INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAG002

Thèse de doctorat

Sampling with the Langevin Monte-Carlo

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale n°574 Ecole doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 1er Juin 2021, par

Avetik Karagulyan

Composition du Jury :

Alexandre Tsybakov École nationale de la statistique et de l'administration économique (CREST)	Président du jury
Francis Bach INRIA	Rapporteur
Sébastien Gadat Université Toulouse 1 Capitole (MADS)	Rapporteur
Pierre Alquier RIKEN AIP	Examinateur
Éric Moulines École polytechnique (CMAP)	Examinateur
Arnak Dalalyan École nationale de la statistique et de l'administration économique (CREST)	Directeur de thèse

”We feel free because we lack the very language to articulate our unfreedom.”

Slavoj Žižek

Acknowledgements

First and foremost, I would like to thank Arnak, a professor, a friend, and a mentor who believed in me way before I started speaking French. Thank you for talking to me about MVA, for teaching me Statistics, for taking me as your Ph.D. student, for meeting me here with hospitality, for helping whenever I needed, and for the endless conversations that we had during these years. I am proud to have worked with you, and I hope to continue the collaboration with you in the future.

Huge thanks to Francis and Sébastien for reviewing this manuscript. Also, I want to thank the jury members Sasha, Eric and Pierre for taking their time to participate in the defense.

I'd not be here without my professors and teachers. I want to thank my high-school teacher Nairi Sedrakyan for introducing to me the fundamentals of grown-up maths. Also to my professors from YSU Karen Keryan and Michael Poghosyan for neatly prepared lectures, advanced problem-solvings and general guidance into the academic world. I also want to thank my professors from MVA in ENS Cachan. I would like to thank Julien Mairal, Vianney Perchet, Alexander d'Aspremont, Francis, and Arnak for their formative and deep courses.

I was lucky to do my thesis in CREST, a cold grey building filled with warm people. I want to thank Amir, one of the kindest people I know and the best office-mate ever, for bearing with me over these years. Shout out to Arya for always being there for me, for our endless conversations, and for confusing everyone in the lab about who is Iranian and who is Armenian. I want to thank Badr for setting a high bar for the acknowledgments part of my manuscript and for being the friendliest person on the plateau of Saclay. Shout out to Geoffrey for finding me an apartment with a balcony and for the constant borderline jokes during coffee breaks. Special thanks to Lionel for being a good collaborator, and Simo for showing admirable work ethics. Cheers to Gauthier for the endless exciting stories. Kudos to Arshak for the 20 days we spent together in Paris with 50 percent accuracy. I would also like to thank Nicolas and Gabriel, from the far East office, for giving me a helping hand, whenever needed. The economists Yannick and Jeremy, for the interesting discussions about memes, politics, and sports. Lucie for introducing me to French theater and the rare but hilarious jokes. Christophe, for being always kindly responsive to all my demands when he was the math assistant and Jules for being a great successor of Christophe and for his sharp outfits. Phillip, for the evenings in Buttes aux Cailles and Pierre, for the parties at Kfet. Last but not least, I want to thank all the participants of the coffee breaks which helped me to improve my French: Alexandre, Alexis, Anna, Boris, Cristina, Dang, Etienne, Evgenii, Flore, François-Pierre, Gabriel, Guillaume, Jaouad, Julien, Katiya, Lena, Martin, Meyer, Mohammed, Nicolas, Solenne, Suzanne, Tom, Vianney, Victor-Emmanuel and Yannis.

I also want to thank my friends Gevorg, Olgert, Tigran and Zori for being on Telegram for me during good and bad times. Kudos to my friends in Paris Gayane, Lily, Mariam, Marina, Narek, Nushik, Sona, Zara for their willingness to dance to "Ur vor gnas hetd kgam". Special thanks to Anahit, Karen, and Hayk for staying true friends since day one. Cheers to my niece Julieta for her constant support. Huge thanks to my flatmates Sophie, Coralie, Lambert and MLBCV who shared with me the quarantine months last year.

Last but not least I want to thank my family. My parents, for constantly nourishing my interest in mathematics, without whom I would not be the person I am, and my siblings for showing constant love and support from the distance.

Contents

1 Introduction **1**

1.1 High-dimensional Parametric Statistics 1

1.2 Statistiques paramétriques en grande dimension 9

1.3 Convex Optimization 18

1.4 Sampling from a probability distribution 23

1.5 Langevin sampling 29

1.6 Contributions 38

1.7 Résumé substantiel 40

2 Langevin Monte-Carlo with inaccurate gradient **43**

2.1 Introduction 44

2.2 Guarantees in the Wasserstein distance with accurate gradient 47

2.3 Guarantees for the inaccurate gradient version 52

2.4 Guarantees under additional smoothness 55

2.5 Relation with optimization 58

2.6 Conclusion 60

Appendix to Chapter 2 **62**

2.A Proof of Theorem 4 64

2.B Proof of Theorem 5 65

2.C Proof of Theorem 6 66

2.D Proof of Theorem 7 66

2.E Proof of Theorem 8 67

2.F Proof of Theorem 9 68

2.G Proofs of the lemmas 75

3 Langevin Monte-Carlo with convex potentials **85**

3.1 Introduction 86

3.2 Further precisions on the analyzed methods 87

3.3 How to measure the complexity of a sampling scheme? 91

3.4 Overview of main contributions 92

3.5 Prior work 96

3.6 Precision and computational complexity of the LMC 97

3.7 Precision and computational complexity of KLMC and KLMC2 100

3.8 Bounding moments 104

Appendix to Chapter 3 **107**

3.A Proof of Proposition 13 107

3.B Proof of Proposition 14 109

3.C Proof of Proposition 15 111

3.D Proof of Proposition 12 111

3.E Technical lemmas 115

4 Penalized Langevin Dynamics **123**

4.1 Introduction 123

4.2 Convergence of penalized Langevin dynamics 127

4.3 The counterpart in optimization: penalized gradient flow 131

4.4 Prior work and outlook 134

4.5 Conclusion 135

Appendix to Chapter 4 **136**

4.A Proof of Theorem 14 136

4.B Proof of Theorem 15 138

4.C Proofs of the lemmas used in Theorem 14 141

4.D Proof of Proposition 17 146

4.E (Weakly) convex potentials: what is known and what we can hope for . . 146

4.F Proofs of the lemmas used in Theorem 15 148
4.G Penalized Gradient Flow 152
4.H Examples of functions satisfying condition $\mathbf{A}(D, q)$ 156

5 Summary and Perspectives **161**
5.1 Summary of the Thesis 161
5.2 Perspectives 162

Chapter 1

Introduction

Various problems that applied sciences are encountered with can be modeled in two general fashions. The first is to take the model with all its descriptive parameters and features, which makes the subsequent analysis hard and even intractable. The second is to approximate the model, which leads to computationally feasible methods. This dilemma is apparent everywhere in mathematical modeling. As a result, numerous approximation techniques have been proposed, each adapted to a particular problem.

Sampling from probability distributions enters the scope of the mathematical setting above described. Often, we are faced with complex measures that are impossible to sample from in exact manner. Thus, the problem requires an approximative approach. In this manuscript, we focus on one such technique, called the Langevin sampling method, which have their origins in statistical physics.

This introductory chapter presents the mathematical framework and the historical development of the Langevin Monte-Carlo type algorithms. It starts with general notions of Statistics and Optimization and then discusses the exact and the approximate sampling methods. In the end, we provide a quick overview of the main contributions of the thesis, which are described in detail in the subsequent chapters.

1.1 High-dimensional Parametric Statistics

In this section, we briefly present the general setting of the parametric statistical inference. First, we introduce the mathematical formulation of the problem and the main assumptions, required for the subsequent analysis. Next, we discuss the maximum likelihood and Bayesian estimators. In the third subsection, we present the linear regression problem with its regularized versions. We conclude the section with logistic regression models in both, frequentist and Bayesian settings.

1.1.1 General Notions

The general goal of mathematical statistics is to estimate a certain characteristic of the unknown probability law P^n on $\mathbb{R}^{dn}$, using a dataset $\mathcal{Z}_n = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, which is sampled from this distribution. The data points $\boldsymbol{Z}_k$ are d-dimensional random vectors. In the case when they are independent of each other, P^n can be expressed as a product measure:

$$P^n = \mathcal{L}(\boldsymbol{Z}_1) \otimes \ldots \otimes \mathcal{L}(\boldsymbol{Z}_n).$$

Another important assumption about the distribution P^n is that it belongs to a parametric class:

$$\mathcal{P} := \{P^n_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\},$$

where $\Theta \subset \mathbb{R}^p$ is called the set of parameters and it is known in advance. Finally, in most cases it is convenient for us to have an *identifiable* model. That is, for different values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of the parameter, the corresponding distributions are different from each other:

$$\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \implies P^n_{\boldsymbol{\theta}_1} \neq P^n_{\boldsymbol{\theta}_2}.$$

Thus, the unknown distribution P^n is described by one vector which we denote by $\boldsymbol{\theta}^*$ and we call it the true value of the parameter. The latter means, that in order to estimate the unknown distribution P^n, one needs to estimate the true parameter $\boldsymbol{\theta}^*$. Hence, the estimation problem boils down to finding a function $\widehat{\boldsymbol{\theta}}_n(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, that is close to $\boldsymbol{\theta}^*$ in some probabilistic sense.

Now, that we have described the general mathematical setting of the problem, a question that naturally arises is how to choose "a good" or "the best" estimator $\widehat{\boldsymbol{\theta}}_n$? To make a mathematically justified choice one needs to have a criterion or a quantitative method of comparison between two estimators. This is done using a loss function l, which is a positive function defined on $\Theta \times \Theta$. The risk of the estimator $\widehat{\boldsymbol{\theta}}_n$ at point $\boldsymbol{\theta}$ is defined as follows:

$$\mathcal{R}\big(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}\big) := \mathbb{E}_{\boldsymbol{\theta}} \left[l\Big(\widehat{\boldsymbol{\theta}}_n(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n), \boldsymbol{\theta}\Big)\right],$$

where $\mathbb{E}_{\boldsymbol{\theta}}$ means that the expectation is taken over the sample $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n) \sim P^n_{\boldsymbol{\theta}}$. One of the most common choices for $l(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is the quadratic function $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2$. In this case, the risk is called quadratic risk or the mean square error (MSE). A simple calculation shows that there exists no estimator that is universally (i.e., for all parameter values $\boldsymbol{\theta}$) better in terms of risk. Nevertheless, we can hope for certain asymptotic and/or probabilistic qualities. We say that the estimator $\widehat{\boldsymbol{\theta}}_n$ is consistent if $\mathcal{R}\big(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}\big) \to 0$, when $n \to \infty$.

This means that with the growth of n, our estimator (which depends on n) gets closer to $\boldsymbol{\theta}^*$ in terms of the risk. The latter is a significant feature, as we want to estimate better with the growth of the sample size n. Another important aspect is the asymptotic behavior of the estimator. We call the estimator $\widehat{\boldsymbol{\theta}}_n$ *asymptotically normal*, if it satisfies the following condition:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_p(0, \Sigma(\boldsymbol{\theta})), \quad \text{for every} \quad \boldsymbol{\theta} \in \Theta.$$

Here Σ is the asymptotic covariance matrix which is a function of the parameter $\boldsymbol{\theta}$, that depends on the estimator. One may construct different quantitative criteria, based on Σ, that can be used to compare two estimators $\widehat{\boldsymbol{\theta}}^1$ and $\widehat{\boldsymbol{\theta}}^2$ (see e.g. [FK85, Rao]).

1.1.2 Maximum Likelihood and Bayesian Estimators

Suppose now that the distribution $P^n_{\boldsymbol{\theta}}$ is absolutely continuous w.r.t. some σ-additive measure defined on $\mathbb{R}^{dn}$. Thus, $P^n_{\boldsymbol{\theta}}$ can be characterized by its density $f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta})$. We call *likelihood function* the map $\boldsymbol{\theta} \to L_n(\boldsymbol{\theta})$, defined as

$$L_n(\boldsymbol{\theta}) = f(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, \boldsymbol{\theta}), \quad \text{for} \quad \forall \boldsymbol{\theta} \in \Theta,$$

where $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ is the initially observed sample. The *maximum likelihood estimator* is defined as a maximizer of L_n:

$$\widehat{\boldsymbol{\theta}}^{\mathsf{ML}}_n \in \arg\max_{\boldsymbol{\theta}} L_n(\boldsymbol{\theta}). \tag{MLE}$$

One needs to bear in mind that the SGD is not always unique. See [LC90] for some simple, as well as advanced examples of maximum likelihood estimation. The estimator $\widehat{\boldsymbol{\theta}}^{\mathsf{ML}}_n$ can also be defined as a minimizer of the negative logarithm of the likelihood function L_n. The latter is defined in the following manner:

$$l_n(\boldsymbol{\theta}) := -\frac{1}{n} \log\left(L_n(\boldsymbol{\theta})\right), \tag{1.1}$$

and is referred to as the log-likelihood function. If the data points $\boldsymbol{Z}_k$ are i.i.d then the (MLE) takes the following form:

$$\widehat{\boldsymbol{\theta}}^{\mathsf{ML}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \Big\{ -\frac{1}{n} \sum_{k=1}^{n} \log f(\boldsymbol{Z}_k, \boldsymbol{\theta}) \Big\},$$

where $f(\boldsymbol{z}, \boldsymbol{\theta})$ is the marginal density of $\boldsymbol{Z}_1$. If the log-likelihood is convex w.r.t. $\boldsymbol{\theta}$, then the minimization problem (1.1) can be solved using the methods of convex optimization (see Section 1.3). Under mild assumptions, the MLE is proved to be consistent and asymptotically normal (see [LC70, Dan61]).

Remark 1. *The assumption that the unknown distribution belongs to a parametric class* *can be relaxed in the general case. In the recent literature of statistics the true distribution is supposed to be close to some parametric family in a certain probabilistic sense ([Tsy08]) instead of being represented exactly by some value of the parameter.*

In the previous sections, we did not exploit any topological structure of the parameter space Θ. So far, we have considered the true parameter as an unknown constant vector, which belonged to Θ. In the Bayesian approach, the space is equipped with a measure μ, defined on a sigma algebra Σ_Θ. This measure is often called prior distribution in the literature of Bayesian statistics. Again, we are given a dataset $\mathcal{Z}_n = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, sampled from a distribution $P^n_{\boldsymbol{\theta}^*}$, and we want to infer the unknown parameter $\boldsymbol{\theta}^*$. The measure μ is chosen according to the concrete problem and earlier experience of the practitioner, and it allows to universally quantify the quality of an estimator $\widehat{\boldsymbol{\theta}}_n$. The latter is accomplished using the *integrated risk*:

$$\mathcal{R}\big(\widehat{\boldsymbol{\theta}}_n\big) = \int_\Theta \mathcal{R}\big(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n\big)\mu(d\boldsymbol{\theta}) = \int_\Theta \int_{\mathbb{R}^{dn}} l\big(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)\big) P^n_{\boldsymbol{\theta}}(d\boldsymbol{z}_1, \ldots, d\boldsymbol{z}_n)\mu(d\boldsymbol{\theta}).$$

The integrated risk takes the role of the estimator's average error. Therefore, one would be interested in minimizing this error. The Bayesian estimator is defined as the minimizer of the integrated risk:

$$\widehat{\boldsymbol{\theta}}^B_n \in \arg\min_{\widehat{\boldsymbol{\theta}}_n} \mathcal{R}(\widehat{\boldsymbol{\theta}}_n).$$

The minimization here is taken over the class of all measurable functions $\mathcal{M}(\mathbb{R}^{dn}, \Theta)$. The Bayesian estimator depends on the choice of the loss function l. In the case when l is the quadratic loss, $\widehat{\boldsymbol{\theta}}^B_n$ is given as follows:

$$\widehat{\boldsymbol{\theta}}^B_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) := \frac{\int_\Theta \boldsymbol{\theta} f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta})\mu(d\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta})\mu(d\boldsymbol{\theta})}, \tag{1.2}$$

where $p^n_{\boldsymbol{\theta}}$ is the density of $P^n_{\boldsymbol{\theta}}$. One may notice that the integral in the denominator serves as a normalization constant. Thus, $f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta}) / \int_\Theta f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta})\mu(d\boldsymbol{\theta})$ can be interpreted as a density function of a probability distribution defined on (Θ, Σ_Θ). It is often called *posterior distribution*. We will see later, that the normalizing constant can be neglected in many approximate integration techniques. The estimator $\widehat{\boldsymbol{\theta}}^B_n$ is often referred to as *posterior mean*. When the dimension p of the parameter space is large, the computation of the integrals appearing in the last display is generally intractable. To circumvent this problem, one has to resort to approximate calculation methods. One of the main lines of research in this setting is the Monte-Carlo integration and MCMC algorithms. A central ingredient of these algorithms is a method of approximate sampling from a given probability distribution (see Section 1.4.4 below). To complete this subsection, let us note that the behavior of the Bayesian estimator depends on the choice of the loss function l. We refer the reader to [Bro80], for more details on this topic, and to [Rob96, LC06] for a study on intrinsic losses that present some advantages as compared to the quadratic loss.

1.1.3 Multivariate linear regression

In this section, we study the linear regression. Before stating the model itself, let us introduce the notation. We assume that the data points $\boldsymbol{Z}_k = (\boldsymbol{X}_k, Y_k)$ are composite vectors. Here, $\boldsymbol{X}_k \in \mathbb{R}^p$ are called *feature* vectors, whereas $Y_k \in \mathbb{R}$ are called *labels*. We denote by X the *design matrix* of the data $\mathcal{Z}_n$, which is the $d \times n$ matrix, having the features $\boldsymbol{X}_k$ as its rows. Likewise, $\mathrm{Y} := (Y_1, Y_2, \ldots, Y_n)^\top$ is the vector of the labels.

We say that the model is linear, if the labels are a linear transformation of the features up to a mean-zero additive noise term. Mathematically, it has the following form:

$$\mathrm{Y} = \mathrm{X}\,\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}.$$

Here, $\boldsymbol{\varepsilon}$ is a mean-zero random vector in $\mathbb{R}^n$, that is independent of the design matrix X. Let us suppose that $\boldsymbol{\varepsilon}$ is a Gaussian vector: $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then, SGD of $\boldsymbol{\theta}^*$ in this model is written as

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{ML}} \in \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^p} L_n(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^p} \sum_{k=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\big(\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k\big)^2\right).$$

Due to the monotonicity of the exponential function we obtain

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{ML}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{\frac{1}{n}\sum_{k=1}^{n} \big(\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k\big)^2\right\}.$$

This problem is also called *least square estimation*. This is a quadratic minimization problem, which always has a solution. Moreover, if $\operatorname{rank}(\mathrm{X}) = p$, then a closed-form solution is available and it is called ordinary least square (OLS) estimator:

$$\widehat{\boldsymbol{\theta}}^{\mathsf{OLS}} := \big(\mathrm{X}^\top\mathrm{X}\big)^{-1}\mathrm{X}^\top\mathrm{Y}. \tag{OLS}$$

Therefore we have an explicit formula for the solution and it is efficiently computable with modern scientific programming languages. We refer the interested reader to [Kol09] for detailed discussion on this topic, and for the concentration properties of $\widehat{\boldsymbol{\theta}}_n^{\mathsf{OLS}}$.

Regularization

In several frequently encountered situations, the OLS does not have a satisfactory performance from the practitioner's point of view. It suffers from several major issues.

- Although unbiased, OLS has a large variance. Let us denote by Δ the Euclidean distance between the $\widehat{\boldsymbol{\theta}}_n^{\mathsf{ML}}$ and the parameter value $\boldsymbol{\theta}$. Then the expectation and the variance of Δ^2 have the following formula (see [HK70]):

$$\mathbb{E}_{\boldsymbol{\theta}^*}[\Delta^2] = \sum_{k=1}^{p} \frac{2\sigma^2}{\lambda_k} \quad \text{and} \quad \mathrm{var}_{\boldsymbol{\theta}^*}(\Delta^2) = \sum_{k=1}^{p} \frac{2\sigma^4}{\lambda_k^2}.$$

  Here $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\mathrm{X}^\top \mathrm{X}$. We see that the right-hand side of the equation explodes when, the eigenvalues are small. This is often the case in high-dimensional problems.

- Another drawback of the method is the lack of interpretability. OLS does not allow to differentiate between relevant variables and irrelevant variables. One would be interested in sparse estimators. The estimator is called *sparse*, if it has a relatively small number of non-zero entries. Thus, sparsity would lead to a large number of vanishing coordinates for the estimator of $\boldsymbol{\theta}^*$. The latter, in its turn, would yield that the corresponding entries of the feature vector are not useful for predicting the label.

Common approach to solve this issues is to introduce a regularization term $r(\cdot)$. The map $r : \mathbb{R}^+ \to \mathbb{R}^+$ is a monotonically increasing function. The regularized version of least square estimation is the following optimization problem:

$$\widehat{\boldsymbol{\theta}}_n \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{k=1}^{n} (\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k)^2 + \lambda\, r(\|\boldsymbol{\theta}\|) \right\},$$

where λ is a positive parameter and $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^p$. Adding the term $\lambda r(\|\boldsymbol{\theta}\|)$ we thus penalize the norm of the prediction vector. This simple trick allows to have smaller norms, which yields to less complex estimated models. The choice of the norm is based on the complexity notion we are interested in. The regularization is chosen to reflect our prior knowledge on the desirable model parameters. The most popular regularized least square methods, Lasso and ridge, are briefly presented below.

LASSO

In the case when $r(a) = a$, for all $a > 0$, and l_1-norm is taken, the solution of the following minimization problem is called a LASSO estimator:

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{LASSO}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k)^2 + \lambda \left\|\boldsymbol{\theta}\right\|_1 \right\}. \tag{LASSO}$$

This is a convex optimization problem that always has a solution. LASSO was proposed by [Tib96]. The non-smooth behavior of the l_1-norm favors *sparse* minimizers of (LASSO). There is a close relation between the Lasso and soft thresholding, we refer the reader to [DJHS92]. This idea is also exploited in another sparse recovery method called Dantzig selector (see [CT07]). A detailed comparison and a unified analysis of both methods can be found in [BRT+09]. The complexity of LASSO has been extensively studied, see [BTW07, vdG07, BRT+09, AH12, vdGL13]. In [DHL17], the authors show that $\lambda = \sqrt{2\log(p)/n}$ yields to optimal excess risk of order $O(\mathrm{rank}(\mathrm{X})\log(p)/n)$ for (LASSO). Here X is the design matrix of the feature vectors. To gain more insight on the above-mentioned methods, see the survey paper by [Kol09].

Ridge

Ridge regression is the linear regression with the l_2-type regularization term:

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{Ridge}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k)^2 + \lambda \left\|\boldsymbol{\theta}\right\|_2^2 \right\}. \tag{Ridge}$$

The solution to (Ridge) has an explicit formula, similar to (OLS). In the case of ridge regression, however, the design matrix X can have any rank:

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{Ridge}} = \left(\mathrm{X}^\top\mathrm{X} + \lambda I_n\right)^{-1}\mathrm{X}^\top\mathrm{Y}.$$

Since $\lambda > 0$ the inverse of the matrix is always well-defined. The first articles that introduce this method include [Cro72, Hor62]. As mentioned previously, the regularization induces certain bias to reduce the variance in the learning problem (bias-variance trade-off). In this phenomenon, the tuning parameter λ plays a major role. For results on the estimation of λ, the interested reader is referred to [HK70, Nor82, Wen00, Kib03, ST+99a].

1.1.4 Logistic regression

So far, we have discussed only the case, where the labels admit real values. Yet, there is a vast collection of problems when the labels are categorical. These problems are called classification problems. In this section, we will discuss only the case with two classes, that is -1 and 1. Hence, our labels are discrete random variables: $Y_i \in \{-1, 1\}$. The logistic model targets to estimate the conditional distribution of Y knowing the feature vector $\boldsymbol{X}$. The main assumption is that the features can be separated by a hyperplane in $\mathbb{R}^p$, which is described by its normal $\boldsymbol{\theta}^*$: $Y = \operatorname{sign}(\boldsymbol{\theta}^{*\top}\boldsymbol{X})$. The latter yields

$$\mathcal{L}(Y \mid X) = \delta(\boldsymbol{\theta}^{*\top}\boldsymbol{X}), \tag{1.3}$$

where $\delta_{\boldsymbol{x}}$ is the Dirac measure at point $\boldsymbol{x} \in \mathbb{R}^p$. As previously, we would like to perform maximum likelihood estimation. This model is not dominated; therefore, the maximum likelihood approach can not be applied. Logistic regression proposes a relaxation of (1.3) to overcome this issue:

$$P(Y = y \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \exp(-y\boldsymbol{\theta}^{*\top}\boldsymbol{x})}. \tag{1.4}$$

We observe that the conditional probability is greater than $1/2$ when the signs of y and $\boldsymbol{\theta}^{*\top}\boldsymbol{x}$ coincide and vice-versa. We also notice that, the larger the norm of $\boldsymbol{x}$, the more precise is the proposed model about the label. Therefore, the model indeed replicates the initial assumption. In addition, the relaxed probability is smooth and log-concave (see Section 1.5.1). Summing up, MLE takes the following form:

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{logit}} = \widehat{\boldsymbol{\theta}}_n^{\mathsf{MLE}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \log\big(1 + \exp\big(-Y_i\boldsymbol{\theta}^\top\boldsymbol{X}_i\big)\big).$$

Logistic regression is widely used in classification problems that appear in many fields (see [Wri95, HJLS13]).

Bayesian binary logistic regression

In the Bayesian settings, the logistic regression is of particular interest to us, as it includes all the general concepts that we discuss in this manuscript. Similar to the previous case, we have independent and identically distributed data points $\boldsymbol{Z}_k = (\boldsymbol{X}_k, Y_k)$, where $\boldsymbol{X}_k \in \mathbb{R}^p$ and labels in $Y_k \in \{-1, +1\}$. Also, the conditional distribution is given as in (1.4). The Bayesian approach to the logistic model assumes that the parameter space is the Euclidean space $\mathbb{R}^p$, and that it is equipped with a Gaussian prior measure μ:

$$\mu := \mathcal{N}\left(0, \frac{1}{\lambda}\mathbf{\Sigma}_{\boldsymbol{X}}^{-1}\right), \quad \text{where} \quad \mathbf{\Sigma}_{\boldsymbol{X}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top.$$

The hyperparameter λ is a positive number, that is usually specified by the practitioner. The Bayesian estimator is calculated for the quadratic loss. Using the formula (1.2), we obtain the following:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_n^B(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n) = C_n \int_{\mathbb{R}^p} \boldsymbol{\theta} \exp\Bigg(- \sum_{i=1}^{n} Y_i \boldsymbol{\theta}^\top \boldsymbol{X}_i - \sum_{i=1}^{n} \log\left\{1 + \exp\left(-\boldsymbol{\theta}^\top \boldsymbol{X}_i\right)\right\} \\ - \frac{\lambda}{2}\left\|\mathbf{\Sigma}_{\mathbf{X}}^{1/2}\boldsymbol{\theta}\right\|_2^2\Bigg) d\boldsymbol{\theta}.\end{aligned}$$

Here, C_n is a constant in terms of $\boldsymbol{\theta}$ and it depends on $\mathcal{Z}_n$. The first two terms in the exponential correspond to the log-likelihood of the logistic model, whereas the last term comes from the log-density of the prior and can be seen as a penalty term. An important remark about the function in the exponent, is that it is concave in terms of $\boldsymbol{\theta}$, and therefore, the integrand is log-concave (see Section 1.5.1).

We observe here, that the estimator is not computable for rather simple data sets. To construct an estimator for Y, the classical approaches resort to approximate methods. Solutions to this problem were proposed by [CD98] and [HH06], who use *sampling* based algorithms. See Section 1.4.4 for more details.

1.2 Statistiques paramétriques en grande dimension

Dans cette section, nous présentons brièvement le cadre général de l'inférence statistique paramétrique. Tout d'abord, nous Tout d'abord, nous introduisons la formulation mathématique du problème et les principales hypothèses requises pour l'analyse suivante. analyse ultérieure. Ensuite, nous discutons des estimateurs de maximum de vraisemblance et bayésiens. Dans la troisième sous-section, nous présentons le problème de régression linéaire avec ses versions régularisées. Nous concluons la section avec les modèles de régression logistique dans les deux cadres, fréquentiste et bayésien.

1.2.1 Notions générales

Le but général de la statistique mathématique est d'estimer une certaine caractéristique de la loi de probabilité inconnue inconnue P^n sur $\mathbb{R}^{dn}$, en utilisant un ensemble de données $\mathcal{Z}_n = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, qui est échantillonné à partir de cette distribution. Les points de données $\boldsymbol{Z}_k$ sont des vecteurs aléatoires à d dimensions. Dans le cas où ils sont indépendants les uns des autres, P^n peut être exprimé comme une mesure de produit:

$$P^n = \mathcal{L}(\boldsymbol{Z}_1) \otimes \ldots \otimes \mathcal{L}(\boldsymbol{Z}_n).$$

Une autre hypothèse importante concernant la distribution P^n est qu'elle appartient à une classe paramétrique:

$$\mathcal{P} := \{P^n_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\},$$

où le sous-ensemble $\Theta \subset \mathbb{R}^p$ est appelé l'ensemble des paramètres et il est connu à l'avance. Enfin, dans la plupart des cas, il est pratique pour nous d'avoir un modèle *identifiable*. C'est-à-dire que, pour différentes valeurs de $\boldsymbol{\theta}_1$ et de $\boldsymbol{\theta}_2$, nous avons besoin d'un modèle de type "textuel". et $\boldsymbol{\theta}_2$ du paramètre, les distributions correspondantes sont différentes les unes des autres:

$$\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \implies P^n_{\boldsymbol{\theta}_1} \neq P^n_{\boldsymbol{\theta}_2}.$$

Ainsi, la distribution inconnue inconnue P^n est décrite par un vecteur que nous désignons par $\boldsymbol{\theta}^*$ et nous l'appelons la valeur réelle du paramètre. Cela signifie que pour estimer la distribution inconnue P^n, on doit estimer le vrai paramètre $\boldsymbol{\theta}^*$. Par conséquent, le problème de l'estimation se résume à trouver une fonction $\widehat{\boldsymbol{\theta}}_n(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, qui est proche de $\boldsymbol{\theta}^*$ dans un certain sens probabiliste.

Maintenant que nous avons décrit le cadre mathématique général du problème, une question qui se pose naturellement est la suivante comment choisir "un bon" ou "le meilleur" estimateur $\widehat{\boldsymbol{\theta}}_n$? Pour faire un choix mathématiquement justifié, il faut avoir un critère ou une méthode quantitative de comparaison entre deux estimateurs. Pour ce faire, on utilise une fonction de perte l, qui est une fonction positive définie sur $\Theta \times \Theta$. Le risque de l'estimateur $\widehat{\boldsymbol{\theta}}_n$ au point $\boldsymbol{\theta}$ est défini comme suit :

$$\mathcal{R}\big(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}\big) := \mathbb{E}_{\boldsymbol{\theta}}\left[l\Big(\widehat{\boldsymbol{\theta}}_n(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n), \boldsymbol{\theta}\Big)\right],$$

où $\mathbb{E}_{\boldsymbol{\theta}}$ signifie que l'espérance est prise sur l'échantillon $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n) \sim P^n_{\boldsymbol{\theta}}$. L'un des choix les plus courants pour $l(\boldsymbol{\theta}, \boldsymbol{\theta}')$ est la fonction quadratique $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2$. Dans ce cas, le risque est appelé risque quadratique ou l'erreur quadratique moyenne (EQM). Un simple calcul montre qu'il n'existe pas d'estimateur qui soit universellement (c'est-à-dire pour toutes les valeurs des paramètres $\boldsymbol{\theta}$) meilleur en termes de risque. Néanmoins, nous pouvons espérer certaines qualités asymptotiques et/ou probabilistes. Nous disons que l'estimateur $\widehat{\boldsymbol{\theta}}_n$ est cohérent si $\mathcal{R}\big(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}\big) \to 0$, lorsque $n \to \infty$.

Cela signifie qu'avec la croissance de n, notre estimateur (qui dépend de n) se rapproche de $\boldsymbol{\theta}^*$ en termes de du risque. Ce dernier point est une caractéristique importante, car nous voulons améliorer l'estimation avec la croissance de la taille de l'échantillon n. taille de l'échantillon n. Un autre aspect important est le comportement asymptotique de l'estimateur. Nous appelons l'estimateur $\widehat{\boldsymbol{\theta}}_n$ *est asymptotiquement normal*, s'il satisfait à la condition suivante :

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_p(0, \Sigma(\boldsymbol{\theta})), \quad \text{pour tout} \quad \boldsymbol{\theta} \in \Theta.$$

Ici, Σ est la matrice de covariance asymptotique qui est une fonction du paramètre $\boldsymbol{\theta}$, qui dépend de l'estimateur. On peut construire différents critères quantitatifs, basés sur Σ, qui peuvent être utilisés pour comparer deux estimateurs $\widehat{\boldsymbol{\theta}}^1$ et $\widehat{\boldsymbol{\theta}}^2$ (voir par exemple [FK85, Rao]).

1.2.2 Estimateurs du maximum de vraisemblance et estimateurs bayésiens

Supposons maintenant que la distribution $P^n_{\boldsymbol{\theta}}$ est absolument continue par rapport à une certaine mesure additive de σ. définie sur $\mathbb{R}^{dn}$. Ainsi, $P^n_{\boldsymbol{\theta}}$ peut être caractérisé par sa densité $f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta})$. Nous appelons *fonction de vraisemblance* la carte $\boldsymbol{\theta} \to L_n(\boldsymbol{\theta})$, définie comme suit

$$L_n(\boldsymbol{\theta}) = f(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, \boldsymbol{\theta}), \quad \text{pour} \quad \forall \boldsymbol{\theta} \in \Theta,$$

où $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ est l'échantillon initialement observé. L'estimateur du maximum de vraisemblance est défini comme un maximisant de L_n :

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{ML}} \in \arg\max_{\boldsymbol{\theta}} L_n(\boldsymbol{\theta}). \tag{1.5}$$

Il faut garder à l'esprit que le SGD n'est pas toujours unique. Voir [LC90] pour quelques exemples simples, ainsi que des exemples avancés d'estimation par maximum de vraisemblance. L'estimateur $\widehat{\boldsymbol{\theta}}_n^{\mathsf{ML}}$ peut également être défini comme un minimiseur du logarithme négatif de la fonction de vraisemblance L_n. Cette dernière est définie de la manière suivante :

$$l_n(\boldsymbol{\theta}) := -\frac{1}{n} \log\left(L_n(\boldsymbol{\theta})\right), \tag{1.6}$$

et est appelée fonction de log-vraisemblance. Si les points de données $\boldsymbol{Z}_k$ sont i.i.d. alors la (1.5) prend la forme suivante :

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{ML}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \Big\{ -\frac{1}{n} \sum_{k=1}^{n} \log f(\boldsymbol{Z}_k, \boldsymbol{\theta}) \Big\},$$

où $f(\boldsymbol{z}, \boldsymbol{\theta})$ est la densité marginale de $\boldsymbol{Z}_1$. Si la log-vraisemblance est convexe en ce qui concerne $\boldsymbol{\theta}$, alors le problème de minimisation (1.6) peut être résolu à l'aide des méthodes d'optimisation convexe (voir Section 1.3). Sous des hypothèses légères, la MLE s'avère cohérente et asymptotiquement normale (cf. [LC70, Dan61]).

Remark 2. *L'hypothèse selon laquelle la distribution inconnue appartient à une classe paramétrique peut être relâchée dans le cas général. Dans la littérature récente de la statistique, la vraie distribution est supposée être proche d'une certaine famille paramétrique dans un certain sens probabiliste ([Tsy08]) au lieu d'être représentée exactement par une certaine valeur du paramètre.*

Dans les sections précédentes, nous n'avons exploité aucune structure topologique de l'espace des paramètres Θ. Jusqu'à présent, nous avons considéré le vrai paramètre comme un vecteur constant inconnu, qui appartenait à Θ. Dans l'approche bayésienne, l'espace est doté d'une mesure μ, définie sur une algèbre sigma Σ_Θ. Cette mesure est souvent appelée distribution distribution préalable dans la littérature des statistiques bayésiennes. Une fois encore, nous disposons d'un ensemble de données $\mathcal{Z}_n = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, échantillonné à partir d'une distribution $P^n_{\boldsymbol{\theta}^*}$, et nous voulons déduire le paramètre inconnu $\boldsymbol{\theta}^*$. paramètre inconnu $\boldsymbol{\theta}^*$. La mesure μ est choisie en fonction du problème concret et de l'expérience du praticien. problème concret et de l'expérience antérieure du praticien, et elle permet de quantifier universellement la qualité d'un estimateur $\widehat{\boldsymbol{\theta}}_n$. Pour ce faire, on utilise le *risque intégré*:

$$\mathcal{R}\big(\widehat{\boldsymbol{\theta}}_n\big) = \int_\Theta \mathcal{R}\big(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n\big)\mu(d\boldsymbol{\theta}) = \int_\Theta \int_{\mathbb{R}^{dn}} l\big(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)\big) P^n_{\boldsymbol{\theta}}(d\boldsymbol{z}_1, \ldots, d\boldsymbol{z}_n)\mu(d\boldsymbol{\theta}).$$

Le risque intégré joue le rôle de l'erreur moyenne de l'estimateur. Par conséquent, on s'intéresse à la minimisation de cette erreur. L'estimateur bayésien est défini comme le minimiseur du risque intégré :

$$\widehat{\boldsymbol{\theta}}^B_n \in \arg\min_{\widehat{\boldsymbol{\theta}}_n} \mathcal{R}(\widehat{\boldsymbol{\theta}}_n).$$

La minimisation est ici effectuée sur la classe de toutes les fonctions mesurables $\mathcal{M}(\mathbb{R}^{dn}, \Theta)$. L'estimateur bayésien dépend du choix de la fonction de perte l. estimateur bayésien dépend du choix de la fonction de perte l. Dans le cas où l est la perte quadratique, $\widehat{\boldsymbol{\theta}}^B_n$ est donné comme suit :

$$\widehat{\boldsymbol{\theta}}^B_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) := \frac{\int_\Theta \boldsymbol{\theta} f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta})\mu(d\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta})\mu(d\boldsymbol{\theta})}, \tag{1.7}$$

où $p^n_{\boldsymbol{\theta}}$ est la densité de $P^n_{\boldsymbol{\theta}}$. On peut remarquer que l'intégrale au dénominateur sert de constante de normalisation. Ainsi, $f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta}) / \int_\Theta f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{\theta})\mu(d\boldsymbol{\theta})$ peut être interprété comme une fonction de densité d'une distribution de probabilité définie sur (Θ, Σ_Θ). On l'appelle souvent *distribution postérieure*. Nous verrons plus tard que la constante de normalisation peut être négligée dans de nombreuses techniques d'intégration approximative. L'estimateur $\widehat{\boldsymbol{\theta}}_n^B$ est souvent appelé *moyenne postérieure*. Lorsque la dimension p de l'espace des paramètres est grande, le calcul des intégrales apparaissant dans le dernier affichage est généralement intraitable. Pour contourner ce problème, il faut recourir à des méthodes de calcul approximatif. L'un des principaux axes de recherche dans ce domaine est l'intégration de Monte-Carlo et les algorithmes MCMC. Un ingrédient central de ces algorithmes est une méthode d'échantillonnage approximatif à partir d'une distribution de probabilité donnée (voir Section 1.4.4 ci-dessous). Pour compléter cette sous-section, notons que le comportement de l'estimateur bayésien dépend du choix de la fonction de perte l. Nous renvoyons le lecteur à [Bro80], pour plus de détails sur ce sujet, et à [Rob96, LC06] pour une étude sur les pertes intrinsèques qui présentent certains avantages par rapport à la perte quadratique.

1.2.3 Régression linéaire multivariée

Dans cette section, nous étudions la régression linéaire. Avant d'énoncer le modèle lui-même, introduisons la notation. Nous supposons que les points de données $\boldsymbol{Z}_k = (\boldsymbol{X}_k, Y_k)$ sont des vecteurs composites. Ici, $\boldsymbol{X}_k \in \mathbb{R}^p$ sont appelés vecteurs de caractéristiques, tandis que $Y_k \in \mathbb{R}$ sont appelés étiquettes. Nous désignons par X la *matrice de conception* des données $\mathcal{Z}_n$, qui est la matrice $d \times n$, ayant pour lignes les caractéristiques $\boldsymbol{X}_k$. De même, $\mathrm{Y} := (Y_1, Y_2, \ldots, Y_n)^\top$ est le vecteur des étiquettes.

Nous disons que le modèle est linéaire si les étiquettes sont une transformation linéaire des caractéristiques jusqu'à une moyenne. linéaire des caractéristiques jusqu'à un terme de bruit additif de moyenne nulle. Mathématiquement, il a la forme suivante :

$$\mathrm{Y} = \mathrm{X}\,\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}.$$

Ici, $\boldsymbol{\varepsilon}$ est un vecteur aléatoire de moyenne nulle dans $\mathbb{R}^n$, qui est indépendant de la matrice de conception X. Supposons que $\boldsymbol{\varepsilon}$ soit un vecteur gaussien : $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Alors, la SGD de $\boldsymbol{\theta}^*$ dans ce modèle s'écrit comme suit

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{ML}} \in \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^p} L_n(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{k=1}^n \exp\left(-\frac{1}{2\sigma^2}\big(\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k\big)^2\right).$$

En raison de la monotonicité de la fonction exponentielle, on obtient

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{ML}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{\frac{1}{n} \sum_{k=1}^n \big(\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k\big)^2\right\}.$$

Ce problème est également appelé "estimation des moindres carrés". Il s'agit d'un problème de minimisation quadratique, qui a toujours une solution. De plus, si $\mathrm{rank}(\mathrm{X}) = p$, alors une solution à forme fermée est disponible et elle est appelée estimateur des moindres carrés ordinaires (1.8) estimateur :

$$\widehat{\boldsymbol{\theta}}^{\mathsf{OLS}} := \left(\mathrm{X}^\top \mathrm{X}\right)^{-1} \mathrm{X}^\top \mathrm{Y}. \tag{1.8}$$

Nous avons donc une formule explicite pour la solution solution et elle est calculable efficacement avec les langages de programmation scientifique modernes. L'estimateur des moindres carrés peut également être considéré comme une tentative de prédire l'étiquette Y à l'aide d'une fonction linéaire du vecteur caractéristique $\boldsymbol{Y}$. Cette interprétation des MCO est apparente dans le problème de l'apprentissage supervisé. Nous renvoyons le lecteur intéressé à [Kol09] pour une discussion détaillée sur ce sujet, et pour la concentration propriétés de concentration de $\widehat{\boldsymbol{\theta}}_n^{\mathsf{OLS}}$.

Regularisation

Dans plusieurs situations fréquemment rencontrées, le OLS n'a pas une performance satisfaisante du du point de vue du praticien. Il souffre de plusieurs problèmes majeurs.

