in \mathbb{R}^{d'},\ T \cdot h^{\otimes 2k} = \sum_{i=1}^{q} \lambda_i (v^i)^{\otimes 2k} \cdot h^{\otimes 2k} = \sum_{i=1}^{q} \lambda_i \langle v^i, h \rangle^{2k}.$$

For $k = 2$ and $T = \nabla^{2k} f(x^\star)_{|_F}$, since $x^\star$ is a local minimum, we have, identifying F and $\mathbb{R}^{d'}$,

$$\forall h \in \mathbb{R}^{d'},\ T \cdot h^{\otimes 2k} \geq 0$$

Then, we could think it implies that for all i, $\lambda_i \geq 0$, and then

$$T \cdot h^{\otimes 2k} = 0 \implies \forall i,\ \langle v^i, h \rangle = 0$$

which would give a linear caracterization of $\{h \in \mathbb{R}^{d'} : T \cdot h^{\otimes 2k} = 0\}$ and in this case, F_2 would be a subspace of F. However this reasoning is not correct in general as we do not have necessarily that for all i, $\lambda_i \geq 0$.

We can build counter-examples as follows. Since T is a non-negative symmetric tensor, T can be seen as a non-negative homogeneous polynomial of degree $2k$ with d' variables. A counter-example at order $2k = 4$ is $T(X,Y,Z) = ((X-Y)(X-Z))^2$, which is a non-negative polynomial of order 4, but $\{T = 0\} = \{X = Y \text{ or } X = Z\}$, which is not a vector space.

Another counterexample given in [Mot67] at order $2k = 6$ is the following. We define

$$T(X, Y, Z) = Z^6 + X^4Y^2 + X^2Y^4 - 3X^2Y^2Z^2$$

By the arithmetic-geometric mean inequality and its equality case, T is non-negative and $T(x, y, z) = 0$ if and only if $z^6 = x^4y^2 = x^2y^4$, so that

$$\{T = 0\} = \mathbb{R}\begin{pmatrix}1\\1\\1\end{pmatrix} \cup \mathbb{R}\begin{pmatrix}-1\\1\\1\end{pmatrix} \cup \mathbb{R}\begin{pmatrix}1\\-1\\1\end{pmatrix} \cup \mathbb{R}\begin{pmatrix}1\\1\\-1\end{pmatrix}.$$

Hence, $\{T = 0\}$ is not a subspace of $\mathbb{R}^3$. In particular T cannot be written as $\sum_i \lambda_i (v^i)^{\otimes 2k}$ with $\lambda_i \geq 0$.

In fact, this problem is linked with the Hilbert's seventeenth problem that we recall below.

Problem 1 (Hilbert's seventeeth problem). Let P be a non-negative polynomial with d' variables, homogeneous of even degree $2k$. Find polynomials $P_1, \ldots, P_r$ with d' variables, homogeneous of degree k, such that $P = \sum_{i=1}^r P_i^2$

Hilbert proved in 1888 [Hil88] that there does not always exist a solution. In general $\{T = 0\}$ is not even a submanifold of $\mathbb{R}^{d'}$. Indeed, taking $T : h \mapsto \nabla^{2k} f(x^\star) \cdot h^{\otimes 2k}$, we have $\partial_h T \cdot h = 2k\nabla^{2k} f(x^\star) \cdot h^{\otimes 2k-1}$ is not surjective in $h = 0$, so the surjectivity condition for $\{T = 0\}$ to be a submanifold is not fulfilled.

4.4.6 Proof of Theorem 4.4.1 for $p = 3$

We slightly change our strategy of proof developed in Section 4.4.4. For $k \geq 2$, we define F_k recursively as

$$F_k := \{h \in F_{k-1} : \ \forall h' \in F_{k-1}, \ \nabla^{2k} f(x^\star) \cdot h \otimes h'^{\otimes 2k-1} = 0\}, \tag{4.4.9}$$

instead of $\{h \in F_{k-1} : \ \nabla^{2k} f(x^\star) \cdot h^{\otimes 2k} = 0\}$. Then, by construction, F_k is a vector subspace of $\mathbb{R}^d$.

Theorem 4.4.4. *Let $f \in \mathscr{A}_3(x^\star)$. Then there exist orthogonal subspaces of $\mathbb{R}^d$, E_1, E_2 and F_2, such that*

$$\mathbb{R}^d = E_1 \oplus E_2 \oplus F_2,$$

and such that for all $h \in \mathbb{R}^d$,

$$\begin{aligned}&\frac{1}{t}\Big[f(x^\star + t^{1/2}p_{E_1}(h) + t^{1/4}p_{E_2}(h) + t^{1/6}p_{F_2}(h)) - f(x^\star)\Big] \\ &\xrightarrow[t\to 0]{} \frac{1}{2}\nabla^2 f(x^\star) \cdot p_{E_1}(h)^{\otimes 2} + \frac{1}{2}\nabla^3 f(x^\star) \cdot p_{E_1}(h) \otimes p_{E_2}(h)^{\otimes 2} + \frac{1}{4!}\nabla^4 f(x^\star) \cdot p_{E_2}(h)^{\otimes 4} \\ &+ \frac{4}{4!}\nabla^4 f(x^\star) \cdot p_{E_1}(h) \otimes p_{F_2}(h)^{\otimes 3} + \frac{10}{5!}\nabla^5 f(x^\star) \cdot p_{E_2}(h)^{\otimes 2} \otimes p_{F_2}(h)^{\otimes 3} + \frac{1}{6!}\nabla^6 f(x^\star) \cdot p_{F_2}(h)^{\otimes 6}.\end{aligned} \tag{4.4.10}$$

Moreover, if $f \in \mathscr{A}_3^\star(x^\star)$, then this is a solution to the problem (4.3.2)*, with $E_3 = F_2$, α defined in* (4.4.5)*, B adapted to the previous decomposition and g defined in* (4.4.4)*.*

Remark: The set of 3-tuples (i_1, i_2, i_3) such that $\frac{i_1}{2} + \frac{i_2}{4} + \frac{i_3}{6} = 1$, are $(2, 0, 0)$, $(1, 2, 0)$, $(0, 4, 0)$, $(1, 0, 3)$, $(0, 2, 3)$, $(0, 0, 6)$, which gives the terms appearing in (4.4.2).

Proof. We consider the subspace

$$F_1 := \{h \in \mathbb{R}^d : \ T_2 \cdot h^{\otimes 2} = 0\} = \{h \in \mathbb{R}^d : \ T_2 \cdot h = 0^{\otimes 1}\},$$

since $T_2 \geq 0$. Then, let E_1 be the orthogonal complement of F_1 in $\mathbb{R}^d$ and consider the vector subspace of F_1 defined by

$$F_2 = \{h \in F_1 : \ \forall h' \in F_1, \ T_4 \cdot h \otimes h'^{\otimes 3} = 0\}.$$

Let E_2 be the orthogonal complement of F_2 in F_1. Then we have

$$\mathbb{R}^d = E_1 \oplus F_1 = E_1 \oplus E_2 \oplus F_2.$$

For $h \in \mathbb{R}^d$ we expand the left term of (4.4.10) using the Taylor formula up to order 6 and the multinomial formula (4.2.1), giving

$$\sum_{k=2}^{6} \frac{1}{k!} \sum_{\substack{i_1,i_2,i_3 \in \{0,\dots,k\} \\ i_1+i_2+i_3=k}} \binom{k}{i_1, i_2, i_3} t^{\frac{i_1}{2}+\frac{i_2}{4}+\frac{i_3}{6}-1} T_k \cdot p_{E_1}(h)^{\otimes i_1} \otimes p_{E_2}(h)^{\otimes i_2} \otimes p_{F_2}(h)^{\otimes i_3} + o(1),$$

and we prove the convergence stated in (4.4.10).

All the terms with coefficient t^a where $a > 0$ are $o(1)$ as $t \to 0$.

Order 2: we have $T_2 \cdot p_{E_2}(h) = 0^{\otimes 1}$ and $T_2 \cdot p_{F_2}(h) = 0^{\otimes 1}$ so the only term for $k = 2$ is $\frac{1}{2} T_2 \cdot p_{E_1}(h)^{\otimes 2}$.

Order 3: ▷ Since $x^\star$ is a local minimum and $T_2 \cdot p_{F_1}(h)^{\otimes 2} = 0$, we have $T_3 \cdot p_{F_1}(h)^{\otimes 3} = 0$. Then, using property Proposition 4.8.1, if the factor $p_{E_1}(h)$ does not appear as an argument in T_3, then the corresponding term is zero.

▷ Let us prove that

$$T_3 \cdot p_{E_1}(h) \otimes p_{F_2}(h)^{\otimes 2} = 0.$$

Using Theorem 4.4.3 with $E = E_1$, $F = E_2 \oplus F_2$, we have in particular that for all $h \in \mathbb{R}^d$,

$$\frac{1}{2} T_2 \cdot p_E(h)^{\otimes 2} + \frac{1}{2} T_3 \cdot p_E(h) \otimes p_F(h)^{\otimes 2} + \frac{1}{4!} T_4 \cdot p_F(h)^{\otimes 4} \geq 0. \tag{4.4.11}$$

Then taking $h \in E_1 \oplus F_2$ so that $h = p_{E_1}(h) + p_{F_2}(h)$ and with

$$\Big[T_4 \cdot p_{F_2}(h)\Big]_{|F_1} \equiv 0^{\otimes 3}, \tag{4.4.12}$$

we may rewrite (4.4.11) as

$$\frac{1}{2} T_2 \cdot p_{E_1}(h)^{\otimes 2} + \frac{1}{2} T_3 \cdot p_{E_1}(h) \otimes p_{F_2}(h)^{\otimes 2} \geq 0.$$

Now, considering $h' = \lambda h$, we have that for all $\lambda \in \mathbb{R}$,

$$\lambda^2 \left(\frac{1}{2} T_2 \cdot p_{E_1}(h)^{\otimes 2} + \frac{\lambda}{2} T_3 \cdot p_{E_1}(h) \otimes p_{F_2}(h)^{\otimes 2} \right) \geq 0,$$

so that necessarily $T_3 \cdot p_{E_1}(h) \otimes p_{F_2}(h)^{\otimes 2} = 0$.

▷ Let us prove that

$$T_3 \cdot p_{E_1}(h) \otimes p_{E_2}(h) \otimes p_{F_2}(h) = 0.$$

We use again (4.4.11), with $p_F(h) = p_{E_2}(h) + p_{F_2}(h)$, so that

$$\frac{1}{2}T_2 \cdot p_{E_1}(h)^{\otimes 2} + \frac{1}{2}T_3 \cdot p_{E_1}(h) \otimes \Big(p_{E_2}(h) + p_{F_2}(h)\Big)^{\otimes 2} + \frac{1}{4!}T_4 \cdot \Big(p_{E_2}(h) + p_{F_2}(h)\Big)^{\otimes 4} \geq 0.$$

But using (4.4.12) and that $T_3 \cdot p_{E_1}(h) \otimes p_{F_2}(h)^{\otimes 2} = 0$, we obtain

$$\frac{1}{2}T_2 \cdot p_{E_1}(h)^{\otimes 2} + \frac{1}{2}T_3 \cdot p_{E_1}(h) \otimes p_{E_2}(h)^{\otimes 2} + T_3 \cdot p_{E_1}(h) \otimes p_{E_2}(h) \otimes p_{F_2}(h) + \frac{1}{4!}T_4 \cdot p_{E_2}(h)^{\otimes 4} \geq 0.$$

Now, considering $h' = p_{E_1}(h) + p_{E_2}(h) + \lambda p_{F_2}(h)$, we have that for all $\lambda \in \mathbb{R}$,

$$\frac{1}{2}T_2 \cdot p_{E_1}(h)^{\otimes 2} + \frac{1}{2}T_3 \cdot p_{E_1}(h) \otimes p_{E_2}(h)^{\otimes 2} + \lambda T_3 \cdot p_{E_1}(h) \otimes p_{E_2}(h) \otimes p_{F_2}(h) + \frac{1}{4!}T_4 \cdot p_{E_2}(h)^{\otimes 4} \geq 0,$$

so necessarily $T_3 \cdot p_{E_1}(h) \otimes p_{E_2}(h) \otimes p_{F_2}(h) = 0$.

▷ The last remaining term for $k = 3$ is $\frac{1}{2}T_3 \cdot p_{E_1}(h) \otimes p_{E_2}(h)^{\otimes 2}$.

Order 4: If the factor $p_{E_1}(h)$ does not appear and if the factor $p_{F_2}(h)$ appears at least once, then using (4.4.12) the corresponding term is zero. If $p_{E_1}(h)$ appears, the only term with a non-positive exponent of t is $\frac{4}{4!}T_4 \cdot p_{E_1}(h) \otimes p_{F_2}(h)^{\otimes 3}$. So the only terms for $k = 4$ are $\frac{1}{4!}T_4 \cdot p_{E_2}(h)^{\otimes 4}$ and $\frac{4}{4!}T_4 \cdot p_{E_1}(h) \otimes p_{F_2}(h)^{\otimes 3}$.

Order 5: ▷ The terms where $p_{E_1}(h)$ appears at least once have a coefficient t^a with $a > 0$ so are $o(1)$ when $t \to 0$.

▷ We have $T_2 \cdot p_{F_2}(h)^{\otimes 2} = 0$, $T_3 \cdot p_{F_2}(h)^{\otimes 3} = 0$, $T_4 \cdot p_{F_2}(h)^{\otimes 4} = 0$ and since $x^\star$ is a local minimum, we have

$$T_5 \cdot p_{F_2}(h)^{\otimes 5} = 0.$$

▷ Let us prove that

$$T_5 \cdot p_{E_2}(h) \otimes p_{F_2}(h)^{\otimes 4} = 0.$$

Let $h \in \mathbb{R}^d$. We have

$$\frac{1}{t^{11/12}}\Big[f(x^\star + t^{1/4}p_{E_2}(h) + t^{1/6}p_{F_2}(h)) - f(x^\star)\Big] \underset{t\to 0}{\longrightarrow} \frac{1}{4!}T_5 \cdot p_{E_2}(h) \otimes p_{F_2}(h)^{\otimes 4} \geq 0.$$

Hence, considering $h' = \lambda h$, we have for every $\lambda \in \mathbb{R}$,

$$\lambda^5 T_5 \cdot p_{E_2}(h) \otimes p_{F_2}(h)^{\otimes 4} \geq 0,$$

which yields the desired result.

▷ The only remaining term for $p = 5$ is

$$\frac{10}{5!}T_5 \cdot p_{E_2}(h)^{\otimes 2} \otimes p_{F_2}(h)^{\otimes 3}.$$

Order 6: The only term for $k = 6$ is $\frac{1}{6!}T_6 \cdot p_{F_2}(h)^{\otimes 6}$; the other terms have a coefficient t^a with $a > 0$, so are $o(1)$ when $t \to 0$. □

Remark : As in Theorem 4.4.3 and the remark that follows, the remaining odd cross-terms cannot be proved to be zero using the same method of proof, and may be actually not zero. For example, consider:

$$\begin{array}{rccc} f: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & x^4 + y^6 + x^2y^3, \end{array}$$

which satisfies $h \mapsto \nabla^5 f(x^\star) \cdot p_{E_2}(h)^{\otimes 2} \otimes p_{F_2}(h)^{\otimes 3} \not\equiv 0$.

Order 2	$(2,0,0,0)$
Order 3	$(2,1,0,0)$
Order 4	$(0,4,0,0),\ (1,1,0,2),\ (1,0,3,0)$
Order 5	$(1,0,0,4),\ (0,2,3,0),\ (0,3,0,2)$
Order 6	$(0,1,3,2),\ (0,2,0,4),\ (0,0,6,0)$
Order 7	$(0,1,0,6),\ (0,0,3,4)$
Order 8	$(0,0,0,8)$

Table 4.1: Terms expressed as 4-tuples in the development (4.4.13)

<table>
<tr><td>E_1</td><td colspan="3">F_1</td></tr>
<tr><td rowspan="5">$T_2 \geq 0$</td><td colspan="3">$T_2 = 0$</td></tr>
<tr><td>E_2</td><td colspan="2">F_2</td></tr>
<tr><td rowspan="3">$T_4 \geq 0$</td><td colspan="2">$T_4 = 0$</td></tr>
<tr><td>E_3</td><td>F_3</td></tr>
<tr><td>$T_6 \geq 0$</td><td>$T_6 = 0$</td></tr>
</table>

Table 4.2: Illustration of the subspaces

4.4.7 Proof of Theorem 4.4.1 for $p = 4$

Theorem 4.4.5. *Let $f \in \mathscr{A}_4(x^\star)$. Then there exist orthogonal subspaces of $\mathbb{R}^d$, E_1, E_2, E_3 and F_3 such that*

$$\mathbb{R}^d = E_1 \oplus E_2 \oplus E_3 \oplus F_3,$$

and for all $h \in \mathbb{R}^d$,

$$\begin{aligned}&\frac{1}{t}\left[f(x^\star + t^{1/2}p_{E_1}(h) + t^{1/4}p_{E_2}(h) + t^{1/6}p_{E_3}(h) + t^{1/8}p_{F_3}(h)) - f(x^\star)\right]\\ \xrightarrow[t\to 0]{} &\sum_{k=2}^{8}\frac{1}{k!}\sum_{\substack{i_1,\dots,i_4\in\{0,\dots,k\}\\ i_1+\cdots+i_4=k}}\binom{k}{i_1,\dots,i_4}T_k\cdot p_{E_1}(h)^{\otimes i_1}\otimes p_{E_2}(h)^{\otimes i_2}\otimes p_{E_3}(h)^{\otimes i_3}\otimes p_{F_3}(h)^{\otimes i_4}.\end{aligned} \tag{4.4.13}$$

These terms are summarized as tuples $(i_1,\dots,i_4)$ in Table 4.1. Moreover, if $f \in \mathscr{A}_4^\star(x^\star)$, then this is a solution to (4.3.2), with $E_4 = F_3$, α defined in (4.4.5), B adapted to the previous decomposition and g defined in (4.4.4).

Proof. As before, we define the subspaces $F_0 := \mathbb{R}^d$ and by induction:

$$F_k = \left\{h \in F_{k-1} :\ \forall h' \in F_{k-1},\ T_{2k}\cdot h\otimes h'^{\otimes 3} = 0\right\}$$

for $k = 1,2,3$. We define E_k as the orthogonal complement of F_k in F_{k-1} for $k = 1,2,3$, so that

$$\mathbb{R}^d = E_1 \oplus E_2 \oplus E_3 \oplus F_3.$$

Then we apply a Taylor expansion up to order 8 to the left side of (4.4.13) and the multinomial formula (4.2.1), which reads

$$\sum_{k=2}^{8}\frac{1}{k!}\sum_{\substack{i_1,\dots,i_4\in\{0,\dots,k\}\\ i_1+\cdots+i_4=k}}\binom{k}{i_1,\dots,i_4}t^{\frac{i_1}{2}+\cdots+\frac{i_4}{8}-1}T_k\cdot p_{E_1}(h)^{\otimes i_1}\otimes p_{E_2}(h)^{\otimes i_2}\otimes p_{E_3}(h)^{\otimes i_3}\otimes p_{F_3}(h)^{\otimes i_4}+o(1).$$

▷ If $\frac{i_1}{2}+\cdots+\frac{i_4}{8}>1$ then the corresponding term is in $o(1)$ when $t\to 0$.
▷ If $\frac{i_1}{2}+\cdots+\frac{i_4}{8}<1$ then the corresponding term diverges when $t\to 0$, so we need to prove that actually

$$T_k\cdot p_{E_1}(h)^{\otimes i_1}\otimes p_{E_2}(h)^{\otimes i_2}\otimes p_{E_3}(h)^{\otimes i_3}\otimes p_{F_3}(h)^{\otimes i_4}=0. \tag{4.4.14}$$

– If $\frac{i_1}{2}+\frac{i_2}{4}+\frac{i_3}{6}+\frac{i_4}{8}<1$ but if we also have $\frac{i_1}{2}+\frac{i_2}{4}+\frac{i_3}{6}+\frac{i_4}{6}<1$, then by applying the property at the order 6 (Theorem 4.4.4) with the 3-tuple (i_1,i_2,i_3+i_4), we get (4.4.14).

– So we only need to consider 4-tuples such that $\frac{i_1}{2}+\frac{i_2}{4}+\frac{i_3}{6}+\frac{i_4}{8}<1$ and $\frac{i_1}{2}+\frac{i_2}{4}+\frac{i_3}{6}+\frac{i_4}{6}\geq 1$. We can remove all the terms which are null by the definitions of the subspaces $E_1,\ E_2,\ E_3,\ F_3$. The remaining terms are:

For $k=4$: $\frac{t^{21/24}}{6}T_4\cdot p_{E_1}(h)\otimes p_{F_3}(h)^{\otimes 3}$, $\frac{t^{11/12}}{2}T_4\cdot p_{E_1}(h)\otimes p_{E_3}(h)\otimes p_{F_3}(h)^{\otimes 2}$, $\frac{t^{23/24}}{2}T_4\cdot p_{E_1}(h)\otimes p_{E_3}(h)^{\otimes 2}\otimes p_{F_3}(h)$.

For $k=5$: $\frac{t^{21/24}}{12}T_5\cdot p_{E_2}(h)^{\otimes 2}\otimes p_{F_3}(h)^{\otimes 3}$, $\frac{t^{11/12}}{4}T_5\cdot p_{E_2}(h)^{\otimes 2}\otimes p_{E_3}(h)\otimes p_{F_3}(h)^{\otimes 2}$, $\frac{t^{23/24}}{4}T_5\cdot p_{E_2}(h)^{\otimes 2}\otimes p_{E_3}(h)^{\otimes 2}\otimes p_{F_3}(h)$.

For $k=6$: $\frac{t^{21/24}}{5!}T_6\cdot p_{E_2}(h)\otimes p_{F_3}(h)^{\otimes 5}$, $\frac{t^{11/12}}{4!}T_6\cdot p_{E_2}(h)\otimes p_{E_3}(h)\otimes p_{F_3}(h)^{\otimes 4}$, $\frac{t^{23/24}}{12}T_6\cdot p_{E_2}(h)\otimes p_{E_3}(h)^{\otimes 2}\otimes p_{F_3}(h)^{\otimes 3}$.

First, we note that

$$\begin{aligned}&\frac{1}{t^{21/24}}\left[f(x^\star+t^{1/2}p_{E_1}(h)+t^{1/4}p_{E_2}(h)+t^{1/6}p_{E_3}(h)+t^{1/8}p_{F_3}(h))-f(x^\star)\right]\\ \xrightarrow[t\to 0]{}\ &\frac{1}{6}T_4\cdot p_{E_1}(h)\otimes p_{F_3}(h)^{\otimes 3}+\frac{1}{12}T_5\cdot p_{E_2}(h)^{\otimes 2}\otimes p_{F_3}(h)^{\otimes 3}+\frac{1}{5!}T_6\cdot p_{E_2}(h)\otimes p_{F_3}(h)^{\otimes 5}\geq 0.\end{aligned}$$

Then, considering $h'=\lambda p_{E_1}(h)+p_{E_2}(h)+p_{E_3}(h)+p_{F_3}(h)$, we have that for all $\lambda\in\mathbb{R}$,

$$\frac{\lambda}{6}T_4\cdot p_{E_1}(h)\otimes p_{F_3}(h)^{\otimes 3}+\frac{1}{12}T_5\cdot p_{E_2}(h)^{\otimes 2}\otimes p_{F_3}(h)^{\otimes 3}+\frac{1}{5!}T_6\cdot p_{E_2}(h)\otimes p_{F_3}(h)^{\otimes 5}\geq 0,$$

so necessarily

$$T_4\cdot p_{E_1}(h)\otimes p_{F_3}(h)^{\otimes 3}=0.$$

Then, considering $h'=p_{E_2}(h)+\lambda p_{F_3}(h)$ for $\lambda\in\mathbb{R}$, we get successively that the two other terms are null.

Likewise, we prove successively that the terms in $t^{11/12}$ are null, and then that the terms in $t^{23/24}$ are null. This yields the convergence stated in (4.4.13). □

4.4.8 Counter-example and proof of Theorem 4.4.1 with $p\geq 5$ under the hypothesis (4.4.6)

Algorithm 1 may fail to yield such expansion of f for orders no lower than 10 if the hypothesis (4.4.6) is not fulfilled. Indeed for $p\geq 5$, there exist p-tuples $(i_1,\ldots,i_p)$ such that $\frac{i_1}{2}+\cdots+\frac{i_p}{2p}<1$ and $i_1,\ \ldots,\ i_p$ are all even. Such tuples do not appear at orders 8 and lower, but they do appear at orders 10 and higher, for example $(0,2,0,0,4)$ for $k=6$. In such a case, we cannot use the positiveness argument to prove that the corresponding term $T_k\cdot p_{E_1}(h)^{\otimes i_1}\otimes\cdots\otimes p_{E_p}(h)^{\otimes i_p}$ is zero, and in fact, it may be not zero.

Let us give a counter example. Consider

$$\begin{array}{rccc} f: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & x^4+y^{10}+x^2y^4. \end{array}$$

Then $f\in\mathscr{A}_5^\star(0)$ and we have $E_1=\{0\}$, $E_2=\mathbb{R}\cdot(1,0)$, $E_3=\{0\}$, $E_4=\{0\}$, $F_4=\mathbb{R}\cdot(0,1)$. But

$$\frac{1}{t}f(t^{1/4},t^{1/10})=\frac{1}{t}\left(t+t+t^{9/10}\right)$$

goes to $+\infty$ when $t \to 0$.

Now, let us give the proof of Theorem 4.4.1 for $p \geq 5$. In this proof, we assume that the subspaces $E_1, \ldots, E_p$ given in Algorithm 1 satisfy the hypothesis (4.4.6).

Proof. We develop (4.4.1), which reads:

$$\sum_{k=2}^{2p} \frac{1}{k!} \sum_{\substack{i_1,\ldots,i_p \in \{0,\ldots,k\} \\ i_1+\cdots+i_p=k}} \binom{k}{i_1,\ldots,i_p} t^{\frac{i_1}{2}+\cdots+\frac{i_p}{2p}-1} T_k \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} + o(1) =: S.$$

The terms such that $\frac{i_1}{2} + \cdots + \frac{i_p}{2p} < 1$ may diverge when $t \to 0$, so let us prove that they are in fact null. Let

$$\alpha := \inf \Big\{ \frac{i_1}{2} + \cdots + \frac{i_p}{2p} : h \longmapsto \sum_{k=2}^{2p} \frac{1}{k!} \sum_{\substack{i_1,\ldots,i_p \in \{0,\ldots,k\} \\ i_1+\cdots+i_p=k \\ \frac{i_1}{2}+\cdots+\frac{i_p}{2p}=\alpha}} \binom{k}{i_1,\ldots,i_p} T_k \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} \not\equiv 0 \Big\},$$

and assume by contradiction that $\alpha < 1$. Then we have for all $h \in \mathbb{R}^d$:

$$t^{1-\alpha} S \underset{t\to 0}{\longrightarrow} \left(\sum_{k=2}^{2p} \frac{1}{k!} \sum_{\substack{i_1,\ldots,i_p \in \{0,\ldots,k\} \\ i_1+\cdots+i_p=k \\ \frac{i_1}{2}+\cdots+\frac{i_p}{2p}=\alpha}} \binom{k}{i_1,\ldots,i_p} T_k \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} \right) \geq 0,$$

by the local minimum property. Then, considering $h' = \lambda_1 p_{E_1}(h) + \cdots + \lambda_p p_{E_p}(h)$, we have, for all $h \in \mathbb{R}^d$ and $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$,

$$\sum_{k=2}^{2p} \frac{1}{k!} \sum_{\substack{i_1,\ldots,i_p \in \{0,\ldots,k\} \\ i_1+\cdots+i_p=k \\ \frac{i_1}{2}+\cdots+\frac{i_p}{2p}=\alpha}} \lambda_1^{i_1} \ldots \lambda_p^{i_p} \binom{k}{i_1,\ldots,i_p} T_k \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} \geq 0. \tag{4.4.15}$$

Now, we fix $h \in \mathbb{R}^d$ such that the polynomial in (4.4.15) in the variables $\lambda_1, \ldots, \lambda_p$ is not identically zero, and we consider $k_{\max}$ its highest homogeneous degree, so that we have

$$\sum_{\substack{i_1,\ldots,i_p \in \{0,\ldots,k_{\max}\} \\ i_1+\cdots+i_p=k_{\max} \\ \frac{i_1}{2}+\cdots+\frac{i_p}{2p}=\alpha}} \lambda_1^{i_1} \ldots \lambda_p^{i_p} \binom{k_{\max}}{i_1,\ldots,i_p} T_{k_{\max}} \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} \geq 0.$$

If $k_{\max}$ is odd, this yields a contradiction, taking $\lambda_1 = \cdots = \lambda_p =: \lambda \to \pm\infty$. If $k_{\max}$ is even, we consider the index l_1 such that $i_{l_1} =: a_1$ is maximal and the coefficients in the above sum with $i_{l_1} = a_1$ are not all zero. Then fixing all the λ_l for $l \neq l_1$ and taking $\lambda_{l_1} \to \infty$, we have

$$\sum_{\substack{i_1,\ldots,i_p \in \{0,\ldots,k_{\max}\} \\ i_1+\cdots+i_p=k_{\max} \\ \frac{i_1}{2}+\cdots+\frac{i_p}{2p}=\alpha \\ i_{l_1}=a_1}} \lambda_1^{i_1} \ldots \lambda_p^{i_p} \binom{k_{\max}}{i_1,\cdots,i_p} T_{k_{\max}} \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} \geq 0.$$

Thus, if a_1 is odd, this yields a contradiction. If a_1 is even, we carry on this process by induction : knowing $l_1,\ \dots,\ l_r$, we choose the index l_{r+1} such that $l_{r+1} \notin \{l_1, \dots, l_r\}$, the corresponding term

$$\sum_{\substack{i_1,\dots,i_p \in \{0,\dots,k_{\max}\} \\ i_1+\cdots+i_p = k_{\max} \\ \frac{i_1}{2}+\cdots+\frac{i_p}{2p}=\alpha \\ i_{l_1}=a_1,\dots,i_{l_{r+1}}=a_{r+1}}} \lambda_1^{i_1} \dots \lambda_p^{i_p} \binom{k_{\max}}{i_1, \dots, i_p} T_{k_{\max}} \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p}$$

is not identically null and such that $i_{l_{r+1}} =: a_{r+1}$ is maximal. Necessarily, a_{r+1} is even. In the end we will find a non-zero term whose exponents i_ℓ are all even which contradicts assumption (4.4.6). □

4.4.9 Proofs of the uniform convergence and of the non-constant property

In this section we prove the additional properties claimed in Theorem 4.4.1 : the uniform convergence with respect to h on every compact set and the fact that the function g is not constant in any of its variables $h_1,\ \dots,\ h_d$.

Proof. First, let us prove that the convergence is uniform with respect to h on every compact set. Let $\varepsilon > 0$ and let $R > 0$. By the Taylor formula at order $2p$, there exists $\delta > 0$ such that for $||h|| < \delta$,

$$\left| f(x^\star + h) - f(x^\star) - \sum_{k=2}^{2p} \frac{1}{k!} \sum_{i_1+\cdots+i_p=k} \binom{k}{i_1, \dots, i_p} T_k \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} \right| \leq \varepsilon ||h||^{2p}.$$

Now, let us consider $t \to 0$ and $h \in \mathbb{R}^d$ with $||h|| \leq R$. Then we have:

$$\forall t \leq \max\left(1, \left(\frac{\delta}{R}\right)^{1/(2p)}\right),\ ||t^{1/2} p_{E_1}(h) + \cdots + t^{1/(2p)} p_{E_p}(h)|| \leq \delta,$$

so that

$$\left| \frac{1}{t} \left[\ f(x^\star + t^{1/2} p_{E_1}(h) + \cdots + t^{1/(2p)} p_{E_p}(h)) - f(x^\star) \right] - \sum_{k=2}^{2p} \frac{1}{k!} \sum_{i_1+\cdots+i_p=k} \binom{k}{i_1, \dots, i_p} \right.$$
$$\left. \cdot t^{\frac{i_1}{2}+\cdots+\frac{i_p}{2p}-1} T_k \cdot p_{E_1}(h)^{\otimes i_1} \otimes \cdots \otimes p_{E_p}(h)^{\otimes i_p} \right| \leq \frac{\varepsilon}{t} ||t^{1/2} p_{E_1}(h) + \cdots + t^{1/(2p)} p_{E_p}(h)||^{2p}.$$

We proved or assumed that the terms such that $\frac{i_1}{2} + \cdots + \frac{i_p}{2p} < 1$ are zero. We denote by $g_1(h)$ the sum in the last equation with the terms such that $\frac{i_1}{2} + \cdots + \frac{i_p}{2p} = 1$ and by $g_2(h)$ the sum with the terms such that $\frac{i_1}{2} + \cdots + \frac{i_p}{2p} > 1$. We also define a as the smallest exponent of t appearing in $g_2(h)$:

$$a := \min\left\{ \frac{i_1}{2} + \cdots + \frac{i_p}{2p}\ :\ i_1, \dots, i_p \in \{0, \dots, 2p\},\ i_1 + \cdots + i_p \leq 2p,\ \frac{i_1}{2} + \cdots + \frac{i_p}{2p} > 1 \right\} > 1.$$

So that:

$$\left| \frac{1}{t} \left[f(x^\star + t^{1/2} p_{E_1}(h) + \cdots + t^{1/(2p)} p_{E_p}(h)) - f(x^\star) \right] - g_1(h) \right| \tag{4.4.16}$$
$$\leq t^{a-1} |g_2(h)| + \frac{\varepsilon}{t} ||t^{1/2} p_{E_1}(h) + \cdots + t^{1/(2p)} p_{E_p}(h)||^{2p}.$$

We remark that $h \mapsto g_2(h)$ is a polynomial function so is bounded on every compact set. We also have:

$$\frac{\varepsilon}{t}||t^{1/2}p_{E_1}(h) + \cdots + t^{1/(2p)}p_{E_p}(h)||^{2p} \leq \frac{\varepsilon(t^{1/(2p)})^{2p}}{t}||h||^{2p} = \varepsilon||h||^{2p}.$$

So (4.4.16) converges to 0 as $t \to 0$, uniformly with respect to h on every compact set.

Now let us assume that $f \in \mathscr{A}_p^\star(x^\star)$; we prove that the function g defined in (4.3.6) is not constant in any of its variables in the sense of (4.3.3). Let $B \in \mathcal{O}_d(\mathbb{R})$ adapted to the decomposition $\mathbb{R}^d = E_1 \oplus \cdots \oplus E_p$. We have:

$$\frac{1}{t}\left[f\left(x^\star + B \cdot (t^\alpha * h)\right) - f(x^\star)\right] \underset{t\to 0}{\longrightarrow} g(h).$$

Let $i \in \{1, \ldots, p\}$ and k such that $v_i := B \cdot e_i \in E_k$. Let us assume by contradiction that g does not depend on the i^{th} coordinate. Considering the expression of g in (4.3.6) and setting all the variables outside E_k to 0, we have:

$$\forall h \in E_k,\ \lambda \in \mathbb{R} \mapsto T_{2k} \cdot (h + \lambda v_i)^{\otimes 2k}$$

is constant. Then applying (4.2.1), we have:

$$\forall h \in E_k,\ T_{2k} \cdot v_i \otimes h^{\otimes 2k-1} = 0.$$

Moreover, for $h \in F_{k-1}$, let us write $h = h' + h''$ where $h' \in E_k$ and $h'' \in F_k$, so that

$$T_{2k} \cdot v_i \otimes h^{\otimes 2k-1} = T_{2k} \cdot v_i \otimes h'^{\otimes 2k-1} = 0,$$

where we used that

$$\forall h^{(3)} \in F_{k-1},\ T_{2k} \cdot h'' \otimes \left(h^{(3)}\right)^{\otimes 2k-1} = 0$$

following (4.4.9), and Proposition 4.8.1. Considering the definition of E_k as the orthogonal complement of F_k, which is defined in (4.4.9), the last equation contradicts that $v_i \in E_k$. □

4.4.10 Non coercive case

The function g we obtain in Algorithm 1 is a non-negative polynomial function which is constant in none of its variables. However, this does not always guarantee that $e^{-g} \in L^1(\mathbb{R}^d)$, or even that g is coercive. Indeed, g can be null on an unbounded continuous polynomial curve, while the polynomial degree of the minimum $x^\star$ of f is higher than the degree of g in these variables. For example, let us consider

$$\begin{aligned} f\colon\ \mathbb{R}^2 &\to \mathbb{R} \\ (x, y) &\mapsto (x - y^2)^2 + x^6. \end{aligned} \tag{4.4.17}$$

Then $f \in \mathscr{A}_3^\star(0)$ and using Algorithm 1, we get

$$g(x, y) = (x - y^2)^2,$$

which does not satisfy $e^{-g} \in L^1(\mathbb{R}^d)$. In fact this case is highly degenerate, as, with

$$f_\varepsilon(x, y) := f(x, y) + \varepsilon x y^2 = x^2 + y^4 - (2 - \varepsilon)xy^2 + x^6,$$

we have that $g_\varepsilon(x, y) = x^2 + y^4 - (2 - \varepsilon)xy^2$ satisfies $e^{-g_\varepsilon} \in L^1(\mathbb{R}^d)$ for every $\varepsilon \in (0, 4)$ and that $x^\star$ is not the global minimum of f_ε for every $\varepsilon \in (-\infty, 0) \cup (4, \infty)$.

We now prove that instead of assuming $e^{-g} \in L^1(\mathbb{R}^d)$, we can only assume that g is coercive, which is justified in the following proposition. More specific conditions for g to be coercive can be found in [BS15] and [BS19].

Proposition 4.4.6. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be the polynomial function obtained from Algorithm 1. If g is coercive, then $e^{-g} \in L^1(\mathbb{R}^d)$.*

Proof. Let

$$A_k := \text{Span}\,(e_i : \ i \in \{\dim(E_1) + \cdots + \dim(E_{k-1}) + 1, \ldots, \dim(E_1) + \cdots + \dim(E_k)\})$$

for $k \in \{1, \ldots, p\}$. By construction of g, note that for all $t \in [0, +\infty)$,

$$g\left(\sum_{k=1}^{p} t^{1/2k} p_{A_k}(h)\right) = tg(h).$$

Since g is coercive, there exists $R \geq 1$ such that for every h with $||h|| \geq R$, $g(h) \geq 1$. Then, for every $h \in \mathbb{R}^d$, we have:

$$\begin{aligned} g(h) = g\left(\sum_{k=1}^{p} p_{A_k}(h)\right) &= g\left(\sum_{k=1}^{p} \frac{||h||^{1/2k}}{R^{1/2k}} p_{A_k}\left(R^{1/2k}\frac{h}{||h||^{1/2k}}\right)\right) \\ &= \frac{||h||}{R} g\left(\sum_{k=1}^{p} p_{A_k}\left(R^{1/2k}\frac{h}{||h||^{1/2k}}\right)\right). \end{aligned}$$

Then, for $||h|| \geq R$,

$$\left|\left|\sum_{k=1}^{p} p_{A_k}\left(R^{1/2k}\frac{h}{||h||^{1/2k}}\right)\right|\right|^2 = \sum_{k=1}^{p} \frac{R^{1/k}}{||h||^{1/k}} ||p_{A_k}(h)||^2 \geq \frac{R}{||h||}||h||^2 = R||h|| \geq R^2 \geq R,$$

so that $g(h) \geq \frac{||h||}{R}$ which in turn implies $e^{-g} \in L^1(\mathbb{R}^d)$. □

We now deal with the simplest configuration where the function g is not coercive, as described in (4.4.18), by dealing with the case where f is given by (4.4.17), which is an archetype of such configuration. However, dealing with the general case is more complicated and to give a general formula for the rate of convergence of the measure π_t in this case is not our current objective.

Proposition 4.4.7. *Let the function f be given by* (4.4.17)*. Then, if $(X_t, Y_t) \sim C_t e^{-f(x,y)/t} dxdy$, we have:*

$$\left(\frac{X_t}{t^{1/6}}, \frac{Y_t^2 - X_t}{t^{1/2}}\right) \underset{t\to 0}{\longrightarrow} C\frac{e^{-x^6}}{\sqrt{x}}\frac{e^{-y^2}}{\sqrt{\pi}}\mathbb{1}_{x\geq 0}dxdy,$$

where $C = \left(\int_0^\infty \frac{e^{-x^6}}{\sqrt{x}}dx\right)^{-1}$.

Proof. First, let us consider the normalizing constant C_t. We have :

$$\begin{aligned} C_t^{-1} &= \int_{\mathbb{R}^2} e^{-\frac{(x-y^2)^2+x^6}{t}} dxdy = 2t^{3/4}\int_{-\infty}^{\infty} e^{-t^2x^6}\int_0^\infty e^{-(y^2-x)^2}dy\ dx \\ &= t^{3/4}\int_{-\infty}^{\infty} e^{-t^2x^6}\int_{-x}^{\infty}\frac{e^{-u^2}}{\sqrt{u+x}}dy\ dx = t^{7/12}\int_{-\infty}^{\infty} e^{-x^6}\int_{-t^{-1/3}x}^{\infty}\frac{e^{-u^2}}{\sqrt{t^{1/3}u+x}}du\ dx \\ &\underset{t\to 0}{\sim} t^{7/12}\int_0^\infty \frac{e^{-x^6}}{\sqrt{x}}\int_{-\infty}^{\infty} e^{-u^2}du\ dx, \end{aligned}$$

where the convergence is obtained by dominated convergence and where we performed the change of variables $x' = t^{-1/6}x$ and $u = t^{-1/2}(y^2 - x)$. Then we consider, for $a_1 < b_1$ and $a_2 < b_2$,

$$\mathbb{P}\left(\frac{X_t}{t^{1/6}} \in [a_1, b_1],\ \frac{Y^2 - X}{t^{1/2}} \in [a_2, b_2]\right).$$

Performing the same changes of variables and using the above equivalent of C_t completes the proof. □

More generally, if the function g is not coercive and if we can write, up to a change of basis,

$$g(h_1,\dots,h_d) = Q_1(h_1,h_2)^2 + Q_2(h_3,h_4)^2 + \cdots + Q_r(h_{2r-1},h_{2r})^2 + \widetilde{g}(h_{2r+1},\dots,h_d), \quad (4.4.18)$$

where the Q_i are polynomials with two variables null on an unbounded curve (for example, $Q_i(x,y) = (x-y^2)$, $Q_i(x,y) = (x^2-y^3)$, $Q_i(x,y) = x^2y^2$), and where $\widetilde{g}$ is a non-negative coercive polynomial, then

$$\begin{aligned}&\Big(a_1\left((X_t)_1,(X_t)_2,t\right),\dots,a_r\left((X_t)_{2r-1},(X_t)_{2r},t\right),\\ &\qquad\left(\frac{1}{t^{\alpha_{2r+1}}},\dots,\frac{1}{t^{\alpha_d}}\right) * \left(\widetilde{B}\cdot\left((X_t)_{2r+1},\dots,(X_t)_d\right)\right)\Big)\\ &\underset{t\to 0}{\longrightarrow} b_1(x_1,x_2)\dots b_r(x_{2r-1},x_{2r}) C e^{-\widetilde{g}(x_{2r+1},\dots,x_d)} dx_1\dots dx_{2r}dx_{2r+1}\cdots dx_d,\end{aligned}$$

where C is a normalization constant, $\widetilde{B}\in\mathcal{O}_{d-2r-1}(\mathbb{R})$ is an orthogonal transformation and for all $k=1,\dots,r$, $a_k:\mathbb{R}^2\times(0,+\infty)\to\mathbb{R}^2$ and b_k is a density on $\mathbb{R}^2$. Such a_k and b_k can be obtained by applying the same method as in Proposition 4.4.7. Algorithm 1 yields the first change of variable for this method, given by the exponents (α_i) (in the proof of Proposition 4.4.7, the first change of variable is $t^{-1/2}x$ and $t^{-1/4}y$) and thus seems to be the first step of a more general procedure in this case. However, we do not give a general formula as the general case is cumbersome. Moreover, we do not give a method where the non coercive polynomials Q_i depend on more than two variables, like

$$Q(x,y,z) = (x-y^2)^2 + (x-z^2)^2.$$

The method sketched in Proposition 4.4.7 cannot be direclty applied to this case.

4.5 Proofs of Theorem 4.3.3 and Theorem 4.3.5 using Theorem 4.4.1

4.5.1 Single well case

We now prove Theorem 4.3.3.

Proof. Using Theorem 4.4.1, we have for all $h\in\mathbb{R}^d$:

$$\frac{1}{t}[f(x^\star + B\cdot(t^\alpha * h)) - f(x^\star)] \underset{t\to 0}{\longrightarrow} g(h).$$

To simplify the notations, assume that there is no need of a change of basis i.e. $B = I_d$. We want to apply Theorem 4.3.2 to the function f. However the condition

$$\int_{\mathbb{R}^d} \sup_{0<t<1} e^{-\frac{f(x^\star+(t^{\alpha_1}h_1,\dots,t^{\alpha_d}h_d))-f(x^\star)}{t}} dh_1\dots dh_d < \infty$$

is not necessarily true. Instead, let $\varepsilon>0$ and we apply Theorem 4.3.2 to $\widetilde{f}$, where $\widetilde{f}$ is defined as:

$$\widetilde{f}(h) = \begin{cases} f(h) & \text{if } h\in\mathcal{B}(x^\star,\delta)\\ ||h-x^\star||^2 + f(x^\star) & \text{else,}\end{cases}$$

and where $\delta>0$ will be fixed later. Then $\widetilde{f}$ satisfies the hypotheses of Theorem 4.3.2. The only difficult point to prove is the last condition of Theorem 4.3.2. If $t\in(0,1]$ and $h\in\mathbb{R}^d$ are such that $(t^{\alpha_1}h_1,\dots,t^{\alpha_d}h_d)\notin\mathcal{B}(0,\delta)$, then

$$\frac{\widetilde{f}(x^\star+(t^{\alpha_1}h_1,\dots,t^{\alpha_d}h_d)) - f(x^\star)}{t} = \frac{||(t^{\alpha_1}h_1,\dots,t^{\alpha_d}h_d)||^2}{t} \geq ||h||^2,$$

because for all i, $\alpha_i \leq \frac{1}{2}$. If t and h are such that $(t^{\alpha_1}h_1, \dots, t^{\alpha_d}h_d) \in \mathcal{B}(0,\delta)$, then choosing δ such that for all $(t^{\alpha_1}h_1, \dots, t^{\alpha_d}h_d) \in \mathcal{B}(0,\delta)$,

$$\left| \frac{f(x^\star + (t^{\alpha_1}h_1, \dots, t^{\alpha_d}h_d)) - f(x^\star)}{t} - g(h) \right| \leq \varepsilon,$$

which is possible because of the uniform convergence on every compact set (see Section 4.4.9), we derive that

$$\frac{f(x^\star + (t^{\alpha_1}h_1, \dots, t^{\alpha_d}h_d)) - f(x^\star)}{t} \geq g(h) - \varepsilon.$$

Hence

$$\int_{\mathbb{R}^d} \sup_{0<t<1} e^{-\frac{\widetilde{f}(x^\star + t^{\alpha_1}h_1, \dots, t^{\alpha_d}h_d)) - f(x^\star)}{t}} dh_1 \dots dh_d \leq \int_{\mathbb{R}^d} e^{-||h||^2} dh + e^{\varepsilon} \int_{\mathbb{R}^d} e^{-g(h)} dh.$$

Since g is coercive, using Proposition 4.4.6 we have $e^{-g} \in L^1(\mathbb{R}^d)$ and it follows from Theorem 4.3.2 that if $\widetilde{X}_t$ has density $\widetilde{\pi}_t(x) := \widetilde{C}_t e^{-\widetilde{f}(x)/t}$, then

$$\left(\frac{(\widetilde{X}_t - x^\star)_1}{t^{\alpha_1}}, \dots, \frac{(\widetilde{X}_t - x^\star)_d}{t^{\alpha_d}} \right) \xrightarrow{\mathscr{L}} X \quad \text{as } t \to 0,$$

where X has density proportional to $e^{-g(x)}$.

Now, let us prove that if X_t has density proportional to $e^{-f(x)/t}$, then we also have

$$\left(\frac{(X_t - x^\star)_1}{t^{\alpha_1}}, \dots, \frac{(X_t - x^\star)_d}{t^{\alpha_d}} \right) \xrightarrow{\mathscr{L}} X \quad \text{as } t \to 0. \tag{4.5.1}$$

Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be continuous with compact support. Then

$$\begin{aligned}
&\mathbb{E}\left[\varphi\left(\frac{(X_t)_1}{t^{\alpha_1}}, \dots, \frac{(X_t)_d}{t^{\alpha_d}} \right) - \varphi\left(\frac{(\widetilde{X}_t)_1}{t^{\alpha_1}}, \dots, \frac{(\widetilde{X}_t)_d}{t^{\alpha_d}} \right) \right] \\
&= \int_{\mathbb{R}^d} \varphi\left(\frac{x_1}{t^{\alpha_1}}, \dots, \frac{x_d}{t^{\alpha_d}} \right) \left(C_t e^{-\frac{f(x_1,\dots,x_d)}{t}} - \widetilde{C}_t e^{-\frac{\widetilde{f}(x_1,\dots,x_d)}{t}} \right) dx_1 \dots dx_d =: I_1 + I_2,
\end{aligned}$$

where I_1 is the integral on the set $\mathcal{B}(x^\star, \delta)$ and I_2 on $\mathcal{B}(x^\star, \delta)^c$. We have then:

$$|I_2| \leq ||\varphi||_\infty (\pi_t(\mathcal{B}(x^\star, \delta)^c) + \widetilde{\pi}_t(\mathcal{B}(x^\star, \delta)^c)) \underset{t\to 0}{\longrightarrow} 0,$$

where we used Proposition 4.3.1. On the other hand, we have $f = \widetilde{f}$ on $\mathcal{B}(x^\star, \delta)$, so that

$$|I_1| \leq ||\varphi||_\infty |C_t - \widetilde{C}_t| \int_{\mathcal{B}(x^\star,\delta)} e^{-\frac{f(x)}{t}} dx \leq ||\varphi||_\infty \left| 1 - \frac{\widetilde{C}_t}{C_t} \right|.$$

And we have:

$$\frac{\widetilde{C}_t}{C_t} = \frac{\int e^{-\frac{f(x)}{t}} dx}{\int e^{-\frac{\widetilde{f}(x)}{t}} dx} = \frac{\int_{\mathcal{B}(x^\star,\delta)} e^{-\frac{f(x)}{t}} dx + \int_{\mathcal{B}(x^\star,\delta)^c} e^{-\frac{f(x)}{t}} dx}{\int_{\mathcal{B}(x^\star,\delta)} e^{-\frac{f(x)}{t}} dx + \int_{\mathcal{B}(x^\star,\delta)^c} e^{-\frac{\widetilde{f}(x)}{t}} dx}.$$

By Proposition 4.3.1, we have when $t \to 0$

$$\begin{aligned}
\int_{\mathcal{B}(x^\star,\delta)^c} e^{-\frac{\widetilde{f}(x)}{t}} dx &= o\left(\int_{\mathcal{B}(x^\star,\delta)} e^{-\frac{\widetilde{f}(x)}{t}} dx \right) \\
\int_{\mathcal{B}(x^\star,\delta)^c} e^{-\frac{f(x)}{t}} dx &= o\left(\int_{\mathcal{B}(x^\star,\delta)} e^{-\frac{f(x)}{t}} dx \right),
\end{aligned}$$

so that $\widetilde{C}_t / C_t \to 1$, so $I_1 \to 0$, which then implies (4.5.1). □

4.5.2 Multiple well case

We now prove Theorem 4.3.5.

Proof. The first point is a direct application of Theorem 4.3.4. For the second point, we remark that X_{it} has a density proportional to $e^{-f_i(x)/t}$, where

$$f_i(x) := \begin{cases} f(x) \text{ if } x \in \mathcal{B}(x_i^\star, \delta) \\ +\infty \text{ else.} \end{cases}$$

We then consider $\widetilde{f}_i$ as in Section 4.5.1:

$$\widetilde{f}_i(x) = \begin{cases} f_i(x) & \text{if } x \in \mathcal{B}(x_i^\star, \delta) \\ ||x - x_i^\star||^2 + f^\star & \text{else.} \end{cases}$$

and still as in Section 4.5.1, we apply Theorem 4.3.2 to $\widetilde{f}_i$ and then prove that random variables with densities proportional to $e^{-\widetilde{f}_i(x)/t}$ and $e^{-f_i(x)/t}$ respectively have the same limit in law. □

4.6 Infinitely flat minimum

In this section, we deal with an example of infinitely flat global minimum, where we cannot use a Taylor expansion.

Proposition 4.6.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ such that*

$$\forall x \in \mathcal{B}(0,1),\ f(x) = e^{-\frac{1}{||x||^2}}$$

and

$$\forall x \notin \mathcal{B}(0,1),\ f(x) > a$$

for some $a > 0$. Furthermore, assume that f is coercive and $e^{-f} \in L^1(\mathbb{R}^d)$. Then, if X_t has density π_t,

$$\log^{1/2}\left(\frac{1}{t}\right) \cdot X_t \xrightarrow{\mathscr{L}} X \quad \text{as } t \to 0,$$

where $X \sim \mathcal{U}(\mathcal{B}(0,1))$.

Proof. Noting that $\int_{||x||>1} e^{-f(x)/t} dx \to 0$ as $t \to 0$ by dominated convergence, we have

$$C_t \underset{t\to 0}{\sim} \left(\int_{\mathcal{B}(0,1)} e^{-e^{-\frac{1}{||x||^2}}/t} dx\right)^{-1} = \log^{d/2}\left(\frac{1}{t}\right) \left(\underbrace{\int_{\mathcal{B}(0,\sqrt{\log(1/t)})} e^{-t^{\frac{1}{||x||^2}-1}} dx}_{\underset{t\to 0}{\to} \mathrm{Vol}(\mathcal{B}(0,1))}\right)^{-1},$$

where the convergence of the integral is obtained by dominated convergence. Then we have, for $-1 < a_i < b_i < 1$ and $\sum_i a_i^2 < 1$, $\sum_i b_i^2 < 1$:

$$\mathbb{P}\left(\log^{1/2}\left(\frac{1}{t}\right) \cdot X_t \in \prod_{i=1}^d [a_i, b_i]\right) = \frac{C_t}{\log^{d/2}\left(\frac{1}{t}\right)} \int_{(a_i)}^{(b_i)} e^{-t^{\frac{1}{|x|^2}-1}} dx \xrightarrow[t\to 0]{} \frac{\prod_{i=1}^d (b_i - a_i)}{\mathrm{Vol}(\mathcal{B}(0,1))}.$$

□

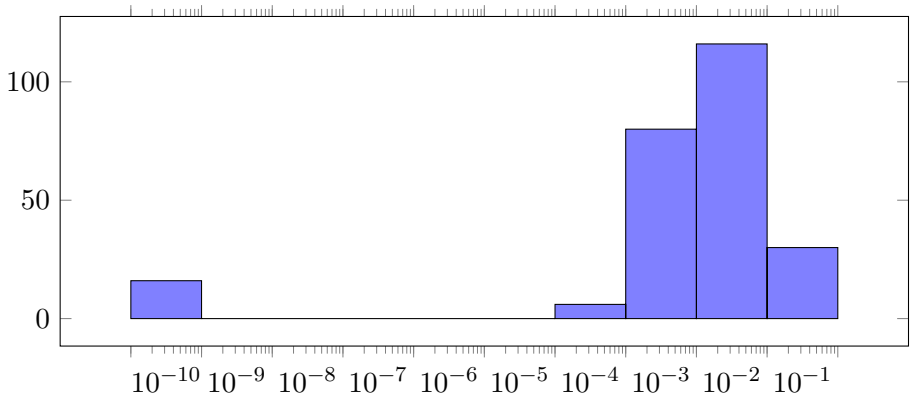


Figure 4.1: Distribution of the eigenvalues of the Hessian matrix at the end of training.

4.7 Simulations: computing high-order expansion of the loss with singular Hessian matrix

In this section, we present simulations illustrating our theoretical results for degenerate cases and showing how we can compute such high-order expansions in practice. Our code and a demonstration notebook are available at

`https://github.com/Bras-P/gibbs-measures-with-singular-hessian`.

We consider the function f to be the loss function associated to the training of an artificial neural network. We train a feedforward neural network on the MNIST dataset [LBBH98], which is composed of grayscale images of size 28×28 of handwritten digits (from 0 to 9). The neural network is composed of two hidden layers with 16 units each and with ReLU activation. After training the neural network on the data, we compute the Hessian matrix of the loss with respect to the kernel of the second Dense layer, as in practice we cannot compute the whole Hessian matrix with respect to all the variables. The Hessian matrix is singular; we give the distribution of its eigenvalues in Figure 4.1 where a gap appears between non-zero eigenvalues and zero eigenvalues.

We then compute the expansion of the loss function as in (4.4.2) up to the order 4 on the subspace of dimension 2 which is spanned by two eigenvectors, one with non-zero eigenvalue (subspace E) and one with zero eigenvalue (subspace F).

In TensorFlow, higher order derivatives can be computed the same way as first-order derivatives by stacking several `tf.GradientTape`. We use `tf.GradientTape.jacobian` to compute the Jacobian tensor of a tensor-valued function, since `tf.GradientTape.gradient` only computes the gradient of a scalar-valued function.

```
def compute_hessian(model, inputs, targets):
    with tf.GradientTape() as g1:
        with tf.GradientTape() as g2:
            loss_value = model.loss(inputs, targets)
        grads = g2.gradient(loss_value, model.trainable_variables)
    hessian_matrix = g1.jacobian(grads, model.trainable_variables)
    return hessian_matrix
```

Since the dimension of the variable may be very large for neural network, we compute derivative tensors only with respect to some variables. To do so, we need to create a new `Layer` object where the reduced kernel only includes these variables. We give the implementation of a dense layer modified in order to compute $(t_1, \dots, t_r) \in \mathbb{R}^r \mapsto \ell(x^\star + t_1 v_1 + \cdots + t_r v_r)$ where ℓ

stands for the output of the layer, where $x^\star \in \mathbb{R}^d$ is fixed and where $v_1, \ldots, v_r$ are vectors. The new layer class is defined as subclass of the base class `tf.keras.layers.Layer`. The method `build` initializes the weights of the layer when the layer is called for the first time; each weight is defined through the method `add_weight`. The method `call` defines the output of the layer depending on the inputs and its weights.

```
class CustomLayer(tf.keras.layers.Layer):
  def __init__(self, x_star, direction_vectors, activation=None):
    super(CustomLayer, self).__init__()
    self.x_star = x_star # x_star[0] is the kernel matrix and x_star[1]
   is the bias
    self.direction_vectors = tf.reshape(direction_vectors, [
   direction_vectors.shape[0], x_star[0].shape[0], x_star[0].shape[1]])
    self.activation = activation

  def build(self, input_shape):
    self.kernel = self.add_weight("kernel", shape = [direction_vectors.
   shape[0],])

  def call(self, inputs):
    outputs = tf.matmul(inputs, self.x_star[0] + tf.tensordot(self.
   kernel, self.direction_vectors, axes=[[0],[0]])) + self.x_star[1]
    if self.activation is not None:
        outputs = self.activation(outputs)
    return outputs
```

We then obtain the following expansion up to the order 4; if we denote by $x^\star$ the (empirical) minimum of the loss function, by x_1 the eigenvector with non-zero eigenvalue and by x_2 the eigenvector with zero eigenvalue, for $\lambda_1, \lambda_2 \in \mathbb{R}$ we have

$$\begin{aligned}&\frac{1}{t}\Big[f(x^\star + t^{1/2}\lambda_1 x_1 + t^{1/4}\lambda_2 x_2) - f(x^\star)\Big]\\ &\quad\xrightarrow[t\to 0]{} \frac{\lambda_1^2}{2}\nabla^2 f(x^\star)\cdot x_1^{\otimes 2} + \frac{\lambda_1\lambda_2^2}{2}\nabla^3 f(x^\star)\cdot x_1 \otimes x_2^{\otimes 2} + \frac{\lambda_2^4}{4!}\nabla^4 f(x^\star)\cdot x_2^{\otimes 4}\end{aligned}$$

where the three coefficients $\nabla^2 f(x^\star)\cdot x_1^{\otimes 2}$, $\nabla^3 f(x^\star)\cdot x_1 \otimes x_2^{\otimes 2}$ and $\nabla^4 f(x^\star)\cdot x_2^{\otimes 4}$ are computed in the notebook by stacking several `tf.GradientTape` as in `compute_hessian`.

```
(<tf.Tensor: shape=(), dtype=float32, numpy=0.026905548>,
 <tf.Tensor: shape=(), dtype=float32, numpy=-0.0007531713>,
 <tf.Tensor: shape=(), dtype=float32, numpy=-3.2213422e-05>)
```

Acknowledgements

I would like to thank Gilles Pagès for insightful discussions.

4.8 Appendix: Properties of tensors

Proposition 4.8.1. *Let T_k be a symmetric tensor of order k in $\mathbb{R}^d$. Let E be a subspace of $\mathbb{R}^d$. Assume that*

$$\forall h \in E,\ T_k \cdot h^{\otimes k} = 0.$$

Then we have

$$\forall h_1, \ldots, h_k \in E,\ T_k \cdot h_1 \otimes \cdots \otimes h_k = 0.$$

Proof. Using (4.2.1), we have for $h_1, \ldots, h_k \in E$ and $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$,

$$T_k \cdot (\lambda_1 h_1 + \cdots + \lambda_k h_k)^{\otimes k} = \sum_{i_1+\cdots+i_k=k} \binom{k}{i_1, \ldots, i_k} \lambda_1^{i_1} \ldots \lambda_k^{i_k} T_k \cdot h_1^{\otimes i_1} \otimes \cdots \otimes h_k^{\otimes i_k} = 0,$$

which is an identically null polynomial in the variables $\lambda_1, \ldots, \lambda_k$, so every coefficient is null, in particular

$$\forall h_1, \ldots, h_k \in E, \ T_k \cdot h_1 \otimes \cdots \otimes h_k = 0.$$

□

Part II

Adaptive Langevin algorithms for deep Neural Networks

Chapter 5

Stochastic algorithms for artificial neural networks with simulations in TensorFlow

5.1 Artificial neural networks

5.1.1 Calibration of artificial neural networks as a stochastic optimization problem

Neural networks are a powerful tool for a wide range of problems (regression, classification, etc) involving inferring a model from a huge amount of data. Neural networks are functions with high-dimensional parametrization thus allowing to fit to many different problems and data. With neural networks, the user does not need to devise a very specific model fitting to the data by himself anymore.

Let us consider data samples $u_i \in \mathbb{R}^{d_{\text{in}}}$ and $y_i \in \mathbb{R}^{d_{\text{out}}}$ for $1 \leq i \leq M$ associated to a regression problem, where (u_i) are the inputs and where (y_i) are the outputs. That is, we look for a function $\psi : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ which fits to the data i.e.

$$\forall\ 1 \leq i \leq M,\ \psi(u_i) \approx y_i.$$

In other words, the objective is to extract a model from the empirical data. We look for a function ψ in a family of functions parametrized by a finite-dimensional parameter: $\{\psi_x,\ x \in \mathbb{R}^d\}$. For $L : \mathbb{R}^{d_{\text{out}}} \to \mathbb{R}$ a loss function which measures the error between the prediction $\psi_x(u_i)$ and the true data y_i, the regression problem then becomes the minimization of the average loss over the data and can be written as the following optimization problem:

$$\underset{x \in \mathbb{R}^d}{\text{Minimize}}\ V(x) := \frac{1}{M} \sum_{i=1}^{M} L(\psi_x(u_i) - y_i). \tag{5.1.1}$$

The output of a fully connected neural network ψ_x is defined as follows. Let us choose an activation function $\varphi : \mathbb{R} \to \mathbb{R}$, which is a non-linear sigmoid-type or ReLU-type function (see Table 5.1). Let $K + 1$, $K \in \mathbb{N}$, be the number of layers of the neural network and for $k = 0$, $\ldots$, K let $d_k \in \mathbb{N}$ be the size of the k^{th} layer with $d_0 = d_{\text{in}}$ and $d_K = d_{\text{out}}$; the output is defined recursively as:

$$\begin{aligned}
&u^{(0)} := u \in \mathbb{R}^{d_{\text{in}}},\\
&u^{(k)} = \varphi(\alpha_k \cdot u^{(k-1)} + \beta_k) \in \mathbb{R}^{d_k}, \quad 1 \leq k \leq K-1,\\
&\psi_x(u) = u^{(K)} = \alpha_K \cdot u^{(K-1)} + \beta_K.
\end{aligned}$$

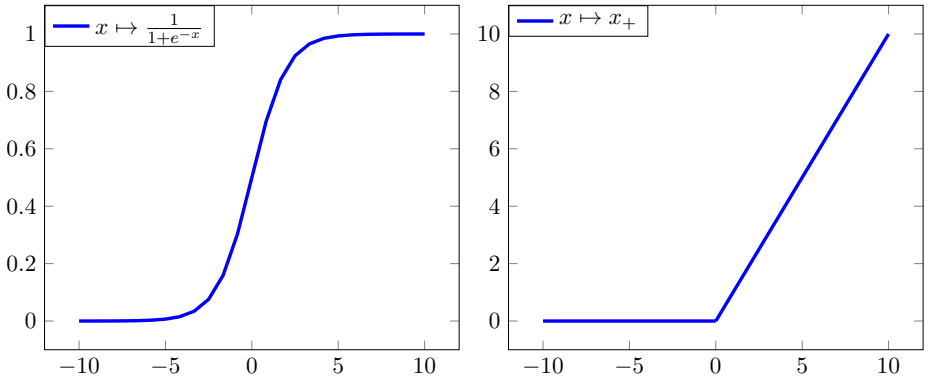


Figure 5.1: The Sigmoid and ReLU functions

The second equation should be understood for each coordinate of $u^{(k)}$. For every $1 \leq k \leq K$, $\alpha_k \in \mathcal{M}_{d_k,d_{k-1}}(\mathbb{R})$ and $\beta_k \in \mathbb{R}^{d_k}$ and the parameter x of the neural network is

$$x = (\alpha_1, \beta_1, \ldots, \alpha_K, \beta_K) \in \mathcal{M}_{d_1,d_0}(\mathbb{R}) \times \mathbb{R}^{d_1} \times \cdots \times \mathcal{M}_{d_K,d_{K-1}}(\mathbb{R}) \times \mathbb{R}^{d_K}.$$

Neural networks have first gained interest as universal approximators since Cybenko's theorem [Cyb89] stating that if $d_{\text{out}} = 1$, if $K = 2$ and if $\varphi : \mathbb{R} \to \mathbb{R}$ is continuous and such that

$$\varphi(t) \underset{t\to-\infty}{\longrightarrow} 0 \quad \text{and} \quad \varphi(t) \underset{t\to+\infty}{\longrightarrow} 1$$

then the set

$$\left\{ u \mapsto \sum_{j=1}^{d_1} \alpha_2^j \varphi([\alpha_1 \cdot u + \beta_1]_j),\ d_1 \in \mathbb{N},\ \alpha_1 \in \mathcal{M}_{d_0,d_1}(\mathbb{R}),\ \beta_1 \in \mathbb{R}^{d_1},\ \alpha_2 \in \mathbb{R}^{d_1} \right\}$$

is dense in $\mathcal{C}^0([0,1]^{d_{\text{in}}}, \mathbb{R})$ for the uniform norm. In fact, no monotonicity assumption is required for φ.

5.1.2 Neural networks architectures for tasks other than regression

5.1.2.1 Classification

Let us consider a classification problem: the output to predict y is not a "continuous" data having values in $\mathbb{R}^{d_{\text{out}}}$ but is a classification label in $\{1, \ldots, N_{\text{labels}}\}$. In this case, the output y is mapped to $\mathbb{R}^{N_{\text{labels}}}$ as a one-hot vector i.e. for $1 \leq j \leq N_{\text{labels}}$, the j^{th} coordinate is the probability to belong to the j^{th} class. Similarly to the logistic classification, the loss function used for classification problems is the categorical cross-entropy: if the data $y = i \in \{1, \ldots, N_{\text{labels}}\}$ and the prediction vector is $z \in \mathbb{R}^{N_{\text{labels}}}$, then the resulting loss is $\log(1 + e^{-z_i})$. For more details we refer to [TKV10].