- Bien que sans biais, les OLS ont une grande variance. Désignons par Δ la distance Euclidienne entre la $\widehat{\boldsymbol{\theta}}_n^{\mathsf{ML}}$ et la valeur du paramètre $\boldsymbol{\theta}$. Ensuite, l'espérance et la variance de Δ^2 ont la formule suivante (voir [HK70]) :

$$\mathbb{E}_{\boldsymbol{\theta}^*}[\Delta^2] = \sum_{k=1}^{p} \frac{2\sigma^2}{\lambda_k} \quad \text{et} \quad \mathrm{var}_{\boldsymbol{\theta}^*}(\Delta^2) = \sum_{k=1}^{p} \frac{2\sigma^4}{\lambda_k^2}.$$

  Ici, $\lambda_1, \ldots, \lambda_p$ sont les valeurs propres de $\mathrm{X}^\top\mathrm{X}$. On voit que le côté droit de l'équation explose lorsque les valeurs propres sont petites. C'est souvent le cas dans les problèmes à haute dimension.

- Un autre inconvénient de la méthode est le manque d'interprétabilité. Les OLS ne permettent pas permet pas de différencier les variables pertinentes des variables non pertinentes. On serait intéressés par les estimateurs épars. L'estimateur est appelé *sparse*, s'il a un nombre relativement faible d'entrées non nulles. d'entrées non nulles. Ainsi, la rareté conduirait à un grand nombre de coordonnées de disparition pour l'estimateur de $\boldsymbol{\theta}^*$. Ce dernier, à son tour, produirait que les entrées correspondantes du vecteur de caractéristiques ne sont pas utiles pour prédire l'étiquette.

Une approche courante pour résoudre ce problème consiste à introduire un terme de régularisation $r(\cdot)$. La carte $r : \mathbb{R}^+ \Rrightarrow \mathbb{R}^+$. est une fonction monotone croissante. La version régularisée de l'estimation des moindres carrés est le problème d'optimisation suivant :

$$\widehat{\boldsymbol{\theta}}_n \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{n} \sum_{k=1}^{n} (\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k)^2 + \lambda\, r(\|\boldsymbol{\theta}\|) \right\},$$

où λ est un paramètre positif et $\|\cdot\|$ est une norme arbitraire sur $\mathbb{R}^p$. En ajoutant le terme $\lambda r(\|\boldsymbol{\theta}\|)$ on pénalise ainsi la norme du vecteur de prédiction. Cette astuce simple permet d'avoir normes plus petites, ce qui permet d'obtenir des modèles estimés moins complexes. Le choix de la norme est basé sur la notion de complexité qui nous intéresse. de complexité qui nous intéresse. La régularisation est choisie pour refléter nos connaissances préalables sur les paramètres souhaitables du modèle. souhaitables du modèle. Les méthodes de moindres carrés régularisées les plus populaires, Lasso et Ridge, sont brièvement présentées ci-dessous.

LASSO

Dans le cas où $r(a) = a$, pour tous les $a > 0$, et où la norme l_1 est prise, la solution du problème de minimisation suivant est appelée un estimateur LASSO :

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{LASSO}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{n} \sum_{k=1}^{n} (\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k)^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\}. \tag{1.9}$$

Il s'agit d'un problème d'optimisation convexe qui a toujours une solution. LASSO a été proposé par [Tib96]. Le comportement non lisse de la norme l_1 favorise *sparse* minimisateurs de (1.9). Il existe une relation étroite entre le Lasso et le seuillage doux, nous renvoyons le lecteur à [DJHS92]. Cette idée est également exploitée dans une autre méthode de récupération éparse appelée sélecteur de Dantzig (voir [CT07]). Une comparaison détaillée comparaison détaillée et une analyse unifiée des deux méthodes se trouvent dans [BRT+09]. La complexité de LASSO a été largement étudiée, voir [BTW07, vdG07, BRT+09, AH12, vdGL13]. Dans [DHL17], les auteurs montrent que $\lambda = \sqrt{2\log(p)/n}$ conduit à un excès de risque optimal d'ordre $O(\mathrm{rank}(\mathrm{X})\log(p)/n)$ pour (1.9). Ici, X est la matrice de conception des vecteurs de caractéristiques. Pour en savoir plus sur les méthodes méthodes susmentionnées, consultez l'article de synthèse de [Kol09].

Ridge

La régression ridge est la régression linéaire avec le terme de régularisation de type l_2 :

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{Ridge}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{n} \sum_{k=1}^{n} (\boldsymbol{\theta}^\top \boldsymbol{X}_k - Y_k)^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \right\}. \tag{1.10}$$

La solution de (1.10) a une formule explicite, similaire à (1.8). Dans le cas de la régression ridge, cependant, la matrice de conception X peut avoir un rang quelconque :

$$\widehat{\boldsymbol{\theta}}_n^{\text{Ridge}} = \left(\mathrm{X}^\top \mathrm{X} + \lambda I_n\right)^{-1} \mathrm{X}^\top \mathrm{Y}.$$

Puisque $\lambda > 0$ l'inverse de la matrice est toujours bien défini. Les premiers articles qui présentent cette méthode sont : [Cro72, Hor62]. Comme Comme mentionné précédemment, la régularisation induit un certain biais pour réduire la variance du problème d'apprentissage (compromis biais-variance). d'apprentissage (compromis biais-variance).Dans ce phénom Dans ce phénomène, le paramètre d'ajustement λ joue un rôle majeur. Pour des résultats sur l'estimation de λ, le lecteur intéressé est renvoyé à [HK70, Nor82, Wen00, Kib03, ST+99a].

1.2.4 Régression logistique

Jusqu'à présent, nous n'avons abordé que le cas où les étiquettes admettent des valeurs réelles. Pourtant, il existe une vaste collection de problèmes lorsque les étiquettes sont catégoriques. Ces problèmes sont appelés problèmes de classification. Dans cette Dans cette section, nous n'aborderons que le cas de deux classes, à savoir -1 et 1. Par conséquent, nos étiquettes sont des variables aléatoires discrètes variables aléatoires discrètes : $Y_i \in \{-1, 1\}$. Le modèle logistique a pour objectif d'estimer la distribution conditionnelle de Y en connaissant le vecteur de caractéristiques $\boldsymbol{X}$. L'hypothèse principale est que les caractéristiques peuvent être séparées par un hyperplan en $\mathbb{R}^p$, qui est décrit par sa normale $\boldsymbol{\theta}^*$: $Y = \operatorname{sign}(\boldsymbol{\theta}^{*\top} \boldsymbol{X})$. Cette dernière donne

$$\mathcal{L}(Y \mid X) = \delta(\boldsymbol{\theta}^{*\top} \boldsymbol{X}), \tag{1.11}$$

où $\delta_{\boldsymbol{x}}$ est la mesure de Dirac au point $\boldsymbol{x} \in \mathbb{R}^p$. Comme précédemment, nous souhaitons effectuer une estimation par maximum de vraisemblance. Ce modèle n'est pas dominé ; par conséquent, l'approche du maximum de vraisemblance ne peut pas être appliquée. Cependant, les méthodes classiques d'optimisation ne peuvent pas traiter les densités de Dirac. La régression logistique propose une relaxation de (1.11) pour surmonter ce problème:

$$P(Y = y \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \exp(-y \boldsymbol{\theta}^{*\top} \boldsymbol{x})}. \tag{1.12}$$

Nous observons que la probabilité conditionnelle est supérieure à $1/2$ lorsque les signes de y et de $\boldsymbol{\theta}^{*\top} \boldsymbol{x}$ coïncident et vice-versa. Nous remarquons également que, plus la norme de $\boldsymbol{x}$, plus le modèle proposé est précis quant à l'étiquette. Par conséquent, le modèle reproduit bien l'hypothèse initiale. initiale. De plus, la probabilité relaxée est lisse et log-concave (voir Section 1.5.1). En résumé, MLE prend la forme suivante :

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{logit}} = \widehat{\boldsymbol{\theta}}_n^{\mathsf{MLE}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \log\big(1+\exp\big(-Y_i\boldsymbol{\theta}^\top \boldsymbol{X}_i\big)\big).$$

La régression logistique est largement utilisée dans les problèmes de classification qui apparaissent dans de nombreux domaines (voir [Wri95, HJLS13]).

Régression logistique binaire bayésienne

Dans les paramètres bayésiens, la régression logistique présente un intérêt particulier pour nous, car elle inclut tous les concepts généraux dont nous discutons dans ce manuscrit. concepts généraux que nous abordons dans ce manuscrit. Comme dans le cas précédent, nous avons des points de données indépendants et identiquement distribués $\boldsymbol{Z}_k = (\boldsymbol{X}_k, Y_k)$, où $\boldsymbol{X}_k \in \mathbb{R}^p$ et des étiquettes dans $Y_k \in \{-1,+1\}$. De plus, la distribution conditionnelle est donnée comme dans (1.12). L'approche bayésienne du modèle logistique suppose que l'espace des paramètres est l'espace euclidien $\mathbb{R}^p$, et qu'il est doté d'une mesure préalable gaussienne mesure gaussienne μ :

$$\mu := \mathcal{N}\left(0, \frac{1}{\lambda}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\right), \quad \text{where} \quad \boldsymbol{\Sigma}_{\boldsymbol{X}} := \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top.$$

L'hyperparamètre λ est un nombre positif, qui est généralement spécifié par le praticien. L'estimateur bayésien est calculé pour la perte quadratique. En utilisant la formule (1.7), on obtient ce qui suit:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_n^B(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_n) = C_n \int_{\mathbb{R}^p} \boldsymbol{\theta}\exp\Bigg(-\sum_{i=1}^n Y_i\boldsymbol{\theta}^\top\boldsymbol{X}_i - \sum_{i=1}^n \log\left\{1+\exp\left(-\boldsymbol{\theta}^\top\boldsymbol{X}_i\right)\right\} \\ - \frac{\lambda}{2}\left\|\boldsymbol{\Sigma}_{\mathbf{X}}^{1/2}\boldsymbol{\theta}\right\|_2^2\Bigg)d\boldsymbol{\theta}.\end{aligned}$$

Ici, C_n est une constante en termes de $\boldsymbol{\theta}$ et elle dépend de $\mathcal{Z}_n$. Les deux premiers termes de l'exponentielle correspondent à la log-vraisemblance du modèle logistique, tandis que le dernier terme provient de la log-densité de l'antériorité et peut être considéré comme un terme de pénalité.

Nous observons ici que l'estimateur n'est pas calculable pour des ensembles de données plutôt simples. Pour construire un estimateur pour Y, les approches classiques ont recours à des à des méthodes approximatives. Des solutions à ce problème ont été proposées par [CD98] et [HH06], qui utilisent des algorithmes basés sur *sampling*. Voir Section 1.4.4 pour plus de détails.

1.3 Convex Optimization

As we have seen in the previous section, convex optimization is of vital importance for the solution of inference problems. In this section, we describe the general mathematical settings of optimization.

A set $\mathcal{C} \subset \mathbb{R}^p$ is called *convex* , if it contains all the segments between its two elements:

$$\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in \mathcal{C}, \quad \text{for every} \quad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C} \quad \text{and} \quad \alpha \in (0,1).$$

Examples of convex sets are halfspaces, balls and hypercubes. A function $F : \mathcal{C} \to \mathbb{R}$, where $\mathcal{C}$ is a convex set, is called *convex*, if the following condition is satisfied:

$$F(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha F(\boldsymbol{x}) + (1-\alpha)F(\boldsymbol{y}), \quad \text{for every} \quad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C} \quad \text{and} \quad \alpha \in (0,1).$$

The concept of the convexity is closely related to the Hessian matrices. Suppose $F \in C^2(\mathbb{R}^p)$. Then F is convex, if and only if its Hessian $\nabla^2 F(\boldsymbol{x})$ is positive semidefinite for all $\boldsymbol{x} \in \mathcal{C}$. For more details on equivalent formulations of convexity, see [Nes04]. The general convex optimization problem is formulated as follows:

$$\begin{aligned} \text{minimize} \quad & F(\boldsymbol{x}), \\ \text{subject to} \quad & \boldsymbol{x} \in \mathcal{C}, \end{aligned} \tag{CO}$$

where $\mathcal{C}$ is a convex set and $F : \mathcal{C} \to \mathbb{R}$ is a convex function. This general form of the problem is very useful. Indeed, in the inference problems (linear regression, maximum likelihood, logistic regression) mentioned above it comes to the minimization of some convex function on the Euclidean space $\mathbb{R}^p$. The advantage of convexity is the control over the local behavior. In particular, an important property of the convex functions is that all the stationary points are global minimizers. Thus, solving (CO) amounts to finding the zeros of the gradient. This observation forms the intuition of the gradient descent, one of the principal algorithms of convex optimization.

1.3.1 Gradient Descent

Suppose we have a convex function $F : \mathbb{R}^p \to \mathbb{R}$, which is continuously differentiable at any point $\boldsymbol{x} \in \mathbb{R}^p$. Then, the following iterative algorithm is called the *gradient descent*:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \nabla F(\boldsymbol{x}_k), \tag{GD}$$

where h_k is a parameter of the method referred to as the *step-size* or the *learning rate*. The choice of the *starting point* $\boldsymbol{x}_0$, as well as the step-size are up to the user. The method was first proposed by [C+47]. In this section, we are going to discuss only the case of functions with Lipschitz continuous gradients, as it is the one that is studied later in the manuscript. For non-smooth optimization, we refer the reader to [Bub15] or [Nes04]. The function F is called M-smooth or M-gradient Lipschitz if

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\|_2 \leq M\|\boldsymbol{x} - \boldsymbol{y}\|_2, \quad \text{for every} \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p.$$

Thus, the gradients of the function have at most linear growth, which induces at most quadratic growth on the function itself. This idea is formulated in the following proposition.

Proposition 1. *Suppose $F \in C^2(\mathbb{R}^p)$. Then F is M-smooth if and only if the following inequality is satisfied:*

$$\nabla^2 F(\boldsymbol{x}) \preceq M\, I_p, \quad \textit{for every} \quad \boldsymbol{x} \in \mathcal{C},$$

where $A \preceq B$ means that $B - A$ is a positive semidefinite matrix.

The proof of the proposition can be found in [Nes04, Theorem 2.1.6]. Based on this result, the theorem below upper bounds the error of convergence of the sequence $F(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ to the minimum $F(\boldsymbol{x}_*)$.

Theorem 1. *Let F be convex and M-smooth on $\mathbb{R}^p$. Then a gradient descent with a constant step-size $h_k = h < \frac{1}{M}$ satisfies*

$$F\left(\boldsymbol{x}_K\right) - F\left(\boldsymbol{x}_*\right) \leq \frac{2\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|_2^2}{hK}.$$

We refer the reader to [Bub15, Theorem 3.3] for the proof. Let us now comment the results of the theorem:

- We have a polynomial convergence of order $O(1/K)$ for the *value error*.
- The theorem does not provide guarantees on the distance between $\boldsymbol{x}_k$ and the minimum point $\boldsymbol{x}_*$.
- The result is dimension-free: the only term in the upper-bound that may depend on the dimension d is the square distance between the initial point $\boldsymbol{x}_0$ and the optimum $\boldsymbol{x}_*$.
- The term $\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2$ is the initial distance from the minimum point. We see that the error grows when $\boldsymbol{x}_0$ is placed further from the optimal point $\boldsymbol{x}_*$. The result also implies that the initial conditions are forgotten in polynomial time.

1.3.2 Strongly convex functions

In the theory of convex optimization, the strong convexity plays a significant role. Let m be a positive number. The continuously differentiable function $F : \mathcal{C} \to \mathbb{R}$, where $\mathcal{C} \subset \mathbb{R}^p$ is a convex set, is called *m-strongly convex* if

$$F(\boldsymbol{x}) \geq F(\boldsymbol{y}) + \nabla F(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \tag{1.13}$$

for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$. The following proposition (see [Nes04, Theorem 2.1.11]) establishes the link between strong convexity and Hessian matrices for C^2 functions.

Proposition 2. *Suppose $F : \mathcal{C} \to \mathbb{R}$ is a C^2 function, where the domain $\mathcal{C}$ is an open convex set. Then f is m-strongly convex, if and only if the following condition is satisfied:*

$$\nabla^2 F(\boldsymbol{x}) \succeq m\, I_p, \quad \textit{for every} \quad \boldsymbol{x} \in \mathcal{C}.$$

The strong-convexity essentially means that the function is lower-bounded by a quadratic function. In the case, when the constant m in (1.13) is equal to zero, we recover the convexity condition. The following theorem (see [Nes04, Theorem 2.1.15]) provides us with an upper bound on the convergence in the strongly convex case.

Theorem 2. *Suppose F is m-strongly convex M-smooth. We define by $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, the sequence generated using the algorithm (GD) with a constant step-size h. Then, if $0 < h \leq \frac{2}{m+M}$, we have the following convergence bounds:*

$$\begin{aligned} \|\boldsymbol{x}_k - \boldsymbol{x}_*\|_2^2 &\leq (1 - mh)^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2\,; \\ F(\boldsymbol{x}_k) - F(\boldsymbol{x}_*) &\leq \frac{M}{2}(1 - mh)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2\,. \end{aligned}$$

Let us comment the results of the theorem.

- We obtain exponential convergence bounds for the distance of each iteration $\boldsymbol{x}_k$ from the optimal point $\boldsymbol{x}_*$.
- The second inequality shows that the strong convexity improves the polynomial convergence of the convex case to an exponential convergence of order $\exp(-mh)$.
- We observe also that the contraction coefficient becomes zero, when $m = 0$, which means that the current result is not applicable in the general convex case.

Readers interested in a more detailed account on convex analysis are referred to [Roc70] and [Nes04]; see also [Bub15] for a review of the topics of convex optimization which are relevant to machine learning.

1.3.3 Stochastic Gradient Descent

As we have seen before in the context of statistics (Section 1.1.1), the function to minimize is often sum-decomposable:

$$F(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\theta}).$$

The component functions g_i are usually assumed to be "similar". In this context, the similarity means to have the same strong convexity and smoothness constants m and M. Hence, the gradient descent for this function is the following iterative algorithm:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \nabla F(\boldsymbol{x}_k) = \boldsymbol{x}_k - \frac{h_k}{n}\sum_{i=1}^{n} \nabla g_i(\boldsymbol{x}_k).$$

However, the implementation of these steps is computationally expensive when we have a large dataset. That is when n and p are large. Therefore, a modification of each iteration that would avoid to do n gradient computations could be appealing. Stochastic Gradient Descent (introduced by [RM51]) proposes to uniformly sample an index $\boldsymbol{i}_k$ from $\{1, 2, \ldots, n\}$ and approximate the empirical mean of the gradients by replacing with the term corresponding to $\boldsymbol{i}_k$:

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - h_k \nabla g_{\boldsymbol{i}_k}(\boldsymbol{X}_k) = \boldsymbol{X}_k - h_k\big(\nabla F(\boldsymbol{X}_k) + \boldsymbol{\xi}_k\big), \tag{SGD}$$

where $\boldsymbol{\xi}_k$, in this case, is a mean zero noise vector. In particular, for the sum-decomposable setting $\boldsymbol{\xi}_k = \nabla g_{\boldsymbol{i}_k}(\boldsymbol{X}_k) - \nabla F(\boldsymbol{X}_k)$. The convergence of this stochastic algorithm has been extensively studied. In particular, if $\sum_k h_k^2 < \infty$ and $\sum_k h_k = +\infty$, then under mild conditions, SGD converges almost surely to the minimum point ([Bot10]). For a more thorough review on SGD, we refer the reader to [BB11, BM13, DFB17, GP20].

1.3.4 Gradient Flows

In the previous section, we have presented the basic concepts in convex optimization. We have discussed iterative algorithms that provide us with convergence of up to an exponential rate. It turns out that these methods have their continuous alternatives.
Suppose we have a smooth function F. The *gradient flow* is defined as a curve $\boldsymbol{x}(t)$ in $\mathbb{R}^p$ that starts at a point $\boldsymbol{x}(0) = \boldsymbol{x}_0$. At each time moment t is pointed to the direction of the fastest minimization of F, that is the negative gradient. It is given as the solution of the

following Cauchy problem:

$$\begin{cases} \dot{\boldsymbol{x}}(t) = -\nabla F(\boldsymbol{x}(t)), & \text{for} \quad t > 0, \\ \boldsymbol{x}(t) = \boldsymbol{x}_0. \end{cases} \tag{GF}$$

Here the dot is the derivate with respect to time. When the gradient ∇F is Lipschitz continuous, Cauchy's theorem guarantees the existence and the uniqueness of the solution. The next proposition shows that for a convex function F, the solution is also unique, and moreover, there is a contraction effect. The proof can be found in [San17].

Proposition 3. *Suppose that F is convex and let $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ be two solutions of* (GF) *with possibly different starting points. Then the application $t \mapsto \|\boldsymbol{x}^1(t) - \boldsymbol{x}^2(t)\|_2$ is non-increasing.*

The proof is a simple consequence of a fact, that convex functions have monotonically non-decreasing gradients. Another important property of the gradient flow is that $F(\boldsymbol{x}(t))$ is monotonically decreasing and convergent when $t \to +\infty$. Indeed,

$$\dot{F}(\boldsymbol{x}(t)) = \nabla F(\boldsymbol{x}(t))^\top \dot{\boldsymbol{x}}(t) = -\|\nabla F(\boldsymbol{x}(t))\|_2^2 \leq 0.$$

Thus, for lower-bounded functions F, the monotone convergence theorem implies the convergence of $F(\boldsymbol{x}(t))$. The next proposition is also based on properties of convexity and it claims the convergence of $F(\boldsymbol{x}(t))$ to the minimum of F.

Proposition 4. *Suppose F is a convex function. Then the curve $\boldsymbol{x}(t)$ satisfying* (GF) *minimizes F, when $t \to +\infty$. Moreover, polynomial convergence is available:*

$$F(\boldsymbol{x}(t)) - F(\boldsymbol{x}_*) \leq \frac{1}{t}\|\boldsymbol{x}(0) - \boldsymbol{x}_*\|_2^2.$$

Thus, we have a convergence of order $O(1/t)$ for the value error of convex functions. However, similar to (GD), the convergence of $\boldsymbol{x}(t)$ to $\boldsymbol{x}_*$ is not guaranteed. In the general convex case, some additional assumptions are required (see [Loj82]). The next proposition provides contraction of exponential order for the flow in the strongly convex case.

Proposition 5. *Suppose that F is m-strongly convex and let $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ be two solutions of* (GF) *with possibly different starting points. Then the following inequality is satisfied for all positive t:*

$$\|\boldsymbol{x}^1(t) - \boldsymbol{x}^2(t)\|_2 \leq \exp(-mt)\,\|\boldsymbol{x}^1(0) - \boldsymbol{x}^2(0)\|_2.$$

An important consequence of this proposition is the convergence to the minimum point $\boldsymbol{x}_*$. Indeed, one can easily verify that the constant $\boldsymbol{x}(t) \equiv \boldsymbol{x}_*$ is a solution to (GF). For more thorough review on the gradient flows, we refer the reader to [AGS08, San17].

1.3.5 Relation with the Gradient Descent

Suppose that $\boldsymbol{x}(t)$ is a solution of (GF). Then, it can also be described as the solution of the following integral equation:

$$\boldsymbol{x}(t) = \boldsymbol{x}_0 - \int_0^t \nabla F(\boldsymbol{x}(s))ds. \tag{1.14}$$

Under mild assumptions, this function converges to $\boldsymbol{x}_*$. One could hope for that a "decent" discrete approximation could also converge to the minimum point, thus providing us with a feasible computational method of minimization. It turns out, that the simplest approach works in this case. Let us now discretize the time axis by dividing it into segments of length $h > 0$. That is we divide it into intervals of form $[kh, (k+1)h]$, where $k \in \mathbb{N}$. First, in view of (1.14) we deduce

$$\boldsymbol{x}_{(k+1)h} = \boldsymbol{x}_{kh} - \int_{kh}^{(k+1)h} \nabla F(\boldsymbol{x}(s))ds. \tag{1.15}$$

In view of (1.15), we construct a sequence $\boldsymbol{y}_k$, which is designed to estimate $\boldsymbol{x}(kh)$, by replacing each component on both sides of the equation by its approximation:

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \int_{kh}^{(k+1)h} \nabla F(\boldsymbol{y}_k)ds = \boldsymbol{y}_k - h\nabla F(\boldsymbol{y}_k).$$

We approximated the integral on the interval $[kh, (k+1)h]$, by $h\nabla f(\boldsymbol{x}(kh)$. The latter, in its turn, is approximated by $h\boldsymbol{y}_k$. Thus, we recognize the already seen formula of (GD), which indeed converges exponentially in the strongly convex case. To sum up we see that the gradient descent is a piecewise-linear interpolation of the gradient flow. We will observe a similar phenomenon, when discretizing random processes in the problem of sampling (see Section 1.5).

1.4 Sampling from a probability distribution

In this section, we introduce some classical methods of sampling. We start with the algorithms of inverse transform and rejection-acceptance. Then we present the problem of approximate integration and the importance sampling method. We conclude the section with Metropolis-Hastings algorithm.

1.4.1 Random variable generation

Generation of random variables has been one of the pivotal questions in computer science and applied mathematics since the early 20th century. The goal is to describe a method or an algorithm that constructs random variables which are distributed according to a certain probability measure π. In certain cases the latter can be unknown or known only partially.

One of the first problems that was investigated in this area was the generation of i.i.d. samples from $\mathcal{U}(0,1)$, the uniform distribution on $(0,1)$, using deterministic sequences. However, Gödel's theorem claims that this is unfortunately impossible. The connection between the two is established using the algorithmic information theory (see e.g. [Cha82, Cha77, Cha88]). In [VN51], Von Neumann writes the following: *"Any one who considers arithmetical methods for reproducing random digits is, of course, in a state of sin. As has been pointed out several times, there is no such thing as a random number. There are only methods of producing random numbers, and a strict arithmetic procedure of course is not such a method."* Nevertheless, many deterministic algorithms can mimic the statistical behavior of independently drawn samples of $\mathcal{U}(0,1)$.

A *pseudo-random number generator* is often an iterative algorithm, that starts at an initial point v_0. At every step, a transformation T is applied to the current value: $v_{k+1} := \mathsf{T}(v_k)$. For all K, the values $(v_1, v_2, \dots, v_K)$ have similar behavior with a uniform sample $(U_1, U_2, \dots, U_K) \sim \mathcal{U}([0,1]^K)$, when compared using certain set of tests. For example, one may perform statistical tests such as the Kolmogorov-Smirnov test (see [Kol33, MJ51]), to compare the empirical CDF of the drawn variables with the actual CDF of $\mathcal{U}(0,1)$. The practice shows, that most generative methods approximate well the cumulative distribution function. The independence of the iterates is harder to achieve, in view of the sequential nature of the algorithm. Several methods use time-series based analysis, such as ARMA (see [HA01]), others use non-parametric tests (see [LD75]). Nowadays, all major scientific softwares provide efficient tools for pseudo-random number generation. *In the rest of the manuscript, we will thus assume that we have access to random samples independently drawn from $\mathcal{U}(0,1)$.*

1.4.2 Inverse sampling

Inverse transformation method is one of the easiest sampling methods. Suppose that π is a probability measure on $\mathbb{R}$ and let $F_\pi : \mathbb{R} \to \mathbb{R}$ be its cumulative distribution function. The *generalized inverse* of F_π is defined as below:

$$F_\pi^{-1}(y) := \inf\{x : F_\pi(x) \geq y\}.$$

The following simple proposition is the key result on which the method is based.

Proposition 6. *If $U \sim \mathcal{U}(0,1)$ then $F_\pi^{-1}(U) \sim \pi$.*

Therefore, using the Proposition 6, we immediately obtain an exact sampling method: we sample from a uniformly distributed U and then calculate $F^{-1}(U)$. This method can be used to generate exponential random variables. However, in most cases, the generalized inverse is computationally infeasible, which renders the algorithm inapplicable. To read more on inverse sampling, see e.g. [RC13]

1.4.3 Rejection - Acceptance

As mentioned in the previous section, the inverse transform method does not provide feasible solutions for the laws in high dimension. Moreover, the exact analytic form of the CDF is often not available (e.g. Bayesian posteriors). Instead, we are provided with the density function[1], which is known up to a constant. The key idea is to use an auxiliary distribution $\widehat{\pi}$ that is easier to sample from. In this section, we introduce the rejection-acceptance method, which is based on the Fundamental Theorem of Simulation.

Theorem 3 (Fundamental Theorem of Simulation)**.** *Simulating a random variable $\boldsymbol{X} \sim \pi$ is equivalent to sampling a vector $(\boldsymbol{X}, U) \sim \mathcal{U}\{(\boldsymbol{x}, u) : u < \pi(\boldsymbol{x})\}$.*

We deduce from the theorem, that the auxiliary dimension reduces sampling from any distribution to a uniform sampling problem in the extended space. However, uniform sampling from the set given by the theorem can itself be very challenging. One would approach this problem by generating the variables one by one using conditional law. However, taken into account the initial settings, this approach is not helpful. Indeed, sampling from $\mathcal{L}(\boldsymbol{X} \mid U)$ can be as complex as π. Thus, the variables need to be sampled jointly. Let us state a simple proposition to this matter.

Proposition 7. *Suppose we have two sets A and B, such that $A \subset B \subset \mathbb{R}^p$. Suppose also that we can generate uniform samples $\boldsymbol{X}$ from B. Then Algorithm 1 results uniform samples from A.*

This algorithm provides us with a general scheme: if we can sample from a density $\widehat{\pi}$ such that $\pi(\boldsymbol{x}) < \widehat{\pi}(\boldsymbol{x})$, then we can sample uniformly from the corresponding set $\{(\boldsymbol{x}, u), u < \widehat{\pi}(\boldsymbol{x})\}$. The latter contains the set $\{(\boldsymbol{x}, u), u < \pi(\boldsymbol{x})\}$, and therefore, Algorithm 1 can be applied. Based on this idea [VN51] proposed the following algorithm called the rejection-acceptance algorithm.

[1]Throughout the manuscript, we will often use the same notation for the probabilistic laws and their densities.

Algorithm 1 Sampling from A using samples from B

> **repeat**
>   Sample $\boldsymbol{X} \sim \mathcal{U}(B)$
> **until** $\boldsymbol{X} \notin A$ (rejection)
> $\boldsymbol{Y} \leftarrow \boldsymbol{X}$ (acceptance)

Proposition 8. *Suppose that $\pi(\boldsymbol{x}) < C\widehat{\pi}(\boldsymbol{x})$ for some distribution π that we are able to sample from. Then the Algorithm 2 returns samples from the distribution π.*

Algorithm 2 Rejection-Acceptance sampling

> **repeat**
>   Sample $\boldsymbol{X} \sim \widehat{\pi}$ and $U \sim \mathcal{U}(0,1)$
> **until** $U > \pi(\boldsymbol{X})/C\widehat{\pi}(\boldsymbol{X})$ (rejection)
> $\boldsymbol{Y} \leftarrow \boldsymbol{X}$ (acceptance)

The rejection-acceptance algorithm can be applied for sampling various distributions, such as $\mathcal{N}(0,1)$, $\mathcal{B}(\alpha,\beta)$, $\Gamma(\alpha,\beta)$. However, it lacks of efficiency in the high dimensional setting. The reason is that the number of iterations required to reach the acceptance step is a geometric distribution with a parameter that is proportional to $1/C$. The latter, in some cases, is exponentially small in terms of dimension. There are various modifications of Algorithm 2, such as Envelope Rejection-Acceptance, ARS (see e.g. [Dev06, Dev86, GW92, GS91]), which overcome this issue. These methods, however, fall out of the scope of the this manuscript.

1.4.4 Monte-Carlo integration

As seen in the previous chapters, the problems of statistical inference mainly arrive to two destinations: optimization or integration. In this section, we present Monte-Carlo integration method, which is based on sampling. Essentially, it consists of estimating the integral w.r.t. a probability measure with an empirical mean. The Monte-Carlo integration method was proposed by [MU49]. It found its early applications in physics in the 1950s [Ula52, VNR46]. See [Ben16, Met87] on the history of the method.

The problem can generally be formulated as the computation of the following integral:

$$\mathbb{E}_\pi[g(\boldsymbol{X})] = \int_{\mathbb{R}^p} g(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x}.$$

Suppose π is a probability distribution on $\mathbb{R}^p$ and let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independent samples drawn from π. Then, the law of large numbers hints at estimating the above integral with the empirical mean of the values:

$$\widehat{I}_n := \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i), \quad \text{where} \quad \boldsymbol{X}_i \sim \pi, \quad \text{for all} \quad i.$$

This algorithm is called *Monte-Carlo integration method*. The estimator $\widehat{I}_n$ is consistent. Moreover, using the central limit theorem we deduce that it is asymptotically normal and we can also show that the convergence is of order $O(1/\sqrt{n})$. The Monte-Carlo integrator proves to be efficient in most cases, however it can be improved (see [Rip09, BMP+94]). A modification of the Monte-Carlo algorithm, called *importance sampling* ([Mar54]), proposes to sample from a law ν, called *instrumental* distribution, instead of sampling π. The importance sampling estimator is thus defined as follows:

$$\widehat{I}_n^{\mathsf{imp}} := \frac{1}{n}\sum_{i=1}^{n} \frac{g(\boldsymbol{X}_i)}{\nu(\boldsymbol{X}_i)}\pi(\boldsymbol{X}_i), \quad \text{where} \quad \boldsymbol{X}_i \sim \nu, \quad \text{for all} \quad i.$$

The advantage of importance sampling over the Monte-Carlo integration, is the ability to choose the instrumental distribution. It allows to avoid the possible complex problem of sampling from π, by replacing the latter with a simpler distribution. The algorithm has found its early applications in sampling from normal distributions [Tuk56] and in statistical physics in 1950s (see [HM54, RR55]). It was shown later (see [RK16]), that the best choice in terms of the variance of $\widehat{I}_n^{\mathsf{imp}}$, is attained when

$$\nu^*(\boldsymbol{x}) = \frac{|g(\boldsymbol{x})|\pi(\boldsymbol{x})}{\int_{\mathbb{R}^p} |g(\boldsymbol{u})|\pi(\boldsymbol{u})d\boldsymbol{u}}.$$

However, the result is of a purely theoretical interest, as it requires the knowledge of the integral we need to calculate in the first place. On the other hand, using an approximation of the denominator, we get an estimate of ν^* which has a good performance (see [VDK83, RC13]). In the end, it is worth to mention that the accept-reject method and the importance sampling often have comparable behavior. To see more on this matter, we refer the reader to the following papers [CR98, Liu96].

The importance sampling method independently draws from the instrumental distribution ν. Although, $I_n^{\mathsf{imp}} \xrightarrow{a.s} \mathbb{E}_\pi(g(\boldsymbol{X}))$, the convergence can be slow. The latter happens because of the second order moment of the weights $\mathbb{E}_\nu[\pi^2(\boldsymbol{X}_i)/\nu^2(\boldsymbol{X}_i)]$. If the moment is large, then the importance sampling method is not stable and it may diverge in certain cases (see e.g. [Gew91, Gew89, RC13]). Thus, the choice of the instrumental distribution requires a scrupulous approach (see [Hes95, OZ00, CMMR12]).

1.4.5 Metropolis-Hastings algorithm

In this section, we propose another method of integration to approximate $\mathbb{E}_\pi[g(\boldsymbol{X})]$. Contrary to the previous methods, it does not require independent sampling at every step. Instead, it is based on generating an ergodic Markov-Chain $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ with π as its unique stationary distribution. The latter ensures the convergence of $\mathcal{L}(\boldsymbol{X}_n)$ to π. We want to stress that iterates of such a Markov chain are random vectors which approximately follow the law π. The more iterations we do, the closer we get to π, also referred to as the target distribution. Thus, this type of Markov chains can also be seen as approximate sampling methods. The MCMC estimator is defined as the empirical mean of the first n elements of the Markov-Chain $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$:

$$\widehat{I}_n^{\mathsf{MCMC}} = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i). \tag{MCMC}$$

The importance sampling method and the MCMC both satisfy $O(1/\sqrt{n})$ convergence rate. However, the instrumental distribution ν, that is sampled from at every iteration, can cause high computational cost, when compared to the MCMC iterates. The advantage of the MCMC is the Markov chain structure of the algorithm, which implies the dependence of each iterate on the previous ones. In addition, as mentioned above, the complexity of the importance largely depends on the proper choice of ν. In particular, it gets harder to choose the instrumental distribution when the dimension gets large. This phenomenon is an instance of “the curse of the dimensionality”.

Now, that we have described the general scheme of the MCMC integration, we proceed to the construction of such a Markov-Chain. The Metropolis-Hastings method was introduced by Metropolis [MRR+53] and later generalized by Hastings [Has70]. They describe a general method of calibrating the kernel of the Markov chain, such that π is its stationary distribution. The algorithm requires mild assumptions on π, which renders it applicable to a wide class of targets.

We assume that the target distribution is absolutely continuous w.r.t. some σ-additive measure. To implement the Metropolis-Hastings method, one needs to know its density function up to a constant factor. Then, an instrumental conditional density $\nu(\cdot \mid \boldsymbol{x})$ is defined for all $\boldsymbol{x} \in \mathbb{R}^p$. We initialize the chain at $\boldsymbol{X}_0$. For every $k \geq 0$, the $(k+1)$-th iteration of the chain starts with drawing a random vector $\boldsymbol{Y}$ from $\nu(\cdot \mid \boldsymbol{X}_k)$. Usually, ν belongs to some class of distributions that is easy to sample from. Afterwards, the drawn vector $\boldsymbol{Y}$ is accepted with a probability $\rho(\boldsymbol{X}_k, \boldsymbol{Y})$. The function ρ is defined as follows:

$$\rho(\boldsymbol{x},\boldsymbol{y}) := \min\left(\frac{\pi(\boldsymbol{y})\nu(\boldsymbol{x} \mid \boldsymbol{y})}{\pi(\boldsymbol{x})\nu(\boldsymbol{y} \mid \boldsymbol{x})}, 1\right).$$

The Metropolis-Hastings method is summarized in Algorithm 3. When the chain is

---

Algorithm 3 Metropolis-Hastings

Require: initial value $\boldsymbol{X}_0$
  for all $k = 1$ to $k = n$ **do**
    Sample independently $\boldsymbol{Y} \sim \nu(\cdot \mid \boldsymbol{X}_k)$ and $U \sim \mathcal{U}(0, 1)$
    if $U < \rho(\boldsymbol{X}_k, \boldsymbol{Y})$ **then**
      $\boldsymbol{X}_{k+1} \leftarrow \boldsymbol{Y}$
    else
      $\boldsymbol{X}_{k+1} \leftarrow \boldsymbol{X}_k$
    end if
  end for

---

irreducible and aperiodic, convergence to the stationary distribution in TV is available (see e.g. [MT12, RT96b]):

$$\|\mathcal{L}(\boldsymbol{X}_n) - \pi\|_{\mathsf{TV}} \xrightarrow[n\to\infty]{} 0.$$

Here, we want to underline the fact that the convergence does not depend on the initial state $\boldsymbol{X}_0$. Another important remark about the algorithm is that the density π can be known up to a constant, as it vanishes in the definition of $\rho(\cdot, \cdot)$. The rejection step, however, leaves certain samples $\boldsymbol{Y}$ out of the final estimation. Thus, there we "lose" some part of the drawn vectors. We had a similar issue, while applying the rejection-acceptance sampling. Nevertheless, the comparison of these methods (see [RC13, Liu96]) shows that the number of unused samples in proportion is lower for the Metropolis-Hastings algorithm.

The general form of the algorithm provides us with a powerful procedure, that can adapted to the problem in hand. Metropolis-Hastings algorithm, with a translation invariant conditional density $\nu(\boldsymbol{y} \mid \boldsymbol{x}) = \nu(\boldsymbol{y} - \boldsymbol{x})$ is called a Metropolis random walk. This method is discussed in several works (see e.g. [RT96b, MT$^+$96, Vem05, DCWY18]). Another popular particular case is the independent Metropolis-Hastings algorithm, which corresponds to the conditional distribution, that is independent of the current state: $\nu(\boldsymbol{y} \mid \boldsymbol{x}) = \nu(\boldsymbol{y})$. The convergence of this method was studied by [Tie94], while its application to the Gibbs sampler can be found in [GR93]. Finally, to accelerate the algorithm, techniques like Rao-Cramerization (see [GS94, MW00]) are applied.

1.5 Langevin sampling

In this chapter, we introduce the Langevin sampling algorithms, their origins, development and state-of-the-art complexity results. The goal is to sample from a probability distribution

π, defined on $\mathbb{R}^p$. The Langevin-type algorithms are based on a discretization of some stochastic differential equation, called the Langevin dynamics. The important property of this equation is that its solution has the target π as its invariant distribution. Moreover, under mild conditions, the solution is a geometrically ergodic continuous-time Markov process. This hints at using the process for sampling purposes. The idea to use the Langevin dynamics as a continuous-time simulation method was proposed by [GM94], in the context of pattern theory. The seminal work [RT96a] established the convergence of MALA (see the section below), which launched a line of research on the properties of the Langevin sampling algorithms (e.g. [ST99b, ST99c, RS02, AFMP11, GC11, XSL+14, Dal17b, Dal17a, DM17, CB18, DK19]). Before getting into details let us state the main assumptions and establish the notation for the rest of the chapter.

1.5.1 Assumptions and notation

We will assume that the target distribution π, defined on $\mathbb{R}^p$, has a density that is given by

$$\pi(\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta})),$$

where f is called the *potential function*. In addition, the potential function is assumed to be m-strongly convex (unless specified otherwise), differentiable and its gradient is assumed to be M-Lipschitz continuous See Section 1.3 for the definitions of these conditions.

Complexity and distances

To compare different algorithms, one needs a rigorous mathematical criterion. Suppose that we have some distance D on the space of probability measures $\mathcal{M}_q(\mathbb{R}^p)$. The latter is the space of probability distributions that have finite second order moment. Then, the *mixing-time* or the complexity of an iterative algorithm can be measured as

$$\mathcal{K}_\varepsilon := \inf\{k \mid \mathrm{D}(\mathcal{L}(\boldsymbol{X}_k), \pi) \leq \varepsilon\},$$

where $\boldsymbol{X}_k$ is the k-th iterate and ε is a positive number. In other words, $\mathcal{K}_\varepsilon$ is the number of iterations required to get an ε error in terms of the distance D. The complexity defined in this manner depends on D. Throughout the manuscript, we will encounter repetitively the following probability measure distances:

- Total variation distance

$$\|\nu - \nu'\|_{\mathsf{TV}} := \sup_{A \in \mathcal{B}(\mathbb{R}^p)} |\nu(A) - \nu'(A)|;$$

- Wasserstein-q distance

$$W_q(\nu, \nu') := \inf \Big\{ \mathbb{E}[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^q]^{1/q} : \boldsymbol{\theta} \sim \nu \text{ and } \boldsymbol{\theta}' \sim \nu' \Big\}; \tag{1.16}$$

- Kullback divergence

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

The TV distance is very common in the context of Markov processes and it can be found in the classical literature (e.g. [Ula52, Nel67]). It controls the probabilities, but does not provide information about the moments. Wasserstein-q distances, also known as the optimal transport cost, play an important role in the theory of optimal transport (see [Vil08]). An important property of W_p, is that the minimum is attained on the right-hand side of (1.16): there exists a distribution Γ defined on $\mathbb{R}^p \times \mathbb{R}^p$, such that it has ν and ν' as its marginals and that

$$W_p(\nu, \nu') = \mathbf{E}_{(\boldsymbol{\theta}, \boldsymbol{\theta}') \sim \Gamma} \big[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^p\big]^{1/p}.$$

Another property of W_p-distances, that is advantageous over TV-distance, is the control over the moments, which impacts the quality of statistical estimation. Finally, KL measures the information that is lost, when ν is used to approximate ν.

Log-concave measures

Log-concave measures gain more and more attention in statistics, probability and other branches of mathematics. In this section, we are going to present the definition and the basic properties of these measures, that are relevant in the problem of sampling.

We consider the probability distributions defined on $\mathbb{R}^p$. We say that π is a log-concave measure if it has a density, that is proportional to the exponent of a concave function:

$$\pi(\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta})), \quad \text{where} \quad f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \quad \text{is convex.}$$

The function f is called the *potential function* and it is usually assumed to be upper semi-continuous. One of the important features of these densities, is that they have sub-exponential tails (see e.g. [Bor83]). Similarly, we define m-strongly log-concave densities, for positive values of m. We say that π is m-strongly log-concave if has m-strongly convex potential:

$$\pi(\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta})), \quad \text{where} \quad f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \quad \text{is } m\text{-strongly convex}.$$

Many classical distribution are log-concave or strongly log-concave. Here is a list of examples:

- *Gaussian vectors* with parameters $(0, \Sigma)$ is σ^2-strongly convex, where σ^2 is the largest eigenvalue of the covariance matrix Σ.
- *Logistic density*, which is defined as $\pi_{\mathsf{log}}(\boldsymbol{\theta}) := \exp(-\boldsymbol{\theta})/(1 + \exp(-\boldsymbol{\theta}))^2$, is log-concave.
- *The Gamma distribution* with a degree of freedom $\alpha > 1$ is log-concave.
- *The Beta distribution* $B_{\alpha,\beta}$ with parameters $\alpha > 1$ and $\beta > 1$ is log-concave.

For more interesting examples, we refer the reader to [DJD88]. Log-concave measures are widely used in different ares, such as geometry [KLS95], game theory [CN91], and functional analysis [BBC$^+$08, Bob99]. Different properties of log-concave measures are extensively studied by many authors. The cornerstone of the interest towards the log-concave measures in probability theory, are the preservation properties.

Proposition 9 (Preservation). *The log-concave measures satisfy the following claims:*

i) Affine transformations: *Suppose $\boldsymbol{\xi}$ is a p-dimensional random vector with a log-concave density π and $A : \mathbb{R}^p \to \mathbb{R}^q$ is an injective affine map. Then the q-dimensional vector $A\boldsymbol{\xi}$ is also log-concave.*

ii) Products: *Suppose we have two log-concave measures π_1 and π_2 defined on $\mathbb{R}^p$. Then $\pi_1 \otimes \pi_2$ is a log-concave measure defined on $\mathbb{R}^p \times \mathbb{R}^p$.*

iii) Marginalization: *Suppose that $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is a log-concave measure on $\mathbb{R}^{p+q}$. Then the marginal distribution of the first component defined as $\pi_1(\boldsymbol{\theta}_1) = \int_{\mathbb{R}^q} \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) d\boldsymbol{\theta}_2$ is log-concave as well.*

The proof of the first two claims can be found in [DJD88]. The third property has been proved using different methods (see [Pré73, BL76, BBN$^+$03, SW14]).

We have already seen an example of a log-concave measure in the model of Bayesian logistic regression. As we have mentioned, the posterior distribution in that setting is log-concave. We also stressed that approximate integration techniques are required to estimate the posterior mean $\widehat{\boldsymbol{\theta}}_n^B$. We will give more insight in the coming sections on how to take advantage of log-concavity, when encountered with this problem.

1.5.2 Langevin diffusion

The Langevin diffusion is a stochastic differential equation, that was proposed by [Lan08] to describe the Brownian motion of a particle with a friction. Later, it was formulated rigorously by [Doo42] and [Nel67]. The diffusion has the following mathematical formula:

$$d\boldsymbol{L}_t^{\mathsf{LD}} = -\nabla f(\boldsymbol{L}_t^{\mathsf{LD}})dt + \sqrt{2}d\boldsymbol{W}_t, \tag{1.17}$$

where $\boldsymbol{W}$ is the standard Wiener process, and f is the potential function on $\mathbb{R}^p$. According to [Bha78], the diffusion is irreducible, strong Feller and aperiodic. In view of this, the distribution with density $\pi(\cdot) \propto \exp(-f(\cdot))$ is the stationary distribution of the diffusion (see [IW14, MT93]). Last but not least, for m-strongly log-concave potentials, an exponential convergence is available in TV [RT96a, Theorem 2.1] and Wasserstein distances [Vil08, Chapter 2]. In particular, for Wasserstein distances it is expressed as:

$$W_2(\nu_0 P_t, \pi) \leq \exp(-mt) W_2(\nu_0, \pi),$$

where P_t is the transition kernel of the Langevin diffusion and $\nu_0 = \mathcal{L}(\boldsymbol{L}_0^{\mathsf{LD}})$ is the initial distribution. Therefore, we deduce a continuous sampling scheme for smooth and strongly log-concave distributions π. As in the case of Metropolis-Hastings method, we see that the knowledge of π up to a constant is sufficient, since the normalization term vanishes in the gradient of the drift term. We can consider this process as a continuous-time sampling method. Unfortunately, in practice the implementation is not possible, which yields us to discretized approximations. One of the main approaches to this problem in the context of sampling, is called Euler-Maruyama discretization:

$$\boldsymbol{\vartheta}_{k+1} = \boldsymbol{\vartheta}_k - h\nabla f(\boldsymbol{\vartheta}_k) + \sqrt{2h}\boldsymbol{\xi}_k. \tag{LMC}$$

Here, h is the step-size and $(\boldsymbol{\xi}_k)_{k\in\mathbb{N}}$ is a sequence of standard normal vectors in $\mathbb{R}^p$ that are mutually independent and independent of the initial state $\boldsymbol{\vartheta}_0$. This iterative algorithm is called the Langevin Monte-Carlo (also known as Unadjusted Langevin Algorithm) (see [Par81]). Thus, we obtain a homogeneous Markov chain. We observe that the LMC updates resemble to the updates of the gradient descent and we will see several times in this manuscript how the two methods are connected.

Metropolis Adjusted Langevin Algorithm

Compared to its continuous counterpart, LMC is computationally feasible. However, some important properties of the Langevin diffusion are not transfered to its discretization. In particular, the target distribution π is not the stationary distribution of the obtained

Markov chain. Thus, some bias is generated, when discretizing the diffusion. [RT96a] propose to insert a Metropolis-Hastings rejection-acceptance step after every iteration of the LMC. Hence, in this setting, the Markov chain (LMC) plays the role of the instrumental conditional law ν. In particular, $\nu(\cdot \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} - h\nabla f(\boldsymbol{x}), 2hI_p)$ and

$$\rho^{\mathsf{MALA}}(\boldsymbol{x}, \boldsymbol{y}) = \min\left(1, \frac{\exp(-f(\boldsymbol{y}) - \|\boldsymbol{x} - \boldsymbol{y} + h\nabla f(\boldsymbol{y})\|_2^2/4h)}{\exp(-f(\boldsymbol{x}) - \|\boldsymbol{y} - \boldsymbol{x} + h\nabla f(\boldsymbol{x})\|_2^2/4h)}\right).$$

The method is described in Algorithm 4 and it is called the Metropolis Adjusted Langevin Algorithm (MALA). The authors show that for log-concave target measures π, MALA is uniformly ergodic. For convergence guarantees of MALA, we refer the reader to [ST99b, ST99c, RS02, SFCM13, DCWY18].

Algorithm 4 MALA

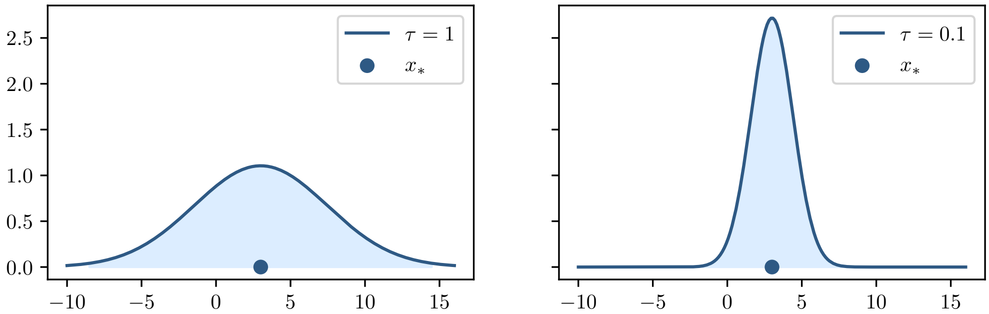
Require: initial value $\boldsymbol{\vartheta}_0$
  for all $k = 0$ to $k = n - 1$ **do**
    Sample independently $\boldsymbol{\xi}_k \sim \mathcal{N}(0, I_p)$ and $U \sim \mathcal{U}(0, 1)$.
    $\boldsymbol{Y}_{k+1} \leftarrow \boldsymbol{\vartheta}_k - h\nabla f(\boldsymbol{\vartheta}_k) + \sqrt{2h}\boldsymbol{\xi}_k$
    $\rho_k \leftarrow \rho^{\mathsf{MALA}}(\boldsymbol{\vartheta}_k, \boldsymbol{Y}_{k+1})$
    if $U < \rho_k$ **then**
      $\boldsymbol{\vartheta}_{k+1} \leftarrow \boldsymbol{Y}_{k+1}$
    else
      $\boldsymbol{\vartheta}_{k+1} \leftarrow \boldsymbol{\theta}_k$
    end if
  end for

1.5.3 Langevin Monte-Carlo

Another contribution of [RT96a] was the analysis of the LMC algorithm. They show that the unadjusted algorithm is very sensitive to choice of the step-size h. If the step-size is relatively large, then the Markov chain is not ergodic and even transient. Hence, the authors underline the importance of the Metropolis adjustment step and recommend not to use the algorithm without it. However, the efficiency of Metropolis-Hastings methods reduces in higher dimensions, as the acceptance probability ρ_k (see Algorithm 4) is inversely scaled in the dimension. The latter slows down the exploration of the space, and therefore, it augments the mixing-time.

The main issue of this approach underlies in the firm requirement to have a method that is consistent. In this settings, it means to have a Markov chain, whose iterates converge to the target π, in terms of some probabilistic measure distance. This results large computational complexity, when the target distribution is complex. However, one

Figure 1.1: Illustration of the concentration around the minimum point. The colored part under the graphs corresponds to the confidence interval of level 0.95.



would be interested in having a chain that does not converge to π, but rather converges to another distribution π_h, that has a low mixing-time. This is an example of a bias-variance trade-off, where the biased algorithm has a significantly lower complexity than its unbiased counterpart.

This idea was implemented by [Dal17b], and polynomial non-asymptotic bounds on the convergence error in TV distance were proved for LMC. It was shown that, for a proper choice of the initial distribution (warm start), the complexity of LMC is of order $\widetilde{O}(\kappa^2 p/\varepsilon^2)$, where $\kappa = M/m$ is the condition number. Later, [DM19] proved that this convergence rate is available without warm start, both in TV and W_2 distances. Overall, these results revived the interest towards the Langevin sampling methods and many articles have since been out that study the convergence of the LMC and its variants in different settings (e.g. [HSR19, DMM19, Wib18, DRD20, CCBJ18, CB18, MCJ$^+$19, SL19]) and their applications ([RRT17, XCZG18, GPP20]).

1.5.4 Relation with optimization

Suppose that we can sample from the distribution π, using (LMC). Then, in view of the scalability of our algorithm, we can also sample from $\pi^\tau(\cdot) \propto \exp(-f(\cdot)/\tau)$, where τ is the temperature parameter (see [Nel67]). When $\tau \to 0$, the distribution π^τ converges to $\delta_{\boldsymbol{x}_*}$. Here, $\boldsymbol{x}_*$ is the minimizer of the potential function f, and $\delta_{\boldsymbol{x}_*}$ is the Dirac measure at the point $\boldsymbol{x}_*$. Thus, when the temperature parameter is small, samples from π^τ concentrate around $\boldsymbol{x}_*$ with high probability. See Figure 1.1, for an illustration of this phenomenon, and Section 2.5, for a detailed explanation.

Let us take a closer look at the formulas of the Langevin Monte-Carlo and the stochastic gradient descent. In (LMC) the noise term $\boldsymbol{\xi}_k$ is multiplied by $\sqrt{h}$, while in the (SGD) by h. In fact, this mismatch yields an entirely different limiting behavior. Indeed, when $h \to 0$, (SGD) tends to the ODE (GF), while the (LMC) converges to the SDE (1.17). Many authors have explored several variants of the LMC which are inspired from analogous algorithms in optimization. For instance, see [CFM+18, DBLJ14] for variance reduction techniques, [Per14, PB14] for proximal methods and [HSR19, PB14] for implicit methods.

The analogy between the SGD and the LMC can also be established by introducing the stochasticity of the gradient to the algorithm of sampling. We assume the availability of an oracle, that provides us with noisy gradients. This setting is of use in the case when the potential function has a sum-decomposable form. Convergence results for this setting were established in [NDH+17, DK19, SKR19, DDB+20]. See Chapter 2 for more details.

Another relation with the optimization is established through the gradient flows in Wasserstein spaces $(\mathcal{M}_2(\mathbb{R}^p), W_2)$. It turns out, that the Langevin diffusion is the gradient flow of the functional $H : \mathcal{M}_2(\mathbb{R}^p) \to \mathbb{R}^+$, with $H(\nu) := \mathsf{KL}(\nu, \pi)$ (see [JKO98, San17] for more details). This means that the Langevin Monte-Carlo can be viewed as a gradient descent method for the functional H. Based on this property convergence rates of the LMC can be derived (see [Wib18, DMM19]).

Finally, there is a line of research, that studies the efficiency of the LMC and its variants in the context of optimization. As mentioned previously, the Gaussian noise can be viewed as a non-vanishing gradient noise. The latter allows the algorithm to escape the local minima of the non-convex potentials. This setting appears in [RRT17, XCZG18].

1.5.5 Higher-order methods

In this section, we discuss a higher-order Langevin based method, called the Kinetic Langevin Monte-Carlo (KLMC). It is based on a system of SDEs called Kinetic Langevin diffusion:

$$\begin{aligned} d\boldsymbol{L}_t^{\mathsf{KLD}} &= \boldsymbol{V}_t^{\mathsf{KLD}} dt; \\ d\boldsymbol{V}_t^{\mathsf{KLD}} &= -\left(\eta \boldsymbol{V}_t^{\mathsf{KLD}} + \nabla f(\boldsymbol{L}_t^{\mathsf{KLD}})\right) dt + \sqrt{2\eta}\boldsymbol{W}_t, \end{aligned}$$

where η is the friction parameter and $\boldsymbol{W}$ is the standard Wiener process. This equation was originally designed to model the movement of a Brownian particle, when the friction is large (see [Kra40, Nel67]). The vector $\boldsymbol{L}_t^{\mathsf{KLD}}$ describes the position at time t, while $\boldsymbol{V}_t^{\mathsf{KLD}}$ corresponds to its velocity. The Langevin diffusion (overdamped) is the limit of its rescaled kinetic (underdamped) counterpart $\bar{\boldsymbol{L}}_t = \boldsymbol{L}_{\eta t}^{\mathsf{KLD}}$, when $\eta \to +\infty$. The reason that the KLD is interesting for sampling, is that it has a stationary distribution $\mathsf{P}(\boldsymbol{\theta}, \boldsymbol{v}) \propto \exp(-f(\boldsymbol{\theta}) - \|\boldsymbol{v}\|_2^2/2)$. Moreover, the process is proved to be ergodic (see e.g. [Vil08, EGZ19, Tal02, MSH02, DRD20]). In particular, [DRD20] have proved that the convergence to the stationary measure in Wasserstein-2 distance is of exponential order, for twice-differentiable and strongly convex potential functions f. Thus, KLD is a continuous sampling scheme and a discretization scheme can make it applicable in practice. One of the first contributions in this area was made by [CCBJ18]. They propose the following discretization of KLD which is called the Kinetic Langevin Monte-Carlo:

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\vartheta}_k) \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\vartheta}_k) \end{bmatrix} + \sqrt{2\gamma}\begin{bmatrix} \boldsymbol{\xi}_{k+1} \\ \boldsymbol{\xi}'_{k+1} \end{bmatrix}. \tag{KLMC}$$

Here the random vectors $\boldsymbol{\xi}_k, \boldsymbol{\xi}'_k$ and the auxiliary function ψ_i satisfy the following conditions:

- the vectors $(\boldsymbol{\xi}_k, \boldsymbol{\xi}'_k)_k$ are a sequence of i.i.d. $2p$-dimensional centered Gaussian vectors that are independent of initial conditions;

- $\psi_0(u) = \exp(-\eta u)$ and $\psi_{j+1} = \int_0^u \psi_j(u)du$, for $j = 0, 1$;

- for every $k \in \mathbb{N}$ the 2-dimensional vectors $((\boldsymbol{\xi}_k)_1, (\boldsymbol{\xi}'_k)_1), ((\boldsymbol{\xi}_k)_2, (\boldsymbol{\xi}'_k)_2), \ldots, ((\boldsymbol{\xi}_k)_p, (\boldsymbol{\xi}'_k)_p)$ are i.i.d. with covariance matrix

$$\mathbf{C} = \int_0^h [\psi_0(t)\psi_1(t)]^\top [\psi_1(t)\psi_2(t)]dt.$$

Despite the complex form of the discretization formula, it has an intuitive interpretation. To understand the essence of this scheme, one needs to discretize the time axis into intervals of size h: $[kh, (k+1)h]$. Afterwards, an analysis similar to the one for LMC, can be performed. In addition, for m-strongly convex and M-gradient Lipschitz potentials f, [CCBJ18] proved tractable polynomial bounds on the Wasserstein error, depending on the parameters m, M and η. Later, [DRD20] have shown that these results can be improved with another discretization scheme, when second-order smoothness of f is available. The virtue of KLMC is that it has smoother trajectories, which result better approximation when discretized, as compared to LMC. This phenomenon manifests itself, when we compare the dependence of mixing-times in dimension p the precision ε of both methods. Indeed, for m-strongly convex and M-smooth potentials KLMC achieves ε Wasserstein error in $\widetilde{O}(\kappa^{3/2}\{\kappa \bigvee (p/m\varepsilon^2)\}^{1/2})$, while LMC requires $\widetilde{O}(\kappa\{1 \bigvee (p/m\varepsilon^2)\})$ iterations. Here $\widetilde{O}$ is the asymptotic order without the logarithmic terms and $\kappa = M/m$ is the condition number. We see, indeed, that KLMC has a much better dependence in p and ε. However, it falls short in terms of κ. The latter means that these methods are not generally comparable since the condition number can be large in certain cases.

1.6 Contributions

Langevin Monte-Carlo is an effective method of sampling from probability distributions in high dimensions. It is a Markov chain, that is a discretization of a stochastic differential equation, called the Langevin diffusion (1.17). The main assumption on the target probability distribution is the absolute continuity w.r.t. some σ-additive measure, with a log-concave density $\pi(\cdot) \propto \exp(-f(\cdot))$. Here, π is the target density, while f is the potential function. In the case, when f is smooth and strongly convex, the solution of the Langevin diffusion is a Markov process, which is ergodic and its distribution converges exponentially to π in the Wasserstein distance. This property hints at the discretization of the process, which yields to an iterative algorithm, called the Langevin Monte-Carlo (LMC). The application of LMC to the problem of sampling was studied by [RT96a]. In order to have a Markov chain converging to the target π, the authors propose to adjust every iteration with a Metropolis-Hastings step (MALA). The convergence of MALA has been extensively studied by many authors (see e.g. [RT96a, RR98, ST99b, ST99c, JH00, RS02]). Later, [Dal17b] proved non-asymptotic polynomial error bounds in the total variation (TV) distance for the Langevin Monte-Carlo, without the adjustment step. The key idea is to have a biased method, that has a smaller variance, and therefore, an overall smaller convergence error. Moreover, under certain conditions (warm start), the convergence rates can be improved. [DM17, DM19] showed that for all initial distributions this improved convergence error bounds are available in TV and Wasserstein distances. This has initiated a line of research, which studies various statistical properties of the Langevin Monte-Carlo and its modifications (see e.g. [DMM19, DM17, DRD20, CB18, CLGL$^+$20, KD20]).

The rest of the manuscript is dedicated to the study of the Langevin sampling methods. It consists of three chapters. Chapter 2 studies the discrete the Langevin algorithm with inaccurate gradients and with higher-order smoothness, while Chapter 3 and 4 focus on the non-strongly convex case for the discrete and continuous schemes, respectively.

In Chapter 2, we study the convergence of LMC, for M-smooth and m-strongly convex functions. [DM19] proved that LMC converges to the target distribution π in Wasserstein distance, with a convergence rate of $\widetilde{O}(p/\varepsilon^2)$. We push further their analysis and generalize the result in two directions. Firstly, we remove the assumption on the availability of the exact evaluations of the gradient of the log-density. Instead, we assume that a stochastic estimation of the gradient is available. A unified framework for handling both deterministic and stochastic approximations is proposed. We subsequently provide an upper bound on the sampling error of the first-order noisy LMC, that quantifies the impact of the gradient evaluation inaccuracies. Secondly, we study the effect of the second-order smoothness of the potential on the convergence analysis. Here, by second-order smoothness we mean the Lipschitz continuity of the Hessian of the potential. Upper bounds on the error of convergence in the Wasserstein distance are established for the first-order noisy LMC, as well as for the LMCO′. The latter is a second-order discretization method, that exploits the Hessian of the log-density. We observe that the additional smoothness on the first-order LMC yields better convergence rates and that the LMC algorithm is adaptive. Chapter 2 is based on the article *"User-friendly guarantees for the Langevin Monte-Carlo with inaccurate gradient"*, which is a joint work with Arnak Dalalyan. It is published in *Stochastic Processes and their Applications*.

In the classical literature of Langevin sampling, the potential function is assumed to be strongly-convex. In practice, however, the density at hand is not strongly log-concave. Chapter 3 focuses on LMC and KLMC in the general log-concave case. The first part of the chapter is dedicated to the study of the Langevin sampling algorithms with a non-strongly convex potential function. To overcome the lack of strong convexity, we introduce a surrogate potential, by adding a fixed square penalty. We then provide explicit non-asymptotic bounds on the Wasserstein error of LMC and KLMC with an explicit dependence on the magnitude of the added penalty. Similar convergence rates are also established for the potentials that satisfy higher-order smoothness assumptions. Our results provide some clear guidance for the calibration of the quadratic penalty, which reduces the sampling error and provides the best known guarantees in the non-strongly convex setting. The rest of the chapter studies the bounds on the second-order moment of log-concave densities in two cases: the potential function is strongly-convex, respectively, inside and outside of some Euclidean ball. Chapter 3 is based on the article *"Bounding the error of the discretized Langevin algorithms for non-strongly log-concave targets"*, which is a joint work with Arnak Dalalyan and Lionel Riou-Durand. The paper is submitted to the *Journal of Machine Learning Research*.

In Chapter 4, we introduce a continuous-time process that converges polynomially to the target distribution π with a non-strongly convex potential. We propose an adjustment of the Langevin diffusion, termed Penalized Langevin dynamics (PLD). It is defined as a continuous-time diffusion-type process, with the negative gradient of the potential plus a vanishing time-dependent penalty (linear in the state variable) as its drift. We provide explicit bounds on the Wasserstein distance between the distribution of the PLD at moment t and the target distribution. This upper bound provides a precise characterization of the influence of the penalty on the approximation error. The described bound, after optimization with respect to the penalty term, also allows to show that the PLD converges to the target π at rate $O(1/\sqrt{T})$. Following a similar logic as described in Section 1.5 (see also [Dal17a]), we consider the gradient flow as the low-temperature limit of the Langevin dynamics. This motivates the application of the proposed penalization scheme to gradient flows, from which we deduce new non-asymptotic guarantees for their convergence.

Chapter 4 is based on the article *"Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets"*, which is a joint work with Arnak Dalalyan. The paper is published in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

1.7 Résumé substantiel

Langevin Monte-Carlo est une méthode efficace d'échantillonnage de lois en grandes dimensions. Cette chaîne de Markov est obtenue par la discretization de l'équation stochastique différentielle de Langevin (1.17). L'hypothèse principale c'est que la loi cible est absolument continue par rapport à une mesure σ-additive et qu'elle a une densité log-concave $\pi(\cdot) \propto \exp(-f(\cdot))$ où π est densité cible et f est la fonction potentielle. Quand f est régulière et fortement convexe, la solution de cette équation est un processus de Markov érgodique et sa loi à l'instant t converge vers π en distance de Wasserstein, quand t tend vers l'infini. Cette propriété fait réference à la discretization du processus, qui nous donne un algorithme itératif, appellé Langevin Monte-Carlo (LMC). L'utilisation du LMC dans le problème d'échantillonnage est étudiée par [RT96a]. Afin d'avoir une chaîne qui converge vers π, les auteurs proposent d'ajouter une étape de Metropolis-Hastings à chaque itération de l'algorithme (MALA). La convergence de MALA a été étudié par plusieurs auteurs (cf. [RT96a, RR98, ST99b, ST99c, JH00, RS02]). Ensuite, [Dal17b] a prouvé des bornes polynomiales non-asymptotiques sur l'erreur en distance de variation totale pour LMC, sans modification du pas de Metropolis-Hastings. L'idée c'est d'avoir une méthode biaisé, mais avec une plus petite variance, et donc avec une erreur de convergence inférieure à celle de MALA. En plus, [Dal17b] a montré que la convergence peut être améliorée, si certaines conditions initiales sont disponible (warm start). Plus tard, [DM17, DM19] ont prouvé que la convergence améliorée, en TV et Wasserstein-2, est atteinte sans warm start. Ces résultats ont initié une série d'articles étudiant l'échantillonnage par les algorithmes de Langevin (cf. [DMM19, DM17, DRD20, CB18, CLGL+20]).

Le reste du manuscrit est dédié à l'étude des méthodes de Langevin pour l'échantillonnage. Il consiste de trois chapitre, dont Chapter 2 est dédié à l'algorithme de Langevin avec des gradients non-précis et avec régularité supplémentaire, alors que Chapter 3 et 4 présentent respectivement des résultats pour les schémas discrets et continus quand la fonction potentielle est faiblement convexe.

Dans le Chapitre 2, on étudie la convergence du LMC dans le cas des fonctions potentielles m-fortement convexe et M-lisse. [DM19] ont prouvé que le LMC converge vers la loi π en distance de Wasserstein à une vitesse $\widetilde{O}(p/\varepsilon^2)$. Nous développons leur analyse et proposons des résultats dans deux diréctions. D'abord nous supprimons l'hypothèse de disponibilité des gradients de log-densité. Nous supposons plutôt que nous disposons une approximation stochastique du gradient. Un cadre d'analyse unifié est proposé pour les approximations stochastiques et déterministes. Nous prouvons des bornes non-asymptotiques pour le LMC de premier ordre, où la dépendance en bruit des calculs des gradients est explicitement donnée. Ensuite, nous étudions l'impact de la régularité de deuxième ordre de la fonction potentielle sur l'analyse de la convergence. Ici, la régularité signifie que la Hessienne est lipschitzienne. Des bornes supérieures sont obtenues pour la convergence en distance de Wasserstein pour le LMC et le LMCO$'$. On observe, en effet, que la régularité supplémentaire améliore la qualité de l'échantillonnage. Chapitre Chapter 2 est basé sur l'article *"User-friendly guarantees for the Langevin Monte-Carlo with inaccurate gradient"*, qui est un travail réalise conjointement avec Arnak Dalalyan. Le travail a été publié dans *Stochastic Processes and their Applications*.

Dans la littérature classique de l'echantillonnage de Langevin, la fonction potentielle est supposée être fortement convexe. Par contre, dans les applications ce n'est pas souvent le cas. Chapitre 3 considère le cas des densités qui sont log-concaves. La première partie du chapitre est consacrée à l'étude de LMC et KLMC pour les fonctions potentielles log-convexes. Afin de revenir au cas fortement convexe, nous proposons un substitut de la fonction potentielle en ajoutant la pénalité quadratique. Des bornes supérieures sont prouvées pour l'erreur de convergence de LMC, KLMC et KLMC-2 avec la potentielle modifiée. Les bornes contiennent aussi la dépendance explicite de la largeur de la pénalité. Nous décrivons l'analyse complète du choix optimal de la pénalité et des paramètres des algorithmes, afin d'obtenir la meilleure vitesse de convergence. La deuxième partie du papier étudie des bornes sur le deuxième moment des densités log-concaves dans les deux cas suivants: la fonction potentielle est fortement convexe à l'intérieure et respectivement, à l'exterieure d'une boule Euclidéenne centrée à 0. Le Chapitre 3 est basé sur l'article *"Bounding the error of the discretized Langevin algorithms for non-strongly log-concave targets"*, qui est un travail réalisé conjointement avec Arnak Dalalyan et Lionel Riou-Durand. Le papier est soumis à *Journal of Machine Learning Research*.

Dans le Chapitre 4, nous introduisons un processus continu qui converge à une vitesse polynomiale ves la loi cible π, qui est log-concave. Nous proposons un ajustement de la diffusion de Langevin, qui s'appelle PLD (Penalized Langevin Dynamics). Il s'agit d'un processus de type diffusion qui a pour drift l'opposé du gradient de la fonction potentielle. La pénalisation proposée est linéaire et tend vers 0 quand $t \to +\infty$. Nous donnons des bornes explicites sur la distance de Wasserstein entre la loi cible et la loi du PLD à l'instant t. Cette borne décrit explicitement la dépendance de la pénalité. Nous montrons qu'après avoir optimisé la pénalité, nous obtenons une convergence en $O(1/\sqrt{T})$. Suivant une logique similaire à celle de la Section 1.5 (voir aussi [Dal17a]), nous considèrons le flow du gradient comme une limite de la diffusion de Langevin. Cela nous permet d'appliquer la même schéma de pénalisation au flow du gradient, duquel on déduit des nouvelles garanties non-asymptotiques de la convergence. Le Chapitre 4 est basée sur l'article *"Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets"*, qui est un travail réaliśe conjointement avec Arnak Dalalyan. Le papier a été publié dans *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Chapter 2

Langevin Monte-Carlo with inaccurate gradient

Abstract

In this paper, we study the problem of sampling from a given probability density function that is known to be smooth and strongly log-concave. We analyze several methods of approximate sampling based on discretizations of the (highly overdamped) Langevin diffusion and establish guarantees on its error measured in the Wasserstein-2 distance. Our guarantees improve or extend the state-of-the-art results in three directions. First, we provide an upper bound on the error of the first-order Langevin Monte Carlo (LMC) algorithm with optimized varying step-size. This result has the advantage of being horizon free (we do not need to know in advance the target precision) and to improve by a logarithmic factor the corresponding result for the constant step-size. Second, we study the case where accurate evaluations of the gradient of the log-density are unavailable, but one can have access to approximations of the aforementioned gradient. In such a situation, we consider both deterministic and stochastic approximations of the gradient and provide an upper bound on the sampling error of the first-order LMC that quantifies the impact of the gradient evaluation inaccuracies. Third, we establish upper bounds for two versions of the second-order LMC, which leverage the Hessian of the log-density. We provide nonasymptotic guarantees on the sampling error of these second-order LMCs. These guarantees reveal that the second-order LMC algorithms improve on the first-order LMC in ill-conditioned settings.
This chapter is based on a joint work with Arnak Dalalyan called "User-friendly guarantees for the Langevin Monte-Carlo with inaccurate gradient". It is published in *Stochastic Processes and their Applications*.

2.1 Introduction

The problem of sampling a random vector distributed according to a given target distribution is central in many applications. In the present paper, we consider this problem in the case of a target distribution having a smooth and log-concave density π and when the sampling is performed by a version of the Langevin Monte Carlo algorithm (LMC). More precisely, for a positive integer p, we consider a continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ satisfying the following assumption: For some positive constants m and M, it holds

$$\begin{cases} f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') - \nabla f(\boldsymbol{\theta}')^\top(\boldsymbol{\theta} - \boldsymbol{\theta}') \geq ({}^m\!/\!{}_2)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2, \\ \|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \leq M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \end{cases} \qquad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p, \tag{2.1}$$

where ∇f stands for the gradient of f and $\|\cdot\|_2$ is the Euclidean norm. The target distributions considered in this paper are those having a density with respect to the Lebesgue measure on $\mathbb{R}^p$ given by

$$\pi(\boldsymbol{\theta}) = \frac{e^{-f(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f(\boldsymbol{u})}\, d\boldsymbol{u}}.$$

We say that the density $\pi(\boldsymbol{\theta}) \propto e^{-f(\boldsymbol{\theta})}$ is log-concave (resp. strongly log-concave) if the function f satisfies the first inequality of (2.1) with $m = 0$ (resp. $m > 0$).

Most part of this work focused on the analysis of the LMC algorithm, which can be seen as the analogue in the problem of sampling of the gradient descent algorithm for optimization. For a sequence of positive parameters $\boldsymbol{h} = \{h_k\}_{k\in\mathbb{N}}$, referred to as the step-sizes and for an initial point $\boldsymbol{\vartheta}_{0,\boldsymbol{h}} \in \mathbb{R}^p$ that may be deterministic or random, the iterations of the LMC algorithm are defined by the update rule

$$\boldsymbol{\vartheta}_{k+1,\boldsymbol{h}} = \boldsymbol{\vartheta}_{k,\boldsymbol{h}} - h_{k+1}\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}}) + \sqrt{2h_{k+1}}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{2.2}$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k, \ldots$ is a sequence of mutually independent, and independent of $\boldsymbol{\vartheta}_{0,\boldsymbol{h}}$, centered Gaussian vectors with covariance matrices equal to identity.

When all the h_k's are equal to some value $h > 0$, we will call the sequence in (2.2) the constant step LMC and will denote it by $\boldsymbol{\vartheta}_{k+1,h}$. When f satisfies assumptions (2.1), if h is small and k is large (so that the product kh is large), the distribution of $\boldsymbol{\vartheta}_{k,h}$ is known to be a good approximation to the distribution with density $\pi(\boldsymbol{\theta})$. An important question is to quantify the quality of this approximation. An appealing approach to address this question is by establishing non asymptotic upper bounds on the error of sampling; this kind of bounds are particularly useful for deriving a stopping rule for the LMC algorithm, as well as for understanding the computational complexity of sampling methods in high dimensional problems. In the present paper we establish such bounds by focusing on their user-friendliness. The latter means that our bounds are easy to interpret, hold under conditions that are not difficult to check and lead to simple theoretically grounded choice of the number of iterations and the step-size.

In the present work, we measure the error of sampling in the Wasserstein-Monge-Kantorovich distance W_2. For two measures μ and ν defined on $(\mathbb{R}^p, \mathscr{B}(\mathbb{R}^p))$, and for a real number $q \geq 1$, W_q is defined by

$$W_q(\mu,\nu) = \Big(\inf_{\varrho\in\varrho(\mu,\nu)} \int_{\mathbb{R}^p\times\mathbb{R}^p} \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^q\, d\varrho(\boldsymbol{\theta},\boldsymbol{\theta}')\Big)^{1/q},$$

where the $\inf$ is with respect to all joint distributions ϱ having μ and ν as marginal distributions. For statistical and machine learning applications, we believe that this distance is more suitable for assessing the quality of approximate sampling schemes than other metrics such as the total variation or the Kullback-Leibler divergence. Indeed, bounds on the Wasserstein distance—unlike the bounds on the total-variation—provide direct guarantees on the accuracy of approximating the first and the second order moments.

Asymptotic properties of the LMC algorithm, also known as Unadjusted Langevin Algorithm (ULA), and its Metropolis adjusted version, MALA, have been studied in a number of papers [RT96a, RR98, ST99b, ST99c, JH00, RS02]. These results do not emphasize the effect of the dimension on the computational complexity of the algorithm, which is roughly proportional to the number of iterations. Non asymptotic bounds on the total variation error of the LMC for log-concave and strongly log-concave distributions have been established by [Dal17b]. If a warm start is available, the results in [Dal17b] imply that after $O(p/\varepsilon^2)$ iterations the LMC algorithm has an error bounded from above by ε. Furthermore, if we assume that in addition to (2.1) the function f has a Lipschitz continuous Hessian, then a modified version of the LMC, the LMC with Ozaki discretization (LMCO), needs $O(p/\varepsilon)$ iterations to achieve a precision level ε. These results were improved and extended to the Wasserstein distance by [DM17, DM19]. More precisely, they removed the condition of the warm start and proved that under the Lipschitz continuity assumption on the Hessian of f, it is not necessary to modify the LMC for getting the rate $O(p/\varepsilon)$. The last result is closely related to an error bound between a diffusion process and its Euler discretization established by [AJKH14].

On a related note, [BEL18] studied the convergence of the LMC algorithm with reflection at the boundary of a compact set, which makes it possible to sample from a compactly supported density (see also [BDMP17]). Extensions to non-smooth densities were presented in [DMP18, LFC17]. [CB18] obtained guarantees similar to those in [Dal17b] when the error is measured by the Kullback-Leibler divergence. Very recently, [CCBJ18] derived non asymptotic guarantees for the kinetic LMC which turned out to improve on the previously known results. Langevin dynamics was used in [ARW16, BDM18] in order to approximate normalizing constants of target distributions. [HZ17] established tight bounds in Wasserstein distance between the invariant distributions of two (Langevin) diffusions; the bounds involve mixing rates of the diffusions and the deviation in their drifts.

The goal of the present work is to push further the study of the LMC and its variants both by improving the existing guarantees and by extending them in some directions. Our main contributions can be summarized as follows:

- We state simplified guarantees in Wasserstein distance with improved constants both for the LMC and the LMCO when the step-size is constant, see Theorem 4 and Theorem 9.
- We propose a varying-step LMC which avoids a logarithmic factor in the number of iterations required to achieve a precision level ε, see Theorem 5.
- We extend the previous guarantees to the case where accurate evaluations of the gradient are unavailable. Thus, at each iteration of the algorithm, the gradient is computed within an error that has a deterministic and a stochastic component. Theorem 7 deals with functions f satisfying (2.1), whereas Theorem 8 requires the additional assumption of the smoothness of the Hessian of f.
- We propose a new second-order sampling algorithm termed LMCO$'$. It has a per-iteration computational cost comparable to that of the LMC and enjoys nearly the same guarantees as the LMCO, when the Hessian of f is Lipschitz continuous, see Theorem 9.
- We provide a detailed discussion of the relations between, on the one hand, the sampling methods and guarantees of their convergence and, on the other hand, optimization methods and guarantees of their convergence (see Section 2.5).

We have to emphasize right away that Theorem 4 is a corrected version of [Dal17a, Theorem 1], whereas Theorem 7 extends [Dal17a, Theorem 3] to more general noise. In particular, Theorem 7 removes the unbiasedness and independence conditions. Furthermore, thanks to a shrewd use of a recursive inequality, the upper bound in Theorem 7 is tighter than the one in [Dal17a, Theorem 3].

As an illustration of the first two bullets mentioned in the above summary of our contributions, let us consider the following example. Assume that $m = 10$, $M = 20$ and we have at our disposal an initial sampling distribution ν_0 satisfying $W_2(\nu_0, \pi) = p+(p/m)$. The main inequalities in Theorem 4 and Theorem 5 imply that after K iterations, the distribution ν_K obtained by the LMC algorithm satisfies

$$W_2(\nu_K,\pi) \leq (1-mh)^K W_2(\nu_0,\pi) + 1.65(M/m)(hp)^{1/2} \tag{2.3}$$

for the constant step LMC and

$$W_2(\nu_K,\pi) \leq \frac{3.5M\sqrt{p}}{m\sqrt{M+m+(2/3)m(K-K_1)}} \tag{2.4}$$

for the varying-step LMC, where K_1 is an integer the precise value of which is provided in Theorem 5. One can compare these inequalities with the corresponding bound in [DM19]: adapted to the constant-step, it takes the form

$$\begin{aligned} W_2^2(\nu_K,\pi) \leq& 2\Big(1-\frac{mMh}{m+M}\Big)^K W_2^2(\nu_0,\pi) \\ &+ \frac{Mhp}{m}(m+M)\Big(h+\frac{m+M}{2mM}\Big)\Big(2+\frac{M^2h}{m}+\frac{M^2h^2}{6}\Big). \end{aligned} \tag{2.5}$$

For any $\varepsilon > 0$, we can derive from these guarantees the smallest number of iterations, K_ε, for which there is a $h > 0$ such that the corresponding upper bound is smaller than ε. The logarithms of these values K_ε for varying $\varepsilon \in \{0.001, 0.005, 0.02\}$ and $p \in \{25, \ldots, 1000\}$ are plotted in Figure 2.1. We observe that for all the considered values of ε and p, the number of iterations derived from (2.4) (referred to as Theorem 5) is smaller than those derived from (2.3) (referred to as Theorem 4) and from (2.5) (referred to as DM bound). The difference between the varying-step LMC and the constant step LMC becomes more important when the target precision level ε gets smaller. In average over all values of p, when $\varepsilon = 0.001$, the number of iterations derived from (2.5) is 4.6 times larger than that derived from (2.4), and almost 3 times larger than the number of iterations derived from (2.3).

2.2 Guarantees in the Wasserstein distance with accurate gradient

The rationale behind the LMC (2.2) is simple: the Markov chain $\{\boldsymbol{\vartheta}_{k,\boldsymbol{h}}\}_{k\in\mathbb{N}}$ is the Euler discretization of a continuous-time diffusion process $\{\boldsymbol{L}_t : t \in \mathbb{R}_+\}$, known as Langevin diffusion. The latter is defined by the stochastic differential equation

$$d\boldsymbol{L}_t = -\nabla f(\boldsymbol{L}_t)\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \qquad t \geq 0, \tag{2.6}$$

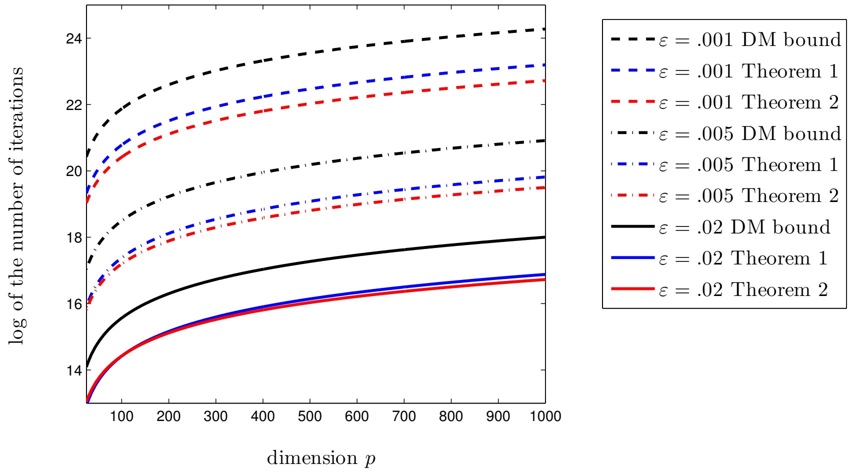


Figure 2.1: Plots showing the logarithm of the number of iterations as function of dimension p for several values of ε. The plotted values are derived from (2.3)-(2.5) using the data $m = 10$, $M = 20$, $W_2^2(\nu_0, \pi) = p + (p/m)$.

where $\{\boldsymbol{W}_t : t \geq 0\}$ is a p-dimensional Brownian motion. When f satisfies condition (2.1), equation (2.6) has a unique strong solution, which is a Markov process. Furthermore, the process $\boldsymbol{L}$ has π as invariant density [Bha78, Thm. 3.5]. Let ν_k be the distribution of the k-th iterate of the LMC algorithm, that is $\vartheta_{k,\boldsymbol{h}} \sim \nu_k$. In what follows, we present user-friendly guarantees on the closeness of ν_k and π, when f is strongly convex.