5.1.2.2 Images and convolutional layers

An image encoded as a matrix can be seen as an input vector, however this does not take into account the spatial properties of the image. Moreover, the original dimension of an image is generally too large to be directly processed using fully-connected layers. A common way to deal with image analysis is to stack a succession of 2D-convolutions which kernel is a trainable parameter in order to extract its spatial features, and of pooling layers which are non-trainable but reduce the dimension of the image by averaging or taking the maximum over small squares

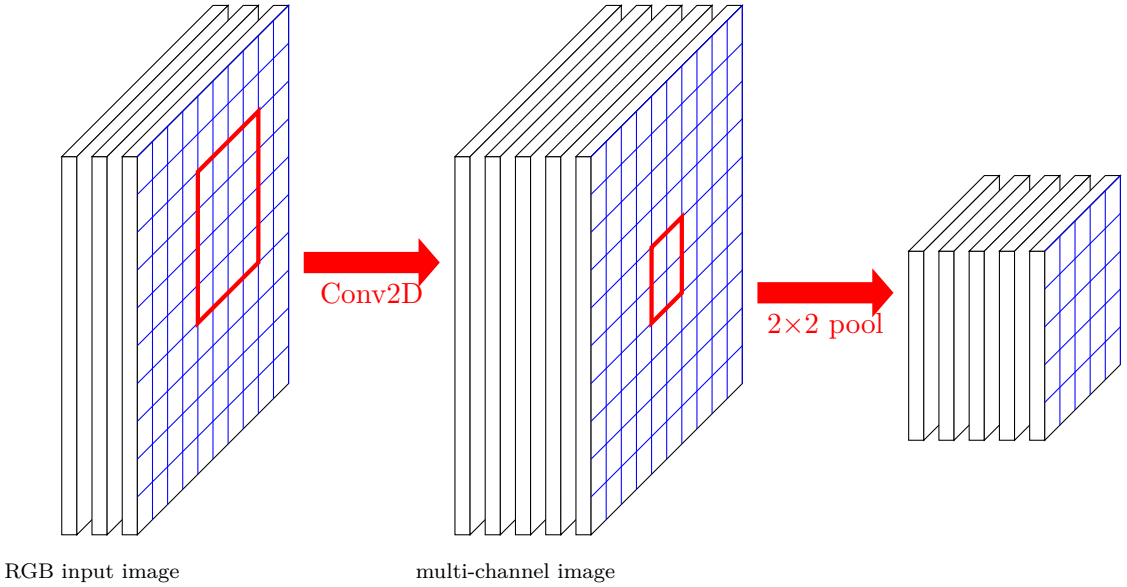


Figure 5.2: 2D-convolution layer followed by max-pooling architecture

[JKRL09]. Once the features are extracted and the dimension is reduced, we use dense layers to extract the role of each feature. An illustration is given in Figure 5.2.

5.1.2.3 Time series

The problem is to predict the value of a time series in the future knowing the previous values over a time window. This problem can be rewritten as a regression problem as follows. For a whole time series $(u_k)_{1\leq k\leq T} \in (\mathbb{R}^{d_1})^T$, the input data is the extracted sub-series $(u_{k+1}, \dots, u_{k+\ell})$ where $\ell \ll T$ and for each k the output is $u_{k+\ell+s}$ where s is the number of time steps in the future for which we predict the value of u.

To design a neural network adapted to time series analysis, we first need layers that extract the features of the data and that take into account their sequential structure. Along with images, we can use 1D-convolutional layers however such layers only extract short-term local features. Recurrent Neural Networks (RNN) are more adapted to the analysis of short but also long-term dependencies. RNN have a "hidden state" which is a time series that is recursively updated according to the data [RHW86]. More specifically, the output of a RNN is the hidden state $(h_{k+1}, \dots, h_{k+\ell})$ that is recursively updated as

$$h_{j+1} = \psi_x(u_j, h_j)$$

where ψ_x is some neural function with trainable parameter x (independent of j). Popular choices for the parametrization of ψ_x are LSTM (Long Short Term Memory) networks [HS96, HS97] and GRU (Gated Recurrent Unit) networks [CvMBB14].

5.1.3 Neural networks, stochastic gradient and automatic differentiation

The stochastic gradient descent for (5.1.1) writes

$$x_{n+1} = x_n - \gamma_{n+1}\nabla_x \left(L(\psi_{x_n}(u_{i_{n+1}}) - y_{i_{n+1}})\right) \tag{5.1.2}$$

where $i_{n+1} \in \{1, \dots, M\}$ is an index chosen uniformly at random and where (γ_n) is a non-increasing step sequence. In practice, SGD is implemented by batches where the gradient is

estimated by the average over a small amount of data:

$$x_{n+1} = x_n - \frac{\gamma_{n+1}}{N_{\text{batch}}} \sum_{i \in \mathcal{I}_{n+1}} \nabla_x \left(L(\psi_{x_n}(u_i) - y_i) \right) \tag{5.1.3}$$

where $\mathcal{I}_{n+1}$ is a subset of $\{1, \dots, M\}$ of size N_{batch} taken uniformly at random with $N_{\text{batch}} \ll M$. The advantages of using batches are that the gradient descent is more stable and that for each iteration, all the gradients $\nabla_x(L(\psi_{x_n}(u_i) - y_i))$ for $i \in \mathcal{I}_{n+1}$ are computed simultaneously by parallelization in particular using GPU devices, so that computing N_{batch} gradients takes the same time as computing one gradient provided that N_{batch} is not too large.

Computing the gradient $\nabla_x \left(L(\psi_{x_n}(u_i) - y_i) \right)$ using the finite difference method is inaccurate and cumbersome in particular when the dimension of the parameter x is large. Instead we use automatic differentiation: knowing that $L(\psi_x(u_i) - y_i)$ is written as a composition of elementary operations (addition, multiplication, etc) and functions (exp, log etc) of the parameter x as it is the case for neural networks, then the gradient with respect to x is explicitly computed using the chain rule.

In a more general framework, let us consider a function $f : \mathbb{R}^d \to \mathbb{R}^q$ of x which can be written as a composition of elementary operations which gradients are analytically known. More specifically, let us write these elementary operations as $f(x_0) = x_N$ with x_N defined recursively as

$$x_0 \in \mathbb{R}^{\ell_0},\ x_{n+1} = \begin{pmatrix} x_n \\ f_{n+1}(x_n) \end{pmatrix} =: F_{n+1}(x_n) \in \mathbb{R}^{\ell_{n+1}},\ n \in \{0, \dots, N-2\},\ x_N = f_N(x_{N-1}) \tag{5.1.4}$$

where $f_{n+1} : \mathbb{R}^{\ell_n} \to \mathbb{R}^{\ell_{n+1} - \ell_n}$ and $F_{n+1} : \mathbb{R}^{\ell_n} \to \mathbb{R}^{\ell_{n+1}}$ and $d = \ell_0 < \ell_1 < \cdots < \ell_{N-1}$ are increasing dimensions and $\ell_N = q$. Thus in order to compute the gradient $\partial x_N / \partial x_0$, writing $\dot{x}_n := \partial x_n / \partial x_0 \in \mathbb{R}^{\ell_n \times \ell_0}$ with $\dot{x}_0 = I_{\ell_0}$ the identity matrix and using the chain rule we have

$$\dot{x}_{n+1} = \nabla F_{n+1}(x_n) \cdot \dot{x}_n. \tag{5.1.5}$$

However (5.1.5) uses the matrix products $\nabla F_{n+1}(x_n) \cdot \dot{x}_n$ involving matrix products of dimensions $(\ell_{n+1} \times \ell_n) \cdot (\ell_n \times \ell_0)$, which is computationally expensive in the cases where ℓ_0 (the number of parameters with respect to which we compute the gradient) is large and ℓ_N is small (typically $\ell_N = 1$). Another method is to consider instead the adjoint

$$\bar{x}_n := \left(\frac{\partial x_N}{\partial x_n} \right)^\top \in \mathbb{R}^{\ell_n \times \ell_N} \tag{5.1.6}$$

with $\bar{x}_N = I_{\ell_N}$ so that we have the backward recursive relation

$$\bar{x}_n = \left(\nabla F_{n+1}(x_n) \right)^\top \cdot \bar{x}_{n+1}, \tag{5.1.7}$$

involving matrix products of dimensions $(\ell_n \times \ell_{n+1}) \cdot (\ell_{n+1} \times \ell_N)$. Thus a method to compute $\dot{x}_N = \bar{x}_0$ is:

1. Forward step: compute x_N and the gradients $\nabla F_{n+1}(x_n)$ using the forward recursive relation (5.1.5).

2. Backward step: compute $\bar{x}_0$ using the backward recursive relation (5.1.7).

The gradients $\nabla F_{n+1}(x_n)$ are directly computed using the known analytical expression of ∇f_{n+1} for f_{n+1} in a finite set of elementary operations and functions. For example, for $f(x) = e^x$ then $\nabla f(x) = e^x$; for $f(x, y) = x + y \in \mathbb{R}^\ell$ then $\nabla f(x, y) = (I_\ell, I_\ell)$; for $f(x, y) = x * y \in \mathbb{R}^\ell$ then $\nabla f(x, y) = (\text{diag}(y), \text{diag}(x))$.

For more details we refer to [Gil07, BPRS17].

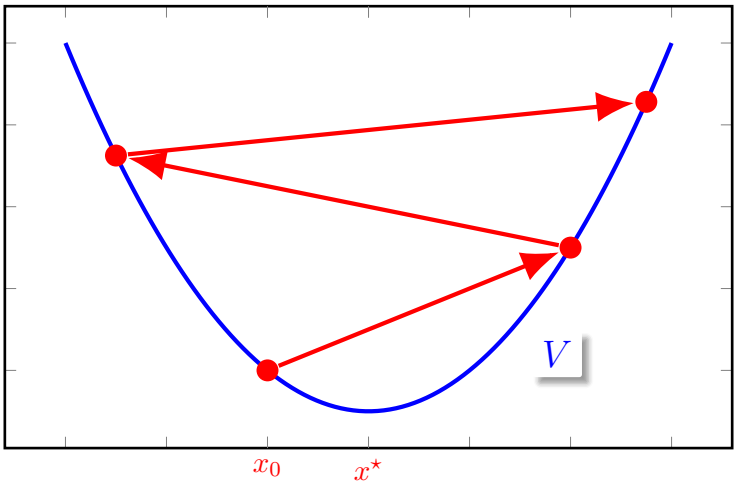


Figure 5.3: Behaviour of SGD when the learning rate is too large.

5.2 Stochastic gradient algorithms

5.2.1 Adaptive stochastic algorithms

Many variants of the the classic SGD algorithm (5.1.2) have been developed in order to make more efficient and more stable algorithms and to accelerate their speed of convergence. One main drawback of (5.1.2) is that the learning rate (γ_n) must be set by the user and may not be adapted simultaneously to any optimization problem (5.1.1) in general. Indeed, if the learning rate is too large then the algorithm may be repelled from minima or diverge as illustrated in Figure 5.3. Other stochastic gradient methods are usually based on adaptive learning rates and for a noisy measurement of the gradient g_{n+1} they read

$$x_{n+1} = x_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1} \tag{5.2.1}$$

where $P_{n+1} \in \mathcal{M}_d(\mathbb{R})$ is positive definite. As the dimension of x is large, P_{n+1} is chosen to be a diagonal matrix, thus yielding an adaptive learning rate for each dimension.

Each different choice for the preconditionner rule P yields a different algorithm; let us present some of the most used stochastic gradient methods in practice.

The RMSprop algorithm [TH12] consists in maintaining a discounted average of the square of gradients for each weight and to divide the gradient by the root of this average. More precisely, given $\alpha \in (0,1)$ close to 1 and $\lambda > 0$ close to 0 the RMSprop update is written in Algorithm 1.

Algorithm 1 RMSprop

for $1 \leq n \leq N_{\text{iter}}$ **do**
$g_{n+1} = \nabla_x \left(L(\psi_{x_n}(u_{i_{n+1}}) - y_{i_{n+1}}) \right)$
$\text{MS}_{n+1} = \alpha\, \text{MS}_n + (1-\alpha) g_{n+1} \odot g_{n+1}$
$P_{n+1} = \text{diag}\left(\mathbb{1} \oslash (\lambda \mathbb{1} + \sqrt{\text{MS}_{n+1}}) \right)$
$x_{n+1} = x_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1}$
end for

The parameter λ is used for numerical stability. The operators $\odot$ and $\oslash$ represent element-wise product and division, respectively. Using a discounted average with rate α instead of the current value allows to counter the unstable mini-batch effects, as evaluating the gradient on mini-batches leads to estimates that may greatly differs from one iteration to the next.

The Adam algorithm [KB15] is based on first and second moment estimates of the gradient. For $\beta_1 \in (0,1)$ and $\beta_2 \in (0,1)$ both close to 1 and for $\lambda > 0$ close to 0, the Adam update is written in Algorithm 2. Note that in the update of x_{n+1}, an average of the gradient $\widehat{M}_{n+1}$ over the past iterations with exponentially decreasing weight is considered. The notations β_1^{n+1} and β_2^{n+1} denote β_1 and β_2 to the power $n+1$ respectively.

Algorithm 2 Adam

for $1 \leq n \leq N_{\text{iter}}$ **do**
$g_{n+1} = \nabla_x \left(L(\psi_{x_n}(u_{i_{n+1}}) - y_{i_{n+1}}) \right)$
$M_{n+1} = \beta_1 M_n + (1-\beta_1) g_{n+1}$
$\mathrm{MS}_{n+1} = \beta_2 \mathrm{MS}_n + (1-\beta_2) g_{n+1} \odot g_{n+1}$
$\widehat{M}_{n+1} = M_{n+1}/(1-\beta_1^{n+1})$
$\widehat{\mathrm{MS}}_{n+1} = \mathrm{MS}_{n+1}/(1-\beta_2^{n+1})$
$P_{n+1} = \mathrm{diag}\left(\mathbb{1} \oslash \left(\lambda \mathbb{1} + \sqrt{\widehat{\mathrm{MS}}_{n+1}} \right) \right)$
$x_{n+1} = x_n - \gamma_{n+1} P_{n+1} \cdot \widehat{M}_{n+1}$.
end for

The Adadelta algorithm [Zei12] uses as preconditioner the moving root mean square of the increments of (x_n) divided by the moving root mean square of the gradients (g_n) and is given in Algorithm 3.

Algorithm 3 Adadelta

for $1 \leq n \leq N_{\text{iter}}$ **do**
$g_{n+1} = \nabla_x \left(L(\psi_{x_n}(u_{i_{n+1}}) - y_{i_{n+1}}) \right)$
$\mathrm{MS}_{n+1} = \beta_1 \mathrm{MS}_n + (1-\beta_1) g_{n+1} \odot g_{n+1}$
$P_{n+1} = \mathrm{diag}\left((\lambda \mathbb{1} + \widehat{MS}_n) \oslash \left(\lambda \mathbb{1} + \sqrt{\widehat{\mathrm{MS}}_n} \right) \right)$
$x_{n+1} = x_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1}$.
$\widehat{MS}_{n+1} = \beta_2 \mathrm{MS}_n + (1-\beta_2)(x_{n+1} - x_n) \odot (x_{n+1} - x_n)$.
end for

Other optimizers used in Machine Learning mainly rely on the same idea which is to adapt the learning rate for each weight using gradient moving averages and are often based on tunes of one of the previous algorithms, such as Adamax [KB15], which is a variant of Adam based on the infinity norm instead of the L^2-norm, and AMSgrad [RKK18].

Likewise, vanilla or preconditioned Langevin algorithms are obtained by adding Gaussian noise to a stochastic gradient method for some choice of preconditioner. More specifically, the Langevin version of (5.2.1) reads

$$x_{n+1} = x_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1} + a_{n+1} \gamma_{n+1}^{1/2} P_{n+1}^{1/2} \cdot \mathcal{N}(0, I_d) + a_{n+1}^2 \gamma_{n+1} \Upsilon_{n+1}, \tag{5.2.2}$$

where a_{n+1} is a scalar positive decreasing or constant sequence controlling the amplitude of the noise and where Υ is a second-order correction term given by

$$[\Upsilon_{n+1}]_{1 \leq i \leq d} = \Big[\sum_{j=1}^{d} \partial_j [P_{n+1}]_{ij} \Big]_{1 \leq i \leq d}. \tag{5.2.3}$$

This formula is given in [MCF15] [LCCC16, Equation (3)] [PP23, Proposition 2.5]. For a non-Langevin algorithm with name *Name*, we denote by L-*Name* its Langevin version i.e. the

Langevin algorithm obtained by adding a preconditioned Gaussian noise and given by the update (5.2.2).

5.2.2 Gradient optimizers in TensorFlow

We use the Machine Learning library `TensorFlow` [AAB+15] with `Python`. Neural networks can be easily built by stacking layers:

```
import tensorflow as tf

model = tf.keras.Sequential([
  tf.keras.layers.Conv2D(8, (4, 4), activation='sigmoid'), # 8 filters
   with convolution kernel of size 4x4
  tf.keras.layers.MaxPooling2D((2, 2)), # Pooling with infinity norm on
   2x2 squares
  tf.keras.layers.Flatten(),
  tf.keras.layers.Dense(32, activation='sigmoid'), # 32 units
  tf.keras.layers.Dense(10)
])
```

Listing 5.1: Example of a 2D-convolutional neural network in TensorFlow

Once the model is built, we compile it with a loss function and an optimizer:

```
model.compile(
  optimizer='SGD',
  loss=tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True)
)
```

Then the model is trained and its weights are fitted to the data:

```
model.fit(train_features, train_labels, batch_size=256, epochs=20,
   validation_data=(val_features, val_labels))
```

Some gradient algorithms are already implemented in TensorFlow and can be found in the module `tf.keras.optimizers`, such as `tf.keras.optimizers.SGD`, `tf.keras.optimizers.Adam`, `tf.keras. optimizers.RMSprop`. We implement Langevin algorithms as subclasses of the base class `tf.keras. optimizers.Optimizer`. The method `_resource_apply_dense(self, grad, var)` defines the algorithm update rule applied to the weight of the model `var` knowing the derivative of the loss `grad` with respect to `var`. For example, for the basic stochastic gradient algorithm, the update rule is simply

```
def _resource_apply_dense(self, grad, var):
  var.assign(var - self.learning_rate*grad)
```

Auxiliary variables of the optimizers can be defined in `create_slots(self, var_list)`, for example the gradient square moving average MS_n for `RMSprop`, see Algorithm 1.

We give the code using these methods for the L-RMSprop algorithm in Listing 5.2. The code for other L-algorithms can be found at

`https://github.com/Bras-P/langevin-simulated-annealing/blob/main/optimizers.py`.

We first create a subclass `LangevinOptimizer` of the base TensorFlow optimizer class `tf.keras.optimizers.Optimizer` in order to handle the possible decreasing of `sigma` which is either a `float` value either a `tf.keras.optimizers. schedules. LearningRateSchedule`:

```
class LangevinOptimizer(tf.keras.optimizers.Optimizer):
    def __init__(self, learning_rate, sigma, name='LangevinOptimizer',
   **kwargs):
```

```
        super().__init__(name, **kwargs)
        self._set_hyper("learning_rate", kwargs.get("lr", learning_rate)
    )
        self._set_hyper("sigma", kwargs.get("sigma", sigma))

    def _decayed_sigma(self, var_dtype):
        sigma = self._get_hyper("sigma", var_dtype)
        if isinstance(sigma, tf.keras.optimizers.schedules.
    LearningRateSchedule):
            local_step = tf.cast(self.iterations, var_dtype)
            sigma = tf.cast(sigma(local_step), var_dtype)
        return sigma

    def _prepare_local(self, var_device, var_dtype, apply_state):
        super()._prepare_local(var_device, var_dtype, apply_state)
        if "sigma" in self._hyper:
            sigma = tf.identity(self._decayed_sigma(var_dtype))
            apply_state[(var_device, var_dtype)]["sigma"] = sigma

    def _get_coefficients(self, grad, var, apply_state=None):
        var_device, var_dtype = var.device, var.dtype.base_dtype
        coefficients = (apply_state or {}).get(
            (var_device, var_dtype)
        ) or self._fallback_apply_state(var_device, var_dtype)
        lr_t, sigma = coefficients["lr_t"], coefficients["sigma"]
        return coefficients, lr_t, sigma
```

Then the L-RMSprop algorithm is given as a subclass of `LangevinOptimizer` by:

```
class LRMSprop(LangevinOptimizer):
    def __init__(self, learning_rate=0.001, sigma=0.001, alpha=0.9,
    diagonal_bias=1e-6, name="LRMSprop", **kwargs):
        super().__init__(learning_rate, sigma, name, **kwargs)
        self.alpha = alpha
        self.diagonal_bias = diagonal_bias

    def _create_slots(self, var_list):
        for var in var_list:
            self.add_slot(var, "rms")

    @tf.function
    def _resource_apply_dense(self, grad, var, apply_state=None):
        coefficients, lr_t, sigma = self._get_coefficients(grad, var,
    apply_state)
        rms_var = self.get_slot(var, "rms")
        new_rms = self.alpha*rms_var + (1-self.alpha)*tf.square(grad)
        preconditioner = 1./(self.diagonal_bias + tf.math.sqrt(new_rms))
        stddev = sigma*tf.math.sqrt(lr_t*preconditioner)
        new_var = var - lr_t*preconditioner*grad + tf.random.normal(
    shape=tf.shape(grad), stddev=stddev)
        rms_var.assign(new_rms)
        var.assign(new_var)
```

Listing 5.2: Implementation of the L-RMSprop optimizer in TensorFlow

The correction term Υ which includes second-order derivatives can be computed using automatic differentiation as same as for the computation of first-order derivatives with the function `diag_jacobian` from the module `tensorflow_probability.python.math`. However in general

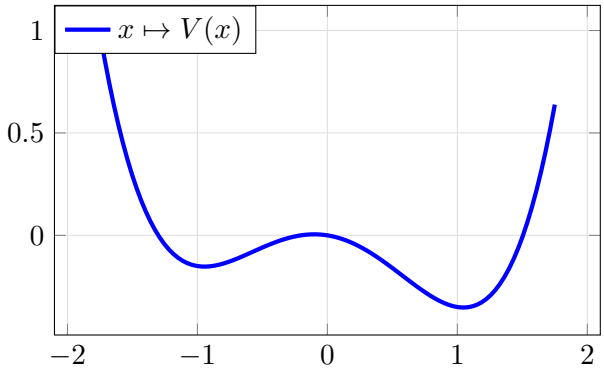


Figure 5.4: Double well potential

Algorithm	SGLD	L-RMSprop	L-Adam
Mean escape time	217.8	90.40	87.13

Table 5.1: Mean escape time of the local minimum of a double-well function. The values of the parameters are $c = 0.1$, $x_0 = -1.5$ (initial point), $\gamma = 0.1$. The escape times are averaged over 1000 runs. The escape time is defined as the first iteration k such that $x_k \geq 1$.

we do not include the correction term Υ in the Langevin optimizers as computing second-order derivatives highly increases the computation time and including the term Υ does not lead to any visible improvement in practice in comparison with the computation time. Moreover the time step coefficient of Υ_n in (5.2.2) is $a_n^2\gamma_n$ with $a_n \ll 1$ and $\gamma_n \ll 1$, in comparison with γ_n for the gradient and $a_n\gamma_n^{1/2}$ for the Gaussian noise, and if the regularization coefficient α is close to 1 then the preconditioner slowly varies at each iteration and then the true correction term Υ is close to 0.

5.3 Simulations

We compare the performances of various adaptive (preconditioned) Langevin algorithms with vanilla SGLD, which is the Langevin algorithm with constant (additive) σ (P_n), as well as with standard non-Langevin algorithms. The algorithms are tested on diverse optimization and inference problems with real-life data.

5.3.1 Double well potential

Let us start with a simple one-dimensional example. Let $V : \mathbb{R} \to \mathbb{R}$ be a coercive $\mathcal{C}^2$ "double-well"-type function i.e. with two local minima $x_1^\star$ and $x_2^\star$ and such that $V(x_1^\star) \geq V(x_2^\star)$. In this section we consider

$$V(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2 - cx,$$

see Figure 5.4. Starting in the the neighbourhood of $x_1^\star = -1$, we simulate how many iterations are needed to escape the first local minimum and to reach $x_2^\star = 1$. The Gaussian noise added to the gradient descent indeed allows to escape from local minima and to explore the whole space. The results are given in Table 5.1, where we compare the performances of preconditioned Langevin algorithms with vanilla SGLD. Preconditioned Langevin algorithms show better performances in exploring the state space with a lower mean escape time.

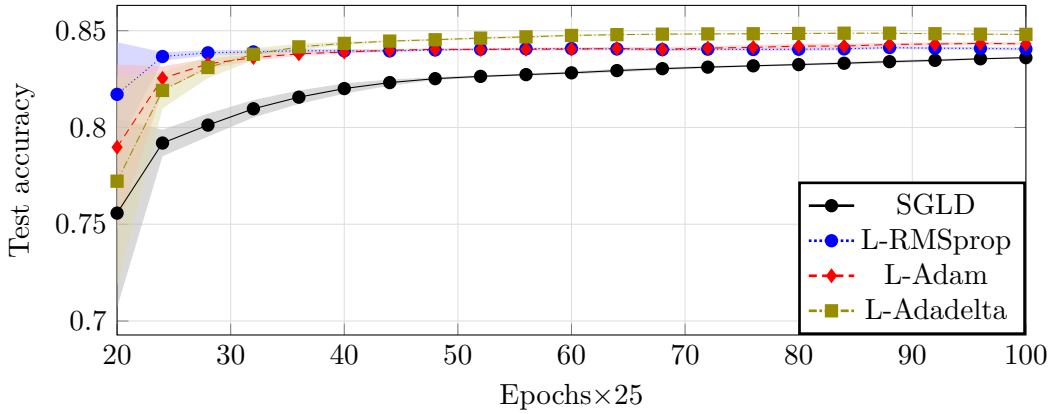


Figure 5.5: Performance of preconditioned Langevin algorithms compared with vanilla SGLD on the `Adult` dataset. The values of the hyperparameters are $a = 0.01$ and $\gamma = 0.01$ for SGLD, L-RMSprop and L-Adam and $\gamma = 0.5$ for L-Adadelta, batch size is 32 and the window for the posterior is $N = 800$. 95% confidence intervals are indicated.

5.3.2 Bayesian Logistic Regression

We compare vanilla SGLD with preconditioned Langevin algorithms on a Bayesian logistic regression problem. Given a binary classification problem $(u_i, y_i)_{1\leq i\leq N}$ where $y_i \in \{\pm 1\}$, the probability of the i^{th} output is

$$p(y_i|u_i, x) = \frac{1}{1 + \exp\left(-y_i\langle x, u_i\rangle\right)},$$

corresponding to the binary entropy loss. We use the Adult dataset `a9a` [Koh96] where the binary classification task is to determine whether a person earns over 50k$ per year knowing 14 features such as age, education etc. We use a Bayesian approach for this problem and we simulate the parameter x according to the Gibbs distribution of density proportional to $e^{-2V(x)/a^2}$. The empirical posterior distribution is given by the N last values of the weights in the algorithm. The advantage of the Bayesian approach is to be able to quantify the uncertainty about the inferred values of the weights and about the underlying process.

The results are given in Figure 5.5, showing that the decrease of the loss function is faster for precondtioned Langevin algorithms than for vanilla SGLD.

5.3.3 Image classification (1)

We train a neural network on the MNIST dataset [LBBH98], which is composed of grayscale images of size 28×28 of handwritten digits (from 0 to 9). The goal is to recognize the handwritten digit and to classify the images. 60.000 images are used for training and 10.000 images are used for test. Examples of predictions are given in Figure 5.6.

We first compare the performances of various Langevin algorithms for the training of a feedforward neural network, composed of two hidden dense layers with 128 units each and with ReLU activation. The results are given in Figure 5.7. The preconditioned Langevin algorithms show significant improvement compared with the vanilla SGLD algorithm. Their convergence is faster and they achieve a lower error on the test set. We also display the value of the loss function on the train set during the training to show that the better performances of the preconditioned algorithms are not due to some overfitting effect.

We then compare preconditioned Langevin algorithms with their respective non-Langevin counterparts. For shallow neural networks, adding an exogenous noise does not seem to improve

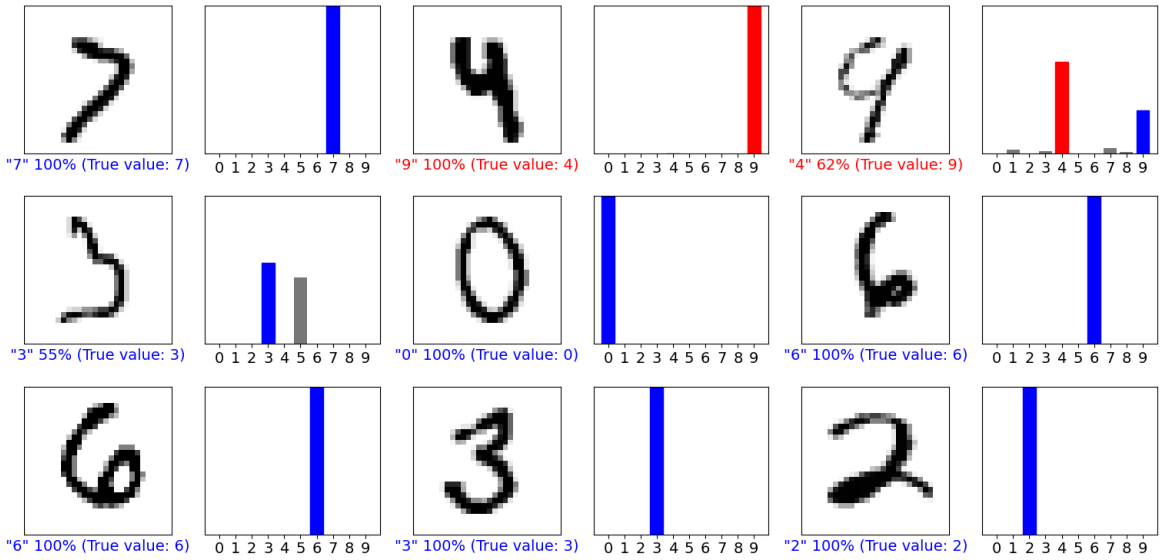


Figure 5.6: Examples of probability predictions for some images from the MNIST dataset

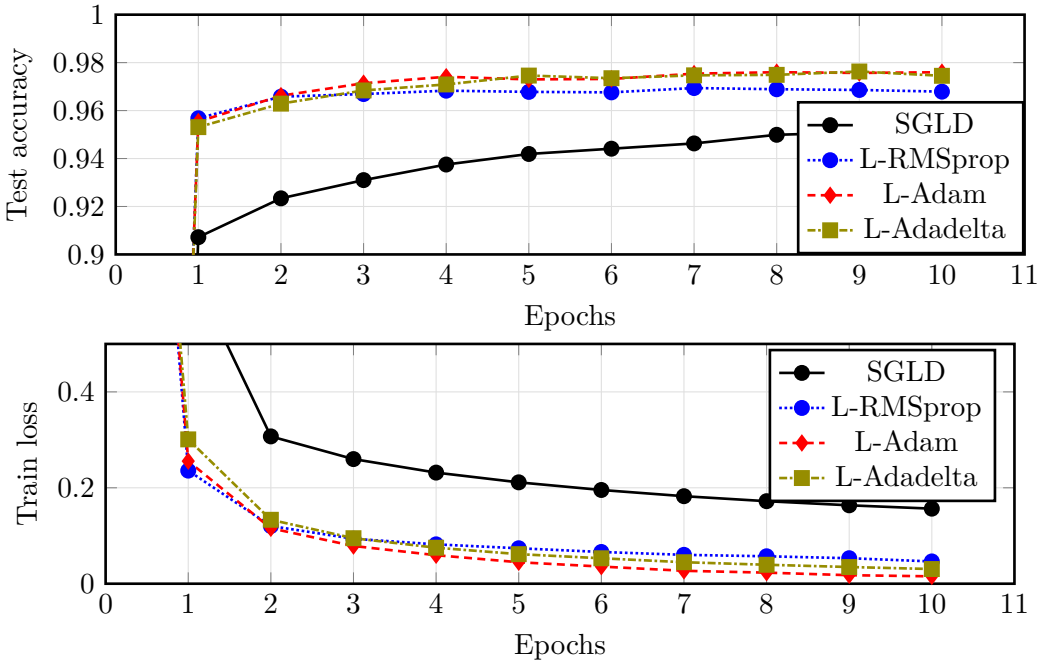


Figure 5.7: Performance of preconditioned Langevin algorithms compared with vanilla SGLD on the MNIST dataset. The values of the hyperparameters are $a(n) = A\log^{-1/2}(c_1 n + e)$ with $A = 2.10^{-3}$ and where $c_1 n = 1$ after 5 epochs; $\gamma_n = \gamma_1/(1 + c_2 n)$ where $c_2 n = 1$ after 5 epochs and where for SGLD, $\gamma_1 = 0.001$ for L-RMSprop and L-Adam and $\gamma_1 = 0.1$ for L-Adadelta.

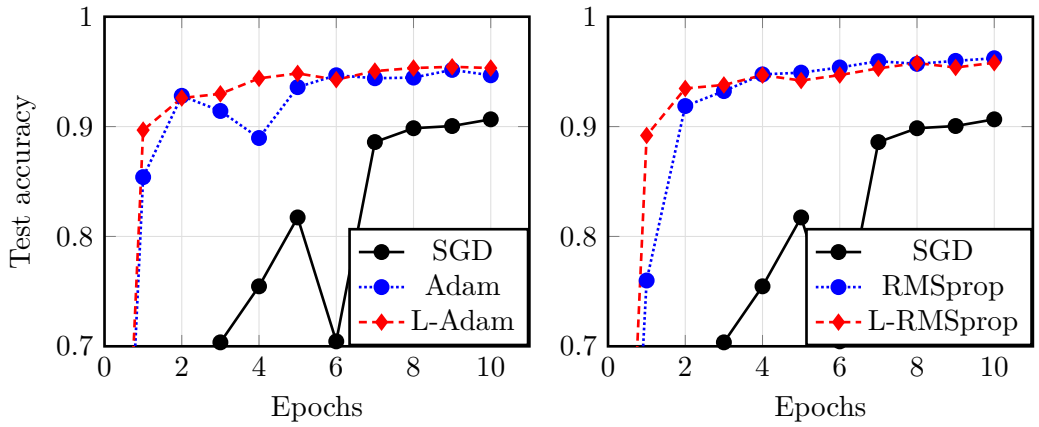


Figure 5.8: Side-by-side comparison of optimization algorithms with their respective Langevin counterparts for the training of a deep neural network on the MNIST dataset. We display the performance of SGD for reference. The values of the hyperparameters are $a(n) = A\log^{-1/2}(c_1 n + e)$ with $A = 1.10^{-3}$ for L-Adam and $A = 5.10^{-4}$ for L-RMSprop and where $c_1 n = 1$ after 5 epochs; $\gamma_n = \gamma_1/(1 + c_2 n)$ where $c_2 n = 1$ after 5 epochs and where $\gamma_1 = 0.01$ for SGLD and $\gamma_1 = 0.001$ for the others.

the performances of the optimization algorithm, although it adds some regularization. In particular, we could not reproduce the good results of [LCCC16] for feedforward neural networks, as it is also noted in a footnote in [MO17]. However, for deep neural networks, which are highly non-linear and which loss function has many local minima, the Langevin version is competitive with the currently widely used non-Langevin algorithms and can even lead to improvements. The results are given in Figure 5.8 where we used a deep neural network with 20 hidden layers of 32 units each and with ReLU activation. This aspect shall be developed in a Chapter 6.

5.3.4 Image classification (2)

We train a convolutional neural network on the CIFAR-10 dataset [KH09], which is composed of RGB images of size 32×32 belonging to 10 different classes: airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. 50000 images are used for training and 10000 images for test. We train a deep neural network composed of two convolutional layers with 4×4 kernel size and 32 channels for each; 2×2 max-pooling is used after each convolutional layer, which is a standard network configuration [JKRL09]. These layers are followed by 20 hidden dense layers with 64 units each and with ReLU activation. Since the images in the CIFAR10 dataset do not have a good resolution, we cannot expect a very high accuracy on the test set.

We proceed to side-by-side comparison of Langevin algorithms with their respective non-Langevin counterparts. The results are given in Figure 5.9 and show that in the case of a deep neural network with a large number of hidden layers, preconditioned Langevin optimizers achieve competing or even faster convergence speed than non-Langevin optimizers. For more developments we refer to Chapter 6.

5.3.5 Time series analysis and prediction

We use the Jena Climate dataset which is a weather time series dataset recorded at the weather station of the Max Planck Institute for Biogeochemistry in Jena, Germany. This dataset contains 14 different features such as air temperature, atmospheric pressure, humidity and wind. We also pre-process the dataset and add some relevant features for weather prediction such as the the

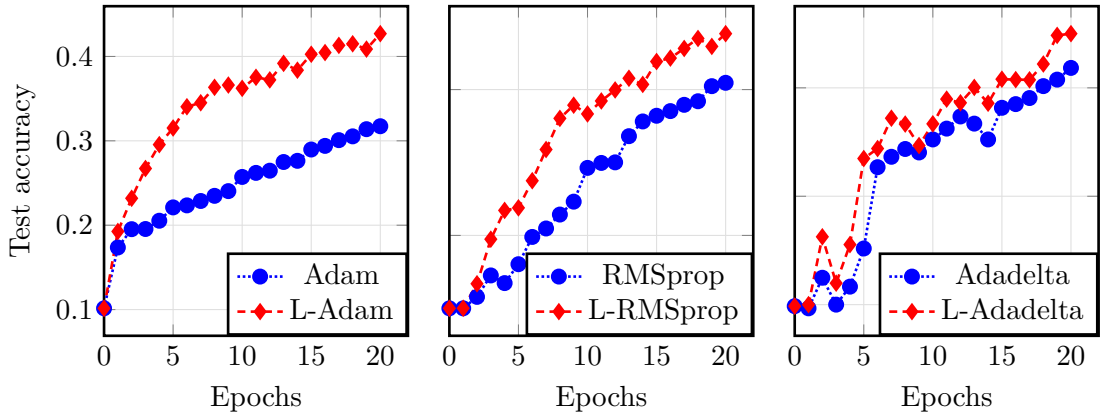


Figure 5.9: Side-by-side comparison of optimization algorithms with their respective Langevin counterparts for the training of a deep neural network on the CIFAR-10 dataset. The values of the hyperparameters are $a(n) = A\log^{-1/2}(c_1 n + e)$ with $A = 5.10^{-4}$ and where $c_1 n = 1$ after 5 epochs; $\gamma_n = \gamma_1/(1 + c_2 n)$ where $c_2 n = 1$ after 5 epochs and where $\gamma_1 = 1\,\mathrm{e}{-3}, 2\,\mathrm{e}{-3}, 1\,\mathrm{e}{-1}$ for Adam, RMSprop and Adadelta respectively

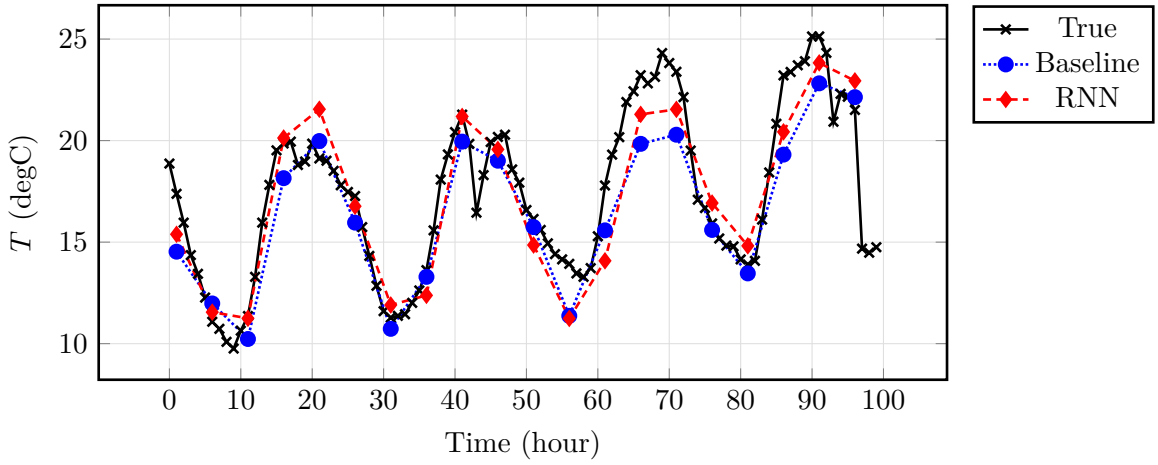


Figure 5.10: Example of hourly temperature prediction

hour in the day and the day in the year. We consider a recurrent neural network with one LSTM layer with 64 filters and with tanh activation and sigmoid recurrent activation. The loss function is the classic L^2-mean square loss.

We compare the performance of the RNN model with the performance of the baseline model which is the model obtained by simply predicting the future value by the current value. An example of prediction of the temperature using a RNN is given in Figure 5.10.

We compare the performances of the preconditioned Langevin algorithms with the vanilla SGLD algorithm in Figure 5.11 and in Table 5.2, showing that preconditioned optimizers achieve a faster convergence speed.

Preconditioner	SGLD	L-RMSprop	L-Adam	L-Adadelta	Baseline model
Best test loss	0.739	0.434	0.433	0.467	0.545

Table 5.2: Best accuracy performance on the JENA weather test set after 10 epochs.

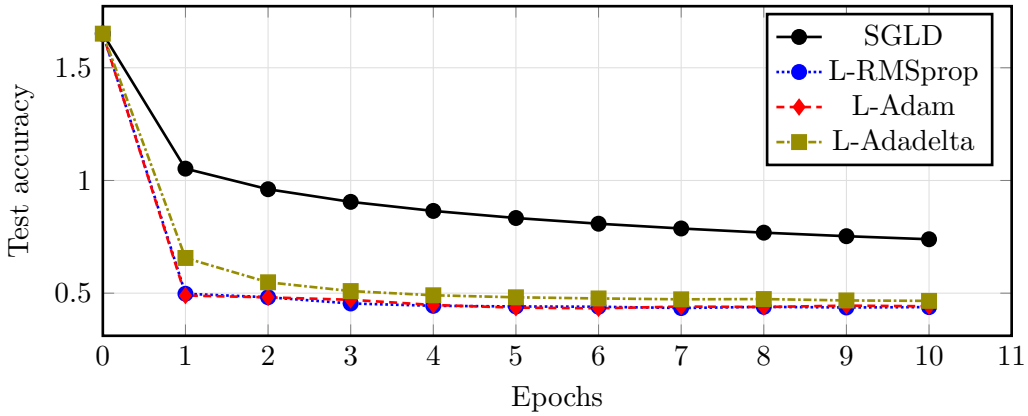


Figure 5.11: Performance of preconditioned Langevin algorithms compared with vanilla SGLD on the JENA weather dataset. The values of the hyperparameters are $a(n) = A\log^{-1/2}(c_1 n + e)$ with $A = 2.10^{-3}$ and where $c_1 n = 1$ after 5 epochs; $\gamma_n = \gamma_1/(1 + c_2 n)$ where $c_2 n = 1$ after 5 epochs and where for SGLD, $\gamma_1 = 0.001$ for L-RMSprop and L-Adam and $\gamma_1 = 0.1$ for L-Adadelta.

5.3.6 Optimal Quantization

We consider the problem of vector optimal quantization as stated in Section 1.1.2.5, which does not involve neural networks. For μ some probability distribution in $L^2(\mathbb{R}^q)$ and for fixed $K \in \mathbb{N}$, the objective is to minimize the distortion function:

$$\min_{x=(x^1,\dots,x^K)\in(\mathbb{R}^q)^K} V(x) := \frac{1}{2}\int_{\mathbb{R}^q} \min_{1\le k\le K} |\xi - x^k|^2 \mu(d\xi). \tag{5.3.1}$$

Then defining the Voronoï partition of $\mathbb{R}^q$:

$$V_k(x^1,\dots,x^K) := \left\{\xi \in \mathbb{R}^q |\ \forall\ 1 \le j \le K,\ |\xi - x^k| \le |\xi - x^j|\right\},\ 1 \le k \le K,$$

the corresponding gradient descent algorithm reads with $X_n = (X_n^1,\dots,X_n^K) \in (\mathbb{R}^q)^K$ and $Y_n \sim \mu$ and iid:

$$X_{n+1} = X_n - \gamma_{n+1}\left[\mathbb{1}_{Y_{n+1}\in V_k(X_n)}(X_n^k - Y_{n+1})\right]_{1\le k\le K}. \tag{5.3.2}$$

Optimal quantization problems involve the minimization of a potential function (distortion) which is highly non-convex and with many local minima and saddle points. Indeed, if $x = (x^1,\dots,x^K) \in (\mathbb{R}^q)^K$ is a global minimizer, then by symmetry and for every $\tau \in \mathfrak{S}_K$ the group of permutations of $\{1,\dots,K\}$, $(x^{\tau(1)},\dots,x^{\tau(K)})$ is also a minimizer. It follows from the Mountain Pass theorem [AR73] the existence of critical points which are not global minimizer. Moreover, the set of global minimizers is even larger if the law μ to be quantified has its own symmetry properties. Typically for $\mu = \mathcal{N}(0, I_q)$, if $x = (x^1,\dots,x^K) \in (\mathbb{R}^q)^K$ is a global minimizer then for every linear isometry B, $(Bx^1,\dots,Bx^K)$ is also a global minimizer. We also refer to [GL00, Section 5] for a discussion on the uniqueness of optimal quantizers (up to permutation) in the one-dimensional case. An illustration in a simple case is given in Figure 5.13.

Therefore, Langevin algorithms may be suitable for improving the CLVQ algorithm. In Figure 5.12, we consider the optimal quantization problem with $\mu = \mathcal{N}(0, I_q)$ in dimension $q = 20$ with $K = 1\,\mathrm{e}\,4$ quantization points and we proceed to side-by-side comparison of Langevin and non-Langevin optimizers, showing that Langevin methods improve the performances of classic gradient descent methods for optimal quantization. We plot the two-dimensional projection of the quantizers in Figure 5.14.

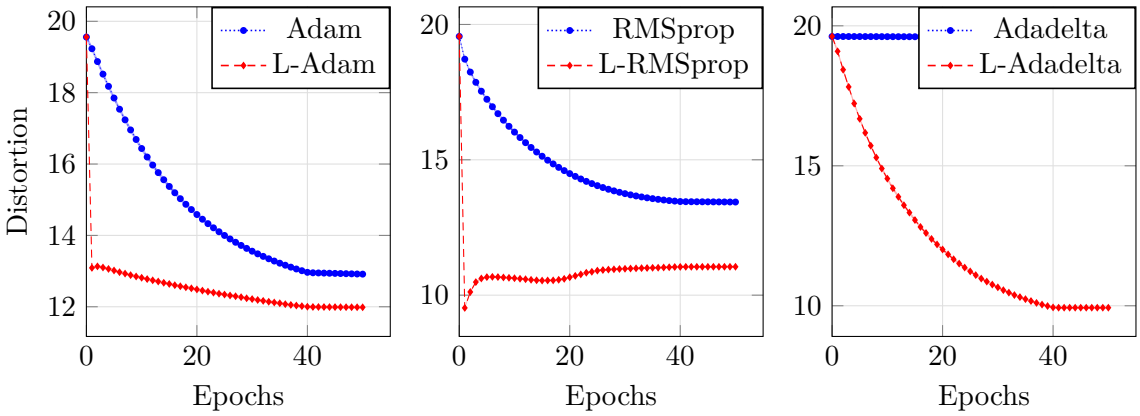


Figure 5.12: Side-by-side comparison of optimization algorithms with their respective Langevin counterparts for the quantization of the law $\mathcal{N}(0, I_{20})$ with 10000 quantization points. The batch size is 512 and each epochs consists in 5 batches. At the end of each epoch, the distortion V is evaluated over 25×512 samples and the 95%-confidence intervals are indicated (but may be too small to be visible). The schedules are $\gamma_n = 5\,\mathrm{e}{-3}$ ($2\,\mathrm{e}{-2}$) and $a(n) = 5\,\mathrm{e}{-3}$ ($5\,\mathrm{e}{-2}$) for epochs 0 to 40 and $\gamma_n = 5\,\mathrm{e}{-4}$ ($2\,\mathrm{e}{-3}$) and $a(n) = 0$ beyond for Adam and RMSprop (for Adadelta).

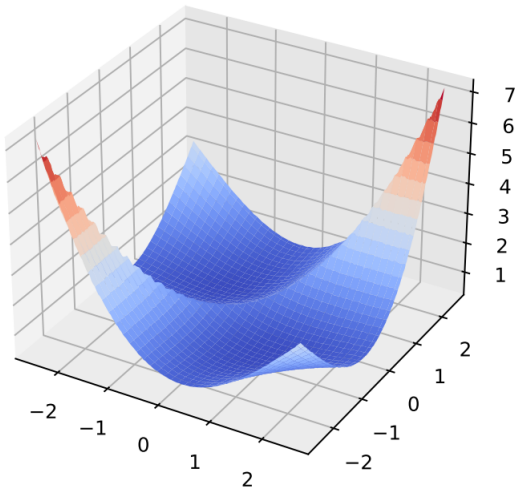


Figure 5.13: Distortion function for $\mu = \mathcal{N}(0, 1)$ and $K = 2$.

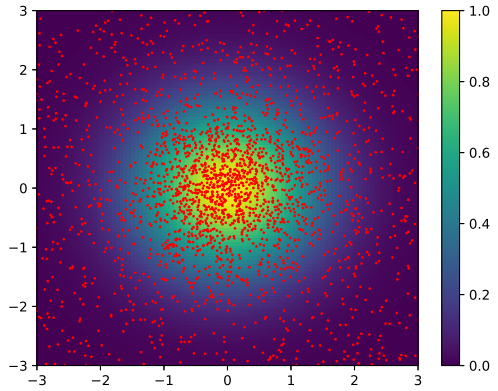


Figure 5.14: Projection on the first two axes of the quantization of $\mathcal{N}(0, I_{20})$ with 10000 quantization points.

Chapter 6

Langevin algorithms for very deep Neural Networks with application to image classification

The results in this chapter have been presented at the *International Neural Network Society Workshop on Deep Learning Innovations and Applications* (INNS DLIA), part of the Machine Learning conference *International Joint Conference on Neural Networks* IJCNN 2023, and will be published in the first edition of the INNS workshop series in *Procedia Computer Science*. An arXiv preprint is available [Bra23].

Abstract

Training a very deep neural network is a challenging task, as the deeper a neural network is, the more non-linear it is. We compare the performances of various preconditioned Langevin algorithms with their non-Langevin counterparts for the training of neural networks of increasing depth. For shallow neural networks, Langevin algorithms do not lead to any improvement, however the deeper the network is and the greater are the gains provided by Langevin algorithms. Adding noise to the gradient descent allows to escape from local traps, which are more frequent for very deep neural networks. Following this heuristic we introduce a new Langevin algorithm called Layer Langevin, which consists in adding Langevin noise only to the weights associated to the deepest layers. We then prove the benefits of Langevin and Layer Langevin algorithms for the training of popular deep residual architectures for image classification.

6.1 Introduction

Langevin algorithms are widely used for the training of neural networks in a Bayesian setting [WT11, VBB+20]. Adding a small exogenous noise adds regularization to the training and allows to quantify the degree of uncertainty on the parameters. In this paper, we consider Langevin algorithms directly used for stochastic optimization of neural networks in a non-Bayesian setting

and compare their performances with non-Langevin stochastic gradient algorithms. As it was noted in [NVL+15, Ani19], adding gradient noise can in fact improve the learning. Similarly, noisy activation functions [GMDB16, SLH+19] may yield better learning for very deep neural networks. Indeed, the noise provides regularization and allows to escape from traps for the gradient descent such as local minima and saddle points [DPG+14]. Moreover, the deeper the neural network is, the more non-linear it is, thus increasing the number of such traps. Non-convex optimization through Langevin algorithms shares heuristics with simulated annealing which consists in sampling with respect to a Gibbs measure where the noise parameter gradually decreases to zero, see [vLA87] and Chapter 2.

Many advances in supervised learning were made possible using very deep neural networks, which are able to tackle much more difficult problems than shallow ones [KSH12, MPCB14, LBH15], in particular as it comes to image classification [SLJ+15, SZ15, HZRS16, HLVDMW17]. Still, deep neural networks which consist in a succession of dense layers are considerably more difficult to train [GB10, DPG+14] and may run into vanishing gradient problems [Hoc91, Han18]. Without proper adaptation or training, they show poor performance. To cope with this issue, highway networks [SGS15] and residual networks [HZRS16] were introduced. Their many successive layers behaves either as a dense layer or as the identity function, allowing the gradient information to propagate trough the successive layers.

We compare the benefits of preconditioned Langevin algorithms [LCCC16] for various architectures and depths of neural networks and we proceed to side-to-side comparison of Langevin algorithms with their respective non-Langevin counterparts. The purpose of our experiments is to compare different methods on the same model architecture, not to achieve state-of-the-art results. For shallow networks, there is no benefit in using Langevin algorithms as it only adds noise to the gradient descent and brings a less accurate estimation of the minimum. However, we observe that the deeper the network is, the greater are the gains provided by Langevin algorithms.

Since the most important non-linearities of the network are contained in the deepest layers, we introduce a new optimization method that we call Layer Langevin algorithm, which consists in training the network by adding Langevin noise only to the training of some layers and not to the other layers. In particular, we choose the Langevin layers to be the k first (deepest) layers for some integer k. We then highlight the possibilities of training acceleration using Langevin and Layer Langevin methods on deep residual networks [HZRS16] and dense convolutional networks [HLVDMW17] for image classification.

Our code for the numerical experiments is available at

`https://github.com/Bras-P/deep-layer-langevin`.

It includes in particular ready-to-use Langevin optimizers and Layer Langevin optimizers as instances of the TensorFlow `Optimizer` base class and a demonstration notebook.

6.2 Very deep neural networks

Training of very deep neural networks is a significantly more challenging task than for shallow networks [GB10, DPG+14]. Let us write the output of a neural network with K layers and with weights $\theta = (\theta^1, \dots, \theta^K)$ as

$$\psi_\theta(x) = \varphi^K_{\theta^K} \circ \cdots \circ \varphi^1_{\theta^1}(x), \tag{6.2.1}$$

where $\varphi^1, \dots, \varphi^K$ are activation function and where $\varphi^k_{\theta^k} : x \mapsto \varphi^k(\theta^k \cdot x)$ at every unit. Denoting

$$\Phi_k(x) := \varphi^k_{\theta^k} \circ \cdots \circ \varphi^1_{\theta^1}(x) \tag{6.2.2}$$

for $1 \leq k \leq K$ and $\Phi_0(x) := x$, then the gradient reads for $1 \leq k \leq K$:

$$\nabla_{\theta^k} \psi_\theta(x) = (\nabla_{\theta^K} \varphi_{\theta^K} \circ \Phi_{K-1}(x)) \cdot \cdots \cdot (\nabla_{\theta^k} \varphi_{\theta^k} \circ \Phi_{k-1}(x)) \,. \tag{6.2.3}$$

Thus heuristically, the deeper the layer is, the more the gradient with respect to the parameters of this layer has annealing points, since more factors appear in (6.2.3), hinting that deep layers show more non-linearities and local traps.

6.3 Langevin algorithms for the training of deep neural networks

6.3.1 Experimental setting

In our experiments we use the following datasets. The MNIST dataset [LBBH98] is composed of 28×28 grayscale images of handwritten digits (from 0 to 9). 60.000 images are used for training and 10.000 images are used for test. The CIFAR-10 and the CIFAR-100 datasets [KH09] consist in RGB images of size 32×32 belonging to 10 and 100 different classes respectively. For both datasets 50.000 images are used for training and 10.000 images are used for test.

The neural networks are trained using preconditioned Langevin algorithms with per-dimension adaptive stepsize [LCCC16] with different choices of preconditioner. That is, for a preconditioner rule (P_n) the Langevin update reads

$$g_{n+1} = \nabla_\theta V(\theta_n; \mathcal{D}_{n+1}) \tag{6.3.1}$$

$$\theta_{n+1} = \theta_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1} + \sigma \sqrt{\gamma_{n+1}} \mathcal{N}(0, P_{n+1}), \tag{6.3.2}$$

where $\sigma \in (0, \infty)$ controls the amount of injected noise, (γ_n) is the non-increasing learning rate sequence, V denotes the objective function and where $\nabla_\theta V(\theta_n; \mathcal{D}_n)$ stands for the mean gradient computed on a subset $\mathcal{D}_n$ of the dataset. The corresponding preconditioned non-Langevin algorithm follows the same update as in (6.3.2) without Gaussian noise. In our experiments we use the RMSprop [DHS11, LCCC16], the Adam [KB15] and the Adadelta [Zei12] preconditioners and we call the Langevin version of these algorithms as L-RMSprop, L-Adam and L-Adadelta respectively. The preconditioner rules are given in Algorithms 4, 5, 6 respectively. Note that depending on the algorithm version, in the update (6.3.2) the gradient g_{n+1} can be replaced by an averaged gradient over the past iterations as this in the case in Adam (Algorithm 5) i.e. momentum gradient is used. While comparing some preconditioned method with its Langevin counterpart, we ensure that both training procedures start with the same initial weights.

Algorithm 4 RMSprop update

Parameters: $\alpha, \lambda > 0$

$\mathrm{MS}_{n+1} = \alpha \, \mathrm{MS}_n + (1-\alpha) g_{n+1} \odot g_{n+1}$

$P_{n+1} = \mathrm{diag}\left(\mathbb{1} \oslash (\lambda \mathbb{1} + \sqrt{\mathrm{MS}_{n+1}})\right)$

$\theta_{n+1} = \theta_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1}$

Algorithm 5 Adam update

Parameters: $\beta_1, \beta_2, \lambda > 0$

$M_{n+1} = \beta_1 M_n + (1-\beta_1) g_{n+1}$

$\mathrm{MS}_{n+1} = \beta_2 \, \mathrm{MS}_n + (1-\beta_2) g_{n+1} \odot g_{n+1}$

$\widehat{M}_{n+1} = M_{n+1} / (1 - \beta_1^{n+1})$

$\widehat{\mathrm{MS}}_{n+1} = \mathrm{MS}_{n+1} / (1 - \beta_2^{n+1})$

$P_{n+1} = \mathrm{diag}\left(\mathbb{1} \oslash (\lambda \mathbb{1} + \sqrt{\widehat{\mathrm{MS}}_{n+1}})\right)$

$\theta_{n+1} = \theta_n - \gamma_{n+1} P_{n+1} \cdot \widehat{M}_{n+1}.$

6.3.2 Plain and convolutional networks

We first train fully connected feedforward neural networks on the MNIST dataset. The networks are composed of 3, 20, 30 and 40 hidden dense layers respectively with 64 units each and with

Algorithm 6 Adadelta update

Parameters: $\beta_1, \beta_2, \lambda > 0$
$\mathrm{MS}_{n+1} = \beta_1 \mathrm{MS}_n + (1-\beta_1) g_{n+1} \odot g_{n+1}$
$P_{n+1} = \mathrm{diag}\left(\widehat{MS}_n + \lambda \mathbb{1} \oslash (\lambda \mathbb{1} + \sqrt{\widehat{\mathrm{MS}}_n})\right)$
$\theta_{n+1} = \theta_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1}$.
$\widehat{MS}_{n+1} = \beta_2 \mathrm{MS}_n + (1-\beta_2)(\theta_{n+1} - \theta_n) \odot (\theta_{n+1} - \theta_n)$.

ReLU activation, followed by one dense output layer. The results are given in Figure 6.1. We observe that for shallow neural networks, Langevin algorithms do not outperform their respective non-Langevin counterparts; they add noise to the gradient descent thus giving a less accurate estimate of the minimum value. In particular and as noted in the footnote in [MO17], we could not reproduce the good results from [LCCC16] for plain networks with two hidden layers. However, the deeper the network is, the greater the gains induced by Langevin algorithms compared with their respective non-Langevin counterparts are. We also display the value of the loss function on the training set in order to highlight that the better performances of the Langevin algorithms are not due to some overfitting effect. Langevin algorithms indeed show improvements on 20-layer deep networks; beyond 30-layer deep networks, the gains are significant. The training of 40-layer deep networks with non-Langevin algorithms may run into the vanishing gradient problem, whereas such problem is avoided by Langevin algorithms. In the latter case of very deep training, preconditioned Langevin algorithms not only add noise preventing the vanishing of the gradient, they also help starting up the training in the right directions. To obtain better results with Langevin algorithms, we recommend using a small coefficient σ, empirically ranging from $1\,\mathrm{e}{-3}$ to $5\,\mathrm{e}{-5}$.

We then perform simulations in a similar setup on convolutional architectures that are more adapted to image recognition [JKRL09] followed by a large number of hidden dense layers. More specifically, we train neural networks composed of two convolutional layers with 4×4 kernel size and 32 channels for each; 2×2 max-pooling is used after each convolutional layer. These layers are followed by respectively 10 and 30 hidden dense layers with 64 units each and by one dense output layer. Since the images in the CIFAR-10 dataset do not have a good resolution, we cannot expect a very high accuracy on the test set. Instead, we focus on comparing different algorithms on the same model architecture. The results are given in Figure 6.2 and we make similar observations: Langevin algorithms show improvements with 10 hidden dense layers and for 30 dense layers, non-Langevin algorithms run into vanishing gradient issues which is not the case for Langevin algorithms.

6.3.3 Highway networks

We now perform the same simulations on Highway networks, in a setting very similar to [SGS15]. Comparably to residual networks, the output of a highway layer is a convex combination of the output of a dense layer and the output of a identity layer; the parameter controlling the convex combination is itself trainable. For a layer with weights (θ_D, θ_T), the output reads

$$y = T_{\theta_T}(x) \cdot D_{\theta_D}(x) + (1 - T_{\theta_T}(x)) \cdot x, \tag{6.3.3}$$

where T and D are dense layers and where T has sigmoid output.

We give the implementation of a Highway layer as a subclass of the base TensorFlow class `tf.keras. layers.Layer`. The method `build` initializes the weights of the layer when the layer is called for the first time; each weight is defined through the method `add_weight`. The method `call` defines the output of the layer depending on the inputs and its weights; we need to resort

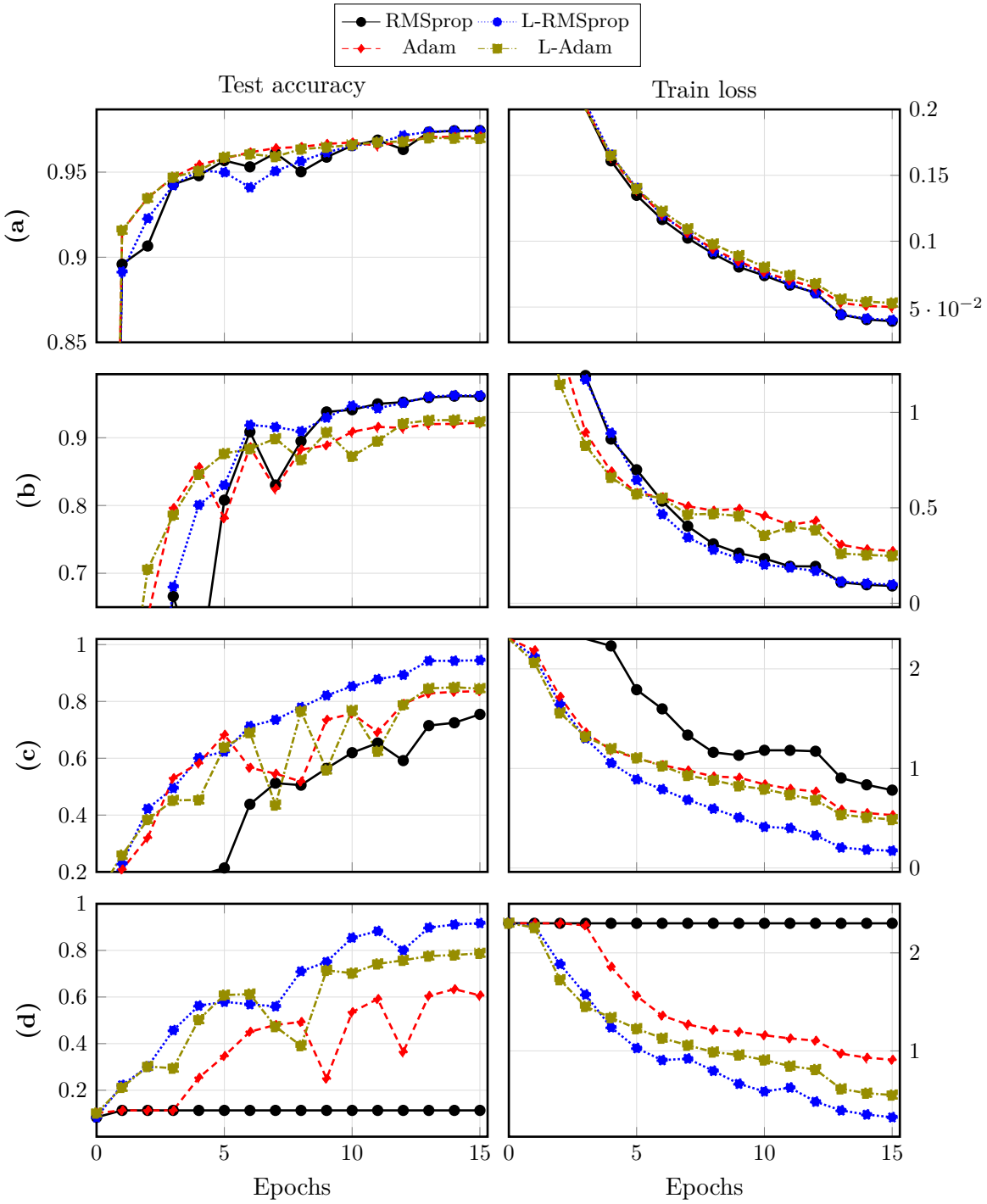


Figure 6.1: Training of neural networks of various depths on the MNIST dataset using Langevin algorithms compared with their non-langevin counterparts. (a): 3 hidden layers, (b): 20 hidden layers, (c): 30 hidden layers, (d): 40 hidden layers. The batch size is 512. The schedules are $\gamma_n = 1\,\mathrm{e}\,{-3}$ and $\sigma = 5\,\mathrm{e}\,{-4}$ for epochs 1 to 12 and $\gamma_n = 1e-4$ and $\sigma = 0$ beyond.

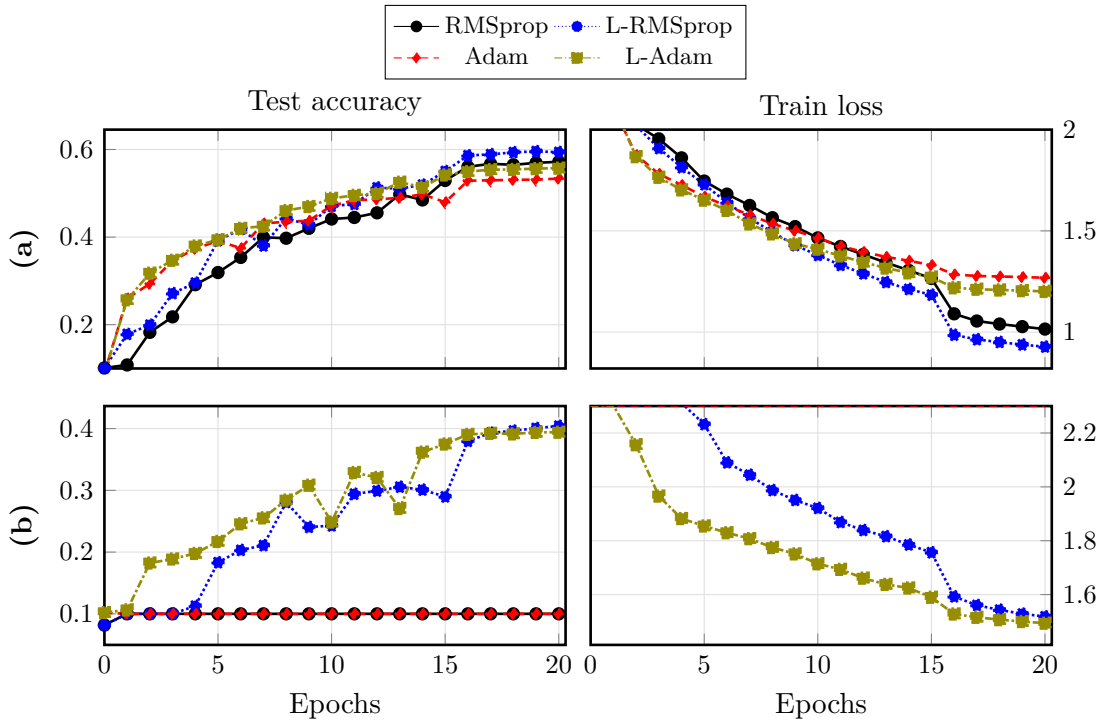


Figure 6.2: Training of convolutional neural networks on the CIFAR-10 dataset. (a): 10 hidden dense layers, (b): 30 hidden layers. The batch size is 512. The schedules are $\gamma_n = 1\,\mathrm{e}{-3}$ and $\sigma = 2e-4$ for epochs 1 to 15 and $\gamma_n = 1e-4$ and $\sigma = 0$ beyond.

to TensorFlow functions so that the gradient with respect to the layer weights is computed using the TensorFlow automatic differentiation.

```
class Highway(tf.keras.layers.Layer):

    def __init__(self, activation=None, tgBias=-1.):
        super(Highway, self).__init__()
        self.activation = tf.keras.activations.get(activation)
        self.tgActivation = tf.keras.activations.get('sigmoid')
        self.tgBias_init = tgBias

    def build(self, input_shape):
        dim = input_shape[-1]
        self.kernel = self.add_weight("kernel", shape=[dim, dim])
        self.bias = self.add_weight("bias", shape=[dim,])
        self.tgKernel = self.add_weight("tgKernel", shape=[dim,dim])
        self.tgBias = self.add_weight("tgBias", shape=[dim], initializer
   =tf.keras.initializers.Constant(self.tgBias_init))
        self.built = True

    def call(self, inputs):
        outputs = tf.matmul(inputs, self.kernel)
        outputs = tf.nn.bias_add(outputs, self.bias)
        if self.activation is not None:
            outputs = self.activation(outputs)
        transform_gate = tf.matmul(inputs, self.tgKernel)
        transform_gate = tf.nn.bias_add(transform_gate, self.tgBias)
        transform_gate = self.tgActivation(transform_gate)
```

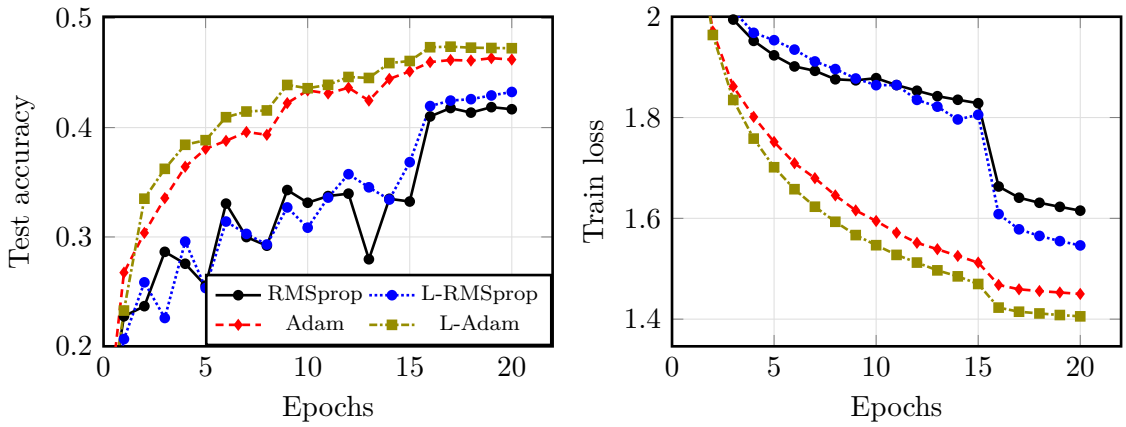


Figure 6.3: Training of a highway neural network with 80 dense hidden layers. The schedules are $\gamma_n = 1\,\mathrm{e}{-3}$ and $\sigma = 1e-4$ for epochs 1 to 15 and $\gamma_n = 1e-4$ and $\sigma = 0$ beyond.

```
        outputs = tf.math.multiply(outputs, transform_gate) + tf.math.
    multiply(inputs, tf.math.add(1.,-transform_gate))
        return outputs
```

We observe that Langevin algorithms become effectively faster than non-Langevin algorithms only from a larger depth than for plain networks. In Figure 6.3 we plot the results for the training of a network composed of 80 dense hidden layers with 64 units each and ReLU activation on the CIFAR-10 dataset, showing the possibilities of acceleration through Langevin algorithms, even in a residual (highway) architecture.