2.2.1 Reminder on guarantees for the constant-step LMC

When the function f is m-strongly convex and M-gradient Lipschitz, upper bounds on the sampling error measured in Wasserstein distance of the LMC algorithm have been established in [DM19, Dal17a]. We state below a slightly adapted version of their result, which will serve as a benchmark for the bounds obtained in this work.

Theorem 4. *Assume that $h \in (0, {}^2\!/\!{}_M)$ and f satisfies condition* (2.1). *The following claims hold:*

(a) *If $h \leq {}^2\!/\!{}_{(m+M)}$ then* $W_2(\nu_K, \pi) \leq (1 - mh)^K W_2(\nu_0, \pi) + 1.65(\frac{M}{m})(hp)^{1/2}$.

(b) *If $h \geq {}^2\!/\!{}_{(m+M)}$ then* $W_2(\nu_K, \pi) \leq (Mh - 1)^K W_2(\nu_0, \pi) + \dfrac{1.65Mh}{2 - Mh}(hp)^{1/2}$.

We refer the readers interested in the proof of this theorem either to [Dal17a] or to Section 2.6, where the latter is obtained as a direct consequence of Theorem 7. The factor 1.65 is obtained by upper bounding $7\sqrt{2}/6$.

In practice, a relevant approach to getting an accuracy of at most ϵ is to minimize the upper bound provided by Theorem 4 with respect to h, for a fixed K. Then, one can choose the smallest K for which the obtained upper bound is smaller than ϵ. One useful observation is that the upper bound of case (b) is an increasing function of h. Its minimum is always attained at $h = 2/(m+M)$, which means that one can always look for a step-size in the interval $(0, 2/(m+M)]$ by minimizing the upper bound in (a). This can be done using standard line-search methods such as the bisection algorithm.

Note that if the initial value $\boldsymbol{\vartheta}_0 = \boldsymbol{\theta}_0$ is deterministic then, using the notation $\boldsymbol{\theta}^* = \arg\min_{\theta\in\mathbb{R}^p} f(\boldsymbol{\theta})$, in view of [DM19, Proposition 1], we have

$$W_2(\nu_0,\pi)^2 = \int_{\mathbb{R}^p} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_2^2 \pi(d\boldsymbol{\theta}) \le \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2^2 + p/m. \tag{2.7}$$

Finally, let us remark that if we choose h and K so that

$$h \le {}^{2}/_{(m+M)}, \qquad e^{-mhK} W_2(\nu_0,\pi) \le \varepsilon/2, \quad 1.65(M/m)(hp)^{1/2} \le \varepsilon/2, \tag{2.8}$$

then we have $W_2(\nu_K, \pi) \le \varepsilon$. In other words, conditions (2.8) are sufficient for the density of the output of the LMC algorithm after K iterations to be within the precision ε of the target density when the precision is measured using the Wasserstein distance. This readily yields

$$h \le \frac{m^2\varepsilon^2}{11M^2p} \wedge \frac{2}{m+M} \quad \text{ and } \quad hK \ge \frac{1}{m}\log\Big(\frac{2(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2^2 + p/m)^{1/2}}{\varepsilon}\Big)$$

Assuming m, M and $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2^2/p$ to be constants, we can deduce from the last display that it suffices $K = C(p/\varepsilon^2)\log(p/\varepsilon^2)$ number of iterations in order to reach the precision level ε. This fact has been first established in [Dal17b] for the LMC algorithm with a warm start and the total-variation distance. It was later improved by [DM17, DM19], where the authors showed that the same result holds for any starting point and established similar bounds for the Wasserstein distance. Theorem 4 above can be seen as a user-friendly version of the corresponding result established by [DM19].

Remark 3. *Although (2.7) is relevant for understanding the order of magnitude of $W_2(\nu_0, \pi)$, it has limited applicability since the distance $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|$ might be hard to evaluate. As mentioned in [Dal17a], an attractive alternative to that bound is given by the inequality* [1]

$$\begin{aligned} mW_2(\nu_0,\pi)^2 &\le m\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2^2 + p \\ &\le 2\big(f(\boldsymbol{\theta}_0) - f(\boldsymbol{\theta}^*)\big) + p. \end{aligned}$$

[1]The second line follows from strong convexity whereas the third line is a consequence of the fact that $\boldsymbol{\theta}^*$ is a stationary point of f.

If f is lower bounded by some known constant, for instance if $f \geq 0$, the last inequality provides the computable upper bound $W_2(\nu_0, \pi)^2 \leq \big(2f(\boldsymbol{\theta}_0) + p\big)/m$.

2.2.2 Guarantees under strong convexity for the varying step LMC

The result of previous section provides a guarantee for the constant step LMC. One may wonder if using a variable step sizes $\boldsymbol{h} = \{h_k\}_{k \in \mathbb{N}}$ can improve the convergence. Note that in [DM19, Theorem 5], guarantees for the variable step LMC are established. However, they do not lead to a clear message on the choice of the step-sizes. The next result fills this gap by showing that an appropriate selection of step-sizes improves on the constant step LMC with an improvement factor logarithmic in p/ϵ^2.

Theorem 5. *Let us consider the LMC algorithm with varying step-size h_{k+1} defined by*

$$h_{k+1} = \frac{2}{M + m + (2/3)m(k - K_1)_+}, \qquad k = 1, 2, \ldots \tag{2.9}$$

where K_1 is the smallest non-negative integer satisfying[2]

$$K_1 \geq \frac{\ln\big(W_2(\nu_0, \pi)/\sqrt{p}\big) + \ln(m/M) + (1/2)\ln(M + m)}{\ln(1 + 2m/M-m)}. \tag{2.10}$$

If f satisfies (2.1), then for every $k \geq K_1$, we have

$$W_2(\nu_k, \pi) \leq \frac{3.5M\sqrt{p}}{m\sqrt{M + m + (2/3)m(k - K_1)}}. \tag{2.11}$$

The step size (2.9) has two important advantages as compared to the constant steps. The first advantage is that it is independent of the target precision level ϵ. The second advantage is that we get rid of the logarithmic terms in the number of iterations required to achieve the precision level ϵ. Indeed, it suffices $K = K_1 + (27M^2/2m^3)(p/\varepsilon^2)$ iterations to get the right hand side of (2.11) smaller than ε, where K_1 depends neither on the dimension p nor on the precision level ε.

Since the choice of h_{k+1} in (2.9) might appear mysterious, we provide below a quick explanation of the main computations underpinning this choice. The main step of the proof of upper bounds on $W_2(\nu_k, \pi)$, is the following recursive inequality (see Proposition 11 in Section 2.6)

$$W_2(\nu_{k+1}, \pi) \leq (1 - mh_{k+1})W_2(\nu_k, \pi) + 1.65M\sqrt{p}\, h_{k+1}^{3/2}.$$

[2]Combining the definition of K_1 and the upper bound in (2.7), one easily checks that if $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_\infty$ is bounded, then K_1 is upper bounded by a constant that does not depend on the dimension p.

Using the notation $B_k = \frac{2(m/3)^{3/2}}{1.65M\sqrt{p}} W_2(\nu_k, \pi)$, this inequality can be rewritten as

$$B_{k+1} \leq (1 - mh_{k+1})B_k + 2(mh_{k+1}/3)^{3/2}.$$

Minimizing the right hand side with respect to h_{k+1}, we find that the minimum is attained at the stationary point

$$h_{k+1} = \frac{3}{m} B_k^2. \tag{2.12}$$

With this h_{k+1}, one checks that the sequence B_k satisfies the recursive inequality

$$B_{k+1}^2 \leq B_k^2(1 - B_k^2)^2 \leq \frac{B_k^2}{1 + B_k^2}.$$

The function $g(x) = x/(1+x)$ being increasing in $(0, \infty)$, we get

$$B_{k+1}^2 \leq \frac{B_k^2}{1 + B_k^2} \leq \frac{\frac{B_{k-1}^2}{1+B_{k-1}^2}}{1 + \frac{B_{k-1}^2}{1+B_{k-1}^2}} = \frac{B_{k-1}^2}{1 + 2B_{k-1}^2}.$$

By repetitive application of the same argument, we get

$$B_{k+1}^2 \leq \frac{B_{K_1}^2}{1 + (k + 1 - K_1)B_{K_1}^2}.$$

The integer K_1 was chosen so that $B_{K_1}^2 \leq \frac{2m}{3(M+m)}$, see (2.26). Inserting this upper bound in the right hand side of the last display, we get

$$B_{k+1}^2 \leq \frac{2m}{3(M + m) + 2m(k + 1 - K_1)}.$$

Finally, replacing in (2.12) B_k^2 by its upper bound derived from the last display, we get the suggested value for h_{k+1}.

2.2.3 Extension to mixtures of strongly log-concave densities

We describe here a simple setting in which a suitable version of the LMC algorithm yields efficient sampling algorithm for a target function which is not log-concave. Indeed, let us assume that

$$\pi(\boldsymbol{\theta}) = \int_H \pi_1(\boldsymbol{\theta}|\boldsymbol{\eta})\, \pi_0(d\eta),$$

where H is an arbitrary measurable space, π_0 is a probability distribution on H and $\pi_1(\cdot|\cdot)$ is a Markov kernel on $\mathbb{R}^p \times H$. This means that $\pi_2(d\boldsymbol{\theta}, d\boldsymbol{\eta}) = \pi_1(\boldsymbol{\theta}|\boldsymbol{\eta})\,\pi_0(d\boldsymbol{\eta})d\boldsymbol{\theta}$ defines a probability measure on $\mathbb{R}^p \times H$ of which π is the first marginal.

Theorem 6. *Assume that $\pi_1(\boldsymbol{\theta}|\boldsymbol{\eta}) = \exp\{-f_{\boldsymbol{\eta}}(\boldsymbol{\theta})\}$ so that for every $\boldsymbol{\eta} \in H$, $f_{\boldsymbol{\eta}}$ satisfies assumption (2.1). Define the mixture LMC (MLMC) algorithm as follows: sample $\boldsymbol{\eta} \sim \pi_0$ and choose an initial value $\boldsymbol{\vartheta}_0 \sim \nu_0$, then compute*

$$\boldsymbol{\vartheta}_{k+1}^{\mathrm{MLMC}} = \boldsymbol{\vartheta}_k^{\mathrm{MLMC}} - h_{k+1}\nabla f_{\boldsymbol{\eta}}(\boldsymbol{\vartheta}_k^{\mathrm{MLMC}}) + \sqrt{2h_{k+1}}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots$$

where h_k is defined by (2.9) and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k, \ldots$ is a sequence of mutually independent, and independent of $(\boldsymbol{\eta}, \boldsymbol{\vartheta}_0)$, centered Gaussian vectors with covariance matrices equal to identity. It holds that, for every positive integer $k \geq K_1$ (see eq. (2.10) for the definition of K_1),

$$W_2(\nu_k, \pi) \leq \frac{3.5M\sqrt{p}}{m\sqrt{M + m + (2/3)m(k - K_1)}}.$$

This result extends the applicability of Langevin based techniques to a wider framework than the one of strongly log-concave distributions. The proof, postponed to Section 2.6, is a straightforward consequence of Theorem 5.

2.3 Guarantees for the inaccurate gradient version

In some situations, the precise evaluation of the gradient $\nabla f(\boldsymbol{\theta})$ is computationally expensive or practically impossible, but it is possible to obtain noisy evaluations of ∇f at any point. This is the setting considered in the present section. More precisely, we assume that at any point $\boldsymbol{\vartheta}_{k,h} \in \mathbb{R}^p$ of the LMC algorithm, we can observe the value

$$\boldsymbol{Y}_{k,h} = \nabla f(\boldsymbol{\vartheta}_{k,h}) + \boldsymbol{\zeta}_k,$$

where $\{\boldsymbol{\zeta}_k : k = 0, 1, \ldots\}$ is a sequence of random (noise) vectors. The noisy LMC (nLMC) algorithm is defined as

$$\boldsymbol{\vartheta}_{k+1,h} = \boldsymbol{\vartheta}_{k,h} - h\boldsymbol{Y}_{k,h} + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{2.13}$$

where $h > 0$ and $\boldsymbol{\xi}_{k+1}$ are as in (2.2). The noise $\{\boldsymbol{\zeta}_k : k = 0, 1, \ldots\}$ is assumed to satisfy the following condition.

Condition N: for some $\delta > 0$ and $\sigma > 0$ and for every $k \in \mathbb{N}$,

- (bounded bias) $\mathbf{E}\big[\big\|\mathbf{E}(\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,h})\big\|_2^2\big] \leq \delta^2 p$,

- (bounded variance) $\mathbf{E}[\|\boldsymbol{\zeta}_k - \mathbf{E}(\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,h})\|_2^2] \le \sigma^2 p$,
- (independence of updates) $\boldsymbol{\xi}_{k+1}$ in (2.13) is independent of $(\boldsymbol{\zeta}_0, \ldots, \boldsymbol{\zeta}_k)$.

We emphasize right away that the random vectors $\boldsymbol{\zeta}_k$ are not assumed to be independent, as opposed to what is done in [Dal17a]. The next theorem extends the guarantees of Theorem 4 to the inaccurate-gradient setting and to the nLMC algorithm.

Theorem 7. *Let $\boldsymbol{\vartheta}_{K,h}$ be the K-th iterate of the nLMC algorithm* (2.13) *and ν_K be its distribution. If the function f satisfies condition* (2.1) *and $h \le {}^{2}/_{(m+M)}$ then*

$$W_2(\nu_K, \pi) \le (1 - mh)^K W_2(\nu_0, \pi) + 1.65(M/m)(hp)^{1/2} \\ + \frac{\delta\sqrt{p}}{m} + \frac{\sigma^2(hp)^{1/2}}{1.65M + \sigma\sqrt{m}} . \tag{2.14}$$

To the best of our knowledge, the first result providing guarantees for sampling from a distribution in the scenario when precise evaluations of the log-density or its gradient are not available has been established in [Dal17a]. Prior to that work, some asymptotic results has been established in [AFEB16]. The closely related problem of computing an average value with respect to a distribution, when the gradient of its log-density is known up to an additive noise, has been studied by [TTV16, VZ15, NDH$^+$17, CDC15]. Note that these settings are of the same flavor as those of stochastic approximation, an active area of research in optimization and machine learning.

As compared to the analogous result in [Dal17a], Theorem 7 above has several advantages. First, it extends the applicability of the result to the case of a biased noise. In other words, it allows for $\boldsymbol{\zeta}_k$ with nonzero means. Second, it considerably relaxes the independence assumption on the sequence $\{\boldsymbol{\zeta}_k\}$, by replacing it by the independence of the updates. Third, and perhaps the most important advantage of Theorem 7 is the improved dependence of the upper bound on σ. Indeed, while the last term in the upper bound in Theorem 7 is $O(\sigma^2)$, when $\sigma \to 0$, the corresponding term in [Dal17a, Th. 3] is only $O(\sigma)$.

To understand the potential scope of applicability of Theorem 7, let us consider a generic example in which $f(\boldsymbol{\theta})$ is the average of n functions defined through independent random variables $X_1, \ldots, X_n$:

$$f(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, X_i).$$

When the gradient of $\ell(\boldsymbol{\theta}, X_i)$ with respect to parameter $\boldsymbol{\theta}$ is hard to compute, one can replace the evaluation of $\nabla f(\boldsymbol{\vartheta}_{k,h})$ at each step k by that of $Y_k = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\vartheta}_{k,h}, X_{N_k})$, where

N_k is a random variable uniformly distributed in $\{1, \ldots, n\}$ and independent of $\boldsymbol{\vartheta}_{k,h}$. Under suitable assumptions, this random vector satisfies the conditions of Theorem 7 with $\delta = 0$ and constant σ^2. Therefore, if we analyze the upper bound provided by (2.14), we see that the last term, due to the subsampling, is of the same order of magnitude as the second term. Thus, using the subsampled gradient in the LMC algorithm does not cause a significant deterioration of the precision while reducing considerably the computational burden.

Note that Theorem 7 allows to handle situations in which the approximations of the gradient are biased. This bias is controlled by the parameter δ. Such a bias can appear when using deterministic approximations of integrals or differentials. For instance, in statistical models with latent variables, the gradient of the log-likelihood has often an integral form. Such integrals can be approximated using quadrature rules, yielding a bias term, or Monte Carlo methods, yielding a variance term.

In the preliminary version [Dal17a] of this work, we made a mistake by claiming that the stochastic gradient version of the LMC, introduced in [WT11] and often referred to as Stochastic Gradient Langevin Dynamics (SGLD), has an error of the same order as the non-stochastic version of it. This claim is wrong, since when $f(\boldsymbol{\theta}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, X_i)$ with a strongly convex function $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{\theta}, x)$ and iid variables $X_1, \ldots, X_n$, we have m and M proportional to n. Therefore, choosing $Y_k = n\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\vartheta}_{k,h}, X_{N_k})$ as a noisy version of the gradient (where N_k is a uniformly over $\{1, \ldots, n\}$ distributed random variable independent of $\boldsymbol{\vartheta}_{k,h}$), we get $\delta = 0$ but σ^2 proportional to n^2. Therefore, the last term in (2.14) is of order $(nhp)^{1/2}$ and dominates the other terms. Furthermore, replacing Y_k by $Y_k = \frac{n}{s}\sum_{j=1}^s \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\vartheta}_{k,h}, X_{N_k^j})$ with iid variables $N_k^1, \ldots, N_k^s$ does not help, since then σ^2 is of order n^2/s and the last term in (2.14) is of order $(nhp/s)^{1/2}$, which is still larger than the term $(M/m)(hp)^{1/2}$. This discussion shows that Theorem 7 applied to SGLD is of limited interest. For a more in-depth analysis of the SGLD, we refer the reader to [NDH+17, RRT17, XCZG18].

It is also worth mentioning here that another example of approximate gradient—based on a quadratic approximation of the log-likelihood of the generalized linear model—has been considered in [HZ17, Section 5]. It corresponds, in terms of condition N, to a situation in which the variance σ^2 vanishes but the bias δ is non-zero. An important ingredient of the proof of Theorem 7 is the following simple result, which can be useful in other contexts as well (for a proof, see Lemma 7 in Section 2.G below).

Lemma 1. *Let A, B and C be given non-negative numbers such that $A \in (0, 1)$. Assume that the sequence of non-negative numbers $\{x_k\}_{k=0,1,2,\ldots}$ satisfies the recursive inequality*

$$x_{k+1}^2 \leq [(1 - A)x_k + C]^2 + B^2$$

for every integer $k \geq 0$. Then, for all integers $k \geq 0$,

$$x_k \leq (1-A)^k x_0 + \frac{C}{A} + \frac{B^2}{C + \sqrt{A}\,B}.$$

Thanks to this lemma, the upper bound on the Wasserstein distance provided by (2.14) is sharper than the one proposed in [Dal17a].

2.4 Guarantees under additional smoothness

When the function f has Lipschitz continuous Hessian, one can get improved rates of convergence. This has been noted by [Dal17b], where the author proposed to use a modified version of the LMC algorithm, the LMC with Ozaki discretization, in order to take advantage of the smoothness of the Hessian. On the other hand, it has been proved in [AJKH14, AJKH15] that the boundedness of the third order derivative of f (equivalent to the boundedness of the second-order derivative of the drift of the Langevin diffusion) implies that the Wasserstein distance between the marginals of the Langevin diffusion and its Euler discretization are of order $h\sqrt{\log(1/h)}$. Note however, that in [AJKH15] there is no evaluation of the impact of the dimension on the quality of the Euler approximation. This evaluation has been done by [DM19] by showing that the Wasserstein error of the Euler approximation is of order hp. This raises the following important question: is it possible to get advantage of the Lipschitz continuity of the Hessian of f in order to improve the guarantees on the quality of sampling by the standard LMC algorithm. The answer of this question is affirmative and is stated in the next theorem.

In what follows, for any matrix $\mathbf{M}$, we denote by $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_F$, respectively, the spectral norm and the Frobenius norm of $\mathbf{M}$. We write $\mathbf{M} \preceq \mathbf{M}'$ or $\mathbf{M}' \succeq \mathbf{M}'$ to indicate that the matrix $\mathbf{M}' - \mathbf{M}$ is positive semi-definite.

Condition F: the function f is twice differentiable and for some positive numbers m, M and M_2,

- (strong convexity) $\nabla^2 f(\boldsymbol{\theta}) \succeq m\mathbf{I}_p$, for every $\boldsymbol{\theta} \in \mathbb{R}^p$,
- (bounded second derivative) $\nabla^2 f(\boldsymbol{\theta}) \preceq M\mathbf{I}_p$, for every $\boldsymbol{\theta} \in \mathbb{R}^p$,
- (further smoothness) $\|\nabla^2 f(\boldsymbol{\theta}) - \nabla^2 f(\boldsymbol{\theta}')\| \leq M_2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$, for every $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$.

Theorem 8. *Let $\boldsymbol{\vartheta}_{K,h}$ be the K-th iterate of the nLMC algorithm (2.13) and ν_K be its distribution. Assume that conditions* **F** *and* **N** *are satisfied. Then, for every $h \leq 2/(m+M)$, we* *have*

$$W_2(\nu_K,\pi) \le (1-mh)^K W_2(\nu_0,\pi) + \frac{M_2hp}{2m} + \frac{11Mh\sqrt{Mp}}{5m} + \frac{\delta\sqrt{p}}{m} + \frac{2\sigma^2\sqrt{hp}}{M_2\sqrt{hp}+2\sigma\sqrt{m}}.$$

In the last inequality, $11/5$ is an upper bound for $0.5 + 2\sqrt{2/3} \approx 2.133$.

When applying the nLMC algorithm to sample from a target density, the user may usually specify four parameters: the step-size h, the number of iterations K, the tolerated precision δ of the deterministic approximation and the precision σ of the stochastic approximation. An attractive feature of Theorem 8 is that the contributions of these four parameters are well separated, especially if we upper bound the last term by $2\sigma^2/M_2$. As a consequence, in order to have an error of order ε in Wasserstein distance, we might choose: σ at most of order $\sqrt{\varepsilon}$, δ at most of order $m\varepsilon/\sqrt{p}$, h of order ε/p and K of order $(p/m\varepsilon)\log(p/\varepsilon)$. Akin to Theorem 5, one can use variable step-sizes to avoid the logarithmic factor; we leave these computations to the reader.

Note that if we instantiate Theorem 8 to the case of accurate gradient evaluations, that is when $\sigma = \delta = 0$, we recover the constant step-size version of [DM19, Theorem 8], with optimized constants. Indeed, for contant step-size, [DM19, Theorem 8] yields

$$W_2(\nu_K,\pi) \le \Big\{2(1-\bar{m}\,h)^K W_2(\nu_0,\pi)^2 + 2ph^2\Big(\frac{M^2}{\bar{m}} + \frac{M^4}{3m\bar{m}^2} + \frac{M_2^2 p}{3\bar{m}^2} + O(h)\Big)\Big\}^{1/2}, \tag{2.15}$$

where $\bar{m} = \frac{mM}{m+M} \in [m/2, m)$ and the term $O(h)$ can be given explicitly. A visual comparison of the optimal number of iterations obtained from this bound to that obtained from Theorem 8 (with $\delta = \sigma = 0$) is provided in Figure 2.2.

Under the assumption of Lipschitz continuity of the Hessian of f, one may wonder whether second-order methods that make use of the Hessian in addition to the gradient are able to outperform the standard LMC algorithm. The most relevant candidate algorithms for this are the LMC with Ozaki discretization (LMCO) and a variant of it, LMCO′, a slightly modified version of an algorithm introduced in [Dal17b]. The LMCO is a recursive algorithm the update rule of which is defined as follows: for every $k \ge 0$, we set $\mathbf{H}_k = \nabla^2 f(\boldsymbol{\vartheta}_{k,h}^{\rm LMCO})$, which is an invertible $p \times p$ matrix since f is strongly convex, and define

$$\begin{aligned} &\mathbf{M}_k = \big(\mathbf{I}_p - e^{-h\mathbf{H}_k}\big)\mathbf{H}_k^{-1}, \qquad \boldsymbol{\Sigma}_k = \big(\mathbf{I}_p - e^{-2h\mathbf{H}_k}\big)\mathbf{H}_k^{-1},\\ &\boldsymbol{\vartheta}_{k+1,h}^{\rm LMCO} = \boldsymbol{\vartheta}_{k,h}^{\rm LMCO} - \mathbf{M}_k\nabla f\big(\boldsymbol{\vartheta}_{k,h}^{\rm LMCO}\big) + \boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\xi}_{k+1}, \end{aligned} \tag{2.16}$$

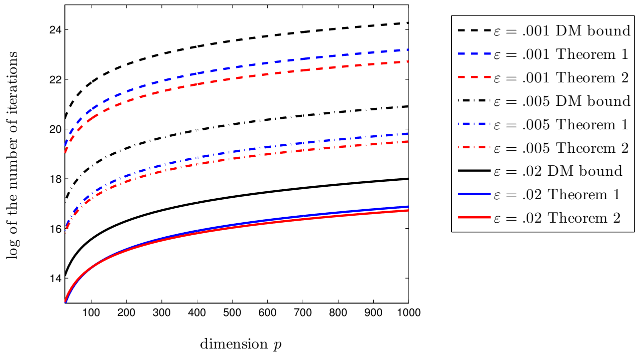


Figure 2.2: Plots showing the logarithm of the number of iterations as function of dimension p for several values of ε. The plotted values are derived from Theorem 8 and (2.15) (referred to as DM bound) using the data $m = 10$, $M = 50$, $M_2 = 1$, $W_2^2(\nu_0, \pi) = p + (p/m)$, $\delta = \sigma = 0$.

where $\{\boldsymbol{\xi}_k : k \in \mathbb{N}\}$ is a sequence of independent random vectors distributed according to the $\mathcal{N}_p(0, \mathbf{I}_p)$ distribution. The LMCO$'$ algorithm is based on approximating the matrix exponentials by linear functions, more precisely, for $\mathbf{H}'_k = \nabla^2 f(\boldsymbol{\vartheta}_{k,h}^{\text{LMCO}'})$,

$$\boldsymbol{\vartheta}_{k+1,h}^{\text{LMCO}'} = \boldsymbol{\vartheta}_{k,h}^{\text{LMCO}'} - h\Big(\mathbf{I}_p - \frac{1}{2}h\mathbf{H}'_k\Big)\nabla f\big(\boldsymbol{\vartheta}_{k,h}^{\text{LMCO}'}\big) + \sqrt{2h}\Big(\mathbf{I}_p - h\mathbf{H}'_k + \frac{1}{3}h^2(\mathbf{H}'_k)^2\Big)^{1/2}\boldsymbol{\xi}_{k+1}. \tag{2.17}$$

Let us mention right away that the stochastic perturbation present in the last display can be computed in practice without taking the matrix square-root. Indeed, it suffices to generate two independent standard Gaussian vectors $\boldsymbol{\eta}_{k+1}$ and $\boldsymbol{\eta}'_{k+1}$; then the random vector

$$\big(\mathbf{I}_p - (1/2)h\mathbf{H}'_k\big)\boldsymbol{\eta}_{k+1} + (\sqrt{3}/6)\, h\, \mathbf{H}'_k\boldsymbol{\eta}'_{k+1}$$

has exactly the same distribution as $\big(\mathbf{I}_p - h\mathbf{H}'_k + (1/3)h^2(\mathbf{H}'_k)^2\big)^{1/2}\boldsymbol{\xi}_{k+1}$.

In the rest of this section, we provide guarantees for methods LMCO and LMCO$'$. Note that we consider only the case where the gradient and the Hessian of f are computed exactly, that is without any approximation.

Theorem 9. *Let ν_K^{LMCO} and $\nu_K^{\text{LMCO}'}$ be, respectively, the distributions of the K-th iterate of the LMCO algorithm (2.16) and the LMCO$'$ algorithm (2.17) with an initial distribution ν_0. Assume that conditions* **F** *and* **N** *are satisfied. Then, for every $h \leq m/M^2$,*

$$W_2(\nu_K^{\text{LMCO}}, \pi) \leq (1 - 0.25mh)^K W_2(\nu_0, \pi) + \frac{11.5M_2h(p+1)}{m}. \tag{2.18}$$

If, in addition, $h \le {}^{3m}/_{4M^2}$*, then*

$$W_2(\nu_K^{\mathrm{LMCO}'}, \pi) \le (1 - 0.25mh)^K W_2(\nu_0, \pi) + \frac{1.3M^2h^2\sqrt{Mp}}{m} + \frac{7.3M_2h(p+1)}{m}. \qquad (2.19)$$

A very rough consequence of this theorem is that one has similar theoretical guarantees for the LMCO and the LMCO′ algorithms, since in most situations the middle term in the right hand side of (2.19) is smaller than the last term. On the other hand, the per-iteration cost of the modified algorithm LMCO′ is significantly smaller than the per-iteration cost of the original LMCO. Indeed, for the LMCO′ there is no need to compute matrix exponentials neither to invert matrices, one only needs to perform matrix-vector multiplication for $p \times p$ matrices. Note that for many matrices such a multiplication operation might be very cheap using the fast Fourier transform or other similar techniques. In addition, the computational complexity of the Hessian-vector product is provably of the same order as that of evaluating the gradient, see [Gri93]. Therefore, one iteration of the LMCO′ algorithm is not more costly than one iteration of the LMC. At the same time, the error bound (2.19) for the LMCO′ is smaller than the one for the LMC provided by Theorem 8. Indeed, the term $Mh\sqrt{Mp}$ present in the bound of Theorem 8 is generally of larger order than the term $(Mh)^2\sqrt{Mp}$ appearing in (2.19).

2.5 Relation with optimization

We have already mentioned that the LMC algorithm is very close to the gradient descent algorithm for computing the minimum $\boldsymbol{\theta}^*$ of the function f. However, when we compare the guarantees of Theorem 4 with those available for the optimization problem, we remark the following striking difference. The approximate computation of $\boldsymbol{\theta}^*$ requires a number of steps of the order of $\log(1/\varepsilon)$ to reach the precision ε, whereas, for reaching the same precision in sampling from π, the LMC algorithm needs a number of iterations proportional to $(p/\varepsilon^2)\log(p/\varepsilon)$. The goal of this section is to explain that this, at first sight disappointing behavior of the LMC algorithm is, in fact, consistent with the exponential convergence of the gradient descent. Furthermore, the latter is obtained from the guarantees on the LMC by letting a temperature parameter go to zero.

The main ingredient for the explanation is that the function $f(\boldsymbol{\theta})$ and the function $f_\tau(\boldsymbol{\theta}) = f(\boldsymbol{\theta})/\tau$ have the same point of minimum $\boldsymbol{\theta}^*$, whatever the real number $\tau > 0$. In addition, if we define the density function $\pi_\tau(\boldsymbol{\theta}) \propto \exp\big(-f_\tau(\boldsymbol{\theta})\big)$, then the average value

$$\bar{\boldsymbol{\theta}}_\tau = \int_{\mathbb{R}^p} \boldsymbol{\theta}\, \pi_\tau(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

tends to the minimum point $\boldsymbol{\theta}^*$ when τ goes to zero. Furthermore, the distribution $\pi_\tau(d\boldsymbol{\theta})$ tends to the Dirac measure at $\boldsymbol{\theta}^*$. Clearly, f_τ satisfies (2.1) with the constants $m_\tau = m/\tau$ and $M_\tau = M/\tau$. Therefore, on the one hand, we can apply to π_τ claim (a) of Theorem 4, which tells us that if we choose $h = 1/M_\tau = \tau/M$, then

$$W_2(\nu_K, \pi_\tau) \le \Big(1 - \frac{m}{M}\Big)^K W_2(\delta_{\boldsymbol{\theta}_0}, \pi_\tau) + 1.65\Big(\frac{M}{m}\Big)\Big(\frac{p\tau}{M}\Big)^{1/2}. \tag{2.20}$$

On the other hand, the LMC algorithm with the step-size $h = \tau/M$ applied to f_τ reads as

$$\boldsymbol{\vartheta}_{k+1,h} = \boldsymbol{\vartheta}_{k,h} - \frac{1}{M}\nabla f(\boldsymbol{\vartheta}_{k,h}) + \sqrt{\frac{2\tau}{M}}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{2.21}$$

When the parameter τ goes to zero, the LMC sequence (2.21) tends to the gradient descent sequence $\boldsymbol{\theta}_k$. Therefore, the limiting case of (2.20) corresponding to $\tau \to 0$ writes as

$$\|\boldsymbol{\theta}^{(K)} - \boldsymbol{\theta}^*\|_2 \le \Big(1 - \frac{m}{M}\Big)^K \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2,$$

which is a well-known result in Optimization. This clearly shows that Theorem 4 is a natural extension of the results of convergence from optimization to sampling.

Such an analogy holds true for the Newton method as well. Its counterpart in sampling is the LMCO algorithm. Indeed, one easily checks that if f is replaced by f_τ with τ going to zero, then, for any fixed step-size h, the matrix $\boldsymbol{\Sigma}_k$ in (2.16) tends to zero. This implies that the stochastic perturbation vanishes. On the other hand, the term $\mathbf{M}_{k,\tau}\nabla f_\tau(\boldsymbol{\vartheta}_{k,h}^{\rm LMCO})$ tends to $\{\nabla^2 f(\boldsymbol{\vartheta}_{k,h}^{\rm LMCO})\}^{-1}\nabla f(\boldsymbol{\vartheta}_{k,h}^{\rm LMCO})$, as $\tau \to 0$. Thus, the updates of the Newton algorithm can be seen as the limit case, when τ goes to zero, of the updates of the LMCO.

However, if we replace f by f_τ in the upper bounds stated in Theorem 9 and we let τ go to zero, we do not retrieve the well-known guarantees for the Newton method. The main reason is that Theorem 9 describes the behavior of the LMCO algorithm in the regime of small step-sizes h, whereas Newton's method corresponds to (a limit case of) the LMCO with a fixed h. Using arguments similar to those employed in the proof of Theorem 9, one can establish the following result, the proof of which is postponed to Section 2.6.

Proposition 10. *Let $\nu_K^{\rm LMCO}$ be the distributions of the K-th iterate of the LMCO algorithm (2.16) with an initial distribution ν_0. Assume that condition* **F** *is satisfied. Then, for every $h > 0$ and $K \in \mathbb{N}$,*

$$W_2(\nu_K^{\rm LMCO}, \pi) \le \frac{2m}{M_2}\big(w_K \exp(v_K w_K^{-2^K})\big)^{2^K} \tag{2.22}$$

with

$$w_K = \frac{M_2 W_{2^{K+1}}(\nu_0,\pi)}{2m} + \frac{1}{2}e^{-mh}, \text{ and } v_K = \frac{2M_2 M^{3/2}\sqrt{2p+2^K}}{m^3} + e^{-mh}.$$

If we replace in the right hand side of (2.22) the quantities m, M and M_2, respectively, by $m_\tau = m/\tau$, $M_\tau = M/\tau$ and $M_{2,\tau} = M_2/\tau$, and we let τ go to zero, then it is clear that the term v_K vanishes. On the other hand, if ν_0 is the Dirac mass at some point $\boldsymbol{\theta}_0$, then w_K converges to $M_2\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2/(2m)$. Therefore, for Newton's algorithm as a limiting case of (2.22) we get

$$\|\boldsymbol{\theta}_K^{\text{Newton}} - \boldsymbol{\theta}^*\|_2 \le \frac{2m}{M_2}\bigg(\frac{M_2\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2}{2m}\bigg)^{2^K}.$$

The latter provides the so called quadratic rate of convergence, which is a well-known result that can be found in many textbooks; see, for instance, [CZ13, Theorem 9.1].

A particularly promising remark made in Section 2.2.3 is that all the results established for the problem of approximate sampling from a log-concave distribution can be carried over the distributions that can be written as a mixture of (strongly) log-concave distributions. The only required condition is to be able to sample from the mixing distribution. This provides a well identified class of (posterior) distributions for which the problem of finding the mode is difficult (because of nonconvexity) whereas the sampling problem can be solved efficiently.

There are certainly other interesting connections to uncover between sampling and optimization. In particular, in [MCJ+19], it was shown that in the case of mixture distributions, sampling algorithms scale linearly with the model dimension, as opposed to those of optimization, which have exponential scaling. One can think of lower bounds for sampling or finding a sampling counterpart of Nesterov acceleration. Some recent advances on the gradient flow [WWJ16] might be useful for achieving these goals.

2.6 Conclusion

We have presented easy-to-use finite-sample guarantees for sampling from a strongly log-concave density using the Langevin Monte-Carlo algorithm with a fixed step-size and extended it to the case where the gradient of the log-density can be evaluated up to some error term. Our results cover both deterministic and random error terms. We have also demonstrated that if the log-density f has a Lipschitz continuous second-order derivative, then one can choose a larger step-size and obtain improved convergence rate.

We have also uncovered some analogies between sampling and optimization. The underlying principle is that an optimization algorithm may be seen as a limit case of a sampling algorithm. Therefore, the results characterizing the convergence of the optimization schemes should have their counterparts for sampling strategies. We have described these analogues for the steepest gradient descent and for the Newton algorithm. However, while in the optimization the relevant characteristics of the problem are the dimension p, the desired accuracy ε and the condition number M/m, the problem sampling involves an additional characteristic which is the scale given by the strong-convexity constant m. Indeed, if we increase m by keeping the condition number M/m constant, the number of iterations for the LMC to reach the precision ϵ will decrease. In this respect, we have shown that the LMC with Ozaki discretization, termed LMCO, has a better dependence on the overall scale of f than the original LMC algorithm. However, the weakness of the LMCO is the high computational cost of each iteration. Therefore, we have proposed a new algorithm, LMCO$'$, that improves the LMC in terms of its dependence on the scale and each iteration of LMCO$'$ is computationally much cheaper than each iteration of the LMCO.

Another interesting finding is that, in the case of accurate gradient evaluations (*i.e.*, when there is no error in the gradient computation), a suitably chosen variable step-size leads to logarithmic improvement in the convergence rate of the LMC algorithm.

Interesting directions for future research are establishing lower bounds in the spirit of those existing in optimization, obtaining user-friendly guarantees for computing the posterior mean or for sampling from a non-smooth density. Some of these problems have already been tackled in several papers mentioned in previous sections, but we believe that the techniques developed in the present work might be helpful for revisiting and deepening the existing results.

Appendix to Chapter 2

The basis of the proofs of all the theorems stated in previous sections is a recursive inequality that upper bounds the error at the step $k+1$, $W_2(\nu_{k+1},\pi)$, by an expression involving the error of the previous step, $W_2(\nu_k,\pi)$. To this end, we use the fact that for a suitably chosen Langevin diffusion, $\boldsymbol{L}$, in stationary regime, we have $W_2(\nu_k,\pi)^2 = \mathbf{E}[\|\vartheta_k - \boldsymbol{L}_{kh}\|_2^2]$ and $W_2(\nu_{k+1},\pi)^2 \le \mathbf{E}[\|\vartheta_{k+1} - \boldsymbol{L}_{(k+1)h}\|_2^2]$. The goal is then to upper bound the latter by an expression that involves the former and some suitably controlled remainder terms. This leads to a recursive inequality and the last step of the proof is to unfold the recursion. Since different chains $\vartheta_{k,h}$ are considered in this paper, we get different recursive inequalities. Lemma 7 and Lemma 8 are the new technical tools that are used for solving the encountered recursive inequalities. The remainder terms appearing in the recursive inequalities are evaluated by using stochastic calculus and the smoothness properties of f. The main building blocks for these evaluations are Lemma 3, Lemma 4 and Lemma 6, the latter being used only in the results assuming the Hessian-Lipschitz condition.

We will also make repeated use of the Minkowski inequality and its integral version

$$\left\{\mathbf{E}\left[\left(\int_a^b X_t\,dt\right)^p\right]\right\}^{1/p} \le \int_a^b \left\{\mathbf{E}\big[|X_t|^p\big]\right\}^{1/p} dt, \qquad \forall p \in \mathbb{N}^*, \tag{2.23}$$

where X is a random process almost all paths of which are integrable over the interval $[a,b]$. Furthermore, for any random vector $\boldsymbol{X}$, we define the norm $\|\boldsymbol{X}\|_{L_2} = (\mathbf{E}[\|\boldsymbol{X}\|_2^2])^{1/2}$.

The next result is the central ingredient of the proofs of Theorems 4, 5 and 7. Readers interested only in the proof of Theorems 4 and 5, are invited—in the next proof—to consider the random vectors $\boldsymbol{\zeta}_k$ as equal to $\mathbf{0}$ and $\boldsymbol{Y}_{k,\boldsymbol{h}}$ as equal to $\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}})$. This implies, in particular, that $\sigma = \delta = 0$.

Proposition 11. *Let us introduce $\varrho_{k+1} = \max(1 - mh_{k+1}, Mh_{k+1} - 1)$ (since $h \in (0, {}^2\!/\!{}_M)$, this value ϱ satisfies $0 < \varrho < 1$). If f satisfies (2.1) and $h_{k+1} \le 2/M$, then*

$$W_2(\nu_{k+1},\pi)^2 \le \left\{\varrho_{k+1} W_2(\nu_k,\pi) + \alpha M (h_{k+1}^3 p)^{1/2} + h_{k+1}\delta\sqrt{p}\right\}^2 + \sigma^2 h_{k+1}^2 p,$$

with $\alpha = 7\sqrt{2}/6 \le 1.65$.

Proof. To simplify notation, and since there is no risk of confusion, we will write h instead of h_{k+1}. The main steps of the proof are the following. We use a synchronous coupling for approximating the distribution of the LMC sequence by that of a continuous-time Langevin diffusion. We then take advantage of the strong convexity of f for showing that, for h small enough, the error at round $k+1$ is upper bounded, up to a additive remainder term, by the error at round k multiplied by a factor strictly smaller than one, see Lemma 2. The smoothness of the gradient of f ensures that the aforementioned remainder term is small, see Lemma 3 and Lemma 4 below.

Let $\boldsymbol{L}_0$ be a random vector drawn from π such that $W_2(\nu_k, \pi) = \|\boldsymbol{L}_0 - \boldsymbol{\vartheta}_{k,\boldsymbol{h}}\|_{L_2}$ and $\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}, \boldsymbol{L}_0] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]$. Let $\boldsymbol{W}$ be a p-dimensional Brownian Motion independent of $(\boldsymbol{\vartheta}_{k,\boldsymbol{h}}, \boldsymbol{L}_0, \boldsymbol{\zeta}_k)$, such that $\boldsymbol{W}_h = \sqrt{h}\,\boldsymbol{\xi}_{k+1}$. We define the stochastic process $\boldsymbol{L}$ so that

$$\boldsymbol{L}_t = \boldsymbol{L}_0 - \int_0^t \nabla f(\boldsymbol{L}_s)\,ds + \sqrt{2}\,\boldsymbol{W}_t, \qquad \forall\, t > 0. \tag{2.24}$$

It is clear that this equation implies that

$$\begin{aligned}
\boldsymbol{L}_h &= \boldsymbol{L}_0 - \int_0^h \nabla f(\boldsymbol{L}_s)\,ds + \sqrt{2}\,\boldsymbol{W}_h \\
&= \boldsymbol{L}_0 - \int_0^h \nabla f(\boldsymbol{L}_s)\,ds + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}.
\end{aligned}$$

Furthermore, $\{\boldsymbol{L}_t : t \ge 0\}$ is a diffusion process having π as the stationary distribution. Since the initial value $\boldsymbol{L}_0$ is drawn from π, we have $\boldsymbol{L}_t \sim \pi$ for every $t \ge 0$.

Let us denote $\boldsymbol{\Delta}_k = \boldsymbol{L}_0 - \boldsymbol{\vartheta}_{k,\boldsymbol{h}}$ and $\boldsymbol{\Delta}_{k+1} = \boldsymbol{L}_h - \boldsymbol{\vartheta}_{k+1,\boldsymbol{h}}$. We have

$$\begin{aligned}
\boldsymbol{\Delta}_{k+1} &= \boldsymbol{\Delta}_k + h\boldsymbol{Y}_{k,\boldsymbol{h}} - \int_0^h \nabla f(\boldsymbol{L}_t)\,dt \\
&= \boldsymbol{\Delta}_k - h\big(\underbrace{\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}} + \boldsymbol{\Delta}_k) - \nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}})}_{:=\boldsymbol{U}}\big) + h\boldsymbol{\zeta}_k \\
&\qquad \underbrace{- \int_0^h \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0)\big)\,dt}_{:=\boldsymbol{V}}\,.
\end{aligned} \tag{2.25}$$

Using the equalities $\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\Delta}_k, \boldsymbol{U}, \boldsymbol{V}] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}, \boldsymbol{L}_0, \boldsymbol{W}] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}, \boldsymbol{L}_0] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]$, we get

$$\begin{aligned}
\|\boldsymbol{\Delta}_{k+1}\|_{L_2}^2 &= \big\|\boldsymbol{\Delta}_k - h\boldsymbol{U} - \boldsymbol{V} + h\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]\big\|_{L_2}^2 + h^2\big\|\boldsymbol{\zeta}_k - \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]\big\|_{L_2}^2 \\
&\le \big\{\|\boldsymbol{\Delta}_k - h\boldsymbol{U}\|_{L_2} + h\delta\sqrt{p} + \|\boldsymbol{V}\|_{L_2}\big\}^2 + \sigma^2 h^2 p.
\end{aligned}$$

We need now three technical lemmas. Lemma 2 and Lemma 3 are borrowed from [Dal17a], whereas Lemma 4 is an improved version of [Dal17a, Lemma 3]. For the sake of self-containedness, we provide proofs of these lemmas in Section 2.G.

Lemma 2. *Let f be m-strongly convex and the gradient of f be Lipschitz with constant M. If $h < 2/M$, then the mapping $(\mathbf{I}_p - h\nabla f)$ is a contraction in the sense that*

$$\big\|\boldsymbol{x} - \boldsymbol{y} - h\big(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\big)\big\|_2 \le \big\{(1 - mh) \vee (Mh - 1)\big\}\|\boldsymbol{x} - \boldsymbol{y}\|_2,$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$. In particular, using notations in (2.25), it holds that $\|\boldsymbol{\Delta}_k - h\boldsymbol{U}\|_2 \le \varrho\|\boldsymbol{\Delta}_k\|_2$.

Lemma 3. *If the function f is continuously differentiable and the gradient of f is Lipschitz with constant M, then $\int_{\mathbb{R}^p} \|\nabla f(\boldsymbol{x})\|_2^2\, \pi(\boldsymbol{x})\, d\boldsymbol{x} \le Mp$.*

Lemma 4. *If the function f and its gradient is Lipschitz with constant M, $\boldsymbol{L}$ is the Langevin diffusion (2.24) and $\boldsymbol{V}(a) = \int_a^{a+h} \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_a)\big)\, dt$ for some $a \ge 0$, then*

$$\|\boldsymbol{V}(a)\|_{L_2} \le \frac{1}{2}\big(h^4M^3p\big)^{1/2} + \frac{2}{3}(2h^3p)^{1/2}M.$$

Using Lemma 2 and Lemma 4 above, as well as the inequality $W_2(\nu_{k+1}, \pi)^2 \le \mathbf{E}[\|\boldsymbol{\Delta}_{k+1}\|_2^2]$, we get the recursion

$$\begin{aligned}
W_2(\nu_{k+1}, \pi)^2 &\le \big\{\varrho W_2(\nu_k, \pi) + (1/2)\big(h^4M^3p\big)^{1/2} + (2/3)(2h^3p)^{1/2}M + h\delta\sqrt{p}\big\}^2 + \sigma^2h^2p\\
&\overset{(a)}{\le} \big\{\varrho W_2(\nu_k, \pi) + (1/2)\big(2h^3M^2p\big)^{1/2} + (2/3)(2h^3p)^{1/2}M + h\delta\sqrt{p}\big\}^2 + \sigma^2h^2p\\
&\overset{(b)}{\le} \big\{\varrho W_2(\nu_k, \pi) + \alpha M\big(h^3p\big)^{1/2} + h\delta\sqrt{p}\big\}^2 + \sigma^2h^2p,
\end{aligned}$$

where in (a) we have used the condition $h \le 2/M$ whereas in (b) we have put $\alpha = 7\sqrt{2}/6 \le 1.65$. □

2.A Proof of Theorem 4

Using Proposition 11 with $\sigma = \delta = 0$, we get $W_2(\nu_{k+1}, \pi) \le \varrho\, W_2(\nu_k, \pi) + \|\boldsymbol{V}\|_{L_2}$ for all $k \in \mathbb{N}$. In view of Lemma 4, this yields

$$W_2(\nu_{k+1}, \pi) \le \varrho\, W_2(\nu_k, \pi) + \alpha M(h^3p)^{1/2}.$$

Using this inequality repeatedly for $k+1, k, k-1, \ldots, 1$, we get

$$\begin{aligned} W_2(\nu_{k+1}, \pi) &\leq \varrho^{k+1}\, W_2(\nu_0, \pi) + \alpha M (h^3 p)^{1/2} (1 + \varrho + \ldots + \varrho^k) \\ &\leq \varrho^{k+1}\, W_2(\nu_0, \pi) + \alpha M (h^3 p)^{1/2} (1-\varrho)^{-1}. \end{aligned}$$

This completes the proof.

2.B Proof of Theorem 5

Recall that $\alpha = 7\sqrt{2}/6 \leq 1.65$. Theorem 4 implies that using the step-size $h_k = 2/(M+m)$ for $k = 1, \ldots, K_1$, we get

$$\begin{aligned} W_2(\nu_{K_1}, \pi) &\leq \Big(1 + \frac{2m}{M-m}\Big)^{-K_1} W_2(\nu_0, \pi) + \frac{\alpha M}{m}\Big(\frac{2p}{m+M}\Big)^{1/2} \\ &\leq \frac{3.5M}{m}\Big(\frac{p}{M+m}\Big)^{1/2}. \end{aligned} \tag{2.26}$$

Starting from this iteration K_1, we use a decreasing step-size

$$h_{k+1} = \frac{2}{M + m + (2/3)m(k - K_1)}.$$

Let us show by induction over k that

$$W_2(\nu_k, \pi) \leq \frac{3.5M}{m}\bigg(\frac{p}{M + m + (2/3)m(k-K_1)}\bigg)^{1/2}, \qquad \forall\, k \geq K_1. \tag{2.27}$$

For $k = K_1$, this inequality is true in view of (2.26). Assume now that (2.27) is true for some k. For $k+1$, we have

$$\begin{aligned} W_2(\nu_{k+1}, \pi) &\leq (1 - m h_{k+1}) W_2(\nu_k, \pi) + \alpha M \sqrt{p}\, h_{k+1}^{3/2} \\ &\leq (1 - m h_{k+1}) \frac{3.5M\sqrt{p}\,(h_{k+1}/2)^{1/2}}{m} + \alpha M \sqrt{p}\, h_{k+1}^{3/2} \\ &\leq (1 - \frac{1}{3} m h_{k+1}) \frac{3.5M\sqrt{p}\,(h_{k+1}/2)^{1/2}}{m}. \end{aligned}$$

One can check that

$$(1 - \frac{1}{3} m h_{k+1})(h_{k+1}/2)^{1/2} = \frac{\sqrt{3}\,[m + 3M + 2m(k-K_1)]}{[3m + 3M + 2m(k-K_1)]^{3/2}}$$

$$
\begin{aligned}
&\leq \frac{\sqrt{3}\,[m+3M+2m(k-K_1)]^{1/2}}{3m+3M+2m(k-K_1)}\\
&\leq \frac{\sqrt{3}}{[3m+3M+2m(k+1-K_1)]^{1/2}}.
\end{aligned}
$$

This completes the proof of the theorem.

2.C Proof of Theorem 6

Let us denote by $\nu_k(\cdot|\boldsymbol{x})$ the conditional distribution of $\vartheta_k^{\mathrm{MLMC}}$ given $\boldsymbol{\eta} = \boldsymbol{x}$. In view of Theorem 5, we have

$$
W_2\big(\nu_k(\cdot|\boldsymbol{x}), \pi_1(\cdot|\boldsymbol{x})\big) \leq \frac{3.5M\sqrt{p}}{m\sqrt{M+m+(2/3)m(k-K_1)}}, \qquad \forall \boldsymbol{x} \in H.
$$

This readily yields

$$
\int_H W_2\big(\nu_k(\cdot|\boldsymbol{x}), \pi_1(\cdot|\boldsymbol{x})\big)\,\pi_0(d\boldsymbol{x}) \leq \frac{3.5M\sqrt{p}}{m\sqrt{M+m+(2/3)m(k-K_1)}}.
$$

The last step is to apply the convexity of the Wasserstein distance, which means that for any probability measure π_0, we have

$$
\int_H W_2\big(\nu_k(\cdot|\boldsymbol{x}), \pi_1(\cdot|\boldsymbol{x})\big)\,\pi_0(d\boldsymbol{x}) \geq W_2\bigg(\int_H \nu_k(\cdot|\boldsymbol{x})\,\pi_0(d\boldsymbol{x}), \int_H \pi_1(\cdot|\boldsymbol{x})\,\pi_0(d\boldsymbol{x})\bigg) = W_2(\nu_k, \pi).
$$

2.D Proof of Theorem 7

As explained in Section 2.3, the main new ingredient of the proof is Lemma 1, that has to be combined with Proposition 11. We postpone the proof of Lemma 1 to Section 2.G and do it in a more general form (see Lemma 7).

In view of Proposition 11, we have

$$
W_2(\nu_{k+1}, \pi)^2 \leq \big\{(1-mh)W_2(\nu_k, \pi) + \alpha M(h^3 p)^{1/2} + h\delta\sqrt{p}\big\}^2 + \sigma^2 h^2 p.
$$

We apply now Lemma 1 with $A = mh$, $B = \sigma h\sqrt{p}$ and $C = \alpha M(h^3p)^{1/2} + h\delta\sqrt{p}$, which implies that $W_2(\nu_k, \pi)$ is less than or equal to

$$
(1-mh)^k W_2(\nu_0, \pi) + \frac{\alpha M(hp)^{1/2} + \delta\sqrt{p}}{m} + \frac{\sigma^2 h\sqrt{p}}{\alpha M h^{1/2} + \delta + (mh)^{1/2}\,\sigma}.
$$

This completes the proof of the theorem.

2.E Proof of Theorem 8

Using the same construction and the same definitions as in the proof of Proposition 11, for $\boldsymbol{\Delta}_k = \boldsymbol{L}_0 - \boldsymbol{\vartheta}_{k,\boldsymbol{h}}$, we have

$$\begin{aligned}
\boldsymbol{\Delta}_{k+1} - \boldsymbol{\Delta}_k &= h\boldsymbol{Y}_{k,\boldsymbol{h}} - \int_{I_k} \nabla f(\boldsymbol{L}_t)\,dt \\
&= -h\big(\underbrace{\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}} + \boldsymbol{\Delta}_k) - \nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}})}_{:=\boldsymbol{U}}\big) \\
&\quad - \sqrt{2}\underbrace{\int_0^h \int_0^t \nabla^2 f(\boldsymbol{L}_s) d\boldsymbol{W}_s\,dt}_{:=\boldsymbol{S}} + h\boldsymbol{\zeta}_k \\
&\quad - \underbrace{\int_0^h \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0) - \sqrt{2}\int_0^t \nabla^2 f(\boldsymbol{L}_s) d\boldsymbol{W}_s\big)\,dt}_{:=\bar{\boldsymbol{V}}}.
\end{aligned}$$

Using the following equalities of conditional expectations $\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\Delta}_k, \boldsymbol{U}, \bar{\boldsymbol{V}}] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}, \boldsymbol{L}_0, \boldsymbol{W}] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}, \boldsymbol{L}_0] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]$ and $\mathbf{E}[\boldsymbol{S}_h|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}, \boldsymbol{L}_0] = 0$, we get

$$\begin{aligned}
\|\boldsymbol{\Delta}_{k+1}\|_{L_2}^2 &\le \big\|\boldsymbol{\Delta}_k - h\boldsymbol{U} - \bar{\boldsymbol{V}} - \sqrt{2}\boldsymbol{S}_h + h\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]\big\|_{L_2}^2 + \sigma^2 h^2 p \\
&\le \big\{\big(\|\boldsymbol{\Delta}_k - h\boldsymbol{U}\|_{L_2}^2 + 2\|\boldsymbol{S}_h\|_{L_2}^2\big)^{1/2} + h\delta\sqrt{p} + \|\bar{\boldsymbol{V}}\|_{L_2}\big\}^2 + \sigma^2 h^2 p.
\end{aligned}$$

In addition, we have

$$\begin{aligned}
\|\boldsymbol{S}_h\|_{L_2}^2 &= \Big\|\int_0^h (h-s)\nabla^2 f(\boldsymbol{L}_s)\,d\boldsymbol{W}_s\Big\|_{L_2}^2 \\
&= \int_0^h (h-s)^2 \mathbf{E}[\|\nabla^2 f(\boldsymbol{L}_s)\|_F^2]\,ds \le (1/3)\,M^2 h^3 p.
\end{aligned}$$

Setting $x_k = \|\boldsymbol{\Delta}_k\|_{L_2} = W_2(\nu_k, \pi)$ and using Lemma 2, this yields

$$x_{k+1}^2 \le \big\{\big((1-mh)^2 x_k^2 + (2/3)\,M^2 h^3 p\big)^{1/2} + h\delta\sqrt{p} + \|\bar{\boldsymbol{V}}\|_{L_2}\big\}^2 + \sigma^2 h^2 p.$$

Let us define $A = mh$, $F = (2/3)\,M^2 h^3 p$, $G = \sigma^2 h^2 p$ and[3]

$$C = h\delta\sqrt{p} + 0.5 M_2 h^2 p + 0.5 M^{3/2} h^2 \sqrt{p}.$$

[3]In view of Lemma 6 in Section 2.G, we have $h\delta\sqrt{p} + \|\bar{\boldsymbol{V}}\|_{L_2} \le C$.

Then

$$x_{k+1}^2 \leq \big\{\big((1-A)^2 x_k^2 + F\big)^{1/2} + C\big\}^2 + G.$$

One can deduce from this inequality that $x_{k+1}^2 \leq \big((1-A)x_k + C\big)^2 + F + G + 2C\sqrt{F}$. Therefore, using (2.46) of Lemma 7 below, we get

$$\begin{aligned}
x_k &\leq (1-A)^k x_0 + \frac{C}{A} + \frac{F + G + 2C\sqrt{F}}{C + \big(A(F+G+2C\sqrt{F})\big)^{1/2}} \\
&\leq (1-A)^k x_0 + (C/A) + 2(F/A)^{1/2} + \frac{G}{C + \sqrt{AG}}.
\end{aligned}$$

Replacing A, C, F and G by their respective expressions, we get the claim of the theorem.

2.F Proof of Theorem 9

To ease notation, throughout this proof, we will write ν_k and ν_k' instead of ν_k^{LMCO} and $\nu_k^{\text{LMCO}'}$, respectively.

Let $\boldsymbol{D}_0 \sim \nu_k$ and $\boldsymbol{L}_0 \sim \pi$ be two random variables such that $\|\boldsymbol{D}_0 - \boldsymbol{L}_0\|_{L_2}^2 = W_2(\nu_k, \pi)$. Let $\boldsymbol{W}$ be a p-dimensional Brownian motion independent of $(\boldsymbol{D}_0, \boldsymbol{L}_0)$. We define $\boldsymbol{L}$ to be the Langevin diffusion process (2.24) driven by $\boldsymbol{W}$ and starting at $\boldsymbol{L}_0$, whereas $\boldsymbol{D}$ is the process starting at $\boldsymbol{D}_0$ and satisfying the stochastic differential equation

$$d\boldsymbol{D}_t = -[\nabla f(\boldsymbol{D}_0) + \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{D}_t - \boldsymbol{D}_0)]\, dt + \sqrt{2}\, d\boldsymbol{W}_t, \quad t \geq 0. \tag{2.28}$$

This is an Ornstein-Uhlenbeck process. It can be expressed explicitly as a function of $\boldsymbol{D}_0$ and $\boldsymbol{W}$. The corresponding expression implies that $\boldsymbol{D}_h \sim \nu_{k+1}$ and, hence, $W_2(\nu_{k+1}, \pi) \leq \|\boldsymbol{D}_h - \boldsymbol{L}_h\|_{L_2}^2$.

An important ingredient of our proof is the following version of the Gronwall lemma, the proof of which is postponed to Section 2.G.

Lemma 5. *Let $\boldsymbol{\alpha} : [0,T] \times \Omega \to \mathbb{R}^p$ be a continuous semi-martingale and $\mathbf{H} : [0,T] \times \Omega \to \mathbb{R}^{p\times p}$ be a random process with continuous paths in the space of all symmetric $p \times p$ matrices such that $\mathbf{H}_s\mathbf{H}_t = \mathbf{H}_t\mathbf{H}_s$ for every $s, t \in [0,T]$. If $\boldsymbol{x} : [0,T] \times \Omega \to \mathbb{R}^p$ is a semi-martingale satisfying the identity*

$$\boldsymbol{x}_t = \boldsymbol{\alpha}_t - \int_0^t \mathbf{H}_s \boldsymbol{x}_s\, ds, \qquad \forall t \in [0,T], \tag{2.29}$$

then, for every $t \in [0, T]$*,*

$$\boldsymbol{x}_t = \exp\Big\{-\int_0^t \mathbf{H}_s\, ds\Big\}\boldsymbol{\alpha}_0 + \int_0^t \exp\Big\{-\int_s^t \mathbf{H}_u\, du\Big\} d\boldsymbol{\alpha}_s. \tag{2.30}$$

We denote $\boldsymbol{X}_t = \boldsymbol{L}_t - \boldsymbol{L}_0 - (\boldsymbol{D}_t - \boldsymbol{D}_0)$, where $\boldsymbol{D}_t$ is the random process defined in (2.28) and $\boldsymbol{L}_t$ is the Langevin diffusion driven by the same Wiener process $\boldsymbol{W}$ and with initial condition $\boldsymbol{L}_0 \sim \pi$. It is clear that

$$\begin{aligned}
\boldsymbol{X}_t &= -\int_0^t \nabla f(\boldsymbol{L}_s)\, ds + \int_0^t [\nabla f(\boldsymbol{D}_0) + \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{D}_s - \boldsymbol{D}_0)]\, ds \\
&= -\int_0^t \big\{\nabla f(\boldsymbol{L}_s) - \nabla f(\boldsymbol{D}_0) - \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{L}_s - \boldsymbol{L}_0)\big\}\, ds - \int_0^t \nabla^2 f(\boldsymbol{D}_0)\boldsymbol{X}_s\, ds.
\end{aligned}$$

Using Lemma 5, we get

$$\begin{aligned}
\boldsymbol{X}_t &= -\int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}\big\{\nabla f(\boldsymbol{L}_s) - \nabla f(\boldsymbol{D}_0) - \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{L}_s - \boldsymbol{L}_0)\big\}\, ds \\
&= \int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}\, ds[\nabla f(\boldsymbol{D}_0) - \nabla f(\boldsymbol{L}_0)] \\
&\quad - \int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}\big\{\nabla f(\boldsymbol{L}_s) - \nabla f(\boldsymbol{L}_0) - \nabla^2 f(\boldsymbol{L}_0)(\boldsymbol{L}_s - \boldsymbol{L}_0)\big\}\, ds \\
&\quad - \int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}[\nabla^2 f(\boldsymbol{D}_0) - \nabla^2 f(\boldsymbol{L}_0)] \int_0^s \nabla f(\boldsymbol{L}_u)\, du\, ds \\
&\quad + \sqrt{2}\int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}[\nabla^2 f(\boldsymbol{D}_0) - \nabla^2 f(\boldsymbol{L}_0)]\boldsymbol{W}_s\, ds.
\end{aligned} \tag{2.31}$$

Let us set $\boldsymbol{\Delta}_t = \boldsymbol{L}_t - \boldsymbol{D}_t$. We have $\boldsymbol{X}_t = \boldsymbol{\Delta}_t - \boldsymbol{\Delta}_0 = A_t - B_t - C_t + S_t$, where A_t, B_t, C_t and S_t stand for the four integrals in (2.31). We now evaluate these terms separately. For the first one, using the notation $\mathbf{H}_0 = \nabla^2 f(\boldsymbol{D}_0)$ and the identity $\nabla f(\boldsymbol{L}_0) - \nabla f(\boldsymbol{D}_0) = \int_0^1 \nabla^2 f(\boldsymbol{D}_0 + x\boldsymbol{\Delta}_0)\, dx \boldsymbol{\Delta}_0$, we get

$$\begin{aligned}
\|\boldsymbol{\Delta}_0 + A_t\|_2 &\le \|\boldsymbol{\Delta}_0 - t\big(\nabla f(\boldsymbol{L}_0) - \nabla f(\boldsymbol{D}_0)\big)\|_2 \\
&\qquad + \int_0^t \|\mathbf{I} - e^{-s\mathbf{H}_0}\|\, ds \big\|\nabla f(\boldsymbol{L}_0) - \nabla f(\boldsymbol{D}_0)\big\|_2 \\
&\le (1 - mt + 0.5M^2t^2)\|\boldsymbol{\Delta}_0\|_2.
\end{aligned} \tag{2.32}$$

For the term B_t with $t \le h \le m/M^2 \le 1/M$, we can apply (2.44) to infer that

$$\|B_t\|_{L_2}^2 \le 0.88 M_2 t^2 (p^2 + 2p)^{1/2}. \tag{2.33}$$

As for C_t, in view of the inequality $\|\nabla^2 f(\boldsymbol{L}_0) - \nabla^2 f(\boldsymbol{D}_0)\| \le M_2\|\boldsymbol{\Delta}_0\|_2 \wedge M \le \sqrt{MM_2\|\boldsymbol{\Delta}_0\|_2}$, we have

$$\begin{aligned}\|C_t\|_2 &\le \sqrt{MM_2\|\boldsymbol{\Delta}_0\|_2}\int_0^t\int_0^s \|\nabla f(\boldsymbol{L}_u)\|_2\,du\,ds\\ &\le \mu\|\boldsymbol{\Delta}_0\|_2 + (4\mu)^{-1}MM_2\bigg(\int_0^t (t-u)\|\nabla f(\boldsymbol{L}_u)\|_2\,du\bigg)^2.\end{aligned}$$

On the other hand, the fact that $\mathbf{E}[\|\nabla f(\boldsymbol{L}_u)\|_2^4] \le M^2(p^2+2p)$ yields

$$\bigg(\int_0^t (t-u)(\mathbf{E}[\|\nabla f(\boldsymbol{L}_u)\|_2^4])^{1/4}\,du\bigg)^2 \le \frac{Mt^4(p^2+2p)^{1/2}}{4}. \tag{2.34}$$

This implies the inequality

$$\|C_t\|_{L_2} \le \mu W_2(\nu_k,\pi) + (16\mu)^{-1}M^2M_2t^4(p+1). \tag{2.35}$$

Finally, using the integration by parts formula for semi-martingales, one can easily write S_t as a stochastic integral with respect to $\boldsymbol{W}$ and derive from that representation the inequality

$$\begin{aligned}\|S_t\|_{L_2}^2 &\le 2\mathbf{E}\bigg[\int_0^t \bigg\|\int_u^t e^{-s\mathbf{H}_0}\,ds\big(\nabla^2 f(\boldsymbol{L}_0) - \nabla^2 f(\boldsymbol{D}_0)\big)\bigg\|_F^2\,du\bigg]\\ &\le 2p\mathbf{E}[(M_2\|\boldsymbol{\Delta}_0\|_2\wedge M)^2]\int_0^t (t-u)^2\,du \le (2/3)M_2Mpt^3\|\boldsymbol{\Delta}_0\|_{L_2}^2.\end{aligned} \tag{2.36}$$

Putting all these pieces together, taking the expectation, using the Minkowski inequality, the equality $\mathbf{E}[(\boldsymbol{\Delta}_0 + A_h)^\top S_h] = 0$ and the inequality $\sqrt{a^2+b} \le a + b/(2a)$, we get

$$\begin{aligned}\|\boldsymbol{\Delta}_h\|_{L_2}^2 &= \|\boldsymbol{\Delta}_0 + A_h - B_h - C_h + S_h\|_{L_2}^2\\ &\le \big(\|\boldsymbol{\Delta}_0 + A_h\|_{L_2}^2 + \|S_h\|_{L_2}^2\big)^{1/2} + \|B_h\|_{L_2}^2 + \|C_h\|_{L_2}^2\\ &\le \big(1 - mh + 0.5M^2h^2 + \mu\big)\|\boldsymbol{\Delta}_0\|_{L_2}^2 + \frac{M_2Mph^3}{3(1-mh+0.5M^2h^2)}\\ &\quad + 0.88M_2h^2(p^2+2p)^{1/2} + \frac{M^2M_2h^4}{16\mu}(p+1).\end{aligned} \tag{2.37}$$

Let μ be any real number smaller than $0.5h(m - 0.5M^2h)$; Eq. (2.37) and the inequality

$p^2 + 2p \leq (p+1)^2$ yield

$$W_2(\nu_{k+1}, \pi) \leq (1-\mu) W_2(\nu_k, \pi) + \frac{M_2 M p h^3}{3(1-2\mu)} + 0.88 M_2 h^2 (p+1) + \frac{M^2 M_2 h^4}{16\mu}(p+1).$$

Since $h \leq m/M^2$, we can choose $\mu = 0.25mh$ so that $1 - 2\mu = 1 - 0.5mh \geq 0.5$ and

$$\begin{aligned} W_2(\nu_{k+1}, \pi) &\leq (1 - 0.25mh) W_2(\nu_k, \pi) + \frac{2M_2 M p h^3}{3} + 0.88 M_2 h^2 (p+1) \\ &\quad + \frac{M^2 M_2 h^3}{4m}(p+1) \\ &\leq (1 - 0.25mh) W_2(\nu_k, \pi) + 1.8 M_2 h^2 (p+1). \end{aligned}$$

This recursion implies the inequality

$$\begin{aligned} W_2(\nu_k, \pi) &\leq (1 - 0.25mh)^k W_2(\nu_0, \pi) + \frac{1.8 M_2 h(p+1)}{0.25m} \\ &= (1 - 0.25mh)^k W_2(\nu_0, \pi) + \frac{7.2 M_2 h(p+1)}{m}. \end{aligned}$$

This completes the proof of claim (2.18) of the theorem.

To establish inequality (2.19), we follow the same steps as in the proof of (2.18), with a slightly different choice of the process $\boldsymbol{D}$. More precisely, we define $\boldsymbol{D}$ by

$$\boldsymbol{D}_t - \boldsymbol{D}_0 = -(t\mathbf{I}_p - 0.5t^2 \nabla^2 f(\boldsymbol{D}_0)) \nabla f(\boldsymbol{D}_0) + \sqrt{2} \int_0^t (\mathbf{I} - (t-u)\nabla^2 f(\boldsymbol{D}_0))\, d\boldsymbol{W}_u.$$

One can check that the conditional distribution of $\boldsymbol{D}_h$ given $\boldsymbol{D}_0 = \boldsymbol{x}$ coincides with the conditional distribution of $\boldsymbol{\vartheta}^{\mathrm{LMCO}'}_{k+1,h}$ given $\boldsymbol{\vartheta}^{\mathrm{LMCO}'}_{k,h} = \boldsymbol{x}$. Therefore, if $\boldsymbol{D}_0 \sim \nu'_k$, then $\boldsymbol{D}_h \sim \nu'_{k+1}$ and, consequently, $W_2(\nu'_{k+1}, \pi)^2 \leq \mathbf{E}[\|\boldsymbol{D}_h - \boldsymbol{L}_h\|_2^2]$.

To ease notation, we set $\mathbf{H}_0 = \nabla^2 f(\boldsymbol{D}_0)$. The process $\boldsymbol{D}$ satisfies the SDE

$$d\boldsymbol{D}_t = -\big[(\mathbf{I}_p - t\nabla^2 f(\boldsymbol{D}_0))\nabla f(\boldsymbol{D}_0) + \sqrt{2}\,\mathbf{H}_0 \boldsymbol{W}_t\big]\, dt + \sqrt{2}\, d\boldsymbol{W}_t,$$

which implies that

$$\begin{aligned} d\boldsymbol{D}_t = &- \big[\nabla f(\boldsymbol{D}_0) + \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{D}_t - \boldsymbol{D}_0)\big]\, dt + \sqrt{2}\, d\boldsymbol{W}_t \\ &- 0.5t^2 \mathbf{H}_0^2 \nabla f(\boldsymbol{D}_0)\, dt - \sqrt{2}\,\mathbf{H}_0^2 \int_0^t (t-u)\, d\boldsymbol{W}_u\, dt. \end{aligned}$$

Proceeding in the same way as for getting (2.31), we arrive at the decomposition $\boldsymbol{X}_t =$

$\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_0 = A_t - B_t - C_t + S_t - E_t - F_t$, where A_t, B_t, C_t and S_t stand for the four integrals in (2.31) whereas E_t and F_t are

$$E_t = 0.5 \int_0^t e^{-s\mathbf{H}_0} s^2\, ds\, \mathbf{H}_0^2 \nabla f(\boldsymbol{D}_0)$$
$$F_t = \sqrt{2}\, \mathbf{H}_0^2 \int_0^t e^{-s\mathbf{H}_0} \int_0^s (s-u)\, d\boldsymbol{W}_u\, ds.$$

Using the properties of the stochastic integral, we get

$$\begin{aligned}
\mathbf{E}[\|F_h\|_2^2] &= 2\mathbf{E}\Big[\Big\|\mathbf{H}_0^2 \int_0^h e^{-s\mathbf{H}_0} \int_0^s (s-u)\, d\boldsymbol{W}_u\, ds\Big\|_2^2\Big] \\
&= 2\mathbf{E}\Big[\Big\|\int_0^h \int_u^h \mathbf{H}_0^2 e^{-s\mathbf{H}_0}(s-u)\, ds\, d\boldsymbol{W}_u\Big\|_2^2\Big] \\
&= 2\int_0^h \Big\|\int_u^h \mathbf{H}_0^2 e^{-s\mathbf{H}_0}(s-u)\, ds\Big\|_F^2\, du \\
&\leq 2M^4 p \int_0^h \Big(\int_u^h (s-u)\, ds\Big)^2 du = \frac{M^4 h^5 p}{10}.
\end{aligned} \tag{2.38}$$

On the other hand,

$$\|E_h\|_2 \leq 0.5 M^2 \int_0^h s^2\, ds \|\nabla f(\boldsymbol{D}_0)\|_2 \leq \frac{M^2 h^3}{6}\big(\|\nabla f(\boldsymbol{L}_0)\|_2 + M\|\boldsymbol{\Delta}_0\|_2\big),$$

which, in view of Lemma 3, implies that

$$\|E_h\|_{L_2}^2 \leq \frac{M^2 h^3}{6}\big(\sqrt{Mp} + M W_2(\nu_k', \pi)\big). \tag{2.39}$$

Proceeding as in (2.37) and using (2.34), we get

$$\begin{aligned}
\|\boldsymbol{\Delta}_h\|_{L_2} &= \|\boldsymbol{\Delta}_0 + A_h - B_h - C_h + S_h - E_h - F_h\|_{L_2} \\
&\leq \|\boldsymbol{\Delta}_0 + A_h + S_h - F_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|E_h\|_{L_2} \\
&\leq (\|\boldsymbol{\Delta}_0 + A_h\|_{L_2}^2 + \|S_h - F_h\|_{L_2}^2)^{1/2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|E_h\|_{L_2}.
\end{aligned} \tag{2.40}$$

Using the last but one estimate in (2.36), in conjunction with (2.38), we get inequalities

$$\|S_h\|_{L_2}^2 \leq (2/3) M_2 M h^3 p W_2(\nu_k', \pi)$$
$$|\mathbf{E}[S_h^\top F_h]| \leq (1/\sqrt{15}) M^2 M_2 h^4 p W_2(\nu_k', \pi),$$

which, for $h \leq 3m/(4M^2)$, imply that $\|S_h - F_h\|_{L_2}^2$ is less than or equal to

$$(2/3)M_2Mh^3pW_2(\nu_k', \pi) + (2/\sqrt{15})M^2M_2h^4pW_2(\nu_k', \pi) + (1/10)M^4h^5p \\ \leq 1.06M_2Mh^3pW_2(\nu_k', \pi) + 0.1M^4h^5p.$$

Injecting this bound, (2.32), (2.33), (2.35) and (2.39) in (2.40), we arrive at

$$\|\mathbf{\Delta}_h\|_{L_2} \leq \big\{\big[(1 - mh + 0.5M^2h^2)^2W_2(\nu_k', \pi)^2 + 1.06M_2Mh^3pW_2(\nu_k', \pi) + 0.1M^4h^5p\big\}^{1/2} \\ + 0.88M_2h^2(p+1) + \Big(\mu + \frac{M^3h^3}{6}\Big)W_2(\nu_k', \pi) + \frac{M^2M_2h^4(p+1)}{16\mu} + \frac{M^{5/2}h^3\sqrt{p}}{6}.$$

In view of the inequality $\sqrt{a^2 + b + c} \leq \sqrt{a^2 + c} + (b/2a)$, the last display leads to

$$W_2(\nu_{k+1}', \pi) \leq \big\{\big[(1 - mh + 0.5M^2h^2)^2W_2(\nu_k', \pi)^2 + 0.1M^4h^5p\big\}^{1/2} \\ + \frac{0.53M_2Mh^3p}{1 - mh + 0.5M^2h^2} + 0.88M_2h^2(p+1) + \Big(\mu + \frac{M^3h^3}{6}\Big)W_2(\nu_k', \pi) \\ + \frac{M^2M_2h^4(p+1)}{16\mu} + \frac{M^{5/2}h^3\sqrt{p}}{6}.$$

For $h \leq 3m/(4M^2)$ and $\mu = 0.25mh$, we can use the inequality $1 - mh + 0.5M^2h^2 \geq 17/32$ and simplify the last display as follows:

$$W_2(\nu_{k+1}', \pi) \leq \big\{\big[(1 - mh + 0.5M^2h^2)^2W_2(\nu_k', \pi)^2 + 0.1M^4h^5p\big\}^{1/2} \\ + \frac{0.3975M_2h^2(p+1)}{1 - mh + 0.5M^2h^2} + 0.88M_2h^2(p+1) + \Big(\mu + \frac{M^3h^3}{6}\Big)W_2(\nu_k', \pi) \\ + \frac{3M_2h^2(p+1)}{16} + \frac{M^{5/2}h^3\sqrt{p}}{6} \\ \leq \big\{(1 - mh + 0.5M^2h^2)^2W_2(\nu_k', \pi)^2 + 0.1M^4h^5p\big\}^{1/2} \\ + \Big(0.25mh + \frac{M^3h^3}{6}\Big)W_2(\nu_k', \pi) + 1.82M_2h^2(p+1) + \frac{M^{5/2}h^3\sqrt{p}}{6}.$$

We apply Lemma 9 to the sequence $x_k = W_2(\nu_k', \pi)$ with $A = mh - 0.5M^2h^2$ and $D = 0.25mh + M^3h^3/6$. For $h \leq 3m/(4M^2)$ we have $A - D = 0.75mh - 0.5M^2h^2 - (Mh)^3/6 \geq 0.25mh$ and $A + D \leq 1.25mh - (3/8)M^2h^2 \leq 0.727$. This yields

$$W_2(\nu_{k+1}', \pi) \leq (1 - 0.25mh)^kW_2(\nu_0', \pi) + \frac{7.28M_2h(p+1)}{m} + \frac{2M^{5/2}h^2\sqrt{p}}{3m} + \frac{2\sqrt{0.1}\,M^2h^2\sqrt{p}}{\sqrt{1.273m}} \\ \leq (1 - 0.25mh)^kW_2(\nu_0', \pi) + \frac{7.28M_2h(p+1)}{m} + \frac{1.23M^{5/2}h^2\sqrt{p}}{m}.$$

This completes the proof of (2.19) and that of the theorem.

Proof of Proposition 10. Let us denote $\mathbf{M}_k = \int_0^h e^{-s\mathbf{H}_k}\,ds \int_0^1 \nabla^2 f(\boldsymbol{D}_{kh} + x\boldsymbol{\Delta}_k)\,dx$. From (2.31), we have $\boldsymbol{\Delta}_{k+1} = \boldsymbol{\Delta}_k + A_{k,h} + G_{k,h}$ with

$$\begin{aligned}
A_{k,h} &= \int_0^h e^{-s\mathbf{H}_k}\,ds\big(\nabla f(\boldsymbol{D}_{kh}) - \nabla f(\boldsymbol{L}_{kh})\big) = -\mathbf{M}_k\boldsymbol{\Delta}_k,\\
G_{k,h} &= \int_0^h e^{-s\mathbf{H}_k}\big(\nabla f(\boldsymbol{L}_{kh}) - \nabla f(\boldsymbol{L}_s) + \mathbf{H}_k(\boldsymbol{L}_s - \boldsymbol{L}_{kh})\big)\,ds.
\end{aligned}$$

Using the fact that

$$\bigg\|\int_0^1 \nabla^2 f(\boldsymbol{D}_{kh} + x\boldsymbol{\Delta}_k)\,dx - \mathbf{H}_k\bigg\| \le \int_0^1 \big\|\nabla^2 f(\boldsymbol{D}_{kh} + x\boldsymbol{\Delta}_k) - \mathbf{H}_k\big\|\,dx \le \frac{M_2}{2}\,\|\boldsymbol{\Delta}_k\|_2,$$

we get $\|\boldsymbol{\Delta}_k + A_{k,h}\|_2 = \|(\mathbf{I} - \mathbf{M}_k)\boldsymbol{\Delta}_k\|_2 \le \frac{M_2}{2m}\,\|\boldsymbol{\Delta}_k\|_2^2 + e^{-mh}\|\boldsymbol{\Delta}_k\|_2$. This further leads to the recursive inequality

$$\|\boldsymbol{\Delta}_{k+1}\|_2 \le \frac{M_2}{2m}\,\|\boldsymbol{\Delta}_k\|_2^2 + e^{-mh}\|\boldsymbol{\Delta}_k\|_2 + \|G_{k,h}\|_2.$$

In view of the Minkowski inequality, this yields

$$(\mathbf{E}[\|\boldsymbol{\Delta}_{k+1}\|_2^q])^{1/q} \le \frac{M_2}{2m}\,\mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^{2q}]^{1/q} + e^{-mh}\mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^{2q}]^{1/2q} + \mathbf{E}[\|G_{k,h}\|_2^q]^{1/q}. \tag{2.41}$$

We choose some $K \in \mathbb{N}$ and define the sequence $\{x_0, \ldots, x_K\}$ by setting $x_k^{2^{K+1-k}} = \mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^{2^{K+1-k}}]$. Choosing in (2.41) $q = 2^{K-k}$, we get

$$x_{k+1} \le \frac{M_2}{2m}\,x_k^2 + e^{-mh}x_k + \mathbf{E}[\|G_{k,h}\|_2^{2^{K-k}}]^{2^{k-K}}, \quad k = 0, 1, \ldots, K-1.$$

We are in a position to apply Lemma 8 to the sequence $\{x_k\}_{k=0,\ldots,K}$. This yields

$$x_K \le \frac{2m}{M_2}\bigg(\frac{M_2 x_0}{2m} + \frac{1}{2}e^{-mh}\bigg)^{2^K} \exp\bigg\{2^K\,\frac{M_2 \max_k \mathbf{E}[\|G_{k,h}\|_2^{2^K}]^{2^{-K}} + me^{-mh}}{m(\frac{M_2 x_0}{2m} + \frac{1}{2}e^{-mh})^{2^K}}\bigg\}, \tag{2.42}$$

where $\max_k$ is a short notation for $\max_{k=0,1,\ldots,K-1}$. It suffices now to upper bound the moments of $\|G_{k,h}\|_2$. We have

$$\begin{aligned}
\mathbf{E}[\|G_{k,h}\|_2^q]^{1/q} &\le M\int_0^h e^{-sm}\big(\mathbf{E}[\|\boldsymbol{L}_{kh+s} - \boldsymbol{L}_{kh}\|_2^q]\big)^{1/q}\,ds\\
&\le M\int_0^h e^{-sm}\Big\{\big(\mathbf{E}[\|\int_0^s \nabla f(\boldsymbol{L}_{kh+u})\,du\|_2^q]\big)^{1/q} + \sqrt{2}\big(\mathbf{E}[\|\boldsymbol{W}_s\|_2^q]\big)^{1/q}\Big\}\,ds\\
&\le M\int_0^h e^{-sm}s\,ds\big(\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^q]\big)^{1/q} + M\sqrt{2p+q-2}\int_0^s e^{-sm}\sqrt{s}\,ds.
\end{aligned}$$

Thus,

$$\mathbf{E}[\|G_{k,h}\|_2^q]^{1/q} \le \frac{M}{m^2}\big(\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^q]\big)^{1/q} + \frac{M}{2m^{3/2}}\sqrt{(2p+q-2)\pi}.$$

On the other hand, by integration by parts, for every $q \in 2\mathbb{N}$, we have

$$\begin{aligned}
\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^q] &= -\int_{\mathbb{R}^p} \|\nabla f(\boldsymbol{x})\|_2^{q-2}\, \nabla f(\boldsymbol{x})^\top d\pi(\boldsymbol{x}) \\
&= \sum_{\ell=1}^{p} \int_{\mathbb{R}^p} \partial_\ell\Big(\|\nabla f(\boldsymbol{x})\|_2^{q-2}\, \partial_\ell f(\boldsymbol{x})\Big)\pi(\boldsymbol{x})\, d\boldsymbol{x} \\
&\le M(p+q-2)\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^{q-2}].
\end{aligned}$$

This yields $(\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^q])^{1/q} \le \sqrt{M(p+0.5q-1)}$. Combining all these estimates, we arrive at

$$\mathbf{E}[\|G_{k,h}\|_2^q]^{1/q} \le \frac{1.6M^{3/2}\sqrt{2p+q-2}}{m^2}.$$

Combining this inequality with (2.42) and replacing x_K by $(\mathbf{E}[\|\boldsymbol{\Delta}_K\|_2^2])^{1/2}$, we get

$$(\mathbf{E}[\|\boldsymbol{\Delta}_K\|_2^2])^{1/2} \le \frac{2m}{M_2}\left(\frac{M_2 x_0}{2m} + \frac{1}{2}e^{-mh}\right)^{2^K} \exp\left\{2^K \frac{1.6M_2M^{3/2}\sqrt{2p+2^{K-1}-2}+m^3e^{-mh}}{m^3(\frac{M_2x_0}{2m}+\frac{1}{2}e^{-mh})^{2^K}}\right\}.$$

This completes the proof of the proposition. $\square$

2.G Proofs of the lemmas

Here we provide the proofs of Lemma 2, Lemma 3 and Lemma 4.