6.4 Layer Langevin algorithm

We introduce a new Langevin algorithm for stochastic optimization of deep neural networks that we call Layer Langevin algorithm. Choosing a preconditioner rule P, some weights are updated following the Langevin rule while the other weights are updated following the non-Langevin rule. Denoting $\theta_n^{(i)}$ the ith weight at step n, we have for every i:

$$\theta_{n+1}^{(i)} = \theta_n^{(i)} - \gamma_{n+1}[P_{n+1} \cdot g_{n+1}]^{(i)} + \mathbb{1}_{i\in\mathcal{J}}\sigma\sqrt{\gamma_{n+1}}[\mathcal{N}(0, P_{n+1})]^{(i)}, \tag{6.4.1}$$

where $\mathcal{J}$ is a subset of weight indices and where P_n denotes the preconditioner. To simplify the choice of $\mathcal{J}$, we choose $\mathcal{J}$ as the subset of indexes of weights belonging to some layers. However, a finer control over the subset $\mathcal{J}$ remains possible. To implement this method in practice, we simply assign before the training an attribute equals to $\mathbb{1}_{i\in\mathcal{J}}$ to every trainable variable of the network.

We compare the performances of Layer Langevin algorithms with the Adam preconditioner for different choices of the subset of layers. The results are given in Figure 6.4 for the training of a dense network with 30 hidden dense layers on the MNIST dataset in a setting similar to Figure 6.1. For some optimizer *Name*, we denote LL-*Name* p% the corresponding Layer Langevin algorithm where the subset $\mathcal{J}$ is the first p% layers of the network. We observe that we obtain significant gains in comparison with the vanilla Langevin algorithm and that the best performances are obtained when choosing the subset $\mathcal{J}$ as being the first ℓ layers for some $\ell \in \mathbb{N}$, in particular all the layers of the network except the few last ones.

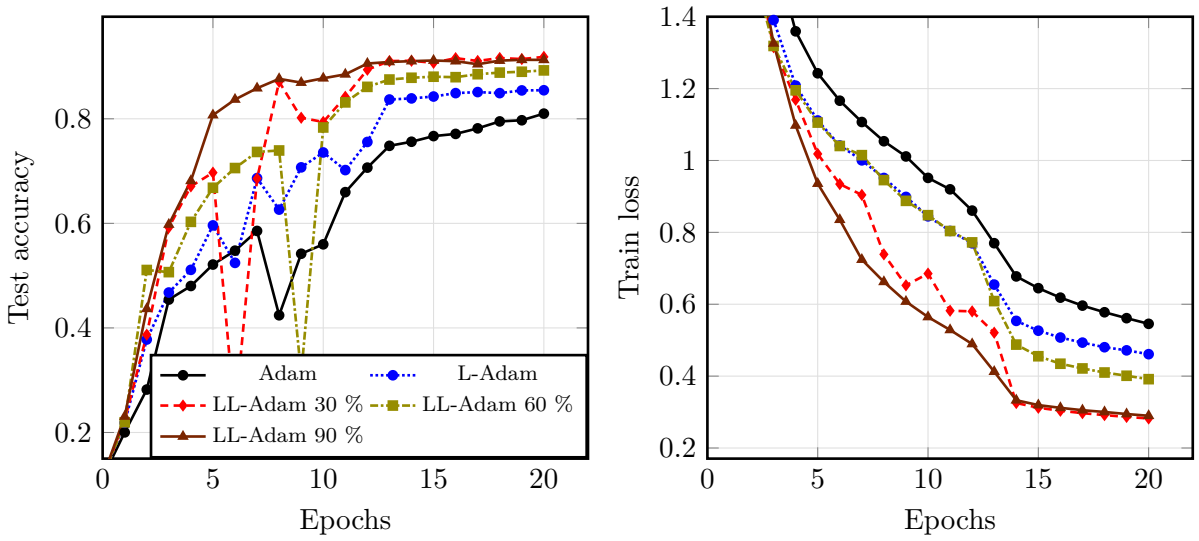


Figure 6.4: Layer Langevin method comparison on a dense neural network with 30 hidden layers. The schedules are $\gamma_n = 1\,\mathrm{e}\,{-3}$ and $\sigma = 5e-4$ for epochs 1 to 13 and $\gamma_n = 1e-4$ and $\sigma = 0$ beyond.

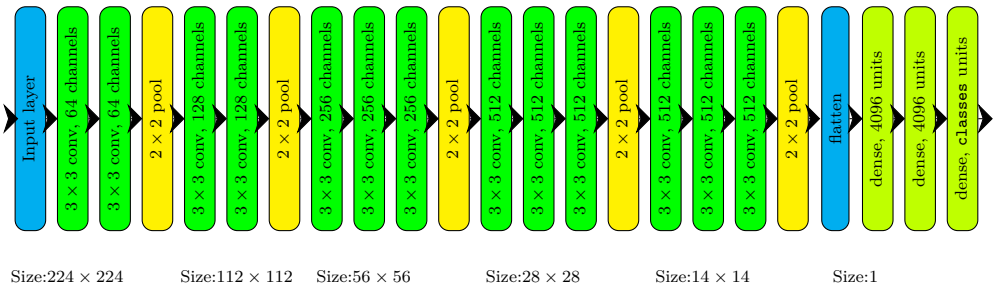


Figure 6.5: Architecture of the VGG-16 network for an input image of size 224×224. The architecture is composed of 5 convolution blocks; inside each block the number of channels is kept constant. At each transition between two blocks, the height and the width of the image are both divided by 2 while the number of channels is multiplied by 2 (except for the last block). Two hidden fully connected layers are used at the end.

6.5 Application to deep architectures for image recognition

We now test the Layer Langevin algorithm to speed up the training of neural networks with very deep architectures that are popular in image recognition.

Many advances in supervised learning were made possible using very deep neural networks, which are able to tackle much more difficult problems than shallow ones. In particular, let us focus on very deep convolutional networks for image recognition. VGG (Visual Geometry Group) networks were introduced in [SZ15] and then became a standard for deep convolutional neural networks. The VGG architecture consists in a succession of 2D convolutional layers with kernel size 3×3 and with ReLU activation; a batch normalization layer is applied before each convolutional layer. The dimensions of the input image are gradually reduced using 2×2 pooling layers while the number of channels is increased. At the end of the network, dense layers are applied for class prediction. An illustration of an example of VGG architecture is given in Figure 6.5.

However, as this is the case for fully connected neural networks, very deep convolutional

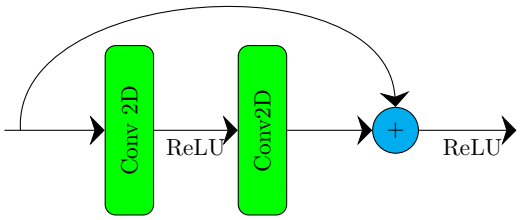


Figure 6.6: ResNet elementary residual block

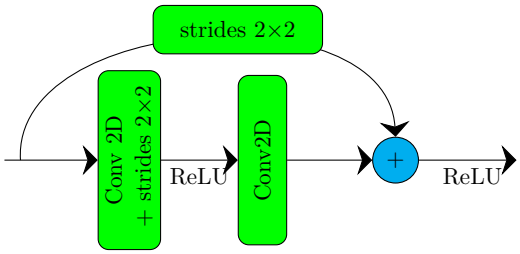


Figure 6.7: ResNet size reduction block

networks are considerably more difficult to train. To cope with this issue, residual networks (ResNets) were introduced in [HZRS16]. In this architecture, in order to propagate the gradient information through the many layers a residual learning is implemented every few stacked layers. More precisely, ResNets are composed of a succession of blocks as described in Figure 6.6. Each of these blocks includes a shortcut connection and operates in part as a simple identity layer. Contrary to a highway connection, the coefficient of the shortcut connection is 1 and is non-trainable. Similarly to VGG nets, the size of the image is gradually reduced using convolutional layers with 2×2 strides, see Figure 6.7. An example of ResNet architecture is given in Figure 6.8. Residual networks allow substantially deeper architectures and gain accuracy from increased depth [HZRS16].

An implementation of residual networks using the TensorFlow functional API is given in Listing 6.1. The (main) arguments are:

- `block_layers`: list of the number of blocks between two size reductions. For example, `[2,3,3]` yields the architecture in Figure 6.8.
- `filters`: initial number of channels which is multiplied by 2 at every size reduction.
- `mode`: either `'resent'` or `'vgg'`, in order to instantiate either the ResNet of the VGG

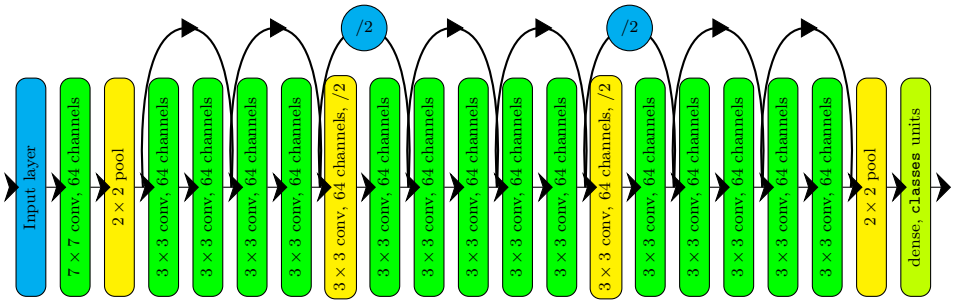


Figure 6.8: Example of (reduced) residual network architecture.

architecture.

An implementation of the VGG architecture can be directly obtained from Listing 6.1 by skipping the shortcut connections in `identity_block` and `convolutional_block`.

```
class ResNet(ModelBuilder): # can also instantiate the VGG model
    def __init__(
        self,
        input_shape = (32,32,3),
        filters = 32,
        block_layers = [4,4,4],
        classes = 10,
        zero_padding = (0,0),
        mode = 'resnet' # equals to 'resnet' or to 'vgg'
    ):
        self.input_shape = input_shape
        self.filters = filters
        self.block_layers = block_layers
        self.classes = classes
        self.zero_padding = zero_padding
        self.mode = mode

    def identity_block(self, x, filter_size):
        x_skip = x
        x = tf.keras.layers.Conv2D(filter_size, (3,3), padding = 'same')
    (x)
        x = tf.keras.layers.BatchNormalization(axis=3)(x)
        x = tf.keras.layers.Activation('relu')(x)
        x = tf.keras.layers.Conv2D(filter_size, (3,3), padding = 'same')
    (x)
        x = tf.keras.layers.BatchNormalization(axis=3)(x)
        if not self.mode=='vgg':
            x = tf.keras.layers.Add()([x, x_skip])
        x = tf.keras.layers.Activation('relu')(x)
        return x

    def convolutional_block(self, x, filter_size):
        x_skip = x
        x = tf.keras.layers.Conv2D(filter_size, (3,3), padding = 'same',
     strides = (2,2))(x)
        x = tf.keras.layers.BatchNormalization(axis=3)(x)
        x = tf.keras.layers.Activation('relu')(x)
        x = tf.keras.layers.Conv2D(filter_size, (3,3), padding = 'same')
    (x)
        x = tf.keras.layers.BatchNormalization(axis=3)(x)
        if not self.mode=='vgg':
            x_skip = tf.keras.layers.Conv2D(filter_size, (1,1), strides
    = (2,2))(x_skip)
            x = tf.keras.layers.Add()([x, x_skip])
        x = tf.keras.layers.Activation('relu')(x)
        return x


    def getModel(self):
        filter_size = self.filters
        x_input = tf.keras.layers.Input(self.input_shape)
        x = tf.keras.layers.ZeroPadding2D(self.zero_padding)(x_input)
```

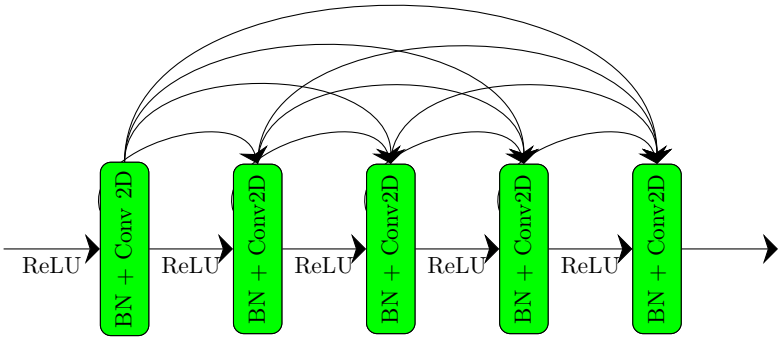


Figure 6.9: Dense convolutional network block

```
48        x = tf.keras.layers.Conv2D(filter_size, kernel_size=3, padding='
    same')(x) # kernel_size is 3 for CIFAR-10, 7 in general
49        x = tf.keras.layers.BatchNormalization()(x)
50        x = tf.keras.layers.Activation('relu')(x)
51
52        for i in range(len(self.block_layers)):
53            if i == 0:
54                # For sub-block 1 Residual/Convolutional block not
    needed
55                for j in range(self.block_layers[i]):
56                    x = self.identity_block(x, filter_size)
57            else:
58                filter_size = filter_size*2
59                x = self.convolutional_block(x, filter_size)
60                for j in range(self.block_layers[i] - 1):
61                    x = self.identity_block(x, filter_size)
62
63        x = tf.keras.layers.AveragePooling2D((2,2), padding = 'same')(x)
64        x = tf.keras.layers.Flatten()(x)
65        x = tf.keras.layers.Dense(self.classes)(x)
66
67        model = tf.keras.models.Model(inputs = x_input, outputs = x)
68        return model
```

Listing 6.1: ResNet implementation

Dense convolutional networks (DenseNets), introduced in [HLVDMW17], are composed of convolutional layers; within every DenseNet blocks, each block is connected to all the following layers inside the block, as described in Figure 6.9. Contrary to residual networks where simply adding the features implies some loss of information, the layers are not connected by summation but by concatenation, i.e. for $y_0, \dots, y_\ell$ the outputs of the previous layers and H_ℓ the ℓ^{layer} in the block which is composed of 2D convolution, activation and batch normalization, the output of the ℓ^{th} layer is $H_\ell([x_0, \dots, x_\ell])$ where $[x_0, \dots, x_\ell]$ denotes the concatenation of the (same sized) multi-channel images $x_0, \dots, x_\ell$, whereas the output for a resnet architecture is $H_\ell(x_\ell) + x_{\ell-1}$. This architecture allowing feature reuse and improves the flow of information throughout the network , which helps training of deeper architectures. Likewise VGG and residual networks, the different blocks are concatenated using a convolutional layer and a pooling layer so that the size of the image is progressively reduced. Dense convolutional networks obtained significant improvements over previous architectures while also reducing the number of parameters [HLVDMW17].

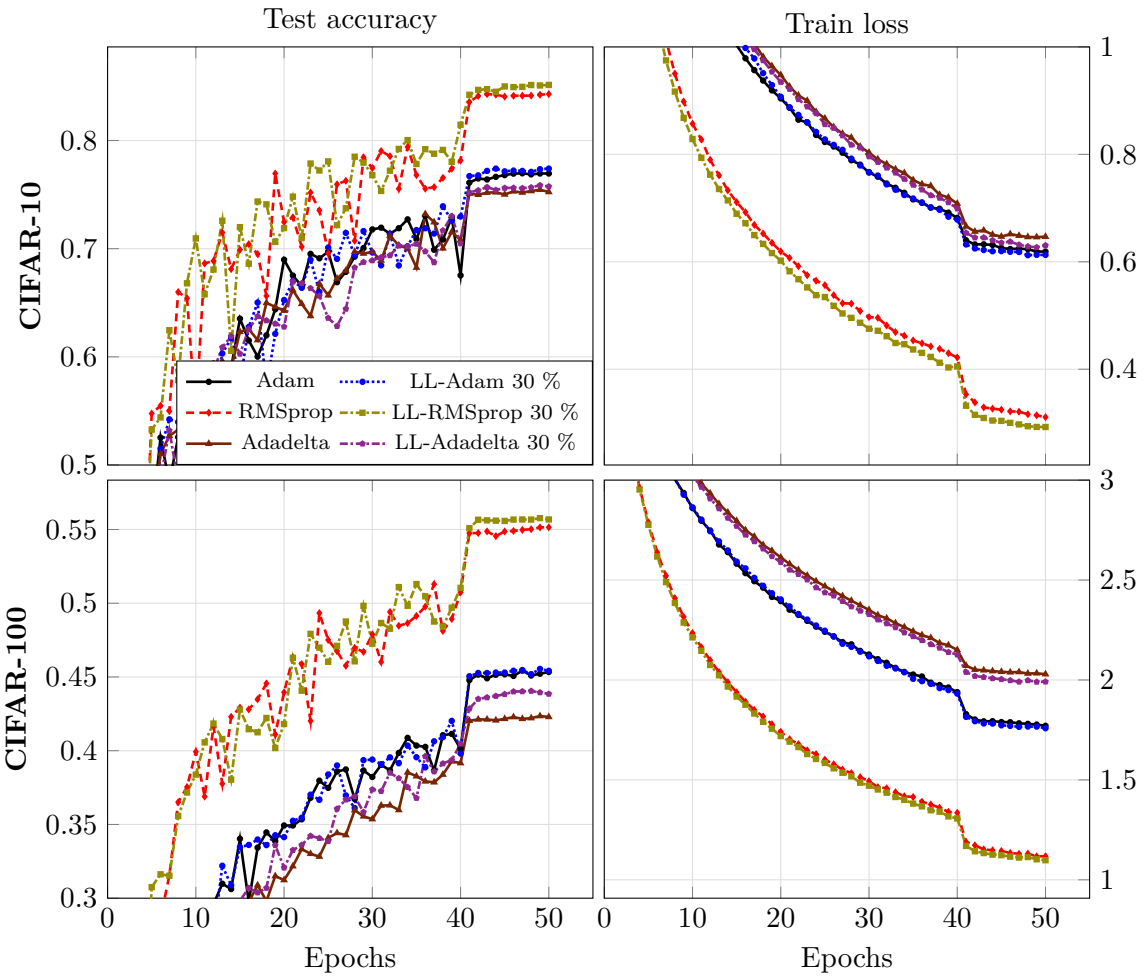


Figure 6.10: Layer Langevin method comparison for the training of ResNet-20. The initial number of channels is 16. The schedules are $\gamma_n = 1\,\mathrm{e}{-3}$ ($2\,\mathrm{e}{-1}$ for Adadelta) and $\sigma = 5\,\mathrm{e}{-4}$ ($5\,\mathrm{e}{-3}$ for Adadelta) for epochs 1 to 40 γ_n is divided by 10 and σ is set to 0 beyond.

We train on the CIFAR-10 dataset a ResNet architecture composed of 2 blocks with 5 residual layers each; each block is followed by a size reduction layer. This architecture is given as ResNet-20 in [HZRS16, Section 4.2]. Similarly, we train a DenseNet architecture with depth $L = 22$ and growth order $k = 5$ [HLVDMW17] on the CIFAR-10 dataset. We apply usual data augmentation to both CIFAR-10 and CIFAR-100 datasets [LXG+15, HZRS16]: 4 pixels are padded on each side and a 32×32 crop is randomly sampled from the padded image or its horizontal flip. The results are given in figure 6.10 and tables 6.1 and 6.2. Experiments show that Layer Langevin algorithms (in this case LL-Adam 30%) yield improvements in comparison with non-Langevin methods, even on residual architectures adapted to very deep learning. The train loss is also plotted, showing that the better performances of Layer Langevin is not only due to some overfitting effect.

Acknowledgements

I would like to thank Gilles Pagès for insightful discussions.

Table 6.1: Final test accuracy values obtained in Figures 6.10.

	Adam	LL-Adam	RMSprop	LL-RMSprop	Adadelta	LL-Adadelta
CIFAR-10	76.95 %	77.39 %	84.29 %	85.14 %	75.23 %	75.74 %
CIFAR-100	45.33 %	45.41 %	55.15 %	55.68 %	42.28 %	43.84 %

Table 6.2: Final test accuracy values on the CIFAR-10 dataset with DenseNet architecture. The schedules are $\gamma_n = 5\,\mathrm{e}\,{-3}$ ($2\,\mathrm{e}\,{-1}$ for Adadelta) and $\sigma = 5\,\mathrm{e}\,{-4}$ ($5\,\mathrm{e}\,{-3}$ for Adadelta) for epochs 1 to 40 and γ_n is divided by 10 and σ is set to 0 for epochs 41 to 50.

	Adam	LL-Adam	RMSprop	LL-RMSprop	Adadelta	LL-Adadelta
CIFAR-10	87.81 %	88.16 %	57.59 %	57.56 %	71.64 %	72.72 %

Chapter 7

Langevin algorithms for Markovian Neural Networks and Deep Stochastic control

The results in this chapter have been presented at the Machine Learning conference *International Joint Conference on Neural Networks* IJCNN 2023 and are published in the conference proceedings as a joint work with Gilles Pagès [BP23b].

Abstract

Stochastic Gradient Descent Langevin Dynamics (SGLD) algorithms, which add noise to the classic gradient descent, are known to improve the training of neural networks in some cases where the neural network is very deep. In this paper we study the possibilities of training acceleration for the numerical resolution of stochastic control problems through gradient descent, where the control is parametrized by a neural network. If the control is applied at many discretization times then solving the stochastic control problem reduces to minimizing the loss of a very deep neural network. We numerically show that Langevin algorithms improve the training on various stochastic control problems like hedging and resource management, and for different choices of gradient descent methods.

Keywords– Langevin algorithm, SGLD, Markovian neural network, Stochastic control, Deep neural network, Stochastic optimization.

7.1 Introduction

Stochastic Optimal Control (SOC), which consists in optimizing a functional of a trajectory of a controlled Stochastic Differential Equation (SDE) has applications in a wide range of problems: management of resources, queuing systems, epidemic and population processes, pricing of financial derivatives, portfolio allocation... In comparison with classic optimal control, SOC models include a random noise with known probability distribution that affects the evolution or the observation of the system. SOC also aims at managing the risk induced by this noise.

SOC problems are usually solved using specific strategies, such as Forward-Backward SDEs (FBSDEs) [PW99], or by solving Hamilton-Jacobi-Bellman (HJB) optimality conditions [Bel57] through partial differential equations methods using appropriate numerical schemes or by stochastic dynamic programming [KD01]. Such problems can also be solved using Neural Networks calibrated by SGD techniques [GM05, HE16, WLP+19, CL21, BHLP22].

More specifically, in this article we consider the numerical resolution of a SOC problem where the control is parametrized by a neural network calibrated by gradient descent. This method implies to compute the path-wise derivatives along the trajectory of the SDE of the objective function with respect to the parameters of the neural network, as introduced in [GG05, Gil07]. Stochastic gradient descent is a very general approach that can be applied to a wide range of problems and which does not need to be specifically adapted to each problem under consideration. Moreover, SGD scales very efficiently to high-dimensional problems, in contrast with HJB-based methods, and has proved its efficiency on highly non-convex problems [DdVB15].

However, if the neural control is applied at many time steps as it is the case for the Euler-Maruyama scheme where the (discrete) control is taken as an approximation of a continuous control, then the SOC problem reads as the optimization of a very deep neural network, which is roughly as deep as the number of instants at which the control can be applied (see Figures 7.1 and 7.2). Very deep neural networks are known to be considerably more difficult to train [GB10, DPG+14] and may run into vanishing gradient problems [Hoc91, Han18]. Indeed, the deeper the neural network is, the more non-linear it is, thus increasing the number of local traps for the gradient descent such as local (but not global) and saddle points. In image analysis where very deep convolutional neural networks are commonly used, residual [HZRS16] and convolutional dense [HLVDMW17] networks were introduced to deal with this issue. These networks are based on architectures with residual connections to propagate the gradient information through the numerous successive layers.

As it comes to deep SOC, we cannot freely change the structure of the neural network since it is fixed by the equations defining the SOC problem and therefore we cannot directly use residual connections. We can only freely choose the structure of the neural network returning the control, for which a few layers is often enough (see for example [BGTW19, BHLP22]).

A way to improve the learning is to replace SGD algorithms by Stochastic Gradient Langevin Dynamics (SGLD) algorithms. Such optimizers add an exogenous white noise to the gradient descent, providing regularization and allowing to escape from traps. It has indeed been observed that adding noise improves the learning for very deep neural networks [NVL+15, Ani19, GMDB16, SLH+19]. Moreover, Chapter 6 compares side-by-side Langevin with non-Langevin algorithms on networks with increasing depth and shows that for shallow neural networks, Langevin algorithms do nothing else than adding noise to the gradient descent, however the deeper the network is, the greater the gains provided by Langevin algorithms are. We also refer to Chapter 5.

In the present article we study the performances of Langevin optimizers on SOC problems where the number of discretization times where the control is applied is large enough. We use the preconditioned versions of SGD and SGLD [LCCC16] for various choices of preconditioners. We compare side-by-side Langevin and non-Langevin algorithms and we show that Langevin optimizers can significantly improve the training procedure on various problems: fishing quotas [LPP23], deep hedging [BGTW19], oil drilling and resource management [GGKL21]. We mainly consider two different approaches for numerical resolution of SOCs. In the first approach, the control is a single neural network which is applied to every time step and which may depend on the running time t (see Figure 7.1). This approach leads to a model with fewer trainable parameters, which is critical in some data-driven financial applications where the amount of data is limited, and which is more able to capture the specific Markovian features of the problem. In

the second approach, a different neural network is used for each control time (see Figure 7.2). This last setting is also suitable for applying the Layer Langevin algorithm, which is a variant of the Langevin algorithm introduced in Chapter 6 and which proved to be more adapted to the training of very deep neural networks than the Langevin algorithm itself.

We observe that the gains of Langevin algorithms depend on the preconditioner however. While the Adam [KB15] and the Adadelta [Zei12] algorithms can be substantially accelerated by Langevin training, the gains are more limited or sometimes null for RMSprop [TH12].

The code for the numerical experiments is available at

`https://github.com/Bras-P/langevin-for-stochastic-control`.

It includes in particular ready-to-use Langevin optimizers and Layer Langevin optimizers as instances of the TensorFlow `Optimizer` base class, a framework for algorithm comparison in a SOC setting with GPU support and a demonstration notebook.

In the following, on top of the notations defined from page 1, we consider multivariate $(\mathcal{F}_t)$-Brownian motions W and B defined on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}), \mathbb{P})$.

7.2 Stochastic control through gradient descent

7.2.1 Stochastic optimal control

We consider the following SOC problem in continuous time:

$$\min_u J(u) := \mathbb{E}\left[\int_0^T G(X_t)dt + F(X_T)\right], \tag{7.2.1}$$

$$dX_t = b(X_t, u_t)dt + \sigma(X_t, u_t)dW_t,\ t \in [0,T] \tag{7.2.2}$$

where $b : \mathbb{R}^{d_1} \times \mathbb{R}^{d_3} \to \mathbb{R}^{d_1}$, $\sigma : \mathbb{R}^{d_1} \times \mathbb{R}^{d_3} \to \mathcal{M}_{d_1,d_2}(\mathbb{R})$, W is a $\mathbb{R}^{d_2}$-valued Brownian motion and u is a $\mathbb{R}^{d_3}$-valued continuous $(\mathcal{F}_t)$-adapted process, $T > 0$, $G : [0,T]\times\mathbb{R}^{d_1} \to \mathbb{R}$ and $F : \mathbb{R}^{d_1} \to \mathbb{R}$.

We first approximate the continuous SDE $(X_t)_{t\in[0,T]}$ with its Euler-Maruyama scheme and the control u_t with a discrete-time control. For $N \in \mathbb{N}$ being the number of discretization times, we consider the regular subdivision of $[0,T]$:

$$t_k := kT/N,\ k \in \{0,\dots,N\}, \quad h := T/N \tag{7.2.3}$$

and we approximate the control applied at times $t_0,\dots,t_{N-1}$ as the output of a single neural network depending on t, or as the output of N neural networks, one for each discretization instant t_k:

$$u_{t_k} = \bar{u}_\theta(t_k, X_{t_k}, H_{t_k}) \quad \text{or} \quad u_{t_k} = \bar{u}_{\theta^k}(X_{t_k}, H_{t_k}) \tag{7.2.4}$$

where

$$H_t := \int_0^t G(X_s)ds,$$

where $\bar{u}_\theta$ is a neural function with finite-dimensional parameter $\theta \in \mathbb{R}^d$. Indeed, since (7.2.2) defines a Markovian process, we can assume that u_t depends only on t, X_t and H_t instead of t and $(X_s)_{s\in[0,t]}$.

The SOC problem (7.2.1) is numerically approximated by:

$$\min_\theta \bar{J}(\bar{u}_\theta) := \mathbb{E}\Big[\sum_{k=0}^{N-1}(t_{k+1}-t_k)G(\bar{X}^\theta_{t_{k+1}}) + F(\bar{X}^\theta_{t_N})\Big], \tag{7.2.5}$$

$$\bar{X}^\theta_{t_{k+1}} = \bar{X}^\theta_{t_k} + (t_{k+1}-t_k)b(\bar{X}^\theta_{t_k}, \bar{u}_{k,\theta}(\bar{X}^\theta_{t_k}, \bar{H}^\theta_{t_k})) + \sqrt{t_{k+1}-t_k}\sigma(\bar{X}^\theta_{t_k}, \bar{u}_{k,\theta}(\bar{X}^\theta_{t_k}, \bar{H}^\theta_{t_k}))\xi_{k+1}, \tag{7.2.6}$$
$$\bar{H}^\theta_{t_{k+1}} = \bar{H}^\theta_{t_k} + (t_{k+1}-t_k)G(\bar{X}^\theta_{t_{k+1}}),$$
$$\xi_k \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_2})$$

where $\theta \in \mathbb{R}^d$ and $\bar{u}_{k,\theta} = \bar{u}_\theta(t_k, \cdot)$ in the first case of (7.2.4) and $\theta = (\theta^0, \dots, \theta^{N-1}) \in (\mathbb{R}^d)^N$ and $\bar{u}_{k,\theta} = \bar{u}_{\theta^k}$ in the second case of (7.2.4).

For every θ, $\nabla_\theta \bar{J}$ can be computed by automatic differentiation as the gradient w.r.t. to θ is tracked all along the trajectory through the recursive relation (7.2.6) [GG05, Gil07]. Then the SGD algorithm reads

$$\theta_{n+1} = \theta_n - \gamma_{n+1}\frac{1}{n_{\text{batch}}}\sum_{i=1}^{n_{\text{batch}}} \nabla_\theta \bar{J}(\bar{u}_{\theta_n}, (\xi_k^{i,n+1})_{1\le k\le N}) =: \theta_n - \gamma_{n+1}g_{n+1} \tag{7.2.7}$$

where $(\xi_k^{i,n})_{1\le k\le N, 1\le i\le n_{\text{batch}}, n\in\mathbb{N}}$ is an array of i.i.d. random vectors $\mathcal{N}(0, I_{d_2})$-distributed, $(\gamma_n)_{n\in\mathbb{N}}$ is a non-increasing positive step sequence and where the dependence of $\bar{J}$ in $(\xi_k^{i,n})$ is made explicit.

If the number of Euler-Maruyama steps N is large, then the optimization problem in (7.2.5) consists in the training of a very deep neural network that can be difficult to train directly (see the Introduction). Both cases are illustrated in Figures 7.1 and 7.2.

7.2.2 Preconditioned stochastic gradient Langevin dynamics

We consider preconditioned stochastic gradient algorithms i.e. for (P_n) a preconditioner rule the update reads

$$\theta_{n+1} = \theta_n - \gamma_{n+1}P_{n+1}\cdot g_{n+1} \tag{7.2.8}$$

where g_{n+1} is defined in (7.2.7). We use the Adam [KB15], the RMSprop [TH12] and the Adadelta [Zei12] preconditioners, which are detailed in Algorithms 7, 8 and 9 respectively. For some algorithm *name*, the corresponding Langevin algorithm denoted L-*name* reads

$$\theta_{n+1} = \theta_n - \gamma_{n+1}P_{n+1}\cdot g_{n+1} + \sigma_{n+1}\sqrt{\gamma_{n+1}}\mathcal{N}(0, P_{n+1}) \tag{7.2.9}$$

where (σ_n) is a constant or non-decreasing sequence controlling the amount of injected noise.

Algorithm 7 Adam update

Parameters: $\beta_1, \beta_2, \lambda > 0$
$M_{n+1} = \beta_1 M_n + (1-\beta_1)g_{n+1}$
$\mathrm{MS}_{n+1} = \beta_2 \mathrm{MS}_n + (1-\beta_2)g_{n+1}\odot g_{n+1}$
$\widehat{M}_{n+1} = M_{n+1}/(1-\beta_1^{n+1})$
$\widehat{\mathrm{MS}}_{n+1} = \mathrm{MS}_{n+1}/(1-\beta_2^{n+1})$
$P_{n+1} = \mathrm{diag}\left(\mathbb{1} \oslash (\lambda\mathbb{1} + \sqrt{\widehat{\mathrm{MS}}_{n+1}})\right)$
$\theta_{n+1} = \theta_n - \gamma_{n+1}P_{n+1}\cdot \widehat{M}_{n+1}$.

Algorithm 8 RMSprop update

Parameters: $\alpha, \lambda > 0$
$\mathrm{MS}_{n+1} = \alpha\,\mathrm{MS}_n + (1-\alpha)g_{n+1}\odot g_{n+1}$
$P_{n+1} = \mathrm{diag}\left(\mathbb{1} \oslash (\lambda\mathbb{1} + \sqrt{\mathrm{MS}_{n+1}})\right)$
$\theta_{n+1} = \theta_n - \gamma_{n+1}P_{n+1}\cdot g_{n+1}$

The Layer Langevin algorithm, introduced in Chapter 6 Section 6.4 consists in updating with Langevin noise only some layers of the network. It relies on the heuristic that for a deep neural network, the non-linearities of the network are mostly contained in the deepest layers and adds Langevin noise to these layers only.

Choosing a preconditioner rule P called *name*, the Layer Langevin algorithm denoted LL-*name* reads

$$\theta^{(i)}_{n+1} = \theta^{(i)}_n - \gamma_{n+1}[P_{n+1}\cdot g_{n+1}]^{(i)} + \mathbb{1}_{i\in\mathcal{J}}\sigma_{n+1}\sqrt{\gamma_{n+1}}[\mathcal{N}(0, P_{n+1})]^{(i)}, \tag{7.2.10}$$

Algorithm 9 Adadelta update

Parameters: $\beta_1, \beta_2, \lambda > 0$
$\mathrm{MS}_{n+1} = \beta_1 \mathrm{MS}_n + (1-\beta_1) g_{n+1} \odot g_{n+1}$
$P_{n+1} = \mathrm{diag}\left((\lambda \mathbb{1} + \widehat{\mathrm{MS}}_n) \oslash (\lambda \mathbb{1} + \sqrt{\widehat{\mathrm{MS}}_n})\right)$
$\theta_{n+1} = \theta_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1}.$
$\widehat{\mathrm{MS}}_{n+1} = \beta_2 \mathrm{MS}_n + (1-\beta_2)(\theta_{n+1} - \theta_n) \odot (\theta_{n+1} - \theta_n).$

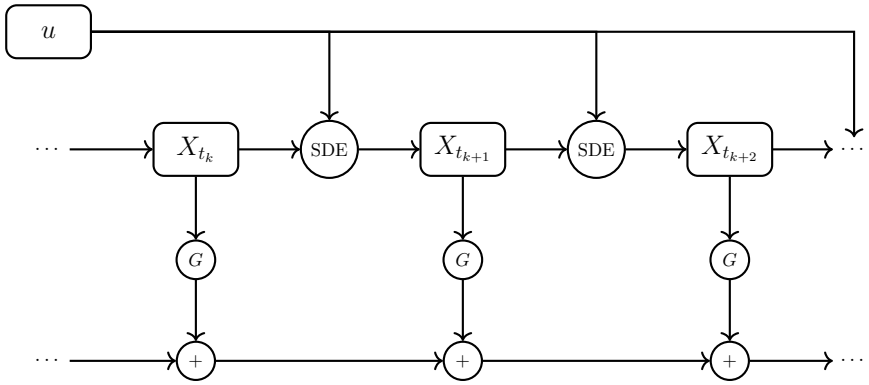


Figure 7.1: Depth of Markovian neural networks. The control u acts on X_{t_k}, which itself acts on $X_{t_{k+1}}$, $X_{t_{k+2}}$, ..., X_{t_N}, hence the depth of the network.

where $\mathcal{J}$ is a subset of weight indices. In particular, we denote LL-*name* p% the Layer Langevin *name* algorithm where the Langevin layers are chosen to be the first p% layers.

7.2.3 Experimental setting

We proceed to side-by-side comparison of Langevin algorithms (7.2.9) with their non-Langevin counterparts (7.2.8) on various SOC problems.

We consider a first case where the control is given by only one neural network depending on t and a second case where a different neural network is used for each control time (7.2.4). In this second case, since we can expect two consecutive control networks to have close parameters, one usual way of performing the training procedure is to first train networks for a small amount of control times, then to perform the whole training through transfer learning. We do not expect

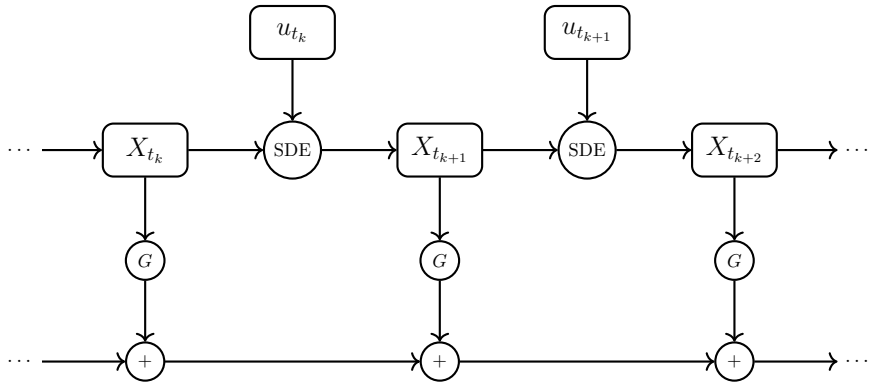


Figure 7.2: Markovian neural network with one control for every time step.

Langevin algorithms to be suitable for the fine tuning, but since the first step of the training still consists in training a deep neural network, we analyse the benefits of Langevin algorithms for this first step.

Unless stated otherwise, the batch size is set to $n_{\text{batch}} = 512$ i.e. stochastic gradient iterations are performed by averaging the gradient over 512 random trajectories. In the plot, each epoch consists in 5 batches i.e. the average loss $J(u_\theta)$ is plotted every 5 iterations of the stochastic gradient. After each epoch, $J(u_\theta)$ is estimated over 25×512 trajectories and 95% confidence intervals are indicated, although for some plots the intervals are too small to be visible. While comparing some algorithm with its Langevin or Layer Langevin counterpart, we ensure that both training procedures start with the same initial weights.

7.3 Fishing quotas

We consider the fishing quota problem introduced in [LPP23]. Let $X_t \in \mathbb{R}^{d_1}$ be the fish biomass for every fish species; we wish to keep it close to an ideal state $\mathcal{X}_t \in \mathbb{R}^{d_1}$. The dynamics of X are given by

$$dX_t = X_t * ((r - u_t - \kappa X_t)dt + \eta dW_t)\,,\ t \in [0,T], \tag{7.3.1}$$

where $r \in \mathbb{R}^{d_1}$ is the growth rate for each species, $u_t \in \mathbb{R}^{d_1}$ is the controlled fishing (with $d_3 = d_1$), $\kappa \in \mathcal{M}_{d_1,d_1}(\mathbb{R})$ is the interaction matrix between the fish species, $\eta \in \mathcal{M}_{d_1,d_2}(\mathbb{R})$, W is a $\mathbb{R}^{d_2}$-valued Brownian motion. The control u is constrained to take its values in $[u_m, u_M]^{d_1}$. The objective is

$$J(u) = \mathbb{E}\left[\int_0^T (|X_t - \mathcal{X}_t|^2 - \langle \alpha, u_t\rangle)dt + \beta[u]^{0,T}\right], \tag{7.3.2}$$

where $\alpha \in \mathbb{R}^{d_1}$, $\beta \in \mathbb{R}^+$, $[u]^{0,T}$ denotes the quadratic variation of u on $[0,T]$. The term $\langle \alpha, u\rangle$ penalizes small fishing quotas while the term $\beta[u]^{0,T}$ penalizes too many daily changes.

In the experiments, following [LPP23] we choose

$$d_1 = d_2 = 5,\ T = 1,\ \mathcal{X} \equiv \mathbf{1},\ r = 2 * \mathbf{1},\ \eta = 0.1 * I_{d_1},\ \alpha = 0.01 * \mathbf{1},\ \beta = 0.1,\ u_m = 0.1,\ u_M = 1 \tag{7.3.3}$$

and

$$\kappa = \begin{pmatrix} 1.2 & -0.1 & 0 & 0 & -0.1 \\ 0.2 & 1.2 & 0 & 0 & -0.1 \\ 0 & 0.2 & 1.2 & -0.1 & 0 \\ 0 & 0 & 0.1 & 1.2 & 0 \\ 0.1 & 0.1 & 0 & 0 & 1.2 \end{pmatrix}. \tag{7.3.4}$$

The initial state X_0 is randomly generated following $\mathcal{N}(\mathbf{1}, (1/2)I_{d_1})$ clipped to $[0.2, 2]^{d_1}$. The quadratic variation $[u]^{0,T}$ is approximated in the discretized setting by

$$[u]^{0,T} \simeq \sum_{k=0}^{N-1} |u_{t_{k+1}} - u_{t_k}|^2. \tag{7.3.5}$$

Each control u_θ is given by a feedforward neural network with two hidden layers with 32 units each and with ReLU activation while the output layer has sigmoid activation in order to fulfil the constraint on u. An example of controlled trajectory is plotted in Figure 7.3.

The results are given in Figure 7.4 for the Adam optimizer with an increasing number of Euler-Maruyama steps and with one single control, in Figure 7.5 for the RMSprop and L-Adadelta optimizers and with one single control and in Figure 7.6 for the training with multiple neural networks.

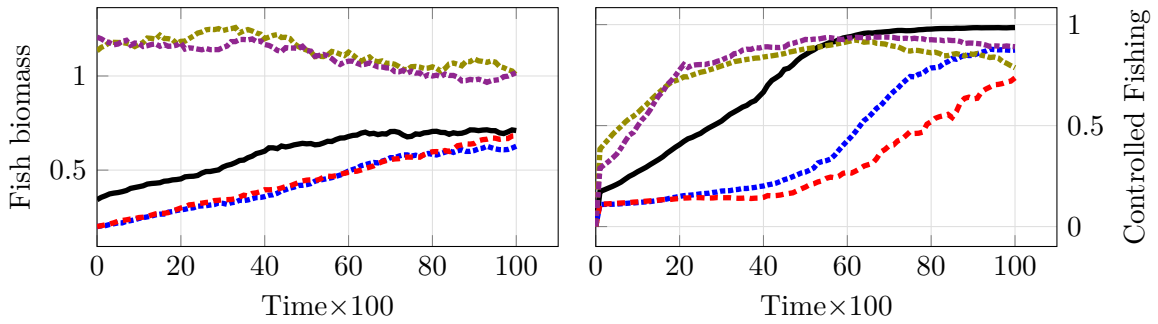


Figure 7.3: Example of a trajectory of $X_t \in \mathbb{R}^5$ along with the controlled fishing u_t with $N = 100$. We recall that the objective biomass is $\mathcal{X}_t \equiv \mathbf{1}$.

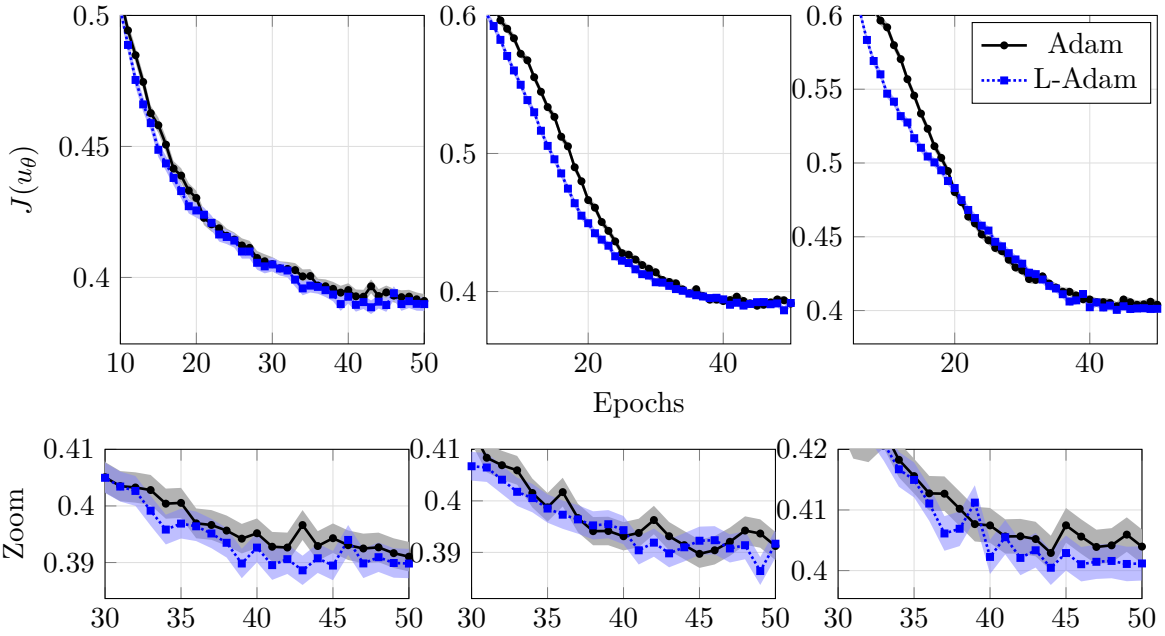


Figure 7.4: Comparison of Adam et L-Adam algorithms during the training for the fishing control problem with $N = 20, 50, 100$ respectively. The schedules are $\gamma_n = 2\,\mathrm{e}{-3}$ and $\sigma_n = 1\,\mathrm{e}{-3}$ ($5\,\mathrm{e}{-3}$ for $N = 100$) for epochs 0 to 40 and $\gamma_n = 2\,\mathrm{e}{-4}$ and $\sigma_n = 0$ beyond. At the end of each epoch, J is estimated over 50×512 trajectories. A zoom on the last epochs is given.

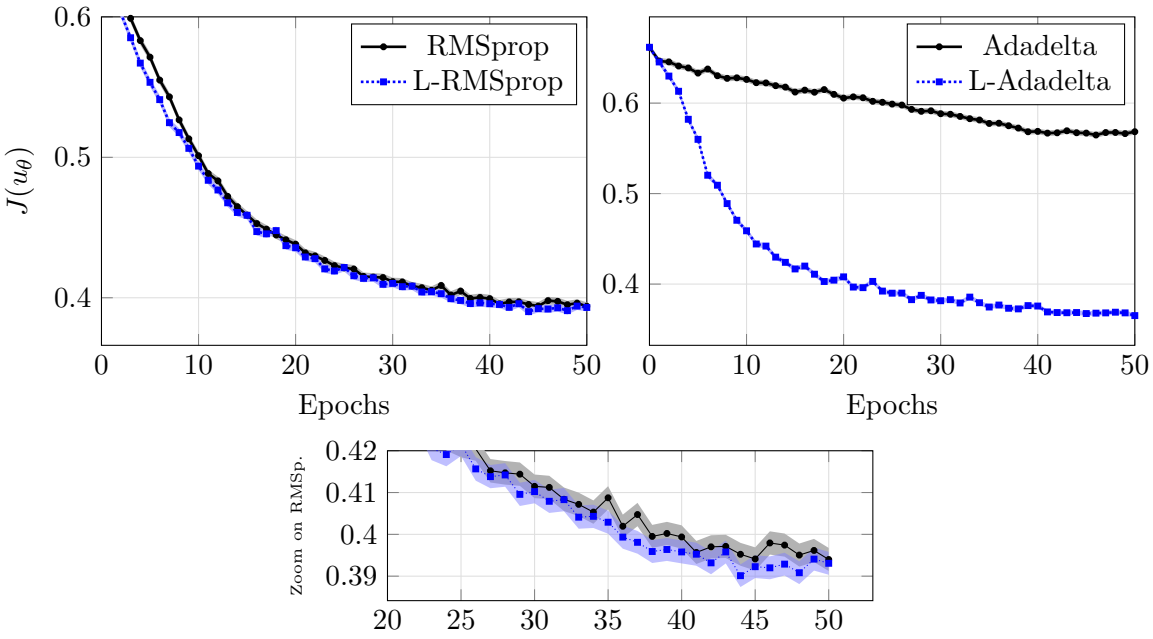


Figure 7.5: Comparison of Langevin algorithms with their non-Langevin counterparts during the training for the fishing control problem with $N = 50$ respectively. The schedules are $\gamma_n = 2\,\mathrm{e}{-3}$ ($5\,\mathrm{e}{-1}$) and $\sigma_n = 5\,\mathrm{e}{-3}$ ($1\,\mathrm{e}{-2}$) for RMSprop (Adadelta resp.) for epochs 0 to 40 and γ_n is divided by 10 and σ_n is set to 0 beyond. At the end of each epoch, J is estimated over 50×512 trajectories. A zoom on the last epochs for RMSprop is given.

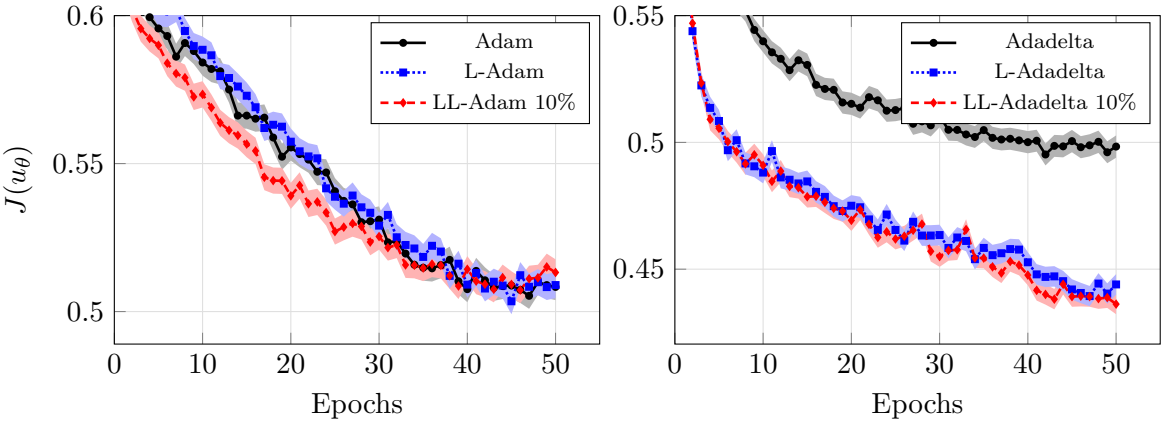


Figure 7.6: Training of the fishing problem with multiple controls with $N = 10$. The schedules are $\gamma_n = 2\,\mathrm{e}{-3}$ and $\sigma_n = 2\,\mathrm{e}{-3}$ for Adam and $\gamma_n = 5\,\mathrm{e}{-1}$ and $\sigma_n = 5\,\mathrm{e}{-3}$ for Adadelta, for epochs 0 to 40 γ_n is divided by 10 and σ_n is set to 0 beyond.

7.4 Deep hedging

We consider the problem of hedging portfolio of derivatives as a SOC problem as in [BGTW19]. We aim to replicate a $\mathcal{F}_T$-measurable payoff Z defined on some portfolio $S_t \in \mathbb{R}^{d_1}$ by trading (at least some of) the assets contained in S_t at times (t_k). The control is given by $u_t \in \mathbb{R}^{d_1}$ representing the amount held for each asset. The objective is

$$J(u) = \nu\left(-Z + \sum_{k=0}^{N-1} \langle u_{t_k}, S_{t_{k+1}} - S_{t_k}\rangle - \sum_{k=0}^{N} \langle c_{tr}, S_{t_k} * |u_{t_k} - u_{t_{k-1}}|\rangle\right) \tag{7.4.1}$$

where $\nu : L^1(\Omega) \to \mathbb{R}$ is a convex risk measure (see [BGTW19, Definition 3.1]), $c_{tr} \in \mathbb{R}^{d_1}$ represents proportional transaction costs and we fix $u_{t_{-1}} = u_{t_N} = 0$, implying full liquidation in T. We furthermore assume that ν can be written as

$$\nu(X) = \inf_{w \in \mathbb{R}} (w + \mathbb{E}[\ell(-X - w)]) \tag{7.4.2}$$

where the loss function $\ell : \mathbb{R} \to \mathbb{R}$ is continuous, non-decreasing and convex. This is the case in particular for the entropic risk measure where $\ell(x) = -\exp(-\lambda x)$ and the conditional value at risk measure where $\ell(x) = (1-\alpha)^{-1}\max(x, 0)$. Then (7.4.1) can be rewritten as

$$\inf_{u,w} J(u, w) := \mathbb{E}\left[w + \ell\left(Z - \sum_{k=0}^{N-1} \langle u_{t_k}, S_{t_{k+1}} - S_{t_k}\rangle + \sum_{k=0}^{N} \langle c_{tr}, S_{t_k} * |u_{t_k} - u_{t_{k-1}}|\rangle - w\right)\right]. \tag{7.4.3}$$

In the numerical experiments, we analyse the problem of hedging in a Heston model as described in [BGTW19, Section 5]. For even d_1, we consider $d_1' := d_1/2$ independent Heston models where the price and volatility processes are described by the following SDEs for $1 \le i \le d_1'$:

$$dS_t^{1,i} = \sqrt{V_t^i} S_t^{1,i} dB_t^i, \quad S_0^{1,i} = s_0^i, \tag{7.4.4}$$

$$dV_t^i = a^i(b^i - V_t^i)dt + \eta^i \sqrt{V_t^i} dW_t^i, \quad V_0^i = v_0^i, \tag{7.4.5}$$

where $a, b, \eta, s_0, v_i \in (\mathbb{R}^+)^{d_1'}$ and for each $1 \le i \le d_1'$, B^i and W^i are standard Brownian motions with correlation $\rho^i \in [-1, 1]$. The volatility V itself is not tradable directly but only through options on variance modelled by the following variance swap:

$$S_t^{2,i} := \mathbb{E}\left[\int_0^T V_s^i ds \Big| \mathcal{F}_t\right] = \int_0^t V_s^i ds + L^i(t, V_t^i), \tag{7.4.6}$$

$$L^i(t, v) := \frac{v - b^i}{a^i}\left(1 - e^{a^i(T-t)}\right) + b^i(T - t). \tag{7.4.7}$$

The payoff is given by

$$Z = \sum_{i=1}^{d_1'} \left(S_T^{1,i} - K^i\right)_+$$

where $K \in (\mathbb{R}^+)^{d_1'}$. We consider the convex risk measure associated to the value-at-risk i.e. associated to the loss function

$$\ell(x) = \frac{1}{1-\alpha}\max(x, 0).$$

In the experiments we choose

$$d_1' = 5,\ T = 1,\ a = \mathbf{1},\ b = 0.04 * \mathbf{1},\ \eta = 2 * \mathbf{1},\ \rho = -0.7 * \mathbf{1},\ \alpha = 0.9, \tag{7.4.8}$$

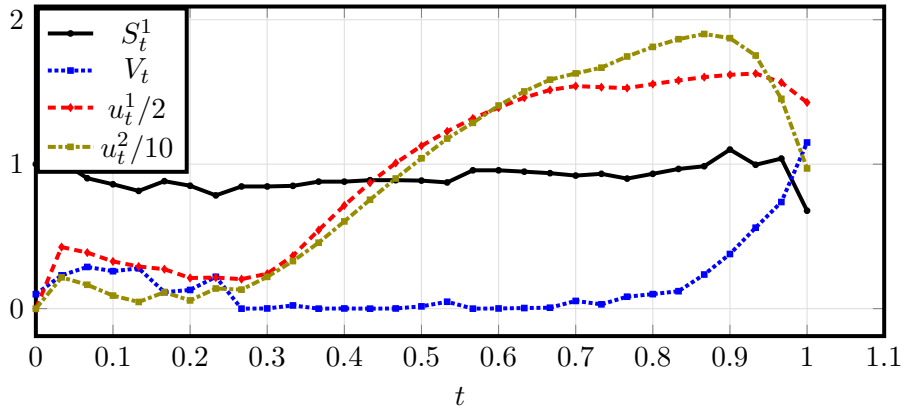


Figure 7.7: Example of trajectory for the deep hedging problem with $N = 30$.

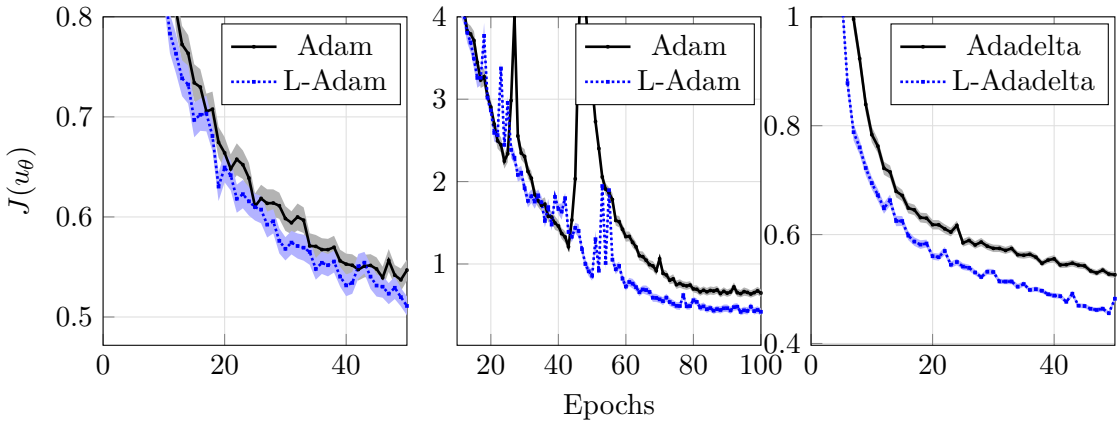


Figure 7.8: Comparison of algorithms during the training for the deep hedging control problem with $N = 30, 50, 50$ respectively. The schedules are $\gamma_n = 2\,\mathrm{e}{-3}$ $(5\,\mathrm{e}{-1})$ and $\sigma_n = 2\,\mathrm{e}{-3}$ $(5\,\mathrm{e}{-3})$ for Adam (resp. Adadelta) for epochs 0 to 80 and γ_n is divided by 10 and σ_n is set to 0 beyond.

$$s_0 = K = \mathbf{1},\ v_0 = 0.1 * \mathbf{1},\ c_{\mathrm{tr}} = 5\,\mathrm{e}{-4} * \mathbf{1}. \tag{7.4.9}$$

Each control u_θ is given by a feedforward neural network with two hidden layers with 32 units each and with ReLU activation while the output layer has ReLU activation too in order to forbid short-selling. As recommended in [BGTW19], since transaction costs are involved the control u_θ at time t_k is a function of $\log(S^1_{t_k})$, V_{t_k} and $u_{t_{k-1}}$. An example of controlled trajectory showing only one of the five Heston models is plotted in Figure 7.7.

The results are given in Figure 7.8 for the comparison of Langevin and non-Langevin algorithms with a single control and in Figure 7.9 for the training with multiple controls.

7.5 Resource management

We consider the control problem in the management of natural resources applied to oil drilling introduced in [GKL18] and extended in [GGKL21]. The objective is for an oil driller, to balance the costs of extraction, storage in a volatile energy market. The oil price $P_t \in \mathbb{R}$ is assumed to be a Black-Scholes process:

$$dP_t = \mu P_t dt + \eta P_t dW_t. \tag{7.5.1}$$

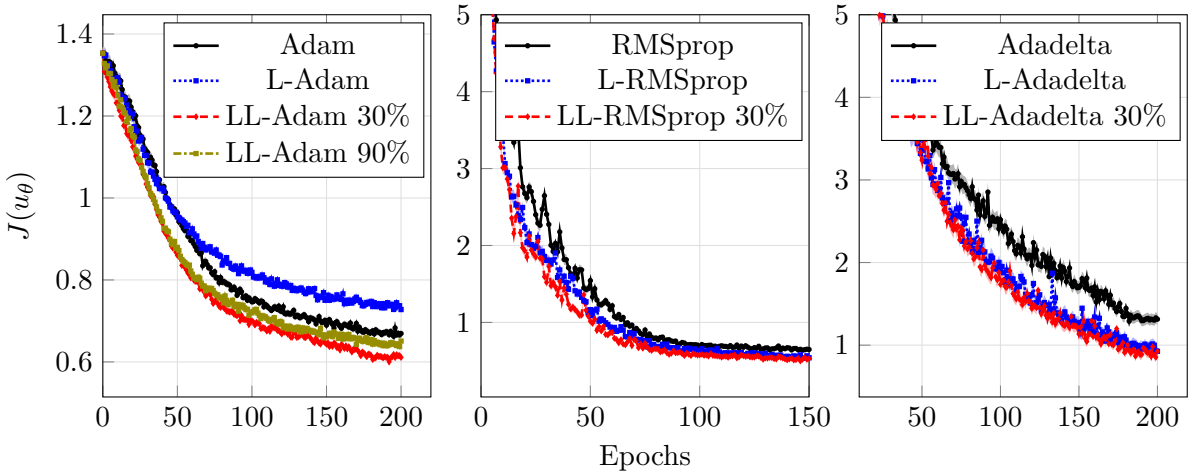


Figure 7.9: Training of the deep hedging problem with multiple controls with $N = 10$. The schedules are $\gamma_n = 2\,\mathrm{e}{-3}$ $(5\,\mathrm{e}{-1})$ and $\sigma_n = 2\,\mathrm{e}{-3}$ $(5\,\mathrm{e}{-3})$ for Adam and RMSprop (resp. Adadelta) for epochs 0 to 180 and γ_n is divided by 10 and σ_n is set to 0 beyond.

The control is given by $q_t = (q_t^v, q_t^s, q_t^{v,s}) \in \mathbb{R}^3$ where q_t^v is the quantity of extracted oil immediately sold on the market per time unit, q_t^s is the quantity of extracted oil that is stored per time unit, $q_t^{v,s}$ is the quantity of stored oil that is sold per time unit. The cumulated quantities of extracted and stored oil at time t are respectively given by

$$E_t = \int_0^t (q_r^v + q_r^s)dr, \quad S_t = \int_0^t (q_r^s - q_r^{v,s})dr. \tag{7.5.2}$$

The extraction and storage prices are respectively given by

$$c_e(E_t) = \exp\left(\xi_e E_t\right), \quad c_s(S_t) = \exp\left(\xi_s S_t\right) - 1. \tag{7.5.3}$$

The constraints on the control are the following:

$$q_t^v,\ q_t^s,\ q_t^{v,s} \geq 0, \quad q_t^{v,s} \leq q^S, \quad q_t^v + q_t^s \leq K_0, \quad 0 \leq S_t \leq Q^S, \tag{7.5.4}$$

where q^S, K_0 and Q^S are operational bounds. The objective is

$$J(q) = -\mathbb{E}\left[\int_0^T e^{-\rho r} U\Big(q_r^v P_r + q_r^{v,s}(1-\varepsilon)P_r - (q_r^v + q_r^s)c_e(E_r) - c_s(S_r)\Big)dr\right], \tag{7.5.5}$$

where $U : \mathbb{R} \to \mathbb{R}$ is the utility function.

In the experiments we take

$$\begin{aligned} &T = 1,\ \mu = 0.01,\ \eta = 0.2,\ \rho = 0.01,\ \varepsilon = 0,\ K_0 = 5, \\ &\xi_e = 1\,\mathrm{e}{-2},\ \xi_s = 5\,\mathrm{e}{-3},\ q^S = 10,\ P_0 = 1,\ U(x) = x. \end{aligned} \tag{7.5.6}$$

The control q_t is given by a feedforward neural network with two hidden layers with 32 units and with ReLU activation while the output layer has several ReLU activations such that the constraints on q (7.5.4) are fulfilled[1]. An example of controlled trajectory is given in Figure 7.10.

The results are given in Figure 7.11 for the comparison of Langevin and non-Langevin algorithms with a single control. We do not display the results for the training with multiple controls however as we could not obtain satisfying results neither with Langevin nor-with non-Langevin methods.

[1]We remark that $\max(q, K) = K - \mathrm{ReLU}(-q + K)$.

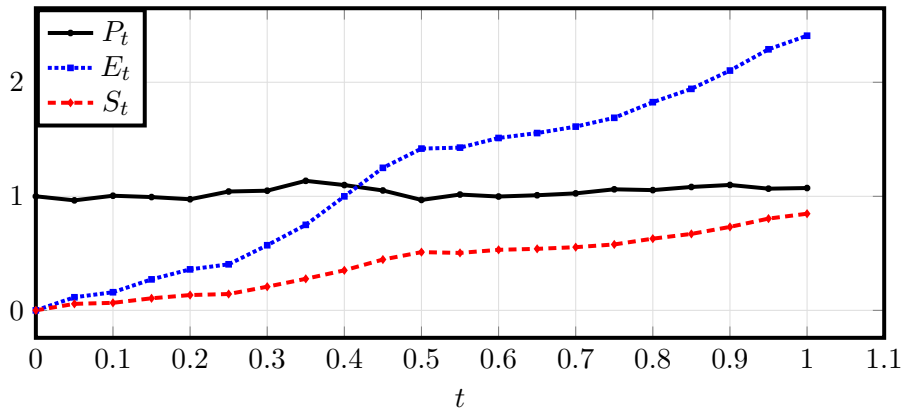


Figure 7.10: Example of trajectory for the oil drilling problem with $N = 20$.

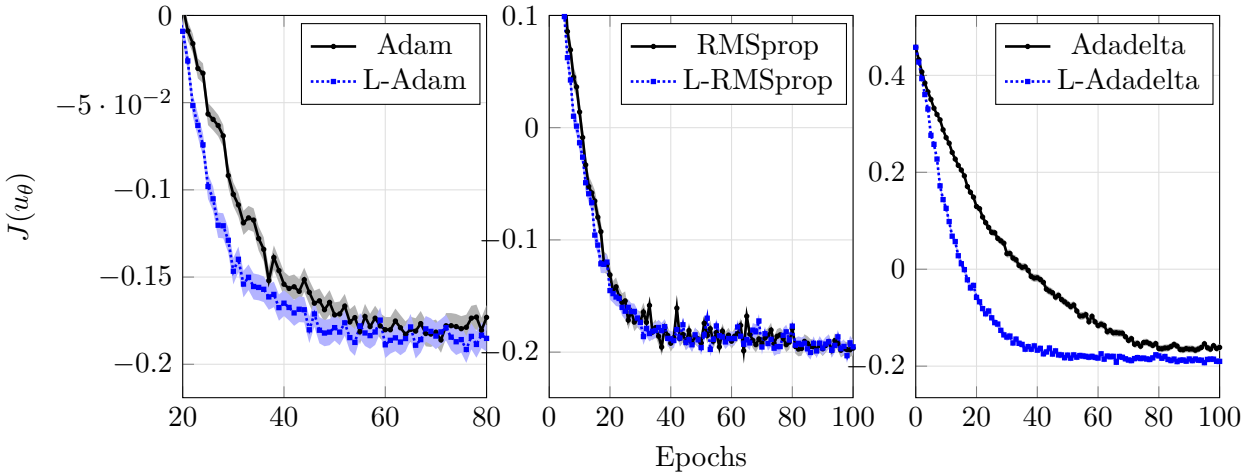


Figure 7.11: Comparison of algorithms during the training for the deep hedging control problem with $N = 50$. The schedules are $\gamma = 2\,\mathrm{e}{-3}$ ($2\,\mathrm{e}{-3}$, $5\,\mathrm{e}{-1}$) and $\sigma = 1\,\mathrm{e}{-3}$ ($2\,\mathrm{e}{-3}$, $5\,\mathrm{e}{-3}$) for Adam (resp. RMSprop, Adadelta) for epochs 0 to 60 (resp. 80, 80) and γ is divided by 10 and σ is set to 0 beyond.