Proof of Lemma 2. We start by recalling the following inequality [Nes04, Theorem 2.12], true for any m-strongly convex and M-gradient Lipschitz function f:

$$(\boldsymbol{y}-\boldsymbol{x})^\top(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})) \ge \frac{mM}{m+M}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + \frac{1}{m+M}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2^2,$$

for all vectors $\boldsymbol{x}$ and y from $\mathbb{R}^p$. This yields

$$\begin{aligned}
&\|\boldsymbol{y}-\boldsymbol{x}-h(\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x}))\|_2^2 \\
&\qquad = \|\boldsymbol{y}-\boldsymbol{x}\|_2^2 - 2h(\boldsymbol{y}-\boldsymbol{x})^\top(\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x})) + h^2\|\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x})\|_2^2 \\
&\qquad \le \left(1-\frac{2hmM}{m+M}\right)\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + h\left(h-\frac{2}{m+M}\right)\|\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x})\|_2^2.
\end{aligned}$$

Since f is m-strongly convex, we have ([Nes04], Theorem 2.1.9)

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2 \geq m\|\boldsymbol{y} - \boldsymbol{x}\|_2.$$

In the case $h \leq \frac{2}{m+M}$, applying the previous result to the second summand, we get

$$\|\boldsymbol{y} - \boldsymbol{x} - h(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))\|_2^2 \leq (1 - hm)^2 \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

In the case when $h \geq \frac{2}{m+M}$, we use the Lipschitz continuity of ∇f, which leads to

$$\|\boldsymbol{y} - \boldsymbol{x} - h(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))\|_2^2 \leq (hM - 1)^2 \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

Summing up, for all $h \in (0, 2/M)$ we have shown

$$\|\boldsymbol{y} - \boldsymbol{x} - h(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))\|_2^2 \leq \left\{(1 - hm)^2 \vee (hM - 1)^2\right\} \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

This completes the proof. □

Proof of Lemma 3. We start the proof with the case $p = 1$. The function $x \mapsto f'(x)$ being Lipschitz continuous is almost surely differentiable. Furthermore, it is clear that $|f''(x)| \leq M$ for every x for which this second derivative exists. The result of [Rud87, Theorem 7.20] implies that

$$f'(x) - f'(0) = \int_0^x f''(y)\, dy.$$

Therefore, using the relation $f'(x)\,\pi(x) = -\pi'(x)$, we get

$$\begin{aligned}
\int_{\mathbb{R}} f'(x)^2\, \pi(x)\, dx &= f'(0) \int_{\mathbb{R}} f'(x)\, \pi(x)\, dx + \int_{\mathbb{R}} \Big(\int_0^x f''(y)\, dy \Big) f'(x)\, \pi(x)\, dx \\
&= -f'(0) \int_{\mathbb{R}} \pi'(x)\, dx - \int_{\mathbb{R}} \Big(\int_0^x f''(y)\, dy \Big) \pi'(x)\, dx \\
&= -\int_0^\infty \int_0^x f''(y)\, \pi'(x)\, dy\, dx + \int_{-\infty}^0 \int_x^0 f''(y)\, \pi'(x)\, dy\, dx.
\end{aligned}$$

In view of Fubini's theorem, we arrive at

$$\int_{\mathbb{R}} f'(x)^2\, \pi(x)\, dx = \int_0^\infty f''(y)\, \pi(y)\, dy + \int_{-\infty}^0 f''(y)\, \pi(y)\, dy \leq M. \tag{2.43}$$

Now let us return to the multidimensional case:

$$\int_{\mathbb{R}^p} \|\nabla f(\boldsymbol{x})\|_2^2\, \pi(\boldsymbol{x})\, d\boldsymbol{x} = \sum_{k=1}^p \int_{\mathbb{R}^p} \Big(\frac{\partial f}{\partial x_k}(\boldsymbol{x}) \Big)^2 \pi(\boldsymbol{x})\, d\boldsymbol{x}.$$

We will show that each of the summands is less than M, thus the sum is less than Mp. Let us prove it for $k = 1$. The proof is similar for the case $k > 1$. Using Fubini's theorem, we have

$$\int_{\mathbb{R}^p} \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x})\right)^2 \pi(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left(\frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_p)\right)^2 \pi(x_1, x_2, \ldots, x_p)\, dx_1 dx_2 \ldots dx_p.$$

Let us fix the $(p-1)$-tuple $(x_2, x_3, \ldots, x_p)$ and define functions g and η as $g(t) = f(t, x_2, \ldots, x_p)$ and $\eta(t) = \pi(t, x_2, \ldots, x_p)$, respectively. It is easy to verify that η is an integrable log-concave function, with g as its potential. The latter is also differentiable and its derivative is Lipschitz-continuous with constant M. Thus we have

$$\int_{\mathbb{R}} \left(\frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_p)\right)^2 \pi(x_1, x_2, \ldots, x_p)\, dx_1 = \int_{\mathbb{R}} \left(g'(t)\right)^2 \eta(t) dt.$$

From the definition one can verify that $\int_{\mathbb{R}} \eta(t) dt = \pi_1(x_2, \ldots, x_p)$, where π_1 is the marginal distribution of all the coordinates except the first. Therefore,

$$\int_{\mathbb{R}} g'(t)^2 \eta(t) dt = \pi_1(x_2, \ldots, x_p) \int_{\mathbb{R}} g'(t)^2 \frac{\eta(t)}{\pi_1(x_2, \ldots, x_p)} dt \leq M \pi_1(x_2, \ldots, x_p).$$

The last inequality is true due to (2.43). Returning to our initial integral, we obtain

$$\int_{\mathbb{R}^p} \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x})\right)^2 \pi(\boldsymbol{x})\, d\boldsymbol{x} \leq M \int_{\mathbb{R}^{p-1}} \pi_1(x_2, \ldots, x_p) dx_2 \ldots dx_p = M.$$

This completes the proof. □

Proof of Lemma 4. Since the process $\boldsymbol{L}$ is stationary, $V(a)$ has the same distribution as $V(0)$. For this reason, it suffices to prove the claim of the lemma for $a = 0$ only. Using the Cauchy-Schwarz inequality and the Lipschitz continuity of f, we get

$$\begin{aligned}
\|\boldsymbol{V}(0)\|_{L_2} &= \Big\| \int_0^h \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0)\big)\, dt \Big\|_{L_2} \\
&\leq \int_0^h \big\|\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0)\big\|_{L_2} dt \\
&\leq M \int_0^h \big\|\boldsymbol{L}_t - \boldsymbol{L}_0\big\|_{L_2} dt.
\end{aligned}$$

Combining this inequality with the definition of $\boldsymbol{L}_t$, we arrive at

$$\|\boldsymbol{V}(0)\|_{L_2} \leq M \int_0^h \big\| - \int_0^t \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\, \boldsymbol{W}_t \big\|_{L_2} dt.$$

Therefore,

$$\begin{aligned}\|\boldsymbol{V}(0)\|_{L_2} &\le M\int_0^h \Big\|\int_0^t \nabla f(\boldsymbol{L}_s)\,ds\Big\|_{L_2} dt + M\int_0^h \big\|\sqrt{2}\,\boldsymbol{W}_t\big\|_{L_2} dt\\ &\le M\int_0^h\int_0^t \|\nabla f(\boldsymbol{L}_s)\|_{L_2}\,ds\,dt + M\int_0^h \sqrt{2pt}\,dt.\end{aligned}$$

In view of the stationarity of $\boldsymbol{L}_t$, we have $\|\nabla f(\boldsymbol{L}_s)\|_{L_2} = \|\nabla f(\boldsymbol{L}_0)\|_{L_2}$, which leads to

$$\|\boldsymbol{V}(0)\|_{L_2} \le (1/2)Mh^2\big\|\nabla f(\boldsymbol{L}_0)\big\|_{L_2} + (2/3)M\sqrt{2p}\,h^{3/2}.$$

To complete the proof, it suffices to apply Lemma 3. □

Lemma 6. *Let us denote*

$$\begin{aligned}\widetilde{\boldsymbol{V}} &= \int_0^h \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0) - \nabla^2 f(\boldsymbol{L}_0)(\boldsymbol{L}_t - \boldsymbol{L}_0)\big)\,dt,\\ \bar{\boldsymbol{V}} &= \int_0^h \Big\{\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0) - \sqrt{2}\int_0^t \nabla^2 f(\boldsymbol{L}_s)d\boldsymbol{W}_s\Big\}\,dt,\end{aligned}$$

with f satisfying **Condition F** *and $h \le 1/M$, then*

$$(\mathbf{E}[\|\widetilde{\boldsymbol{V}}\|_2^2])^{1/2} \le 0.877M_2h^2(p^2+2p)^{1/2}, \tag{2.44}$$

$$\|\bar{\boldsymbol{V}}\|_{L_2} \le (1/2)(M^{3/2}\sqrt{p} + M_2p)h^2. \tag{2.45}$$

Proof. We first note that we have

$$\begin{aligned}\|\widetilde{\boldsymbol{V}}\|_2 &\le \int_0^h \Big\|\int_0^1 \big(\nabla^2 f(\boldsymbol{L}_0 + x(\boldsymbol{L}_t - \boldsymbol{L}_0)) - \nabla^2 f(\boldsymbol{L}_0)\big)\,dx(\boldsymbol{L}_t - \boldsymbol{L}_0)\Big\|_2\,dt\\ &\le 0.5M_2\int_0^h \|\boldsymbol{L}_t - \boldsymbol{L}_0\|_2^2\,dt.\end{aligned}$$

In view of (2.23), this implies that $(\mathbf{E}[\|\widetilde{\boldsymbol{V}}\|_2^2])^{1/2} \le 0.5M_2\int_0^h(\mathbf{E}[\|\boldsymbol{L}_t - \boldsymbol{L}_0\|_2^4])^{1/2}\,dt$. Using the triangle inequality and integration by parts (precise details of the computations are omitted in the interest of saving space), we arrive at

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{L}_t - \boldsymbol{L}_0\|_2^4] &\le \mathbf{E}\Big[\Big\|\int_0^t \nabla f(\boldsymbol{L}_s)\Big\|_2^4\Big] + 4\mathbf{E}[\|\boldsymbol{W}_t\|_2^4]\\ &\qquad + 12\Big(\mathbf{E}\Big[\Big\|\int_0^t \nabla f(\boldsymbol{L}_s)\Big\|_2^4\Big]\mathbf{E}[\|\sqrt{2}\boldsymbol{W}_t\|_2^4]\Big)^{1/2}\\ &\le t^4M^2p(2+p) + 12t^3Mp(2+p) + 4t^2p(2+p)\\ &= p(2+p)t^2(t^2M^2 + 12tM + 4).\end{aligned}$$

Integrating this inequality, we get

$$
\begin{aligned}
(\mathbf{E}[\|\widetilde{\boldsymbol{V}}\|_2^2])^{1/2} &\le 0.5M_2(p^2+2p)^{1/2}\int_0^h t(t^2M^2+12tM+4)^{1/2}\,dt\\
&\le \frac{0.5M_2(p^2+2p)^{1/2}}{M^2}\int_0^{Mh} t(t^2+12t+4)^{1/2}\,dt\\
&\le 0.5M_2h^2(p^2+2p)^{1/2}\sup_{x\in(0,2]}\frac{1}{x^2}\int_0^x t(t^2+12t+4)^{1/2}\,dt\\
&= \frac{0.5M_2h^2(p^2+2p)^{1/2}}{4}\int_0^2 t(t^2+12t+4)^{1/2}\,dt\\
&\le 1.16M_2h^2(p^2+2p)^{1/2}.
\end{aligned}
$$

This completes the proof of (2.44). To prove (2.45), we first assume that f is three times continuously differentiable and apply the Ito formula:

$$
\nabla f(\boldsymbol{L}_t)-\nabla f(\boldsymbol{L}_0)=\int_0^t \nabla^2 f(\boldsymbol{L}_s)\,d\boldsymbol{L}_s+\int_0^t \Delta[\nabla f(\boldsymbol{L}_s)]\,ds.
$$

Let us check that $\|\Delta[\nabla f(\boldsymbol{x})]\|_2=\|\nabla[\Delta f(\boldsymbol{x})]\|_2\le M_2p$ for every $\boldsymbol{x}\in\mathbb{R}^p$. Indeed, let us introduce the function $g:\mathbb{R}^p\to\mathbb{R}$ defined by $g(\boldsymbol{x})=\Delta f(\boldsymbol{x})=\mathrm{tr}[\nabla^2 f(\boldsymbol{x})]$. The third item of condition F implies that $|g(\boldsymbol{x}+t\boldsymbol{u})-g(\boldsymbol{x})|\le pM_2|t|$ for every $t\in\mathbb{R}$ and every unit vector $\boldsymbol{u}\in\mathbb{R}^p$. Therefore, letting t go to zero, we get $|\boldsymbol{u}^\top\nabla g(\boldsymbol{x})|\le pM_2$ for every unit vector $\boldsymbol{u}$. Choosing $\boldsymbol{u}$ proportional to $\nabla g(\boldsymbol{x})$, we get the inequality $\|\nabla g(\boldsymbol{x})\|_2=\|\nabla[\Delta f(\boldsymbol{x})]\|_2\le pM_2$. This leads to

$$
\|\bar{\boldsymbol{V}}\|_{L_2}\le\int_0^h\int_0^t \big\|\nabla^2 f(\boldsymbol{L}_s)\nabla f(\boldsymbol{L}_s)-\Delta[\nabla f(\boldsymbol{L}_s)]\big\|_{L^2}\,ds\,dt
$$

$$
\begin{aligned}
&\le \int_0^h\int_0^t \big(M\big\|\nabla f(\boldsymbol{L}_s)\big\|_{L^2}+M_2p\big)\,ds\,dt\\
&= (1/2)(M^{3/2}\sqrt{p}+M_2p)h^2.
\end{aligned}
$$

This completes the proof of the lemma in the case of three times continuously differentiable functions f. If f is two-times differentiable with a second-order derivative satisfying the Lipschitz condition, then we can choose an arbitrarily small $\delta>0$ and apply the previous result to the smoothed function $f_\delta=f*\varphi_\delta$. Here, φ_δ denotes the density of the Gaussian distribution $\mathcal{N}_p(0,\delta^2\mathbf{I}_p)$ and "$*$" is the convolution operator. The formula $\nabla^2 f_\delta=(\nabla^2 f)*\varphi_\delta$ implies that f_δ satisfies the required smoothness assumptions with the same constants M and M_2 as the function f. Thus, defining $\bar{\boldsymbol{V}}_\delta$ in the same way as $\bar{\boldsymbol{V}}$

with f_δ instead of f, we get

$$\|\bar{\boldsymbol{V}}_\delta\|_{L_2} \leq (1/2)(M^{3/2}\sqrt{p} + M_2 p)h^2.$$

On the other hand, setting $g_\delta = f - f_\delta$, we get

$$\begin{aligned}
\|\bar{\boldsymbol{V}}_\delta - \bar{\boldsymbol{V}}\|_{L_2} &\leq \int_0^h \Big\| \nabla g_\delta(\boldsymbol{L}_t) - \nabla g_\delta(\boldsymbol{L}_0) - \sqrt{2}\int_0^t \nabla^2 g_\delta(\boldsymbol{L}_s) d\boldsymbol{W}_s \Big\|_{L^2} dt \\
&\leq \int_0^h \big\| \nabla g_\delta(\boldsymbol{L}_t) - \nabla g_\delta(\boldsymbol{L}_0) \big\|_{L^2} dt \\
&\quad + \sqrt{2p}\int_0^h \Big(\int_0^t \mathbf{E}\|\nabla^2 g_\delta(\boldsymbol{L}_s)\|^2 ds \Big)^{1/2} dt.
\end{aligned}$$

Using the Lipschitz continuity of ∇f and $\nabla^2 f$, one easily checks that

$$\begin{aligned}
\|\nabla g_\delta(\boldsymbol{x})\|_2 &\leq \int_{\mathbb{R}^p} \|\nabla f(\boldsymbol{x}-\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2 \varphi_\delta(\boldsymbol{y})\, d\boldsymbol{y} \\
&\leq M \int_{\mathbb{R}^p} \|\boldsymbol{y}\|_2 \varphi_\delta(\boldsymbol{y})\, d\boldsymbol{y} \leq M\delta\sqrt{p}, \\
\|\nabla^2 g_\delta(\boldsymbol{x})\| &\leq \int_{\mathbb{R}^p} \|\nabla^2 f(\boldsymbol{x}-\boldsymbol{y}) - \nabla^2 f(\boldsymbol{x})\| \varphi_\delta(\boldsymbol{y})\, d\boldsymbol{y} \\
&\leq M_2 \int_{\mathbb{R}^p} \|\boldsymbol{y}\|_2 \varphi_\delta(\boldsymbol{y})\, d\boldsymbol{y} \leq M_2\delta\sqrt{p}.
\end{aligned}$$

This implies that the limit, when δ tends to zero, of $\|\bar{\boldsymbol{V}}_\delta - \bar{\boldsymbol{V}}\|_{L_2}$ is equal to zero. As a consequence,

$$\begin{aligned}
\|\bar{\boldsymbol{V}}\|_{L_2} &\leq \lim_{\delta\to 0} \big(\|\bar{\boldsymbol{V}}_\delta\|_{L_2} + \|\bar{\boldsymbol{V}}_\delta - \bar{\boldsymbol{V}}\|_{L_2} \big) \\
&\leq (1/2)(M^{3/2}\sqrt{p} + M_2 p)h^2 + \lim_{\delta\to 0} \|\bar{\boldsymbol{V}}_\delta - \bar{\boldsymbol{V}}\|_{L_2} \\
&\leq (1/2)(M^{3/2}\sqrt{p} + M_2 p)h^2.
\end{aligned}$$

This completes the proof of the lemma. □

Lemma 7. *Let A, B and C be given non-negative numbers such that $A \in (0,1)$. Assume that the sequence of non-negative numbers $\{x_k\}_{k\in\mathbb{N}}$ satisfies the recursive inequality*

$$x_{k+1}^2 \leq [(1-A)x_k + C]^2 + B^2$$

for every integer $k \geq 0$. Let us denote E and D respectively by

$$E = \frac{(1-A)C + \big\{C^2 + (2A - A^2)B^2\big\}^{1/2}}{2A - A^2} \geq \frac{(1-A)C}{A(2-A)} + \frac{B}{\sqrt{A(2-A)}}$$

and

$$D = \left\{[(1-A)E + C]^2 + B^2\right\}^{1/2} - (1-A)E \le C + \frac{B^2 A}{C + \sqrt{A(2-A)}\,B}.$$

Then

$$x_k \le (1-A)^k x_0 + \frac{D}{A} \le (1-A)^k x_0 + \frac{C}{A} + \frac{B^2}{C + \sqrt{A(2-A)}\,B} \tag{2.46}$$

for all integers $k \ge 0$.

Proof. We will repeatedly use the fact that $D = EA$. Let us introduce the sequence y_k defined as follows: $y_0 = x_0 + E$ and

$$y_{k+1} = (1-A)y_k + D, \quad k = 0, 1, 2, \ldots$$

We will first show that $y_k \ge x_k \vee E$ for every $k \ge 0$. This can be done by mathematical induction. For $k = 0$, this claim directly follows from the definition of y_0. Assume that for some k, we have $x_k \le y_k$ and $y_k \ge E$. Then, for $k+1$, we have

$$\begin{aligned}
x_{k+1} &\le \left([(1-A)x_k + C]^2 + B^2\right)^{1/2} \\
&\le \left([(1-A)y_k + C]^2 + B^2\right)^{1/2} \\
&= (1-A)y_k + \left([(1-A)y_k + C]^2 + B^2\right)^{1/2} - (1-A)y_k \\
&\le (1-A)y_k + \left([(1-A)E + C]^2 + B^2\right)^{1/2} - (1-A)E = y_{k+1}
\end{aligned}$$

and, since $D = EA$, $y_{k+1} = (1-A)y_k + D \ge (1-A)E + EA = E$. Thus, we have checked that the sequence x_k is dominated by the sequence y_k. It remains to establish an upper bound on y_k. This is an easy task since y_k satisfies a first-order linear recurrence relation. We get

$$\begin{aligned}
y_k &= (1-A)^{k-1} y_1 + \sum_{j=0}^{k-2} (1-A)^j D \\
&= (1-A)^{k-1}\Big(x_1 + \frac{D}{A}\Big) + \frac{D}{A}\big(1 - (1-A)^{k-1}\big) \\
&= (1-A)^{k-1} x_1 + \frac{D}{A}.
\end{aligned}$$

This completes the proof of (2.46). □

Proof of Lemma 5. Let us introduce the $\mathbb{R}^p$-valued random process $\boldsymbol{v}_t = -\exp\left\{\int_0^t \mathbf{H}_u\,du\right\}\int_0^t \mathbf{H}_s \boldsymbol{x}_s\,ds$.

The time derivative of this process satisfies

$$\boldsymbol{v}'_t = -\exp\Big\{\int_0^t \mathbf{H}_u\,du\Big\}\mathbf{H}_t\boldsymbol{\alpha}_t.$$

This implies that $\boldsymbol{v}_t = -\int_0^t \exp\{\int_0^s \mathbf{H}_u\,du\}\mathbf{H}_s\boldsymbol{\alpha}_s\,ds$. Using the definition of $\boldsymbol{v}_t$, we can check that $\int_0^t \mathbf{H}_s\boldsymbol{x}_s\,ds = -\exp\{-\int_0^t \mathbf{H}_u\,du\}\boldsymbol{v}_t = \int_0^t \exp\{-\int_s^t \mathbf{H}_u\,du\}\mathbf{H}_s\boldsymbol{\alpha}_s\,ds$. Substituting this in (2.29), we get

$$\boldsymbol{x}_t = \boldsymbol{\alpha}_t - \int_0^t \exp\{-\int_s^t \mathbf{H}_u\,du\}\mathbf{H}_s\boldsymbol{\alpha}_s\,ds. \tag{2.47}$$

On the other hand—using the notation $\mathbf{M}_t = \exp\{\int_0^t \mathbf{H}_u\,du\}$ and the integration by parts formula for semi-martingales—the second integral on the right hand side of (2.30) can be modified as follows:

$$\begin{aligned}
\int_0^t \exp\Big\{-\int_s^t \mathbf{H}_u\,du\Big\}d\boldsymbol{\alpha}_s &= \mathbf{M}_t^{-1}\int_0^t \mathbf{M}_s d\boldsymbol{\alpha}_s \\
&= \mathbf{M}_t^{-1}\Big(\mathbf{M}_t\boldsymbol{\alpha}_t - \mathbf{M}_0\boldsymbol{\alpha}_0 - \int_0^t d\mathbf{M}_s\,\boldsymbol{\alpha}_s\Big) \\
&= \boldsymbol{\alpha}_t - \exp\Big\{-\int_0^t \mathbf{H}_u\,du\Big\}\boldsymbol{\alpha}_0 \\
&\quad - \int_0^t \exp\Big\{-\int_s^t \mathbf{H}_u\,du\Big\}\mathbf{H}_s\boldsymbol{\alpha}_s\,ds.
\end{aligned}$$

Combining this equation with (2.47), we get the claim of the lemma. □

Lemma 8. *Let A and B be given positive numbers and $\{C_k\}_{k\in\mathbb{N}}$ be a given sequence of real numbers. Assume that the sequence $\{x_k\}_{k\in\mathbb{N}}$ satisfies the recursive inequality*

$$x_{k+1} \le Ax_k^2 + 2Bx_k + C_k, \qquad \forall k \in \mathbb{N}.$$

Then, for all $k \in \mathbb{N}$,

$$x_k \le \frac{1}{A}\big(Ax_0 + B\big)^{2^k}\exp\Big\{\sum_{j=0}^{k-1} 2^{k-1-j}\,\frac{AC_j + B(1-B)}{(Ax_0+B)^{2^{j+1}}}\Big\}.$$

Proof. Let us introduce the sequences $\{y_k\}_{k\in\mathbb{N}}$ and $\{z_k\}_{k\in\mathbb{N}}$ defined by the relations $y_0 = x_0$,

$$y_{k+1} = Ay_k^2 + 2By_k + C_k$$

and

$$z_k = (Ax_0 + B)^{2^k} \exp\left\{ \sum_{j=0}^{k-1} 2^{k-1-j} \frac{AC_j + B(1-B)}{(Ax_0 + B)^{2^{j+1}}} \right\}.$$

Using mathematical induction, one easily shows that inequalities

$$x_k \leq y_k \qquad \text{and} \qquad (Ax_0 + B)^{2^k} \leq Ay_k + B \leq z_k$$

hold for every $k \in \mathbb{N}$. As a consequence, we get

$$x_k \leq \frac{Ax_k + B}{A} \leq \frac{Ay_k + B}{A} \leq \frac{z_k}{A}.$$

This completes the proof of the lemma. □

Lemma 9. *Let A, B, C, D be positive numbers satisfying $D < A < 1$ and $\{x_k\}_{k\in\mathbb{N}}$ be a sequence of positive numbers satisfying the inequality*

$$x_{k+1} \leq \left((1-A)^2 x_k^2 + B^2\right)^{1/2} + C + Dx_k.$$

Then, for every $k \geq 0$, we have

$$x_k \leq (1-A+D)^k x_0 + \frac{C}{A-D} + \frac{B}{\sqrt{(A-D)(2-A-D)}}.$$

Proof. We start by setting

$$E = \frac{B}{\sqrt{(A-D)(2-A-D)}}, \qquad F = C + (A-D)E$$

and by defining a new sequence $\{y_k\}_{k\in\mathbb{N}}$ by $y_0 = x_0 + E$ and

$$y_{k+1} = (1-A+D)y_k + F.$$

Our goal is to prove that $y_k \geq x_k \vee E$ for every k. This claim is clearly true for $k = 0$. Let us assume that it is true for the value k and prove its validity for $k+1$. Since the function $x \mapsto \sqrt{x^2 + a^2} - x$ is decreasing, we have

$$\begin{aligned}
x_{k+1} &\leq \sqrt{(1-A)^2 y_k^2 + B^2} + C + Dy_k \\
&\leq (1-A+D)y_k + C + \sqrt{(1-A)^2 y_k^2 + B^2} - (1-A)y_k \\
&\leq (1-A+D)y_k + C + \sqrt{(1-A)^2 E^2 + B^2} - (1-A)E = y_{k+1}.
\end{aligned}$$

On the other hand,

$$\begin{aligned} y_{k+1} &\geq (1 - A + D)y_k + (A - D)E \\ &\geq (1 - A + D)E + (A - D)E = E. \end{aligned}$$

This implies, in particular, that $x_k \leq y_k$ for every $k \in \mathbb{N}$. Since $\{y_k\}$ satisfies a first-order linear recursion, we get $y_k = (1 - A + D)^k y_0 + F(1 - (1 - A + D)^k)/(A - D)$. $\square$

Chapter 3

Langevin Monte-Carlo with convex potentials

Abstract

In this paper, we provide non-asymptotic upper bounds on the error of sampling from a target density using three schemes of discretized Langevin diffusions. The first scheme is the Langevin Monte Carlo (LMC) algorithm, the Euler discretization of the Langevin diffusion. The second and the third schemes are, respectively, the kinetic Langevin Monte Carlo (KLMC) for differentiable potentials and the kinetic Langevin Monte Carlo for twice-differentiable potentials (KLMC2). The main focus is on the target densities that are smooth and log-concave on $\mathbb{R}^p$, but not necessarily strongly log-concave. Bounds on the computational complexity are obtained under two types of smoothness assumption: the potential has a Lipschitz-continuous gradient and the potential has a Lipschitz-continuous Hessian matrix. The error of sampling is measured by Wasserstein-q distances. We advocate for the use of a new dimension-adapted scaling in the definition of the computational complexity, when Wasserstein-q distances are considered. The obtained results show that the number of iterations to achieve a scaled-error smaller than a prescribed value depends only polynomially in the dimension.
This chapter is based on a joint work with Arnak Dalalyan and Lionel Riou-Durand entitled “Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets”. It is submitted to *Journal of Machine Learning Research*.

3.1 Introduction

The two most popular techniques for defining estimators or predictors in statistics and machine learning are the M estimation, also known as empirical risk minimization, and the Bayesian method (leading to posterior mean, posterior median, etc.). In practice, it is necessary to devise a numerical method for computing an approximation of these estimators. Optimization algorithms are used for approximating an M-estimator, while Monte Carlo algorithms are employed for approximating Bayesian estimators. In statistical learning theory, over past decades, a concentrated effort was made for getting non asymptotic guarantees on the error of an optimization algorithm. For smooth optimization, sharp results were obtained in the case of strongly convex and convex cases [Bub15], the case of non-convex smooth optimization being much more delicate [JK17]. As for Monte Carlo algorithms, past three years or so witnessed considerable progress on theory of sampling from strongly log-concave densities. Some results for non strongly convex densities were obtained as well. However, to the best of our knowledge, there is no paper providing a systematic account on the error bounds for sampling from non strongly concave densities. The main goal of this paper is to fill this gap.

A good starting point for accomplishing the aforementioned task is perhaps a result from [DMM19] for the sampling error measured by the Kullback-Leibler divergence. The result is established for the Langevin Monte Carlo (LMC) algorithm, which is the "sampling analogue" of the gradient descent. Let $\pi : \mathbb{R}^p \to [0, +\infty)$ be a probability density function (with respect to Lebesgue's measure) given by

$$\pi(\boldsymbol{\theta}) = \frac{e^{-f(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f(\boldsymbol{v})} d\boldsymbol{v}}.$$

for a potential function f. The goal of sampling is to generate a random vector in $\mathbb{R}^p$ having a distribution close to the target distribution defined by π. In the sequel, we will make repeated use of the moments $\mu_k(\pi) = \mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\|\boldsymbol{\vartheta}\|_2^k]$, where $\|\boldsymbol{v}\|_q = (\sum_j |v_j|^q)^{1/q}$ is the usual ℓ_q-norm for any $q \geq 1$. When there is no risk of confusion, we will write μ_k instead of $\mu_k(\pi)$.

To define the LMC algorithm, we need a sequence of positive parameters $\boldsymbol{h} = \{h_k\}_{k\in\mathbb{N}}$, referred to as the step-sizes and an initial point $\boldsymbol{\vartheta}_{0,\boldsymbol{h}} \in \mathbb{R}^p$ that may be deterministic or random. The successive iterations of the LMC algorithm are given by the update rule

$$\boldsymbol{\vartheta}_{k+1,\boldsymbol{h}} = \boldsymbol{\vartheta}_{k,\boldsymbol{h}} - h_{k+1}\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}}) + \sqrt{2h_{k+1}}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{3.1}$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k, \ldots$ is a sequence of independent, and independent of $\boldsymbol{\vartheta}_{0,\boldsymbol{h}}$, centered Gaussian vectors with identity covariance matrices. Let ν_K denote the distribution of the

K-th iterate of the LMC algorithm, assuming that all the step-sizes are equal ($h_k = h$ for every $k \in \mathbb{N}$) and the initial point is $\boldsymbol{\vartheta}_{0,h} = \mathbf{0}_p$. We will also define the distribution $\bar\nu_K = (1/K)\sum_{k=1}^K \nu_k$, obtained by choosing uniformly at random one of the elements of the sequence $\{\boldsymbol{\vartheta}_{1,\boldsymbol{h}}, \ldots, \boldsymbol{\vartheta}_{K,\boldsymbol{h}}\}$. It is proved in [DMM19, Cor. 7] that if the gradient ∇f is Lipschitz continuous with the Lipschitz constant M, then for every $K \in \mathbb{N}$, the Kullback-Leibler divergence between $\bar\nu_K$ and π satisfies

$$D_{\rm KL}(\bar\nu_K \| \pi) \le \frac{\mu_2(\pi)}{2Kh} + Mph, \qquad D_{\rm KL}(\bar\nu_K^{\rm opt} \| \pi) \le \sqrt{\frac{2Mp\mu_2(\pi)}{K}}.$$

Note that the second inequality above is obtained from the first one by using the step-size $h_{\rm opt} = (2KMp/\mu_2(\pi))^{-1/2}$ obtained by minimizing the right hand side of the first inequality. Therefore, if we assume that the second order moment μ_2 of π satisfies the condition $M\mu_2 \le \kappa p^\beta$, for some dimension-free positive constants β and κ, we get

$$D_{\rm KL}(\bar\nu_K^{\rm opt} \| \pi) \le \sqrt{\frac{2\kappa p^{1+\beta}}{K}}.$$

A natural measure of complexity of the LMC with averaging is, for every $\varepsilon > 0$, the number of gradient evaluations that is sufficient for getting a sampling error bounded from above by ε. From the last display, taking into account the Pinsker inequality, $d_{\rm TV}(\bar\nu_K, \pi) \le \sqrt{D_{\rm KL}(\bar\nu_K, \pi)/2}$ and the fact that each iterate of the LMC requires one evaluation of the gradient of f, we obtain the following result. The number of gradient evaluations $K_{\rm LMCa,TV}(p,\varepsilon)$ sufficient for the total-variation-error of the LMC with averaging (hereafter, LMCa) to be smaller than ε is

$$K_{\rm LMCa,TV}(p,\varepsilon) = \frac{\kappa p^{1+\beta}}{2\varepsilon^4}.$$

The main goal of the present work is to provide this type of bounds on the complexity of various versions of the Langevin algorithm under different measures of the quality of sampling. The most important feature that we wish to uncover is the explicit dependence of the complexity $K(\varepsilon)$ on the dimension p, the inverse-target-precision $1/\varepsilon$ and the parameter κ. We will focus only on those measures of quality of sampling that can be directly used for evaluating the quality of approximating expectations.

3.2 Further precisions on the analyzed methods

Since our main motivation for considering the sampling problem comes from applications in statistics and machine learning, we will focus on the Monge-Kantorovich-Wasserstein

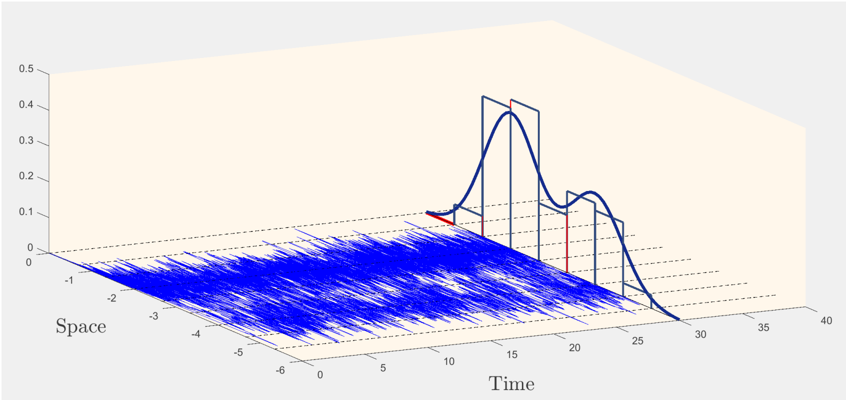


Figure 3.1: Illustration of Langevin dynamics. The blue lines represent different paths of a Langevin process. We see that the histogram of the state at time $t = 30$ is close to the target density (the dark blue line).

distances W_q defined by

$$W_q(\nu, \nu') = \inf\Big\{\mathbf{E}[\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2^q]^{1/q} : \boldsymbol{\vartheta} \sim \nu \text{ and } \boldsymbol{\vartheta}' \sim \nu'\Big\}, \qquad q \geq 1.$$

The infimum above is over all the couplings between ν and ν'. In view of the Hölder inequality, the mapping $q \mapsto W_q(\nu, \nu')$ is increasing for every pair (ν, ν').

Our main contributions are upper bounds on quantities of the form $W_q(\nu_K, \pi)$where π is a log-concave target distribution and ν_K is the distribution of the Kth iterate of various discretization schemes of Langevin diffusions. More precisely, we consider two types of Langevin processes: the kinetic Langevin diffusion and the vanilla Langevin diffusion. The latter is the highly overdamped version of the former, see [Nel67]. The Langevin diffusion, having π as invariant distribution, is defined as a solution[1] to the stochastic differential equation

$$d\boldsymbol{L}_t^{\mathsf{LD}} = -\nabla f(\boldsymbol{L}_t^{\mathsf{LD}})\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \qquad t \geq 0, \tag{3.2}$$

where $\boldsymbol{W}$ is a p-dimensional standard Brownian motion independent of the initial value $\boldsymbol{L}_0$. An illustration of this process is given in Figure 3.1. The LMC algorithm presented in (3.1) is merely the Euler-Maruyama discretization of the process $\boldsymbol{L}$. The kinetic Langevin diffusion $\{\boldsymbol{L}_t^{\mathsf{KLD}} : t \geq 0\}$, also known as the second-order Langevin process, is defined by

$$d\begin{bmatrix}\boldsymbol{V}_t\\ \boldsymbol{L}_t^{\mathsf{KLD}}\end{bmatrix} = \begin{bmatrix}-(\gamma\boldsymbol{V}_t + \nabla f(\boldsymbol{L}_t^{\mathsf{KLD}}))\\ \boldsymbol{V}_t\end{bmatrix} dt + \sqrt{2\gamma}\begin{bmatrix}\mathbf{I}_p\\ \mathbf{0}_{p\times p}\end{bmatrix} d\boldsymbol{W}_t, \qquad t \geq 0, \tag{3.3}$$

[1]Under the conditions imposed on the function f throughout this paper, namely the convexity and the Lipschitzness of the gradient, all the considered stochastic differential equations have unique strong solutions. Furthermore, all conditions (see, for instance, [Pav14]) ensuring that π and p^* are invariant densities of, respectively, processes (3.2) and (3.3) are fulfilled.

where $\gamma > 0$ is the friction coefficient. The process $\boldsymbol{V}_t$ is often called the velocity process since the second row in (3.3) implies that $\boldsymbol{V}_t$ is the time derivative of $\boldsymbol{L}_t^{\mathsf{KLD}}$. The continuous-time Markov process $(\boldsymbol{L}_t^{\mathsf{KLD}}, \boldsymbol{V}_t)$ is positive recurrent and has a unique invariant distribution, which has the following density with respect to the Lebesgue measure on $\mathbb{R}^{2p}$:

$$p_*(\boldsymbol{\theta}, \boldsymbol{v}) \propto \exp\Big\{ - f(\boldsymbol{\theta}) - \frac{1}{2}\|\boldsymbol{v}\|_2^2 \Big\}, \qquad \boldsymbol{\theta} \in \mathbb{R}^p,\ \boldsymbol{v} \in \mathbb{R}^p.$$

If $(\boldsymbol{L}, \boldsymbol{V})$ is a pair of random vectors drawn from the joint density p_*, then $\boldsymbol{L}$ and $\boldsymbol{V}$ are independent, $\boldsymbol{L}$ is distributed according to the target π, whereas $\boldsymbol{V}$ is a standard Gaussian vector. Therefore, at equilibrium, the random variable $\boldsymbol{L}_t^{\mathsf{KLD}}$ has the target distribution π.

Time-discretized versions of Langevin diffusion processes (3.2) and (3.3) are used for (approxi-mate) sampling from π. In order to guarantee that the discretization error is not too large, as well as that the process $\{\boldsymbol{L}_t\}$ converges fast enough to its invariant distribution, we need to impose some assumptions on f. In the present work, we will assume that either Conditions 1, 2 or Conditions 1, 2, 3 presented below are satisfied.

Condition 1. *The function f is continuously differentiable on $\mathbb{R}^p$ and its gradient ∇f is M-Lipschitz for some $M > 0$: $\|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \le M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$.*

From now on, we will always assume that the Langevin (vanilla or kinetic) diffusion under consideration has the initial point $\boldsymbol{L}_0 = \mathbf{0}$. Some of the conditions presented below implicitly require that this initialization is not too far away from the "center" of the target distribution π. In many statistical problems where π is the Bayesian posterior, one can come close to these assumptions by shifting the distribution using a simple initial estimator.

Condition 2. *The function f is convex on $\mathbb{R}^p$. Furthermore, for some positive constants D and β, we have $\mu_2(\pi) = \mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\|\boldsymbol{\vartheta}\|_2^2] \le Dp^\beta$.*

Under Condition 2, the centered second moment of π scales polynomially with the dimension with power $\beta > 0$, while the flatness of the distribution is controlled by the parameter $D > 0$. Remarkably, Condition 2 implies that all the moments $\{\mu_q(\pi)\}_{q\ge1}$ scale polynomially with p, provided that $\|\mathbf{E}_{\vartheta\sim\pi}[\boldsymbol{\vartheta}]\|_2$ also does. This fact is a consequence of Borell's lemma [GBVV14, Theorem 2.4.6], which states that for any $q \ge 1$, there is a numerical constant B_q that depends only on q such that $\mu_q(\pi)^{1/q} \le B_q\mu_2(\pi)^{1/2}$. An attempt to provide optimized constants in this inequality is stated in Lemma 14.

In the sequel, we show that the smoothness and the flatness of π have a combined impact on the sampling error considered. It turns out that the important parameter with respect to the hardness of the sampling problem is the product

$$\kappa := MD.$$

For m-strongly convex functions f, Condition 2 is satisfied with $D = 1/m$ and $\beta = 1$, according to Brascamp-Lieb inequality [BL76]. In this case the parameter $\kappa = M/m$ is known as the condition number. We will show that Condition 2 is also satisfied for functions f that are convex everywhere and strongly convex inside a ball, as well as for functions f that are convex everywhere and strongly convex only outside a ball.

In the next assumption, we use notation $\|\mathbf{M}\|$ for the spectral norm (the largest singular value) of a matrix $\mathbf{M}$.

Condition 3. *The function f is twice differentiable in $\mathbb{R}^p$ with a M_2-Lipschitz Hessian $\nabla^2 f$ for some $M_2 > 0$: $\|\nabla^2 f(\boldsymbol{\theta}) - \nabla^2 f(\boldsymbol{\theta}')\| \leq M_2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$.*

Condition 3 ensures further smoothness of the potential f. When it holds, the Lipschitz continuity of the Hessian and the flatness of π also have a combined impact on the sampling error. A second important parameter with respect to the hardness of the sampling problem in such a case is the product

$$\kappa_2 := M_2^{2/3} D.$$

The case of an m-strongly convex function f has been studied in several recent papers. As a matter of fact, global strong convexity implies exponentially fast mixing of processes (3.2) and (3.3), with dimension-free rates e^{-mt} and $e^{-mt/(M+m)^{1/2}}$, respectively. When only simple convexity is assumed, such results do not hold in general. Therefore, the strategy we adopt here consists in sampling from a distribution that is provably close to the target, but has the advantage of being strongly log-concave.

More precisely, for some small positive α, the surrogate potential is defined by $f_\alpha(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + \alpha\|\boldsymbol{\theta}\|_2^2/2$. Therefore, the corresponding surrogate distribution has the density

$$\pi_\alpha(\boldsymbol{\theta}) := \frac{e^{-f_\alpha(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f_\alpha(\boldsymbol{v})} d\boldsymbol{v}}.$$

We stress the fact that the quadratic penalty $\alpha\|\boldsymbol{\theta}\|_2^2/2$ added to the potential f is centered at the origin. This is closely related to the fact that the diffusion is assumed to have the origin as initial point, and also to the fact that the origin is assumed here to be a good guess of the "center" of π. The parameter α, together with the step-size $\boldsymbol{h}$, is considered as a tuning parameter of the algorithms to be calibrated. Large values of α will result in fast convergence to π_α but a poor approximation of π by π_α. On the other hand, smaller values of α will lead to a small approximation error but also slow convergence. The next result quantifies the approximation of π by π_α, for different distances.

Proposition 12. *For any $\alpha \geq 0$ and $q \in [1, +\infty)$, there is a numerical constant C_q depending only on q such that*

$$d_{\mathrm{TV}}(\pi, \pi_\alpha) \leq \alpha\mu_2(\pi)$$
$$W_q^q(\pi, \pi_\alpha) \leq C_q \alpha \mu_2(\pi)^{(q+2)/2}.$$

Here the constant C_q can be bounded for every q. In particular, $C_1 \leq 22$ and $C_2 \leq 111$.

This result allows us to control the bias induced by replacing the target distribution by the surrogate one and paves the way for choosing the "optimal" α by minimizing an upper bound on the sampling error. We draw the attention of the reader to the fact that, for W_q distance, the dependence on α of the upper bound is $\alpha^{1/q}$, which slows down when q increases (recall that α is a small parameter). This explains a deterioration with increasing q of the complexity bounds presented in forthcoming sections. In the rest of the paper, we define the constant

$$C_q = \inf\{C : W_q^q(\pi, \pi_\alpha) \leq C\alpha\mu_2(\pi)^{(q+2)/2},\ \forall\, \pi \text{ log-concave}\}. \tag{3.4}$$

This constant will repeatedly appear in the statements of the theorems.

3.3 How to measure the complexity of a sampling scheme?

We have already introduced the notation $K_{\mathsf{Alg},\mathsf{Crit}}(p, \varepsilon)$, the number of iterations that guarantee that algorithm Alg has an error—measured by criterion Crit—smaller than ε. If we choose a criterion, this quantity can be used to compare two methods, the iterates of which have comparable computational complexity. For example, LMC and KLMC being discretized versions of the Langevin process (3.2) and the kinetic Langevin process (3.3), respectively, are such that one iteration requires one evaluation of ∇f and generation of one realization of a Gaussian vector of dimension p or $2p$. Thus, the iterations are of comparable computational complexity and, therefore, it is natural to prefer LMC if $2K_{\mathrm{LMC},\mathsf{Crit}}(p, \varepsilon) \leq K_{\mathrm{KLMC},\mathsf{Crit}}(p, \varepsilon)$ and to prefer KLMC if the opposite inequality is true.

A delicate question that has not really been discussed in literature is a notion of complexity that allows to compare the quality of a given sampling method for two different criteria. To be more precise, assume that we are interested in the LMC algorithm and wish to figure out whether it is "more difficult" to perform approximate sampling for the TV-distance or for the Wasserstein distance. It is a well-known fact that the TV-distance induces the uniform strong convergence of measures whereas the Wasserstein distances induce the weak convergence. Therefore, at least intuitively, approximate sampling for the

TV-distance should be harder than approximate sampling for the Wasserstein distance[2]. However, under Condition 1 and m-strong convexity of f, the available results for the LMC provide the same order of magnitude, p/ε^2, both for $K_{\mathrm{LMC,TV}}$ [Dal17b, DM19] and K_{LMC,W_2} [DM19, DT09]. The point we want to put forward is that the origin of this discrepancy between the intuitions and mathematical results is the inappropriate scaling of the target accuracy in the definition of K_{LMC,W_2}.

To further justify the importance of choosing the right scaling of the target accuracy, let us make the following observation. The total-variation distance serves to approximate probabilities, which are adimensional and scale-free quantities belonging to the interval $[0, 1]$. The Wasserstein distances are useful for approximating moments[3] which depend on both the dimension and the scale. For this reason, we suggest the following definition of the analogue of K in the case of Wasserstein distances:

$$K_{\mathsf{Alg},W_q}(p,\varepsilon) = \min\{k \in \mathbb{N} : W_q(\nu_k^{\mathsf{Alg}}, \pi) \le \varepsilon\sqrt{\mu_2(\pi)},\ \forall \pi \in \mathscr{P}\}, \tag{3.5}$$

where Alg is a Markov Chain Monte Carlo or another method of sampling, k is generally the number of calls to the oracle and $\mathscr{P}$ is a class of target distributions. Examples of oracle call are the evaluation of the gradient of the potential at a given point or the computation of the product of the Hessian of f at a given point and a given vector. Note also that $\sqrt{\mu_2(\pi)}$ is the W_2 distance between the Dirac mass at the origin and the target distribution.

Definition (3.5), as opposed to those used in prior work, has the advantage of being scale invariant and reflecting the fact that we deal with objects that might be large if the dimension is large. Note that the idea of scaling the error in order to make the complexity measure scale-invariant has been recently used in [TSV18] as well. Indeed, in the context of m-strongly log-concave distributions, [TSV18] propose to find the smallest k such that $W_2(\nu_k^{\mathsf{Alg}}, \pi) \le \varepsilon/\sqrt{m}$. This is close to our proposal, since in the case of m-strongly log-concave distributions, it follows from the Brascamp-Lieb inequality that $\sup_\pi \sqrt{\mu_2(\pi)} = \sqrt{p/m}$ (the sup is attained for Gaussian distributions).

3.4 Overview of main contributions

In this work, we analyze three methods, LMC, KLMC [CCBJ18] and KLMC2 [DRD20], applied to the strong-convexified potential $f_\alpha(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + (\alpha/2)\|\boldsymbol{\theta}\|_2^2$ in order to cope with

[2] We underline here that the aforementioned hardness argument is based only on the topological argument, since it is not possible, in general, to upper bound the Wasserstein distance W_q, for $q \ge 1$ by the TV-distance or a function of it.

[3] Recall that by the triangle inequality, one has $(\mathbf{E}_{\boldsymbol{\vartheta}\sim\nu}[\|\boldsymbol{\vartheta}\|_2^q])^{1/q} - (\mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\|\boldsymbol{\vartheta}\|_2^q])^{1/q} \le W_q(\nu, \pi)$.

the lack of strong convexity. We briefly recall these algorithms and present a summary of the main contributions of this work.

3.4.1 Considered Markov chain Monte-Carlo methods

We first recall the definition of the Langevin Monte Carlo algorithms. For the LMC algorithm introduced in (3.1), we will only use the constant step-size form, the update rule of which is given by

$$\boldsymbol{\vartheta}_{k+1} = (1-\alpha h)\boldsymbol{\vartheta}_k - h\nabla f(\boldsymbol{\vartheta}_k) + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{α-LMC}$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k, \ldots$ is a sequence of mutually independent, independent of $\boldsymbol{\vartheta}_0$, centered Gaussian vectors with covariance matrices equal to identity. We will refer to this version of the LMC algorithm as α-LMC.

We now recall the definition of the first and second-order Kinetic Langevin Monte Carlo algorithms. We suppose that, for some initial distribution ν_0 chosen by the user, both KLMC and KLMC2 algorithms start from $(\boldsymbol{v}_0, \boldsymbol{\vartheta}_0) \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) \otimes \nu_0$. Before stating the update rules, we specify the structure of the random perturbation generated at each step. In what follows, $\{(\boldsymbol{\xi}_k^{(1)}, \boldsymbol{\xi}_k^{(2)}, \boldsymbol{\xi}_k^{(3)}, \boldsymbol{\xi}_k^{(4)}) : k \in \mathbb{N}\}$ will stand for a sequence of iid $4p$-dimensional centered Gaussian vectors, independent of the initial condition $(\boldsymbol{v}_0, \boldsymbol{\vartheta}_0)$.

To specify the covariance structure of these Gaussian variables, we define two sequences of functions (ψ_k) and (φ_k) as follows. For every $t > 0$, let $\psi_0(t) = e^{-\gamma t}$, then for every $k \in \mathbb{N}$, define $\psi_{k+1}(t) = \int_0^t \psi_k(s)\,ds$ and $\varphi_{k+1}(t) = \int_0^t e^{-\gamma(t-s)}\psi_k(s)\,ds$. Now, let us denote by $\xi_{k,j}$ for the j-th component of the vector $\boldsymbol{\xi}_k$ (a scalar), and assume that for any fixed k, the 4-dimensional random vectors $\{(\xi_{k,j}^{(1)}, \xi_{k,j}^{(2)}, \xi_{k,j}^{(3)}, \xi_{k,j}^{(4)}) : 1 \le j \le p\}$ are iid with the covariance matrix

$$\mathbf{C}_{h,\gamma} = \int_0^h [\psi_0(t);\ \psi_1(t);\ \varphi_2(t);\ \varphi_3(t)]^\top [\psi_0(t);\ \psi_1(t);\ \varphi_2(t);\ \varphi_3(t)]\,dt.$$

The KLMC algorithm, introduced by [CCBJ18], is a sampler derived from a suitable time-discretization of the kinetic diffusion. When applied to the strong-convexified potential f_α, for a step-size $h > 0$, its update rule reads as follows

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)(\nabla f(\boldsymbol{\vartheta}_k) + \alpha\boldsymbol{\vartheta}_k) \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)(\nabla f(\boldsymbol{\vartheta}_k) + \alpha\boldsymbol{\vartheta}_k) \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \boldsymbol{\xi}_{k+1}^{(1)} \\ \boldsymbol{\xi}_{k+1}^{(2)} \end{bmatrix}. \tag{α-KLMC}$$

Roughly speaking, this formula is obtained from (3.3) by replacing the function $t \mapsto \nabla f(\boldsymbol{L}_t)$ by a piecewise constant approximation. Such an approximation is made possible by the fact that f is gradient-Lipschitz.

It is natural to expect that further smoothness of f may allow one to improve upon the aforementioned piecewise constant approximation. This is done by the KLMC2 algorithm, introduced by [DRD20], which takes advantage of the existence and smoothness of the Hessian of f in order to use a local-linear approximation. At any iteration $k \in \mathbb{N}$ with a current value $\boldsymbol{\vartheta}_k$, define the gradient $\boldsymbol{g}_{k,\alpha} = \nabla f(\boldsymbol{\vartheta}_k) + \alpha\boldsymbol{\vartheta}_k$ and the Hessian $\mathbf{H}_{k,\alpha} = \nabla^2 f(\boldsymbol{\vartheta}_k) + \alpha\mathbf{I}_p$. When applied to the modified strongly convex potential f_α, for $h > 0$, the update rule of the KLMC2 algorithm is

$$\begin{bmatrix}\boldsymbol{v}_{k+1}\\ \boldsymbol{\vartheta}_{k+1}\end{bmatrix} = \begin{bmatrix}\psi_0(h)\boldsymbol{v}_k - \psi_1(h)\boldsymbol{g}_{k,\alpha} - \varphi_2(h)\mathbf{H}_{k,\alpha}\boldsymbol{v}_k\\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\boldsymbol{g}_{k,\alpha} - \varphi_3(h)\mathbf{H}_{k,\alpha}\boldsymbol{v}_k\end{bmatrix} + \sqrt{2\gamma}\begin{bmatrix}\boldsymbol{\xi}^{(1)}_{k+1} - \mathbf{H}_{k,\alpha}\boldsymbol{\xi}^{(3)}_{k+1}\\ \boldsymbol{\xi}^{(2)}_{k+1} - \mathbf{H}_{k,\alpha}\boldsymbol{\xi}^{(4)}_{k+1}\end{bmatrix}. \tag{α-KLMC2}$$

Notice that if we apply KLMC2 with $\mathbf{H}_{k,\alpha} = 0$, we recover the KLMC algorithm. These two algorithms, derived from the kinetic Langevin diffusion, will be referred to as α-KLMC and α-KLMC2.

3.4.2 Summary of the obtained complexity bounds

Without going into details here, we mention in the tables below the order of magnitude of the number of iterations required by different algorithms for getting an error bounded by ε for various metrics. For improved legibility, we do not include logarithmic factors and report the order of magnitude of $K_{\square,\square}(p,\varepsilon)$ in the case when the parameter β in Condition 2 is fixed to a particular value. We present hereafter the case where $\beta = 1$, which is of particular interest as discussed in Section 3.8. In this table, $\widetilde{\kappa} = \kappa_2 + \kappa p^{-1/3}$.

$\beta = 1$	LMCa	α-LMC		α-KLMC	α-KLMC2
Cond.	1-2	1-2	1-3	1-2	1-3
W_2	–	$\kappa p^2/\varepsilon^6$	$(\kappa_2^{1.5}p^{0.5} + \kappa^{1.5})p^2/\varepsilon^5$	$\kappa^{1.5}p^2/\varepsilon^5$	$\kappa^{0.5}\kappa_2{}^{1.5}p^2/\varepsilon^4$
W_1	–	$\kappa p^2/\varepsilon^4$	$(\kappa_2^{1.5}p^{0.5} + \kappa^{1.5})p^2/\varepsilon^3$	$\kappa^{1.5}p^2/\varepsilon^3$	$\kappa^{0.5}\kappa_2{}^{1.5}p^2/\varepsilon^2$
d_{TV}	p^2/ε^4 △	p^3/ε^4 □	–	–	–

The results indicated by △ describe the behavior of the Langevin Monte Carlo with averaging established in [DMM19]. To date, these results have the best known dependence (under conditions 1 and 2 only) on p. The results indicated by □ summarize the behavior of the Langevin Monte Carlo established in [Dal17b]. All the remaining cells of the table are filled in by the results obtained in the present work. One can observe that the results for W_1 are strictly better than those for W_2. Similar hierarchy was already reported in [MMS18, Remark 1.9]. It is also worth mentioning here, that using Metropolis-Hastings adjustment of the LMC (termed MALA), [DCWY18] obtained the complexity

$$K_{\text{MALA,TV}}(p,\varepsilon) = O\bigg(\frac{p^3\kappa^{3/2}}{\varepsilon^{3/2}}\log^{3/2}\left(p\kappa/\varepsilon\right)\bigg).$$

It is still an open question whether this type of result can be proved for Wasserstein distances.

We would also like to comment on the relation between the third and the fourth columns of the table, corresponding to the α-LMC algorithm under different sets of assumptions. The result for the more constrained Hessian-Lipschitz case is not always better than the result when only gradient Lipschitzness is assumed. For instance, for W_2, the latter is better than the former when $\kappa \lesssim (\kappa_2^{1.5}p^{0.5} + \kappa^{1.5})\varepsilon$, which is equivalent to $M \lesssim (M_2 p^{1/2} + M^{3/2})D^{1/2}\varepsilon$. At a very high level, this reflects the fact that when the condition number is large, the Hessian-Lipschitzness does not help to get an improved result.

3.4.3 The general approach based on a log-strongly-concave surrogate

We have already mentioned that the strategy we adopt here is the one described in [Dal17b], consisting of replacing the potential of the target density by a strongly convex surrogate. Prior to instantiating this approach to various sampling algorithms under various conditions and error measuring distances, we provide here a more formal description of it. In the remaining dist is a general distance on the set of all probability measures.

We will denote by $\nu_{k,\alpha}^{\mathsf{Alg}}$ the distribution of the random vector obtained after performing k iterations of the algorithm Alg with the surrogate potential $f_\alpha(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + \alpha\|\boldsymbol{\theta}\|_2^2/2$. Our first goal is to establish an upper bound on the distance between the sampling distribution $\nu_{k,\alpha}^{\mathsf{Alg}}$ and the target π. The methods we analyze here depend on the step-size h, as they are discretizations of continuous-time diffusion processes. Thus, the obtained bound will depend on h. This bound should be so that one can make it arbitrarily small by choosing small α and h and a large value of k. In a second stage, the goal is to exploit the obtained error-bounds in order to assess the order of magnitude of the computational complexity K, defined in Section 3.3, as a function of p, ε and the condition number κ.

To achieve this goal, we first use the triangle inequality

$$\mathsf{dist}(\nu_{k,\alpha}^{\mathsf{Alg}}, \pi) \leq \mathsf{dist}(\nu_{k,\alpha}^{\mathsf{Alg}}, \pi_\alpha) + \mathsf{dist}(\pi_\alpha, \pi).$$

Then, the second term of the right hand side of the last displayed equation is bounded using Proposition 12. Finally, the distance between the sampling density $\nu_{k,\alpha}^{\mathsf{Alg}}$ and the surrogate π_α is bounded using the prior work on sampling for log-strongly-concave distributions. Optimizing over α leads to the best bounds on precision and complexity.

3.5 Prior work

Mathematical analysis of MCMC methods defined as discretizations of diffusion processes is an active area of research since several decades. Important early references are [RT96a, RR98, RS02, DMR04] and the references therein. Although those papers do cover the multidimensional case, the guarantees they provide do not make explicit the dependence on the dimension. In a series of work analyzing ball walk and hit-and-run MCMCs, [LV06a, LV06b] put forward the importance of characterizing the dependence of the number of iterations on the dimension of the state space.

More recently, [Dal17b] advocated for analyzing MCMCs obtained from continuous time diffusion processes by decomposing the error into two terms: a non-stationarity error of the continuous-time process and a discretization error. A large number of works applied this kind of approach in various settings. [BEL18, DM17, DM19, DMP18] improved the results obtained by Dalalyan and extended them in many directions including non-smooth potentials and variable step-sizes. While previous work studied the sampling error measured by the total variation and Waserstein distances, [CB18] proved that similar results hold for the Kulback-Leibler divergence. [CCBJ18, CCA$^+$18, DRD20] investigated the case of a kinetic Langevin diffusion, showing that it leads to improved dependence on the dimension. A promising line of related research, initiated by [Wib18, Ber18], is to consider the sampling distributions as a gradient flow in a space of measures. The benefits of this approach were demonstrated in [DMM19, MCC$^+$19].

Motivated by applications in Statistics and Machine Learning, many recent papers developed theoretical guarantees for stochastic versions of algorithms, based on noisy gradients, see [BFFN19, CFM$^+$18, DK19, Dal17a, ZXG19, RRT17] and the references therein. A related topic is non-asymptotic guarantees for the Hamiltonian Monte Carlo (HMC). There is a growing literature on this in recent years, see [MS17, TSV18, CV19, MS19] and the references therein.

In all these results, the dependence of the number of iterations on the inverse precision is polynomial. [DCWY18, CDWY20, MV19a] proved that one can reduce this dependence to logarithmic by using Metropolis adjusted versions of the algorithms.

3.6 Precision and computational complexity of the LMC

In this section, we present non-asymptotic upper bounds in the non-strongly convex case for the suitably adapted LMC algorithm for Wasserstein and bounded-Lipschitz error measures under two sets of assumptions: Conditions 1-2 and Conditions 1-3. To refer to these settings, we will call them "Gradient-Lipschitz" and "Hessian-Lipschitz", respectively. The main goal is to provide a formal justification of the rates included in columns 2 and 3 of the table presented in Section 3.4.2. To ease notation, and since there is no risk of confusion, we write μ_2 instead $\mu_2(\pi)$.

3.6.1 The Gradient-Lipschitz setting

First we consider the Gradient-Lipschitz setting and give explicit conditions on the parameters α, h and K to have a theoretical guarantee on the sampling error, measured in the Wasserstein distance, of the LMC algorithm.

Theorem 10. *Suppose that the potential function f is convex and satisfies Condition 1. Let $q \in [1, 2]$. Then, for every $\alpha \le M/20$ and $h \le 1/(M + \alpha)$, we have*

$$W_q(\nu_K^{\alpha\text{-LMC}}, \pi) \le \underbrace{\sqrt{\mu_2}(1 - \alpha h)^{K/2}}_{\substack{\text{error due to the}\\ \text{time finiteness}}} + \underbrace{(2.1hMp/\alpha)^{1/2}}_{\text{discretization error}} + \underbrace{\Big(C_q \alpha \mu_2^{(q+2)/2}\Big)^{1/q}}_{\substack{\text{error due to the lack}\\ \text{of strong-convexity}}},$$

where C_q is a dimension free constant given by (3.4).

The proof of this result is postponed to the end of this section. Let us consider its consequences in the cases $q = 1$ and $q = 2$ presented in the table of Section 3.4.2. The general strategy is to choose the value of α by minimizing the sum of the discretization error and the error caused by the lack of strong convexity. Then, the parameter h is chosen so that the sum of the two aforementioned errors is smaller than 99% of the target precision $\varepsilon\sqrt{\mu_2}$. Finally, the number of iterations K is selected in such a way that the error due to the time finiteness is also smaller than 1% of the target precision.

Implementing this strategy for $q = 1$ and $q = 2$, we get the optimized value of α and the corresponding value of h,

$q = 1$		$q = 2$	
$\alpha = \dfrac{(2.1hMp)^{1/3}}{(44)^{2/3}\mu_2}$	$h = \dfrac{\varepsilon^3}{322Mp}$	$\alpha = \dfrac{(2.1hMp)^{1/2}}{(111)^{1/2}\mu_2}$	$h = \dfrac{\varepsilon^4}{3900Mp}$

These values of α and h satisfy the conditions imposed in Theorem 10. They imply that the computational complexity of the method, for $\nu_0 = \delta_0$ (the Dirac mass at the origin), is given by

$$K_{\alpha\text{-LMC},W_1}(p,\varepsilon) \le \frac{2}{\alpha h}\log\left(\frac{100W_2(\nu_0,\pi_\alpha)}{\varepsilon\sqrt{\mu_2}}\right) \le 4.3\times 10^4 M\frac{\mu_2 p}{\varepsilon^4}\log(100/\varepsilon)$$
$$K_{\alpha\text{-LMC},W_2}(p,\varepsilon) \le \frac{2}{\alpha h}\log\left(\frac{100W_2(\nu_0,\pi_\alpha)}{\varepsilon\sqrt{\mu_2}}\right) \le 3.6\times 10^6 M\frac{\mu_2 p}{\varepsilon^6}\log(100/\varepsilon).$$

In both inequalities, the second passage is due to the monotone behaviour of the function $\gamma \mapsto \mu_2(\pi_\gamma)$. This property is formulated in the Lemma 10. Combining Condition 2, $M\mu_2(\pi) \le \kappa p^\beta$ with the last display, we check that $K_{\alpha\text{-LMC},W_1}(p,\varepsilon) \le C\kappa(p^{1+\beta}/\varepsilon^4)\log(100/\varepsilon)$ and $K_{\alpha\text{-LMC},W_2}(p,\varepsilon) \le C\kappa(p^{1+\beta}/\varepsilon^6)\log(100/\varepsilon)$. For $\beta = 1$, this matches well with the rates reported in the table of Section 3.4.2. Unfortunately, the numerical constant C, just like the factors 4.3×10^4 and 2×10^7 in the last display, is way too large to be useful for practical purposes. Getting similar bounds with better numerical constants is an open question. The same remark applies to all the results presented in the subsequent sections.

Proof of Theorem 10. To ease notation, we write ν_K instead of $\nu_K^{\alpha\text{-LMC}}$. The triangle inequality and the monotony of W_q with respect to q imply that

$$W_q(\nu_K,\pi) \le W_2(\nu_K,\pi_\alpha) + W_q(\pi_\alpha,\pi).$$

Recall that π_α is α-strongly log-concave and have f_α as its potential function. By definition, f_α has also a Lipschitz continuous gradient with the Lipschitz constant at least $M+\alpha$. As we assume the condition $h \le 1/(M+\alpha)$ is satisfied, we can apply [DMM19, Theorem 9]. It implies that

$$\begin{aligned}W_2(\nu_K,\pi_\alpha) &\le (1-\alpha h)^{K/2}W_2(\nu_0,\pi_\alpha) + (2h(M+\alpha)p/\alpha)^{1/2}\\ &\le (1-\alpha h)^{K/2}W_2(\nu_0,\pi_\alpha) + (2.1hMp/\alpha)^{1/2}.\end{aligned}$$

The latter is true due to the fact that $\alpha \le M/20$. The remaining term is bounded using Proposition 12. We obtain

$$\begin{aligned}W_q(\nu_K,\pi) &\le (1-\alpha h)^{K/2}W_2(\nu_0,\pi_\alpha) + (2.1hMp/\alpha)^{1/2} + W_q(\pi_\alpha,\pi)\\ &\le \sqrt{\mu_2(\pi_\alpha)}(1-\alpha h)^{K/2} + (2.1hMp/\alpha)^{1/2} + \left(C_q\alpha\mu_2^{(q+2)/2}\right)^{1/q}.\end{aligned}$$

Thus, applying Lemma 10, we conclude the proof. □

3.6.2 The Hessian-Lipschitz setting

It has been noticed by [DM19], see also [DK19, Theorem 5], that if the potential f has a Lipschitz-continuous Hessian matrix, then the LMC algorithm, without any modification, is more accurate than in the Gradient-Lipschitz setting. These improvements were obtained under the condition of strong convexity of the potential, showing that the computational complexity drops down from p/ε^2 to p/ε. The goal of this section is to understand how this additional smoothness assumption impacts the computational complexity of the α-LMC algorithm.

Theorem 11. *Suppose that the potential function f satisfies conditions 1 and 3. Let $q \in [1, 2]$. For every $\alpha \leq M/20$ and $h \leq 1/(M + \alpha)$, we have*

$$W_q(\nu_K^{\alpha\text{-LMC}}, \pi) \leq \underbrace{\sqrt{\mu_2}(1-\alpha h)^K}_{\substack{\text{error due to the}\\ \text{time finiteness}}} + \underbrace{\frac{M_2 hp}{2\alpha} + \frac{2.8 M^{3/2} hp^{1/2}}{\alpha}}_{\text{discretization error}} + \underbrace{\Big(C_q \alpha \mu_2^{(q+2)/2}\Big)^{1/q}}_{\substack{\text{error due to the lack}\\ \text{of strong-convexity}}},$$

where C_q is a dimension free constant given by (3.4).

In order to provide more insight on the complexity bounds implied by the latter result, let us instantiate it for $q = 1$ and $q = 2$. Optimizing the sum of the two last error terms with respect to α, then choosing this sum to be equal to $0.99\varepsilon\sqrt{\mu_2}$, we arrive at the following values

$q = 1$		$q = 2$	
$\alpha = \left(\dfrac{hQp}{44\mu_2^{3/2}}\right)^{1/2}$	$h = \dfrac{\varepsilon^2}{45\mu_2^{1/2}Qp}$	$\alpha = \dfrac{(hQp)^{2/3}}{(111\mu_2^2)^{1/3}}$	$h = \dfrac{\varepsilon^3}{387\mu_2^{1/2}Qp}$

Here Q is defined as $(M_2 + 5.6M^{3/2}p^{-1/2})$. These values of α and h satisfy the conditions imposed in Theorem 11. They imply that the computational complexity of the method, for $\nu_0 = \delta_0$ (the Dirac mass at the origin), is given by

$$K_{\alpha\text{-LMC},W_1}(p,\varepsilon) \leq \frac{2}{\alpha h}\log\Big(\frac{100 W_2(\nu_0,\pi_\alpha)}{\varepsilon\sqrt{\mu_2}}\Big) \leq 2\times 10^3 \mu_2^{3/2} Q(p/\varepsilon^3)\log(100/\varepsilon)$$
$$K_{\alpha\text{-LMC},W_2}(p,\varepsilon) \leq \frac{2}{\alpha h}\log\Big(\frac{100 W_2(\nu_0,\pi_\alpha)}{\varepsilon\sqrt{\mu_2}}\Big) \leq 9.9\times 10^4 \mu_2^{3/2} Q(p/\varepsilon^5)\log(100/\varepsilon).$$

Combining Condition 2 and the last display, we check that

$$K_{\alpha\text{-LMC},W_1}(p,\varepsilon) \leq C\varepsilon^{-3}(\kappa_2^{3/2}p^{(2+3\beta)/2} + \kappa^{3/2}p^{(1+3\beta)/2})\log(100/\varepsilon),$$
$$K_{\alpha\text{-LMC},W_2}(p,\varepsilon) \leq C\varepsilon^{-5}(\kappa_2^{3/2}p^{(2+3\beta)/2} + \kappa^{3/2}p^{(1+3\beta)/2})\log(100/\varepsilon).$$

The latter is true, since by definition κ_2 is equal to $M_2^{2/3}D$. For $\beta = 1$, this matches well with the rates reported in the table of Section 3.4.2.

Proof. Theorem 11 We repeat the same steps as in the proof of Theorem 10, except that instead of [DMM19, Theorem 9] we use [DK19, Theorem 5]. To ease notation, we write ν_K instead of $\nu_K^{\alpha\text{-LMC}}$. One easily checks that π_α is α-strongly log-concave with potential function f_α. Furthermore, the latter is $(M+\alpha)$-gradient-Lipschitz and M_2-Hessian-Lipschitz. Therefore, for $h \le 2/(M+\alpha)$, Theorem 5 from [DK19] implies that

$$\begin{aligned} W_2(\nu_K, \pi_\alpha) &\le (1-\alpha h)^K W_2(\nu_0, \pi_\alpha) + \frac{M_2 hp}{2\alpha} + \frac{13(M+\alpha)^{3/2} hp^{1/2}}{5\alpha} \\ &\le (1-\alpha h)^K W_2(\nu_0, \pi_\alpha) + \frac{M_2 hp}{2\alpha} + \frac{2.8 M^{3/2} hp^{1/2}}{\alpha}. \end{aligned}$$

where the second inequality follows from the fact that $\alpha \le M/20$. The triangle inequality and the monotony of W_q with respect to q yield $W_q(\nu_K, \pi) \le W_2(\nu_K, \pi_\alpha) + W_q(\pi_\alpha, \pi)$, which leads to

$$W_q(\nu_K, \pi) \le \sqrt{\mu_2(\pi_\alpha)}(1-\alpha h)^K + \frac{M_2 hp}{2\alpha} + \frac{2.8 M^{3/2} hp^{1/2}}{\alpha} + W_q(\pi, \pi_\alpha).$$

Replacing the last term above by its upper bound provided by Proposition 12 and applying Lemma 10, we get the claimed result. □

3.7 Precision and computational complexity of KLMC and KLMC2

Several recent studies showed that for some classes of targets, including the strongly log-concave densities, the sampling error of discretizations of the kinetic Langevin diffusion scales better with the large dimension than discretizations of the Langevin diffusion. However, the dependence of the available bounds on the condition number is better for the Langevin diffusion. In this section we show a similar behavior in the case of non-strongly log-concave densities. This is done by providing quantitative upper bounds on the error of sampling using the kinetic Langevin process.

Theorem 12. *Suppose that the potential function f satisfies Condition 1. Let $q \in [1, 2]$.*

Then for every $\alpha \le M/20$, $\gamma \ge \sqrt{M+2\alpha}$ and $h \le \alpha/(4\gamma(M+\alpha))$, we have

$$W_q(\nu_K^{\alpha\text{-KLMC}}, \pi) \le \underbrace{\sqrt{2\mu_2}\left(1 - \frac{3\alpha h}{4\gamma}\right)^K}_{\text{error due to the time finiteness}} + \underbrace{1.5Mp^{1/2}(h/\alpha)}_{\text{discretization error}} + \underbrace{\left(C_q \alpha \mu_2^{(q+2)/2}\right)^{1/q}}_{\substack{\text{error due to the lack}\\ \text{of strong-convexity}}}.$$

where C_q is a dimension free constant given by (3.4).

The proof of this result is postponed to the end of this section. The contraction rate is an increasing function of γ, therefore we choose its lowest possible value achieved for $\gamma = \sqrt{M+2\alpha}$. Then the strategy is the same as for the previous section, that is to choose the value of α by minimizing the sum of the discretization error and the error caused by the lack of strong convexity. Then, the parameter h is chosen so that the sum of the two aforementioned errors is smaller than 99% of the target precision $\varepsilon\sqrt{\mu_2}$. The number of iterations K is selected in such a way that the error due to the time finiteness is also smaller than 1% of the target precision.

Implementing this strategy for $q = 1$ and $q = 2$, we get the optimized value of α and the corresponding value of h,

$q = 1$		$q = 2$	
$\alpha = \dfrac{(1.5hMp^{1/2})^{1/2}}{(22\mu_2^{3/2})^{1/2}}$	$h = \dfrac{\varepsilon^2}{143M(\mu_2 p)^{1/2}}$	$\alpha = \dfrac{(3hMp^{1/2})^{2/3}}{(111\mu_2^2)^{1/3}}$	$h = \dfrac{\varepsilon^4}{1200M(\mu_2 p)^{1/2}}$

These values of α and h satisfy the conditions imposed in Theorem 12. They imply that the computational complexity of the method, for $\nu_0 = \delta_0$ (the Dirac mass at the origin), is given by

$$K_{\alpha\text{-KLMC},W_1}(p,\varepsilon) \le \frac{4\gamma}{3\alpha h}\log\left(\frac{150}{\varepsilon}\right) \le 9.2\times 10^3 (M\mu_2)^{3/2}(p^{1/2}/\varepsilon^3)\log(150/\varepsilon)$$
$$K_{\alpha\text{-KLMC},W_2}(p,\varepsilon) \le \frac{4\gamma}{3\alpha h}\log\left(\frac{150}{\varepsilon}\right) \le 4.4\times 10^5 (M\mu_2)^{3/2}(p^{1/2}/\varepsilon^5)\log(150/\varepsilon).$$

Recall that Condition 2 implies $M\mu_2 \le \kappa p^\beta$. Combining this inequality with the last display, we check that

$$K_{\alpha\text{-KLMC},W_q}(p,\varepsilon) \le C\kappa^{3/2}(p^{(1+3\beta)/2}/\varepsilon^{2q+1})\log(150/\varepsilon), \qquad q = 1,2.$$

For $\beta = 1$, this matches well with the rates reported in the table of Section 3.4.2.

Proof. Theorem 12 To ease notation, we write ν_K instead of $\nu_K^{\alpha\text{-KLMC}}$. The triangle inequality and the monotony of W_q with respect to q imply that

$$W_q(\nu_K, \pi) \le W_2(\nu_K, \pi_\alpha) + W_q(\pi_\alpha, \pi).$$

Recall that π_α is α-strongly log-concave and have f_α as its potential function. By definition, f_α has also a Lipschitz continuous gradient with the Lipschitz constant at least $M + \alpha$. As we assume the conditions $\alpha \le M/20$, $\gamma \ge \sqrt{M + 2\alpha}$ and $h \le \alpha/(4\gamma(M+\alpha))$ are satisfied, we can apply [DRD20, Theorem 2]. It implies that

$$\begin{aligned} W_2(\nu_K, \pi_\alpha) &\le \sqrt{2}(1 - 3\alpha h/(4\gamma))^K W_2(\nu_0, \pi_\alpha) + \sqrt{2}(M+\alpha)p^{1/2}(h/\alpha) \\ &\le \sqrt{2\mu_2(\pi_\alpha)}(1 - 3\alpha h/(4\gamma))^K + 1.5Mp^{1/2}(h/\alpha). \end{aligned}$$

The latter is true thanks to the fact that $\alpha \le M/20$. Thus, Lemma 10 yields

$$W_q(\nu_K, \pi) \le \sqrt{2\mu_2(\pi)}(1 - 3\alpha h/(4\gamma))^K + 1.5Mp^{1/2}(h/\alpha) + W_q(\pi_\alpha, \pi).$$

The remaining term is bounded using Proposition 12. $\square$

The rest of this section is devoted to the results for the KLMC2 algorithm, which assumes that accurate evaluations of the Hessian of the potential function f can be performed at each given point.

Theorem 13. *Suppose that the potential function f satisfies conditions 1 and 3. Let $q \in [1, 2]$ and $Q = (M_2 + M^{3/2}p^{-1/2})$. Then for every $\alpha, h, \gamma > 0$ such that*

$$\alpha \le \frac{M}{20}, \qquad \gamma \ge \sqrt{M + 2\alpha}, \qquad h \le \frac{\alpha}{5\gamma(M+\alpha)} \bigvee \frac{\alpha}{4M_2\sqrt{5p}}$$

we have

$$W_q(\nu_K^{\alpha\text{-KLMC2}}, \pi) \le \underbrace{\sqrt{2\mu_2}\left(1 - \frac{\alpha h}{4\gamma}\right)^K}_{\substack{\text{error due to the} \\ \text{time finiteness}}} + \underbrace{\frac{2h^2Qp}{\alpha} + \frac{1.6}{\sqrt{M}}\exp\left\{-\frac{(\alpha/h)^2}{160M_2^2}\right\}}_{\text{discretization error}} + \underbrace{\left(C_q\alpha\mu_2^{(q+2)/2}\right)^{1/q}}_{\substack{\text{error due to the lack} \\ \text{of strong-convexity}}},$$

where C_q is a dimension free constant given by (3.4).

The proof of this result is postponed to the end of this section. The contraction rate is an increasing function of γ, therefore we choose its lowest possible value achieved for $\gamma = \sqrt{M + 2\alpha}$. In this case the strategy for finding h and α is slightly different from the previous ones. Here we first choose the parameter h so that the two terms of the discretization error are respectively bounded by 1% and 2% of the target precision $\varepsilon\sqrt{\mu_2}$.

This yields the following choice for the time step h:

$$h = \alpha\left(160 M_2^2 \log\left(\frac{160}{\varepsilon\sqrt{M\mu_2}}\right) \bigvee \frac{100\alpha Q p}{\varepsilon\sqrt{\mu_2}}\right)^{-1/2}.$$

The parameter α is then chosen so that the error due to the lack of strong convexity is lower than 96% of the target precision. Implementing this strategy for $q = 1$ and $q = 2$, we get the following value for α

$q = 1$	$q = 2$
$\alpha = \dfrac{\varepsilon}{23\mu_2}$	$\alpha = \dfrac{\varepsilon^2}{116\mu_2}$

Finally, the number of iterations K is selected in such a way that the error due to the time finiteness is also smaller than 1% of the target precision. This yields, that

$$K = \frac{4\gamma}{\alpha h}\log\left(\frac{142}{\varepsilon}\right)$$

is sufficient to reach the target precision. The values of γ, α and h imply that the computational complexity of the method is given by

$$K_{\alpha\text{-KLMC2},W_1}(p,\varepsilon) = 2.2\times 10^4\,\frac{M^{1/2}M_2\mu_2^2}{\varepsilon^2}\left\{1.6\log\left(\frac{160}{\varepsilon\sqrt{M\mu_2}}\right)\bigvee\frac{Qp}{23M_2^2\mu_2^{3/2}}\right\}^{1/2}\log\left(\frac{142}{\varepsilon}\right)$$
$$K_{\alpha\text{-KLMC2},W_2}(p,\varepsilon) = 5.4\times 10^6\,\frac{M^{1/2}M_2\mu_2^2}{\varepsilon^4}\left\{1.6\log\left(\frac{160}{\varepsilon\sqrt{M\mu_2}}\right)\bigvee\frac{\varepsilon Qp}{116M_2^2\mu_2^{3/2}}\right\}^{1/2}\log\left(\frac{142}{\varepsilon}\right).$$

Since according to Condition 2 $\mu_2 \le Dp^\beta$, the last display implies that up to logarithmic factors $K_{\alpha\text{-KLMC2},W_1}(p,\varepsilon)$ scales as $\kappa^{1/2}\kappa_2^{3/2}p^{2\beta}/\varepsilon^2$ and $K_{\alpha\text{-KLMC2},W_2}(p,\varepsilon)$ scales as $\kappa^{1/2}\kappa_2^{3/2}p^{2\beta}/\varepsilon^4$. For $\beta = 1$, this matches well with the rates reported in the table of Section 3.4.2.

Proof. Theorem 13 To ease notation, we write ν_K instead of $\nu_K^{\alpha\text{-KLMC2}}$. As already checked in the proof of Theorem 11 the distribution π_α is α-strongly log-concave with potential function f_α. Furthermore, the latter is $(M+\alpha)$-gradient-Lipschitz and M_2-Hessian-Lipschitz. We apply [DRD20, Theorem 3] which ensures that, if the parameters $\alpha, \gamma, h > 0$ are such that

$$\alpha \le \frac{M}{20}, \qquad \gamma \ge \sqrt{M + 2\alpha}, \qquad h \le \frac{\alpha}{5\gamma(M+\alpha)} \bigwedge \frac{\alpha}{4M_2\sqrt{5p}}$$

then the distribution of the KLMC2 sampler after k iterates satisfies

$$\begin{aligned}
W_2(\nu_k, \pi_\alpha) &\leq \sqrt{2\mu_2(\pi_\alpha)}\left(1 - \frac{\alpha h}{4\gamma}\right)^k + \frac{2h^2 M_2 p}{\alpha} + \frac{h^2(M+\alpha)^{3/2}\sqrt{2p}}{\alpha} \\
&\qquad\qquad + \frac{8h(M+\alpha)}{\alpha}\exp\left\{-\frac{\alpha^2}{160 M_2^2 h^2}\right\} \\
&\leq \sqrt{2\mu_2(\pi_\alpha)}\left(1 - \frac{\alpha h}{4\gamma}\right)^K + \frac{2h^2(M_2 p + M^{3/2}p^{1/2})}{\alpha} + \frac{1.6}{\sqrt{M}}\exp\left\{-\frac{(\alpha/h)^2}{160 M_2^2}\right\},
\end{aligned}$$

where the second inequality follows from the fact that $\alpha \leq M/20$ and $h \leq \alpha/(5\gamma(M+\alpha))$. The triangle inequality and the monotony of W_q with respect to q yields $W_q(\nu_K, \pi) \leq W_2(\nu_K, \pi_\alpha) + W_q(\pi_\alpha, \pi)$, which leads to

$$W_q(\nu_K, \pi) \leq \sqrt{2\mu_2(\pi_\alpha)}\left(1 - \frac{\alpha h}{4\gamma}\right)^K + \frac{2h^2 Q p}{\alpha} + \frac{1.6}{\sqrt{M}}\exp\left\{-\frac{(\alpha/h)^2}{160 M_2^2}\right\} + W_q(\pi, \pi_\alpha).$$

Replacing the last term above by its upper bound provided by Proposition 12 and applying Lemma 10, we conclude the proof of the theorem. □

3.8 Bounding moments

From the user's perspective, the choice of α and h requires the computation of the second moment of the distribution π. In most cases, this moment is an intractable integral. However, when some additional information on π is available, this moment can be replaced in some cases by a tractable upper bound. In this section, we provide upper bounds on the moments

$$\mu_a^* := \int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}, \qquad a > 0,$$

centered at the minimizer of the potential $\boldsymbol{\theta}^* \in \text{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} f(\boldsymbol{\theta})$. The knowledge of the second moment is enough to compute the mixing times presented in Section 3.6 and Section 3.7. However, providing bounds on general moments may be of interest in order to get sharp numerical constants. For instance, the proof of Proposition 12 shows that results for the W_1 and W_2 metrics essentially rely on some bounds over the third and fourth moments of π, which could be better understood in some specific contexts.