7.6 Comments on the numerical experiments

We observe that in many cases and in various SOC problems, Langevin and Layer Langevin algorithms show improvement when compared with their respective non-Langevin counterparts, provided that N is large enough, which is remarkable for randomized algorithms. Langevin algorithms converge faster and/or toward a lower loss value. This is particularly visible for the Adadelta method. The gains are limited in some cases (see Figures 7.5 and 7.11 for RMSprop) but still the optimization procedure is improved.

The gains for the L-RMSprop algorithm remain limited however. In particular, we did not observe any significant improvement for fishing SOC with multiple controls, for deep hedging SOC with a single control and for oil drilling SOC. We do not have explanation for this fact.

The gains brought by Langevin algorithm increase with the depth of the network as shown in Figures 7.4 and 7.8. However, contrary to Chapter 6, we did not observe overwhelming gains as N becomes very large. We believe that this is due (in part) to the particular structure of the deep SOC problem where the same control is repeated all along the trajectory.

As for SOC with multiple controls, we observe that Layer Langevin algorithms with a small number of Langevin layers (10%-30%) generally outperforms Vanilla Langevin methods while Vanilla Langevin may bring limited gains or be less efficient than the standard non-Langevin methods, see Figures 7.6 and 7.9 for Adam.

Acknowledgements

The authors thank Idris Kharroubi for helpful discussions.

Part III

Numerical simulation of stochastic processes

Chapter 8

Total variation distance between two diffusions in small time with unbounded drift: application to the Euler-Maruyama scheme

This chapter corresponds to the article [BPP22] published in *Electronic Journal of Probability* as a joint work with Gilles Pagès and Fabien Panloup.

Abstract

We give bounds for the total variation distance between the solutions to two stochastic differential equations starting at the same point and with close coefficients, which applies in particular to the distance between an exact solution and its Euler-Maruyama scheme in small time. We show that for small t, the total variation distance is of order $t^{r/(2r+1)}$ if the noise coefficient σ of the SDE is elliptic and $\mathcal{C}_b^{2r}$, $r \in \mathbb{N}$ and if the drift is C^1 with bounded derivatives, using multi-step Richardson-Romberg extrapolation. We do not require the drift to be bounded. Then we prove with a counterexample that we cannot achieve a bound better than $t^{1/2}$ in general.

Keywords– Stochastic Differential Equation, Euler scheme, Total Variation, Richardson – Romberg extrapolation, Aronson's bounds.

8.1 Introduction

The convergence properties of Euler-Maruyama schemes to approximate the solution of a Stochastic Differential Equation (SDE) have been extensively studied, in particular for L^p distances. However, the literature seems to lack some results about the convergence in total variation in small time. More specifically, in this paper we consider the two following SDEs in $\mathbb{R}^d$ starting at the same point:

$$X_0^x = x \in \mathbb{R}^d, \qquad dX_t^x = b_1(t, X_t^x)dt + \sigma_1(t, X_t^x)dW_t,$$

$$Y_0^x = x, \qquad dY_t^x = b_2(t, Y_t^x)dt + \sigma_2(t, Y_t^x)dW_t,$$

where W is a Brownian motion. We generally assume that for $i = 1, 2$, b_i is Lipschitz continuous and that σ_i is elliptic, bounded and Lipschitz continuous, but we do not assume that b_i is bounded. Our objective is to give bounds of the total variation distance between the law of X_t^x and the law of Y_t^x, denoted $\mathrm{d_{TV}}(X_t^x, Y_t^x)$, as $t \to 0$. In particular, we apply our results to the case where $Y^x = \bar{X}^x$ is the one-step Euler-Maruyama scheme associated to the SDE X, given by

$$dY_t^x = b_1(0, x)dt + \sigma_1(0, x)dW_t.$$

Such bounds are well known for L^p distances and their associated Wasserstein distances and are known to be of order t as $t \to 0$. Yet the literature seems to lack results as it comes to $\mathrm{d_{TV}}$. If $\sigma_1 = \sigma_2$ is constant, then it is classical background that $\mathrm{d_{TV}}(X_t^x, Y_t^x)$ is of order t, using a Girsanov change of measure (see for example [PP23, Proposition 4.1]) but this strategy cannot be applied to non-constant σ. The difficulty of the total variation distance in small time is the following: considering its representation formula and comparing it with the L^1-Wasserstein distance, if x and $y \in \mathbb{R}^d$ are close to each other and if $f : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous, then we can bound $|f(x) - f(y)|$ by $[f]_{\mathrm{Lip}}|x - y|$; whereas if f is simply measurable and bounded, then we cannot directly bound $|f(x) - f(y)|$ in terms of $|x - y|$. Moreover the regularizing properties of the semi-group cannot be used in small time for the total variation distance.

Results in the literature focus on the Euler-Maruyama scheme. In [BT96] is proved the convergence for a fixed time horizon $T > 0$ and as $N \to \infty$, where N is the number of steps in the Euler-Maruyama scheme on a finite horizon. More precisely, if σ_1 is elliptic and if b_1 and σ_1 are $\mathcal{C}^\infty$ with bounded derivatives (but b_1 and σ_1 are not supposed bounded themselves), then ([BT96, Theorem 3.1])

$$\forall x \in \mathbb{R}^d,\ \mathrm{d_{TV}}(X_T^x, \bar{X}_T^{x,N}) \leq \frac{K(T)(1 + |x|^Q)}{NT^q},$$

where $\bar{X}_T^{x,N}$ stands for the Euler scheme with N steps, where Q and q are positive exponents and where K is a non-decreasing function depending on b_1 and σ_1. The common strategy of proof for such bounds is to use Malliavin calculus in order to perform an integration by parts and to use bounds on the derivatives of the density. However, we cannot infer a bound as $T \to 0$ since we do not know whether $K(T)/T^q \to 0$ as $T \to 0$ in general. In [GL08] are given bounds in small time and as $N \to \infty$. Assuming that σ_1 is uniformly elliptic and that b_1 and σ_1 are bounded with bounded derivatives up to order 3, then ([GL08, Theorem 2.3])

$$\forall t \in (0, T],\ \forall x, y \in \mathbb{R}^d,\ |p(t, x, y) - \bar{p}^N(t, x, y)| \leq \frac{K(T)T}{Nt^{(d+1)/2}} e^{-C|x-y|^2/t},$$

where p and $\bar{p}^N$ denote the transition densities of X^x and $\bar{X}^{x,N}$ respectively and where C is a positive constant depending on d and on the bounds on b_1 and σ_1 and on their derivatives. However, we cannot directly use this result for the total variation distance: taking $N = 1$ yields

$$\mathrm{d_{TV}}(X_t^x, \bar{X}_t^x) = \int_{\mathbb{R}^d} |p(t, x, y) - \bar{p}^N(t, x, y)|dy \leq K(T)Tt^{-1/2} \int_{\mathbb{R}^d} \frac{1}{t^{d/2}} e^{-C|x-y|^2/t},$$

giving a bound in $t^{-1/2}$ which does not converge to 0 as $t \to 0$. Moreover, [GL08] assumes that b_1 is bounded. [BJ22] focuses on the case where b_1 is bounded and measurable but not necessarily regular and where σ_1 is constant; it proves that the convergence in total variation of the Euler scheme on a finite horizon which is regularized with respect to the irregular drift b_1 and with step h, is of order $\sqrt{h}$.

In the present paper, we first prove a convergence rate of order $t^{1/3}$ for $\mathrm{d_{TV}}(X_t^x, Y_t^x)$, provided that for $i = 1, 2$, σ_i is elliptic, σ_i and b_i are Lipschitz-continuous with respect to their time variable and that σ_i is $\mathcal{C}_b^2$ and b_i is $\mathcal{C}^1$ and Lipschitz-continuous with respect to their space variable. More generally, if we furthermore assume that σ is $\mathcal{C}_b^{2r}$, then we obtain a convergence rate of order $t^{r/(2r+1)}$. Letting $r \to \infty$, we also prove that if $\sigma \in \mathcal{C}_b^\infty$ with some technical condition on the derivatives of the densities of the random variables X_t^x and Y_t^x, then the convergence rate is of order $t^{1/2}\exp(C\sqrt{\log(1/t)})$ which is "almost" $t^{1/2}$. Moreover, we provide an example using the geometric Brownian motion where the convergence rate is exactly $t^{1/2}$, thus showing that we cannot achieve better bounds in general. To prove the bound in $t^{r/(2r+1)}$, we use a multi-step Richardson-Romberg extrapolation [RG11, Gil08, LP17], which is a method imported from numerical analysis that we use in our case for theoretical purposes. It relies on a Taylor expansion with null coefficients up to some high order. Such method can be used in more general settings with regularization arguments in order to improve the convergence rates (in our case, we improve $t^{1/3}$ into $t^{r/(2r+1)}$).

Interestingly, the difference between the drift coefficients $b_1 - b_2$ does not need to be small for our bounds to be valid. This is because the dominant term in $\mathrm{d_{TV}}(X_t^x, Y_t^x)$ comes from the the diffusion part.

Recent results (see [Cle21]) establish a convergence in small time at rate $t^{1/2}$ for the Euler scheme of certain classes of diffusions driven by stable Lévy processes, not directly including the Brownian case. This approach relies on Malliavin calculus techniques. In this work the "standard" drift is replaced in the Euler scheme by the flow of the associated (noiseless) ODE. This seems to be specific to Lévy driven SDEs. Adapting this approach to our general continuous framework is not as straightforward as could be expected and would deserve further investigations for future work.

The total variation distance is closely related to the estimation of the density of the solution to an SDE and this density satisfies a Fokker-Planck Partial Differential Equation PDE (8.3.2). If the drift is bounded, then the density and its partial derivatives admit sub-gaussian Aronson's bounds (see [Fri64] and Section 8.3.1). However, giving estimates and bounds for the solution of the PDE in the case of unbounded drift appears to be more difficult, see [Lun97, Cer00, BL05]. Recent improvements have been made in [MPZ21] using the parametrix method. Studying this case is useful to study the convergence in total variation of SDE's with unbounded drift, in particular for the Langevin equation, very popular in stochastic optimization, and which reads

$$dX_t = -\nabla V(X_t)dt + \sigma(X_t)dW_t,$$

where in many cases, $V : \mathbb{R}^d \to \mathbb{R}$ has quadratic growth and ∇V has linear growth (see for example Chapter 2).

In order to deal with unbounded b_i, we propose two different methods. First, we use a localization argument and "cut" the drift b_i into $\tilde{b}_i$ outside a compact set, so that we can use bounds from [Fri64] for the bounded drift case. We use the Girsanov formula to explicit the dependence of these bounds in $\|\tilde{b}_i\|_\infty$. A second method consists in using the density estimates from [MPZ21, Section 4] to improve the dependency with respect to the bounds in x. However this second approach relies on advanced parametrix methods which require further regularity assumptions on the coefficients of the SDE and which are not fully detailed for higher order derivatives. Our first method is clearly much more elementary, starting from a quite general bound established for any pair of integrable random vectors (see Theorem 8.2.7) and calling upon a standard regularization strategy which combined with a multistep procedure, seems to be at least quasi-optimal in a very general framework.

Notations

On top of the notations defined from page 1, we also use the following notations.

For $k \in \mathbb{N}$ and if $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^k$, we denote by $\nabla^k f : \mathbb{R}^d \to (\mathbb{R}^d)^{\otimes k}$ its differential of order k. If f is Lipschitz continuous, we denote by $[f]_{\mathrm{Lip}}$ its Lipschitz constant. If $f : (t,x) \in \mathbb{R} \times \mathbb{R}^d \mapsto f(t,x)$ is $\mathcal{C}^k$ with respect to x, we still denote by $\nabla^k f$ its differential with respect to x.

If Z is a Markov process with values in $\mathbb{R}^d$, we denote, when it exists, its transition probability from x to $y \in \mathbb{R}^d$ between times $s < t$, $p_Z(s,t,x,y)$.

8.2 Main results

We consider the two following SDEs in $\mathbb{R}^d$:

$$X_0^x = x \in \mathbb{R}^d, \qquad dX_t^x = b_1(t, X_t^x)dt + \sigma_1(t, X_t^x)dW_t, \ t \in [0,T], \tag{8.2.1}$$

$$Y_0^x = x, \qquad dY_t^x = b_2(t, Y_t^x)dt + \sigma_2(t, Y_t^x)dW_t, \ t \in [0,T], \tag{8.2.2}$$

where T is a finite time horizon, $b_i : \mathbb{R}^d \to \mathbb{R}^d$, $\sigma_i : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$, $i = 1,2$, are Borel functions and W is a standard $\mathbb{R}^d$-valued Brownian motion defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The one-step Euler-Maruyama scheme of X, denoted $\bar{X}$, is defined by $\bar{X}^x = Y^x$ when Y^x reads

$$dY_t^x = b_1(0,x)dt + \sigma_1(0,x)dW_t, \ t \in [0,T]. \tag{8.2.3}$$

To allievate notations, we also define

$$\Delta b(x) := |b_1(0,x) - b_2(0,x)|, \quad \Delta\sigma(x) := |\sigma_1(0,x) - \sigma_2(0,x)|. \tag{8.2.4}$$

Let us remark that if $Y = \bar{X}$, then $\Delta b(x) = 0$ and $\Delta\sigma(x) = 0$. For $g : (t,x) \in [0,T] \times \mathbb{R}^d \mapsto g(t,x) \in \mathbb{R}^q$ and $r \in \mathbb{N}$, let us define the following assumptions:

- $\mathrm{Lip}_t(g)$: g is Lipschitz continuous with respect to t, uniformly in x.
- $g \in \mathcal{C}^r$: g is differentiable with respect to x with continuous partial derivatives up to the order r.
- $g \in \mathcal{C}_b^r$: $g \in \mathcal{C}^r$ and is bounded with bounded partial derivatives up to the order r.
- $g \in \widetilde{\mathcal{C}}_b^r$: $g \in \mathcal{C}^r$ and has partial bounded derivatives up to the order r, but we do not assume that g is bounded itself.
- For $\sigma : [0,T] \times \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$, we say that σ is (uniformly) elliptic if

$$\exists \underline{\sigma}_0 > 0, \ \forall x \in \mathbb{R}^d, \ \forall t \in [0,T], \ \sigma(t,x)\sigma(t,x)^\top \geq \underline{\sigma}_0^2 I_d. \tag{8.2.5}$$

Theorem 8.2.1. *Let X and Y be the solutions of the SDEs* (8.2.1) *and* (8.2.2)*. For $i = 1,2$, assume* $\mathrm{Lip}_t(b_i)$, $\mathrm{Lip}_t(\sigma_i)$, $\sigma_i \in \mathcal{C}_b^2$, $b_i \in \widetilde{\mathcal{C}}_b^1$ *and σ_i is elliptic. Then*

$$\forall t \in [0,T], \ \forall x \in \mathbb{R}^d, \ \mathrm{d_{TV}}(X_t^x, Y_t^x) \leq C(t^{1/2} + \Delta\sigma(x))^{2/3} + Ce^{c|x|^2}t^{1/2}, \tag{8.2.6}$$

where the positive constants C and c only depend on d, T, $\underline{\sigma}_0$, $\|\sigma_i\|_\infty$, $[b_i]_{\mathrm{Lip}}$, $[\sigma_i]_{\mathrm{Lip}}$ and $\|\nabla^2\sigma_i\|_\infty$. In particular, if $Y = \bar{X}$ we have

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \leq Ct^{1/3} + Ce^{c|x|^2}t^{1/2}.$$

Theorem 8.2.2. *Let X and Y be the solutions of the SDEs* (8.2.1) *and* (8.2.2)*. For $i = 1,2$, assume* $\mathrm{Lip}_t(b_i)$, $\mathrm{Lip}_t(\sigma_i)$, *that $\sigma_i \in \mathcal{C}_b^{2r}$, $b_i \in \widetilde{\mathcal{C}}_b^1$ and that σ_i is elliptic. Then*

$$\forall t \in [0,T], \ \forall x \in \mathbb{R}^d, \ \mathrm{d_{TV}}(X_t^x, Y_t^x) \leq C(t^{1/2} + \Delta\sigma(x))^{2r/(2r+1)} + Ce^{c|x|^2}t^{1/2}, \tag{8.2.7}$$

where the positive constants C and c only depend on d, T, $\underline{\sigma}_0$, on $\|\sigma_i\|_\infty$, on the bounds on the derivatives of b_i and σ_i and on their Lipschitz constants. In particular, if $Y = \bar{X}$ we have

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \leq Ct^{r/(2r+1)} + Ce^{c|x|^2}t^{1/2}.$$

Remark 8.2.3. In Theorems 8.2.1 and 8.2.2, we can actually improve the dependency of the constants in x in small time since we have more precisely:

$$\forall t \in [0, Ce^{-2(2r+1)c|x|^2}], \quad \mathrm{d_{TV}}(X_t^x, Y_t^x) \leq C(t^{1/2} + \Delta\sigma(x))^{2r/(2r+1)}.$$

Remark 8.2.4. We can adapt the framework of Theorem 8.2.2 to the study of SDEs with homogeneous vanishing noise, i.e. where for $a > 0$,

$$dX_t^x = b_1(t, X_t^x)dt + a\sigma_1(X_t^x), \quad dY_t^x = b_2(t, Y_t^x) + a\sigma_2(Y_t^x)$$

and we identify the dependency in a as $a \to 0$. Namely, with the same assumptions, for $a > 0$ small enough we have

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \leq Ce^{ca^{-1}|x|^2}t^{1/2} + Ca^{-(d+r)}(t^{1/2} + \Delta\sigma(x))^{2r/(2r+1)}, \tag{8.2.8}$$

where the constant C does not depend on a. This bound is obtained adapting the proof of Theorem 8.2.2 in Section 8.3.4 and using that if Z^x is the martingale $dZ_t^x = a\sigma(Z_t^x)dW_t$ then $(Z_t^x) \sim (\tilde{Z}_{a^2t}^x)$ where $\tilde{Z}_t^x = \sigma(\tilde{Z}_t^x)dW_t$ which does not depend on a.

We can also improve the dependency in the initial condition $x \in \mathbb{R}^d$ using [MPZ21], however at the expense of further regularity assumptions on b_i and σ_i, $i = 1, 2$.

Theorem 8.2.5. *Let X and Y be the solutions of the SDEs* (8.2.1) *and* (8.2.2). *For $i = 1, 2$, assume* $\mathrm{Lip}_t(b_i)$, $\mathrm{Lip}_t(\sigma_i)$, $\sigma_i \in \mathcal{C}_b^{2r+1}$, $b_i \in \widetilde{\mathcal{C}}_b^{2r}$ *and σ_i is elliptic. Then*

$$\forall t \in [0, T],\ \forall x \in \mathbb{R}^d,\ \mathrm{d_{TV}}(X_t^x, Y_t^x) \leq C\left(t^{1/2}(1 + \Delta b(x)) + \Delta\sigma(x) + t(|b_1| + |b_2|)(0, x)\right)^{2r/(2r+1)}, \tag{8.2.9}$$

where the positive constants C and c only depend on d, T, $\underline{\sigma}_0$, on $\|\sigma_i\|_\infty$, on the bounds on the derivatives of b_i and σ_i and on their Lipschitz constants. In particular, if $Y = \bar{X}$, we have

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \leq Ct^{r/(2r+1)}\left(1 + t^{1/2}(|b_1| + |b_2|)(0, x)|\right)^{2r/(2r+1)}.$$

Remark 8.2.6. Choosing Y not to be the Euler-Maruyama scheme of X but a general SDE and expressing the bounds in the Theorems in terms of $\Delta b(x)$ and $\Delta\sigma(x)$ allows to extend our results to more general couples of diffusions with "close" coefficients, for example to SDE solvers other than the genuine Euler-Maruyama scheme. Also, it is helpful to study perturbed SDEs, for example if we consider

$$dX_t^x = b_1(t, X_t^x)dt + a_1(t)\sigma(X_t^x)dW_t, \quad dY_t^x = b_2(t, Y_t^x)dt + a_2(t)\sigma(Y_t^x)dW_t \tag{8.2.10}$$

where $|a_1(t) - a_2(t)| \to 0$ as $t \to \infty$. Then we have

$$\mathrm{d_{TV}}(X_{t+s}^x, Y_{t+s}^x) \leq Ce^{c|x|^2}s^{1/2} + C(s^{1/2} + |a_1(t) - a_2(t)|)^{2r/(2r+1)}$$

and we obtain different convergence rates as $t \to \infty$ and $s \to 0$, depending on (t, s). A noticeable example of this is the Langevin-simulated annealing SDE, see Chapter 2.

Furthermore we remark that in Theorems 8.2.1 and 8.2.2, the bounds do not depend on Δb, enhancing that the dominant term in the total variation comes from the diffusion part.

To improve the rate of convergence from $t^{1/3}$ in Theorem 8.2.1 to $t^{r/(2r+1)}$ in Theorem 8.2.2, we rely on a Richardson-Romberg extrapolation [RG11, Gil08, LP17]; this argument can also be applied in a more general framework. The following proposition gives bounds on the total variation between two random vectors, knowing bounds on the L^1-Wasserstein distance and bounds on the partial derivatives of the densities up to some order.

Theorem 8.2.7. *Let Z_1 and Z_2 be two random vectors in $L^1(\mathbb{R}^d)$ and admitting densities p_1 and p_2 respectively with respect to the Lebesgue measure. Assume furthermore that p_1 and p_2 are $\mathcal{C}^{2r}$ with $r \in \mathbb{N}$ and that $\nabla^k p_i \in L^1(\mathbb{R}^d)$ for $i = 1, 2$ and $k = 1, \dots, 2r$. Then we have*

$$\mathrm{d_{TV}}(Z_1, Z_2) \leq C_{d,r} \mathcal{W}_1(Z_1, Z_2)^{2r/(2r+1)} \left(\int_{\mathbb{R}^d} \left(\|\nabla^{2r} p_1(\xi)\| + \|\nabla^{2r} p_2(\xi)\| \right) d\xi \right)^{1/(2r+1)} \tag{8.2.11}$$

where the constant $C_{d,r}$ depends only on d and on r.

If $\sigma \in \mathcal{C}_b^\infty$, then we also prove that we can "almost" get a convergence rate of order $t^{1/2}$.

Theorem 8.2.8. *Let X and Y be the solutions of the SDEs (8.2.1) and (8.2.2). For $i = 1, 2$, assume $\mathrm{Lip}_t(b_i)$, $\mathrm{Lip}_t(\sigma_i)$, that $\sigma_i \in \mathcal{C}_b^{2r}$ for every $r \in \mathbb{N}$, $b_i \in \widetilde{\mathcal{C}}_b^1$, that σ_i is elliptic and that $\Delta\sigma(x) = 0$. Assume furthermore that if Z and V are the martingales $dZ_t = \sigma_1(t, Z_t)dW_t$ and $dV_t = \sigma_2(t, Z_t)dW_t$, then*

$$\forall r \in \mathbb{N},\ \forall t \in (0, T], \forall x, y \in \mathbb{R}^d,\ \|\nabla_y^{2r} p_{_Z}(0, t, x, y)\| + \|\nabla_y^{2r} p_{_V}(0, t, x, y)\| \leq \frac{C_{2r}}{t^{(d+2r)/2}} e^{-c_{2r}|y-x|^2/t}$$

$$\text{with } \limsup_{r \to \infty} \left(C_{2r} c_{2r}^{-d/2} \right)^{1/(2r)} < \infty. \tag{8.2.12}$$

(see Theorem 8.3.1). Then

$$\forall t \in (0, T],\ \forall x \in \mathbb{R}^d,\ \mathrm{d_{TV}}(X_t^x, Y_t^x) \leq C e^{c|x|^2} t^{1/2} + C t^{1/2} e^{c\sqrt{\log(1/t)}}, \tag{8.2.13}$$

where the positive constants C and c only depend on d, T, $\underline{\sigma}_0$, on $\|\sigma_i\|_\infty$, on the bounds on the derivatives of b_i and σ_i and on their Lipschitz constants.

Remark 8.2.9. Assumption (8.2.12) is satisfied in the case of a Brownian motion, which suggests that this assumption is satisfied in general provided that σ is "regular enough". Indeed, if $dZ_t = \sigma dW_t$ with $\sigma \in \mathcal{M}_d(\mathbb{R})$ being non degenerate, then with $\Sigma := \sigma\sigma^\top$ we have for $t > 0$ and $x, y \in \mathbb{R}^d$:

$$p_{_Z}(0, t, x, y) = \frac{1}{\sqrt{\det(\Sigma)} t^{d/2}} \Phi\left(\Sigma^{-1/2} \frac{y - x}{\sqrt{t}} \right), \quad \Phi(u) := \frac{1}{(2\pi)^{d/2}} e^{-|u|^2/2}.$$

Moreover for every $r \in \mathbb{N}$ and $u \in \mathbb{R}^d$ we have

$$\left\| \frac{d^r}{du^r} \Phi(u) \right\| \leq \frac{1}{(2\pi)^{d/2}} |\operatorname{He}_r(|u|)| e^{-|u|^2/2}$$

where He_r is the r^{th} probabilist Hermite polynomial. Following [Kra04] we have

$$\forall u \geq 0,\ |\operatorname{He}_{2r}(u)| e^{-u^2/2} \leq C 2^{-r} \sqrt{r} \frac{(2r)!^2}{r!^2} \leq C 2^r \sqrt{r},$$

using the Stirling formula for the last inequality. On the other hand, using [AS64, 22.14.15] we have

$$\forall u \geq 0,\ |\operatorname{He}_{2r}(u)| e^{-u^2/4} \leq 2^{r+1} r!.$$

Then, for for every $\varepsilon \in (0, 1]$,

$$\begin{aligned} |\operatorname{He}_{2r}(u)| e^{-u^2/2} &= \left| \operatorname{He}_{2r}(u) e^{-u^2/4} \right|^{\varepsilon} \left| \operatorname{He}_{2r}(u) e^{-u^2/2} \right|^{1-\varepsilon} e^{-\varepsilon u^2/4} \\ &\leq C \left(2^r r! \right)^{\varepsilon} \left(2^r r^{1/2} \right)^{1-\varepsilon} e^{-\varepsilon u^2/2}. \end{aligned}$$

Then if we choose $\varepsilon_r = \log^{-1}(r)$ for $r \geq 3$, we have $(r!)^{\varepsilon_r} \leq e^{\varepsilon_r r \log(r)} = e^r$ so that

$$\left\| \frac{d^r}{du^r} \Phi(u) \right\| \leq C 2^{r\varepsilon_r} e^r 2^r r^{1/2} e^{-\varepsilon_r |u|^2/2} =: A_r e^{-\varepsilon_r |u|^2/2}$$

and then for $r \geq 3$ we have

$$\begin{aligned} \|\nabla_y^{2r} p_{_Z}(0,t,x,y)\| &\leq \frac{\|\Sigma^{-1/2}\|^{2r}}{\sqrt{\det(\Sigma)} t^{(d+2r)/2}} \frac{d^{2r}}{du^{2r}} \Phi\left(\Sigma^{-1/2} \frac{y-x}{\sqrt{t}}\right) \\ &\leq \frac{\|\Sigma^{-1/2}\|^{2r}}{\sqrt{\det(\Sigma)} t^{(d+2r)/2}} A_r e^{-\varepsilon_r \|\Sigma^{-1}\| |y-x|^2/(2t)} \end{aligned}$$

where $\left(\left(\|\Sigma^{-1/2}\|^{2r} A_r \varepsilon_r^{-d/2} \right)^{1/(2r)} \right)$ is bounded. Thus Assumption (8.2.12) is satisfied.

Remark 8.2.10. For the Euler-Maruyama scheme (8.2.3), with a slight abuse of notation, x is used both for the starting point and in the definition of the drift and diffusion coefficients. The transition density should be considered for constant drift and diffusion coefficients in this case. However the results remain valid as the Euler scheme is simply a Brownian process.

8.3 Proof of the Theorems

8.3.1 Recalls on density estimates for SDEs with bounded drift

We recall results on the bounds for the density of the solution of the SDE using the theory of partial differential equations. Let us consider a generic SDE:

$$Z_0^x = x \in \mathbb{R}^d, \quad dZ_t^x = b_{_Z}(t, Z_t^x)dt + \sigma_{_Z}(t, Z_t^x)dW_t, \ t \in [0,T]. \tag{8.3.1}$$

Then under regularity assumptions on b_Z and on σ_Z, the transition probability $p_{_Z}$ exists and is solution of the backward Kolmogorov PDE:

$$\begin{aligned} &p_{_Z}(t,t,x,\cdot) = \delta_x, \ t \in [0,T], \\ &\partial_s p_{_Z}(s,t,x,y) = \langle b_{_Z}(s,x), \nabla_x p_{_Z}(s,t,x,y) \rangle + \frac{1}{2} \mathrm{Tr}\left(\sigma_{_Z}^\top(s,x) \nabla_x^2 p_{_Z}(s,t,x,y) \sigma_{_Z}(s,x) \right) \end{aligned} \tag{8.3.2}$$

for $s < t \in [0,T]$. Moreover, $p_{_Z}$ and its derivatives satisfy sub-gaussian Aronson's bounds:

Theorem 8.3.1 ([Fri64], Chapter 9, Theorem 7)**.** *Let Z be the solution of* (8.3.1) *and let $T > 0$. Assume* $\mathrm{Lip}_t(b_{_Z})$ *and* $\mathrm{Lip}_t(\sigma_{_Z})$*, that $b_{_Z}, \sigma_{_Z} \in \mathcal{C}_b^r$ and that $\sigma_{_Z}$ is elliptic. Then for every $m_0 = 0, 1$ and for every $0 \leq m_1 + m_2 \leq r$, $\nabla_x^{m_0+m_1} \nabla_y^{m_2} p_{_Z}$ exists and*

$$\forall s<t \in [0,T], \forall x, y \in \mathbb{R}^d, \ \|\nabla_x^{m_0+m_1} \nabla_y^{m_2} p_{_Z}(s,t,x,y)\| \leq \frac{C}{(t-s)^{(d+m_0+m_1+m_2)/2}} e^{-c|y-x|^2/(t-s)}, \tag{8.3.3}$$

where the constants C and c only depend on the bounds on $b_{_Z}$ and $\sigma_{_Z}$ and on their derivatives and their Lipschitz constants, on the modulus of ellipticity of $\sigma_{_Z}$, on d and on T.

Let us also recall the recent result from [MPZ21] giving Aronson's bounds of the partial derivatives with respect to y in the case where b_Z is unbounded. Considering [MPZ21, Section 4] with [MPZ21, (3.1)], we have the following result.

Theorem 8.3.2. *Let Z be the solution of* (8.3.1) *and let $T > 0$. Assume* $\mathrm{Lip}_t(b_{_Z})$ *and* $\mathrm{Lip}_t(\sigma_{_Z})$*, that $b_{_Z} \in \widetilde{\mathcal{C}}_b^r$, $\sigma_{_Z} \in \mathcal{C}_b^{r+1}$ and that $\sigma_{_Z}$ is elliptic. Then for every $0 \leq m \leq r$, $\nabla_y^m p_{_Z}$ exists and*

$$\forall s < t \in [0,T], \ \forall x, y \in \mathbb{R}^d, \ \|\nabla_y^m p_{_Z}(s,t,x,y)\| \leq \frac{C}{(t-s)^{(d+m)/2}} e^{-c|y-x|^2/(t-s)}, \tag{8.3.4}$$

where the constants C and c only depend on the bounds on $b_{_Z}$ and $\sigma_{_Z}$ and on their derivatives and on their Lipschitz constants, on the modulus of ellipticity of $\sigma_{_Z}$, on d and on T.

8.3.2 Preliminary results

In order to apply the bounds on the densities from Theorem 8.3.1 to Theorem 8.2.7, we first "cut" the drifts b_1 and b_2 on a compact set. That is, we instead consider the processes $\widetilde{X}$ and $\widetilde{Y}$ defined by

$$d\widetilde{X}_t^x = \tilde{b}_1^x(t, \widetilde{X}_t^x)dt + \sigma_1(t, \widetilde{X}_t^x)dW_t,\ t \in [0,T], \tag{8.3.5}$$
$$d\widetilde{Y}_t^x = \tilde{b}_2^x(t, \widetilde{Y}_t^x)dt + \sigma_2(t, \widetilde{Y}_t^x)dW_t,\ t \in [0,T], \tag{8.3.6}$$

where $\tilde{b}_i$, $i = 1, 2$ is defined as follows. We choose $R > 0$ and we consider a $\mathcal{C}^\infty$ decreasing function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\psi = 1$ on $[0, R^2]$ and $\psi = 0$ on $[(R+1)^2, \infty)$ and we define $\tilde{b}_i^x(t,y) := b_i(t,y)\psi(|y-x|^2)$, so that $\tilde{b}_i^x$ is bounded:

$$\forall y \in \mathbb{R}^d,\ \forall t \in [0,T],\ |\tilde{b}_i^x(t,y)| \leq \sup_{z \in \mathbf{B}(x,R+1)} |b_i(t,z)| \leq C(1+|x|), \tag{8.3.7}$$

because b_i is Lipschitz continuous.

Lemma 8.3.3. *Assume* $\mathrm{Lip}_t(b_1)$, $\mathrm{Lip}_t(\sigma_1)$, $b_1 \in \widetilde{\mathcal{C}}_b^1$, $\sigma_1 \in \mathcal{C}_b^1$. *Then for every* $x \in \mathbb{R}^d$ *and* $t \in [0,T]$,

$$\mathrm{d_{TV}}(X_t^x, \widetilde{X}_t^x) \leq C(1+|b_1(0,x)|^2)t. \tag{8.3.8}$$

Proof. Let $f : \mathbb{R}^d \to \mathbb{R}$ be measurable and bounded. We remark that on the event $\{\sup_{s\in[0,t]} |X_s^x - x|^2 \leq R^2\}$, we have $\widetilde{X}_t^x = X_t^x$, so that

$$|\mathbb{E}f(\widetilde{X}_t^x) - \mathbb{E}f(X_t^x)| \leq 2\|f\|_\infty \mathbb{P}\left(\sup_{s\in[0,t]} |X_s^x - x|^2 > R^2\right).$$

But using the inequality $|u+v|^2 \leq 2|u|^2 + 2|v|^2$ we have

$$\begin{aligned}
|X_t^x - x|^2 &\leq 2\left|\int_0^t b_1(s, X_s^x)ds\right|^2 + 2\left|\int_0^t \sigma_1(s, X_s^x)dW_s\right|^2 \\
&\leq 2t\int_0^t |b_1(s, X_s^x)|^2 ds + 2\left|\int_0^t \sigma_1(s, X_s^x)dW_s\right|^2 \\
&\leq 4t[b_1]_{\mathrm{Lip}}^2\left(\int_0^t |X_s^x - x|^2 ds + \frac{1}{3}t^3\right) + 4t^2|b_1(0,x)|^2 + 2\left|\int_0^t \sigma_1(s, X_s^x)dW_s\right|^2
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{E}\sup_{s\in[0,t]} |X_s^x - x|^2 \leq 4t[b_1]_{\mathrm{Lip}}^2 \int_0^t \left(\mathbb{E}\sup_{u\in[0,s]} |X_u^x - x|^2\right) ds + \frac{4}{3}[b_1]_{\mathrm{Lip}}^2 t^4 + 4t^2|b_1(0,x)|^2 \\
+ 2\mathbb{E}\sup_{s\in[0,t]} \left|\int_0^s \sigma_1(u, X_u^x)dW_u\right|^2.
\end{aligned}$$

Moreover using Doob's martingale inequality we have

$$\mathbb{E}\sup_{s\in[0,t]} \left|\int_0^s \sigma_1(u, X_u^x)dW_u\right|^2 \leq 4\mathbb{E}\left|\int_0^t \sigma_1(u, X_u^x)dW_u\right|^2 = 4\mathbb{E}\int_0^t \sigma_1^2(u, X_u^x)du \leq 4\|\sigma_1\|_\infty^2 t.$$

Then we define the non-decreasing deterministic process $S_t := \mathbb{E}\sup_{s\in[0,t]} |X_s^x - x|^2$ and we get the differential inequality (using $t \leq T$)

$$S_t \leq 4t\left(T|b_1(0,x)|^2 + \frac{1}{3}[b_1]_{\mathrm{Lip}}^2 T^3 + 2\|\sigma_1\|_\infty^2\right) + 4t[b_1]_{\mathrm{Lip}}^2 \int_0^t S_s ds,$$

so the Gronwall lemma yields

$$S_t \leq 4t\left(T|b_1(0,x)|^2 + \frac{1}{3}[b_1]^2_{\mathrm{Lip}}T^3 + 2\|\sigma_1\|^2_\infty\right)e^{2t^2[b_1]^2_{\mathrm{Lip}}} \leq C(1+|b_1(0,x)|^2)t.$$

Using Markov's inequality, we have then

$$|\mathbb{E}f(\widetilde{X}^x_t) - \mathbb{E}f(X^x_t)| \leq 2\|f\|_\infty \mathbb{P}\left(\sup_{s\in[0,t]}|X^x_s - x|^2 > R^2\right) \leq 2\|f\|_\infty \frac{C(1+|b_1(0,x)|^2)t}{R^2}.$$

□

We can now apply Theorem 8.3.1 to $\widetilde{X}$ and to $\widetilde{Y}$ however the constants arising depend on the bound on $\|\tilde{b}^x_i\|_\infty$ and thus on x. In order to deal with the dependency in $\|\tilde{b}^x_i\|_\infty$, we apply the Girsanov formula and reduce to the null drift case.

Proposition 8.3.4. *Let Z^x be the solution of*

$$Z^x_0 = x, \quad dZ^x_t = \sigma_1(t, Z^x_t)dW_t, \ t \in [0,T]. \tag{8.3.9}$$

Assume $\mathrm{Lip}_t(b_1)$, $\mathrm{Lip}_t(\sigma_1)$, $b_1 \in \widetilde{C}^1_b$, $\sigma_1 \in \mathcal{C}^1_b$ *and σ_1 is elliptic. Then we have for every $t \in [0,T]$, $x, y \in \mathbb{R}^d$,*

$$p_{\widetilde{X}}(0,t,x,y) = p_Z(0,t,x,y) + \int_0^t \mathbb{E}\left[U^x_s\langle \tilde{b}^x_1(s,Z^x_s), \nabla_x p_Z(s,t,Z^x_s,y)\rangle\right]ds, \tag{8.3.10}$$

where $\widetilde{X}$ is defined in (8.3.5) *and U^x is defined as*

$$U^x_s = \exp\left(\int_0^s \langle g(u,Z^x_u)\tilde{b}^x_1(u,Z^x_u), dZ^x_u\rangle - \frac{1}{2}\int_0^s \langle g(u,Z^x_u)\tilde{b}^x_1(u,Z^x_u), \tilde{b}^x_1(u,Z^x_u)\rangle du\right), \tag{8.3.11}$$

$$g = (\sigma_1\sigma_1^\top)^{-1}. \tag{8.3.12}$$

Proof. First, note that since σ_1 is elliptic and since $\tilde{b}^x_1, \sigma_1 \in \mathcal{C}^1_b$, then $p_{\widetilde{X}}$ and p_Z exist as well as $\nabla_x p_Z$ (Theorem 8.3.1). We then use [QZ04, Theorem 2.4] extended to non-homogeneous diffusion processes. Following [QZ04, Remark 2.5], since σ_1 is elliptic and bounded, the assumptions of [QZ04, Theorem 2.4] hold. □

We also have the following bounds on the process U^x.

Lemma 8.3.5. *With the same assumptions as in Proposition 8.3.4, for every $p \geq 2$, $x \in \mathbb{R}^d$ and $t \in [0,T]$ we have*

$$\mathbb{E}\left[\sup_{s\in[0,t]}|U^x_s|^p\right] \leq e^{Cp^2\|\tilde{b}^x_1\|^2_\infty t}. \tag{8.3.13}$$

Proof. We recall that for every $q \geq 1$, the process $(U^x)^q$ is a martingale with

$$d(U^x_s)^q = q(U^x_s)^q\langle g(s,Z^x_s)\tilde{b}^x_1(s,Z^x_s), \sigma_1(s,Z^x_s)dW_s\rangle.$$

Thus, Doob's martingale inequality yields

$$\mathbb{E}\left[\sup_{s\in[0,t]}|U^x_s|^{2q}\right] \leq Cq^2\|\tilde{b}^x_1\|^2_\infty \mathbb{E}\int_0^t |U^x_s|^{2q}ds \leq Cq^2\|\tilde{b}^x_1\|^2_\infty \int_0^t \mathbb{E}\left[\sup_{u\in[0,s]}|U^x_s|^{2q}\right]ds.$$

So with $U^x_0 = 1$ we obtain

$$\mathbb{E}\left[\sup_{s\in[0,t]}|U^x_s|^{2q}\right] \leq e^{Cq^2\|\tilde{b}^x_1\|^2_\infty t}.$$

□

Lemma 8.3.6. *With the same assumptions as in Proposition 8.3.4, we have for every $x, y \in \mathbb{R}^d$ and $t \in [0,T]$,*

$$\left| \int_0^t \mathbb{E}\left[U_s^x \langle \tilde{b}_1^x(s, Z_s^x), \nabla_x p_Z(s,t,Z_s^x,y) \rangle \right] ds \right| \leq C e^{C\|\tilde{b}_1^x\|_\infty^2} \frac{e^{-c|y-x|^2/t}}{t^{(d-1)/2}}. \tag{8.3.14}$$

Proof. We use Theorem 8.3.1 on the process Z^x, which yields bounds with constants depending on σ_1 but not on $\tilde{b}_1^x$. We obtain for every $q \geq 1$ and for every $s \in [0,t]$:

$$\begin{aligned}
\mathbb{E}\,|\nabla_x p_Z(s,t,Z_s^x,y)|^q &= \int_{\mathbb{R}^d} |\nabla_x p_Z(s,t,\xi,y)|^q p_Z(0,s,x,\xi) d\xi \\
&\leq \frac{C_q}{(t-s)^{(q+(q-1)d)/2}} \int_{\mathbb{R}^d} \frac{1}{(s(t-s))^{d/2}} \exp\left(-c_q \left(\frac{|y-\xi|^2}{t-s} + \frac{|\xi - x|^2}{s} \right) \right) d\xi \\
&\leq \frac{C_q}{t^{d/2}} \frac{1}{(t-s)^{(q+(q-1)d)/2}} e^{-c_q|y-x|^2/t},
\end{aligned}$$

where we used Lemma 8.5.1 in the appendix. Then for $p^{-1} + q^{-1} = 1$ and $p \geq 2$, using the Hölder inequality we have

$$\begin{aligned}
&\left| \int_0^t \mathbb{E}\left[U_s^x \langle \tilde{b}_1^x(s, Z_s^x), \nabla_x p_Z(s,t,Z_s^x,y) \rangle \right] ds \right| \\
&\quad \leq \|\tilde{b}_1^x\|_\infty \Big(\sup_{s\in[0,t]} \mathbb{E}|U_s^x|^p \Big)^{1/p} \int_0^t \left(\mathbb{E}\,|\nabla_x p_Z(s,t,Z_s^x,y)|^q \right)^{1/q} ds \\
&\quad \leq \|\tilde{b}_1^x\|_\infty e^{Cp\|\tilde{b}_1^x\|_\infty^2 t} \frac{C_q e^{-c_q|y-x|^2/t}}{t^{d/(2q)}} \int_0^t \frac{ds}{(t-s)^{(1+(1-q^{-1})d)/2}}.
\end{aligned}$$

The integral in ds converges under the condition $q < d/(d-1)$ if $d > 1$, and for any value of $q > 1$ if $d = 1$. Then performing the change of variable $s = tu$ we obtain

$$\begin{aligned}
\left| \int_0^t \mathbb{E}\left[U_s^x \langle \tilde{b}_1^x(s, Z_s^x), \nabla_x p_Z(s,t,Z_s^x,y) \rangle \right] ds \right| &\leq \|\tilde{b}_1^x\|_\infty e^{Cp\|\tilde{b}_1^x\|_\infty^2 T} \frac{C_q e^{-c_q|y-x|^2/t}}{t^{(d-1)/2}} \\
&\leq C e^{C\|\tilde{b}_1^x\|_\infty^2} \frac{e^{-c|y-x|^2/t}}{t^{(d-1)/2}}.
\end{aligned}$$

□

8.3.3 Proof of Theorem 8.2.1

Lemma 8.3.7. *We have for every $x \in \mathbb{R}^d$ and $t \in [0,T]$:*

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \leq \mathrm{d_{TV}}(Z_t^x, V_t^x) + C e^{C|x|^2} t^{1/2},$$

where $dZ_t^x = \sigma_1(t, Z_t^x) dW_t$ and $dV_t^x = \sigma_2(t, V_t^x) dW_t$.

Proof. Let us write

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \leq \mathrm{d_{TV}}(X_t^x, \widetilde{X}_t^x) + \mathrm{d_{TV}}(\widetilde{X}_t^x, Z_t^x) + \mathrm{d_{TV}}(Z_t^x, V_t^x) + \mathrm{d_{TV}}(V_t^x, \widetilde{Y}_t^x) + \mathrm{d_{TV}}(\widetilde{Y}_t^x, Y_t^x),$$

where $\widetilde{X}$ and $\widetilde{Y}$ are defined in (8.3.5) and (8.3.6). Using Lemma 8.3.3 with (8.3.7), we have

$$\mathrm{d_{TV}}(X_t^x, \widetilde{X}_t^x) + \mathrm{d_{TV}}(\widetilde{Y}_t^x, Y_t^x) \leq C(1 + |x|^2)t.$$

Using the formula (8.3.10) and the inequality (8.3.14), we have

$$\begin{aligned}\mathrm{d_{TV}}(\widetilde{X}_t^x, Z_t^x) &= \int_{\mathbb{R}^d} |p_{\widetilde{X}}(0,t,x,y) - p_Z(0,t,x,y)| dy \\ &= \int_{\mathbb{R}^d} \left| \int_0^t \mathbb{E}\left[U_s^x \langle \tilde{b}_1^x(s, Z_s^x), \nabla_x p_Z(s,t,Z_s^x,y) \rangle \right] ds \right| dy \\ &\leq C e^{C\|\tilde{b}_1^x\|_\infty^2} t^{1/2} \int_{\mathbb{R}^d} \frac{e^{-c|x-y|^2/t}}{t^{d/2}} dy \leq C e^{C|x|^2} t^{1/2},\end{aligned}$$

where we used (8.3.7). The term $\mathrm{d_{TV}}(\widetilde{Y}_t^x, V_t^x)$ is treated likewise. □

We now prove Theorem 8.2.1.

Proof. Let us introduce an artificial regularization. For $\varepsilon > 0$ and using Lemma 8.3.7 we have

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \leq C e^{C|x|^2} t^{1/2} + \mathrm{d_{TV}}(Z_t^x, Z_t^x + \sqrt{\varepsilon}\zeta) + \mathrm{d_{TV}}(Z_t^x + \sqrt{\varepsilon}\zeta, V_t^x + \sqrt{\varepsilon}\zeta) \tag{8.3.15}$$

$$+ \mathrm{d_{TV}}(V_t^x + \sqrt{\varepsilon}\zeta, V_t^x) \tag{8.3.16}$$

where $\zeta \sim \mathcal{N}(0, I_d)$ and is independent of the Brownian motion W.

• Let $f : \mathbb{R}^d \to \mathbb{R}$ be measurable and bounded and let us define

$$\varphi : y \in \mathbb{R}^d \mapsto \mathbb{E} f(Z_t^x + y) = \int_{\mathbb{R}^d} f(\xi + y) p_Z(0,t,x,\xi) d\xi = \int_{\mathbb{R}^d} f(\xi) p_Z(0,t,x,\xi - y) d\xi. \tag{8.3.17}$$

Then φ is $\mathcal{C}^2$ with

$$\nabla^2 \varphi(y) = \int_{\mathbb{R}^d} f(\xi) \nabla_y^2 p_Z(0,t,x,\xi - y) d\xi.$$

Moreover, using Theorem 8.3.1, we have

$$\|\nabla_y^2 p_Z(0,t,x,\xi - y)\| \leq \frac{C}{t^{(d+2)/2}} e^{-c|x-\xi+y|^2/t},$$

where the constants C and c do not depend on $\tilde{b}_1^x$. This implies that for every $y \in \mathbb{R}^d$,

$$\|\nabla^2 \varphi(y)\|_\infty \leq C\|f\|_\infty t^{-1} \int_{\mathbb{R}^d} \frac{1}{t^{d/2}} e^{-c|x-\xi+y|^2/t} d\xi \leq C\|f\|_\infty t^{-1}.$$

Then using the Taylor formula, for every $y \in \mathbb{R}^d$ there exists $\tilde{y} \in (0, y)$ such that

$$\varphi(y) = \varphi(0) + \nabla\varphi(0) \cdot y + \frac{1}{2} \nabla^2 \varphi(\tilde{y}) \cdot y^{\otimes 2}$$

and then for some random $\tilde{\zeta} \in (0, \zeta)$ we have

$$\begin{aligned}|\mathbb{E} f(Z_t^x + \sqrt{\varepsilon}\zeta) - \mathbb{E} f(Z_t^x)| = |\mathbb{E}\varphi(\sqrt{\varepsilon}\zeta) - \varphi(0)| &= \left| \sqrt{\varepsilon}\mathbb{E}[\nabla\varphi(0) \cdot \zeta] + \frac{\varepsilon}{2}\mathbb{E}[\nabla^2 \varphi(\sqrt{\varepsilon}\tilde{\zeta}) \cdot \zeta^{\otimes 2}] \right| \\ &\leq C\varepsilon\|f\|_\infty t^{-1},\end{aligned}$$

where we used that $\mathbb{E}[\nabla\varphi(0) \cdot \zeta] = \nabla\varphi(0) \cdot \mathbb{E}[\zeta] = 0$. This way we obtain

$$\mathrm{d_{TV}}(Z_t^x, Z_t^x + \sqrt{\epsilon}\zeta) \leq C\varepsilon t^{-1}.$$

The term $\mathrm{d_{TV}}(V_t^x, V_t^x + \sqrt{\epsilon}\zeta)$ is treated likewise.

• Let $f : \mathbb{R}^d \to \mathbb{R}$ be measurable and bounded and let us define

$$f_\varepsilon : y \mapsto \mathbb{E}f(y + \sqrt{\varepsilon}\zeta) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(y + \sqrt{\varepsilon}\xi) e^{-|\xi|^2/2} d\xi = \frac{1}{(2\pi\varepsilon)^{d/2}} \int_{\mathbb{R}^d} f(\xi) e^{-|\xi-y|^2/(2\varepsilon)} d\xi. \tag{8.3.18}$$

Then f_ε is $\mathcal{C}^1$ with

$$\begin{aligned}\nabla f_\varepsilon(y) &= \frac{1}{(2\pi\varepsilon)^{d/2}} \int_{\mathbb{R}^d} f(\xi) \frac{\xi - y}{\varepsilon} e^{-|\xi-y|^2/(2\varepsilon)} d\xi = \frac{\varepsilon^{-1/2}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(y + \sqrt{\varepsilon}\xi) \xi e^{-|\xi|^2/2} d\xi \\ &= \varepsilon^{-1/2} \mathbb{E}[f(y + \sqrt{\varepsilon}\zeta)\zeta]\end{aligned}$$

and then

$$[f_\varepsilon]_{\mathrm{Lip}} \le \|f\|_\infty \varepsilon^{-1/2} \mathbb{E}|\zeta| \le C\|f\|_\infty \varepsilon^{-1/2}. \tag{8.3.19}$$

So that

$$\begin{aligned}&|\mathbb{E}f(Z_t^x + \sqrt{\varepsilon}\zeta) - \mathbb{E}f(V_t^x + \sqrt{\varepsilon}\zeta)| = |\mathbb{E}f_\varepsilon(Z_t^x) - \mathbb{E}f_\varepsilon(V_t^x)| \\ &\quad \le \frac{C\|f\|_\infty}{\sqrt{\varepsilon}} \|Z_t^x - V_t^x\|_1 \le \frac{C\|f\|_\infty}{\sqrt{\varepsilon}} (t + t^{1/2}\Delta\sigma(x)),\end{aligned} \tag{8.3.20}$$

where we used Lemma 8.5.3 in the Appendix. This implies that

$$\mathrm{d_{TV}}(Z_t^x + \sqrt{\varepsilon}\zeta, V_t^x + \sqrt{\varepsilon}\zeta) \le C\varepsilon^{-1/2}(t + t^{1/2}\Delta\sigma(x)).$$

• **Conclusion :** Considering (8.3.15), we have

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \le Ce^{C|x|^2} t^{1/2} + C\varepsilon t^{-1} + C\varepsilon^{-1/2}(t + t^{1/2}\Delta\sigma(x)).$$

We now choose $\varepsilon = \left[t(t + t^{1/2}\Delta\sigma(x))\right]^{2/3}$, so that

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \le Ce^{C|x|^2} t^{1/2} + C(t^{1/2} + \Delta\sigma(x))^{2/3}.$$

□

8.3.4 Proof of Theorem 8.2.2 using Theorem 8.2.7

We first prove Theorem 8.2.7.

Proof. Let $f : \mathbb{R}^d \to \mathbb{R}$ be measurable and bounded, let $\varepsilon > 0$ and let $\zeta \sim \mathcal{N}(0, I_d)$ be independent of (Z_1, Z_2). We have

$$\begin{aligned}|\mathbb{E}f(Z_1) - \mathbb{E}f(Z_2)| \le &\left|\mathbb{E}f(Z_1) - \sum_{i=1}^r w_i \mathbb{E}f_{\varepsilon/n_i}(Z_1)\right| + \left|\sum_{i=1}^r w_i \mathbb{E}f_{\varepsilon/n_i}(Z_1) - \sum_{i=1}^r w_i \mathbb{E}f_{\varepsilon/n_i}(Z_2)\right| \\ &+ \left|\sum_{i=1}^r w_i \mathbb{E}f_{\varepsilon/n_i}(Z_2) - \mathbb{E}f(Z_2)\right|,\end{aligned} \tag{8.3.21}$$

where f_ε is defined as in (8.3.18) and where the n_i's and the w_i's will be defined later.

Let φ be as defined in (8.3.17) replacing Z_t^x by Z_1. Then, φ is differentiable up to the order $2r$ and for all $k = 0, 1, \ldots, 2r$:

$$\nabla^k \varphi(y) = (-1)^k \int_{\mathbb{R}^d} f(\xi) \nabla^k p_1(\xi - y) d\xi.$$

Using the Taylor formula up to order $2r$, for every $y \in \mathbb{R}^d$ there exits $\tilde{y} \in (0, y)$ such that

$$\varphi(y) = \varphi(0) + \sum_{k=1}^{2r-1} \frac{\nabla^k \varphi(0)}{k!} \cdot y^{\otimes k} + \frac{\nabla^{2r}\varphi(\tilde{y})}{(2r)!} \cdot y^{\otimes 2r}.$$

Moreover, we have

$$\left|\nabla^{2r}\varphi(\tilde{y}) \cdot y^{\otimes 2r}\right| \leq C\|f\|_\infty |y|^{2r} \int_{\mathbb{R}^d} \|\nabla^{2r} p_1(\xi)\| d\xi. \tag{8.3.22}$$

Then there exists a random $\tilde{\zeta} \in (0, \zeta)$ such that

$$\begin{aligned}
\mathbb{E}f(Z_1 + \sqrt{\varepsilon}\zeta) - \mathbb{E}f(Z_1) &= \mathbb{E}\varphi(\sqrt{\varepsilon}\zeta) - \varphi(0) &(8.3.23)\\
&= \sum_{k=1}^{2r-1} \frac{\nabla^k \varphi(0)}{k!} \varepsilon^{k/2} \cdot \mathbb{E}[\zeta^{\otimes k}] + \frac{\mathbb{E}[\nabla^{2r}\varphi(\sqrt{\varepsilon}\tilde{\zeta}) \cdot \zeta^{\otimes 2r}]}{(2r)!}\varepsilon^r \\
&= \sum_{k=1}^{r-1} \frac{\nabla^{2k}\varphi(0)}{(2k)!} \varepsilon^k \cdot \mathbb{E}[\zeta^{\otimes 2k}] + \frac{\mathbb{E}[\nabla^{2r}\varphi(\sqrt{\varepsilon}\tilde{\zeta}) \cdot \zeta^{\otimes 2r}]}{(2r)!}\varepsilon^r =: \sum_{k=1}^{r-1} \beta_k(t)\varepsilon^k + \tilde{\beta}_r(t,\varepsilon)\varepsilon^r, &(8.3.24)
\end{aligned}$$

because if k is odd, then $\mathbb{E}[\zeta^{\otimes k}] = 0$. We now rely on a multi-step Richardson-Romberg extrapolation [LP17, Appendix A]. Let us denote the refiners $n_i = 2^{i-1}$ and the auxiliary sequences and weights

$$u_k := \left(\prod_{\ell=1}^{k-1}(1 - 2^{-\ell})\right)^{-1}, \; v_k := (-1)^k 2^{-k(k+1)/2} u_{k+1}, \; w_k := u_k v_{r-k}, \; k = 1, \dots, r. \tag{8.3.25}$$

These weights are the unique solution to the $r \times r$ Vandermonde system

$$\sum_{i=1}^{r} w_i n_i^{-k} = \begin{cases} 1 & \text{if } k = 0, \\ 0 & \text{else.} \end{cases}, \quad k = 0, 1, \dots, r-1. \tag{8.3.26}$$

Then we have

$$\begin{aligned}
\sum_{i=1}^{r} w_i \left(\mathbb{E}f(Z_1 + \sqrt{\varepsilon/n_i}\zeta) - \mathbb{E}f(Z_1)\right) &= \sum_{i=1}^{r} w_i \sum_{k=1}^{r-1} \beta_k(t)\varepsilon^k n_i^{-k} + \sum_{i=1}^{r} w_i \tilde{\beta}_r(t, \varepsilon/n_i)\varepsilon^r n_i^{-r} \\
&= \sum_{k=1}^{r-1} \varepsilon^k \beta_k(t) \sum_{i=1}^{r} w_i n_i^{-k} + \varepsilon^r \sum_{i=1}^{r} \tilde{\beta}_r(t, \varepsilon/n_i) w_i n_i^{-r} \\
&= \varepsilon^r \sum_{i=1}^{r} \tilde{\beta}_r(t, \varepsilon/n_i) w_i n_i^{-r}, &(8.3.27)
\end{aligned}$$

where we used (8.3.26) in the last equation. Now, using (8.3.22) we have

$$\left|\sum_{i=1}^{r} \tilde{\beta}_r(t, \varepsilon/n_i) w_i n_i^{-r}\right| \leq C\|f\|_\infty \left(\int_{\mathbb{R}^d} \|\nabla^{2r} p_1(\xi)\| d\xi\right) \sum_{i=1}^{r} |w_i| n_i^{-r}.$$

Since $u_k \to u_\infty = \prod_{\ell \geq 1}(1 - 2^{-\ell})^{-1} < \infty$, the weights satisfy

$$|w_i| \leq u_\infty^2 2^{-(r-i)(r-i+1)/2}, \quad i = 1, \dots, r,$$

so that

$$\sum_{i=1}^{r} \frac{|w_i|}{n_i^r} \leq u_\infty^2 \sum_{i=1}^{r} 2^{-(r-i)(r-i+1)/2} \leq u_\infty^2 \sum_{i=1}^{r} 2^{(r-i)/2} = u_\infty^2 \sum_{i=0}^{r-1} 2^{-i/2} \leq C. \tag{8.3.28}$$

As a consequence and since $\sum_{i=1}^r w_i = 1$, we may write from (8.3.27)

$$\left|\mathbb{E}f(Z_1) - \sum_{i=1}^r w_i \mathbb{E}f_{\varepsilon/n_i}(Z_1)\right| \leq C\|f\|_\infty \varepsilon^r \int_{\mathbb{R}^d} \|\nabla^{2r} p_1(\xi)\| d\xi. \tag{8.3.29}$$

The same way, we obtain

$$\left|\mathbb{E}f(Z_2) - \sum_{i=1}^r w_i \mathbb{E}f_{\varepsilon/n_i}(Z_2)\right| \leq C\|f\|_\infty \varepsilon^r \int_{\mathbb{R}^d} \|\nabla^{2r} p_2(\xi)\| d\xi.$$

On the other side, using (8.3.19) we have

$$\left|\sum_{i=1}^r w_i \mathbb{E}f_{\varepsilon/n_i}(Z_1) - \sum_{i=1}^r w_i \mathbb{E}f_{\varepsilon/n_i}(Z_2)\right| \leq \frac{C\|f\|_\infty}{\sqrt{\varepsilon}} \mathcal{W}_1(Z_1, Z_2) \left(\sum_{i=1}^r |w_i| 2^{(i-1)/2}\right). \tag{8.3.30}$$

Moreover, for every $i = 1, \ldots, r$,

$$|w_i| 2^{(i-1)/2} \leq u_\infty^2 2^{-(r-i)(r-i+1)/2 + (i-1)/2}$$

and then

$$\sum_{i=1}^r |w_i| 2^{(i-1)/2} \leq u_\infty^2 \sum_{i=1}^r 2^{(i-1)/2} \leq u_\infty^2 2^r. \tag{8.3.31}$$

Thus considering (8.3.21), we obtain for every $\varepsilon > 0$,

$$\mathrm{d_{TV}}(Z_1, Z_2) \leq C\varepsilon^r \int_{\mathbb{R}^d} \left(\|\nabla^{2r} p_1(\xi)\| + \|\nabla^{2r} p_2(\xi)\|\right) d\xi + C\varepsilon^{-1/2} \mathcal{W}_1(Z_1, Z_2).$$

Optimizing in ε gives

$$\varepsilon_\star = \left(\mathcal{W}_1(Z_1, Z_2)/(2r \int_{\mathbb{R}^d} \left(\|\nabla^{2r} p_1(\xi)\| + \|\nabla^{2r} p_2(\xi)\|\right) d\xi\right)^{2/(2r+1)}$$

and then

$$\mathrm{d_{TV}}(Z_1, Z_2) \leq C_{d,r} \mathcal{W}_1(Z_1, Z_2)^{2r/(2r+1)} \left(\int_{\mathbb{R}^d} \left(\|\nabla^{2r} p_1(\xi)\| + \|\nabla^{2r} p_2(\xi)\|\right) d\xi\right)^{1/(2r+1)}.$$

□

We now prove Theorem 8.2.2.

Proof. Using Lemma 8.3.7, we have

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \leq Ce^{C|x|^2} t^{1/2} + \mathrm{d_{TV}}(Z_t^x, V_t^x) \tag{8.3.32}$$

We now apply Theorem 8.2.7 with the random vectors $Z_1 = Z_t^x$ and $Z_2 = V_t^x$. Assuming that σ_1 is C_b^{2r} and using Theorem 8.3.1, $\nabla_y^k p_Z$ exists for $k = 0, 1, \ldots, 2r$ and

$$\forall k = 0, 1, \ldots, 2r,\ \forall t \in (0, T],\ \forall x, y \in \mathbb{R}^d,\ \|\nabla_y^k p_Z(0, t, x, y)\| \leq \frac{C}{t^{(d+k)/2}} e^{-c|y-x|^2/t}.$$

Then we have

$$\int_{\mathbb{R}^d} \nabla_y^{2r} p_Z(0, t, x, \xi) d\xi \leq Ct^{-r} \int_{\mathbb{R}^d} \frac{1}{t^{d/2}} e^{-c|x-\xi+y|^2/t} d\xi \leq Ct^{-r}.$$

The same way we have

$$\int_{\mathbb{R}^d} \nabla_y^{2r} p_V(0, t, x, \xi) d\xi \leq Ct^{-r}.$$

Applying Theorem 8.2.7 with Lemma 8.5.3 yields

$$\mathrm{d_{TV}}(Z_t^x, V_t^x) \leq C(\sqrt{t} + \Delta\sigma(x))^{2r/(2r+1)}.$$

□

8.3.5 Proof of Theorem 8.2.5

For the proof of Theorem 8.2.5, we do not use Lemma 8.3.7; instead we directly apply Theorem 8.2.7. Using Theorem 8.3.2, $\nabla_y^k p_X$ and $\nabla_y^k p_Y$ exist for $k = 0, 1, \dots, 2r$ and satisfy the same bounds as previously. Then using Theorem 8.2.7 with Lemma 8.5.3 we obtain

$$\mathrm{d_{TV}}(X_t^x, Y_t^x) \le C(\sqrt{t}(1 + \Delta b(x)) + \Delta\sigma(x) + t|b(x)|)^{2r/(2r+1)}.$$

8.3.6 Proof of Theorem 8.2.8

Proof. We use Lemma 8.3.7 again and rework the bound on $\mathrm{d_{TV}}(Z_t^x, V_t^x)$ by paying attention to the dependency of the constants in r in the proof of Theorem 8.2.7 with $Z_1 := Z_t^x$ and $Z_2 := V_t^x$. Since $\sigma_1 \in \mathcal{C}_b^{2r}$ for every $r \in \mathbb{N}$, we write (8.3.24) for any $r \in \mathbb{N}$ and we have

$$|\tilde{\beta}_r(t, \varepsilon)| \le \widetilde{C}_{2r} \|f\|_\infty t^{-r} \frac{\mathbb{E}[|\zeta|^{2r}]}{(2r)!}, \quad \widetilde{C}_{2r} := C_{2r} c_{2r}^{-d/2},$$

where C_{2r} and c_{2r} are defined in (8.2.12) and where

$$\mathbb{E}[|\zeta|^{2r}] = \frac{2^r \Gamma(d/2 + r)}{\Gamma(d/2)} = \prod_{i=0}^{r-1} (d + 2i).$$

Using (8.3.28) we get

$$\left| \sum_{i=1}^r \tilde{\beta}_r(t, \varepsilon/n_i) w_i n_i^{-r} \right| \le C\widetilde{C}_{2r} \|f\|_\infty t^{-r} \frac{\prod_{i=0}^{r-1}(d + 2i)}{(2r)!}$$

and we obtain as in (8.3.29):

$$\left| \mathbb{E} f(Z_t^x) - \sum_{i=1}^r w_i \mathbb{E} f_{\varepsilon/n_i}(Z_t^x) \right| \le \frac{1}{2} \kappa_1 \|f\|_\infty \varepsilon^r t^{-r}, \quad \kappa_1 := C\widetilde{C}_{2r} \frac{\prod_{i=0}^{r-1}(d + 2i)}{(2r)!}$$
$$\left| \mathbb{E} f(V_t^x) - \sum_{i=1}^r w_i \mathbb{E} f_{\varepsilon/n_i}(V_t^x) \right| \le \frac{1}{2} \kappa_1 \|f\|_\infty \varepsilon^r t^{-r}.$$

On the other hand, considering (8.3.30) and (8.3.31) with Lemma 8.5.3 with $\Delta\sigma(x) = 0$ we have

$$\left| \sum_{i=1}^r w_i \mathbb{E} f_{\varepsilon/n_i}(V_t^x) - \sum_{i=1}^r w_i \mathbb{E} f_{\varepsilon/n_i}(Z_t^x) \right| \le \kappa_2 \frac{\|f\|_\infty}{\sqrt{\varepsilon}} t, \quad \kappa_2 := C 2^r. \tag{8.3.33}$$

We now minimize $\kappa_1 \varepsilon^r t^{-r} + \kappa_2 \varepsilon^{-1/2} t$ in ε, giving

$$\varepsilon_\star = \frac{t^{(2r+2)/(2r+1)}}{(2r\kappa_1)^{2/(2r+1)}} \kappa_2^{2/(2r+1)}$$

and then

$$\kappa_1 \varepsilon_\star^r t^{-r} + \kappa_2 \varepsilon_\star^{-1/2} t \le C \kappa_2^{2r/(2r+1)} \kappa_1^{1/(2r+1)} t^{r/(2r+1)}$$

with as $r \to \infty$:

$$\kappa_2^{2r/(2r+1)} \kappa_1^{1/(2r+1)} \sim \widetilde{C}_{2r}^{1/(2r+1)} \left(\prod_{i=0}^{r-1} (d + 2i) \right)^{1/(2r+1)} \frac{1}{(2r)!^{1/(2r+1)}} 2^{2r^2/(2r+1)}$$

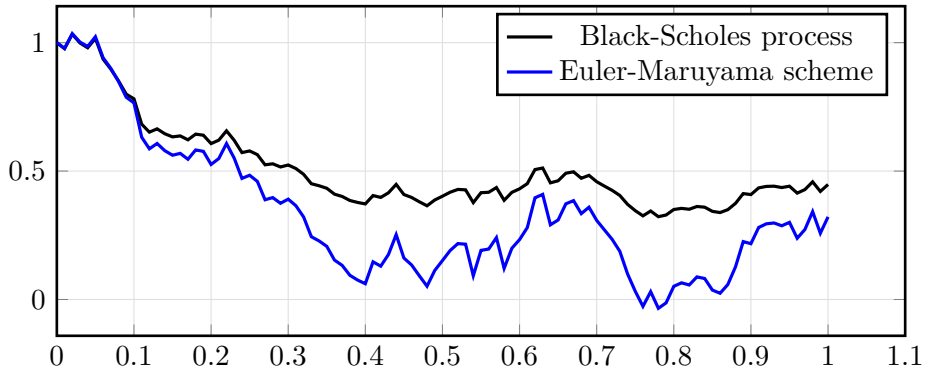


Figure 8.1: Example of trajectory of a Black-Scholes process (8.4.1) with $x = 1$ and $\sigma = 0.5$ along with its one-step Euler-Maruyama scheme (8.4.2).

with

$$\left(\prod_{i=0}^{r-1}(d+2i)\right)^{\frac{1}{(2r+1)}} = \exp\left(\frac{r}{2r+1}\frac{1}{r}\sum_{i=0}^{r-1}\log(d+2i)\right) \le \exp\left(\frac{r}{2r+1}\log(d+(r-1))\right) \le \sqrt{d+r-1},$$
$$\frac{1}{(2r)!^{1/(2r+1)}} \sim \frac{e}{2r}, \quad \limsup_{r\to\infty} \widetilde{C}_{2r}^{1/(2r+1)} < \infty$$

where we used Assumption (8.2.12), so that

$$\kappa_2^{2r/(2r+1)}\kappa_1^{1/(2r+1)} \le C\sqrt{d+r-1}\frac{e}{2r}2^r.$$

Then we have $\mathrm{d_{TV}}(Z_t^x, V_t^x) \le C2^r r^{-1/2} t^{r/(2r+1)}$ and we choose $r(t) = \lfloor \log^{1/2}(1/t) \rfloor$ so that as $t \to 0$,

$$\mathrm{d_{TV}}(Z_t^x, V_t^x) \le Ct^{1/2}\exp\left(C\sqrt{\log(1/t)}\right).$$

□

8.4 Counterexample

In this section we give a counter-example showing that we cannot achieve a bound better than $t^{1/2}$ in general. More specifically, we show that we cannot achieve a bound better than $t^{1/2}$ for the total variation between an SDE and its Euler-Maruyama-scheme in general. For $x > 0$ and $\sigma > 0$, let us consider the one-dimensional process

$$Y_t^x = xe^{\sigma W_t}, \tag{8.4.1}$$

where W is a standard Brownian motion. The process Y is solution of the SDE $dY_t^x = (\sigma^2/2)Y_t^x dt + \sigma Y_t^x dW_t$ and its associated Euler-Maruyama schemes reads

$$\bar{Y}_t^x = x + (\sigma^2/2)xt + \sigma x W_t \sim \mathcal{N}\left(x(1+t\sigma^2/2), \sigma^2 x^2 t\right). \tag{8.4.2}$$

An example of trajectory is given in Figure 8.1.

Proposition 8.4.1. *Let Y be the process defined in* (8.4.1)*. Then for small enough t we have*

$$\mathrm{d_{TV}}(Y_t^x, \bar{Y}_t^x) \ge C_x t^{1/2}. \tag{8.4.3}$$

Proof. We have

$$p_Y(t,x,y) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \frac{\exp\left(-\frac{1}{2\sigma^2 t}\log^2(y/x)\right)}{y} \mathbb{1}_{y\geq 0} \tag{8.4.4}$$

so that

$$\begin{aligned}
\mathrm{d_{TV}}(Y_t^x, \bar{Y}_t^x) &= \frac{1}{\sqrt{2\pi\sigma^2 t}} \int_{\mathbb{R}} \left| \exp\left(-\frac{\log^2(y/x)}{2\sigma^2 t}\right) y^{-1} \mathbb{1}_{y\geq 0} - \exp\left(-\frac{(y-x-xt\sigma^2/2)^2}{2\sigma^2 x^2 t}\right) x^{-1} \right| dy \\
&\geq \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-x/\sqrt{t}}^{\infty} \left| \frac{1}{x+\sqrt{t}y} \exp\left(-\frac{\log^2(1+\sqrt{t}y/x)}{2\sigma^2 t}\right) - \frac{1}{x} \exp\left(-\frac{(y-x\sqrt{t}\sigma^2/2)^2}{2\sigma^2 x^2}\right) \right| dy.
\end{aligned}$$

But we have as $(t,y) \to 0$:

$$\begin{aligned}
&\frac{1}{1+\sqrt{t}y/x} \exp\left(-\frac{\log^2(1+\sqrt{t}y/x)}{2\sigma^2 t}\right) - \exp\left(-\frac{(y-x\sqrt{t}\sigma^2/2)^2}{2\sigma^2 x^2}\right) \\
&\quad = (1-\sqrt{t}y/x + O(ty^2)) \exp\left(-\frac{1}{2\sigma^2 t}\left(\frac{ty^2}{x^2} - \frac{t^{3/2}y^3}{x^3} + O(t^2y^4)\right)\right) \\
&\qquad - \exp\left(-\frac{y^2}{2\sigma^2 x^2} - \frac{t\sigma^2}{8} + \frac{\sqrt{t}y}{2\sigma^2 x}\right) \\
&\quad = e^{-\frac{y^2}{2\sigma^2 x^2}} \left[(1-\sqrt{t}y/x + O(ty^2)) \left(1 + \frac{\sqrt{t}y^3}{2\sigma^2 x^3} + O(ty^4)\right) \right. \\
&\qquad \left. - \left(1 + \frac{\sqrt{t}y}{2\sigma^2 x} - \frac{t\sigma^2}{8} + O(t^2) + O(ty^2)\right) \right] \\
&\quad = e^{-\frac{y^2}{2\sigma^2 x^2}} \left[-\frac{\sqrt{t}y}{x} - \frac{\sqrt{t}y}{2\sigma^2 x} + \frac{\sqrt{t}y^3}{2\sigma^2 x^3} + \frac{t\sigma^2}{8} + O(ty^2) + O(t^2) \right].
\end{aligned}$$

Thus there exists $\epsilon > 0$ and t_0 such that for every $t \leq t_0$:

$$\mathrm{d_{TV}}(Y_t^x, \bar{Y}_t^x) \geq \frac{1}{\sqrt{2\pi\sigma^2 x^2}} e^{-\frac{\varepsilon^2}{2\sigma^2 x^2}} \frac{\sqrt{t}}{2} \int_{-\varepsilon}^{\varepsilon} \left| -\frac{y}{x} - \frac{y}{2\sigma^2 x} + \frac{y^3}{2\sigma^2 x^3} \right| dy,$$

so that $\mathrm{d_{TV}}(Y_t^x, \bar{Y}_t^x)$ is of order $t^{1/2}$ as $t \to 0$. □

However, the process Y does not satisfy the assumptions of Theorem 8.2.2 as its noise coefficient is not elliptic neither bounded on $(0,\infty)$. We then prove the following result.

Proposition 8.4.2. *There exists a diffusion process X on $\mathbb{R}$ with $\mathcal{C}^1$ and Lipschitz continuous drift, with $\mathcal{C}_b^\infty$ and elliptic diffusion coefficient σ and there exists $T > 0$ and $\varepsilon \in (0,1)$ such that*

$$\forall t \in [0,T],\ \forall x \in (\varepsilon, \varepsilon^{-1}), \quad \mathrm{d_{TV}}(X_t^x, \bar{X}_t^x) \geq C_x t^{1/2}$$

where $\bar{X}$ is the Euler-Maruyama scheme of X and where the positive constant C_x depends on x.

Proof. We construct from the geometric Brownian motion Y defined in (8.4.1), a process X with elliptic and bounded drift and such that $\mathrm{d_{TV}}(X_t^x, \bar{X}_t^x) \geq C_x t^{1/2}$. For $\varepsilon \in (0, 1/2)$, let us consider $\psi : \mathbb{R} \to \mathbb{R}^+$ a $\mathcal{C}_b^\infty$ approximation of

$$\widetilde{\psi} : x \in \mathbb{R} \longmapsto \begin{cases} x & \text{if } x \in [\varepsilon, \varepsilon^{-1}], \\ \varepsilon & \text{if } x \leq \varepsilon \\ \varepsilon^{-1} & \text{if } x \in [\varepsilon^{-1}, \infty) \end{cases}$$

such that $\psi = \widetilde{\psi}$ on $[2\varepsilon, \varepsilon^{-1}/2] \cup (-\infty, \varepsilon/2] \cup [2\varepsilon^{-1}, \infty)$. Then we define the process with elliptic and bounded noise coefficient

$$dX_t^x = -\frac{\sigma^2}{2} X_t^x dt + \sigma\psi(X_t^x) dW_t.$$

Then for $x \in (2\varepsilon, \varepsilon^{-1}/2)$ we have $\bar{X}_t^x = \bar{Y}_t^x$ and

$$\mathbb{P}(Y_t^x \neq X_t^x) \leq \mathbb{P}\left(\sup_{s\in[0,t]} Y_s^x \geq \varepsilon^{-1}/2\right) + \mathbb{P}\left(\inf_{s\in[0,t]} Y_s^x \leq 2\varepsilon\right).$$

With a proof similar to the proof of Lemma 8.3.3, we show that

$$\mathbb{P}\left(\sup_{s\in[0,t]} Y_s^x \geq \varepsilon^{-1}/2\right) \leq C_{x,\varepsilon} t.$$

Moreover, we remark that $(Y^x)^{-1} \sim x^{-2} Y^x$ in law so

$$\mathbb{P}\left(\inf_{s\in[0,t]} Y_s^x \leq 2\varepsilon\right) = \mathbb{P}\left(\sup_{s\in[0,t]} (Y_s^x)^{-1} \geq \varepsilon^{-1}/2\right) = \mathbb{P}\left(\sup_{s\in[0,t]} Y_s^x \geq x^2\varepsilon^{-1}/2\right) \leq C_{x,\varepsilon} t.$$

Then we obtain

$$\mathrm{d_{TV}}(X_t^x, \bar{X}_t^x) \geq \mathrm{d_{TV}}(Y_t^x, \bar{Y}_t^x) - \mathrm{d_{TV}}(X_t^x, Y_t^x) \geq C_x \sqrt{t}.$$

□

Remark 8.4.3. We could also consider the process X with "cut" bounded drift $\tilde{b}$ and get the same bounds, proving then that we cannot achieve better bounds in general than the ones established in Theorem 8.2.8 even if we assume that b is bounded.

8.5 Appendix

Lemma 8.5.1 ([Fri64], Chapter 9, Lemma 7)**.** *For $a > 0$, $0 < u < t \leq T$, $x \in \mathbb{R}^d$, $\xi \in \mathbb{R}^d$, let*

$$I_a := \int_{\mathbb{R}^d} \frac{1}{(u(t-u))^{d/2}} \exp\left(-a\left(\frac{|x-y|^2}{t-u} + \frac{|y-\xi|^2}{u}\right)\right) dy.$$

Then there exists a constant $C > 0$ depending only on d and T such that for every $0 < \varepsilon < 1$,

$$I_a \leq \frac{C}{(\varepsilon a t)^{d/2}} \exp\left(-a(1-\varepsilon)\frac{|x-\xi|^2}{t}\right).$$

Let us recall [PP23, Lemma 3.4(a)], with an immediate adaptation to the non-homogeneous case.

Lemma 8.5.2. *Let Z be solution to the generic SDE:*

$$Z_0^x = x \in \mathbb{R}^d, \quad dZ_t^x = b(t, Z_t^x)dt + \sigma(t, Z_t^x)dW_t, \ t \in [0,T],$$

where b and σ are Lipschitz continuous in (t,x) and where σ is bounded. Then for $p \geq 1$,

$$\forall t \in [0,T], \ \forall x \in \mathbb{R}^d, \quad \|Z_t^x - x\|_p \leq C(p, T, [b]_{\mathrm{Lip}}, [\sigma]_{\mathrm{Lip}}, \|\sigma\|_\infty)\left(t|b(0,x)| + t^{1/2}\right).$$

Lemma 8.5.3. *Let X and Y be the solution to the two general SDEs* (8.2.1) *and* (8.2.2)*. Assume that b and σ are Lipschitz continuous in (t,x) and that σ is bounded. Then for every $p \geq 1$,*

$$\forall t \in [0,T], \ \forall x \in \mathbb{R}^d, \ \|X_t^x - Y_t^x\|_p \leq C\left(t(1 + \Delta b(x)) + t^{3/2}(|b_1| + |b_2|)(0,x) + \Delta\sigma(x) t^{1/2}\right) \tag{8.5.1}$$

Proof. We first deal with the case $p \geq 2$. We have

$$\begin{aligned}\|X_t^x - Y_t^x\|_p &\leq \left\|\int_0^t (b_1(s, X_s^x) - b_2(s, Y_s^x))ds\right\|_p + \left\|\int_0^t (\sigma_1(s, X_s^x) - \sigma_2(s, Y_s^x))dW_s\right\|_p \\ &\leq \left\|\int_0^t (b_1(s, X_s^x) - b_1(0, x))ds\right\|_p + t\Delta b(x) + \left\|\int_0^t (b_2(s, Y_s^x) - b_2(0, x))ds\right\|_p \\ &\quad + \left\|\int_0^t (\sigma_1(s, X_s^x) - \sigma_1(0, x))dW_s\right\|_p + \left\|\int_0^t \Delta\sigma(x)dW_s\right\|_p + \left\|\int_0^t (\sigma_2(s, Y_s^x) - \sigma_2(0, x))dW_s\right\|_p\end{aligned}$$

But using the Burkholder-Davis-Gundy and the generalized Minkowski inequalities, we have

$$\begin{aligned}\left\|\int_0^t (\sigma_1(s, X_s^x) - \sigma_1(0, x))dW_s\right\|_p &\leq C_p^{\mathrm{BDG}}[\sigma_1]_{\mathrm{Lip}} \left\|\int_0^t |(s, X_s^x) - (0, x)|^2 ds\right\|_{p/2}^{1/2} \\ &\leq C_p^{\mathrm{BDG}}[\sigma_1]_{\mathrm{Lip}} \left(\int_0^t \|(s, X_s^x) - (0, x)\|_p^2 ds\right)^{1/2} \leq C(t + t^{3/2}|b_1(0, x)|),\end{aligned}$$

where C_p^{BDG} is a constant which only depends on p and where we used Lemma 8.5.2. So that

$$\begin{aligned}\|X_t^x - Y_t^x\|_p &\leq [b_1]_{\mathrm{Lip}} \int_0^t \|(s, X_s^x) - (0, x)\|_p ds + [b_2]_{\mathrm{Lip}} \int_0^t \|(s, Y_s^x) - (0, x)\|_p + t\Delta b(x) \\ &\quad + C(t + t^{3/2}(|b_1| + |b_2|)(0, x)) + \Delta\sigma(x)\sqrt{t}\|W_1\|_p \\ &\leq C\left(t(\Delta b(x) + 1) + t^{3/2}(|b_1| + |b_2|)(0, x) + \Delta\sigma(x)\sqrt{t}\right)\end{aligned}$$

which completes the proof for $p \geq 2$. For $p \in [1, 2)$, the inequality is still true remarking that $\|\cdot\|_p \leq \|\cdot\|_2$. □

Chapter 9

Weak error rates for numerical schemes of non-singular Stochastic Volterra equations with application to option pricing under path-dependent volatility

The results presented in this chapter are a joint work with Masaaki Fukasawa. They have been submitted to *SIAM Journal on Financial Mathematics* (*SIFIN*) and are currently in revision for possible publication.

Abstract

We study the weak error rate for the Euler-Maruyama scheme for Stochastic Volterra equations (SVE) with application to pricing under stochastic volatility models. SVEs are non-Markovian stochastic differential equations with memory kernel. We assume in particular that the kernel is non-singular and $\mathcal{C}^4$. We show that the weak error rate is of order $O(1/N)$ where N is the number of steps of the Euler-Maruyama scheme, thus giving the same weak error rate as for SDEs. Our proof consists in adapting the classic weak error proof for Markov processes to SVEs; to this end we rely on infinite dimensional functionals and on their derivatives. Our work opens the way to study rough SVEs, which shall be investigated in a next paper.

Keywords– Stochastic Volterra equation, Rough volatility, Euler-Maruyama scheme, Weak error rate.

9.1 Introduction

Stochastic Volterra equations (SVE) have recently attracted much attention in the mathematical finance community in the context of rough volatility modelling, which is more able to reproduce some features of asset prices [ALV07, GJR18, EEFR18, JR20, Fuk17, Fuk21]. SVEs have also

been introduced with regular (non-singular) kernel for modelling in population dynamics, biology and physics [Moh98], in order to generalize modelling to non-Markovian stochastic systems with some memory effect. They where also motivated in particular by the physics of heat transfer [GLS90] and have been mathematically studied since [BM80, Pro85]. Applications, e.g. pricing options in financial practice, require numerical methods to simulate the solution of the SVE, such as simulation through Euler-Maruyama schemes.

In the present paper, we give bounds for the weak error of the Euler-Maruyama scheme with N steps of an SVE on a finite time interval $[0, T]$ in the case where the kernel is non-singular. We consider two different Euler schemes: one where the kernel is not discretized, thus requiring the simulation of a large Gaussian matrix with covariance, and another one where the kernel is discretized, thus requiring only the simulation of (independent) Brownian increments.

A first bound on the weak error can be obtained from bounds on the strong error, however this is sub-optimal in general. For example, for Stochastic Differential Equations (SDE) the strong error is of order $O(1/\sqrt{N})$ but the weak error is of order $O(1/N)$. Such bounds get even worse in the case of SVE with fractional kernel, giving a weak error bounded by the strong error which is order $O(N^{-H})$, where $H \in (0, 1/2)$ is the Hurst parameter of the fractional kernel and is small ($H \simeq 0.1$) in many financial applications. In [RTY21] are given bounds for the weak error for the multi-level Euler-Maruyama scheme, however the authors only assume that the weak error is bounded by the strong error (see [RTY21, Section 2.3]). In [FU21] is shown that the strong error is exactly of order H and the authors give the expression of the limit law of the (rescaled) strong error. In [Gas23] are given weak error rates from some rough volatility models and are proved to be of order $(3H + 1/2) \wedge 1$, yielding significantly better bounds in the case where H is close to 0. However the results are valid only for some special cases (semilinear or cubic test function), hinting that obtaining general results for fractional processes is difficult. More recently, [FSW22] also proved that the weak error rate is of order $(3H + 1/2) \wedge 1$ for some class of stochastic rough volatility models, including the rough Bergomi and the rough Stein-Stein models. We recall that a rough (resp. Volterra) stochastic volatility model is a special case of a singular (resp. non-singular) two-dimensional SVE, where the first process is an asset price satisfying $dS_t = S_t\sqrt{V_t}dB_t$ for some Brownian motion B and where the second process (V_t) is the volatility satisfying some rough (resp. Volterra) stochastic equation, then giving for the joint process a matrix kernel K being diagonal and constant on its first coordinate.

Our main result states that for SVEs with non-singular kernel, the weak error of the Euler-Maruyama scheme is of order $O(1/N)$, which is the same rate as in the classic SDE case, under regularity assumptions on the coefficients and the kernel: we assume that the kernel is defined on the whole interval $[0, T]$ and is $\mathcal{C}^2$, that the drift and the diffusion coefficients are $\mathcal{C}^4$, bounded with bounded derivatives. Our strategy of proof consists in adapting the domino method from [TT90] to the SVE case. In the SDE case and for a function $f : \mathbb{R}^d \to \mathbb{R}$, the domino strategy consists in a step-by-step decomposition of the weak error to produce an upper bound as follows:

$$\begin{aligned} |\mathbb{E}f(\bar{X}_T^x) - \mathbb{E}f(X_T^x)| &= |\bar{P}_h \circ \cdots \circ \bar{P}_h f(x) - P_{t_N} f(x)| \\ &\leq \sum_{k=1}^{N} \left| \bar{P}_h \circ \cdots \circ \bar{P}_h \circ (\bar{P}_h - P_h) \circ P_{t_N - t_k} f(x) \right|, \end{aligned} \tag{9.1.1}$$

where $h = T/N$, $t_k = kh$, X is the solution to the SDE and $\bar{X}$ is the corresponding Euler-Maruyama scheme, P and $\bar{P}$ are the semi-group operators associated to X and $\bar{X}$ respectively. Then showing that with $g := P_{t_n - t_k} f$, the short term weak error $(\bar{P}_h - P_h)g(x)$ is of order $O(1/N^2)$, the sum in (9.1.1) is then of order $O(1/N)$. An elementary proof in the SDE case using the domino method can be found in [Pag18, Section 7.6]. This strategy fundamentally relies on a Markov semi-group and thus cannot be directly applied to the SVE case, as the

future trajectory of X depends on the whole previous trajectory in this last case. Instead, we consider the solution of an SVE as a Markov process on the infinite dimensional state space of trajectories Ω and then we define a infinite-dimensional Markov semi-group, allowing us to use the previously introduced domino method.

We then show that the weak error in small time is of order $O(1/N^2)$ by establishing a Itō type formula for functionals $g : \Omega \to \mathbb{R}$, involving the Fréchet derivatives of g. Such approach involving the derivatives of functionals on infinite dimensional state space and establishing Itō formula for SVEs was developed in [Dup19] and [VZ19], however in our case we only consider the infinite dimensional state space of continuous paths instead of càdlàg paths, since we consider a different semi-group.

Using a hybrid approach, combining ideas from both finite and infinite dimensional settings, the Itō formula with a finite dimensional Brownian motion on the one side, and the Fréchet derivatives of path functionals and the Markov property on an infinite dimensional state space on the other side, we establish the weak convergence rate of the Euler-Maruyama for SVEs. Studying the non-singular case by adapting the classic domino method to path-dependent setting constitutes a first step for studying weak error rates for fractional SVEs as we believe our method can be adapted to the singular case. We shall conduct such study in a next paper.

We give numerical evidence of the convergence rate we obtained on a Monte Carlo option pricing problem with a stochastic volatility model where the volatility follows some non-singular SVE. Proving weak error rates allows to design weighted and unweighted multi-level Richardson-Romberg extrapolation estimators [Gil08, GS14, LP17] that exploit the faster convergence of the weak error in comparison with the strong error; such application is particularly critical for the Monte Carlo simulation of Volterra and rough stochastic equations where the Vanilla Euler-Maruyama scheme has large time complexity (N^2).

The article is organized as follows. In Section 9.2 we give the precise setting and assumptions of the problem we consider, in particular the regularity assumptions on the coefficients and on the kernel of the SVE, and we state our main theorem. In Section 9.3, we give general results on random paths $(\varphi_u^t)_{u\geq 0, t\in[0,T]}$ which are adapted with respect to t, in particular we establish an Itō formula for infinite dimensional functionals $g : \Omega \to \mathbb{R}$. The proof of the theorem is given in Section 9.4. Considering (9.1.1), the proof of our main result is decomposed into three parts: we prove that the short term error is of order $O(1/N^2)$, applying Itō formula for a regular functional $g : \Omega \to \mathbb{R}$ as in the classic proof in the SDE case. Secondly we show that with g being the concatenation of discrete kernels applied to f, then the functional g is indeed differentiable in the Fréchet meaning and with bounded Fréchet derivatives. Lastly, in Section 9.5, we empirically check the weak convergence rate for some SVE model with non-singular kernel.

9.2 Setting and main results

9.2.1 Setting

Let us consider the following SVE in $\mathbb{R}^d$, $d \in \mathbb{N}$:

$$X_t = X_0 + \int_0^t K_1(t,s)b(X_s)ds + \int_0^t K_2(t,s)\sigma(X_s)dW_s, \quad t \in [0,T], \tag{9.2.1}$$

where (W_t) is a standard Brownian motion in $\mathbb{R}^{q_3}$ defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where

$$b : \mathbb{R}^d \to \mathbb{R}^{q_1}, \quad K_1 : [0,T]^2 \to \mathcal{M}_{d,q_1}(\mathbb{R}), \quad \sigma : \mathbb{R}^d \to \mathcal{M}_{q_2,q_3}(\mathbb{R}), \quad K_2 : [0,T]^2 \to \mathcal{M}_{d,q_2}(\mathbb{R}),$$

and where $q_1, q_2, q_3 \in \mathbb{N}$ and for $a, b \in \mathbb{N}$, $\mathcal{M}_{a,b}(\mathbb{R})$ denotes the set of $a \times b$ matrices with coefficients in $\mathbb{R}$. We denote by $(\mathcal{F}_t)_{t\in[0,T]}$ the filtration generated by the Brownian motion.

The starting point $X_0 \in \mathbb{R}^d$ is fixed. Let us make the following assumptions on the coefficients b and σ and on the kernels K_1 and K_2.

For $(\mathcal{A}, d_{\mathcal{A}})$ and $(\mathcal{B}, d_{\mathcal{B}})$ two metric spaces and $k \in \mathbb{N}$, we consider the following sets of functions from $\mathcal{A}$ to $\mathcal{B}$:

- $\mathcal{C}^k(\mathcal{A}, \mathcal{B})$: functions that are k times differentiable with continuous derivatives,
- $\mathcal{C}_b^k(\mathcal{A}, \mathcal{B})$: functions that are bounded, k times differentiable with continuous and bounded derivatives,
- $\widetilde{\mathcal{C}}_b^k(\mathcal{A}, \mathcal{B})$: functions that are k times differentiable with continuous and bounded derivatives.

When there is no ambiguity on the spaces, we also use the notations $\mathcal{C}^k$, $\mathcal{C}_b^k$ and $\widetilde{\mathcal{C}}_b^k$ respectively.

Assumption 9.2.1. *(i) $K_1 \in \mathcal{C}^2([0,T]^2, \mathcal{M}_{d,q_1}(\mathbb{R}))$ and $K_2 \in \mathcal{C}^4([0,T]^2, \mathcal{M}_{d,q_2}(\mathbb{R}))$, which guarantees that K_1 (resp. K_2) is bounded with bounded derivatives up to order 2 (resp. 4).*

(ii) $b \in \mathcal{C}_b^5(\mathbb{R}^d, \mathbb{R}^{q_1})$ and $\sigma \in \mathcal{C}_b^5(\mathbb{R}^d, \mathcal{M}_{q_3,q_2}(\mathbb{R}))$.

Then Assumption 9.2.1 guarantees that the solution of (9.2.1) is well defined (see for example [AJ18, Lemma 5.29]).

To simplify the notations and for more readability of the proofs, we assume hereafter that all the objects considered are one-dimensional, i.e. that $d = q_1 = q_2 = q_3 = 1$. However the main results in Section 9.2.2 remain valid for any (finite) dimensions d, q_1, q_2 and q_3, re-writing the proofs by replacing the the one-dimensional products by matrix products and writing them as sums over indices.

Let us define the Euler-Maruyama scheme associated to (9.2.1). For $N \in \mathbb{N}$, we define the time step and the regular subdivision

$$h := T/N, \quad t_k := kT/N, \quad k \in \{0, \dots, N\} \tag{9.2.2}$$

and

$$\bar{X}_t = X_0 + \int_0^t K_1(t,s) b(\bar{X}_{\underline{s}}) ds + \int_0^t K_2(t,s)\sigma(\bar{X}_{\underline{s}}) dW_s, \quad t \in [0,T], \tag{9.2.3}$$

where for $s \in [0,T]$, we define

$$\underline{s} = \lfloor s/h \rfloor h.$$

The solution of (9.2.3) can be recursively simulated as

$$\bar{X}_t = X_0 + \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}\wedge t} K_1(t,s) b(\bar{X}_{t_j}) ds + \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}\wedge t} K_2(t,s)\sigma(\bar{X}_{t_j}) dW_s, \quad t \in [t_k, t_{k+1}],$$

where the integrals $(\int_{t_j}^{t_{j+1}} K_2(t,s)dW_s)_j$ can be simulated on the discrete grid $(t_k)_{0\le k\le N}$ by generating the independent sequence of Gaussian vectors

$$\Big(\int_{t_j}^{t_{j+1}} K_2(t_k, s) dW_s \Big)_{k=j,\dots,N}, \quad j = 0, \dots, N-1,$$

using the Cholesky decomposition of the covariance matrix

$$\Big(\int_{t_j}^{t_{j+1}} K_2(t_{k_1}, s) K_2(t_{k_2}, s) ds \Big)_{k_1,k_2=j,\dots,N}.$$

We also define the Euler-Maruyama scheme associated to (9.2.1) with discretization of the kernels as

$$\overrightarrow{X}_t = X_0 + \int_0^t K_1(t,\underline{s})b(\overrightarrow{X}_{\underline{s}})ds + \int_0^t K_2(t,\underline{s})\sigma(\overrightarrow{X}_{\underline{s}})dW_s, \quad t \in [0,T]. \tag{9.2.4}$$

This scheme more convenient to simulate as it only requires the simulation of the Brownian increments $(W_{t_{k+1}} - W_{t_k})_{0\leq k\leq N-1}$.

With no ambiguity, in the proofs we shall use the notation $\bar{\psi}$ for ψ some function defined on $\mathbb{R}^d$, such that for every process Y and every $s \in [0,T]$ we have $\bar{\psi}(Y_s) = \psi(Y_{\underline{s}})$.

We extend the definition of K_1 and K_2 on $\mathbb{R}^+ \times \mathbb{R}^+$ such that for $i = 1, 2$, $K_i(t,s) = 0$ for $(t,s) \notin [0,2T] \times [0,2T]$ and such that K_i is still bounded with bounded derivatives up to order 2.

In this paper, we use the notation C to denote a positive real constant, which may change from line to line. The constant C depends on the parameters of the problem: the coefficients and the kernels of the SVE, the time horizon T.

9.2.2 Main results

Theorem 9.2.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\widetilde{\mathcal{C}}_b^5$ and assume Assumption 9.2.1. Then we have*

$$\mathbb{E}[f(\bar{X}_T)] - \mathbb{E}[f(X_T)] = O\left(1/N\right), \tag{9.2.5}$$

$$\mathbb{E}[f(\overrightarrow{X}_T)] - \mathbb{E}[f(X_T)] = O\left(1/N\right). \tag{9.2.6}$$

Remark 9.2.3.

- Since X, $\bar{X}$ and $\overrightarrow{X}$ satisfy (9.2.1), (9.2.3) and (9.2.4) respectively with K_1, K_2, b and σ being bounded, we obtain

$$\mathbb{E}|X_T| + \mathbb{E}|\bar{X}_T| + \mathbb{E}|\overrightarrow{X}_T| < +\infty.$$

  Since f is Lipschitz-continuous, we get

$$\mathbb{E}|f(X_T)| + \mathbb{E}|f(\bar{X}_T)| + \mathbb{E}|f(\overrightarrow{X}_T)| < +\infty.$$

- The strategy of proof we develop in Section 9.4 does not allow us to give weak error bounds for path-dependent functionals.
- We prove that the weak order of convergence of the Euler-Maruyama scheme for SVEs with regular kernels is the same as for SDEs, however the computation time for this scheme for SVEs is of order N^2, against order N for SDEs.

9.3 Preliminary results on infinite dimensional paths

9.3.1 State space and path derivatives

For $T' \in \mathbb{R}^+$, we consider the infinite dimensional state space $\Omega_{T'}$ being the space of $\mathbb{R}$-valued continuous trajectories on $[0,\infty)$ with support included in $[0,T']$, with the topology of the supremum norm. For $\omega \in \Omega_{T'}$ such that ω is $\mathcal{C}^1$, we denote $\dot{\omega}$ its derivative. If ω is Lipschitz-continuous, we denote $[\omega]_{\text{Lip}}$ its Lipschitz constant.

For $g : \Omega_{T'} \to \mathbb{R}$ and for $\omega \in \Omega_{T'}$, we define, when it exists, $\nabla g(\omega)$ as the Fréchet derivative of g with respect to ω, which is a linear operator on $\Omega_{T'}$:

$$g(\omega + \eta) = g(\omega) + \langle \nabla g(\omega), \eta \rangle + o(\|\eta\|_\infty), \quad \eta \in \Omega_{T'}.$$

More generally, for $\ell \in \mathbb{N}$ we define, when it exists, the derivative of g of order ℓ recursively as the ℓ-multilinear operator on $\Omega_{T'}^{\otimes \ell}$:

$$\langle \nabla^{\ell-1} g(\omega+\eta^1), \bigotimes_{j=2}^{\ell} \eta^j \rangle = \langle \nabla^{\ell-1} g(\omega), \bigotimes_{j=2}^{\ell} \eta^j \rangle + \langle \nabla^{\ell} g(\omega), \bigotimes_{j=1}^{\ell} \eta^j \rangle + o(\|\eta^1\|_\infty),\ \eta^i \in \Omega_{T'},\ i = 1, \dots, \ell.$$

We use the notation $\otimes$ only to enhance the multilinearity of $\nabla^\ell g$.

Remark 9.3.1. The path derivative can be made explicit in some simple cases:

- If $g(\omega) = \widetilde{g}(\omega_{u_0})$ for some fixed $u_0 \in \mathbb{R}^+$ and for some $\widetilde{g} : \mathbb{R} \to \mathbb{R}$, then we have

$$\langle \nabla^{\ell} g(\omega), \bigotimes_{j=1}^{\ell} \eta^j \rangle = \nabla^{\ell} \widetilde{g}(\omega_{u_0}) \cdot \bigotimes_{j=1}^{\ell} \eta^j_{u_0}. \tag{9.3.1}$$

- If $g(\omega) = \int_0^{T'} \widetilde{g}(\omega_u) du$ for some $\widetilde{g} : \mathbb{R} \to \mathbb{R}$, then we have

$$\langle \nabla^{\ell} g(\omega), \bigotimes_{j=1}^{\ell} \eta^j \rangle = \int_0^{T'} \nabla^{\ell} \widetilde{g}(\omega_u) \cdot \bigotimes_{j=1}^{\ell} \eta^j_u du. \tag{9.3.2}$$

9.3.2 Expectation of the supremum of a random path process

Lemma 9.3.2. *Let $F(u,s)_{u \geq 0, s \in [0,T]}$ a $\mathbb{R}$-valued random process adapted to the filtration $\mathcal{F}$ with respect to its second variable, such that for every $s \in [0,T]$ and $u > T'$, $F(u,s) = 0$, and such that $\|F\|_\infty \leq C_1$ almost surely, $\partial_1 F$ exists and $\|\partial_1 F\|_\infty \leq C_2$ almost surely, for some $C_1, C_2 \in \mathbb{R}^+$. and let $(M_s)_{0 \leq s \leq T}$ be a $\mathbb{R}$-valued martingale adapted to $\mathcal{F}$ with $\mathbb{E}\langle M \rangle_T < \infty$. For $r \in [0,T]$ let us define*

$$\varphi_u := \int_0^r F(u,s) dM_s, \quad u \geq 0.$$

Then there exists a continuous modification $\widetilde{\varphi}$ of φ such that

$$\mathbb{E} \sup_{u \geq 0} |\widetilde{\varphi}_u|^2 \leq C(T')^2 \|\partial_1 F\|_\infty^2 \mathbb{E}\langle M \rangle_r. \tag{9.3.3}$$

Proof. For $u_1, u_2 \in [0, T']$ we have

$$\begin{aligned} \mathbb{E}|\varphi_{u_1} - \varphi_{u_2}|^2 &= \mathbb{E}\Big| \int_0^r (F(u_1,s) - F(u_2,s)) dM_s \Big|^2 = \mathbb{E} \int_0^r |F(u_1,s) - F(u_2,s)|^2 d\langle M \rangle_s \\ &\leq |u_1 - u_2|^2 \|\partial_1 F\|_\infty^2 \mathbb{E}\langle M \rangle_r, \end{aligned}$$

so that using the Kolmogorov continuity theorem (9.6.1), there exists a modification $\widetilde{\varphi}$ of φ which is almost surely α-Hölder for every $\alpha \in (0, 1/2)$, and taking for example $\alpha = 1/4$ we have

$$\mathbb{E}\Big[\Big(\sup_{u_1, u_2 \in [0,T'], u_1 \neq u_2} \frac{|\widetilde{\varphi}_{u_1} - \widetilde{\varphi}_{u_2}|}{|u_1 - u_2|^{1/4}} \Big)^2 \Big] \leq C(T')^{3/2} \|\partial_1 F\|_\infty^2 \mathbb{E}\langle M \rangle_r,$$

where C is an universal constant, so that taking $u_1 = u$ and $u_2 = 0$ with $\varphi_0 = 0$ we obtain

$$\mathbb{E} \sup_{u \geq 0} |\widetilde{\varphi}_u|^2 = \mathbb{E} \sup_{u \in [0,T']} |\widetilde{\varphi}_u|^2 \leq C(T')^2 \|\partial_1 F\|_\infty^2 \mathbb{E}\langle M \rangle_r.$$

□

Remark 9.3.3. In the following, we will use Lemma 9.3.2 for families of trajectories $(\varphi^t)_{t\in[0,T]}$ of the form

$$\varphi_u^t = \int_0^t F(u,s)dM_s \quad \text{or} \quad \varphi_u^t = \int_0^t F(t+u,s)dM_s.$$

When the assumptions of Lemma 9.3.2 are checked, the bound (9.3.3) is true up to some modification of φ, i.e. for a family of trajectories $(\widetilde{\varphi}_u^t)$ such that

$$\forall t \in [0,T],\ \forall u \geq 0,\ \mathbb{P}(\varphi_u^t = \widetilde{\varphi}_u^t) = 1 \quad \text{and} \quad \forall t \in [0,T],\ u \mapsto \widetilde{\varphi}_u^t \text{ is continuous.}$$

Without loss of generality, each time we use Lemma 9.3.2 we do not mention explicitly the modification.

9.3.3 A general Itō formula for path-dependent functionals

In this section we prove an extension of the classic Itō formula to processes of the form $G(\varphi^t)$, where $G : \Omega_{T'} \to \mathbb{R}$ and where for every $t \in [0,T]$, φ^t is some $\mathcal{F}_t$-measurable random path.

Theorem 9.3.4. *Let us consider the family of random paths $(\varphi_u^t)_{t\in[0,T],\ u\geq 0}$ such that*

$$\varphi_u^t = \varphi_u^0 + \int_0^t Z_1(u,s)ds + \int_0^t Z_2(u,s)dW_s, \tag{9.3.4}$$

where for every $u \geq 0$, $s \mapsto Z_i(u,s)$, $i = 1,2$, is an adapted $\mathbb{R}$-valued semi-martingale such that $\partial_1 Z_i$ and $\partial_{11}^2 Z_2$ exist almost surely with

$$\mathbb{E}\|Z_1\|_\infty^2 + \mathbb{E}\|\partial_1 Z_1\|_\infty \leq C, \quad \|Z_2\|_\infty + \|\partial_1 Z_2\|_\infty + \|\partial_{11}^2 Z_2\|_\infty \leq C \text{ almost surely} \tag{9.3.5}$$

and such that $Z_i(u,s) = 0$ for $u > T'$, $T' \in \mathbb{R}^+$. We also assume that $\varphi^0 \in \Omega_{T'} \cap \mathcal{C}^1$ and is Lipschitz-continuous. Moreover, let $G : \Omega_{T'} \to \mathbb{R}$ with bounded pathwise derivatives up to order 3. Then we have almost surely

$$\begin{aligned} G(\varphi^t) = G(\varphi^0) &+ \int_0^t \langle \nabla G(\varphi^s), Z_1(\cdot,s)\rangle ds + \int_0^t \langle \nabla G(\varphi^s), Z_2(\cdot,s)\rangle dW_s \\ &+ \frac{1}{2}\int_0^t \langle \nabla^2 G(\varphi^s), Z_2(\cdot,s)^{\otimes 2}\rangle ds. \end{aligned} \tag{9.3.6}$$

Remark 9.3.5. We highlight the fact that in (9.3.4), the values of $Z_i(u,s)$ cannot depend on t. For example, if we consider the SVE

$$\varphi_u^t = \varphi_u^0 + \int_0^t K_2(t+u,s)\sigma(A_s)dW_s$$

for some adapted semi-martingale A, then we need to write φ_u^t as

$$\varphi_u^t = \varphi_u^0 + \int_0^t K_2(s+u,s)\sigma(A_s)dW_s + \int_0^t \Big(\int_0^s \partial_1 K_2(s+u,v)\sigma(A_v)dW_v\Big)ds.$$

Proof. We first remark that if G only depends on a finite number of times $u_1, \dots, u_n \in \mathbb{R}^+$, i.e. if we have

$$\forall \omega \in \Omega_{T'},\ G(\omega) = \widetilde{G}(\omega_{u_1}, \dots, \omega_{u_n}), \quad \widetilde{G} \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R}),$$

then we have

$$\forall \eta \in \Omega,\ \langle \nabla G(\omega), \eta\rangle = \sum_{i=1}^n \partial_i \widetilde{G}(\omega_{u_1}, \dots, \omega_{u_n})\eta_{u_i},$$

$$\forall \eta^1, \eta^2 \in \Omega,\ \langle \nabla^2 G(\omega), \eta^1 \otimes \eta^2 \rangle = \sum_{1 \le i,j \le n} \partial_{ij} \widetilde{G}(\omega_{u_1}, \dots, \omega_{u_n}) \eta^1_{u_i} \eta^2_{u_j},$$

and then (9.3.6) directly comes from the classic Itō formula.

Then, let us define for $n \in \mathbb{N}$ the regular subdivision of $[0, T']$: $u_i^n = iT'/n$, $0 \le i \le n$ and for $\omega \in \Omega_{T'}$ we define ω^n as the affine interpolation of ω on the subdivision $(u_i^n)_i$ i.e. ω^n is equal to the affine interpolation on $[0, T']$ and then ω^n and ω are both equal to 0 on $[T', \infty)$; we also define G^n as for every $\omega \in \Omega_{T'}$, $G^n(\omega) = G(\omega^n)$. Then for every $\omega \in \Omega_{2T}$ we have $G^n(\omega) = \widetilde{G}^n(\omega_{u_1}, \dots, \omega_{u_n})$ where $\widetilde{G}^n : \mathbb{R}^n \to \mathbb{R}$ is the composition of the affine interpolation $\mathscr{L}^n : \mathbb{R}^n \to \Omega_{2T}$, which is a bounded linear operator, and of G, so is $\mathcal{C}^2$ and then (9.3.6) is true for G^n. Moreover for every $\omega, \eta \in \Omega_{T'}$ such that ω is Lipschitz-continuous we have

$$|G(\omega) - G^n(\omega)| = |G(\omega) - G(\omega^n)| \le \|\nabla G\|_\infty \|\omega - \omega^n\|_\infty \le C \|\nabla G\|_\infty [\omega]_{\mathrm{Lip}} / n.$$

Moreover, remarking that the affine interpolation is a bounded linear operator, we get that G^n is also differentiable with

$$\langle \nabla G^n(\omega), \eta \rangle = \langle \nabla G(\omega^n), \eta^n \rangle$$

so that

$$\big|\langle \nabla G(\omega), \eta \rangle - \langle \nabla G^n(\omega), \eta \rangle\big| \le \|\nabla^2 G\|_\infty \|\omega - \omega^n\|_\infty \|\eta\|_\infty + \|\nabla G\|_\infty \|\omega\|_\infty \|\eta - \eta^n\|_\infty. \quad (9.3.7)$$

Moreover for $\eta^1, \eta^2 \in \Omega_{T'}$ we have

$$\langle \nabla^2 G^n(\omega), \eta^1 \otimes \eta^2 \rangle = \langle \nabla^2 G(\omega^n), (\eta^1)^n \otimes (\eta^2)^n \rangle$$

so that

$$\begin{aligned} \big|\langle \nabla^2 G(\omega), \eta^1 \otimes \eta^2 \rangle - \langle \nabla^2 G^n(\omega), \eta^1 \otimes \eta^2 \rangle\big| &\le \|\nabla^3 G\|_\infty \|\omega - \omega^n\|_\infty \|\eta^1\|_\infty \|\eta^2\|_\infty \\ &+ \|\nabla^2 G\|_\infty (\|\omega\|_\infty \|\eta^1 - (\eta^1)^n\|_\infty \|\eta^2\|_\infty + \|\omega\|_\infty \|\eta^1\|_\infty \|\eta^2 - (\eta^2)^n\|_\infty) \end{aligned} \quad (9.3.8)$$

Writing (9.3.6) with G^n gives

$$\begin{aligned} G^n(\varphi^t) = G^n(\varphi^0) &+ \int_0^t \langle \nabla G((\varphi^s)^n), Z_1^n(\cdot, s) \rangle ds + \int_0^t \langle \nabla G((\varphi^s)^n), Z_2^n(\cdot, s) \rangle dW_s \\ &+ \frac{1}{2} \int_0^t \langle \nabla^2 G((\varphi^s)^n), Z_2^n(\cdot, s)^{\otimes 2} \rangle ds, \end{aligned} \quad (9.3.9)$$

and we have

$$\mathbb{E}|G^n(\varphi^t) - G(\varphi^t)|^2 \le 2\|G\|_\infty \mathbb{E}|G^n(\varphi^t) - G(\varphi^t)|^2 \le C \|G\|_\infty \|\nabla G\|_\infty \mathbb{E}[\varphi^t]_{\mathrm{Lip}} n^{-1}$$

with φ^t being $\mathcal{C}^1$ with

$$\dot{\varphi}^t_u = \dot{\varphi}^t_0 + \int_0^t \partial_1 Z_1(u, s) ds + \int_0^t \partial_1 Z_2(u, s) dW_s,$$

where the interchange is ensured by the stochastic Fubini theorem. But following Lemma 9.3.2 with assumption (9.3.5), $\mathbb{E}[\varphi^t]_{\mathrm{Lip}} < \infty$, so that $G^n(\varphi^t)$ converges to $G(\varphi^t)$ in L^2. We proceed the same way for $G^n(\varphi^0)$.

Moreover, we have

$$\mathbb{E}\Big| \int_0^t \langle \nabla G((\varphi^s)^n), Z_1^n(\cdot, s) \rangle ds - \int_0^t \langle \nabla G(\varphi^s), Z_1(\cdot, s) \rangle ds \Big|$$

$$\leq \int_0^t \mathbb{E}|\langle \nabla G(\varphi^s), (Z_1^n(\cdot,s) - Z_1(\cdot,s))\rangle| ds + \int_0^t \mathbb{E}|\langle \nabla G((\varphi^s)^n) - \nabla G(\varphi^s), Z_1^n(\cdot,s))\rangle| ds$$
$$\leq C\|\nabla G\|_\infty \mathbb{E}\|\partial_1 Z_1\|_\infty n^{-1} + C\|\nabla^2 G\|_\infty \int_0^t \mathbb{E}\big[[\varphi^s]_{\mathrm{Lip}}\|Z_1\|_\infty\big] ds\ n^{-1}$$
$$\leq C\|\nabla G\|_\infty \mathbb{E}\|\partial_1 Z_1\|_\infty n^{-1} + C\|\nabla^2 G\|_\infty \int_0^t \big(\mathbb{E}[\varphi^s]^2_{\mathrm{Lip}}\mathbb{E}\|Z_1\|^2_\infty\big)^{1/2} ds\ n^{-1} \underset{n\to\infty}{\longrightarrow} 0,$$

where we used (9.3.7) and that $\mathbb{E}[\varphi^s]^2_{\mathrm{Lip}} \leq C$ with Lemma 9.3.2, where C does not depend on s.

Furthermore, following (9.3.7) we have

$$\begin{aligned}
&\mathbb{E}\Big| \int_0^t \langle \nabla G((\varphi^s)^n), Z_2^n(\cdot,s)\rangle dW_s - \int_0^t \langle \nabla G(\varphi^s), Z_2(\cdot,s)\rangle dW_s\Big|^2 \\
&= \int_0^t \mathbb{E}\big|\langle \nabla G((\varphi^s)^n), Z_2^n(\cdot,s)\rangle - \langle \nabla G(\varphi^s), Z_2(\cdot,s)\rangle\big|^2 ds \\
&\leq 2\|\nabla G\|_\infty \|Z_2\|_\infty \int_0^t \mathbb{E}\big|\langle \nabla G((\varphi^s)^n), Z_2^n(\cdot,s)\rangle - \langle \nabla G(\varphi^s), Z_2(\cdot,s)\rangle\big| ds \\
&\leq 2C\|\nabla G\|_\infty \|Z_2\|_\infty \Big(\|\nabla^2 G\|_\infty n^{-1}\|Z_2\|_\infty \int_0^t \mathbb{E}[\varphi^s]_{\mathrm{Lip}} ds + \|\nabla G\|_\infty n^{-1}\|\partial_1 Z_2\|_\infty \int_0^t \mathbb{E}\|\varphi^s\|_\infty ds\Big)
\end{aligned} \tag{9.3.10}$$

and using Lemma 9.3.2, we have $\mathbb{E}\|\varphi^s\|_\infty \leq C$ and $\mathbb{E}[\varphi^s]_{\mathrm{Lip}} \leq C$ where C does not depend on s, so that the quantity in (9.3.10) converges to 0 as $n \to \infty$.

Last, using (9.3.8) we get

$$\begin{aligned}
&\mathbb{E}\left| \int_0^t \langle \nabla^2 G((\varphi^s)^n), Z_2^n(\cdot,s)^{\otimes 2}\rangle ds - \int_0^t \langle \nabla^2 G(\varphi^s), Z_2(\cdot,s)^{\otimes 2}\rangle ds\right| \\
&\quad \leq C\|\nabla^3 G\|_\infty n^{-1}\|Z_2\|^2_\infty \int_0^t \mathbb{E}[\varphi^s]_{\mathrm{Lip}} ds + C\|\nabla^2 G\|_\infty \|Z_2\|_\infty \|\partial_1 Z_2\|_\infty n^{-1} \int_0^t \mathbb{E}[\|\varphi^s\|_\infty] ds \\
&\quad \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

□

9.4 Proof of Theorem 9.2.2

In this section, we only give the full proof for (9.2.5). The proof of (9.2.6) is similar, as explained in Section 9.4.5.

9.4.1 Definition of the infinite dimensional semi-group and domino strategy

We define the infinite dimensional semi-group that we use for the domino strategy. We do not apply the domino strategy on X directly; instead we define an auxiliary process Y such that X can be induced from Y, as follows. Let us consider the following family of processes:

$$Y_t(u) = \int_0^t K_1(t+u,s) b(X_s) ds + \int_0^t K_2(t+u,s)\sigma(X_s) dW_s, \quad u \geq 0,\ t \in [0,T]. \tag{9.4.1}$$

Following Lemma 9.3.2, for every $u \geq 0$ and $t \in [0,T]$, $Y_t(u)$ is well defined and $Y_t : u \mapsto Y_t(u)$ is continuous. Moreover, for every $t \in [0,T]$, Y_t is $\mathcal{F}_t$-measurable (but the process $(Y_t(u))_{u\geq 0}$ is not adapted w.r.t. u) and using Lemma 9.3.2 again, Y_t is almost surely $\mathcal{C}^1$ and Lipschitz-continuous with

$$\dot{Y}_t(u) = \int_0^t \partial_1 K_1(t+u,s) b(X_s) ds + \int_0^t \partial_1 K_2(t+u,s)\sigma(X_s) dW_s, \tag{9.4.2}$$

where the interchange is ensured by the stochastic Fubini theorem. We also note that

$$Y_t(0) = X_t - X_0, \quad t \in [0,T]. \tag{9.4.3}$$

Then we have

$$X_v = X_0 + Y_t(v-t) + \int_t^v K_1(v,s)b(X_s)ds + \int_t^v K_2(v,s)\sigma(X_s)dW_s, \tag{9.4.4}$$

$$Y_v(u) = Y_t(v-t+u) + \int_t^v K_1(v+u,s)b(X_s)ds + \int_t^v K_2(v+u,s)\sigma(X_s)dW_s, \tag{9.4.5}$$

$0 \le t \le v \le T,\ u \ge 0.$

This leads us to define the following non-homogeneous semi-group $P_{r,t}$ for $t \in [0,T]$ and $r \in [0, T-t]$ on Ω_{2T}:

$$P_{r,t}(\omega)_u = \omega_{r+u} + \int_t^{t+r} K_1(t+r+u,s)b(\widetilde{X}_s)ds + \int_t^{t+r} K_2(t+r+u,s)\sigma(\widetilde{X}_s)dW_s, \quad u \ge 0, \tag{9.4.6}$$

where $(\widetilde{X}_s)_{s\in[t,t+r]}$ is the solution of the following SVE:

$$\widetilde{X}_v = X_0 + \omega_{v-t} + \int_t^v K_1(v,s)b(\widetilde{X}_s)ds + \int_t^v K_2(v,s)\sigma(\widetilde{X}_s)dW_s \tag{9.4.7}$$

and where we omit here the dependency of $\widetilde{X}$ in t and in ω. Since $K_i(u,s) = 0$ for $u \ge 2T$, we have indeed $P_{r,t} : \Omega_{2T} \to \Omega_{2T}$. Then following (9.4.4) and (9.4.5) we have

$$P_{r,t}(Y_t) = Y_{t+r}. \tag{9.4.8}$$

Likewise, we define

$$\bar{Y}_t(u) = \int_0^t K_1(t+u,s)b(\bar{X}_{\underline{s}})ds + \int_0^t K_2(t+u,s)\sigma(\bar{X}_{\underline{s}})dW_s, \quad u \ge 0,\ t \in [0,T] \tag{9.4.9}$$

as well as the semi-group corresponding to the Euler-Maruyama scheme (9.2.3) for $k \in \{0,\dots,N-1\}$ and $r \in [0, T-t_k]$:

$$\bar{P}_{r,t_k}(\omega)_u = \omega_{r+u} + \int_{t_k}^{t_k+r} K_1(t_k+r+u,s)b(X_0+\omega_0)ds + \int_{t_k}^{t_k+r} K_2(t_k+r+u,s)\sigma(X_0+\omega_0)dW_s, \tag{9.4.10}$$

$u \ge 0$,

so that we have

$$\bar{Y}_t(0) = \bar{X}_t - X_0, \quad t \in [0,T] \tag{9.4.11}$$

and for $r \in [0,h]$:

$$\begin{aligned}\bar{P}_{r,t_k}(\bar{Y}_{t_k})_u &= \bar{Y}_{t_k}(r+u) + \int_{t_k}^{t_k+r} K_1(t_k+r+u,s)b(X_0+\bar{Y}_{t_k}(0))ds \\ &\quad + \int_{t_k}^{t_k+r} K_2(t_k+r+u,s)\sigma(X_0+\bar{Y}_{t_k}(0))dW_s \\ &= \bar{Y}_{t_k+r}(u).\end{aligned}$$

By a slight abuse of notation, we use the notations P and $\bar{P}$ also to denote the infinitesimal generators such that for every $g : \Omega_{2T} \to \mathbb{R}$, $\omega \in \Omega_{2T}$, $t \in [0,T]$ and $r \in [0, T-t]$ we have

$$P_{r,t}g(\omega) := \mathbb{E}g(P_{r,t}(\omega)), \quad \bar{P}_{r,t}g(\omega) := \mathbb{E}g(\bar{P}_{r,t}(\omega)).$$

For general semi-groups $Q_1, \dots, Q_r$ we denote their composition as

$$\prod_{k=1}^{r} Q_k := Q_1 \circ \ldots \circ Q_r.$$

Then we obtain $X_T = Y_T(0) + X_0 = P_{T,0}(\widetilde{0})_0 + X_0$ and $\bar{X}_T = \left(\prod_{k=0}^{N-1} \bar{P}_{h,t_{N-1-k}}(\widetilde{0})\right)_0 + X_0$, where $\widetilde{0}$ denotes the path on $\mathbb{R}^+$ constant to 0.

Now for $f : \mathbb{R} \to \mathbb{R}$ being $\widetilde{\mathcal{C}}_b^5$ we define

$$\widetilde{f} : \omega \in \Omega_{2T} \mapsto f(\omega_0 + X_0). \tag{9.4.12}$$

Following Remark 9.3.1, we also have $\widetilde{f} \in \widetilde{\mathcal{C}}_b^5$. Moreover we can write

$$\mathbb{E}f(\bar{X}_T) = \mathbb{E}f\Big(\Big(\prod_{k=0}^{N-1} \bar{P}_{h,t_{N-1-k}}(\widetilde{0})\Big)_0 + X_0\Big) = \mathbb{E}\widetilde{f}\Big(\prod_{k=0}^{N-1} \bar{P}_{h,t_{N-1-k}}(\widetilde{0})\Big) = \prod_{k=0}^{N-1} \bar{P}_{h,t_k}\widetilde{f}(\widetilde{0}).$$

We highlight that in our notations the order of the semi-groups is reversed whether $\prod_k \bar{P}_{h,t_k}$ is applied to some $\omega \in \Omega_{2T}$ or to some $g : \Omega_{2T} \to \mathbb{R}$. We then rewrite the weak error as

$$\begin{aligned}\mathbb{E}f(\bar{X}_T) - \mathbb{E}f(X_T) &= \prod_{k=0}^{N-1} \bar{P}_{h,t_k}\widetilde{f}(\widetilde{0}) - \prod_{k=0}^{N-1} P_{h,t_k}\widetilde{f}(\widetilde{0}) \\ &= \sum_{k=0}^{N-1} P_{kh,0} \circ (\bar{P}_{h,t_k} - P_{h,t_k}) \circ \left(\prod_{j=k+1}^{N-1} \bar{P}_{h,t_j}\right) \widetilde{f}(\widetilde{0}),\end{aligned} \tag{9.4.13}$$

where we used a telescopic sum.

9.4.2 Weak error in small time

In this section, we give a bound on the weak error in small time for the one-step Euler-Maruyama scheme $(\bar{P}_{h,t_k} - P_{h,t_k})g(\omega)$, where $g : \Omega_{2T} \to \mathbb{R}$ is some smooth functional and $\omega \in \Omega_{2T}$ and $\omega \in \mathcal{C}^2$.

Proposition 9.4.1. *Let $\omega \in \Omega_{2T}$ and be $\mathcal{C}^2$, $g : \Omega_{2T} \to \mathbb{R}$ with bounded (pathwise) derivatives up to order 5 and $k \in \{0, \dots, N-1\}$. Then we have*

$$\left|(\bar{P}_{h,t_k} - P_{h,t_k})g(\omega)\right| \leq C(1 + [\omega]_{\mathrm{Lip}})h^2, \tag{9.4.14}$$

where the constant C does not depend on k nor ω nor h.

Proof. We can assume that ω is Lipschitz-continuous without loss of generality.

• Let us consider $(\widetilde{X}_s)_{s\in[t_k,t_k+h]}$ as defined in (9.4.7). Then for $v \in [t_k, t_{k+1}]$ and $\varepsilon \in [0, t_{k+1} - v]$ we have

$$\begin{aligned}\widetilde{X}_{v+\varepsilon} - \widetilde{X}_v = \omega_{v+\varepsilon-t_k} - \omega_{v-t_k} &+ \int_{t_k}^{v} (K_1(v+\varepsilon, s) - K_1(v,s))b(\widetilde{X}_s)ds + \int_{v}^{v+\varepsilon} K_1(v+\varepsilon, s)b(\widetilde{X}_s)ds \\ &+ \int_{t_k}^{v} (K_2(v+\varepsilon, s) - K_2(v,s))\sigma(\widetilde{X}_s)dW_s + \int_{v}^{v+\varepsilon} K_2(v+\varepsilon, s)\sigma(\widetilde{X}_s)dW_s\end{aligned}$$

so that

$$d\widetilde{X}_v = \dot{\omega}_{v-t_k}dv + K_1(v,v)b(\widetilde{X}_v)dv + K_2(v,v)\sigma(\widetilde{X}_v)dW_v \\ + \left(\int_{t_k}^v \partial_1 K_1(v,s)b(\widetilde{X}_s)ds + \int_{t_k}^v \partial_1 K_2(v,s)\sigma(\widetilde{X}_s)dW_s\right)dv.$$

It follows from the classic Itō formula that for $\psi : \mathbb{R} \to \mathbb{R}$ and $v \geq t_k$ we have

$$d\psi(\widetilde{X}_v) = \nabla\psi(\widetilde{X}_v)(K_1(v,v)b(\widetilde{X}_v)dv + \dot{\omega}_{v-t_k}dv + K_2(v,v)\sigma(\widetilde{X}_v)dW_v) \\ + \frac{1}{2}\nabla^2\psi(\widetilde{X}_v)K_2^2(v,v)\sigma^2(\widetilde{X}_v)dv \\ + \nabla\psi(\widetilde{X}_v)\left(\int_{t_k}^v \partial_1 K_1(v,s)b(\widetilde{X}_s)ds + \int_{t_k}^v \partial_1 K_2(v,s)\sigma(\widetilde{X}_s)dW_s\right)dv \tag{9.4.15}$$

In particular we remark that for $r \in [0,h]$ and if ψ is $\mathcal{C}^2$ with bounded derivatives,

$$\begin{aligned}\left|\mathbb{E}\psi(\widetilde{X}_{t_k+r}) - \psi(X_0+\omega_0)\right| &\leq \|\nabla\psi\|_\infty(\|K_1\|_\infty\|b\|_\infty r + [\omega]_{\mathrm{Lip}}) + \frac{1}{2}\|\nabla^2\psi\|_\infty\|K_2\|_\infty^2\|\sigma\|_\infty^2 r \\ &\quad + \|\nabla\psi\|_\infty\|\partial_1 K_1\|_\infty\|b\|_\infty\frac{r^2}{2} + \|\nabla\psi\|_\infty\|\partial_1 K_2\|_\infty\|\sigma\|_\infty\frac{2r^{3/2}}{3} \\ &\leq C(1+[\omega]_{\mathrm{Lip}})r, \end{aligned}\tag{9.4.16}$$

where we bound the last term as follows:

$$\begin{aligned}&\left|\mathbb{E}\int_{t_k}^{t_k+r}\nabla\psi(\widetilde{X}_v)\left(\int_{t_k}^v \partial_1 K_2(v,s)\sigma(\widetilde{X}_s)dW_s\right)dv\right| \\ &\quad\leq \|\nabla\psi\|_\infty\int_{t_k}^{t_k+r}\mathbb{E}\left|\int_{t_k}^v \partial_1 K_2(v,s)\sigma(\widetilde{X}_s)dW_s\right|dv \\ &\quad\leq \|\nabla\psi\|_\infty\|\partial_1 K_2\|_\infty\|\sigma\|_\infty\int_{t_k}^{t_k+r}(v-t_k)^{1/2}dv.\end{aligned}$$

- On the other side for $r \geq 0$ we have

$$\begin{aligned}P_{r+\varepsilon,t_k}(\omega)_u - P_{r,t_k}(\omega)_u &= \omega_{r+\varepsilon+u} - \omega_{r+u} + \int_{t_k+r}^{t_k+r+\varepsilon} K_1(t_k+r+\varepsilon+u,s)b(\widetilde{X}_s)ds \\ &+ \int_{t_k+r}^{t_k+r+\varepsilon} K_2(t_k+r+\varepsilon+u,s)\sigma(\widetilde{X}_s)dW_s \\ &+ \int_{t_k}^{t_k+r}(K_1(t_k+r+\varepsilon+u,s) - K_1(t_k+r+u,s))b(\widetilde{X}_s)ds \\ &+ \int_{t_k}^{t_k+r}(K_2(t_k+r+\varepsilon+u,s) - K_2(t_k+r+u,s))\sigma(\widetilde{X}_s)dW_s\end{aligned}$$

so that we can write

$$dP_{r,t_k}(\omega)_u = \dot{\omega}_{r+u}dr + K_1(t_k+r+u,t_k+r)b(\widetilde{X}_{t_k+r})dr + K_2(t_k+r+u,t_k+r)\sigma(\widetilde{X}_{t_k+r})dW_{t_k+r} \\ + \left(\int_{t_k}^{t_k+r}\partial_1 K_1(t_k+r+u,s)b(\widetilde{X}_s)ds + \int_{t_k}^{t_k+r}\partial_1 K_2(t_k+r+u,s)\sigma(\widetilde{X}_s)dW_s\right)dr$$

so that for $G : \Omega_{2T} \to \mathbb{R}$ being $\widetilde{\mathcal{C}}_b^3$ and using Theorem 9.3.4 (and checking the assumption (9.3.5) with Lemma 9.3.2) we obtain

$$dG(P_{r,t_k}(\omega)) = \langle\nabla G(P_{r,t_k}(\omega)), \dot{\omega}_{r+\cdot}\rangle dr + \langle\nabla G(P_{r,t_k}(\omega)), (K_1(t_k+r+u,t_k+r))_{u\geq 0}\rangle b(\widetilde{X}_{t_k+r})dr$$

$$
\begin{aligned}
&+ \langle \nabla G(P_{r,t_k}(\omega)), (K_2(t_k + r + u, t_k + r))_{u\geq 0} \rangle \sigma(\widetilde{X}_{t_k+r}) dW_{t_k+r} \\
&+ \frac{1}{2} \langle \nabla^2 G(P_{r,t_k}(\omega)), (K_2(t_k + r + u, t_k + r))_{u\geq 0}^{\otimes 2} \rangle \sigma^2(\widetilde{X}_{t_k+r}) dr \\
&+ \Big\langle \nabla G(P_{r,t_k}(\omega)), \Big(\int_{t_k}^{t_k+r} \partial_1 K_1(t_k + r + u, s) b(\widetilde{X}_s) ds \\
&\qquad + \int_{t_k}^{t_k+r} \partial_1 K_2(t_k + r + u, s) \sigma(\widetilde{X}_s) dW_s \Big)_{u\geq 0} \Big\rangle dr. \qquad (9.4.17)
\end{aligned}
$$

In particular, we remark that

$$
\begin{aligned}
\big| \mathbb{E} G(P_{r,t_k}(\omega)) - G(\omega) \big| &\leq \|\nabla G\|_\infty [\omega]_{\mathrm{Lip}} r + \|\nabla G\|_\infty \|K_1\|_\infty \|b\|_\infty r + \frac{1}{2} \|\nabla^2 G\|_\infty \|K_2\|_\infty^2 \|\sigma\|_\infty^2 r \\
&\quad + \|\nabla G\|_\infty \|\partial_1 K_1\|_\infty \|b\|_\infty \frac{r^2}{2} + C \|\nabla G\|_\infty \|\sigma\|_\infty \|\partial_{11}^2 K_2\|_\infty^2 \frac{2r^{3/2}}{3}, \\
&\leq C(1 + [\omega]_{\mathrm{Lip}}) r. \qquad (9.4.18)
\end{aligned}
$$

where we used Lemma 9.3.2 to bound the last term.

Thus for $g : \Omega_{2T} \to \mathbb{R}$ being $\mathcal{C}_b^5$ we have

$$
\begin{aligned}
&\mathbb{E} g(P_{h,t_k}(\omega)) - g(\omega) \\
&= \mathbb{E}\Big[\int_0^h \langle \nabla g(P_{r,t_k}(\omega)), \dot{\omega}_{r+\cdot} \rangle dr + \frac{1}{2} \int_0^h \langle \nabla^2 g(P_{r,t_k}(\omega)), (K_2(t_k + r + u, t_k + r))_{u\geq 0}^{\otimes 2} \rangle \sigma^2(\widetilde{X}_{t_k+r}) dr \\
&\quad + \int_0^h \langle \nabla g(P_{r,t_k}(\omega)), (K_1(t_k + r + u, t_k + r))_{u\geq 0} \rangle b(\widetilde{X}_{t_k+r}) dr \\
&\quad + \int_0^h \Big\langle \nabla g(P_{r,t_k}(\omega)), \Big(\int_{t_k}^{t_k+r} \partial_1 K_1(t_k + r + u, s) b(\widetilde{X}_s) ds \\
&\qquad + \int_{t_k}^{t_k+r} \partial_1 K_2(t_k + r + u, s) \sigma(\widetilde{X}_s) dW_s \Big)_{u\geq 0} \Big\rangle dr \Big] \\
&=: \sum_{i=1}^5 I_i.
\end{aligned}
$$

Likewise, we obtain a similar formula on $\mathbb{E} g(\bar{P}_{h,t_k}(\omega)) - g(\omega)$, replacing b by $\bar{b}$ and σ by $\bar{\sigma}$, and we write

$$
\mathbb{E} g(\bar{P}_{h,t_k}(\omega)) - g(\omega) =: \sum_{i=1}^5 \bar{I}_i.
$$

We shall now inspect the quantity $I_i - \bar{I}_i$ for every $i = 1, \ldots, 5$.

- For fixed $r \in [0, h]$, let $G_r : \eta \mapsto \langle \nabla g(\eta), \dot{\omega}_{r+\cdot} \rangle$. Then $G_r \in \mathcal{C}_b^3$ and we have

$$
\langle \nabla G_r(\eta), \tau \rangle = \langle \nabla^2 g(\eta), \dot{\omega}_{r+\cdot} \otimes \tau \rangle, \quad \langle \nabla^2 G_r(\eta), \tau^1 \otimes \tau^2 \rangle = \langle \nabla^3 g(\eta), \dot{\omega}_{r+\cdot} \otimes \tau^1 \otimes \tau^2 \rangle.
$$

Applying the Itō formula (9.4.17) again to $\alpha \mapsto \mathbb{E} G_r(P_{\alpha,t_k}(\omega))$ for $\alpha \in [0, r]$ and with the estimate (9.4.18), we obtain

$$
|\mathbb{E} G_r(P_{r,t_k}(\omega)) - G_r(\omega)| \leq C r (1 + [\omega]_{\mathrm{Lip}}).
$$

Similarly, we have

$$
\Big| \mathbb{E} G_r(\bar{P}_{r,t_k}(\omega)) - G_r(\omega) \Big| \leq C r (1 + [\omega]_{\mathrm{Lip}}),
$$

and then

$$|I_1-\bar{I}_1| = |\mathbb{E}\Big[\int_0^h G_r(P_{r,t_k}(\omega))dr - \int_0^h G_r(\bar{P}_{h,t_k}(\omega))dr\Big]| \leq C(1+[\omega]_{\mathrm{Lip}})\int_0^h r dr \leq C(1+[\omega]_{\mathrm{Lip}})h^2.$$

- For fixed $r \in [0,h]$, let

$$G_r : \eta \mapsto \frac{1}{2}\langle \nabla^2 g(\eta), (K_2(t_k+r+u,t_k+r))_{u\geq 0}^{\otimes 2}\rangle.$$

Then $G_r \in \mathcal{C}_b^3$ and we have

$$\langle \nabla G_r(\eta), \tau\rangle = \frac{1}{2}\langle \nabla^3 g(\eta), (K_2(t_k+r+u,t_k+r))_{u\geq 0}^{\otimes 2}\otimes \tau\rangle,$$
$$\langle \nabla^2 G_r(\eta), \tau^1\otimes\tau^2\rangle = \frac{1}{2}\langle \nabla^4 g(\eta), (K_2(t_k+r+u,t_k+r))_{u\geq 0}^{\otimes 2}\otimes\tau^1\otimes\tau^2\rangle.$$

Applying the Itō formulae we obtained in (9.4.17) and in (9.4.15) and the classic Itō formula for a product, we get

$$\begin{aligned}
&\mathbb{E}[G_r(P_{r,t_k}(\omega))\sigma^2(\widetilde{X}_{t_k+r})] - G_r(\omega)\sigma^2(X_0+\omega_0)\\
&\quad = \mathbb{E}[\int_0^r d(G_r(P_{\alpha,t_k}(\omega)))\sigma^2(\widetilde{X}_{t_k+\alpha}) + G_r(P_{\alpha,t_k}(\omega))d(\sigma^2(\widetilde{X}_{t_k+\alpha})) + d\langle G_r(P_{\cdot,t_k}(\omega)), \sigma^2(\widetilde{X}_{t_k+\cdot})\rangle_\alpha]\\
&\quad =: A_1 + A_2 + A_3,
\end{aligned}$$

but σ^2 is bounded and following (9.4.18), we obtain that $A_1 \leq C(1+[\omega]_{\mathrm{Lip}})r$; the same way and since G_r is bounded (independently on r) and following (9.4.16) we have $A_2 \leq C(1+[\omega]_{\mathrm{Lip}})r$. Moreover we have

$$\begin{aligned}
d\langle G_r(P_{\cdot,t_k}(\omega)), \sigma^2(\widetilde{X}_{t_k+\cdot})\rangle_\alpha = {}& \nabla\sigma^2(\widetilde{X}_{t_k+\alpha})K_2(t_k+\alpha,t_k+\alpha)\sigma^2(\widetilde{X}_{t_k+\alpha})\\
&\cdot\langle \nabla G_r(P_{\alpha,t_k}(\omega)), (K_2(t_k+\alpha+u,t_k+\alpha))_{u\geq 0}\rangle d\alpha
\end{aligned}$$

so that same way we get $A_3 \leq Cr$. Thus we finally obtain

$$|\mathbb{E}G_r(P_{r,t_k}(\omega))\sigma^2(\widetilde{X}_{t_k+r}) - G_r(\omega)\sigma^2(X_0+\omega_0)| \leq C(1+[\omega]_{\mathrm{Lip}})r. \tag{9.4.19}$$

The same way we have

$$|\mathbb{E}G_r(\bar{P}_{r,t_k}(\omega))\sigma^2(X_0+\omega_0) - G_r(\omega)\sigma^2(X_0+\omega_0)| \leq C(1+[\omega]_{\mathrm{Lip}})r,$$

so that

$$|I_2-\bar{I}_2| = |\mathbb{E}\Big[\int_0^h (G_r(P_{r,t_k}(\omega))\sigma^2(\widetilde{X}_{t_k+r}) - G_r(\bar{P}_{r,t_k}(\omega))\sigma^2(X_0+\omega_0))dr\Big]| \leq C(1+[\omega]_{\mathrm{Lip}})h^2.$$

- For $r \in [0,h]$ and $u \geq 0$ let us define

$$\varphi_u^r := \int_{t_k}^{t_k+r} \partial_1 K_2(t_k+r+u,s)\sigma(\widetilde{X}_s)dW_s$$

and let us write

$$d\varphi_u^r = \partial_1 K_2(t_k+r+u,t_k+r)\sigma(\widetilde{X}_{t_k+r})dW_{t_k+r} + \left(\int_{t_k}^{t_k+r}\partial_{11}^2 K_2(t_k+r+u,s)\sigma(\widetilde{X}_s)dW_s\right)dr. \tag{9.4.20}$$

Moreover, using Lemma 9.3.2, we have $\mathbb{E}\|\varphi^r\|_\infty^2 \le C$. We also define

$$G : (\eta^1, \eta^2) \in \Omega_{2T}^2 \mapsto \langle \nabla g(\eta^1), \eta^2 \rangle.$$

Then $G \in \mathcal{C}_b^3$ with

$$\begin{aligned}
&\langle \nabla G(\eta^1, \eta^2), (\tau^1, \tau^2) \rangle = \langle \nabla^2 g(\eta^1), \tau^1 \otimes \eta^2 \rangle + \langle \nabla g(\eta^1), \tau^2 \rangle, \\
&\langle \nabla^2 G(\eta^1, \eta^2), (\tau^1, \tau^2)^{\otimes 2} \rangle = \langle \nabla^3 g(\eta^1), (\tau^1)^{\otimes 2} \otimes \eta^2 \rangle + 2\langle \nabla^2 g(\eta^1), \tau^1 \otimes \tau^2 \rangle.
\end{aligned}$$

Using (9.4.17) and (9.4.20), for every $r \in [0, h]$ we have

$$\begin{aligned}
&\langle \nabla g(P_{r,t_k}(\omega)), \Big(\int_{t_k}^{t_k+r} \partial_1 K_2(t_k + r + u, s) \sigma(\widetilde{X}_s) dW_s \Big)_{u \ge 0} \rangle \\
&\quad = \int_0^r \Big[\langle \nabla^2 g(P_{\alpha,t_k}(\omega)), \dot{\omega}_{r+\cdot} \otimes \varphi^\alpha \rangle \\
&\qquad + \langle \nabla^2 g(P_{\alpha,t_k}(\omega)), (K_1(t_k + \alpha + u, t_k + \alpha))_{u \ge 0} \otimes \varphi^\alpha \rangle b(\widetilde{X}_{t_k+\alpha}) \\
&\qquad + \frac{1}{2} \langle \nabla^3 g(P_{\alpha,t_k}(\omega)), (K_2(t_k + \alpha + u, t_k + u))_{u \ge 0}^{\otimes 2} \otimes \varphi^\alpha \rangle \sigma^2(\widetilde{X}_{t_k+\alpha}) \\
&\qquad + \langle \nabla^2 g(P_{\alpha,t_k}(\omega)), \Big(\int_{t_k}^{t_k+\alpha} \partial_1 K_1(t_k + \alpha + u, s) b(\widetilde{X}_s) ds \\
&\qquad\quad + \int_{t_k}^{t_k+\alpha} \partial_1 K_2(t_k + \alpha + u, s) \sigma(\widetilde{X}_s) dW_s \Big)_{u \ge 0} \otimes \varphi^\alpha \rangle \Big] d\alpha \\
&\qquad + \int_0^r \langle \nabla^2 g(P_{\alpha,t_k}(\omega)), (K_2(t_k + \alpha + u, t_k + u))_{u \ge 0} \otimes \varphi^\alpha \rangle \sigma(\widetilde{X}_{t_k+\alpha}) dW_{t_k+\alpha} \\
&\qquad + \int_0^r \Big[\langle \nabla g(P_{\alpha,t_k}(\omega)), \Big(\int_{t_k}^{t_k+\alpha} \partial_{11}^2 K_2(t_k + \alpha + u, s) \sigma(\widetilde{X}_s) dW_s \Big)_{u \ge 0} \rangle d\alpha \\
&\qquad + \int_0^r \langle \nabla g(P_{\alpha,t_k}(\omega)), (\partial_1 K_2(t_k + \alpha + u, t_k + \alpha))_{u \ge 0} \rangle \sigma(\widetilde{X}_{t_k+\alpha}) dW_{t_k+\alpha} \\
&\qquad + \int_0^r \langle \nabla^2 g(P_{\alpha,t_k}(\omega)), (K_2(t_k + \alpha + u, t_k + u))_{u \ge 0} \otimes (\partial_1 K_2(t_k + \alpha + u, t_k + \alpha))_{u \ge 0} \rangle \\
&\qquad \cdot \sigma^2(\widetilde{X}_{t_k+\alpha}) d\alpha
\end{aligned}$$

so that

$$\Big| \mathbb{E} \langle \nabla g(P_{r,t_k}(\omega)), \Big(\int_{t_k}^{t_k+r} \partial_1 K_2(t_k + r + u, s) \sigma(\widetilde{X}_s) dW_s \Big)_{u \ge 0} \rangle \Big| \le C(1 + [\omega]_{\mathrm{Lip}}) r.$$

The same way, we obtain

$$\Big| \mathbb{E} \langle \nabla g(\bar{P}_{r,t_k}(\omega)), \Big(\int_{t_k}^{t_k+r} \partial_1 K_2(t_k + r + u, s) \bar{\sigma}(X_0 + \omega(0)) dW_s \Big)_{u \ge 0} \rangle \Big| \le C(1 + [\omega]_{\mathrm{Lip}}) r$$

and then

$$|I_5 - \bar{I}_5| \le C(1 + [\omega]_{\mathrm{Lip}}) h^2.$$

- The arguments to prove that

$$|I_3 - \bar{I}_3| + |I_4 - \bar{I}_4| \le C(1 + [\omega]_{\mathrm{Lip}}) h^2$$

are the same or more simple.

□

9.4.3 Proof that the derivatives of g are bounded

In this section, we prove that if we choose $g : \Omega \to \mathbb{R}$ as in (9.4.13), then g has bounded derivatives up to order 5 so that we can apply Proposition 9.4.1 to g.

Lemma 9.4.2. *Let $k \in \{0, \dots, N-1\}$, let us define*

$$g(\omega) := \prod_{j=k}^{N-1} \bar{P}_{h,t_j} \widetilde{f}(\omega) = \mathbb{E}\left[f((\prod_{j=k}^{N-1} \bar{P}_{h,t_{N-1+k-j}} \cdot \omega)_0) \right],$$

where $\widetilde{f}$ is defined in (9.4.12)*. Then g is five times differentiable with*

$$\|\nabla^\ell g\|_\infty \leq C, \quad \ell \in \{1, \dots, 5\}.$$

Proof. We can rewrite

$$g(\omega) = \mathbb{E}\left[f(\omega_{T-t_k} + \int_{t_k}^T K_1(T,s)\bar{b}(\hat{X}_s^\omega)ds + \int_{t_k}^T K_2(T,s)\bar{\sigma}(\hat{X}_s^\omega)dW_s) \right]$$

where $\hat{X}^\omega$ follows the piecewise SVE:

$$\hat{X}_v^\omega = X_0 + \omega_{v-t_k} + \sum_{\ell=k}^{j} b(\hat{X}_{t_\ell}^\omega) \int_{t_\ell}^{t_{\ell+1}\wedge v} K_1(v,s)ds + \sum_{\ell=k}^{j} \sigma(\hat{X}_{t_\ell}^\omega) \int_{t_\ell}^{t_{\ell+1}\wedge v} K_2(v,s)dW_s,$$
$$v \in [t_j, t_{j+1}],\ j \in \{k, \dots, N-1\}.$$

The process $\hat{X}^\omega$ depends on k, but we omit this dependency in the notation without ambiguity. We define the tangent process $(Z_v^\omega)_{v\in[T-t_k,T]}$ as the process of the linear operators on Ω_{2T} by induction as:

$$\begin{aligned}\langle Z_v^\omega, \eta\rangle = \eta_{v-t_k} &+ \sum_{\ell=k}^{j} \nabla b(\hat{X}_{t_\ell}^\omega) \cdot \langle Z_{t_\ell}^\omega, \eta\rangle \int_{t_\ell}^{t_{\ell+1}\wedge v} K_1(v,s)ds \\ &+ \sum_{\ell=k}^{j} \nabla\sigma(\hat{X}_{t_\ell}^\omega) \cdot \langle Z_{t_\ell}^\omega, \eta\rangle \int_{t_\ell}^{t_{\ell+1}\wedge v} K_2(v,s)dW_s, \quad v \in [t_j, t_{j+1}],\ j \in \{k, \dots, N-1\},\end{aligned} \tag{9.4.21}$$

so that for every fixed v, we have $Z_v^\omega = \nabla \hat{X}_v^\omega$, where ∇ is taken with respect to ω. We now give a bound on $\|Z_v^\omega\|$. For every $\eta \in \Omega_{2T}$ and $v \in [T_k, T]$ we have

$$\langle Z_v^\omega, \eta\rangle = \eta_{v-t_k} + \int_{t_k}^v K_1(v,s)\nabla\bar{b}(\hat{X}_s^\omega)\langle Z_{\underline{s}}^\omega, \eta\rangle ds + \int_{t_k}^v K_2(v,s)\nabla\bar{\sigma}(\hat{X}_s^\omega)\langle Z_{\underline{s}}^\omega, \eta\rangle dW_s.$$

Let us denote

$$\varphi_v := \sup_{s\in[t_k,v]} \mathbb{E}\|Z_s^\omega\|^2, \quad v \in [t_k, T]$$

and we have (using the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$):

$$\varphi_v \leq 3 + 3\int_{t_k}^v (T\|K_1\|_\infty^2\|\nabla b\|_\infty^2 + \|K_2\|_\infty^2\|\nabla\sigma\|_\infty^2)\varphi_s ds$$

so that using the Gronwall inequality:

$$\varphi_v \leq 3\exp\left((v-t_k)(T\|K_1\|_\infty^2\|\nabla b\|_\infty^2 + \|K_2\|_\infty^2\|\nabla\sigma\|_\infty^2)\right) \leq C.$$

Then we have

$$\langle \nabla g(\omega), \eta \rangle = \mathbb{E}\left[\nabla f((\prod_{j=k}^{N-1} \bar{P}_{h,t_{N-1+k-j}} \cdot \omega)_0) \cdot \Big(\eta_{T-t_k} + \int_{t_k}^{T} K_1(T,s) \nabla \bar{b}(\hat{X}_s^\omega) \langle Z_{\underline{s}}^\omega, \eta \rangle ds + \int_{t_k}^{T} K_2(T,s) \nabla \bar{\sigma}(\hat{X}_s^\omega) \langle Z_{\underline{s}}^\omega, \eta \rangle dW_s \Big) \right]$$

implying

$$\|\nabla g\|_\infty \leq \|\nabla f\|_\infty \Big(1 + \int_{t_k}^{T} \|K_1\|_\infty \|\nabla b\|_\infty \varphi_s^{1/2} ds + (\int_{t_k}^{T} \|K_2\|_\infty^2 \|\nabla \sigma\|_\infty^2 \varphi_s ds)^{1/2} \Big)$$

thus implying that ∇g is bounded (independently of k and N).

We prove that the derivatives of g are bounded up to order 5 by following the same method. □

9.4.4 Conclusion: proof of Theorem 9.2.2

Proof. Let us consider (9.4.13) again and for $k \in \{0, \dots, N-1\}$ we set $\omega^k := P_{kh,0}(\widetilde{0})$ and

$$g_{k+1} := \prod_{j=k+1}^{N-1} \bar{P}_{h,t_j} \widetilde{f}.$$

Then following Lemma 9.4.2, we have that $g_{k+1} \in \widetilde{\mathcal{C}}_b^5$. On the other side, we have that ω^k is $\mathcal{C}^2$ with

$$\dot{\omega}_u^k = \int_0^{t_k} \partial_1 K_1(t_k + u, s) b(\widetilde{X}_s) ds + \int_0^{t_k} \partial_1 K_2(t_k + u, s) \sigma(\widetilde{X}_s) dW_s,$$
$$\ddot{\omega}_u^k = \int_0^{t_k} \partial_{11}^2 K_1(t_k + u, s) b(\widetilde{X}_s) ds + \int_0^{t_k} \partial_{11}^2 K_2(t_k + u, s) \sigma(\widetilde{X}_s) dW_s$$

where the interchange is ensured by the stochastic Fubini theorem, and following Lemma 9.3.2, we obtain that $\mathbb{E}[\omega^k]_{\text{Lip}} \leq C$. Then applying Proposition 9.4.1 with $g = g_{k+1}$ and $\omega = \omega^k$ we get

$$P_{kh,0} \circ (\bar{P}_{h,t_k} - P_{h,t_k}) \circ \left(\prod_{j=k+1}^{n-1} \bar{P}_{h,t_j} \right) \widetilde{f}(\widetilde{0}) \leq C h^2.$$

Summing over $k \in \{0, \dots, N-1\}$ yields

$$\mathbb{E}[f(\bar{X}_T)] - \mathbb{E}[f(X_T)] = O\left(\frac{1}{N}\right).$$

□

9.4.5 Proof of weak error for the scheme with discretization of the kernels

The proof of (9.2.6) for the scheme $\overrightarrow{\overline{X}}$ defined in (9.2.4) is similar to the proof for (9.2.5). We define the associated semi-group as

$$\overrightarrow{\overline{P}}_{r,t_k}(\omega)_u = \omega_{r+u} + \int_{t_k}^{t_k+r} K_1(t_k + r + u, t_k) b(X_0 + \omega_0) ds$$

$$+\int_{t_k}^{t_k+r} K_2(t_k+r+u,t_k)\sigma(X_0+\omega_0)dW_s, \quad u\geq 0. \tag{9.4.22}$$

The estimate for the weak error in small time (9.4.14) also holds for $|(\overrightarrow{P}_{h,t_k}-P_{h,t_k})g(\omega)|$; the only necessary adaptation in the proof is for the estimate for I_2 and I_3. Indeed, instead of (9.4.19) we need to prove

$$\left|\mathbb{E}G_r(P_{r,t_k}(\omega))\sigma^2(\widetilde{X}_{t_k+r}) - \frac{1}{2}\langle\nabla^2 g(\omega), K_2(t_k+r+\cdot,t_k)^{\otimes 2}\rangle\sigma^2(X_0+\omega_0)\right| \leq C(1+[\omega]_{\mathrm{Lip}})r.$$

But we have

$$\alpha \mapsto \langle\nabla^2 g(\omega), K_2(t_k+r+\cdot,t_k+\alpha)^{\otimes 2}\rangle$$

is $\mathcal{C}^1$ with derivative

$$\alpha \mapsto 2\langle\nabla^2 g(\omega), K_2(t_k+r+\cdot,t_k)\otimes\partial_2 K_2(t_k+r+\cdot,t_k)\rangle$$

and since g, K_2, $\partial_2 K_2$ and σ are bounded we have

$$\begin{aligned}
&\left|\mathbb{E}G_r(P_{r,t_k}(\omega))\sigma^2(\widetilde{X}_{t_k+r}) - \frac{1}{2}\langle\nabla^2 g(\omega), K_2(t_k+r+\cdot,t_k)^{\otimes 2}\rangle\sigma^2(X_0+\omega_0)\right| \\
&\quad\leq \left|\mathbb{E}G_r(P_{r,t_k}(\omega))\sigma^2(\widetilde{X}_{t_k+r}) - G_r(\omega)\sigma^2(X_0+\omega_0)\right| \\
&\qquad + \left|\mathbb{E}G_r(\omega)\sigma^2(X_0+\omega_0) - \frac{1}{2}\langle\nabla^2 g(\omega), K_2(t_k+r+\cdot,t_k)^{\otimes 2}\rangle\sigma^2(X_0+\omega_0)\right| \\
&\quad\leq C(1+[\omega]_{\mathrm{Lip}})r.
\end{aligned}$$

The argument for the estimate of I_3 is similar.

Having proved the estimate for the weak error in small time, the conclusion of the proof is the same as for $\bar{X}$.

9.5 Simulations

In order to numerically check the convergence rate obtained in Theorem 9.2.2, we empirically measure the weak convergence rate in the case of a stochastic volatility model where the volatility follows some Volterra equation. We consider the following Volterra version of the Stein-Stein model [SS91] where the analogous rough version was introduced in [AJ22]:

$$\begin{cases} dS_t = S_t V_t dB_t, \quad S_0>0, \\ V_t = V_0 + g_0(t) + \kappa\displaystyle\int_0^t K(t-s)V_s ds + \nu\int_0^t K(t-s)dW_s \end{cases} \tag{9.5.1}$$

where the asset price process S and the square volatility process V take their values in $\mathbb{R}$, the function $g_0:\mathbb{R}^+\to\mathbb{R}$ is deterministic and continuous, the processes B and W are standard Brownian motions with correlation $\rho\in[-1,1]$ and the non-singular kernel K is given by

$$K(t) = A_1(A_2+t)^{-1/4},$$

with $A_1,\ A_2>0$. The process $(S_t,V_t)^\top$ in (9.5.1) is a special case of the Volterra equation (9.2.1) with 2×2 matrix kernels

$$K_1(t,s) = K_2(t,s) = \begin{pmatrix} 1 & 0 \\ 0 & K(t-s)\end{pmatrix}.$$

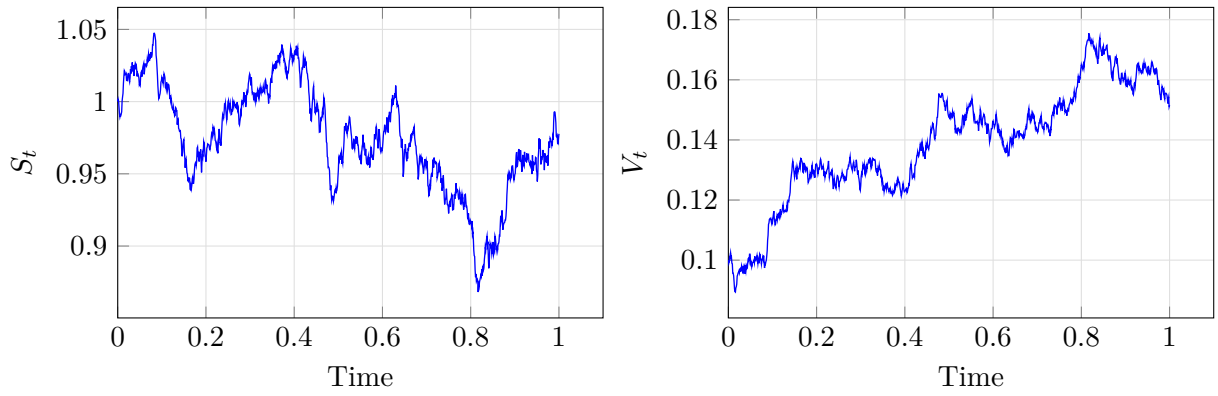


Figure 9.1: Example of trajectory of the asset price and the square volatility processes following the SVE (9.5.1).

We consider the payoff function given by the Call option

$$f(x) = (x - \mathcal{K})_+$$

with strike $\mathcal{K} \geq 0$.

We simulate $(\overrightarrow{S}_T, \overrightarrow{V}_T)$ by discretizing the kernel K using weights matching the second moment, see [Gas23, Section 3] and we plot $\mathbb{E}f(\overrightarrow{S}_T^{\lfloor \beta N \rfloor}) - \mathbb{E}f(\overrightarrow{S}_T^N)$ for some $\beta \in (1, 2]$ and for different values of N, where N is the number of steps in the Euler-Maruyama scheme of the SVE. If $\mathbb{E}f(\overrightarrow{S}_T^N) = \mathbb{E}f(S_T) + O(1/N)$, then we should also have

$$\mathbb{E}f(\overrightarrow{S}_T^{\lfloor \beta N \rfloor}) - \mathbb{E}f(\overrightarrow{S}_T^N) = O(1/N).$$

An example of trajectory is given in Figure 9.1 and the results are given in Figure 9.2 with the following parameters:

$T = 1,\ S_0 = 1,\ \mathcal{K} = 1,\ \kappa = 0.01,\ X_0 = 0.1,\ \rho = -0.7,\ \nu = 0.05,\ A_1 = 0.3,\ A_2 = 0.02,\ \beta = 1.5,$
$g_0 : t \mapsto (4\theta)/(3A_1)t^{3/4},\ \theta = 0.01.$

We empirically obtain a convergence rate for the weak error which is approximatively -1, thus confirming the results in Theorem 9.2.2.

Acknowledgements

The authors thank Professor Gilles Pagès, Sorbonne University, Paris, and Professor Toshihiro Yamada, Hitotsubashi University, Tokyo, for insightful discussions.
The first author thanks the Japanese Society for Promotion of Science JSPS for financial support for a visit to Osaka University and to Hitotsubashi University within the JSPS Summer Program.

9.6 Appendix

We use the following version of the Kolmogorov continuity theorem, giving the precise upper bound constant.

Theorem 9.6.1 (Kolmogorov continuity theorem)**.** *Let $(X_t)_{t\in[0,T]}$ be a $\mathbb{R}^d$-valued random process and assume that for some $p, \epsilon > 0$,*

$$\mathbb{E}[|X_t - X_s|^p] \leq C_0|t - s|^{1+\varepsilon}, \quad t, s \in [0, T].$$

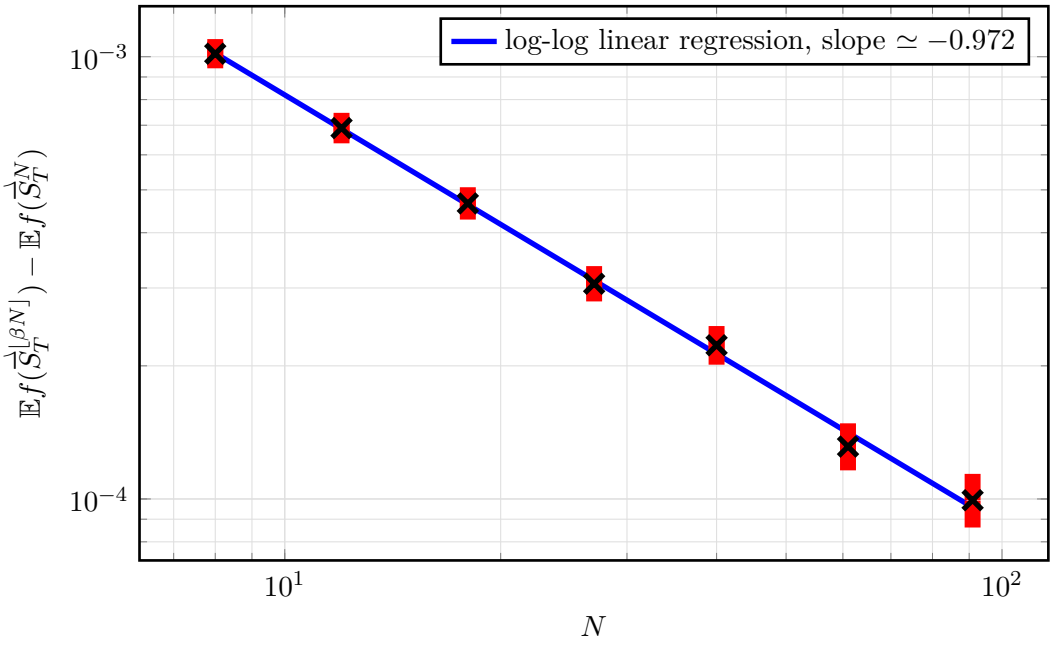


Figure 9.2: Simulation of (9.5.1) and weak error in log-log scale. We give in red the 95% confidence intervals where the number of trajectories is $512 \times 1024 \times 2000$ for each value of N.

Then there exists a modification $\widetilde{X}$ of X which is α-Hölder continuous for every $\alpha \in (0, \varepsilon/p)$ and with

$$\mathbb{E}\left[\left(\sup_{t,s\in[0,T],\ t\neq s} \frac{|\widetilde{X}_t - \widetilde{X}_s|}{|t-s|^\alpha}\right)^p\right] \leq C_0 \left(\frac{2^{1+\alpha}}{1-2^{-\alpha}}\right)^p \frac{T^{1+\varepsilon-\alpha p}}{1-2^{-(\varepsilon-\alpha p)}}.$$

Proof. We refer to the proof of [LG16, Theorem 2.9, Lemma 2.10], with an immediate adaptation if we do not assume $T = 1$. □

Chapter 10

Simulation of Reflected Brownian motion on two dimensional wedges

This chapter corresponds to the article [BKH23] published in *Stochastic Processes and their Applications* as a joint work with Arturo Kohatsu-Higa.

Abstract

We study a correlated Brownian motion in two dimensions, which is reflected, stopped or killed in a wedge represented as the intersection of two half spaces. First, we provide explicit density formulas, hinted by the method of images. These explicit expressions rely on infinite oscillating sums of Bessel functions and may demand computationally costly procedures.

We propose suitable recursive algorithms for the simulation of the laws of reflected and stopped Brownian motion which are based on generalizations of the reflection principle in two dimensions. We study and give bounds for the complexity of the proposed algorithms.

Keywords– Reflected Brownian motion, Wedge, Hitting times, Reflection principle, Method of images, Monte Carlo simulation.

10.1 Introduction

The objective of the present article is to provide exact formulas and algorithms for the simulation of a normally reflected two-dimensional Brownian motion starting at $x_0 \in \mathbb{R}^2$.

The reflected process is denoted by $X \equiv (X_t)_{t\in[0,T]}$ and the domain of reflection is a wedge $\mathcal{D}$, i.e. a subset of $\mathbb{R}^2$ delimited by two (non parallel) lines, so as to give Monte Carlo methods for the estimation of $\mathbb{E}^{x_0}[f(X_T)]$, where $f : \mathcal{D} \to \mathbb{R}$ is a measurable function such that $f(X_T) \in L^1$ and $T > 0$ is the finite time horizon. Our first goal is to obtain an explicit formula for the density of X_t (see Theorem 10.4.2). Unfortunately, this formula involves oscillating infinite sums of Bessel functions. Instead of directly computing these sums, we propose an alternative

simulation method which uses an extension of the reflection principle in two dimensions for a particular type of wedges with angle π/m, $m \in \mathbb{N}$. As a first step, we obtain a simulation method for $\mathbb{E}^{x_0}[f(W_{T\wedge\tau})]$, where $W = (W_t)_{t\geq 0}$ is a two-dimensional Brownian motion and τ stands for the first time the process W touches the boundary of the wedge $\mathcal{D}$. Applying the methodology for the stopped process recursively one obtains an algorithm for the reflected process X. The algorithms we present here rely on an idea which has links to the rectangle method [DL06]: we include a smaller wedge of angle π/m inside the first wedge and simulate the exit from the smaller wedge.

The algorithm for the simulation of the stopped process improves the method proposed in [Met09]. In the reflected case, we prove that although some lower moments are finite, the average number of simulations for this method is infinite. This effect is created due to the large number of reflections that may happen if the process enters the corner of the wedge. We then propose a modified approximation scheme which takes into account the asymptotic behaviour of the density near the corner. With this modification, we show that the expected number of iterations of the algorithm is finite and measure the error of approximation in total variation distance. We then give adaptations of our algorithms to more general Itō processes reflected in a wedge.

The expression of the density of the stopped Brownian motion in a wedge of $\mathbb{R}^2$ has been obtained in [Iye85] using the method of images. However, the author does not provide full arguments for the verification of the initial condition and some exchanges of infinite summation and integrations are not fully explained. For the history of the expression of the density of the stopped Brownian motion on a wedge, see [BnS97].

In [CBM15] and [Met10] the authors correct some mistakes in formulas appearing in [Iye85] and give formulas for a number of other random variables related to the stopped Brownian motion in two dimensions with general correlation coefficients, such as the survival probability and the first-passage time distribution. However, formulas from [Iye85] and [Met10] are not directly applicable to simulation algorithms of stopped processes, as they involve infinite sums and Bessel functions. [Met09, Section 2] gives an approximate simulation algorithm using the semi-analytic expression of the density established in [Iye85]. But this method implies the approximation of an infinite sum of Bessel functions and it leads to bias when simulating the stopping time τ.

More recent papers ([EFW13], [BSCD13]) extend the results of [Iye85] and [Met10] to three dimensions and give algorithms and density formulas for the stopped Brownian motion in three dimensions, using the method of images in $\mathbb{R}^3$. These algorithms which point towards the possibility of larger dimension extensions are however limited to some specific values of correlation. In [KLR18], the authors use the spectral decomposition method which applies in a larger generality but still relies on efficient computational methods for eigenvalue calculations and truncations of infinite sums. These references concentrate on the stopped Brownian motion case and there is no discussion about the simulation of reflected Brownian motion in wedges.

Simulation algorithms for the reflected Brownian motion have been only partly studied, although they share similarities with the stopped case. Simulation algorithms for quantities related to reflected Brownian motion on an orthant in multi-dimensions can be found in [BM18] and [BC15]. They do not use the method of images, but the characterization through the Skorokhod problem and the so-called ε-strong methodology which relies on the regularity of the reflection functional. Still, the method proposed has a theoretically infinite mean running time.

Other approximations methods which are closer to random walk versions of the reflected Brownian motion on tile domains are studied in [Dub04] and [Kag07]. Furthermore, other simulations methods for reflected Brownian motion which do not cover the case of the wedge can be found in [CPS98], [BGT04], [BC08] and [BST10] the references therein.

The reflected Brownian motion and more generally reflected processes have applications in finance (for example, in stochastic models where the process, which can model an interest rates, is constrained to be non-negative [Ha09] or has other constraints like barriers such as in [IKP13]); in queueing models [GNR86], etc. In particular, in the recent years there has been a lot of developments concerning the study of stationary measures of reflected Brownian motion such as [DM09] which is slightly related to the ideas which we use here to construct a simulation method on a wedge. For a review, on this topic and the relationship with queuing theory see e.g. [Die10].

Considering the particular case of a wedge may open the way to new simulation algorithms for reflected processes adapted to non-smooth domains with corner points in two or more dimensions (see Section 10.11).

The paper is organized as follows. In Section 10.3 we state the framework of the problem, giving a parametrization of the wedge in $\mathbb{R}^2$. We also give the formula for the density of the stopped Brownian motion in a wedge of angle $\alpha \in (0, 2\pi)$. The density of the stopped Brownian motion had already been obtained in [Iye85]. A more general case is studied in [BnS97] which provides a detailed history and a full proof that the density formula is the fundamental solution of the associated partial differential equation.

In Section 10.4, we provide the density of the reflected Brownian motion on the wedge. First, we use the method of images in a special type of wedges as it serves to understand how one induces the density of the reflected Brownian motion in the general case. In the Appendix we include the proof in the reflected case on more general cones following [BnS97].

In Section 10.5, we give explicit algorithms for the simulation of the stopped and reflected Brownian motions. To do so, we first derive from [Met10] and the formulas proved in Section 10.4 expressions of densities of the exit time and exit point from the wedge. Then, we give an algorithm for the simulation of the stopped Brownian motion using the explicit expressions in the case $\alpha = \pi/m$. The simulation algorithm is based on the simulation of a random sequence of stopped Brownian motion processes on domains with angle $\alpha = \pi/m$. This first algorithm is useful for the simulation of reflected Brownian motion, but it also has its own interest. Then we give the simulation algorithm for the reflected Brownian motion with unit diffusion matrix.

In Section 10.6, we study the mean number of iterations of the algorithm to finish, denoted $\mathbb{E}[N]$. For the stopped Brownian motion, $\mathbb{E}[N]$ is bounded above by a constant. The reflected case is more difficult to deal with. In this case, N is equal to the number of times a Brownian motion goes from one frontier of a wedge to another. We first give theoretical bounds for $\mathbb{E}[N]$ and prove that although the number of iterations is almost surely finite, $\mathbb{E}[N]$ is infinite, which is due to the events when the Brownian motion comes close to the origin. We then propose a modification of the algorithm: we stop the algorithm when the Brownian motion comes too close to the origin, measured by a parameter ε and use an approximation of the density for the reflected Brownian motion close to the origin, induced by the density formula in Section 10.4. We then prove that for the modified algorithm, $\mathbb{E}[N]$ has an upper bound which grows as ε^{1-p} for all $p \in (1, 2)$, and that the error of approximation is of order ε.

In Section 10.7, we show how the algorithms can directly be applied for the simulation of the reflected and stopped Brownian motions with non-zero drift. Then we give an algorithm for the simulation of a class of Itō diffusion processes, reflected or stopped in a wedge in $\mathbb{R}^2$. This algorithm is in fact a direct application of the precedent algorithms, combined with a Euler-Maruyama scheme.

Finally, in Section 10.8, we perform Monte Carlo simulations for the estimation of reflected and stopped Brownian motions and Itō processes. We show that the bias in the algorithm from [Met09] can be significant compared with our method which is exact. Simulations prove the need for the approximation algorithm for the reflected Brownian motion, as the exact algorithm

takes too much time to be efficient. Clearly, one remaining subject which is not considered here is the reflecting case with non-unit diffusion coefficient. This remains an open problem.

The authors would like to express thanks to the anonymous referee which provided references [BnS97] and [DM09] which improved the results and shortened the proof of Theorem 10.4.2.

10.2 Notations

We give a brief list of notations that are used through the text.

The natural logarithm is denoted with log. We frequently use the notation $\pm$ in order to denote the two rays that define a wedge. Sometimes rather than writing two equations we just write one using the symbol $\pm$ or $\mp$ meaning that we are stating two equations, one using the top symbols and another using the bottom symbols that appear throughout the equation.

I_a stands for the modified Bessel function of the first kind given for $a \geq 0,\ x \geq 0$ by:

$$I_a(x) = \sum_{k=0}^{\infty} \frac{x^{a+2k}}{2^{a+2k} k! \Gamma(k+a+1)}. \tag{10.2.1}$$

This function satisfies the differential equation

$$x^2 \frac{d^2 I_a(x)}{dx^2} + x \frac{dI_a(x)}{dx} - (x^2 + a^2) I_a(x) = 0,\ x \geq 0. \tag{10.2.2}$$

We consider the space $\mathbb{R}^2$ endowed with the Euclidean norm denoted by $|\cdot|$.

For $\mathcal{D} \subset \mathbb{R}^2$ and $k \in \mathbb{N}$, we denote by $\mathcal{C}_b^k(\mathcal{D})$ the set of real-valued functions defined on $\mathcal{D}$ which are bounded and have bounded partial derivatives up to the order k.

Now we define the probabilistic setting as follows. Let $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$ be a filtered probability space satisfying the usual conditions. Let $T > 0$ be the time horizon and let $W = (W_t)_{t\geq 0}$ be a two-dimensional correlated Brownian motion on $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$. We denote by Σ the covariance matrix of W and we assume that Σ is a non-singular matrix. We define a family of probabilities on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0})$ by:

$$\forall x_0 \in \mathbb{R}^2,\ \mathbb{P}^{x_0} := \mathbb{P}_{|W_0 = x_0}.$$

We will denote by $\mathbb{E}^{x_0}$ the expectation under $\mathbb{P}^{x_0}$.

If $W = (W_t)_{t\geq 0}$ is a Brownian motion we will say “stopped Brownian motion" to designate the process $(W_{t\wedge\tau})_{t\geq 0}$, and “killed Brownian motion" to designate the process $(W_t \mathbb{1}_{\tau > T})_{t\in[0,T]}$, where τ is some stopping time defined later.

We will frequently use in the proofs the Bessel process $(R_t)_{t\geq 0}$ in dimension $\delta \geq 2$, i.e. a real-valued process satisfying the stochastic differential equation:

$$R_t = r_0 + B_t + \int_0^t \frac{\delta - 1}{2} \frac{ds}{R_s},$$

where $(B_t)_{t\geq 0}$ is a standard one-dimensional Brownian motion starting from zero. The index of the Bessel process is defined as $\nu = \frac{\delta}{2} - 1$. For a general reference to these matters we refer to [JYC09, Section 6].

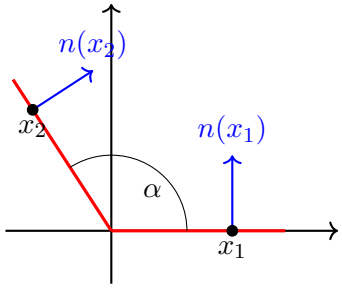


Figure 10.1: Example of a wedge of angle α.

10.3 Setting of the problem

Let $a \in \mathbb{R} \cup \{\pm\}$ and consider the subset $\mathcal{D} \subset \mathbb{R}^2$ defined by one of the four following cartesian equations:

$$\mathcal{D} = \begin{cases} \{(x,y) \in \mathbb{R}^2,\ y \geq 0 \text{ and } y \leq ax\} & \text{with } a > 0, \\ \{(x,y) \in \mathbb{R}^2,\ y \geq 0 \text{ and } y \geq ax\} & \text{with } a < 0, \\ \{(x,y) \in \mathbb{R}^2,\ y \geq 0 \text{ or } y \geq ax\} & \text{with } a > 0, \\ \{(x,y) \in \mathbb{R}^2,\ y \geq 0 \text{ or } y \leq ax\} & \text{with } a < 0. \end{cases} \tag{10.3.1}$$

The set $\mathcal{D}$ is called a wedge[1]. The angle of the wedge is denoted by α and chosen such that $\alpha \in (0, 2\pi)$. An example is given in Figure 10.1. We write the wedge $\mathcal{D}$ in polar coordinates as follows:

$$\mathcal{D} = \{(r\cos(\theta), r\sin(\theta)),\ r \in \mathbb{R}^+,\ \theta \in [0, \alpha]\}.$$

We define its boundary as $\partial\mathcal{D} = \partial\mathcal{D}^- \cup \partial\mathcal{D}^+$, where

$$\begin{aligned} \partial\mathcal{D}^- :=& \{(x,y) \in \mathcal{D}, y = 0\} = \{(r\cos(\theta), r\sin(\theta)),\ r \in \mathbb{R}^+,\ \theta = 0\}, \\ \partial\mathcal{D}^+ :=& \{(x,y) \in \mathcal{D}, y = ax\} = \{(r\cos(\theta), r\sin(\theta)),\ r \in \mathbb{R}^+,\ \theta = \alpha\}. \end{aligned}$$

In general, we use the notation $\mathcal{D} = \langle\alpha\rangle$ to denote a wedge which boundary is determined by the rays $\partial\mathcal{D}^\pm$. In a similar way, we also generalize this notation to a general wedge which boundaries are determined by the rays at the angles $0 \leq \alpha < \beta < 2\pi$ as $\mathcal{V} = \langle\alpha, \beta\rangle$.

We use the following stopping times:

$$\begin{aligned} \tau &= \inf\ \{t > 0, W_t \in \partial\mathcal{D}\}. \\ \tau^\pm &= \inf\ \{t > 0, W_t \in \partial\mathcal{D}^\pm\}. \end{aligned}$$

As in [Met10], we parametrize the covariance matrix of W via its Cholesky decomposition and write $\Sigma = \sigma\sigma^T$ with

$$\sigma = \begin{pmatrix} \sigma_1\sqrt{1-\rho^2} & \sigma_1\rho \\ 0 & \sigma_2 \end{pmatrix},$$

where $\rho \in (-1, 1)$ and $\sigma_1, \sigma_2 \geq 0$. Then assuming that σ is invertible, we consider the process $W' := \sigma^{-1}W$, which is a standard two-dimensional Brownian motion with independent components. Furthermore, using the explicit formula for the inverse matrix, we deduce that

$$W^2 = 0 \iff W'^2 = 0,$$

[1]The cases where $a = 0$ are not considered here because they lead to cases where the problem can be simplified to a one dimensional problem.

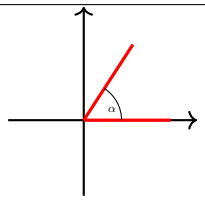
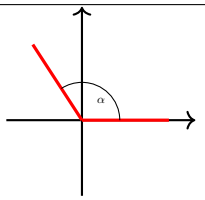
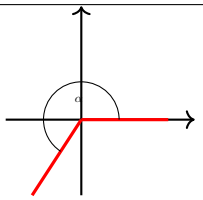
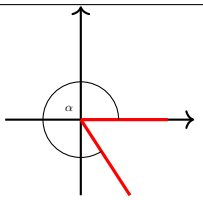
Cartesian Equation	$\{y \geq 0$ and $y \leq ax\}$, $a > 0$	$\{y \geq 0$ and $y \geq ax\}$, $a < 0$	$\{y \geq 0$ or $y \geq ax\}$, $a > 0$	$\{y \geq 0$ or $y \leq ax\}$, $a < 0$
Value of α	$\arctan(a)$	$\pi + \arctan(a)$	$\pi + \arctan(a)$	$2\pi + \arctan(a)$
Wedge for the original problem				
Equation of W' if $\sigma_2 - a\sigma_1\rho > 0$	$\{y \geq 0$ and $y \leq a'x\}$, $a' > 0$	$\{y \geq 0$ and $y \geq a'x\}$, $a' < 0$	$\{y \geq 0$ or $y \geq a'x\}$, $a' > 0$	$\{y \geq 0$ or $y \leq a'x\}$, $a' < 0$
Equation of W' if $\sigma_2 - a\sigma_1\rho < 0$	$\{y \geq 0$ and $y \geq a'x\}$, $a' < 0$	$\{y \geq 0$ and $y \leq a'x\}$, $a' > 0$	$\{y \geq 0$ or $y \leq a'x\}$, $a' < 0$	$\{y \geq 0$ or $y \geq a'x\}$, $a' > 0$

Figure 10.2: Table summing up the four different cases of wedge $\mathcal{D}$

$$W^2 = aW^1 \iff \begin{cases} W'^2 = \frac{a\sigma_1\sqrt{1-\rho^2}}{\sigma_2 - a\sigma_1\rho} W'^1 & \text{if } \sigma_2 - a\sigma_1\rho \neq 0, \\ W'^1 = 0 & \text{if } \sigma_2 - a\sigma_1\rho = 0. \end{cases}$$

In the case that $\sigma_2 - a\sigma_1\rho = 0$, then the problem reduces to two independent one-dimensional Brownian motions being reflected or stopped in the first quadrant, and therefore the problem can be reduced to using two independent reflected or stopped Brownian motions in one dimension. For explicit formulas, see (10.4.2) and (10.4.1). Thus in the following we assume that $\sigma_2 - a\sigma_1\rho \neq 0$. So τ^- is the first passage time of W' through the horizontal axis and τ^+ is the time of the first passage of W' through the line $y = a'x$, where

$$a' = \frac{a\sigma_1\sqrt{1-\rho^2}}{\sigma_2 - a\sigma_1\rho}, \tag{10.3.2}$$

so that $\tau = \tau^- \wedge \tau^+$ is the exit time of W' from the wedge $\mathcal{D}' = \langle \alpha' \rangle$, where α' is defined as $\arctan(a')$ or $\pi + \arctan(a')$ or $2\pi + \arctan(a)$ depending on the cases. The process W' starts at $x'_0 = \sigma^{-1} \cdot x_0$. The possible cases are summed up in Table 10.2.

After the above transformation the original wedge is transformed into a new one according to this table. Therefore mutatis-mutandis, we shall omit the primes in the notation and without loss of generality we directly assume that the process W has the correlation matrix $\Sigma = I_2$ and the wedge is given as in (10.3.1) with angle α.

In some results, we assume that the angle of the wedge satisfies the condition:

$$\alpha = \frac{\pi}{m},$$

for some $m \in \mathbb{N}$. We define adjacent wedges as follows. For $k = 0, 1, ..., 2m-1$, let

$$\mathcal{D}_k := \{(r\cos(\theta), r\sin(\theta)) : \ r \in [0, +\infty), \ \theta \in [k\alpha, (k+1)\alpha]\}.$$

We remark that $\mathcal{D}_0 = \mathcal{D}$ and that $\mathbb{R}^2 = \bigcup_{k=0}^{2m-1} \mathcal{D}_k$. We also define the transformation T_k:

$$\begin{aligned} T_k :& \mathcal{D}_0 \longrightarrow \mathcal{D}_k \\ T_k((r\cos\theta, r\sin\theta)) :=& (r\cos(\vartheta_k), r\sin(\vartheta_k)) \\ \vartheta_k :=& \begin{cases} (k+1)\alpha - \theta; & k \text{ odd}, \\ k\alpha + \theta; & k \text{ even}. \end{cases} \end{aligned}$$

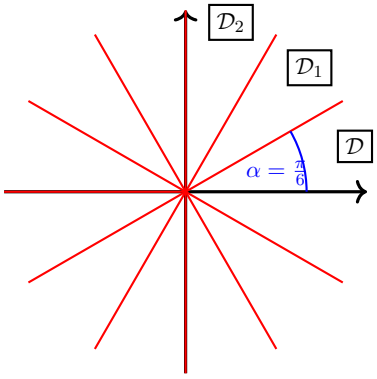


Figure 10.3: Partition of $\mathbb{R}^2$ using $\{\mathcal{D}_k\}_k$

From the definition, it follows that T_k is an isometric bijection between $\mathcal{D}_0$ and $\mathcal{D}_k$. This setting is represented in Figure 10.3.

The density of the killed Brownian motion and some related random quantities appear in [Met10]. We first quote the explicit density result for killed Brownian motion which is fully proven in [BnS97, Lemma 1] for generalized multidimensional cones.

Theorem 10.3.1 (Killed case)**.** *Assume that $\mathcal{D} = \langle \pi/m \rangle$ for some $m \in \mathbb{N}$. Then we have the following formula for the density of killed Brownian motion on the wedge $\mathcal{D}$:*

$$\forall t > 0,\ x, y \in \mathcal{D},\ \mathbb{P}^x(W_t \in dy,\ \tau > t) = \frac{1}{2\pi t} \sum_{k=0}^{2m-1} (-1)^k e^{-\frac{|x - T_k y|^2}{2t}} dy. \tag{10.3.3}$$

In general, let $\alpha \in (0, 2\pi)$, $\mathcal{D} = \langle \alpha \rangle$ and $x = (r_0 \cos(\theta_0), r_0 \sin(\theta_0)) \in \mathcal{D}$. Then for any $y = (r\cos(\theta), r\sin(\theta)) \in \mathcal{D}$, the density of the killed Brownian motion on the wedge $\mathcal{D}$ is given by

$$\mathbb{P}^x(W_t \in dy,\ \tau > t) = \frac{2r}{t\alpha} e^{-\frac{r^2 + r_0^2}{2t}} \sum_{n=1}^{\infty} I_{n\pi/\alpha}\left(\frac{r r_0}{t}\right) \sin\left(\frac{n\pi\theta}{\alpha}\right) \sin\left(\frac{n\pi\theta_0}{\alpha}\right) dr d\theta. \tag{10.3.4}$$

10.4 Analytic formulas for the density of the reflected process

In this section, we give explicit expressions for the densities of the reflected Brownian motion. Before doing this, we review the one dimensional case which will also help explain the motivation for the simulation methods to be introduced later. The one-dimensional case can be treated using the classic reflection principle, which directly gives the density of the killed and reflected Brownian motions. In the one-dimensional case, we can assume that $\mathcal{D} = [0, +\infty)$ and that $\Sigma = 1$. Then for all non-negative measurable functions $f : \mathcal{D} \to \mathbb{R}$ and $x_0 \geq 0$,

$$\mathbb{E}^{x_0}\left[f(W_T)\mathbb{1}_{\tau > T}\right] = \mathbb{E}^{x_0}\left[f(W_T)\mathbb{1}_{W_T > 0}\left(1 - e^{-\frac{x_0(x_0 + W_T)}{T}}\right)\right], \tag{10.4.1}$$

$$\mathbb{E}^{x_0}\left[f(X_T)\right] = \mathbb{E}^{x_0}\left[f(W_T)\mathbb{1}_{W_T > 0}\left(1 + e^{-\frac{x_0(x_0 + W_T)}{T}}\right)\right]. \tag{10.4.2}$$

The above formulas show that the simulation of the killed and reflected Brownian motions in the one-dimensional case can be easily performed using changes of measures on the original unrestricted Brownian motion.

The proof in the reflected case in $\mathbb{R}^2$ shares common arguments with the killed case. Throughout these arguments, one uses results from classical theory of partial differential equations on

wedges. For a general reference on this topic, we refer the reader to [LSU68, chapter IV] and [KZ16] where one can find existence and uniqueness results for the partial differential equations that are related to the problems in this article.

Definition 10.4.1. *We define the normal reflection process of a continuous adapted stochastic process ξ on the subset $\mathcal{D}$, as the unique solution $(X_t, L_t)_{t\geq 0}$ of the following problem (see for example [Pil14]):*

1. *$X_t = \xi_t + L_t$, where $L = (L_t)_{t\geq 0}$ is some adapted process.*
2. *$\forall t \geq 0,\ X_t \in \mathcal{D}$.*
3. *L is a continuous process of bounded variation and $L_0 = 0$.*
4. *$\forall t \geq 0,\ L_t = \int_0^t n(X_s)\mathbb{1}_{X_s\in\partial\mathcal{D}} d|L|_s$, which means that the process L increases only when $X_t \in \partial\mathcal{D}$ and that the reflection is normal to the boundary.*

Therefore the normally reflected Brownian motion is the process that is obtained from the above equation by taking $\xi = W$. In the one dimensional case, L corresponds to the local time of Brownian motion.

The existence and uniqueness of a Brownian motion reflected in a wedge, which is a non-smooth domain, has been proved in the decorrelated case in [VW85]. In this article we will only consider normal reflections in the case of unitary diffusion matrix. Note that if we use the decorrelation step as in (10.3.2) for the reflected Brownian motion, this transformation changes the normal reflected process into an obliquely reflected process. We do not know how to obtain the density of such a process.

Theorem 10.4.2 (Reflected case)**.** *Assume that $\mathcal{D} = \langle \pi/m \rangle$ for some $m \in \mathbb{N}$. Then we have the following formulas for the density of the reflected Brownian motion on the wedge $\mathcal{D} = \langle \pi/m \rangle$:*

$$\forall t > 0,\ x, y \in \mathcal{D},\ \mathbb{P}^x(X_t \in dy) = \frac{1}{2\pi t} \sum_{k=0}^{2m-1} e^{-\frac{|x-T_k y|^2}{2t}} dy. \tag{10.4.3}$$

In the general case for $\alpha \in (0, 2\pi)$, $\mathcal{D} = \langle \alpha \rangle$ and $x = (r_0\cos(\theta_0), r_0\sin(\theta_0)) \in \mathcal{D}$ we have that for any $y = (r\cos(\theta), r\sin(\theta)) \in \mathcal{D}$, the density of the reflected Brownian motion on the wedge $\mathcal{D} = \langle \alpha \rangle$ is given by

$$\mathbb{P}^x(X_t \in dy) = \frac{2r}{t\alpha} e^{-\frac{r^2+r_0^2}{2t}} \left(\frac{1}{2} I_0\left(\frac{rr_0}{t}\right) + \sum_{n=1}^{\infty} I_{n\pi/\alpha}\left(\frac{rr_0}{t}\right) \cos\left(\frac{n\pi\theta}{\alpha}\right) \cos\left(\frac{n\pi\theta_0}{\alpha}\right) \right) dr d\theta. \tag{10.4.4}$$

The proof of the above theorem is given in 10.10. The idea of using the simple case $\mathcal{D} = \langle \pi/m \rangle$ to infer a general result appears in the literature of the image method. It has been used in a non-trivial way in order to deduce stationary measures for reflected processes (see e.g. [DM09]). A formula similar to (10.4.4) appears in [CJ59, page 379 (8)], which is proved using Laplace transform inversion methods. We have decided to include a proof which uses the image method because it gives some intuition on the simulations methods to be introduced later.

10.5 Exact simulation algorithms

Theorems 10.3.1 and 10.4.2 give formulas that can be directly used to simulate the final values of the reflected and stopped processes in the case that the wedge angle is $\alpha = \pi/m$ (see Sections

10.5.2 and 10.5.3). However, direct algorithms that arise due to these results demand the use of approximations and larger computational time in general. Instead, we investigate an alternative simulation method using the tractable case $\alpha = \pi/m$ and give algorithms for the exact simulation of the reflected and stopped processes in any wedge.

10.5.1 Formulas for the simulation of (τ, W_τ) in the case $\alpha = \pi/m$

Before providing the simulation method, we give a formula that will be the key to provide a simplified simulation method which avoids the calculation of Bessel functions.

Theorem 10.5.1. *Let $\mathcal{V} = \langle \alpha^-, \alpha^+ \rangle$ and assume that $\alpha^+ - \alpha^- = \pi/m =: \alpha$ for some $m \in \mathbb{N}$ and let τ be the hitting time on the wedge $\mathcal{V}$ for W which starts at $x_0 = (r_0\cos(\theta_0), r_0\sin(\theta_0)) \in \mathcal{V}$, then*

$$\mathbb{P}^{x_0}(\tau \in dt,\ W_\tau \in dy^\pm) = \frac{r_0}{2\pi t^2} e^{-\frac{r^2+r_0^2}{2t}} \sum_{k=0}^{m-1} \sin\left(\gamma_k^\pm\right) e^{\frac{rr_0}{t}\cos(\gamma_k^\pm)} dr dt, \tag{10.5.1}$$

where $y^\pm = (r\cos(\alpha^\pm), r\sin(\alpha^\pm)) \in \partial\mathcal{V}^\pm$. We use the notation

$$\gamma_k^\pm := \pm\alpha^\pm \pm \frac{2k\pi}{m} \mp \theta_0. \tag{10.5.2}$$

Proof. We can assume that $\alpha^- = 0$ without loss of generality. We use the formulas [Met10, (1.5), (1.6)]. If $y \in \partial\mathcal{D}^-$, writing $y = (r, 0)$ for some $r > 0$, we have:

$$\mathbb{P}^{x_0}(\tau \in dt,\ W_\tau \in dy) = \frac{\pi}{\alpha^2 tr} e^{-\frac{r^2+r_0^2}{2t}} \sum_{n=1}^{\infty} n\sin\left(\frac{n\pi\theta_0}{\alpha}\right) I_{n\pi/\alpha}\left(\frac{rr_0}{t}\right) dr dt. \tag{10.5.3}$$

Using (10.3.4) and switching the order of derivation and integration, which can be justified (see the proof of Theorem 10.4.2 with Corollary 10.10.1 in the Appendix), we obtain for $y \in \partial\mathcal{D}^-$:

$$\mathbb{P}^{x_0}(\tau \in dt,\ W_\tau \in dy) = \frac{1}{2r^2}\left.\frac{\partial}{\partial\theta}\right|_{\theta=0} \mathbb{P}^{x_0}(\tau > t, W_t \in d(r\cos(\theta), r\sin(\theta))).$$

So that in the case $\alpha = \pi/m$, using (10.3.3) in polar coordinates and the definition of ϑ_k:

$$\begin{aligned}\mathbb{P}^{x_0}(\tau \in dt,\ W_\tau \in dy) &= \frac{1}{2r^2}\left.\frac{\partial}{\partial\theta}\right|_{\theta=0} \frac{r}{2\pi t} e^{-\frac{r^2+r_0^2}{2t}} \sum_{k=0}^{2m-1} (-1)^k e^{\frac{1}{t}rr_0\cos(\theta_0-\vartheta_k)} dr dt \\ &= \frac{r_0}{2\pi t^2} e^{-\frac{r^2+r_0^2}{2t}} \sum_{k=0}^{m-1} \sin(\theta_0 - 2k\alpha) e^{\frac{rr_0}{t}\cos(\theta_0-2k\alpha)} dr dt. \end{aligned} \tag{10.5.4}$$

Likewise, we obtain a similar formula for the case $y \in \partial\mathcal{D}^+$. □

From here, we present a method to simulate τ and W_τ in the case of the wedge $\mathcal{V} = \langle \alpha^-, \alpha^+ \rangle$ with $\alpha^+ - \alpha^- = \pi/m$. For the simulation of W_τ, we use the same method which is proposed in [Met09, Section 2.1] but we restrict to the case $\alpha = \pi/m$. On the other hand, we propose an unbiased simulation method for τ in comparison with the method proposed in [Met09] which has a bias difficult to estimate (see [Met09, Proposition 2.1.8] and the remark that follows).

Simulation of W_τ: First, we simulate the selection of the boundary line on which W_τ arrives, with a Bernoulli distribution. These two boundaries are denoted by $\partial\mathcal{V}^\pm = \{(r\cos(\alpha^\pm), r\sin(\alpha^\pm)) : r \geq 0\}$. Following [Met09, Corollary 2.1.5] we have:

$$\mathbb{P}^{x_0}(W_\tau \in \mathcal{V}^+) = \frac{\theta_0 - \alpha^-}{\alpha^+ - \alpha^-}, \tag{10.5.5}$$

We simulate on which frontier W_τ arrives first and then we directly simulate W_τ, without simulating τ yet. Following [Met09, Proposition 2.1.6] one can compute explicit formulas for the distribution function of the radius of exit given that it exits at $\mathcal{V}^+$ or $\mathcal{V}^-$. Also its inverse can be computed and therefore the inverse transformation method for simulation can be applied. Using these formulas, we can simulate the radius r_τ as:

$$r_\tau = \begin{cases} r_0 \left(\cos\left(\frac{\pi\theta_0}{\alpha}\right) - \frac{\sin\left(\frac{\pi\theta_0}{\alpha}\right)}{\tan((\pi - \pi\theta_0/\alpha)(U-1))} \right)^{\alpha/\pi} & \text{if } W_\tau \in \mathcal{V}^-, \\ r_0 \left(-\cos\left(\frac{\pi\theta_0}{\alpha}\right) - \frac{\sin\left(\frac{\pi\theta_0}{\alpha}\right)}{\tan((\pi\theta_0/\alpha)(U-1))} \right)^{\alpha/\pi} & \text{if } W_\tau \in \mathcal{V}^+, \end{cases} \tag{10.5.6}$$

with $U \sim \mathcal{U}([0,1])$.

Simulation of τ: Knowing W_τ, we simulate τ according to the conditional formula (see (10.5.4) and recall that $dy^\pm = (dr\cos(\alpha^\pm), dr\sin(\alpha^\pm))$):

$$\mathbb{P}^{x_0}(\tau \in dt |\ W_\tau = y^\pm) = \left((\mathbb{P}^{x_0}(W_\tau \in dy^\pm))^{-1} dr \right) \frac{r_0}{2\pi t^2} \sum_{k=0}^{m-1} \sin(\gamma_k^\pm) \cdot \exp\left(-\frac{(r - r_0\cos(\gamma_k^\pm))^2 + r_0^2 \sin^2(\gamma_k^\pm)}{2t} \right) dt. \tag{10.5.7}$$

Since some of the coefficients $\sin(\gamma_k^\pm)$ are negative, we cannot directly simulate according to this distribution as a mixture. Instead we select the indexes $k_0^\pm, \ldots, k_{p^\pm}^\pm$ such that for all i, $\sin(\gamma_{k_i^\pm}^\pm) \geq 0$ and write:

$$\mathbb{P}^{x_0}(\tau \in dt |\ W_\tau = y^\pm) \leq \left((\mathbb{P}^{x_0}(W_\tau \in dy^\pm))^{-1} dr \right) \frac{r_0}{2\pi t^2} \sum_{i=0}^{p^\pm} \sin(\gamma_{k_i^\pm}^\pm) \cdot \exp\left(-\frac{(r - r_0\cos(\gamma_{k_i^\pm}^\pm))^2 + r_0^2 \sin^2(\gamma_{k_i^\pm}^\pm)}{2t} \right) dt. \tag{10.5.8}$$

This is the distribution of a discrete mixture of random variables ξ_k for $0 \leq k \leq p^\pm$, where ξ_k is the inverse of an exponential random variable of parameter $((r - r_0\cos(\gamma_{k_i^\pm}^\pm))^2 + r_0^2\sin^2(\gamma_{k_i^\pm}^\pm))/2$. With this information, we build an acceptance-rejection sampling, based on the density on the right hand side of (10.5.8) and accept the sample value t under the condition

$$U < \left(\sum_{k=0}^{m-1} \sin(\gamma_k^\pm) \exp\left(-\frac{(r - r_0\cos(\gamma_k^\pm))^2 + r_0^2\sin^2(\gamma_k^\pm)}{2t} \right) \right) \times \left(\sum_{i=0}^{p^\pm} \sin(\gamma_{k_i^\pm}^\pm) \exp\left(-\frac{(r - r_0\cos(\gamma_{k_i^\pm}^\pm))^2 + r_0^2\sin^2(\gamma_{k_i^\pm}^\pm)}{2t} \right) \right)^{-1},$$

with $U \sim \mathcal{U}([0,1])$.

Simulation in the case W does not exit the wedge: So far, we have provided a methodology in order to simulate the exit location and time for the simpler wedge $\mathcal{V} = \langle \alpha^-, \alpha^+ \rangle$ with $\alpha^+ - \alpha^- = \pi/m$. Now we discuss the case when W does not leave this wedge in the time interval $[0,T]$. To simulate the final value W_T of the Brownian motion starting from x_0 conditionally to the fact that it does not exit the wedge $\mathcal{V}$ on $[0,T]$, we need to simulate according to the density (10.3.3), which is a sum where some terms are negative. We propose the following method.

A slight generalization of (10.3.3) to the wedge $\mathcal{V} = \langle \alpha^-, \alpha^+ \rangle$ with

$$\widetilde{\vartheta}_k := \begin{cases} \vartheta_k + 2\alpha^- & \text{if } k \text{ odd} \\ \vartheta_k & \text{if } k \text{ even,} \end{cases}$$

gives, for $y \in \mathcal{V}$:

$$\begin{aligned} \mathbb{P}^{x_0}(W_T \in dy |\ \tau > T) &= \frac{\mathbb{P}^{x_0}(\tau > T)^{-1}}{2\pi T} \sum_{k=0}^{2m-1} (-1)^k e^{-\frac{|x_0 - \widetilde{T}_k y|^2}{2T}} dy \\ &\leq \frac{\mathbb{P}^{x_0}(\tau > T)^{-1}}{2\pi T} \sum_{k=0}^{m-1} e^{-\frac{|x_0 - \widetilde{T}_{2k} y|^2}{2T}} dy, \end{aligned} \tag{10.5.9}$$

where $\widetilde{T}_k(r\cos(\theta), r\sin(\theta)) = (r\cos(\widetilde{\vartheta}_k), r\sin(\widetilde{\vartheta}_k))$. We then apply acceptance-rejection sampling with the reference density proportional to $\frac{1}{2\pi T}\sum_{k=0}^{m-1} e^{-\frac{|x_0 - \widetilde{T}_{2k} y|^2}{2T}}$ and accept the sample y under the condition

$$U < \left(\sum_{k=0}^{2m-1} (-1)^k e^{-\frac{|x_0 - \widetilde{T}_k y|^2}{2T}} \mathbb{1}_{y \in \mathcal{V}} dy \right) \left(\sum_{k=0}^{m-1} e^{-\frac{|x_0 - \widetilde{T}_{2k} y|^2}{2T}} \right)^{-1}, \tag{10.5.10}$$

with $U \sim \mathcal{U}([0,1])$.

10.5.2 Algorithm for the simulation of the stopped Brownian motion: General case

In this subsection, we give a recursive algorithm to simulate the final value $W_{T\wedge\tau}$ of the stopped process on the wedge $\mathcal{D}$, for any angle $\alpha \in (0, 2\pi)$. This algorithm will be used for the simulation algorithm of the reflected Brownian motion, but it also has its own interest as it can be used as itself for the simulation of the stopped Brownian motion. In what follows, we denote $\mathcal{F}_n$ the sigma algebra generated by the algorithm until the step n, i.e. $\mathcal{F}_n = \sigma(\tau_1, y_1, \ldots, \tau_n, y_n)$, where $\tau_1, \ldots, \tau_n$ and $y_1, \ldots, y_n$ are random variables generated by the algorithm. We also denote by $\mathbb{P}_n$ and $\mathbb{E}_n$ the conditional probability and expectation with respect to $\mathcal{F}_n$.

First, we propose an algorithm for the case where the angle is π/m. At the step n of the algorithm we will define a wedge $\mathcal{V}_n \subset \mathcal{D}$ of angle π/m for some $m \in \mathbb{N}$ and then simulate the process stopped on this wedge. Define $\Theta := \max(\{\pi/m,\ m \in \mathbb{N}\} \cap (0, \alpha])$. Note that since $\alpha \in (0, 2\pi)$, we have $\alpha/2 < \Theta \leq \alpha$.

Algorithm I: Start at the point $x_0 = (r_0\cos(\theta_0), r_0\sin(\theta_0)) =: y_0$ at time $T_0 = 0$. The algorithm follows these steps for $n \in \mathbb{N}$:

1. Given $y_n = (r_n\cos(\theta_n), r_n\sin(\theta_n)) \in \mathcal{V}_n$, define the angles:

$$\beta_{n+1}^- := \begin{cases} 0 & \text{if } \theta_n \in \left[0, \frac{\Theta}{2}\right] \\ \theta_n - \frac{\Theta}{2} & \text{if } \theta_n \in \left[\frac{\Theta}{2}, \alpha - \frac{\Theta}{2}\right] \\ \alpha - \Theta & \text{if } \theta_n \in \left[\alpha - \frac{\Theta}{2}, \alpha\right] \end{cases}$$
$$\beta_{n+1}^+ := \beta_{n+1}^- + \Theta.$$

   By the definition of Θ, we have $\Theta \leq \alpha$ and therefore $0 \leq \Theta/2 \leq \alpha - \Theta/2 \leq \alpha$.

2. Consider the wedge $\mathcal{V}_{n+1} = \langle \beta_{n+1}^-, \beta_{n+1}^+ \rangle$ of angle Θ, which satisfies $\mathcal{V}_{n+1} \subset \mathcal{D}$ and $y_n = (r_n\cos(\theta_n), r_n\sin(\theta_n)) \in \mathcal{V}_{n+1}$.

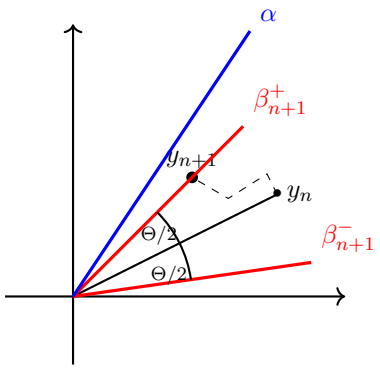


Figure 10.4: Example of the domains of simulation at step $n+1$.

3. Simulate on which one of the two sets $\partial\mathcal{V}^+_{n+1}$ or $\partial\mathcal{V}^-_{n+1}$, the process $(W_t)_{t\geq T_n}$ starting from y_n, at time T_n, exits the wedge $\mathcal{V}_{n+1}$ and define $\theta_{n+1} := \beta^{\pm}_{n+1}$ according to the simulation result. This is done using a Bernoulli variable and the formula (10.5.5).

4. Simulate $y_{n+1} \in \partial\mathcal{V}_{n+1}$, which is the value of W when it reaches $\partial\mathcal{V}_{n+1}$ for the first time after starting at y_n at time T_n, and knowing which one of the two events $y_{n+1} \in \partial\mathcal{V}^{\pm}_{n+1}$ has occurred, using formula (10.5.6). Denote the exit point as

$$y_{n+1} = (r_{n+1}\cos(\theta_{n+1}), r_{n+1}\sin(\theta_{n+1})).$$

5. Simulate τ_{n+1}, the time for the process W starting at $y_n = (r_n\cos(\theta_n), r_n\sin(\theta_n))$ at time T_n to reach $\partial\mathcal{V}_{n+1}$, knowing the event $W_{T_n+\tau_{n+1}} = y_{n+1}$. This can be done using the formula (10.5.7) and the acceptance-rejection procedure described after it. Define

$$T_{n+1} := T_n + \tau_{n+1},$$

so that $W_{T_{n+1}} = y_{n+1}$.

If $T_n + \tau_{n+1} < T$ then the algorithm iterates. This algorithm stops under one of these two conditions:

- Condition 1: $T_n + \tau_{n+1} \geq T$. Then simulate the final value W_T knowing that $W_{T_n} = y_n$ and that for $t \in [T_n, T]$, $W_t \in \mathcal{V}_n$. We do this simulation conditionally to the event $T_n + \tau_{n+1} > T$ and "forget" the exact simulated value of τ_{n+1} as well as the value of $W_{T_n+\tau_{n+1}}$. This is justified in Proposition 10.9.1, using $X := (W_{T\wedge T_{n+1}}, T_{n+1}\mathbb{1}_{T_{n+1}<T})_{|\mathscr{F}_n}$, $Y := (W_{T_{n+1}}, T_{n+1})_{|\mathscr{F}_n}$ and $A = \mathbb{R}^2 \times [0,T]$. We simulate the value of W_T using (10.5.9) and the acceptance-rejection procedure described after it.

- Condition 2: $\theta_{n+1} = 0$ or $\theta_{n+1} = \alpha$. Then we obtain $\tau := T_{n+1} < T$, i.e. the Brownian motion reaches the wedge $\mathcal{D}$ before the time T and the result of the simulation is

$$W_{T\wedge\tau} = (r_{n+1}\cos(\theta_{n+1}), r_{n+1}\sin(\theta_{n+1})).$$

An illustration of this algorithm is given in Figure 10.4. Next, we prove that this algorithm can be carried out in a finite number of steps.

Proposition 10.5.2. *Algorithm I ends in finite time, i.e. the number of required iterations to finish the algorithm, denoted by N, is such that $N < \infty$ almost surely. More precisely, for all $K \in \mathbb{N}$, $\mathbb{P}^{x_0}(N \geq K) \leq 2^{-\lfloor K/2 \rfloor}$.*

Proof. Fix $K \in \mathbb{N}$. For a wedge $\mathcal{V}_n$, $n \leq N$ in Algorithm I, we say that "$\mathcal{V}_n$ intersects the boundary of $\mathcal{D}$" if $\beta_n^+ = \alpha$ or $\beta_n^- = 0$. For $n \leq N$, we have two possible cases:

- $\mathcal{V}_n$ does not intersect the boundary of $\mathcal{D}$. Then, using the fact that $\Theta > \frac{\alpha}{2}$ and the definition of $\beta_{n+1}^{\pm}$, $\mathcal{V}_{n+1}$ intersects the boundary of $\mathcal{D}$, and θ_n is closer to $\beta_{n+1}^{\pm}$ than $\beta_{n+1}^{\mp}$, where here the sign $\pm$ is such that $\partial\mathcal{V}_{n+1}^{\pm} \cap \partial\mathcal{D} \neq \{0\}$. So using (10.5.5), $\mathbb{P}_n(y_{n+1} \notin \partial\mathcal{D}) \leq 1/2$.
- $\mathcal{V}_n$ intersects the boundary of $\mathcal{D}$. Then if $\theta_{n+1} \notin \{0, \alpha\}$, $\mathcal{V}_{n+1}$ does not intersect the boundary of $\mathcal{D}$ and by the same reasoning as above $\mathbb{P}_n(y_{n+2} \notin \partial\mathcal{D}) \leq 1/2$.

Then, by tower property of conditional expectations, we have:

$$\mathbb{P}^{y_0}(N \geq K) \leq \mathbb{P}^{y_0}(y_1 \notin \partial\mathcal{D}, \dots, y_K \notin \partial\mathcal{D}) \leq 2^{-\lfloor \frac{K}{2} \rfloor}.$$

So that:

$$\mathbb{P}^{y_0}(N = \infty) = \lim_{K \to \infty} \mathbb{P}^{y_0}(N \geq K) = 0.$$

□

10.5.3 Algorithm for the simulation of the reflected Brownian motion

We now state an algorithm to simulate the final value of the reflected Brownian motion X_T in a wedge $\mathcal{D} = \langle \alpha \rangle$ of any angle $\alpha \in (0, 2\pi)$.

Algorithm II: Choose $\Theta := \max(\{\pi/m,\ m \in \mathbb{N}\} \cap (0, \alpha])$ as before. Start at the point $x_0 = (r_0 \cos(\theta_0), r_0 \sin(\theta_0)) =: y_0$ at time $T_0 = 0$. In general, suppose that for $n \in \mathbb{N}$ the point $y_n = (r_n \cos(\theta_n), r_n \sin(\theta_n))$ has already been simulated.

1. Define the angles:
$$\beta_{n+1}^- := \theta_n - \frac{\Theta}{2},$$
$$\beta_{n+1}^+ := \theta_n + \frac{\Theta}{2} = \beta_{n+1}^- + \Theta.$$

2. Consider the wedge $\mathcal{V}_{n+1} := \langle \beta_{n+1}^-, \beta_{n+1}^+ \rangle$. Note that although $y_n \in \mathcal{V}_{n+1}$, we do not necessarily have that $\mathcal{V}_{n+1} \subset \mathcal{D}$.

3. Simulate the random variables τ_{n+1} and z_{n+1}, which are respectively the time and the final point of a Brownian motion $Z = (Z_t)_{t \geq 0}$ starting at y_n at time 0 and stopped on the wedge $\mathcal{V}_{n+1}$ using the steps 3-5 of Algorithm I in Section 10.5.2. Note that, since θ_n is in the middle of the wedge $\mathcal{V}_{n+1}$, we have by symmetry $\mathbb{P}^{z_n}(Z_{\tau_{n+1}} \in \partial\mathcal{V}_{n+1}^{\pm}) = 1/2$.

4. If $T_n + \tau_{n+1} < T$, then we define r_{n+1} and $\widetilde{\theta}_{n+1}$ to be respectively the radius and the angle of z_{n+1}. Since it is possible that $z_{n+1} \notin \mathcal{D}$, we define θ_{n+1} as follows:
$$\theta_{n+1} = \begin{cases} \widetilde{\theta}_{n+1} & \text{if } \widetilde{\theta}_{n+1} \in [0, \alpha], \\ -\widetilde{\theta}_{n+1} & \text{if } \widetilde{\theta}_{n+1} = \beta_{n+1}^- \text{ and } \beta_{n+1}^- < 0, \\ 2\alpha - \widetilde{\theta}_{n+1} & \text{if } \widetilde{\theta}_{n+1} = \beta_{n+1}^+ \text{ and } \beta_{n+1}^+ > \alpha. \end{cases}$$
And then we define $y_{n+1} := (r_{n+1} \cos(\theta_{n+1}), r_{n+1} \sin(\theta_{n+1}))$, so that $y_{n+1} \in \mathcal{D}$. Note that if $z_{n+1} \notin \mathcal{D}$, then y_{n+1} is the reflection of z_{n+1} with respect to the line $\{\theta = 0\}$ in the second case or $\{\theta = \alpha\}$ in the third case in the definition of θ_{n+1}. Then define the time:
$$T_{n+1} = T_n + \tau_{n+1}$$
and the algorithm iterates.

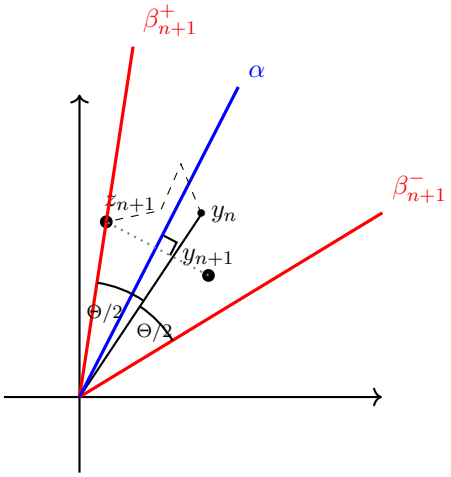


Figure 10.5: Example of domains of simulation at the step $n+1$ in the reflected case.

5. If $T_n + \tau_{n+1} > T$, we simulate $z_{n+1} \in \mathcal{V}_{n+1}$ as the value at time $T - T_n$ of a Brownian motion starting at y_n and conditionally to the fact that it stays in the wedge $\mathcal{V}_{n+1}$ in the time interval $[0, T - T_n]$. This can be done using the acceptance-rejection method given in (10.5.9) and what follows based on Proposition 10.9.1 as it was done in the termination Condition 1 in Algorithm I in Section 10.5.2. Then we define y_{n+1} in function of z_{n+1} as in step 4. Finally, we stop the algorithm, and the resulting value of the simulation is y_{n+1}.

An illustration of this algorithm is given in Figure 10.5.

Proposition 10.5.3. *Define* $N := \inf\{n \in \mathbb{N},\ \tau_1 + \ldots + \tau_n > T\}$. *Then* $N < \infty$ *almost surely. That is, the algorithm terminates in finite time.*

Proof. The stopping times $(\tau_i)_i$ follow the same law as $(\widetilde{\tau}_i)$, defined by $\widetilde{\tau}_0 = 0$ and:

$$\widetilde{\tau}_{i+1} := \inf\left\{t > 0 :\ |\theta(B_{t+\widetilde{\tau}_i}) - \theta(B_{\widetilde{\tau}_i})| \geq \frac{\pi}{2m}\right\}, \tag{10.5.11}$$

where $m = \pi/\Theta \in \mathbb{N}$, $B = (B_t)_{t\geq 0}$ is a standard two-dimensional Brownian motion and $\theta(B_t)$ denotes the angle of the process[2] B. Then:

$$\mathbb{P}^{x_0}\left(\sum_{i=1}^{\infty} \tau_i < \infty\right) = \mathbb{P}^{x_0}\left(\sum_{i=1}^{\infty} \widetilde{\tau}_i < \infty\right).$$

But, if $\mathcal{T} := \sum_{i=1}^{\infty} \widetilde{\tau}_i < \infty$, then there exist two random sequences, (t^1_n) and (t^2_n), increasing and converging to $\mathcal{T}$, such that $\theta(B_{t^i_n}) = k^i_n \pi/(2m) + \theta_0$ for all $n \in \mathbb{N}$ and $i \in \{1, 2\}$, and where $k^i_n \in \mathbb{Z}$ has the same parity as i, so that the difference between $\theta(B_{t^1_n})$ and $\theta(B_{t^1_{n'}})$ is at least $\pi/(2m)$ for all $n, n' \in \mathbb{N}$. Taking $n \to \infty$, we have necessarily that $B_{\mathcal{T}} = 0$, which occurs with probability 0 on any closed time interval. □

10.6 Folding number in a wedge and complexity

In this section we study the complexity of Algorithms I and II in Sections 10.5.2 and 10.5.3 in separate cases. Recall that N denotes the number of iterations that an algorithm requires to

[2]We choose the angle of the process $(\theta(B_t))_{t\geq 0}$ so that it is continuous with respect to t.

finish one simulation, i.e.

$$N := \inf\{n \in \mathbb{N} : \ \tau_1 + \ldots + \tau_n > T\}.$$

For Algorithm I, N is the number of times we touch the boundaries of the corresponding sets $\mathcal{V}_n$, $n \in \mathbb{N}$, before touching one of the boundaries of $\mathcal{D}$ or reaching time T. For Algorithm II, N is the number of times that a Brownian motion reflected in a wedge $\mathcal{V}$ of angle $\pi/(2m)$ goes from one boundary $\partial\mathcal{V}^{\pm}$ to another $\partial\mathcal{V}^{\mp}$ until time T. For this reason, we may call N the number of folds for the simulation of the reflected Brownian motion.

In this section, we give theoretical properties and bounds for N for each algorithm separately.

10.6.1 Majoration of N for the simulation algorithm of the stopped Brownian motion

For Algorithm I, by Proposition 10.5.2, we have for all $K \in \mathbb{N}$, $\mathbb{P}^{x_0}(N \geq K) \leq 2^{-\lfloor K/2 \rfloor}$, so

$$\mathbb{E}^{x_0}[N] = \sum_{K=1}^{\infty} \mathbb{P}^{x_0}(N \geq K) \leq \sum_{K=1}^{\infty} 2^{-\lfloor \frac{K}{2} \rfloor} = 3.$$

So Algorithm I ends in finite time almost surely and its complexity is finite in expectation.

10.6.2 Majoration of the number of folds of the Brownian motion in a wedge

We now investigate the complexity for Algorithm II. We decompose the standard two-dimensional Brownian motion $W = (W_t)_{t\geq 0}$ in polar coordinates and introduce the notations that will be used in this section and the next one as follows. The successive stopping times $(\tau_i)_i$ which appear in Algorithm II, have the same law as $(\tilde{\tau}_i)_i$, defined as in the proof of Proposition 10.5.3. Then we have

$$N \overset{\mathscr{L}}{=} \inf\{n \in \mathbb{N} : \ \tilde{\tau}_1 + \ldots + \tilde{\tau}_n > T\}.$$

Moreover, using the skew-product representation of the Brownian motion (see [RY99, page 194]):

$$B_t = R_t U_{F(t)}, \tag{10.6.1}$$

where $(R_t)_{t\geq 0} = (|B_t|)_{t\geq 0}$ is a Bessel process of dimension 2, $(U_t)_{t\geq 0}$ is a Brownian motion on $\mathbb{S}^1 \subset \mathbb{R}^2$ and F is defined as

$$F(t) = \int_0^t \frac{ds}{R_s^2}, \tag{10.6.2}$$

which is a strictly increasing process almost surely. Moreover, the processes $(R_t)_{t\geq 0}$ and $(U_t)_{t\geq 0}$ are independent. If we let $(\theta(U_t))_{t\geq 0}$ be the angle of the process $(U_t)_{t\geq 0}$ then $(\theta(U_t))_{t\geq 0}$ is a standard one-dimensional Brownian motion.

Define the stopping times $(s_i)_i$ by $s_0 = 0$ and

$$s_{i+1} = \inf\{t > 0 : \ |\theta(U_{t+s_i}) - \theta(U_{s_i})| \geq \pi/(2m)\}.$$

Then, the random variables s_i are i.i.d. and have the law of the time for one-dimensional Brownian motion starting at 0 to reach the double barrier $\pm\pi/(2m)$. Then using (10.5.11) we have that for all $K \in \mathbb{N}$,

$$\sum_{i=1}^{K} s_i \overset{\mathscr{L}}{=} F\left(\sum_{i=1}^{K} \tilde{\tau}_i\right).$$

So the estimation of N can be simplified as

$$\mathbb{P}^{x_0}(N \geq K) = \mathbb{P}^{x_0}\left(\sum_{i=1}^{K} \tilde{\tau}_i \leq T\right) = \mathbb{P}^{x_0}\left(\sum_{i=1}^{K} s_i \leq F(T)\right). \tag{10.6.3}$$

Note that F is a random change of time, independent of $(s_i)_i$. Furthermore, the Laplace transform of s_i is explicitly given in [JYC09, Section 3.5, Proposition 3.5.1.3] as

$$\mathbb{E}[e^{-\lambda s_1}] = \frac{2}{e^{\frac{\sqrt{2}\pi\sqrt{\lambda}}{2m}} + e^{-\frac{\sqrt{2}\pi\sqrt{\lambda}}{2m}}}.$$

In particular, we have:

$$(\mathbb{E}[s_i], \mathrm{Var}(s_i)) = \left(\frac{\pi^2}{4m^2}, \frac{2}{3}\left(\frac{\pi}{2m}\right)^4\right) =: (\mu, \sigma^2).$$

Proposition 10.6.1. *Let $0 < a < \frac{1}{2}$. Then we have $\mathbb{E}^{x_0}[N^a] < \infty$. On the other hand, if $a \geq 1/2$ then $\mathbb{E}^{x_0}[N^a] = \infty$.*

Equation (10.6.3) together with the estimate in (10.9.3) give the result which is a consequence of the fact that the tails of $F(T)$ are large because the probability that the Bessel process R is close to zero is large and therefore due to (10.6.1) the simulation may spend a large time close to the origin.

Proof. We have that for all $K \in \mathbb{N}$ and $\delta > 1$:

$$\begin{aligned}
\mathbb{P}^{x_0}(N^a \geq K) &= \int_0^{\frac{\mu K^{1/a}}{\delta}} \mathbb{P}^{x_0}\left(\sum_{i=1}^{\lceil K^{1/a} \rceil} s_i \leq y\right) \mathbb{P}^{x_0}\left(F(T) \in dy\right) \\
&\quad + \int_{\frac{\mu K^{1/a}}{\delta}}^{\infty} \mathbb{P}^{x_0}\left(\sum_{i=1}^{\lceil K^{1/a} \rceil} s_i \leq y\right) \mathbb{P}^{x_0}\left(F(T) \in dy\right) \\
&=: \mathcal{I}_1(K) + \mathcal{I}_2(K).
\end{aligned}$$

For $\mathcal{I}_1(K)$, we will use the fact that

$$2\exp\left(\frac{\pi^2}{4\delta m^2}\right) < \exp\left(\frac{\sqrt{2}\pi}{2m}\right) + \exp\left(-\frac{\sqrt{2}\pi}{2m}\right)$$

for δ large enough. Then, we obtain that $\mathcal{I}_1(K)$ decreases exponentially fast to zero. In fact, using Markov inequality

$$I_1(K) \leq \mathbb{P}^{x_0}\left(\sum_{i=1}^{\lceil K^{1/a} \rceil} s_i \leq \frac{\mu K^{1/a}}{\delta}\right) \leq e^{\frac{\mu K^{1/a}}{\delta}} \left(\mathbb{E}^{x_0}\left[e^{-s_1}\right]\right)^{\lceil K^{1/a} \rceil} \leq \left(\frac{2e^{\frac{\pi^2}{4\delta m^2}}}{e^{\frac{\sqrt{2}\pi}{2m}} + e^{-\frac{\sqrt{2}\pi}{2m}}}\right)^{\lceil K^{1/a} \rceil}. \tag{10.6.4}$$

Using (10.9.3), we have:

$$I_2(K) \leq \mathbb{P}^{x_0}\left(F(T) \geq \frac{\mu K^{1/a}}{\delta}\right) \underset{K\to\infty}{\sim} CK^{-\frac{1}{2a}}, \quad C = \frac{1}{\sqrt{2}\Gamma(1/2)}\left(\int_{\frac{r_0^2}{2T}}^{\infty} \frac{e^{-u}}{u} du\right)\left(\frac{\mu}{\delta}\right)^{-\frac{1}{2}}. \tag{10.6.5}$$

Since $0 < a < 1/2$, the result follows.

Similarly, in the case $a \geq 1/2$, one has that

$$\mathbb{P}^{x_0}(N^a \geq K) \geq \int_{2\mu K^{1/a}}^{\infty} \mathbb{P}^{x_0}\left(\sum_{i=1}^{\lceil K^{1/a}\rceil} s_i \leq y\right) \mathbb{P}^{x_0}\left(F(T) \in dy\right).$$

For K large enough and $y \geq 2\mu K^{1/a}$, we have $\mathbb{P}^{x_0}(\sum_{i=1}^{\lceil K^{1/a}\rceil} s_i \leq y) \geq 1/2$ by the Central Limit Theorem, and $\mathbb{P}^{x_0}\left(F(T) \geq 2\mu K^{1/a}\right) \sim CK^{-1/(2a)}$ as $K \to \infty$ for some constant $C > 0$ which implies the result in this case. □

From this result, we conclude that the average number of iterations of the algorithm in the reflected case is infinite. A modification of the above algorithm with finite number of iterations in expectation is provided below.

10.6.3 Proposition of modification of Algorithm II

When the simulated radii $r_1, \ldots, r_n$ in Algorithm II of Section 10.5.3 become small, the process $(X_t)_{t\in[0,T]}$ goes from one boundary $\partial\mathcal{V}^{\pm}$ to the other boundary $\partial\mathcal{V}^{\mp}$ many times and the number of iterations becomes high. In Section 10.11 is hinted that this problem is in fact specific to the dimension 2. To remedy this problem, we study the asymptotic behavior of $\mathbb{P}^{x_0}(X_t \in dy)$ when r_0 is close to zero. In fact, if the ratio rr_0/t is small, then the density of the reflected Brownian motion expressed in (10.4.4) can be approximated by the distribution obtained by taking only $n = 0$:

$$\mathbb{P}^{x_0}(X_t \in dy) \simeq \frac{r}{t\alpha} e^{-\frac{r^2+r_0^2}{2t}} I_0\left(\frac{rr_0}{t}\right) dr d\theta. \tag{10.6.6}$$

The function above can be renormalized so that it becomes a density which can be simulated using inequality (10.9.1) and the acceptance-rejection method with the reference density

$$\propto re^{-\frac{(r-r_0)^2}{2t}} dr d\theta = \left((r-r_0)e^{-\frac{(r-r_0)^2}{2t}} dr + r_0 e^{-\frac{(r-r_0)^2}{2t}} dr\right) d\theta. \tag{10.6.7}$$

Then we modify Algorithm II as follows. We choose some small $\varepsilon \in (0, r_0)$ and:

- Simulate the sequence of $(r_n)_n$ and $(\theta_n)_n$ as in Algorithm II in Section 10.5.3.
- If after some iteration n, we have

$$\frac{r_n^2}{T - T_n} < \varepsilon \tag{10.6.8}$$

(note that $T - T_n > 0$ is the remaining time of the simulated process after step n), then we directly simulate the final point $\bar{X}_T$ according to the approximation (10.6.6):

$$\mathbb{P}^{x_0}(\bar{X}_T \in dy |\ X_{T_n} = y_n) = C(\varepsilon)\frac{r}{(T-T_n)\alpha} e^{-\frac{r^2+r_n^2}{2(T-T_n)}} I_0\left(\frac{rr_n}{T-T_n}\right) dr d\theta, \tag{10.6.9}$$

$$C(\varepsilon) := \left(\int_0^{\alpha}\int_0^{\infty} \frac{r}{(T-T_n)\alpha} e^{-\frac{r^2+r_n^2}{2(T-T_n)}} I_0\left(\frac{rr_n}{T-T_n}\right) dr d\theta\right)^{-1}.$$

Using that $I_0(x) \geq 1$ for all $x \geq 0$, one obtains that $C(\varepsilon) \leq e^{r_n^2/(2(T-T_n))}$. This upper bound for $C(\varepsilon)$ is enough in order to implement the acceptance-rejection sampling method. In this case, the algorithm directly ends after one additional iteration.

Proposition 10.6.2. *Denote by $\bar{N}$ the number of iterations of the modified algorithm,* (10.6.9)*, then for any $1 \leq a < p < 2$, there exists a quantity $C(a, p, x_0, T, m) > 0$ such that for ε small enough:*

$$\mathbb{E}^{x_0}[\bar{N}^a] \leq \frac{C(a, p, x_0, T, m)}{\varepsilon^{p-1}}. \tag{10.6.10}$$

Proof. We use a similar argument as in the proof of Proposition 10.6.1. Let $\zeta_\varepsilon := \inf\{t \in [0, T] : R_t^2/(T-t) \leq \varepsilon\}$ and let $K \in \mathbb{N}$ fixed. Using Markov's inequality, we have:

$$\begin{aligned}
\mathbb{P}^{x_0}(\bar{N} - 1 \geq K^{1/a}) &\leq \mathbb{P}^{x_0}\left(\sum_{i=1}^{\lceil K^{1/a}\rceil} s_i \leq F(T \wedge \zeta_\varepsilon),\ F(T \wedge \zeta_\varepsilon) \leq \frac{\mu K^{1/a}}{\delta}\right) \\
&\quad + \mathbb{P}^{x_0}\left(F(T \wedge \zeta_\varepsilon) > \frac{\mu K^{1/a}}{\delta}\right) \\
&\leq e^{\frac{\mu K^{1/a}}{\delta}} \left(\mathbb{E}^{x_0}\left[e^{-s_1}\right]\right)^{\lceil K^{1/a}\rceil} + \frac{\delta^p}{(\mu K^{1/a})^p} \mathbb{E}^{x_0}\left[F(T \wedge \zeta_\varepsilon)^p\right].
\end{aligned}$$

From here the result follows by using Lemma 10.9.5. □

Proposition 10.6.3. *Denote by $\bar{X}_T$, the result obtained from the above approximation algorithm. Then, for $\alpha \in (0, 2\pi)$ and ε small enough, the error made by the approximation in total variation distance satisfies:*

$$d_{\mathrm{TV}}(\bar{X}_T, X_T) \leq C_m \varepsilon^{\min\left(1, \frac{\pi}{2\alpha}\right)},$$

where C_m is a constant which only depends on m.

Proof. If the approximation is not used, then $\bar{X}_T = X_T$. Else, we denote by n the step when the approximation is used, so that we simulate the last step starting from $(r_n \cos(\theta_n), r_n \sin(\theta_n))$ with remaining time $T - T_n =: t'$. Since the ε-condition is reached, we have $r_n^2/t' < \varepsilon$. Then if we denote by $d_{\mathrm{TV}_n}(\bar{X}_T, X_T)$ the total variation distance conditioned to the filtration $\mathbb{F}_n$ up to time T_n, and $d_n := 1/2$ for $n = 0$ and $d_n := 1$ otherwise, we have:

$$\begin{aligned}
d_{\mathrm{TV}_n}(\bar{X}_T, X_T) &\leq \int_{\mathcal{D}} \Big| (C(\varepsilon) - 1 + 1) \frac{r}{t'\alpha} e^{-\frac{r^2 + r_n^2}{2t'}} I_0\left(\frac{r r_n}{t'}\right) \\
&\qquad - \frac{2r}{t'\alpha} e^{-\frac{r^2 + r_n^2}{2t'}} \sum_{k=0}^{\infty} d_n I_{k\pi/\alpha}\left(\frac{r r_n}{t'}\right) \cos\left(\frac{k\pi\theta}{\alpha}\right) \cos\left(\frac{k\pi\theta_n}{\alpha}\right) \Big| dr d\theta \\
&\leq \frac{|C(\varepsilon) - 1|}{C(\varepsilon)} + \int_0^\infty \frac{2r}{t'} e^{-\frac{r^2 + r_n^2}{2t'}} \sum_{k=1}^{\infty} I_{k\pi/\alpha}\left(\frac{r r_n}{t'}\right) dr.
\end{aligned}$$

Using that for all $x \geq 0$, $I_0(x) \geq 1$, we have $C(\varepsilon) \leq e^{r_n^2/(2t')}$. On the other hand, we have

$$C(\varepsilon)^{-1} \leq \frac{1}{t'} \int_0^\infty r e^{-r^2/(2t')} I_0\left(\frac{r r_n}{t'}\right) dr = \int_0^\infty u e^{-u^2/2} I_0\left(\frac{r_n u}{\sqrt{t'}}\right) du.$$

Using (10.2.1) one immediately obtains that $|I_0'(x)| + |I_0''(x)| \leq 2e^x$. Therefore the function $A(\delta) := \int_0^\infty u e^{-u^2/2} I_0(\delta u) du$, $\delta \geq 0$ is twice differentiable. As $I_0(0) = 1$ and $I_0'(0) = 0$, we obtain $A(0) = 1$, $A'(0) = 0$ and then $A(\delta) = 1 + O(\delta^2)$ as $\delta \to 0$ which gives $C(\varepsilon) \geq 1 - C\varepsilon$ for small enough $\varepsilon > 0$. These bounds on $C(\varepsilon)$ imply that $|C(\varepsilon) - 1|/(C(\varepsilon)\varepsilon)$ is bounded. For the second term, using the expression (10.2.1) and a multiplicity property of the number of terms of the type $k\pi/\alpha$ in the interval $[0, k]$, we have

$$\sum_{k=1}^{\infty} I_{k\pi/\alpha}\left(\frac{r r_n}{t'}\right) = \left(\frac{r r_n}{2t'}\right)^{\frac{\pi}{\alpha}} \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \frac{1}{m! \Gamma\left(m + \frac{(k+1)\pi}{\alpha} + 1\right)} \left(\frac{r r_n}{2t'}\right)^{2m + \frac{k\pi}{\alpha}}$$

$$\leq \left(\frac{rr_n}{2t'}\right)^{\frac{\pi}{\alpha}} \sum_{m=0}^{\infty} \frac{1}{m!} \left(\frac{rr_n}{2t'}\right)^{2m} \sum_{k=0}^{\infty} \frac{1}{\Gamma\left(\frac{(k+1)\pi}{\alpha}+1\right)} \left(\frac{rr_n}{2t'}\right)^{\frac{k\pi}{\alpha}} \leq \left(\frac{rr_n}{2t'}\right)^{\frac{\pi}{\alpha}} e^{\left(\frac{rr_n}{2t'}\right)^2} \frac{\alpha}{\pi} e^{\frac{rr_n}{2t'}}.$$

With this inequality and the change of variable $u = rr_n/(2t')$, we obtain

$$d_{\mathrm{TV}_n}(\bar{X}_T, X_T) \leq C\varepsilon + \frac{8\alpha t'}{\pi r_n^2} e^{-\frac{r_n^2}{2t'}} \int_0^{\infty} u^{1+\frac{\pi}{\alpha}} e^{u^2+u} e^{-\frac{2t'}{r_n^2}u^2} du =: C\varepsilon + I(\varepsilon).$$

The integral $I(\varepsilon)$ converges if $\gamma \equiv \gamma(\varepsilon) := 2t'/r_n^2 > 1$, which is satisfied as soon as $\varepsilon < 2$. Then, denoting $\beta := 1 + \pi/\alpha$,

$$\begin{aligned} I(\varepsilon) \leq & \frac{8\alpha t'}{\pi r_n^2} \int_0^{\infty} u^{\beta} e^{-(\gamma-1)u^2+u} du = \frac{8\alpha t'}{\pi r_n^2} e^{\frac{1}{4(\gamma-1)}} \int_0^{\infty} u^{\beta} e^{-(\gamma-1)\left(u-\frac{1}{2(\gamma-1)}\right)^2} du \\ \leq & \frac{8\alpha t'}{\pi r_n^2} e^{\frac{1}{4(\gamma-1)}} \left(\int_{\frac{1}{2(\gamma-1)}}^{\infty} u^{\beta} e^{-(\gamma-1)u^2} du + \left(\frac{1}{2(\gamma-1)}\right)^{\beta} \int_{-\frac{1}{2(\gamma-1)}}^{0} e^{-(\gamma-1)u^2} du \right) \\ \leq & \frac{4\alpha\gamma}{\pi} e^{\frac{1}{4(\gamma-1)}} \left(\frac{1}{\sqrt{\gamma-1}^{1+\beta}} \int_0^{\infty} v^{\beta} e^{-v^2} dv + \frac{1}{2^{\beta}(\gamma-1)^{\beta+1/2}} \int_{-\frac{1}{2\sqrt{\gamma-1}}}^{0} e^{-v^2} dv \right). \end{aligned}$$

The first term converges to zero as $O(\varepsilon^{\pi/(2\alpha)})$ while the second term converges to zero as $O(\varepsilon^{1+\pi/\alpha})$ when $\varepsilon \to 0$. Taking all the elements into account we obtain that

$$d_{\mathrm{TV}}(\bar{X}_T, X_T) \underset{\varepsilon\to 0}{=} O\left(\varepsilon^{\min\left(1, \frac{\pi}{2\alpha}\right)}\right).$$

□

10.7 Adaptation of the algorithms to general processes

10.7.1 Stopped Brownian motion with constant drift

Let $b \in \mathbb{R}^2$ and consider the Brownian motion with drift: $\widetilde{W}_t := W_t + bt$. Then by Girsanov's theorem, $\widetilde{W}$ is a Brownian motion under the probability

$$\mathbb{Q}^{x_0}_{|\mathcal{F}_t} := e^{-b\cdot W_t - b^2 t/2} \mathbb{P}^{x_0}_{|\mathcal{F}_t}, \tag{10.7.1}$$

so that for any bounded test function $f : \mathcal{D} \times [0,T] \to \mathbb{R}$:

$$\mathbb{E}^{x_0}\left[e^{-b\cdot \widetilde{W}_{T\wedge\widetilde{\tau}} + \frac{b^2 T\wedge\widetilde{\tau}}{2}} f(\widetilde{W}_{T\wedge\widetilde{\tau}}, T\wedge\widetilde{\tau}) \right] = \mathbb{E}^{x_0}\left[f(W_{T\wedge\tau}, T\wedge\tau) \right].$$

So we have:

$$\mathbb{E}^{x_0}\left[f(\widetilde{W}_{T\wedge\widetilde{\tau}}, T\wedge\widetilde{\tau}) \right] = \mathbb{E}^{x_0}\left[e^{b\cdot W_{T\wedge\tau} - \frac{b^2 T\wedge\tau}{2}} f(W_{T\wedge\tau}, T\wedge\tau) \right].$$

To simulate the Brownian motion with drift stopped in the wedge $\mathcal{D}$, we proceed as follows:

1. Simulate $W_{T\wedge\tau}$ and $T\wedge\tau$ for the stopped process without drift according to Algorithm I in Section 10.5.2.

2. The result of the simulation becomes

$$e^{-\frac{b^2 T\wedge\tau}{2}} e^{b\cdot W_{T\wedge\tau}} f(W_{T\wedge\tau}) \tag{10.7.2}$$

We can also deduce an explicit formula for the density of the stopped Brownian motion with drift using (10.3.4). In fact, for $y \in \mathcal{D}$, we have

$$\mathbb{P}^{x_0}(\widetilde{W}_t \in dy,\ \widetilde{\tau} > t) = \frac{2r}{t\alpha} e^{-\frac{b^2 t}{2}} e^{b\cdot y} e^{-(r^2+r_0^2)/2t} \sum_{n=1}^{\infty} I_{n\pi/\alpha}\left(\frac{rr_0}{t}\right) \sin\left(\frac{n\pi\theta}{\alpha}\right) \sin\left(\frac{n\pi\theta_0}{\alpha}\right) dr d\theta, \tag{10.7.3}$$

with the same notations as before: $\widetilde{\tau} := \inf\{s :\ W_s + bs \in \partial\mathcal{D}\}$, $x = (r_0\cos(\theta_0), r_0\sin(\theta_0))$ and $y = (r\cos(\theta), r\sin(\theta))$.

10.7.2 Adaptation of the simulation algorithms for Itō processes

In this section, we present some extensions of the above simulation methods applied to the approximation of stopped and reflected diffusions at some fixed time $T > 0$. Let us start, first with the case of killed diffusion.

Consider the following Itō process in $\mathbb{R}^2$:

$$\begin{cases} dY_t = b(Y_t)dt + \sigma(Y_t)dW_t \\ Y_0 = x_0, \end{cases} \tag{10.7.4}$$

where $(W_t)_{t\geq 0}$ is a standard two-dimensional Brownian motion and $x_0 \in \mathcal{D}$. We define τ to be the exit time of $(Y_t)_{t\geq 0}$ from $\mathcal{D}$.

Simulation of $Y_{T\wedge\tau}$: Choose $t_i := Ti/n$, $i = 0, ..., n$, a uniform partition of $[0, T]$ and consider the following Euler-Maruyama scheme. Given $\bar{Y}_{t_k} \in \text{int}(\mathcal{D})$ (with $\bar{Y}_{t_0} = x_0$) and supposing that $\bar{\tau} := \inf\{s; \bar{Y}_s \in \partial\mathcal{D}\} > t_k$, simulate $\bar{Y}_{t_{k+1}}$ as the final value of the following Brownian motion with drift, stopped on the wedge $\mathcal{D}$, $t \in [t_k, t_{k+1}]$:

$$\bar{Y}_{t\wedge\bar{\tau}} := \bar{Y}_{t_k} + b(\bar{Y}_{t_k})(t - t_k) + \sigma(\bar{Y}_{t_k}) \cdot (W_{t\wedge\bar{\tau}} - W_{t_k}), \tag{10.7.5}$$

This is done using the algorithm in Section 10.7.1. More precisely, we simulate $\bar{Y}_{T\wedge\bar{\tau}}$ taking $b \equiv 0$ while keeping track of the Brownian increments; then instead of $f(\bar{Y}_{T\wedge\bar{\tau}})$, the result of the simulation becomes

$$\exp\left(-\frac{1}{2}\int_0^{T\wedge\bar{\tau}} |\sigma^{-1}(\bar{Y}_s)\cdot b(\bar{Y}_s)|^2 ds + \int_0^{T\wedge\bar{\tau}} (\sigma^{-1}(\bar{Y}_s)\cdot b(\bar{Y}_s))\cdot dW_s\right) f(\bar{Y}_{T\wedge\bar{\tau}}). \tag{10.7.6}$$

Note that we have to take into account the decorrelation step and the change of angle as described in Section 10.3 at every step. If the process (10.7.5) exits the wedge $\mathcal{D}$ in the interval $[t_k, t_{k+1}]$ then the algorithm directly stops. This scheme has weak order one. That is, for any $f, b, \sigma \in \mathcal{C}_b^5(\bar{\mathcal{D}})$ one has that there exists a constant $C_f > 0$

$$\left|\mathbb{E}^{x_0}\left[f(Y_{T\wedge\tau})\right] - \mathbb{E}^{x_0}\left[f(\bar{Y}_{T\wedge\bar{\tau}})\right]\right| \leq C_f n^{-1}.$$

The proof is straightforward if one follows the same line of proof as in [Gob01]. In particular, see Section 2.2.1 and note that our case is simpler as the proposed scheme does not have the possibility of touching the boundary before it is stopped or reaches T. For a discussion about the associated partial differential equation with Dirichlet conditions, see the discussion right after (10.10.1).

Since the Girsanov change of measure uses an exponential function, it may not be suitable to processes with large drifts as it may increase the variance. For this reason, we also propose a two-step Euler-Maruyama scheme for the stopped process.

Simulation of $Y_{T\wedge\tau}$: Choose $t_i := Ti/n$, $i = 0, ..., n$, a uniform partition of $[0, T]$ and consider the following two-step Euler-Maruyama scheme. Given $\bar{Y}_{t_k} \in \text{int}(\mathcal{D})$ (with $\bar{Y}_{t_0} = x_0$) and supposing that $\bar{\tau} := \inf\{s; \bar{Y}_s \in \partial\mathcal{D}\} > t_k$ and that $\tilde{\tau} := \inf\{s; \tilde{Y}_s \in \partial\mathcal{D}\} > t_k$, simulate $\bar{Y}_{t_{k+1}\wedge\bar{\tau}}$ in two steps as follows for $t \in [t_k, t_{k+1}]$:

$$\tilde{Y}_{t\wedge\bar{\tau}} := \bar{Y}_{t_k} + b(\bar{Y}_{t_k})(t \wedge \bar{\tau} - t_k) \tag{10.7.7}$$

$$\bar{Y}_{t\wedge\tilde{\tau}} := \tilde{Y}_{t_{k+1}\wedge\bar{\tau}} + (W_{t\wedge\bar{\tau}} - W_{t_k}). \tag{10.7.8}$$

The second step is performed using the algorithm in Section 10.5.2. Note that we have to take into account the decorrelation step and the change of angle as described in Section 10.3. If the process exits the wedge $\mathcal{D}$ in the interval $[t_k, t_{k+1}]$ then the algorithm directly stops. This scheme has weak order one. That is, for any $f, b, \sigma \in \mathcal{C}_b^5(\bar{\mathcal{D}})$, there exists a constant $C_f > 0$ such that

$$\left|\mathbb{E}^{x_0}\left[f(Y_{T\wedge\tau})\right] - \mathbb{E}^{x_0}\left[f(\bar{Y}_{T\wedge\bar{\tau}})\right]\right| \leq C_f n^{-1}. \tag{10.7.9}$$

The proof is similar to the proof for the convergence of the two-step Euler scheme for the reflected processes in Proposition 10.7.1.

Now let $(Z_t)_{t\geq 0}$ to be the solution of a stochastic differential equation with drift coefficient b which is normally reflected on the boundary of the wedge $\mathcal{D}$. That is,

$$dZ_t = b(Z_t)dt + dW_t + dL_t.$$

For details on existence and uniqueness for this equation we refer to [Sai87].

In this case, we do not know how to extend the simulation method using the Girsanov change of measure; instead we propose a two-step scheme.

Simulation of Z_T: Choose $t_i := Ti/n$, $i = 0, ..., n$ a uniform partition of $[0, T]$ and consider the following two-step scheme. Given $\bar{Z}_{t_k}$ (with $\bar{Z}_{t_0} = x_0$), simulate $\bar{Z}_{t_{k+1}}$ as the final value of the reflection on the wedge $\mathcal{D}$ of the Brownian motion and drift in two steps as follows for $t \in [0, t_{k+1} - t_k]$

$$\tilde{Z}_{t+t_k} := \bar{Z}_{t_k} + b(\bar{Z}_{t_k})(t - t_k) + L_t^1 - L_{t_k}^1 \tag{10.7.10}$$

$$\bar{Z}_{t+t_k} := \tilde{Z}_{t_{k+1}} + (W_t - W_{t_k}) + L_t^2 - L_{t_k}^2. \tag{10.7.11}$$

The terms L^i, $i = 1, 2$, denote the respective local time terms for each of the two steps. The simulation of the second step of this algorithm uses the argument described in either Sections 10.5.3 or 10.6.3. The weak rate of convergence for these methods is as follows:

Proposition 10.7.1. *Assume that $f, b \in \mathcal{C}_b^5(\bar{\mathcal{D}})$ then*

$$\left|\mathbb{E}^{x_0}\left[f(Z_T)\right] - \mathbb{E}^{x_0}\left[f(\bar{Z}_T)\right]\right| \leq C_f n^{-1}. \tag{10.7.12}$$

Denoting by $\hat{Z}_T$, the result of the algorithm using the approximation described in Section 10.6.3, we have

$$\left|\mathbb{E}^{x_0}\left[f(Z_T)\right] - \mathbb{E}^{x_0}\left[f(\hat{Z}_T)\right]\right| \leq C_f\left(n^{-1} + \varepsilon^{\min\left(1, \frac{\pi}{2\alpha}\right)}\right). \tag{10.7.13}$$

Proof. The proof of (10.7.13) is a small modification of (10.7.12). Therefore, we prove the rate of convergence for the algorithm provided in Section 10.5.3.

The argument follows as in the classical diffusion case which can be found in the proof of the main result in [TT90]. We use this argument and refer the reader to [TT90] for more details.

First, consider u to be the solution of the following backward partial differential equation with Neumann conditions:

$$\begin{array}{ll} \partial_t u(t,x) + \mathcal{L}u(t,x) = 0, & (t,x) \in [0,T] \times \mathcal{D} \\ u(T,x) = f(x), & x \in \mathcal{D} \\ \nabla u(t,x) \cdot n(x) = 0, & t > 0,\ x \in \partial\mathcal{D}. \end{array} \tag{10.7.14}$$

Here, $\mathcal{L}u(t,x) := b(x)\nabla u(t,x) + \frac{1}{2}\Delta u(t,x)$. Following the same discussion as in (10.10.1), one obtains that $u \in \mathcal{C}^{7/2,7}([0,T] \times \bar{\mathcal{D}})$. This property will be used when doing Taylor's expansions for u. Next, note that

$$\mathbb{E}^{x_0}\left[f(Z_T)\right] - \mathbb{E}^{x_0}\left[f(\bar{Z}_T)\right] = \sum_{i=0}^{n-1}\left(\mathbb{E}^{x_0}\left[u(t_i,\bar{Z}_{t_i}) - u(t_{i+1},\bar{Z}_{t_{i+1}})\right]\right) =: -\sum_{i=0}^{n-1} A_i. \tag{10.7.15}$$

We will now consider each term within the above sum. Using Itō-Tanaka formula (here one uses (10.7.14) to cancel all local time terms), we have

$$\begin{aligned} A_i &= \mathbb{E}^{x_0}\left[u(t_i,\tilde{Z}_{t_{i+1}}) - u(t_i,\bar{Z}_{t_i}) + \int_{t_i}^{t_{i+1}}\left(\partial_t + \frac{1}{2}\Delta\right)u(s,\bar{Z}_s)ds\right] \\ &= \int_{t_i}^{t_{i+1}} \mathbb{E}^{x_0}\left[b(\bar{Z}_{t_i})\nabla u(t_i,\tilde{Z}_s) + \left(\partial_t + \frac{1}{2}\Delta\right)u(s,\bar{Z}_s)\right]ds. \end{aligned}$$

The following step relies on using the Taylor expansion for each of the terms involving $\nabla u(t_i,\tilde{Z}_s)$ and $(\partial_t + (1/2)\Delta)\,u(s,\bar{Z}_s)$ at $(t_i,\bar{Z}_{t_i})$ together with the fact that u solves (10.7.14). Without going into much detail, let us consider the expansion of the term $\partial_t u(s,\bar{Z}_s)$:

$$\begin{aligned} \partial_t u(s,\bar{Z}_s) = \partial_t u(t_i,\bar{Z}_{t_i}) &+ \int_0^1 \nabla\partial_t u(t_i,\alpha\bar{Z}_s + (1-\alpha)\bar{Z}_{t_i})\cdot(\bar{Z}_s - \bar{Z}_{t_i})d\alpha \\ &+ \int_0^1 \partial_t\partial_t u(\alpha s + (1-\alpha)t_i,\bar{Z}_s)\cdot(s-t_i)d\alpha \end{aligned}$$

Each of the above derivatives of u is bounded. In the case of the increments of $\bar{Z}_s - \bar{Z}_{t_i} = (W_s - W_{t_i}) + L_s^2 - L_{t_i}^2 + \tilde{Z}_{t_{i+1}} - \bar{Z}_{t_i}$ one has to take one further step in the Taylor expansion to be able to use the fact that the expectation of the Brownian increment is zero. Namely,

$$\begin{aligned} \int_0^1 \nabla\partial_t u(t_i,\alpha\bar{Z}_s + (1-\alpha)\bar{Z}_{t_i})\cdot(\bar{Z}_s - \bar{Z}_{t_i})d\alpha &= \nabla\partial_t u(t_i,\bar{Z}_{t_i})\cdot(\bar{Z}_s - \bar{Z}_{t_i}) \\ &+ \int_0^1 (1-\alpha)(\bar{Z}_s - \bar{Z}_{t_i})^\top\cdot\nabla^2\partial_t u(t_i,\alpha\bar{Z}_s + (1-\alpha)\bar{Z}_{t_i})\cdot(\bar{Z}_s - \bar{Z}_{t_i})d\alpha. \end{aligned}$$

After this and canceling the first terms of the expansion using (10.7.14) and taking expectations we see that the sum of most terms is of order $O(n^{-2})$. The remaining terms are bounded by $C\int_{t_i}^{t_{i+1}}(L_s^j - L_{t_i}^j)ds$ for $j = 1,2$. Considering these terms within (10.7.15), one obtains for $L_T^j = \sum_{i=0}^{n-1}(L_{t_{i+1}}^j - L_{t_i}^j)$ that these sums are bounded by

$$\sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} \mathbb{E}^{x_0}[L_s^j - L_{t_i}^j]ds \le n^{-1}\mathbb{E}^{x_0}[L_T^j] \le Cn^{-1},$$

where we used that $t \mapsto L_t^j$ is increasing. The last inequality follows from an estimate for the Skorohod problem in convex domains (see [Sai87, Theorem 4.2]). In the case of the algorithm described in Section 10.6.3 one may modify the above arguments adding at each time t_i, the possibility of ending the simulation in the partition interval using the criteria provided in (10.6.8). Using Proposition 10.6.3 one obtains the result. □

	$\mathbb{E}^{x_0}$	95 % interval	Time (s)	MC iterations	$\mathbb{E}^{x_0}[N]$
$\mathbb{E}^{x_0}[f(W_\tau)]$ with [Met09]	3.473	$\pm$ 0.082	0.67	50000	1
$\mathbb{E}^{x_0}[f(W_\tau)]$	3.489	$\pm$ 0.085	1.04	50000	1.45
$\mathbb{E}^{x_0}[\tau]$ with [Met09]	0.742	$\pm$ 0.030	7.06	20000	-
$\mathbb{E}^{x_0}[\tau]$	0.590	$\pm$ 0.024	4.92	20000	-
$\mathbb{E}^{x_0}[f(W_{\tau\wedge T})]$	2.980	$\pm$ 0.049	5.79	10000	1.37
$\mathbb{E}^{x_0}[(W_{\tau\wedge T})\cdot e_1]$	1.441	$\pm$ 0.012	6.26	10000	1.37
$\mathbb{E}^{x_0}[f(X_T)]$, $\varepsilon = 0$	4.138	$\pm$ 0.224	13.2	1000	75.6
$\mathbb{E}^{x_0}[f(X_T)]$, $\varepsilon = 0.03$	4.313	$\pm$ 0.072	28.2	10000	5.11

Table 10.1: Simulations with $\alpha = 0.9$, $r_0 = 1.5$, $\theta_0 = 0.3$, $f(x, y) = x^2 + y^2$, $T = 1$.

	$\mathbb{E}^{x_0}$	95 % interval	Time (s)	MC iterations	$\mathbb{E}^{x_0}[N]$
$\mathbb{E}^{x_0}[f(W_{\tau\wedge T})]$	0.195	$\pm$ 0.003	8.85	10000	1.28
$\mathbb{E}^{x_0}[f(X_T)]$, $\varepsilon = 0.03$	0.117	$\pm$ 0.003	9.26	5000	2.73

Table 10.2: Simulations with $\alpha = 0.58$, $r_0 = 3$, $\theta_0 = 0.4$, $f(r\cos(\theta), r\sin(\theta)) = \sin^2(\theta)$, $T = 1$.

10.8 Simulations

We implement the algorithms[3] described in Sections 10.5.2 and 10.5.3 as well as the approximated version in Section 10.6.3. We also give simulation results obtained with Metzler's algorithm [Met09, Section 2.1]. This last algorithm simulates W_τ (instead of $W_{\tau\wedge T}$) directly for a wedge of general angle α, by:

1. Simulate the radius $|W_\tau|$ using (10.5.6).

2. Conditionally simulate τ by acceptance-rejection, approximating the infinite sum by a partial sum, and using a Cauchy distribution as reference density.

However, Metzler's algorithm cannot be adapted to simulate $W_{\tau\wedge T}$ and is biased for the simulation of τ (see [Met09, Proposition 2.1.8]).

In Tables 10.1 and 10.2, we consider the Monte Carlo estimation of various expectations. We choose positive test functions in order to avoid cancellations and r_0 so that $\mathbb{E}^{x_0}[\tau]$ is of the same order as T, so that the stopped and reflected processes are significantly different than the standard Brownian motion.

We note that for the estimation of $\mathbb{E}[f(W_\tau)]$, our simulation method finds a similar value as the algorithm proposed by Metzler, as both are exact simulation methods. However, for the estimation of $\mathbb{E}[\tau]$, the bias in the method proposed by Metzler seems to be significant.

As a second way to check our algorithm, we also simulated the projection on the first axis of $W_{\tau\wedge T}$. Note that the simulation gives a result close to the initial point $1.5 \cdot \cos(0.3) \simeq 1.4330$ which is reported on the sixth line of Table 10.1. This is because $t \mapsto W_{\tau\wedge t}$ is a martingale.

For the estimation of $\mathbb{E}[f(X_T)]$ (reflected Brownian motion), the exact algorithm takes too much time, as hinted in Proposition 10.6.1, hence the need to use the approximation version from Section 10.6.3. We only give a simulation example with 1000 samples, as we could not get a more proper estimation. Indeed, if we increase the number of samples, there appear trajectories where N becomes large and where the algorithm does not stop in reasonable time.

[3]The programs with a demonstration notebook are available at `https://github.com/Bras-P/simulation_reflected_brownian_motion_wedge`

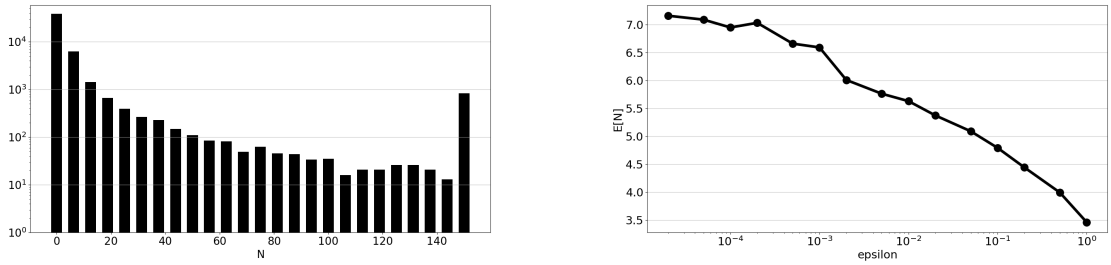


Figure 10.6: (left) Histogram of N for the exact reflected algorithm with 50000 Monte Carlo iterations. The horizontal axis represents the values of N while the vertical axis gives the number of trajectories such that N is in the interval noted in the horizontal axis. The last bar counts all the values greater than 150. On the right, average number of iterations in function of ε. The parameters are the same as in Table 10.1 and for each value of ε, 50000 Monte Carlo iterations are used.

	$\mathbb{E}^{x_0}$	95 % interval	Time (s)	MC iterations	Discretization step
$\mathbb{E}^{x_0}[f(Y_{\tau\wedge T})]$ [Girsanov]	2.83	$\pm$ 0.06	456	5000	T/500
$\mathbb{E}^{x_0}[f(Y_{\tau\wedge T})]$ [Two-step]	2.84	$\pm$ 0.06	457	5000	T/500
$\mathbb{E}^{x_0}[f(Z_T)]$, $\varepsilon = 0.01$	3.77	$\pm$ 0.09	1140	5000	T/500

Table 10.3: Simulation of the process $dY_t = -\mu(Y_t - \kappa) + \sigma dW_t$, with $\alpha = 0.9$, $r_0 = 1.5$, $\theta_0 = 0.3$, $\mu = (0.1, 0.2)$, $\kappa = (0.7, 0.5)$, $\sigma = I_2$, $f(x, y) = x^2 + y^2$, $T = 1$. Z corresponds to the reflected process in the same domain. The first stopped process is simulated using the Girsanov method while the second one using the two-step Euler-Maruyama scheme.

To have a better idea of the behavior of the algorithm, we provide on the left side of Figure 10.6, the histogram of N for the exact Algorithm II from Section 10.5.3; if, for a trajectory, N exceeds 150, then we stop the algorithm. We observe that for most of the trajectories, N is not too large (does not exceed 5). However, for a few iterations (around 0.2%) which corresponds to the case where the radius r becomes small, N is large, which slows down the algorithm. On the right side of Figure 10.6 we trace $\mathbb{E}^{x_0}[N]$ for the approximated reflected algorithm in function of ε. However, we could not trace the bias in function of ε, as it seems negligeable in comparison with the size of the confidence interval. This is why we do not need to lengthen the simulation time of each trajectory by choosing a very small ε; $\varepsilon = 0.02$ or $\varepsilon = 0.05$ with $\mathbb{E}^{x_0}[N] \approx 5$ is sufficient.

In Table 10.3, we implement the algorithms adapted to general Itō processes from Section 10.7.2.

10.9 Appendix: Auxiliary Lemmas

We start with a simple lemma about the simulation of a r.v. X using only partial information of a previously simulated r.v. Y.

Proposition 10.9.1. *Let X and Y be two random variables. We want to simulate X. Let A be a deterministic set. We consider the following algorithm :*

1. *Simulate Y.*
2. *If $Y \in A$, simulate X conditionally to the value of Y in the previous step, and return X.*

3. If $Y \notin A$, simulate X conditionally to the event $Y \notin A$, and return X.

Let us denote $\widetilde{X}$ the value returned by the algorithm. Then:

$$\widetilde{X} \overset{\mathscr{L}}{=} X.$$

Proof. The random variable $\widetilde{X}$ is defined conditionally to Y as:

$$\mathbb{P}(\widetilde{X} \in dx|\ Y = y) = \begin{cases} \mathbb{P}(X \in dx|\ Y = y) & \text{if } y \in A \\ \mathbb{P}(X \in dx|\ Y \notin A) & \text{if } y \notin A \end{cases}$$

Then:

$$\begin{aligned} \mathbb{P}(\widetilde{X} \in dx) &= \int \mathbb{P}(\widetilde{X} \in dx|\ Y = y)\mathbb{P}(Y \in dy) \\ &= \int_{y \in A} \mathbb{P}(X \in dx|\ Y = y)\mathbb{P}(Y \in dy) + \int_{y \notin A} \mathbb{P}(X \in dx|\ Y \notin A)\mathbb{P}(Y \in dy) \\ &= \mathbb{P}(X \in dx,\ Y \in A) + \mathbb{P}(X \in dx|\ Y \notin A)\mathbb{P}(Y \notin A) = \mathbb{P}(X \in dx). \end{aligned}$$

□

10.9.1 Estimates on Bessel processes

Proposition 10.9.2. *For all $x \geq 0$, we have:*

$$1 \leq I_0(x) \leq e^x. \tag{10.9.1}$$

Furthermore, for $x \geq 0$ and $\nu \geq 0$, we have:

$$I_\nu(x) \leq e^x + \frac{1}{\pi(\nu + x)}. \tag{10.9.2}$$

Proof. The inequality follows from the integral representation [Wat44, p.181]:

$$I_\nu(x) = \frac{1}{\pi}\int_0^\pi e^{x\cos\theta}\cos(\nu\theta)d\theta - \frac{\sin(\nu\pi)}{\pi}\int_0^\infty e^{-x\cosh(t)-\nu t}dt.$$

□

Now we provide two results on estimates related to the law of $F(T) = \int_0^T \frac{ds}{R_s^2}$.

Proposition 10.9.3. *Let $(R_t)_{t\geq 0}$ be a Bessel process of dimension 2, or equivalently, of index 0. Then*

$$\mathbb{P}^{r_0}\left[F(T) \geq x\right] \underset{x\to\infty}{\sim} \frac{1}{\sqrt{2}\Gamma(1/2)}\left(\int_{\frac{r_0^2}{2T}}^\infty \frac{e^{-u}}{u}du\right)x^{-1/2}. \tag{10.9.3}$$

Proof. Step 1: We use [JYC09], Proposition 6.2.5.1, with $a = 0$ and $\nu = 0$:

$$\mathbb{E}^{r_0}\left[e^{-\frac{\lambda^2}{2}F(T)}\right] = \frac{r_0^\lambda}{\Gamma(\lambda/2)}\int_0^\infty v^{\frac{\lambda}{2}-1}(1+2vT)^{-(1+\lambda)}e^{-\frac{r_0^2 v}{1+2vT}}dv.$$

Performing the change of variable $w := \frac{Tv}{1+2vT}$, so that $v = \frac{w}{T(1-2w)}$, we have for all $\lambda > 0$:

$$\mathbb{E}^{r_0}\left[e^{-\frac{\lambda^2}{2}F(T)}\right] = \frac{r_0^\lambda}{\Gamma(\lambda/2)T^{\lambda/2}}\int_0^{1/2} w^{\frac{\lambda}{2}-1}(1-2w)^{\lambda/2}e^{-r_0^2 w/T}dw. \tag{10.9.4}$$

Step 2: In fact, we prove the asymptotic expansion:

$$\mathbb{E}^{r_0}\left[e^{-\frac{\lambda^2}{2}F(T)}\right] \underset{\lambda\to 0^+}{=} 1-\frac{\lambda}{2}\int_{\frac{r_0^2}{2T}}^{\infty}\frac{e^{-u}}{u}du \ +O(\lambda^2). \tag{10.9.5}$$

We have for all $x>0$, $\Gamma(x+1)=x\Gamma(x)$ and $\Gamma(1)=1$ so that

$$\Gamma(x) \underset{x\to 0^+}{\sim} \frac{1}{x}.$$

Then, using (10.9.4) and the change of variable $w=Tu/r_0^2$:

$$\mathbb{E}^{r_0}\left[e^{-\frac{\lambda^2}{2}F(T)}\right]=\frac{1}{\Gamma\left(\frac{\lambda}{2}\right)}\int_0^{\frac{r_0^2}{2T}}u^{\frac{\lambda}{2}-1}\left(1-\frac{2Tu}{r_0^2}\right)^{\lambda/2}e^{-u}du.$$

Using the definition of the Gamma function and decomposing the integral in two parts, we have

$$1-\mathbb{E}^{r_0}\left[e^{-\frac{\lambda^2}{2}F(T)}\right]=\frac{1}{\Gamma\left(\frac{\lambda}{2}\right)}\left[\int_0^{\frac{r_0^2}{2T}}u^{\frac{\lambda}{2}-1}e^{-u}\left(1-\left(1-\frac{2Tu}{r_0^2}\right)^{\lambda/2}\right)du+\int_{\frac{r_0^2}{2T}}^{\infty}u^{\frac{\lambda}{2}-1}e^{-u}du\right]=:I_1+I_2.$$

We use the expansion $(1-x)^\alpha=\sum_{k=0}^{\infty}(\alpha\log(1-x))^k/k!$ for $|x|<1$ and $\alpha\to 0$ in order to analyze the first integral. This gives the following asymptotic equivalence

$$I_1 \underset{\lambda\to 0^+}{=} -\frac{\lambda^2}{4}\int_0^{\frac{r_0^2}{2T}}u^{-1}e^{-u}\log\left(1-\frac{2Tu}{r_0^2}\right)du+o(\lambda^2)=O(\lambda^2).$$

And on the other side:

$$I_2=\frac{1}{\Gamma\left(\frac{\lambda}{2}\right)}\int_{\frac{r_0^2}{2T}}^{\infty}u^{\frac{\lambda}{2}-1}e^{-u}du \underset{\lambda\to 0^+}{=} \frac{\lambda}{2}\int_{\frac{r_0^2}{2T}}^{\infty}\frac{e^{-u}}{u}du \ +O(\lambda^2).$$

From here we obtain the asymptotic expansion (10.9.5).

Step 3: We get (10.9.3) from the expansion (10.9.5) and using Karamata Tauberian theorems (see [EKM97], Corollary A3.10):

Proposition 10.9.4. *Let df be a measure on $\mathbb{R}^+$, $F(x):=\int_0^x df$, $\hat{f}(\lambda):=\int_0^\infty e^{-\lambda x}F(x)dx$, $0\le\gamma<1$ and let $L>0$. Then the following are equivalent:*

1. $1-\hat{f}(\lambda)\sim L\lambda^\gamma, \quad \lambda\to 0^+$
2. $1-F(x)\sim\dfrac{L}{\Gamma(1-\gamma)}x^{-\gamma}, \quad x\to\infty.$

□

Next we will give an estimate for $\mathbb{E}[F(T\wedge\zeta_\varepsilon)^p]$ in the case of the modified algorithm for reflected Brownian motion.

Lemma 10.9.5. *We define the stopping time $\zeta_\varepsilon:=\inf\{t\in[0,T]:\ R_t^2/(T-t)\le\varepsilon\}$. Then we have the following bound for $p\in(1,2)$ and $\varepsilon>0$ small enough:*

$$\mathbb{E}^{r_0}\left[F(T\wedge\zeta_\varepsilon)^p\right]\le C\left(1+\log\left(\frac{2T}{r_0^2}\right)\right)(\varepsilon T)^{1-p}. \tag{10.9.6}$$

Proof. By [JYC09, (6.2.3)], the density of R_t in r is

$$\frac{r}{t}\left(\frac{r}{r_0}\right)^{\nu} e^{-\frac{r^2+r_0^2}{2t}} I_\nu\left(\frac{rr_0}{t}\right), \tag{10.9.7}$$

where $\nu = \frac{d}{2} - 1$ and d is the dimension of the Bessel process. Taking $d = 2$, we have:

$$\begin{aligned}
\mathbb{E}^{r_0}\left[F(T \wedge \zeta_\varepsilon)^p\right] &\leq \int_0^T \mathbb{E}^{r_0}\left[\frac{1}{R_t^{2p}} \mathbb{1}_{R_t \geq \sqrt{\varepsilon(T-t)}}\right] dt = \int_0^T \int_{\sqrt{\varepsilon(T-t)}}^\infty \frac{1}{r^{2p-1}t} e^{-\frac{r^2+r_0^2}{2t}} I_0\left(\frac{rr_0}{t}\right) dr dt \\
&= \int_{\sqrt{\varepsilon T}}^\infty \int_0^T \frac{1}{r^{2p-1}t} e^{-\frac{r^2+r_0^2}{2t}} I_0\left(\frac{rr_0}{t}\right) dt dr + \int_0^{\sqrt{\varepsilon T}} \int_{T-r^2/\varepsilon}^T \frac{1}{r^{2p-1}t} e^{-\frac{r^2+r_0^2}{2t}} I_0\left(\frac{rr_0}{t}\right) dt dr \\
&=: I_1 + I_2.
\end{aligned}$$

Then, using (10.9.1) for both I_1 and I_2:

$$\begin{aligned}
I_1 &\leq \int_{\sqrt{\varepsilon T}}^\infty \frac{1}{r^{2p-1}} \int_0^T \frac{1}{t} e^{-\frac{(r-r_0)^2}{2t}} dt dr = \int_{\sqrt{\varepsilon T}}^\infty \frac{1}{r^{2p-1}} \int_{\frac{(r-r_0)^2}{2T}}^\infty \frac{e^{-v}}{v} dv dr \\
&= \int_{\sqrt{\varepsilon T}}^{\sqrt{2T}+r_0} \frac{1}{r^{2p-1}} \int_{\frac{(r-r_0)^2}{2T}}^\infty \frac{e^{-v}}{v} dv dr + \int_{\sqrt{2T}+r_0}^\infty \frac{1}{r^{2p-1}} \int_{\frac{(r-r_0)^2}{2T}}^\infty \frac{e^{-v}}{v} dv dr.
\end{aligned}$$

But we have, for all $a \geq 0$:

$$\int_a^\infty \frac{e^{-x}}{x} dx \leq \begin{cases} e^{-a} & \text{if } a \geq 1, \\ e^{-1} + \log(1/a) & \text{if } a < 1. \end{cases}$$

So that

$$\begin{aligned}
I_1 &\leq \int_{\sqrt{\varepsilon T}}^{\sqrt{2T}+r_0} \frac{dr}{r^{2p-1}} \left(e^{-1} + \log\left(\frac{2T}{(r-r_0)^2}\right)\right) + \int_{\sqrt{2T}+r_0}^\infty \frac{dr}{r^{2p-1}} e^{-\frac{(r-r_0)^2}{2T}} \\
&\underset{\varepsilon \to 0}{\sim} \frac{(\varepsilon T)^{1-p}}{2(p-1)} \left(e^{-1} + \log\left(\frac{2T}{r_0^2}\right)\right).
\end{aligned}$$

For the second integral, we have:

$$\begin{aligned}
I_2 &\leq \int_0^{\sqrt{\varepsilon T}} \frac{1}{r^{2p-1}} \int_{T-r^2/\varepsilon}^T e^{-\frac{(r-r_0)^2}{2t}} \frac{dt}{t} dr \\
&= \int_0^{\sqrt{\varepsilon T}} \frac{1}{r^{2p-1}} \int_{\frac{(r-r_0)^2}{2T}}^{\frac{(r-r_0)^2}{2(T-r^2/\varepsilon)}} \frac{e^{-v}}{v} dv dr \leq \int_0^{\sqrt{\varepsilon T}} \frac{1}{r^{2p-1}} \int_{\frac{(r-r_0)^2}{2T}}^{\frac{(r-r_0)^2}{2(T-r^2/\varepsilon)}} \frac{dv}{v} dr \\
&= -\int_0^{\sqrt{\varepsilon T}} \frac{1}{r^{2p-1}} \log\left(1 - \frac{r^2}{\varepsilon T}\right) dr = -\frac{1}{(\varepsilon T)^{p-1}} \int_0^1 \frac{1}{u^{2p-1}} \log(1-u^2) du,
\end{aligned}$$

where the last integral converges as soon as $p < 2$. □

10.10 Appendix: Proofs in Section 10.4

Proof of Theorem 10.4.2 in the case that $\mathcal{D} = \langle \pi/m \rangle$. We apply the method of images in $\mathbb{R}^2$. Let $f : \mathcal{D} \to [0, +\infty)$ be a non-negative continuous function with compact support. The heat equation with boundary conditions

$$\begin{aligned}
&\partial_t u(t,x) = \tfrac{1}{2}\Delta u(t,x), && (t,x) \in [0,+\infty) \times \mathcal{D} \\
&u(0,x) = f(x), && x \in \mathcal{D} \\
&\nabla u(t,x) \cdot n(x) = 0, && t > 0,\ x \in \partial\mathcal{D},
\end{aligned} \tag{10.10.1}$$

where $n(x)$ denotes the inward unitary orthogonal vector on the boundary, can be rewritten as a partial differential equation in polar coordinates (r, θ) as described in [MNP00, Chapter 1]. Furthermore if we perform the change of variable $z = \log(r)$, we obtain a parabolic problem with non-constant coefficients in a strip with mixed type boundary condition. The existence and uniqueness for this PDE is treated in [LSU68, chapter IV, Theorem 5.3]. From this reference and the Feyman-Kac representation theorem we obtain that

$$u(t,x) := \mathbb{E}^x \left[f(X_t) \right]$$

satisfies the heat equation on $\mathcal{D}$ described above. Now we define the function

$$\begin{array}{ll} \tilde{f} & : \mathbb{R}^2 \to \mathbb{R} \\ y & \mapsto f(T_k^{-1} y) \quad \text{for } y \in \mathcal{D}_k, \end{array}$$

and for $t \geq 0$ and $x \in \mathbb{R}^2$,

$$\tilde{u}(t,x) := \mathbb{E}^x \left[\tilde{f}(W_t) \right].$$

Then $\tilde{u}$ satisfies the heat equation on $\mathbb{R}^2$ with $\tilde{f}$ as initial condition, so that

$$\tilde{u}(t,x) = \frac{1}{2\pi t} \int_{\mathbb{R}^2} \tilde{f}(y) e^{-\frac{|x-y|^2}{2t}} dy = \frac{1}{2\pi t} \sum_{k=0}^{2m-1} \int_{\mathcal{D}_k} \tilde{f}(y) e^{-\frac{|x-y|^2}{2t}} dy = \frac{1}{2\pi t} \int_{\mathcal{D}} f(y) \sum_{k=0}^{2m-1} e^{-\frac{|x-T_k y|^2}{2t}} dy.$$

On the other hand, u and $\tilde{u}$ satisfy the same boundary equation on $\mathcal{D}$. Indeed, for all $x \in \mathcal{D}$, $\tilde{f}(x) = f(x)$ so u and $\tilde{u}$ satisfy the same initial conditions. For $x \in \partial\mathcal{D}^-$, $k = 0, \ \ldots, \ 2m-1$ and $y \in \mathcal{D}$, we have

$$\begin{aligned} (x - T_k y) \cdot n(x) &= -(x - T_{2m-1-k} y) \cdot n(x), \\ |x - T_k y|^2 &= |x - T_{2m-1-k} y|^2, \end{aligned}$$

so that

$$\begin{aligned} \nabla \tilde{u}(t,x) \cdot n(x) &= -\frac{1}{2\pi t^2} \int_{\mathcal{D}} f(y) \sum_{k=0}^{2m-1} (x - T_k y) \cdot n(x) e^{-\frac{|x-T_k y|^2}{2t}} dy \\ &= -\frac{1}{2\pi t^2} \int_{\mathcal{D}} f(y) \sum_{k=0}^{m-1} (x - T_k y) \cdot n(x) e^{-\frac{|x-T_k y|^2}{2t}} dy \\ &\quad + \frac{1}{2\pi t^2} \int_{\mathcal{D}} f(y) \sum_{k=0}^{m-1} (x - T_{2m-1-k} y) \cdot n(x) e^{-\frac{|x-T_{2m-1-k} y|^2}{2t}} dy = 0. \end{aligned}$$

If $x \in \partial D^+$, we get the same result noting that for all $y \in \mathcal{D}$ and $k = 0, \ \ldots, \ 2m-1$,

$$\begin{aligned} (x - T_{(k+1) \bmod 2m} y) \cdot n(x) &= -(x - T_{(2m-k) \bmod 2m} y) \cdot n(x), \\ |x - T_{(k+1) \bmod 2m} y|^2 &= |x - T_{(2m-k) \bmod 2m} y|^2, \end{aligned}$$

So $u(t,x) = \tilde{u}(t,x)$ for all $t \geq 0$ and $x \in \mathcal{D}$, and

$$\mathbb{E}^x \left[f(X_t) \right] = \frac{1}{2\pi t} \int_{\mathcal{D}} f(y) \sum_{k=0}^{2m-1} e^{-\frac{|x-T_k y|^2}{2t}} dy.$$

□

Proof of Theorem 10.4.2 in the general case. As in the above proof we consider

$$u(t,x) = \mathbb{E}^x \left[f(X_t) \right], \tag{10.10.2}$$

where $t \geq 0$, $x \in \mathcal{D}$, and $f : \mathcal{D} \to [0, +\infty)$ is a non-negative continuous function with compact support. Then u is the solution of the partial differential equation (10.10.1). Considering the formula obtained for the case $\alpha = \pi/m$ above, we would like to express it so that it does not depend explicitly on m. We first assume that $\alpha = \pi/m$ for some $m \in \mathbb{N}$. In order to switch to polar coordinates, let $x = (r_0 \cos(\theta_0), r_0 \sin(\theta_0))$, $y = (r\cos(\theta), r\sin(\theta))$ and recall that ϑ_k for $0 \leq k \leq 2m-1$ is the angle of $T_k(y)$ in $[0, 2\pi)$, i.e. $T_k(y) = (r\cos(\vartheta_k), r\sin(\vartheta_k))$ with $\vartheta_{2k} = 2k\alpha + \theta$ and $\vartheta_{2k+1} = 2(k+1)\alpha - \theta$. We then rewrite (10.4.3):

$$\mathbb{P}^x(X_t \in dy) = \frac{r}{2\pi t} e^{-\frac{r^2+r_0^2}{2t}} \sum_{k=0}^{2m-1} e^{\frac{rr_0}{t}\cos(\theta_0 - \vartheta_k)} dr d\theta. \tag{10.10.3}$$

We use the following identity (see [GR07, page 933, (8.511.4)]), valid for $\gamma, z \geq 0$:

$$e^{\gamma z} = I_0(z) + 2\sum_{n=1}^{\infty} T_n(\gamma) I_n(z),$$

where T_n is the n^{th} Tchebychev's polynomial of the first kind and I_n the modified Bessel function of the first kind with order n. Then the result (10.3.4) follows because $T_n(\cos(\theta)) = \cos(n\theta)$ and

$$\sum_{k=0}^{2m-1} \cos(n(\theta_0 - \vartheta_k)) = \begin{cases} 2m\cos(n\theta)\cos(n\theta_0) & \text{if } n \text{ is a multiple of } m, \\ 0 & \text{otherwise.} \end{cases}$$

Using the above properties and $m = \pi/\alpha$, we obtain

$$\mathbb{P}^x(X_t \in dy) = \frac{2r}{t\alpha} e^{-(r^2+r_0^2)/(2t)} \left(\frac{1}{2} I_0\left(\frac{rr_0}{t}\right) + \sum_{n=1}^{\infty} I_{n\pi/\alpha}\left(\frac{rr_0}{t}\right) \cos\left(\frac{n\pi\theta}{\alpha}\right) \cos\left(\frac{n\pi\theta_0}{\alpha}\right) \right) dr d\theta, \tag{10.10.4}$$

which gives a formula which does not depend on m. Now, we have to check that this formula is well defined for any $\alpha \in (0, \pi)$ and that it is the unique solution the partial differential equation (10.10.1) with $\mathcal{D} = \langle \alpha \rangle$.

This is explained in Corollary 10.10.1. In our current situation, $d = 2$, the eigenvalues are $\lambda_j := (j\pi/\alpha)^2$ for $j \geq 0$ and the eigenfunctions are $m_j(\theta) = \sqrt{2/\alpha}\cos(j\pi\theta/\alpha)$, $j \geq 1$ and $m_0 \equiv 1/\sqrt{\alpha}$. □

Corollary 10.10.1. *Consider a general d-dimensional cone generated by all rays emanating from the origin and passing through a compact subset $D \subset \mathbb{S}^{d-1}$ which has smooth boundary. Consider X to be the normally reflected Brownian motion at the boundary of the cone. Then X_t has a density given by*

$$\mathbb{P}^x(X_t \in dy) = \frac{r e^{-\frac{r^2+r_0^2}{2t}}}{t(rr_0)^{d/2-1}} \left(I_{\alpha_0}\left(\frac{rr_0}{t}\right) m_0(\theta) m_0(\theta_0) + \sum_{n=1}^{\infty} I_{\alpha_n}\left(\frac{rr_0}{t}\right) m_n(\theta) m_n(\theta_0) \right) dr d\theta. \tag{10.10.5}$$

To explain the elements in the above formula, denote by $L_{\mathcal{S}^{d-1}}$, the Laplace-Beltrami operator on $\mathbb{S}^{d-1}$. With the above assumptions, there exists a complete set of orthonormal eigenfunctions m_j with corresponding eigenvalues $0 \leq \lambda_0 < \lambda_1 \leq \lambda_2 < \dots$ satisfying

$$\begin{cases} L_{\mathcal{S}^{d-1}} m_j(x) = -\lambda_j m_j(x) & \text{for } x \in D \\ \nabla m_j(x) \cdot n(x) = 0 & \text{for } x \in \partial D, \end{cases}$$

$$\alpha_j = \sqrt{\lambda_j + \left(\frac{d}{2} - 1\right)^2}.$$

Proof. The beginning of the proof is the same as in the statements following (10.10.1) in what refers to the existence and uniqueness of the associated PDE. Existence and uniqueness for the reflected process in the generalized cone can be deduced from [Bas96] (see also [BBC05, Remark 4.1] and the references therein).

In order to prove that (10.10.5) satisfies the associated PDE, one follows a similar proof in [BnS97] for the killed case. The proof in the reflected case follows line by line, the proof in [BnS97], except that the required estimates and properties of the eigenvalues and eigenfunctions of the Laplace-Beltrami operator in the case of Neumann boundary conditions have to be referred to the proper literature (for this see, [GN13], [Gri02] and [Kro92]) although the estimates do not change as far as it relates to the proof of [BnS97, Lemma 1] with the corresponding corrections of typos. In order for the required estimates to be satisfied one needs that the generalized d-dimensional cone is generated by all rays emanating from the origin and passing through a compact subset $D \subset \mathbb{S}^{d-1}$ which has smooth boundary.

We also remark that the density expressions in [BnS97] are written under polar measures which explains why our expressions have an extra r which appears due to the Jacobian of the change of coordinates. □

We also provide a full elementary proof that (10.10.4) satisfies the initial conditions in the Supplementary Material.

10.11 Appendix: A hint for higher dimensions

The following proposition is provided so as to hint at the possibilities in dimension higher than two. It shows that dimension two is the case which is mathematically difficult to treat.

Proposition 10.11.1. *Let $(R_t)_{t\geq 0}$ be a Bessel process in dimension $d \geq 3$ and let $F(T) = \int_0^t ds/R_s^2$. Then:*

$$\mathbb{E}^{r_0}\left[F(T)\right] < +\infty.$$

Proof. The density of R_t in r is given in (10.9.7), so by performing the change of variable $r = ut$, we have

$$\begin{aligned}\mathbb{E}^{r_0}\left[F(T)\right] &= \int_0^T \int_0^\infty \frac{1}{rt}\left(\frac{r}{r_0}\right)^\nu e^{-\frac{r^2+r_0^2}{2t}} I_\nu\left(\frac{rr_0}{t}\right) drdt \\ &= \frac{1}{r_0^\nu}\int_0^T t^{\nu-1}\int_0^\infty I_\nu(ur_0)e^{-\frac{tu^2}{2}-\frac{r_0^2}{2t}}u^{\nu-1}dudt.\end{aligned}$$

Note that for all $\varepsilon > 0$,

$$\int_\varepsilon^T t^{\nu-1}\int_0^\infty I_\nu(ur_0)e^{-\frac{tu^2}{2}-\frac{r_0^2}{2t}}u^{\nu-1}dudt \leq \left(\int_\varepsilon^T e^{-\frac{r_0^2}{2t}}t^{\nu-1}dt\right)\left(\int_0^\infty I_\nu(ur_0)e^{-\varepsilon u^2/2}u^{\nu-1}du\right) < \infty,$$

where for we used the Proposition 10.9.2 for the convergence of the second integral. So to prove the convergence of the integral, we only need to prove the convergence for the integral in t around zero. To do so we use Proposition 10.9.2 again so that for $\nu,\ r_0 > 0$:

$$\mathbb{E}^{r_0}\left[F(\varepsilon)\right] \leq \frac{1}{r_0^\nu}\int_0^\varepsilon t^{\nu-1}e^{-\frac{r_0^2}{2t}}\int_0^\infty\left(e^{ur_0} + \frac{1}{\pi(\nu+ur_0)}\right)e^{-\frac{tu^2}{2}}u^{\nu-1}dudt < \infty.$$

In fact, using that $\nu > 0$, we have

$$I_1 := \frac{1}{r_0^\nu}\int_0^\varepsilon t^{\nu-1}\int_0^\infty e^{-\frac{t}{2}\left(u-\frac{r_0}{t}\right)^2}u^{\nu-1}dudt \le \frac{1}{r_0^\nu}\int_0^\varepsilon t^{\nu-1}\int_{-\infty}^\infty e^{-\frac{t}{2}u^2}\left(|u|+\frac{r_0}{t}\right)^{\nu-1}dudt$$
$$= \frac{2}{r_0^\nu}\int_0^\varepsilon t^{\nu-3/2}\int_0^\infty e^{-u^2/2}\left(\frac{u}{\sqrt{t}}+\frac{r_0}{t}\right)^{\nu-1}dudt \le \frac{2}{r_0^\nu}\int_0^\varepsilon t^{\nu-3/2}\int_0^\infty e^{-u^2/2}\left(\frac{u+r_0}{t}\right)^{\nu-1}dudt$$
$$= \frac{2}{r_0^\nu}\int_0^\varepsilon t^{-1/2}\int_0^\infty e^{-u^2/2}\left(u+r_0\right)^{\nu-1}dudt < \infty.$$

As for the second integral, we have

$$I_2 := \frac{1}{r_0^\nu}\int_0^\varepsilon t^{\nu-1}\int_0^\infty \frac{1}{\pi(\nu+ur_0)}e^{-\frac{tu^2}{2}-\frac{r_0^2}{2t}}u^{\nu-1}dudt \le \frac{1}{r_0^\nu}\int_0^\varepsilon t^{\nu-1}e^{-\frac{r_0^2}{2t}}dt\int_0^\infty \frac{u^{\nu-1}}{\pi(\nu+ur_0)}du < \infty.$$

□

Acknowledgements

The first author thanks École Normale Supérieure, Département de Mathématiques et Applications for providing a financial support for a visit to Ritsumeikan University.

The second author is supported in part by KAKENHI 20K03666.

Both authors thank the anonymous referee for a careful review and for correcting mistakes and providing references.

10.12 Supplementary material

10.12.1 Proof of the convergence to the initial condition when $t \to 0$

In this section, we prove the following result:

Proposition 10.12.1. *The formula* (10.4.4) *satisfies the initial conditions of the heat equation, i.e. for all functions $f : \mathcal{D} \to \mathbb{R}$ continuous with compact support, $r_0 > 0$ and $\theta_0 \in (0, \alpha)$:*

$$\int_{\mathcal{D}} \hat{f}(r,\theta)\frac{2r}{t\alpha}e^{-(r^2+r_0^2)/2t}\left(\frac{1}{2}I_0\left(\frac{rr_0}{t}\right)+\sum_{n=1}^\infty I_{n\pi/\alpha}\left(\frac{rr_0}{t}\right)\cos\left(\frac{n\pi\theta}{\alpha}\right)\cos\left(\frac{n\pi\theta_0}{\alpha}\right)\right)drd\theta \tag{10.12.1}$$

$$\xrightarrow[t\to 0]{} \hat{f}(r_0,\theta_0), \tag{10.12.2}$$

where $\hat{f} : [0,\infty)\times[0,\alpha] \to \mathbb{R}$ denotes the function f expressed in polar coordinates.

For the proof, we use the following representation:

$$\forall x > 0,\ \forall \alpha \ge 0,\ I_\alpha(x) = \frac{1}{\pi}\int_0^\pi e^{x\cos(u)}\cos(\alpha u)du - \frac{\sin(\alpha\pi)}{\pi}\int_0^\infty e^{-x\cosh(u)-\alpha u}du.$$

Let us denote by A_1 and A_2 the two terms appearing in the integral representation in (10.12.2) after replacing with the above formula, with $d_n = \frac{1}{2}$ if $n = 0$ and $d_n = 1$ for $n \ge 1$:

$$A_1 := \frac{2}{\alpha\pi}\int_{\mathcal{D}}\frac{r}{t}\hat{f}(r,\theta)e^{-\frac{r^2+r_0^2}{2t}}\sum_{n=0}^\infty d_n\cos\left(\frac{n\pi\theta}{\alpha}\right)\cos\left(\frac{n\pi\theta_0}{\alpha}\right)\int_0^\pi e^{\frac{rr_0}{t}\cos(u)}\cos\left(\frac{n\pi}{\alpha}u\right)dudrd\theta,$$

$$A_2 := -\frac{2}{\alpha\pi}\int_{\mathcal{D}}\frac{r}{t}\hat{f}(r,\theta)e^{-\frac{r^2+r_0^2}{2t}}\sum_{n=1}^\infty \sin\left(\frac{n\pi^2}{\alpha}\right)\cos\left(\frac{n\pi\theta}{\alpha}\right)\cos\left(\frac{n\pi\theta_0}{\alpha}\right)\int_0^\infty e^{-\frac{rr_0}{t}\cosh(u)-\frac{n\pi}{\alpha}u}dudrd\theta.$$

Proposition 10.12.2. *We have* $A_2 \underset{t\to 0}{\longrightarrow} 0$.

Proof. Step 1: For the moment let us assume that the integration order can be switched. For all $a \in \mathbb{R}$ and $b > 0$, we have:

$$g(a,b) := \sum_{n=0}^{\infty} \sin(na)e^{-nb} = \Im\left(\sum_{n=0}^{\infty} e^{ina-nb}\right) = \frac{e^{-b}\sin(a)}{\left(1-e^{-b}\cos(a)\right)^2 + e^{-2b}\sin^2(a)}. \tag{10.12.3}$$

We use the following trigonometric identity written in compact form

$$4\cos(a)\cos(b)\sin(c) = \sin(a+b+c) + \sin(c-a-b) + \sin(a-b+c) + \sin(c-a+b) \tag{10.12.4}$$

$$A_{jk}(\theta) := \frac{\pi}{\alpha}((-1)^j\theta + (-1)^k\theta_0 + \pi),$$

$$\sum_{n=1}^{\infty} \cos\left(\frac{n\pi\theta}{\alpha}\right)\cos\left(\frac{n\pi\theta_0}{\alpha}\right)\sin\left(\frac{n\pi^2}{\alpha}\right)e^{-\frac{n\pi}{\alpha}u} = \frac{1}{4}\sum_{j,k=0}^{1} g\left(A_{jk}(\theta), \frac{\pi}{\alpha}u\right).$$

Now we prove that for all $j,k \in \{0,1\}$, we have:

$$J_{j,k} := \frac{1}{t}\int_{\mathcal{D}} r\hat{f}(r,\theta)e^{-\frac{r^2+r_0^2}{2t}}\int_0^{\infty} e^{-\frac{rr_0}{t}\cosh(u)} g\left(A_{jk}(\theta), \frac{\pi}{\alpha}u\right) dud\theta dr \underset{t\to 0}{\longrightarrow} 0.$$

We will do the analysis of $J_{j,k}$ dividing the integration region in two: $J^1_{j,k}$ which comprises the integral on the region $\mathcal{D}\times(1,\infty)$ and the remaining which is denoted by $J^2_{j,k}$. With this in mind, note that g is continuous on $[0,\infty)\times(0,\alpha)$ and is locally integrable in $(0,0)$. In fact, for all $a \in \mathbb{R}$ and $b > 0$:

$$g(a,b) = \frac{e^{-b}\sin(a)}{\left(1-e^{-b}\cos(a)\right)^2 + e^{-2b}\sin^2(a)} \underset{a,b\to 0}{\sim} \frac{a}{\left(\left(1-e^{-b}\left(1-\frac{a^2}{2}\right)\right)^2 + a^2\right.} \sim \frac{a}{a^2+b^2}. \tag{10.12.5}$$

Moreover,

$$\int_{[0,1]^2} \frac{a}{a^2+b^2} dadb = \int_0^1 \left[\frac{1}{2}\log(a^2+b^2)\right]_{a=0}^{a=1} db = \frac{1}{2}\int_0^1 \left(\log(1+b^2) - \log(b^2)\right) db < \infty.$$

Next, note that $\partial_b g(a,b)$ is negative, so

$$\forall u \geq 1,\ \forall \theta \in \mathbb{R},\ \left|g\left(A_{jk}(\theta), \frac{\pi}{\alpha}u\right)\right| \leq \frac{e^{-\frac{\pi}{\alpha}}}{(1-e^{\frac{\pi}{\alpha}})^2},$$

and since $u \mapsto u^2/\cosh(u)$ is non-negative and bounded above, there exists $B > 0$ such that

$$|J^1_{jk}| \leq \frac{C_1}{t}\int_0^{\infty} re^{-\frac{r^2+r_0^2}{2t}}\int_1^{\infty} e^{-B\frac{rr_0}{t}u^2} dudr \leq \frac{C_2}{\sqrt{tr_0}} e^{-\frac{r_0^2}{2t}}(2t)^{3/4}\int_0^{\infty}\sqrt{r'}e^{-r'^2}dr' \underset{t\to 0}{\longrightarrow} 0.$$

On the other hand, using (10.12.5) and the continuity of g we obtain

$$|J^2_{jk}| \leq ||f||_{\infty}\frac{1}{t}\int_0^{\infty} re^{-\frac{r^2+r_0^2}{2t}}dr \cdot \int_0^1\int_0^{\alpha}\left|g\left(A_{jk}(\theta), \frac{\pi}{\alpha}u\right)\right| d\theta du \underset{t\to 0}{\longrightarrow} 0.$$

So that

$$|A_2| \leq \frac{1}{2\alpha\pi}\sum_{j,k=0}^{1}|J_{jk}| \underset{t\to 0}{\longrightarrow} 0.$$

Step 2: Now, we prove that one can interchange the order of the integrals and sum in A_2. Note that Fubini's theorem does not apply here because of the factors $\cos(n\pi\theta/\alpha)$ and $\cos(n\pi\theta_0/\alpha)$. A computation similar to (10.12.3) leads to, for all $N \in \mathbb{N}$:

$$g_N(a,b) := \sum_{n=N}^{\infty} \sin(na)e^{-nb} = \Im\left(\sum_{n=N}^{\infty} e^{-nb+ina}\right) = e^{-Nb}\frac{\sin(Na) - e^{-b}\sin((N-1)a)}{(1-e^{-b}\cos(a))^2 + e^{-2b}\sin^2(a)}.$$

Next, we prove that, for all $j,k \in \{0,1\}$ and for all $t>0$, denoting R_{jk}^N the difference between the infinite sum and the partial sum up in A_2 to N,

$$R_{jk}^N := \frac{1}{t}\int_{\mathcal{D}} re^{-\frac{r^2+r_0^2}{2t}}\hat{f}(r,\theta)\int_0^{\infty} e^{-\frac{rr_0}{t}\cosh(u)} g_N\left(A_{jk}(\theta), \frac{\pi}{\alpha}u\right) dud\theta dr \underset{N\to\infty}{\longrightarrow} 0.$$

Let us remark that

$$e^{Nb}g_N(a,b) = \frac{\sin(Na) - e^{-b}\sin((N-1)a)}{(1-e^{-b}\cos(a))^2 + e^{-2b}\sin^2(a)} \underset{a,b\to 0}{\sim} \frac{a}{a^2+b^2},$$

so that $(a,b) \mapsto e^{Nb}g_N(a,b)$ is integrable in $(0,0)$. We denote $R_{jk}^{N,1}$ and $R_{jk}^{N,2}$ the two terms obtained after splitting the integral with respect to u on $(0,1)$ and $(1,\infty)$ respectively. We have then

$$|R_{jk}^{N,2}| \leq \frac{1}{t}||f||_{\infty}\int_0^{\infty} re^{-\frac{r^2+r_0^2}{2t}}dr\int_0^1 e^{-\frac{N\pi u}{\alpha}}\int_0^{\alpha} e^{\frac{N\pi u}{\alpha}}\left|g_N\left(A_{jk}(\theta),\frac{\pi}{\alpha}u\right)\right| d\theta du \underset{N\to\infty}{\longrightarrow} 0,$$

where we use Lemma 10.12.3 for the above convergence. Moreover for all $u \geq 1,\ \theta \in \mathbb{R}$,

$$|g_N(A_{jk}(\theta), \frac{\pi}{\alpha}u)| \leq e^{-N\frac{\pi}{\alpha}u}\frac{2}{(1-e^{-\frac{\pi}{\alpha}})^2},$$

so we have

$$|R_{jk}^{N,1}| \leq \frac{1}{t}||f||_{\infty}\int_0^{\infty} re^{-\frac{r^2+r_0^2}{2t}}dr\int_1^{\infty} e^{-N\frac{\pi}{\alpha}u}\frac{2}{(1-e^{-\frac{\pi}{\alpha}})^2}du \underset{N\to\infty}{\longrightarrow} 0.$$

From the above arguments we obtain the conclusion: $R_{jk}^N \to 0$ for all $j,k \in \{0,1\}$.

□

Lemma 10.12.3. *Let $b>0$ and let $f \in L^1((0,b))$ be non-negative and continuous. Then:*

$$\int_0^b e^{-Nx}f(x)dx \underset{N\to\infty}{\longrightarrow} 0.$$

Proof. Let $\varepsilon > 0$ and choose $\delta > 0$ such that $\int_0^{\delta} f(x)dx \leq \varepsilon$. Then for N big enough:

$$\int_0^b e^{-Nx}f(x)dx \leq \int_0^{\delta} f(x)dx + e^{-N\delta}||f||_1 \leq 2\varepsilon.$$

□

Proposition 10.12.4. *We have $A_1 \underset{t\to 0}{\longrightarrow} \hat{f}(r_0,\theta_0)$.*

Proof. First, let us assume that the integration order can be exchanged. This will be further discussed in Step 5.

Step 1 : We exchange the order of integrals in A_1 as:

$$A_1 = \int_0^{\pi} f_1(u,t) f_2(u,t) du,$$
$$f_1(u,t) := \frac{2}{\sqrt{t}\alpha\pi} e^{-\frac{r_0^2 \sin^2(u)}{2t}},$$
$$f_2(u,t) := \int_0^{\infty} \frac{r e^{-\frac{(r-r_0\cos(u))^2}{2t}}}{\sqrt{t}} \sum_{n=0}^{\infty} d_n \cos\left(\frac{n\pi\theta_0}{\alpha}\right) \cos\left(\frac{n\pi}{\alpha}u\right) \int_0^{\alpha} \hat{f}(r,\theta) \cos\left(\frac{n\pi\theta}{\alpha}\right) d\theta dr.$$

We will study the limit of A_1 in the above integral order. First, we treat the sum inside $f_2(u,t)$ using trigonometric identities for $\cos\left(\frac{n\pi(\theta_0 \pm u)}{\alpha}\right)$. We see that is enough to find the limit for:

$$\frac{1}{2}\sum_{n=0}^{\infty} d_n \left(\cos\left(\frac{n\pi(\theta_0+u)}{\alpha}\right) + \cos\left(\frac{n\pi(\theta_0-u)}{\alpha}\right)\right) \int_0^{\alpha} \hat{f}(r,\theta) \cos\left(\frac{n\pi\theta}{\alpha}\right) d\theta.$$

In our case, we will use the following normalization of the classical Fourier inversion formula:

Theorem 10.12.5 (Fourier Formula)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a periodic and continuous function of period $2L$. Then the following series converges for all $x \in \mathbb{R}$ and:*

$$g(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right),$$

$$\textit{with} \quad a_n = \frac{1}{L}\int_{-L}^{L} g(\theta)\cos\left(\frac{n\pi\theta}{L}\right) d\theta \quad \textit{and} \quad b_n = \frac{1}{L}\int_{-L}^{L} g(\theta)\sin\left(\frac{n\pi\theta}{L}\right) d\theta.$$

Next, we extend the definition of the function $\hat{f}$. For all $r \geq 0$ and $\theta \in [0,\alpha]$, we define $\hat{f}(r,-\theta) := \hat{f}(r,\theta)$, and we make it 2α-periodic by defining $\hat{f}(r,\theta+2k\alpha) = \hat{f}(r,\theta)$. This way, $\hat{f}$ is an even and 2α-periodic function and is still continuous. With this definition, and using the Fourier inversion formula, we get for $\theta_0 \pm u \in (-\alpha,\alpha)$:

$$\begin{aligned}
&\frac{1}{2}\sum_{n=0}^{\infty} d_n \cos\left(\frac{n\pi(\theta_0 \pm u)}{\alpha}\right) \int_0^{\alpha} \hat{f}(r,\theta) \cos\left(\frac{n\pi\theta}{\alpha}\right) d\theta \\
&= \frac{1}{4}\sum_{n=0}^{\infty} d_n \cos\left(\frac{n\pi(\theta_0 \pm u)}{\alpha}\right) \int_{-\alpha}^{\alpha} \hat{f}(r,\theta) \cos\left(\frac{n\pi\theta}{\alpha}\right) d\theta = \frac{\alpha}{4}\hat{f}(r,\theta_0 \pm u). \qquad (10.12.6)
\end{aligned}$$

Step 2 : Next, let us study $f_2(u,t)$ in two separate cases the integral in r:

$$\int_0^{\infty} \hat{f}(r,\theta_0 \pm u) \frac{r e^{-\frac{(r-r_0\cos(u))^2}{2t}}}{\sqrt{t}} dr.$$

Case 1: If $r_0\cos(u) > 0$:

$$\frac{r e^{-\frac{(r-r_0\cos(u))^2}{2t}}}{\sqrt{t}} = \frac{1}{\sqrt{t}}(r - r_0\cos(u)) e^{-\frac{(r-r_0\cos(u))^2}{2t}} + r_0\cos(u)\frac{e^{-\frac{(r-r_0\cos(u))^2}{2t}}}{\sqrt{t}}.$$

The total mass of the first term is

$$\int_0^\infty \frac{1}{\sqrt{t}}(r - r_0\cos(u))e^{-\frac{(r-r_0\cos(u))^2}{2t}}dr = \left[-t\frac{e^{-\frac{(r-r_0\cos(u))^2}{2t}}}{\sqrt{t}}\right]_{r=0}^{\infty} = \sqrt{t}e^{-\frac{r_0^2\cos^2(u)}{2t}} \underset{t\to 0}{\longrightarrow} 0.$$

And the second term is, up to the multiplicative constant $\sqrt{2\pi}$, an approximation of the unity around $r = r_0\cos(u) > 0$, so that in this case, the integral in r converges to

$$\sqrt{2\pi}r_0\cos(u)\hat{f}(r_0\cos(u), \theta_0 \pm u).$$

Case 2: If $r_0\cos(u) \leq 0$: The total mass of $\frac{1}{\sqrt{t}}re^{-(r-r_0\cos(u))^2/(2t)}$ is bounded above by

$$\frac{1}{\sqrt{t}}\int_0^\infty re^{-\frac{r^2}{2t}}dr = \sqrt{t} \underset{t\to 0}{\longrightarrow} 0,$$

so that the integral in r converges to 0. We remark here that convergences in the above two cases are uniform with respect to u within their respective domains.

Step 3 : From the previous step, we consider now the integral with respect to u. Taking into consideration the previous step, we can restrict to the case $\cos(u) > 0$ or equivalently $u \in [0, \pi/2]$. That is, consider

$$I_\pm := \frac{1}{\sqrt{t}}\int_0^{\pi/2} \hat{f}(r_0\cos(u), \theta_0 \pm u)e^{-\frac{r_0^2\sin^2(u)}{2t}}\cos(u)du.$$

Note that for all $\varepsilon > 0$, by dominated convergence,

$$\frac{1}{\sqrt{t}}\int_\varepsilon^{\pi/2} \hat{f}(r_0\cos(u), \theta_0 \pm u)e^{-\frac{r_0^2\sin^2(u)}{2t}}\cos(u)du \underset{t\to 0}{\longrightarrow} 0.$$

Fix $\delta > 0$ and take ε small enough such that $\cos(u) \simeq 1$ and $\sin(u) \simeq u$ for all $u \in [0, \varepsilon]$. Then

$$\frac{1}{\sqrt{t}}\int_0^\varepsilon e^{-\frac{r_0^2\sin^2(u)}{2t}}\cos(u)du \underset{t\to 0}{\simeq} \frac{1}{\sqrt{t}}\int_0^\varepsilon e^{-\frac{r_0^2u^2}{2t}}du \underset{t\to 0}{\longrightarrow} \sqrt{\frac{\pi}{2r_0^2}}.$$

Thus, up to the multiplicative constant $\sqrt{\pi/(2r_0^2)}$, $u \mapsto e^{-r_0^2\sin^2(u)/(2t)}\cos(u)$ is an approximation of the unity around $u = 0$, so that

$$I_+ + I_- \underset{t\to 0}{\longrightarrow} \sqrt{\frac{2\pi}{r_0^2}}\hat{f}(r_0, \theta_0).$$

Step 4 : Now, we put all previous steps together. In Step 1, we proved that for all $t > 0$:

$$f_2(u,t) := \int_0^\infty \frac{re^{-\frac{(r-r_0\cos(u))^2}{2t}}}{\sqrt{t}}\frac{\alpha}{4}(\hat{f}(r, \theta_0 + u) + \hat{f}(r, \theta_0 - u))dr.$$

We have proved in Step 2 that for all u, $f_2(u,t)$ converges when $t \to 0$ to

$$\overline{f}_2(u) := \frac{\alpha}{4}\sqrt{2\pi}r_0\cos(u)\left(\hat{f}(r_0\cos(u), \theta_0 + u) + \hat{f}(r_0\cos(u), \theta_0 - u)\right),$$

and in Step 3 that $\int_0^\pi f_1(u,t)\overline{f}_2(u)du$ converges when $t \to 0$ to

$$\lim_{t\to 0}\frac{2}{\alpha\pi\sqrt{t}}\int_0^{\pi/2}\frac{\alpha}{4}\sqrt{2\pi}r_0(\hat{f}(r_0\cos(u), \theta_0 + u) + \hat{f}(r_0\cos(u), \theta_0 - u))e^{-\frac{r_0^2\sin^2(u)}{2t}}\cos(u)du$$

$$=\frac{2}{\alpha\pi}\frac{\alpha}{4}\sqrt{2\pi}r_0\sqrt{\frac{\pi}{2r_0^2}}2\hat{f}(r_0,\theta_0)=\hat{f}(r_0,\theta_0).$$

To end the proof of the convergence, we have to show that

$$\lim_{t\to 0}\int_0^\pi f_1(u,t)f_2(u,t)du=\lim_{t\to 0}\int_0^\pi f_1(u,t)\overline{f}_2(u)du.$$

We have:

$$\left|\int_0^\pi f_1(u,t)f_2(u,t)du-f_1(u,t)\overline{f}_2(u)du\right|\leq\left(\sup_t\int_0^\pi f_1(u,t)du\right)||f_2(\cdot,t)-\overline{f}_2(\cdot)||_\infty,$$

and $||f_2(\cdot,t)-\overline{f}_2(\cdot)||_\infty\to 0$ since in Step 2, the convergence is uniform with respect to u.

Step 5 : We now prove that for all $r_0,\ \theta_0,\ t$ fixed, the integration order can be switched. Note that for fixed n, by Fubini's theorem, the integration order in r, θ and u can be switched, as

$$\int_0^\infty\int_0^\alpha\left|\hat{f}(r,\theta)e^{-\frac{r^2+r_0^2}{2t}}\cos\left(\frac{n\pi\theta}{\alpha}\right)\cos\left(\frac{n\pi\theta_0}{\alpha}\right)\right|\int_0^\pi\left|e^{\frac{rr_0}{t}\cos(u)}\cos\left(\frac{n\pi u}{\alpha}\right)\right|dud\theta dr$$
$$\leq||f||_\infty\alpha\pi\int_0^\infty e^{-\frac{r^2+r_0^2}{2t}}e^{\frac{rr_0}{t}}dr<\infty.$$

So that the order of integration can be exchanged for every partial sum. Now, for $N\in\mathbb{N}$:

$$\left|\int_0^\pi e^{-\frac{r_0^2\sin^2(u)}{2t}}\int_0^\infty re^{-\frac{(r-r_0\cos(u))^2}{2t}}\sum_{n=0}^N d_n\cos\left(\frac{n\pi\theta_0}{\alpha}\right)\cos\left(\frac{n\pi}{\alpha}u\right)\int_0^\alpha\hat{f}(r,\theta)\cos\left(\frac{n\pi\theta}{\alpha}\right)d\theta drdu\right.$$
$$\left.-\int_0^\pi e^{-\frac{r_0^2\sin^2(u)}{2t}}\int_0^\infty re^{-\frac{(r-r_0\cos(u))^2}{2t}}\sum_{n=0}^\infty d_n\cos\left(\frac{n\pi\theta_0}{\alpha}\right)\cos\left(\frac{n\pi}{\alpha}u\right)\int_0^\alpha\hat{f}(r,\theta)\cos\left(\frac{n\pi\theta}{\alpha}\right)d\theta drdu\right|$$
$$\leq\frac{\alpha}{4}\int_0^\infty re^{-\frac{r^2+r_0^2}{2t}}\left(\int_{-\pi}^\pi e^{\frac{rr_0}{t}\cos(u)}|\hat{f}_N(r,\theta_0+u)-\hat{f}(r,\theta_0+u)|du\right)dr$$

where $\hat{f}_N$ denotes the N^{th} partial Fourier sum of $\hat{f}$. Then, by Cauchy-Schwarz inequality:

$$\leq\frac{\alpha}{4}\int_0^\infty re^{-\frac{r^2+r_0^2}{2t}}\sqrt{\int_{-\pi}^\pi\left(\hat{f}_N(r,\theta_0+u)-\hat{f}(r,\theta_0+u)\right)^2du}\sqrt{\int_{-\pi}^\pi e^{\frac{2rr_0}{t}\cos(u)}du}dr$$
$$\leq\frac{\alpha}{4}\sqrt{\left\lceil\frac{\pi}{\alpha}\right\rceil}\sup_r||\hat{f}_N(r,\cdot)-\hat{f}(r,\cdot)||_2\underbrace{\int_0^\infty re^{-\frac{r^2+r_0^2}{2t}}\sqrt{2\pi}e^{\frac{rr_0}{t}}dr}_{<+\infty}.$$

Using Parseval's equality, we obtain

$$\begin{aligned}||\hat{f}_N(r,\cdot)-\hat{f}(r,\cdot)||_2^2&=\sum_{n=N+1}^\infty\int_{-\alpha}^\alpha\cos\left(\frac{n\pi\xi}{\alpha}\right)^2d\xi\frac{1}{\alpha^2}\left(\int_{-\alpha}^\alpha\hat{f}(r,\theta)\cos\left(\frac{n\pi\theta}{\alpha}\right)d\theta\right)^2\\&=\sum_{n=N+1}^\infty\int_{-\alpha}^\alpha\cos\left(\frac{n\pi\xi}{\alpha}\right)^2d\xi\cdot\frac{1}{n^2\pi^2}\left(\int_{-\alpha}^\alpha\partial_\theta\hat{f}(r,\theta)\sin\left(\frac{n\pi\theta}{\alpha}\right)d\theta\right)^2\\&\leq\sum_{n=N+1}^\infty\frac{4\alpha^3}{n^2\pi^2}||\partial_\theta\hat{f}||_\infty.\end{aligned}$$

Although the extension of $\hat{f}$ on $[-\alpha,\alpha]$ is not an element of $\mathcal{C}^1$, we can perform the integration by parts on each interval $[0,\alpha]$ and $[-\alpha,0]$. Then, $||\hat{f}_N(r,\cdot)-\hat{f}(r,\cdot)||_2\to 0$ as $N\to\infty$ uniformly in r. That way we obtain the convergence to 0 of the difference between the partial sum and the series. □

Bibliography

[AAB+15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[AH10] K. B. Athreya and Chii-Ruey Hwang. Gibbs measures asymptotics. *Sankhya A*, 72(1):191–207, 2010.

[AJ18] Eduardo Abi Jaber. *Stochastic Invariance and Stochastic Volterra Equations*. Theses, Université Paris sciences et lettres, October 2018.

[AJ22] Eduardo Abi Jaber. The characteristic function of Gaussian stochastic volatility models: an analytic expression. *Finance Stoch.*, 26(4):733–769, 2022.

[ALV07] Elisa Alòs, Jorge A. León, and Josep Vives. On the short-time behavior of the implied volatility for jump-diffusion models with stochastic volatility. *Finance Stoch.*, 11(4):571–589, 2007.

[Ani19] Chandrasekaran Anirudh Bhardwaj. Adaptively Preconditioned Stochastic Gradient Langevin Dynamics. *arXiv e-prints*, page arXiv:1906.04324, June 2019.

[AR73] Antonio Ambrosetti and Paul H. Rabinowitz. Dual variational methods in critical point theory and applications. *J. Functional Analysis*, 14:349–381, 1973.

[Aro04] Bouhari Arouna. Adaptative Monte Carlo method, a variance reduction technique. *Monte Carlo Methods Appl.*, 10(1):1–24, 2004.

[AS64] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. National Bureau of Standards Applied Mathematics Series, No. 55. U. S. Government Printing Office, Washington, D. C., 1964.

[Bas96] Richard F. Bass. Uniqueness for the Skorokhod equation with normal reflection in Lipschitz domains. *Electron. J. Probab.*, 1:no. 11, approx. 29 pp. 1996.

[BBC05] Richard F. Bass, Krzysztof Burdzy, and Zhen-Qing Chen. Uniqueness for reflecting Brownian motion in lip domains. *Ann. Inst. H. Poincaré Probab. Statist.*, 41(2):197–235, 2005.

[BC08] Krzysztof Burdzy and Zhen-Qing Chen. Discrete approximations to reflected Brownian motion. *Ann. Probab.*, 36(2):698–727, 2008.

[BC15] Jose Blanchet and Xinyun Chen. Steady-state simulation of reflected Brownian motion and related stochastic networks. *Ann. Appl. Probab.*, 25(6):3209–3250, 2015.

[BD96] Odile Brandière and Marie Duflo. Les algorithmes stochastiques contournent-ils les pièges? *Ann. Inst. H. Poincaré Probab. Statist.*, 32(3):395–427, 1996.

[BDMS19] Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. The tamed unadjusted langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

[Bel57] Richard Bellman. *Dynamic programming*. Princeton University Press, Princeton, N. J., 1957.

[BEL15] Sebastien Bubeck, Ronen Eldan, and Joseph Lehec. Finite-time analysis of projected Langevin Monte Carlo. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[BFN22] Christian Bayer, Masaaki Fukasawa, and Shonosuke Nakahara. Short communication: On the weak convergence rate in the discretization of rough volatility models. *SIAM Journal on Financial Mathematics*, 13(2):SC66–SC73, 2022.

[BFP09] O. Bardou, N. Frikha, and G. Pagès. Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling. *Monte Carlo Methods Appl.*, 15(3):173–210, 2009.

[BGT04] Mireille Bossy, Emmanuel Gobet, and Denis Talay. A symmetrized Euler scheme for an efficient approximation of reflected diffusions. *J. Appl. Probab.*, 41(3):877–889, 2004.

[BGTW19] H. Buehler, L. Gonon, J. Teichmann, and B. Wood. Deep hedging. *Quant. Finance*, 19(8):1271–1291, 2019.

[BHLP22] Achref Bachouch, Côme Huré, Nicolas Langrené, and Huyên Pham. Deep neural networks algorithms for stochastic control problems on finite horizon: numerical applications. *Methodol. Comput. Appl. Probab.*, 24(1):143–178, 2022.

[BHT22] Christian Bayer, Eric Joseph Hall, and Raúl Tempone. Weak error rates for option pricing under linear rough volatility. *International Journal of Theoretical and Applied Finance*, 25(07n08):2250029, 2022.

[BJ22] Oumaima Bencheikh and Benjamin Jourdain. Convergence in total variation of the Euler-Maruyama scheme applied to diffusion processes with measurable drift coefficient and additive noise. *SIAM J. Numer. Anal.*, 60(4):1701–1740, 2022.

[BKH23] Pierre Bras and Arturo Kohatsu-Higa. Simulation of reflected Brownian motion on two dimensional wedges. *Stochastic Process. Appl.*, 156:349–378, 2023.

[BL05] Marcello Bertoldi and Luca Lorenzi. Estimates of the derivatives for parabolic operators with unbounded coefficients. *Trans. Amer. Math. Soc.*, 357(7):2627–2664, 2005.

[BM80] Marc A. Berger and Victor J. Mizel. Volterra equations with Itô integrals. II. *J. Integral Equations*, 2(4):319–337, 1980.

[BM18] Jose Blanchet and Karthyek Murthy. Exact simulation of multidimensional reflected Brownian motion. *J. Appl. Probab.*, 55(1):137–156, 2018.

[BnS97] Rodrigo Bañuelos and Robert G. Smits. Brownian motion in cones. *Probab. Theory Related Fields*, 108(3):299–319, 1997.

[BP21] Pierre Bras and Gilles Pagès. Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise. *arXiv e-prints, To appear in Mathematics of Computation*, page arXiv:2109.11669, September 2021.

[BP23a] Pierre Bras and Gilles Pagès. Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise II: Total Variation. *Monte Carlo Methods and Applications*, 2023.

[BP23b] Pierre Bras and Gilles Pagès. Langevin algorithms for markovian neural networks and deep stochastic control. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023.

[BPP22] Pierre Bras, Gilles Pagès, and Fabien Panloup. Total variation distance between two diffusions in small time with unbounded drift: application to the Euler-Maruyama scheme. *Electron. J. Probab.*, 27:1–19, 2022.

[BPRS17] Atılım Güneş Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *J. Mach. Learn. Res.*, 18:Paper No. 153, 43, 2017.

[Bra22] Pierre Bras. Convergence rates of Gibbs measures with degenerate minimum. *Bernoulli*, 28(4):2431 – 2458, 2022.

[Bra23] Pierre Bras. Langevin algorithms for very deep neural networks with application to image classification. *Procedia Computer Science*, 222:303–310, 2023. International Neural Network Society Workshop on Deep Learning Innovations and Applications (INNS DLIA 2023).

[BS15] Tomáš Bajbar and Oliver Stein. Coercive polynomials and their Newton polytopes. *SIAM J. Optim.*, 25(3):1542–1570, 2015.

[BS19] Tomáš Bajbar and Oliver Stein. Coercive polynomials: stability, order of growth, and Newton polytopes. *Optimization*, 68(1):99–124, 2019.

[BSCD13] Christophette Blanchet-Scalliet, Areski Cousin, and Diana Dorobantu. Hitting time for correlated three-dimensional Brownian motion. *HAL*, 2013. hal-00846450v2.

[BST10] Christian Bayer, Anders Szepessy, and Raúl Tempone. Adaptive weak approximation of reflected and stopped diffusions. *Monte Carlo Methods Appl.*, 16(1):1–67, 2010.

[BT96] Vlad Bally and Denis Talay. The law of the Euler scheme for stochastic differential equations. I. Convergence rate of the distribution function. *Probab. Theory Related Fields*, 104(1):43–60, 1996.

[BV22] Gerardo Barrera Vargas. Limit behavior of the invariant measure for langevin dynamics. *Probability and Mathematical Statistics*, 42:143–162, 06 2022.

[CBM15] Marie Chupeau, Olivier Bénichou, and Satya N. Majumdar. Survival probability of a Brownian motion in a planar wedge of arbitrary angle. *Phys. Rev. E (3)*, 91(3):032106, 8, 2015.

[Cer00] Sandra Cerrai. Analytic semigroups and degenerate elliptic operators with unbounded coefficients: a probabilistic approach. *J. Differential Equations*, 166(1):151–174, 2000.

[CGLM08] Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM J. Matrix Anal. Appl.*, 30(3):1254–1279, 2008.

[CHS87] Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in $\mathbf{R}^n$. *SIAM J. Control Optim.*, 25(3):737–753, 1987.

[CJ59] H. S. Carslaw and J. C. Jaeger. *Conduction of Heat in Solids*. Oxford University Press, second edition, 1959.

[CL21] René Carmona and Mathieu Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games I: The ergodic case. *SIAM J. Numer. Anal.*, 59(3):1455–1485, 2021.

[Cle21] Emmanuelle Clement. Hellinger and total variation distance in approximating Levy driven SDEs. *arXiv e-prints*, page arXiv:2103.09648, March 2021.

[CPS98] C. Costantini, B. Pacchiarotti, and F. Sartoretto. Numerical approximation for functionals of reflecting diffusion processes. *SIAM J. Appl. Math.*, 58(1):73–102, 1998.

[CvMBB14] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Cyb89] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2(4):303–314, 1989.

[Dal17] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):651–676, 2017.

[DdVB15] Yann Dauphin, Harm de Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In *Neural Information Processing Systems*, 2015.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[Die10] A. B. Dieker. Reflected Brownian motion. *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

[DL06] Madalina Deaconu and Antoine Lejay. A random walk on rectangles algorithm. *Methodol. Comput. Appl. Probab.*, 8(1):135–151, 2006.

[DM09] A. B. Dieker and J. Moriarty. Reflected Brownian motion in a wedge: sum-of-exponential stationary densities. *Electron. Commun. Probab.*, 14:1–16, 2009.

[DM17] Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587, 2017.

[DM19] Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv e-prints*, page arXiv:1810.08693, October 2018.

[DMS20] Alain Durmus, Éric Moulines, and Eero Saksman. Irreducibility and geometric ergodicity of Hamiltonian Monte Carlo. *The Annals of Statistics*, 48(6):3545 – 3564, 2020.

[DPG+14] Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2933–2941, Cambridge, MA, USA, 2014. MIT Press.

[Dub04] Julien Dubédat. Reflected planar Brownian motions, intertwining relations and crossing probabilities. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(5):539–552, 2004.

[Dup19] Bruno Dupire. Functional Itô calculus. *Quant. Finance*, 19(5):721–729, 2019.

[Ebe16] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields*, 166(3-4):851–886, 2016.

[EEFR18] Omar El Euch, Masaaki Fukasawa, and Mathieu Rosenbaum. The microstructural foundations of leverage effect and rough volatility. *Finance Stoch.*, 22(2):241–280, 2018.

[EFW13] Marcos Escobar, Sebastian Ferrando, and Xianzhang Wen. Three dimensional distribution of Brownian motion extrema. *Stochastics*, 85(5):807–832, 2013.

[EKM97] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. For insurance and finance.

[FP99] Jean-Claude Fort and Gilles Pagès. Asymptotic behavior of a Markovian stochastic algorithm with constant step. *SIAM J. Control Optim.*, 37(5):1456–1482, 1999.

[Fri64] Avner Friedman. *Partial differential equations of parabolic type*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1964.

[FSW22] Peter K. Friz, William Salkeld, and Thomas Wagenhofer. Weak error estimates for rough volatility models. *arXiv e-prints*, page arXiv:2212.01591, December 2022.

[FU21] Masaaki Fukasawa and Takuto Ugai. Limit distributions for the discretization error of stochastic Volterra equations. *arXiv e-prints, to appear in Annals of Applied Probability*, page arXiv:2112.06471, December 2021.

[Fuk17] Masaaki Fukasawa. Short-time at-the-money skew and rough fractional volatility. *Quant. Finance*, 17(2):189–198, 2017.

[Fuk21] Masaaki Fukasawa. Volatility has to be rough. *Quant. Finance*, 21(1):1–8, 2021.

[Gas23] Paul Gassiat. Weak error rates of numerical schemes for rough volatility. *SIAM Journal on Financial Mathematics*, 14(2):475–496, 2023.

[GB10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[GG05] Michael B. Giles and Paul Glasserman. Smoking adjoints: fast evaluation of Greeks in Monte Carlo calculations. Technical Report NA05/15, Oxford University Computing Laboratory, 2005.

[GGKL21] M'hamed Gaïgi, Stéphane Goutte, Idris Kharroubi, and Thomas Lim. Optimal risk management problem of natural resources: application to oil drilling. *Ann. Oper. Res.*, 297(1-2):147–166, 2021.

[Gil07] Michael B. Giles. Monte Carlo evaluation of sensitivities in computational finance. Technical Report NA07/12, Oxford University Computing Laboratory, 2007.

[Gil08] Michael B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[GJR18] Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. Volatility is rough. *Quant. Finance*, 18(6):933–949, 2018.

[GKL18] Stéphane Goutte, Idris Kharroubi, and Thomas Lim. Optimal management of an oil exploitation. *International Journal of Global Energy Issues*, 41(1/2/3/4):69–85, 2018.

[GL00] Siegfried Graf and Harald Luschgy. *Foundations of quantization for probability distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.

[GL08] Emmanuel Gobet and Céline Labart. Sharp estimates for the convergence of the density of the Euler scheme in small time. *Electron. Commun. Probab.*, 13:352–363, 2008.

[GLS90] G. Gripenberg, S. O. Londen, and O. Staffans. *Volterra Integral and Functional Equations*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1990.

[GM91] Saul B. Gelfand and Sanjoy K. Mitter. Recursive stochastic algorithms for global optimization in $\mathbf{R}^d$. *SIAM J. Control Optim.*, 29(5):999–1018, 1991.

[GM05] Emmanuel Gobet and Rémi Munos. Sensitivity analysis using Itô-Malliavin calculus and martingales, and application to stochastic optimal control. *SIAM J. Control Optim.*, 43(5):1676–1713, 2005.

[GMDB16] Caglar Gulcehre, Marcin Moczulski, Misha Denil, and Yoshua Bengio. Noisy activation functions. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 3059–3068. JMLR.org, 2016.

[GN13] D. S. Grebenkov and B.-T. Nguyen. Geometrical structure of Laplacian eigenfunctions. *SIAM Rev.*, 55(4):601–667, 2013.

[GNR86] V. Giorno, A. G. Nobile, and L. M. Ricciardi. On some diffusion approximations to queueing systems. *Adv. in Appl. Probab.*, 18(4):991–1014, 1986.

[Gob01] Emmanuel Gobet. Euler schemes and half-space approximation for the simulation of diffusion in a domain. *ESAIM Probab. Statist.*, 5:261–297, 2001.

[GR07] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products*. Elsevier/Academic Press, Amsterdam, seventh edition, 2007. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX).

[Gri02] D. Grieser. Uniform bounds for eigenfunctions of the Laplacian on manifolds with boundary. *Comm. Partial Differential Equations*, 27(7-8):1283–1299, 2002.

[GS14] Michael B. Giles and Lukasz Szpruch. Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. *The Annals of Applied Probability*, 24(4):1585 – 1620, 2014.

[Ha09] Wonho Ha. *Applications of the reflected Ornstein-Uhlenbeck process*. ProQuest LLC, Ann Arbor, MI, 2009. Thesis (Ph.D.)–University of Pittsburgh.

[Han18] Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In *NeurIPS*, pages 580–589, 2018.

[Has70] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[HE16] Jiequn Han and Weinan E. Deep Learning Approximation for Stochastic Control Problems. *Deep Reinforcement Learning Workshop, NIPS (2016)*, November 2016.

[Hil88] David Hilbert. Ueber die Darstellung definiter Formen als Summe von Formenquadraten. *Math. Ann.*, 32(3):342–350, 1888.

[HLVDMW17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[Hoc91] Sepp Hochreiter. Investigations on dynamic Neural Networks, Untersuchungen zu dynamischen Neuronalen Netzen. Diploma Thesis, Institut für Informatik, Technische Universität München, 1991.

[HRSS21] Kaitong Hu, Zhenjie Ren, David Siska, and Lukasz 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.

[HS96] Sepp Hochreiter and Jürgen Schmidhuber. LSTM can solve hard long time lag problems. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, page 473–479, Cambridge, MA, USA, 1996. MIT Press.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

[Hwa80] Chii-Ruey Hwang. Laplace's method revisited: weak convergence of probability measures. *Ann. Probab.*, 8(6):1177–1182, 1980.

[Hwa81] Chii Ruey Hwang. A generalization of Laplace's method. *Proc. Amer. Math. Soc.*, 82(3):446–451, 1981.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[IKP13] Tomoyuki Ichiba, Ioannis Karatzas, and Vilmos Prokaj. Diffusions with rank-based characteristics and values in the nonnegative quadrant. *Bernoulli*, 19(5B):2455–2493, 2013.

[Iye85] Satish Iyengar. Hitting lines with two-dimensional Brownian motion. *SIAM J. Appl. Math.*, 45(6):983–989, 1985.

[JKRL09] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, 2009.

[Jor96] Philippe Jorion. *Value at Risk: A New Benchmark for Measuring Derivative Risk*. Irwin Professional Publishing, 1996.

[JR20] Paul Jusselin and Mathieu Rosenbaum. No-arbitrage implies power-law market impact and rough volatility. *Math. Finance*, 30(4):1309–1336, 2020.

[JYC09] Monique Jeanblanc, Marc Yor, and Marc Chesney. *Mathematical methods for financial markets*. Springer Finance. Springer-Verlag London, Ltd., London, 2009.

[Kag07] Wouter Kager. Reflected Brownian motion in generic triangles and wedges. *Stochastic Process. Appl.*, 117(5):539–549, 2007.

[KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[KD01] Harold J. Kushner and Paul Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2001. Stochastic Modelling and Applied Probability.

[KGV83] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[KH09] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[Kha12] Rafail Khasminskii. *Stochastic stability of differential equations*, volume 66 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, second edition, 2012. With contributions by G. N. Milstein and M. B. Nevelson.

[KLR18] Vadim Kaushansky, Alexander Lipton, and Christoph Reisinger. Transition probability of Brownian motion in the octant and its application to default modelling. *Appl. Math. Finance*, 25(5-6):434–465, 2018.

[Koh96] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press, 1996.

[Kra04] Ilia Krasikov. New bounds on the Hermite polynomials. *East J. Approx.*, 10(3):355–362, 2004.

[Kro92] P. Kroger. Upper bounds for the Neumann eigenvalues on a bounded domain in Euclidean space. *J. Funct. Anal.*, 106(2):353–357, 1992.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[KZ16] Steven Kou and Haowen Zhong. First-passage times of two-dimensional Brownian motion. *Adv. in Appl. Probab.*, 48(4):1045–1060, 2016.

[Lam21] Andrew Lamperski. Projected Stochastic Gradient Langevin Algorithms for Constrained Sampling and Non-Convex Learning. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2891–2937. PMLR, 15–19 Aug 2021.

[Lan08] Paul Langevin. Sur la théorie du mouvement Brownien. *Comptes-rendus de l'Académie des Sciences*, 146:530–532, 1908.

[Laz92] V. A. Lazarev. Convergence of stochastic approximation procedures in the case of several roots of a regression equation. *Problemy Peredachi Informatsii*, 28(1):75–88, 1992.

[LBBH98] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

[LCCC16] Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 1788–1794. AAAI Press, 2016.

[Lem05] Vincent Lemaire. *Estimation récursive de la mesure invariante d'un processus de diffusion.* Theses, Université de Marne la Vallée, December 2005.

[LG16] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, volume 274 of *Graduate Texts in Mathematics*. Springer, [Cham], french edition, 2016.

[Llo82] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[LP02] Damien Lamberton and Gilles Pagès. Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, 8(3):367–405, 2002.

[LP03] Damien Lamberton and Gilles Pagès. Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift. *Stoch. Dyn.*, 3(4):435–451, 2003.

[LP10] Vincent Lemaire and Gilles Pagès. Unconstrained recursive importance sampling. *Ann. Appl. Probab.*, 20(3):1029–1067, 2010.

[LP12] Sophie Laruelle and Gilles Pagès. Stochastic approximation with averaging innovation applied to finance. *Monte Carlo Methods Appl.*, 18(1):1–51, 2012.

[LP17] Vincent Lemaire and Gilles Pagès. Multilevel Richardson-Romberg extrapolation. *Bernoulli*, 23(4A):2643–2692, 2017.

[LPP23] Mathieu Laurière, Gilles Pagès, and Olivier Pironneau. Performance of a Markovian Neural Network versus dynamic programming on a fishing control problem. *Probability, Uncertainty and Quantitative Risk*, pages –, 2023.

[LSU68] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva. *Linear and quasilinear equations of parabolic type.* Translations of Mathematical Monographs, Vol. 23. American Mathematical Society, Providence, R.I., 1968. Translated from the Russian by S. Smith.

[Lun97] Alessandra Lunardi. Schauder estimates for a class of degenerate elliptic and parabolic operators with unbounded coefficients in $\mathbf{R}^n$. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 24(1):133–164, 1997.

[LXG+15] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 562–570, San Diego, California, USA, 09–12 May 2015. PMLR.

[MCF15] Yian Ma, Tianqi Chen, and Emily B. Fox. A Complete Recipe for Stochastic Gradient MCMC. In *Neural Information Processing Systems*, 2015.

[Met09] Adam Metzler. *Multivariate first-passage models in credit risk*. ProQuest LLC, Ann Arbor, MI, 2009. Thesis (Ph.D.)–University of Waterloo (Canada).

[Met10] Adam Metzler. On the first passage problem for correlated Brownian motion. *Statist. Probab. Lett.*, 80(5-6):277–284, 2010.

[MFT21] Pierre Monmarché, Nicolas Fournier, and Camille Tardif. Simulated annealing in $\mathbb{R}^d$ with slowly growing potentials. *Stochastic Process. Appl.*, 131:276–291, 2021.

[Mic92] Laurent Miclo. Recuit simulé sur $\mathbf{R}^n$. Étude de l'évolution de l'énergie libre. *Ann. Inst. H. Poincaré Probab. Statist.*, 28(2):235–266, 1992.

[MMS20] Mateusz B. Majka, Aleksandar Mijatovic, and Lukasz Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. *The Annals of Applied Probability*, 30(4):1534 – 1581, 2020.

[MNP00] Vladimir Maz'ya, Serguei Nazarov, and Boris Plamenevskij. *Asymptotic theory of elliptic boundary value problems in singularly perturbed domains. Vol. I*, volume 111 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel, 2000. Translated from the German by Georg Heinig and Christian Posthoff.

[MO17] Gaétan Marceau-Caron and Yann Ollivier. Natural Langevin Dynamics for Neural Networks. *arXiv e-prints*, page arXiv:1712.01076, December 2017.

[Moh98] Salah-Eldin A. Mohammed. Stochastic differential systems with memory: Theory, examples and applications. In Laurent Decreusefond, Bernt Øksendal, Jon Gjerde, and Ali Süleyman Üstünel, editors, *Stochastic Analysis and Related Topics VI*, pages 1–77, Boston, MA, 1998. Birkhäuser Boston.

[Mot67] T. S. Motzkin. The arithmetic-geometric inequality. In *Inequalities (Proc. Sympos. Wright-Patterson Air Force Base, Ohio, 1965)*, pages 205–224. Academic Press, New York, 1967.

[MPCB14] Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2924–2932, Cambridge, MA, USA, 2014. MIT Press.

[MPZ21] Stéphane Menozzi, Antonello Pesce, and Xicheng Zhang. Density and gradient estimates for non degenerate Brownian SDEs with unbounded measurable drift. *J. Differential Equations*, 272:330–369, 2021.

[MRR+53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[Nea96] Radford M. Neal. *Monte Carlo Implementation*, pages 55–98. Springer New York, New York, NY, 1996.

[NVL+15] Arvind Neelakantan, Luke Vilnis, Quoc V. Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding Gradient Noise Improves Learning for Very Deep Networks. *arXiv e-prints*, page arXiv:1511.06807, November 2015.

[NW06] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[Pag15] Gilles Pagès. Introduction to vector quantization and its applications for numerics. In *CEMRACS 2013—modelling and simulation of complex systems: stochastic and deterministic approaches*, volume 48 of *ESAIM Proc. Surveys*, pages 29–79. EDP Sci., Les Ulis, 2015.

[Pag18] Gilles Pagès. *Numerical probability*. Universitext. Springer, Cham, 2018. An introduction with applications to finance.

[Pil14] Andrey Pilipenko. *An introduction to stochastic differential equations with reflection*. Number 1 in Lectures in pure and applied mathematics (1). Universitätsverlag Potsdam, 2014.

[PP09] Gilles Pagès and Fabien Panloup. Approximation of the distribution of a stationary Markov process with application to option pricing. *Bernoulli*, 15(1):146–177, 2009.

[PP23] Gilles Pagès and Fabien Panloup. Unadjusted Langevin algorithm with multiplicative noise: Total variation and Wasserstein bounds. *The Annals of Applied Probability*, 33(1):726 – 779, 2023.

[PR20] Gilles Pagès and Clément Rey. Recursive computation of invariant distributions of Feller processes. *Stochastic Process. Appl.*, 130(1):328–365, 2020.

[Pro85] Philip Protter. Volterra equations driven by semimartingales. *Ann. Probab.*, 13(2):519–530, 1985.

[PT13] Sam Patterson and Yee Whye Teh. Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[PW99] Shige Peng and Zhen Wu. Fully coupled forward-backward stochastic differential equations and applications to optimal control. *SIAM J. Control Optim.*, 37(3):825–843, 1999.

[Qui93] J. Ross Quinlan. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning*, ICML'93, page 236–243, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[QZ04] Zhongmin Qian and Weian Zheng. A representation formula for transition probability densities of diffusions and applications. *Stochastic Process. Appl.*, 111(1):57–76, 2004.

[RG11] Lewis Fry Richardson and Richard Tetley Glazebrook. Ix. the approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 210(459-470):307–357, 1911.

[RHW86] D. Rumelhart, G. Hinton, and R Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[RKK18] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

[Roy89] Gilles Royer. A remark on simulated annealing of diffusion processes. *SIAM J. Control Optim.*, 27(6):1403–1408, 1989.

[RS71] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971)*, pages 233–257. Academic Press, New York, 1971.

[RTY21] Alexandre Richard, Xiaolu Tan, and Fan Yang. Discrete-time simulation of stochastic Volterra equations. *Stochastic Process. Appl.*, 141:109–138, 2021.

[RY99] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.

[Sai87] Yasumasa Saisho. Stochastic differential equations for multidimensional domain with reflecting boundary. *Probab. Theory Related Fields*, 74(3):455–477, 1987.

[SBCR16] Umut Simsekli, Roland Badeau, A. Taylan Cemgil, and Gaël Richard. Stochastic Quasi-Newton Langevin Monte Carlo. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 642–651, 2016.

[SBL16] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond. *arXiv e-prints*, page arXiv:1611.07476, 2016.

[SEU+17] Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks. *arXiv e-prints*, page arXiv:1706.04454, 2017.

[SGS15] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[SLH+19] Kumar Shridhar, Joonho Lee, Hideaki Hayashi, Purvanshi Mehta, Brian Kenji Iwana, Seokjun Kang, Seiichi Uchida, Sheraz Ahmed, and Andreas Dengel.

ProbAct: A Probabilistic Activation Function for Deep Neural Networks. *arXiv e-prints*, page arXiv:1905.10761, May 2019.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[SS91] Elias M. Stein and Jeremy C. Stein. Stock price distributions with stochastic volatility: An analytic approach. *The Review of Financial Studies*, 4(4):727–752, 1991.

[SZ15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Tal90] Denis Talay. Second-order discretization schemes of stochastic differential systems for the computation of the invariant law. *Stochastics and Stochastic Reports*, 29(1):13–36, 1990.

[TH12] T. Tieleman and G. E. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. Coursera: Neural Networks for Machine Learning, 2012.

[TKV10] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. *Mining Multi-label Data*, pages 667–685. Springer US, Boston, MA, 2010.

[TT90] Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, 8(4):483–509 (1991), 1990.

[UR01] Stanislav Uryasev and R. Tyrrell Rockafellar. Conditional value-at-risk: optimization approach. In *Stochastic optimization: algorithms and applications (Gainesville, FL, 2000)*, volume 54 of *Appl. Optim.*, pages 411–435. Kluwer Acad. Publ., Dordrecht, 2001.

[VBB+20] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on Bayesian Neural Networks – a Tutorial for Deep Learning Users. *arXiv e-prints*, page arXiv:2007.06823, July 2020.

[Vil09] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[vLA87] P. J. M. van Laarhoven and E. H. L. Aarts. *Simulated annealing: theory and applications*, volume 37 of *Mathematics and its Applications*. D. Reidel Publishing Co., Dordrecht, 1987.

[VW85] S. R. S. Varadhan and R. J. Williams. Brownian motion in a wedge with oblique reflection. *Comm. Pure Appl. Math.*, 38(4):405–443, 1985.

[VZ19] Frederi Viens and Jianfeng Zhang. A martingale approach for fractional Brownian motions and related path dependent PDEs. *Ann. Appl. Probab.*, 29(6):3489–3540, 2019.

[Wan12] Feng-Yu Wang. Coupling and applications. In *Stochastic analysis and applications to finance*, volume 13 of *Interdiscip. Math. Sci.*, pages 411–424. World Sci. Publ., Hackensack, NJ, 2012.

[Wan20] Feng-Yu Wang. Exponential contraction in Wasserstein distances for diffusion semigroups with negative curvature. *Potential Anal.*, 53(3):1123–1144, 2020.

[Wat44] G. N. Watson. *A Treatise on the Theory of Bessel Functions.* Cambridge University Press, Cambridge, England; Macmillan Company, New York, 1944.

[WLP+19] Ziyi Wang, Keuntaek Lee, Marcus A. Pereira, Ioannis Exarchos, and Evangelos A. Theodorou. Deep forward-backward SDEs for min-max control. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6807–6814, 2019.

[WT11] Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688. Omnipress, 2011.

[Zei12] Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *arXiv e-prints*, page arXiv:1212.5701, December 2012.

[Zit08] Pierre-André Zitt. Annealing diffusions in a potential function with a slow growth. *Stochastic Process. Appl.*, 118(1):76–119, 2008.