In this section, we investigate two particular classes of convex functions: (a) those which are m-strongly convex inside a ball of radius R around the mode $\boldsymbol{\theta}^*$, and (b) those which are m-strongly convex outside a ball of radius R around the mode $\boldsymbol{\theta}^*$. We provide user-friendly bounds on μ_a^* with relatively small constants. If m and R are dimension free, we show that μ_a^* scales as $(p\log p)^a$, respectively $(p\log p)^{a/2}$. Within a poly-log factor, the scaling with the dimension is sharp, and matches Condition 2 with $\beta = 2$ for the class (a) and with $\beta = 1$ for the class (b).

Proposition 13. *Assume that for some positive numbers m and R, we have $\nabla^2 f(\boldsymbol{\theta}) \succeq m\mathbf{I}_p$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \le R$. Then, for every $a > 0$, we have*

$$\mu_a^* \le A \vee B + \frac{2^{a+1}}{(mR)^a \Gamma(p/2)}$$

where[4]

$$A = \left\{ \frac{3}{mR} \left((p+a)\log(p+a) + p\log_+ \left(\frac{2M}{m^2R^2} \right) \right) \right\}^a$$

and

$$B = \left(\frac{p}{m}\right)^{a/2} \left\{ 2^{a-1} \left(1 + (1+a/p)^{a/2-1} \right) \right\}^{\mathbb{1}_{a>2}}.$$

Remark 4. *If the assumptions of Proposition 13 are satisfied, then*

$$\mu_a^* = \widetilde{O}\left(\left(\frac{p}{mR}\right)^a \bigvee \left(\frac{p}{m}\right)^{a/2} \right).$$

In the bound of Proposition 13, the term A is the dominating one when p/m is large as compared to R^2, while B is the dominating term when R^2 is of a higher order of magnitude than p/m. The residual term $2^{a+1}/((mR)^a\Gamma(p/2))$ goes to zero whenever p or R tend to infinity. If m and R are assumed to be dimension free constants, then μ_a^* scales as $(p\log p)^a$. This rate is optimal within a poly-log factor, which is proven in Lemma 13. Note that when R goes to infinity we recover exactly the bound of the strongly convex case proven in Lemma 11.

We now switch to bounding the moments of π under the condition that f is convex everywhere and strongly convex outside the ball of radius R around $\boldsymbol{\theta}^*$.

Proposition 14. *Assume that for some positive m and R, we have $\nabla^2 f(\boldsymbol{\theta}) \succeq m\mathbf{I}_p$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 > R$. If $p \ge 3$, then, for every $a > 0$, we have*

$$\mu_a^* \le \left(1 + \frac{2}{\Gamma(p/2)}\right) \left\{ (4R) \bigvee \left(\frac{4(p+a)}{m} \log\left(\frac{pM}{m}\right) \right)^{1/2} \right\}^a.$$

Remark 5. *Under the assumptions of Proposition 14, we obtain*

$$\mu_a^* = \widetilde{O}\left(R^a \vee (p/m)^{a/2} \right).$$

In the bound of Proposition 14, if m and R are assumed to be dimension free constants,

[4] We denote by $\log_+ x$ the positive part of $\log x$, $\log_+ x = max(0, \log x)$.

then μ_a^* scales as $(p \log p)^{a/2}$. This rate is improved in Proposition 15 below to $p^{a/2}$, which is optimal. However, the bound of Proposition 14 is sharper when R is large.

Proposition 15. *Assume that for some positive m and R, we have $\nabla^2 f(\boldsymbol{\theta}) \succeq m\mathbf{I}_p$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 > R$. Then for every $a > 0$ we have*

$$\mu_a^* \leq e^{mR^2/2} \left(\frac{p}{m}\right)^{a/2} \left\{2^{a-1}\left(1 + (1 + a/p)^{a/2-1}\right)\right\}^{\mathbb{1}_{a>2}}.$$

Note that when R approaches zero, this bound matches the one of the strongly convex case; see, for instance, Lemma 11. To close this section, let us note that in the setting considered in propositions 15 and 14, it is quite likely that an approach based on reflection coupling [MMS18, CCA+18] would give a sharper upper bound than those obtained by adding a quadratic penalty to the potential.

Appendix to Chapter 3

This chapter contains proofs of the propositions stated in previous sections as well as those of some technical lemmas used in the proofs of the propositions.

3.A Proof of Proposition 13

Note that for any $\boldsymbol{\theta} \in \mathbb{R}^p$, $\nabla^2 f(\boldsymbol{\theta}) \succeq m(\|\boldsymbol{\theta}\|_2)\mathbf{I}_p$, for the map $m(r) := m\mathbb{1}_{(0,R)}(r)$. We start by computing the map $\widetilde{m}(r) \ := 2\int_0^1 m(ry)(1-y)dy$. By definition, we get

$$\begin{aligned}
\widetilde{m}(r) &= 2\int_0^1 m\mathbb{1}_{(0,R)}(ry)(1-y)dy \\
&= 2m\int_0^{1\wedge R/r}(1-y)dy \\
&= m\mathbb{1}_{r<R} + m\left(\frac{2R}{r} - \frac{R^2}{r^2}\right)\mathbb{1}_{r\geq R}.
\end{aligned}$$

Let $A \geq R$ and $a > 0$. We assume without loss of generality that $\boldsymbol{\theta}^* = \mathbf{0}_p$. Define $B_A = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\|_2 \leq A\}$. We split the integral into two parts:

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} = \int_{B_A} \|\boldsymbol{\theta}\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} + \int_{B_A^c} \|\boldsymbol{\theta}\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}.$$

Let us bound the second summand. Since $A \geq R$, for any $r > A$, we have $\widetilde{m}(r)r^2/2 = mRr - mR^2/2$. Applying Lemma 12 yields

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} \leq \frac{2(M/2)^{p/2}}{\Gamma(p/2)} e^{mR^2/2}\int_A^{+\infty} r^{p+a-1}e^{-mRr}dr.$$

We now use the following inequality on the incomplete Gamma function from [NP00], (see also [BC+09]). Let $B > 1$, and $q \geq 1$. Then, for all $x \geq (B/(B-1))(q-1)$, we have

$$\int_x^{+\infty} y^{q-1}e^{-y}dy \leq Bx^{q-1}e^{-x}. \tag{3.6}$$

We apply this inequality for $B = 2$ and $q = p + a$. For $A \geq 2(p + a - 1)/(mR)$, we have the following:

$$\begin{aligned}\int_A^{+\infty} r^{p+a-1} e^{-mRr} dr &= \left(\frac{1}{mR}\right)^{a+p} \int_{mRA}^{+\infty} y^{p+a-1} e^{-y} dy \\ &\leq \left(\frac{2}{mR}\right)^{a+p} \left(\frac{mRA}{2}\right)^{p+a-1} e^{-mRA}.\end{aligned}$$

Now, we make use of the fact that $mRA \geq mRA/2$. The latter yields

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \frac{2^{a+1}}{(mR)^a \Gamma(p/2)} \left(\frac{2M}{m^2 R^2}\right)^{p/2} \left(\frac{mRA}{2}\right)^{p+a-1} e^{-mRA/2}.$$

The last bound ensures that the inequality

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \frac{2^{a+1}}{(mR)^a \Gamma(p/2)} \tag{3.7}$$

is fulfilled whenever $\varphi(x) := x - c\log(x) - b \geq 0$, where

$$x = \frac{mRA}{2}, \qquad c = p + a - 1, \qquad b = \frac{p}{2} \log\left(\frac{2M}{m^2 R^2}\right).$$

We now establish for which values of x (or equivalently, A) we have $\varphi(x) \geq 0$. Taylor's expansion around $y_c := 1.5(c + 1)\log(c + 1)$ yields

$$\varphi\big(y_c + 3b_+\big) = \varphi(y_c) + \varphi'(y) \times 3b_+$$

for some $y \geq y_c$. The latter implies that

$$\varphi'(y) = 1 - \frac{c}{y} \geq 1 - \frac{c}{y_c} \geq 1/3.$$

Hence, $\varphi\big(y_c + 3b_+\big) \geq \varphi(y_c) + b_+ \geq y_c - c\log y_c + b_+ - b \geq 0$. Since the map φ is increasing on $[c, +\infty)$ and $y_c + 3b_+ \geq c$, we conclude that (3.7) is fulfilled for any

$$A \geq A^* := \frac{3}{mR}\left((p + a)\log(p + a) + p\log_+\left(\frac{2M}{m^2 R^2}\right)\right).$$

Recall that $A \geq R$ by assumption, this brings two cases to consider. The first is the case when $R < A^*$. Therefore for $A = A^*$ we have

$$\int_{B_A} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq A^a.$$

The second case to consider is $R \geq A^*$. Hence for $A = R$, the map $f(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta})$ is m-strongly convex on the ball $B_A = B_R$. Thus Lemma 11 yields

$$\int_{B_A} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \left(\frac{p}{m}\right)^{a/2} \left\{2^{a-1}\left(1 + (1 + a/p)^{a/2-1}\right)\right\}^{\mathbb{1}_{a>2}}.$$

Since inequality (3.7) is fulfilled in both cases, the claim of Proposition 13 follows.

3.B Proof of Proposition 14

Note that for any $\boldsymbol{\theta} \in \mathbb{R}^p$, $\nabla^2 f(\boldsymbol{\theta}) \succeq m(\|\boldsymbol{\theta}\|_2)\mathbf{I}_p$, where $m(\cdot)$ is defined as below:

$$m(r) = m\mathbb{1}_{(R,+\infty)}(r).$$

We begin by computing the map $\widetilde{m}(r) := 2\int_0^1 m(ry)(1-y)dy$. Using the definition of $\widetilde{m}$, we have:

$$\begin{aligned}
\widetilde{m}(r) &= 2\int_0^1 m\mathbb{1}_{(R,+\infty)}(ry)(1-y)dy \\
&= 2m\mathbb{1}_{r>R}\int_{R/r}^1 (1-y)dy \\
&= m\,(1 - R/r)^2\,\mathbb{1}_{r>R}.
\end{aligned}$$

Let $A \geq 4R$ and $a > 0$. We assume without loss of generality that $\boldsymbol{\theta}^* = \mathbf{0}_p$. Define $B_A = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\|_2 \leq A\}$. We will use the following bound:

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq A^a + \int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}.$$

For the second term, Lemma 12 yields

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \frac{2(M/2)^{p/2}}{\Gamma(p/2)} \int_A^{+\infty} r^{p+a-1} e^{-mr^2/8} dr.$$

This is true due to the fact that, for every $r \geq A \geq 4R$, we have $\widetilde{m}(r) \geq m/2$. We now use inequality (3.6) with $B = 2$, $q = (p+a)/2$ and $mA^2/4 \geq (p+a) - 1/2$:

$$\int_A^{+\infty} r^{p+a-1} e^{-mr^2/8} dr = 2^{-1}\left(\frac{4}{m}\right)^{(p+a)/2} \int_{mA^2/4}^{+\infty} y^{(p+a)/2-1} e^{-y} dy.$$

Thus the following is satisfied

$$\begin{aligned}\int_A^{+\infty} r^{p+a-1} e^{-mr^2/8} dr &\leq \left(\frac{4}{m}\right)^{(p+a)/2} \left(\frac{mA^2}{4}\right)^{(p+a)/2-1} e^{-mA^2/4} \\ &\leq A^a \left(\frac{4}{m}\right)^{p/2} \left(\frac{mA^2}{4}\right)^{p/2-1} e^{-mA^2/4}.\end{aligned}$$

This yields

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \frac{2A^a}{\Gamma(p/2)} \left(\frac{2M}{m}\right)^{p/2} \left(\frac{mA^2}{4}\right)^{p/2-1} e^{-mA^2/4}.$$

The last bound ensures that the inequality

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \frac{2A^a}{\Gamma(p/2)} \tag{3.8}$$

is fulfilled whenever $\varphi(x) := x - c\log(x) - b \geq 0$, where

$$x = \frac{mA^2}{4}, \qquad c = \frac{p}{2} - 1, \qquad b = \frac{p}{2} \log\left(\frac{2M}{m}\right) > 0.$$

Taylor's expansion around $y_c := 2(c+1)\log(c+1)$ yields

$$\varphi(y_c + 2b) = \varphi(y_c) + \varphi^{'}(y) \times 2b$$

for some $y \geq y_c$. The latter implies that

$$\varphi^{'}(y) = 1 - \frac{c}{y} \geq 1 - \frac{c}{y_c} \geq 1/3.$$

We get $\varphi(y_c + 2b) \geq y_c - c\log(y_c) + b - b \geq 0$. Since the map φ is increasing on $[c, +\infty)$ and $y_c + 2b \geq c$, we conclude that (3.8) is fulfilled for any

$$A^2 \geq \frac{4}{m}\left(p\log(p/2) + p\log(2M/m)\right) = \frac{4p}{m} \log\left(\frac{pM}{m}\right).$$

Finally, we choose A such that this inequality and the two additional assumptions: $A \geq 2R$ and $mA^2/4 \geq (p+a) - 1/2$ hold. If $p \geq 3$ we can choose

$$A = (4R) \bigvee \left(\frac{4(p+a)}{m} \log\left(\frac{pM}{m}\right)\right)^{1/2}.$$

This yields the claim of Proposition 14.

3.C Proof of Proposition 15

Define $f = -\log \pi$ and for any $\boldsymbol{\theta} \in \mathbb{R}^p$:

$$\bar{f}(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + \frac{m}{2}\left(\|\boldsymbol{\theta}\|_2 - R\right)^2 \mathbb{1}_{\|\boldsymbol{\theta}\|_2 \leq R}.$$

For any $\boldsymbol{\theta} \in \mathbb{R}^p$, we have $\bar{f}(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}) + mR^2/2$, this yields

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq e^{mR^2/2} \int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a e^{-\bar{f}(\boldsymbol{\theta})} d\boldsymbol{\theta}.$$

Now we define the normalising constant

$$\bar{C} := \int_{\mathbb{R}^p} e^{-\bar{f}(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

and the corresponding probability density $\bar{\pi}(\boldsymbol{\theta}) := e^{-\bar{f}(\boldsymbol{\theta})}/\bar{C}$. The constant $\bar{C} \leq 1$ since $f(\boldsymbol{\theta}) \leq \bar{f}(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$. Therefore we have

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq e^{mR^2/2} \int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \bar{\pi}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}.$$

By construction the density $\bar{\pi}$ is m-strongly log-concave. We apply Lemma 11 on this last term and get the claim of Proposition 15.

3.D Proof of Proposition 12

Without loss of generality we may assume that $\int_{\mathbb{R}^p} \exp(-f(\boldsymbol{\theta}))\, d\boldsymbol{\theta} = 1$. We first derive upper and lower bounds for the normalizing constant of π_α, that is

$$c_\alpha := \int_{\mathbb{R}^p} \pi(\boldsymbol{\theta}) e^{-\alpha\|\boldsymbol{\theta}\|_2^2/2} d\boldsymbol{\theta}.$$

To do so, we introduce the following denotation:

$$r_\alpha := \frac{2}{\alpha} \log \frac{1}{c_\alpha}.$$

One can verify that $c_\alpha \leq 1$. To get a lower bound, we note that c_α is an expectation with respect to the density π, hence it can be lower bounded using Jensen's inequality, applied to the convex map $x \mapsto e^{-x}$. These two facts yield

$$\exp\{-\alpha\mu_2/2\} \leq c_\alpha \leq 1.$$

Therefore, by definition of r_α we have

$$0 \leq r_\alpha \leq \mu_2 \tag{3.9}$$

For any fixed $\boldsymbol{\theta} \in \mathbb{R}^p$, we now split the Euclidean distance between $\pi(\boldsymbol{\theta})$ and $\pi_\alpha(\boldsymbol{\theta})$ between its positive and negative parts:

$$|\pi(\boldsymbol{\theta}) - \pi_\alpha(\boldsymbol{\theta})| = \underbrace{\pi(\boldsymbol{\theta}) \left[1 - e^{-(\alpha/2)(\|\boldsymbol{\theta}\|_2^2 - r_\alpha)}\right] \mathbb{1}_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha}}_{:=(\pi-\pi_\alpha)_+(\boldsymbol{\theta})} + \underbrace{\pi(\boldsymbol{\theta}) \left[e^{-(\alpha/2)(r_\alpha - \|\boldsymbol{\theta}\|_2^2)} - 1\right] \mathbb{1}_{\|\boldsymbol{\theta}\|_2^2 < r_\alpha}}_{:=(\pi-\pi_\alpha)_-(\boldsymbol{\theta})}.$$

In order to bound the positive part, we make use of the inequality $1 - e^{-x} \leq x$ for $x > 0$. Therefore:

$$(\pi - \pi_\alpha)_+(\boldsymbol{\theta}) \leq \frac{\alpha}{2} \pi(\boldsymbol{\theta})(\|\boldsymbol{\theta}\|_2^2 - r_\alpha) \mathbb{1}_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha}. \tag{3.10}$$

The total variation distance between densities π and π_α is twice the integral of the positive part, therefore by definition

$$d_{\mathrm{TV}}(\pi_\alpha, \pi) = 2 \int_{\mathbb{R}^p} (\pi - \pi_\alpha)_+(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Which yields the following

$$d_{\mathrm{TV}}(\pi_\alpha, \pi) \leq \alpha \int_{\mathbb{R}^p} \pi(\boldsymbol{\theta})(\|\boldsymbol{\theta}\|_2^2 - r_\alpha) \mathbb{1}_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha} d\boldsymbol{\theta} \leq \alpha \int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^2 \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

where the first inequality follows from (3.10), and the second inequality follows from (3.9). This yields the first claim of the proposition.

The proof of the bound for Wasserstein distances is inspired from [Vil08] (Theorem 6.15, page 115). We consider the following coupling between π and π_α, defined by keeping fixed the mass shared by π and π_α while distributing the rest of the mass with a product measure. Letting $C := (\pi - \pi_\alpha)_+(\mathbb{R}^p) = (\pi - \pi_\alpha)_-(\mathbb{R}^p)$, we define

$$\gamma(d\boldsymbol{\theta}, d\boldsymbol{\theta}') := (\pi \wedge \pi_\alpha)(d\boldsymbol{\theta}) \delta_{\boldsymbol{\theta}' = \boldsymbol{\theta}} + \frac{1}{C} (\pi - \pi_\alpha)_+(d\boldsymbol{\theta})(\pi - \pi_\alpha)_-(d\boldsymbol{\theta}').$$

The joint distribution γ defines a coupling of π and π_α. Therefore for any $q \geq 1$, by definition of the Wasserstein distance we get

$$\begin{aligned} W_q^q(\mu, \nu) &\leq \int_{\mathbb{R}^p \times \mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^q \gamma(d\boldsymbol{\theta}, d\boldsymbol{\theta}') \\ &= \frac{1}{C} \int_{\mathbb{R}^p \times \mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^q (\pi - \pi_\alpha)_+(d\boldsymbol{\theta})(\pi - \pi_\alpha)_-(d\boldsymbol{\theta}') \end{aligned}$$

Thus,

$$
\begin{aligned}
W_q^q(\mu,\nu) &\le \frac{1}{C}\int_{\mathbb{R}^p\times\mathbb{R}^p} (\|\boldsymbol{\theta}\|_2+\sqrt{r_\alpha})^q\,(\pi-\pi_\alpha)_+(d\boldsymbol{\theta})(\pi-\pi_\alpha)_-(d\boldsymbol{\theta}^{'})\\
&= \int_{\mathbb{R}^p} (\|\boldsymbol{\theta}\|_2+\sqrt{r_\alpha})^q\,(\pi-\pi_\alpha)_+(d\boldsymbol{\theta})
\end{aligned}
$$

where the second inequality follows from the fact that $(\pi-\pi_\alpha)_+(d\boldsymbol{\theta})$, respectively $(\pi-\pi_\alpha)_-(d\boldsymbol{\theta}^{'})$, have positive mass only outside, respectively inside, the ball $\{\|\boldsymbol{\theta}^{'}\|_2\le\sqrt{r_\alpha}\}$. Therefore for any fixed $\boldsymbol{\theta}$ outside the ball, the maximum distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{'}$ is obtained on the boundary of the ball in the opposite direction of $\boldsymbol{\theta}$. We now define the quantity

$$
J_{q,\alpha}(\pi) := \frac{1}{2}\int_{\|\boldsymbol{\theta}\|_2^2>r_\alpha} (\|\boldsymbol{\theta}\|_2+\sqrt{r_\alpha})^q\left(\|\boldsymbol{\theta}\|_2^2-r_\alpha\right)\pi(d\boldsymbol{\theta}),
$$

and remark that inequality (3.10) yields

$$
W_q^q(\mu,\nu)\le \alpha J_{q,\alpha}(\pi).
$$

The claim of the proposition follows from the fact that there is a numerical constant C_q that only depends on q such that

$$
J_{q,\alpha}(\pi)\le C_q\mu_2^{(q+2)/2}.
$$

This is a combined consequence of (3.9) and Lemma 14. More precisely, (3.9) ensures that $J_{q,\alpha}(\pi)$ can be controlled only by the moments of π. On the other hand, Borell's lemma [GBVV14, Theorem 2.4.6] shows that $\mathbb{L}_q$-norms of log-concave distributions are all equivalent, in the sense that for any $q\ge 1$ there is a constant B_q that only depends on q such that $\mu_q^{1/q}\le B_q\mu_2^{1/2}$. Our version of this result is stated in Lemma 14. It is (to the best of our knowledge) the first attempt to provide optimized constants. We compute hereafter the values of C_q for $q=1$ and $q=2$. We have

$$
\begin{aligned}
J_{1,\alpha}(\pi) &\le \frac{1}{2}\int_{\mathbb{R}^p} (\|\boldsymbol{\theta}\|_2+\sqrt{\mu_2})\,\|\boldsymbol{\theta}\|_2^2\pi(d\boldsymbol{\theta})\\
&= (\mu_3+\mu_2^{3/2})/2\\
&\le 22\mu_2^{3/2}
\end{aligned}
$$

and

$$
\begin{aligned}
J_{2,\alpha}(\pi) &= \frac{1}{2}\int_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha} (\|\boldsymbol{\theta}\|_2 + \sqrt{r_\alpha})^2 \left(\|\boldsymbol{\theta}\|_2^2 - r_\alpha\right) \pi(d\boldsymbol{\theta}) \\
&\leq \frac{1}{2}\int_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha} \left(\|\boldsymbol{\theta}\|_2^4 + 2\|\boldsymbol{\theta}\|_2^3\sqrt{r_\alpha}\right) \pi(d\boldsymbol{\theta}).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
J_{2,\alpha}(\pi) &= (\mu_4 + 2\mu_3\mu_2^{1/2})/2 \\
&\leq 262\mu_2^2.
\end{aligned}
$$

In both calculations, inequality (3.9) is used to bound r_α, while the last inequality follows from Lemma 14. It turns out that in the particular case $q = 2$, the constant C_2 can be improved using the following transport inequality [GL10, Corollary 7.2].

Suppose that μ is a probability measure on $\mathbb{R}^p$ that admits a m-strongly log-concave density with respect to Lebesgue measure, then for any probability measure ν on $\mathbb{R}^p$ we have

$$
W_2^2(\nu, \mu) \leq (2/m) D_{\mathrm{KL}}(\nu || \mu).
$$

Applied to $\mu = \pi_\alpha$ and $\nu = \pi$ we get

$$
W_2^2(\pi, \pi_\alpha) \leq (2/\alpha) D_{\mathrm{KL}}(\pi || \pi_\alpha).
$$

The computation of the Kullback-Leibler divergence yields

$$
\begin{aligned}
D_{\mathrm{KL}}(\pi || \pi_\alpha) &= \int_{\mathbb{R}^p} \pi(\boldsymbol{\theta})(\alpha/2)(\|\boldsymbol{\theta}\|_2^2 - r_\alpha)\, d\boldsymbol{\theta} \\
&= \alpha\mu_2/2 + \log c_\alpha.
\end{aligned}
$$

Using the inequality $e^{-x} \leq 1 - x + x^2/2$ for $x > 0$ yields the following upper bound for $c\alpha$:

$$
c_\alpha = \int_{\mathbb{R}^p} \pi(\boldsymbol{\theta}) e^{-(\alpha/2)\|\boldsymbol{\theta}\|_2^2} d\boldsymbol{\theta} \leq 1 - \alpha\mu_2/2 + \alpha^2\mu_4/8.
$$

Since $\log(1 + x) \leq x$ for $x > -1$ we get

$$
D_{\mathrm{KL}}(\pi || \pi_\alpha) \leq \alpha^2\mu_4/8.
$$

Combining this inequality with the bound on μ_4 from Lemma 14, we get

$$
W_2^2(\pi, \pi_\alpha) \leq \alpha\mu_4/4 \leq 111\alpha\mu_2^2.
$$

This shows that for $q = 2$ the constant can be improved to $C_2 = 111$. Therefore we get the claim of the proposition.

3.E Technical lemmas

Lemma 10. *Suppose that π has a finite fourth-order moment. Then $\gamma \mapsto \mu_2(\pi_\gamma)$ is continuously differentiable and non-increasing, when $\gamma \in [0, +\infty)$.*

Proof. For $k \in \mathbb{N} \cup \{0\}$, define

$$h_k(\gamma) = \int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^k \exp\left(-f(\boldsymbol{\theta}) - \gamma\|\boldsymbol{\theta}\|_2^2/2\right) d\boldsymbol{\theta}.$$

If $\pi \in \mathcal{P}_k(\mathbb{R}^p)$ then the function h_k is continuous on $[0; +\infty)$. Indeed, if the sequence $\{\gamma_n\}_n$ converges γ_0, when $n \to +\infty$, then the function $\|\boldsymbol{\theta}\|_2^k \exp\left(-f(\boldsymbol{\theta}) - (1/2)\gamma_n\|\boldsymbol{\theta}\|_2^2\right)$ is upper-bounded by $\|\boldsymbol{\theta}\|_2^k \exp\left(-f(\boldsymbol{\theta})\right)$. Thus in view of the dominated convergence theorem, we can interchange the limit and the integral. Since, by definition,

$$\mu_k(\pi_\gamma) = \frac{h_k(\gamma)}{h_0(\gamma)},$$

we get the continuity of $\mu_2(\pi_\gamma)$ and $\mu_4(\pi_\gamma)$. Let us now prove that $h_k(t)$ is continuously differentiable, when $\pi \in \mathcal{P}_{k+2}(\mathbb{R}^p)$. The integrand function in the definition of h_k is a continuously differentiable function with respect to t. In addition, its derivative is continuous and is as well integrable on $\mathbb{R}^p$, as we supposed that π has the $(k+2)$-th moment. Therefore, Leibniz integral rule yields the following

$$h_k'(\gamma) = -\frac{1}{2}\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^{k+2} \exp\left(-f(\boldsymbol{\theta}) - \gamma\|\boldsymbol{\theta}\|_2^2/2\right) d\boldsymbol{\theta} = -\frac{1}{2}h_{k+2}(t).$$

The latter yields the smoothness of h_k. Finally, in order to prove the monotony of $\mu_2(\pi_\gamma)$, we will simply calculate its derivative

$$\begin{aligned}
(\mu_2(\pi_\gamma))' &= -\frac{1}{2h_0(\gamma)}h_4(\gamma) - \frac{h_0'(\gamma)}{h_0(\gamma)^2}h_2(\gamma) \\
&= -\frac{1}{2}\mu_4(\pi_\gamma) + \frac{h_2^2(\gamma)}{2h_0(\gamma)^2} \\
&= \frac{1}{2}\left(\mu_2^2(\pi_\gamma) - \mu_4(\pi_\gamma)\right).
\end{aligned}$$

Since the latter is always negative, this completes the proof of the lemma. $\square$

Lemma 11. *Let $a > 0$ and $m > 0$. Assume $f = -\log \pi$ is m-strongly convex. Then*

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \left(\frac{p}{m}\right)^{a/2} \left\{2^{a-1}\left(1 + (1 + a/p)^{a/2-1}\right)\right\}^{\mathbb{1}_{a>2}}.$$

Proof. [DM19] proved the following bound on the second moment, centered on the mode

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \frac{p}{m}.$$

The monotony of the $\mathbb{L}_a$-norm directly yields the claim of the Lemma for $a \leq 2$.

Now, let $a > 2$. In this proof we will use the following result from [Har04]. Assume that $X \sim \mathcal{N}_p(\mu, \Sigma)$ with density φ and Y with density $\varphi \cdot \psi$ where ψ is a log-concave function. Then for any convex map $g : \mathbb{R}^p \mapsto \mathbb{R}$ we have

$$\mathbb{E}[g(Y - \mathbb{E}[Y])] \leq \mathbb{E}[g(X - \mathbb{E}[X])].$$

Since $f = -\log \pi$ is m-strongly convex, the particular choice $\mu = \mathbf{0}_p$ and $\Sigma = m\mathbf{I}_p$ yields the log-concavity of π/φ. Applied to the convex map $g : \boldsymbol{\theta} \mapsto \|\boldsymbol{\theta}\|_2^a$, the inequality of [Har04] yields

$$\mathbb{E}_\pi[\|\boldsymbol{\theta} - \mathbb{E}_\pi[\boldsymbol{\theta}]\|_2^a] \leq \mathbb{E}[\|X\|_2^a] = \left(\frac{p}{m}\right)^{a/2} \frac{\Gamma((p+a)/2)}{\Gamma(p/2)(p/2)^{a/2}}$$

using known moments of the chi-square distribution.

For any $y > 0$ the map $x \mapsto x^{-y}\Gamma(x+y)/\Gamma(x)$ goes to 1 when x goes to infinity. For convenience, we use an explicit bound from [QL+12, Theorem 4.3], that is

$$\forall y \geq 1, \qquad x^{-y}\Gamma(x+y)/\Gamma(x) \leq (1 + y/x)^{y-1}.$$

When applied to $x = p/2$ and $y = a/2 > 1$, this yields

$$\mathbb{E}_\pi[\|\boldsymbol{\theta} - \mathbb{E}_\pi[\boldsymbol{\theta}]\|_2^a] \leq \left(\frac{p}{m}\right)^{a/2} (1 + a/p)^{a/2-1}. \tag{3.11}$$

We now bound the distance between the mean and the mode

$$\|\mathbb{E}_\pi[\boldsymbol{\theta}] - \boldsymbol{\theta}^*\|_2 \leq \mathbb{E}_\pi[\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2] \leq (p/m)^{1/2}. \tag{3.12}$$

For any $x, y \geq 0$ we have $(x+y)^a \leq 2^{a-1}(x^a + y^a)$, this yields

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq 2^{a-1} \left(\mathbb{E}[\|\boldsymbol{\theta} - \mathbb{E}_\pi[\boldsymbol{\theta}]\|_2^a] + \|\mathbb{E}_\pi[\boldsymbol{\theta}] - \boldsymbol{\theta}^*\|_2^a\right)$$

Using bounds (3.11) and (3.12) in the last expression yields the claim of the lemma for $a > 2$. $\square$

Lemma 12. *Assume there exists a measurable map $m : [0, +\infty) \mapsto [0, M]$ such that for any $\boldsymbol{\theta} \in \mathbb{R}^p$, $\nabla^2 f(\boldsymbol{\theta}) \succeq m(\|\boldsymbol{\theta}\|_2)\mathbf{I}_p$. Let $a > 0$ and $A > 0$. Define the ball $B_A := \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq A\}$. We have*

$$\int_{(B_A)^c} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \frac{2(M/2)^{p/2}}{\Gamma(p/2)} \int_A^{+\infty} r^{p+a-1} e^{-\widetilde{m}(r) r^2/2} dr$$

where

$$\widetilde{m}(r) = 2 \int_0^1 m(ry)(1-y)dy.$$

Proof. Without loss of generality, we assume that $\boldsymbol{\theta}^* = \mathbf{0}_p$ and $f(\mathbf{0}_p) = 0$. Therefore, the density π is such that $\pi(\boldsymbol{\theta}) = e^{-\boldsymbol{\theta}}/C$ where

$$C = \int_{\mathbb{R}^p} e^{-f(\boldsymbol{\theta})} d\boldsymbol{\theta} \geq \int_{\mathbb{R}^p} e^{-M\|\boldsymbol{\theta}\|_2^2/2} d\boldsymbol{\theta}$$

by the fact that $\nabla^2 f(\boldsymbol{\theta}) \preceq M\mathbf{I}_p$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$.

Now, for any $r > 0$ and any $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\|\boldsymbol{\theta}\|_2 = r$, Taylor's expansion around the minimum $\mathbf{0}_p$ yields

$$\begin{aligned}
f(\boldsymbol{\theta}) - f(\mathbf{0}_p) &= \boldsymbol{\theta}^\top \left(\int_0^1 \int_0^1 \nabla^2 f(st\boldsymbol{\theta}) s dt ds \right) \boldsymbol{\theta} \\
&\geq \|\boldsymbol{\theta}\|_2^2 \int_0^1 \int_0^1 m(st\|\boldsymbol{\theta}\|_2^2) s dt ds \\
&= r^2 \int_0^1 \int_0^s m(yr) dy ds \\
&= \frac{r^2}{2} \times \underbrace{2 \int_0^1 m(yr)(1-y) dy}_{=\widetilde{m}(r)}
\end{aligned}$$

We combine this fact with the lower bound on C to get

$$\begin{aligned}
\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} &\leq C^{-1} \int_{\|\boldsymbol{\theta}\|_2 \geq A} \|\boldsymbol{\theta}\|_2^a e^{-f(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&\leq \left(\int_{\mathbb{R}^p} e^{-M\|\boldsymbol{\theta}\|_2^2/2} d\boldsymbol{\theta} \right)^{-1} \int_{\|\boldsymbol{\theta}\|_2 \geq A} \|\boldsymbol{\theta}\|_2^a e^{-\widetilde{m}(\|\boldsymbol{\theta}\|_2)\|\boldsymbol{\theta}\|_2^2/2} d\boldsymbol{\theta} \\
&= \frac{2(M/2)^{p/2}}{\Gamma(p/2)} \int_A^{+\infty} r^{a+p-1} e^{-\widetilde{m}(r) r^2/2} dr
\end{aligned}$$

where the first equality comes from a change of variables in polar coordinates, where the volume of the sphere cancels out in the ratio. □

Lemma 13. *Assume that* $\pi(\boldsymbol{\theta}) \propto e^{-f(\boldsymbol{\theta})}$*, where*

$$f(\boldsymbol{\theta}) = 0.5\|\boldsymbol{\theta}\|_2^2 \mathbb{1}_{\|\boldsymbol{\theta}\|_2 \leq 1} + \|\boldsymbol{\theta}\|_2 \mathbb{1}_{\|\boldsymbol{\theta}\|_2 > 1}.$$

Then for any $a > 0$ *and any* $p \geq 2 \vee (a-1)$

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \geq (0.1)\Gamma(p+a)/\Gamma(p) \underset{p \to +\infty}{\sim} 0.1 p^a.$$

This proves that, under assumptions of Proposition 13 (here with $m = R = 1$*), the dependence* p^a *is not improvable.*

Proof. Remark first that $f(\boldsymbol{\theta}) = \varphi(\|\boldsymbol{\theta}\|_2)$ where

$$\varphi(r) := 0.5 r^2 \mathbb{1}_{r \leq 1} + r \mathbb{1}_{r > 1}.$$

We compute explicitly the moment by a change of variable in polar coordinates

$$\begin{aligned}
\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} &= \left(\int_0^{+\infty} r^{p-1} e^{-\varphi(r)} dr \right)^{-1} \int_0^{+\infty} r^{p+a-1} e^{-\varphi(r)} dr \\
&= \frac{\Gamma(p+a) + \int_0^1 r^{p+a-1}(e^{-r^2/2} - e^{-r}) dr}{\Gamma(p) + \int_0^1 r^{p-1}(e^{-r^2/2} - e^{-r}) dr}.
\end{aligned}$$

Using the fact that $(0.2)r \leq e^{-r^2/2} - e^{-r} \leq r$ for $0 < r < 1$ yields

$$\begin{aligned}
\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} &\geq \frac{\Gamma(p+a) + 0.2/(p+a+1)}{\Gamma(p) + 1/(p+1)} \\
&\geq \frac{\Gamma(p+a) + 0.1/(p+1)}{\Gamma(p) + 1/(p+1)} \\
&\geq (0.1)\Gamma(p+a)/\Gamma(p)
\end{aligned}$$

where the second inequality follows from the fact that $a \leq p+1$ by assumption, while the last inequality follows from the fact that $\Gamma(.)$ is an increasing function on $[2, +\infty)$. This proves the claim of the Lemma. □

Lemma 14. *Let* $\Gamma(k, x)$ *be the upper incomplete Gamma function. Let* $k > 2$ *be a real number, then* $\mu_k \leq A_k \mu_2^{k/2}$ *where* $A_k = \min_{\lambda > 2, \gamma > 1} A_k(\lambda, \gamma)$ *with*

$$A_k(\lambda, \gamma) = \frac{\sqrt{\lambda - 1}}{\lambda} \left[\frac{2\sqrt{\lambda}}{\log(\lambda - 1)} \right]^k k \Gamma\Big(k, \frac{\gamma^{1/2} \log(\lambda - 1)}{2}\Big) + \frac{k(\gamma\lambda)^{k/2-1} - 2}{k-2}. \tag{3.13}$$

Proof. Let $\lambda > 1$ be fixed throughout the proof and define $\mathcal{A} = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2^2 \leq \lambda \mu_2\}$.

From Markov's inequality we have

$$\pi(\mathcal{A}) \geq 1 - \frac{\mathbb{E}_\pi[\|\boldsymbol{\vartheta}\|_2^2]}{\lambda\mu_2} = 1 - \frac{1}{\lambda}.$$

The set A being symmetric, Proposition 2.14 from [Led01] implies the following inequality:

$$1 - \pi(t\mathcal{A}) \leq \pi(\mathcal{A}) \left(\frac{1-\pi(\mathcal{A})}{\pi(\mathcal{A})}\right)^{(t+1)/2},$$

for every real number t larger than 1. Since the right-hand side is a decreasing function of $\pi(\mathcal{A})$, we obtain the following bound on $\pi(t\mathcal{A}^\complement)$:

$$\pi(t\mathcal{A}^\complement) \leq \frac{1}{\lambda \cdot (\lambda-1)^{(t-1)/2}}.$$

Let us introduce the random variable η as $\|\boldsymbol{\vartheta}\|_2/\sqrt{\mu_2}$, where $\boldsymbol{\vartheta} \sim \pi$. It is clear that (3.13) is equivalent to

$$\mathbb{E}[\eta^k] \leq \frac{\sqrt{\lambda-1}}{\lambda}\left[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\right]^k k\Gamma\Big(k, \frac{\gamma^{1/2}\log(\lambda-1)}{2}\Big) + \frac{k(\gamma\lambda)^{k/2-1}-2}{k-2}.$$

Since $\eta > 0$ almost surely,

$$\mathbb{E}[\eta^k] = \int_{\mathbb{R}} \mathbb{P}(\eta > t)dt = k\int_{\mathbb{R}} t^{k-1}\mathbb{P}(\eta > t)dt.$$

Thus, the proof of the lemma reduces to bound the tail of η. The definition of η yields

$$\mathbb{P}(\eta > t) = \mathbb{P}(\|\boldsymbol{\vartheta}\|_2 > t\sqrt{\mu_2}) = \pi\left(\frac{t}{\sqrt{\lambda}}\cdot\mathcal{A}^\complement\right) \leq \frac{1}{\lambda\cdot(\lambda-1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}},$$

when $t > \sqrt{\lambda}$. We choose $\gamma > 1$ and apply this inequality to $t > \sqrt{\gamma\lambda}$. For the other values of t, that is when $t < \sqrt{\gamma\lambda}$, we apply Markov's inequality to get $\mathbb{P}(\eta > t) \leq 1 \wedge t^{-2}$. Combining these two bounds, we arrive at

$$\begin{aligned}\mathbb{E}[\eta^k] &\leq k\int_{\sqrt{\gamma\lambda}}^\infty \frac{t^{k-1}}{\lambda\cdot(\lambda-1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}}dt + \int_0^{\sqrt{\gamma\lambda}} kt^{k-1}(1\wedge t^{-2})dt\\ &= k\int_{\sqrt{\gamma\lambda}}^\infty \frac{t^{k-1}}{\lambda\cdot(\lambda-1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}}dt + \frac{k(\gamma\lambda)^{k/2-1}-2}{k-2}.\end{aligned}$$

The first integral of the last sum can be calculated using the upper incomplete gamma

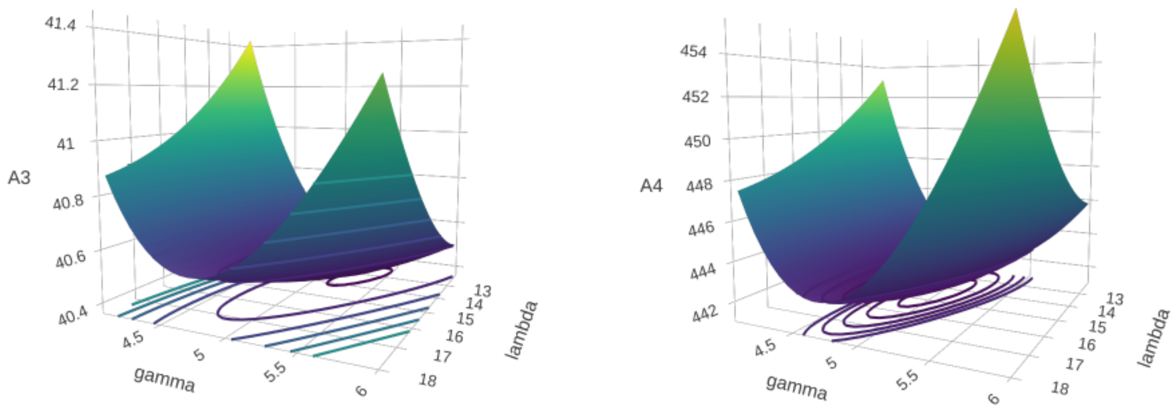


Figure 3.2: Shapes of the surfaces defined by the functions $A_3(\cdot,\cdot)$ and $A_4(\cdot,\cdot)$, see Lemma 14.

function $\Gamma(k,z)$. Indeed, the change of variable $z = t\log(\lambda-1)/(2\sqrt{\lambda})$ yields

$$\begin{aligned}\int_{\sqrt{\gamma\lambda}}^{\infty} \frac{t^{k-1}}{\lambda\cdot(\lambda-1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}}\,dt &= \frac{\sqrt{\lambda-1}}{\lambda}\int_{\sqrt{\gamma\lambda}}^{\infty} t^{k-1}\exp\Big(-\log(\lambda-1)\frac{t}{2\sqrt{\lambda}}\Big)\,dt\\ &= \frac{\sqrt{\lambda-1}}{\lambda}\bigg[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\bigg]^k\int_{\frac{\gamma^{1/2}\log(\lambda-1)}{2}}^{\infty} z^{k-1}e^{-z}\,dz\\ &= \frac{\sqrt{\lambda-1}}{\lambda}\bigg[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\bigg]^k\Gamma\Big(k,\frac{\gamma^{1/2}\log(\lambda-1)}{2}\Big).\end{aligned}$$

Finally, we obtain

$$\mathbb{E}[\eta^k] \le k\cdot\frac{\sqrt{\lambda-1}}{\lambda}\bigg[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\bigg]^k\Gamma\Big(k,\frac{\gamma^{1/2}\log(\lambda-1)}{2}\Big) + \frac{k(\gamma\lambda)^{k/2-1}-2}{k-2}.$$

This concludes the proof. □

Remark 6. *We plotted*[5] *in Figure 3.2 the plots of the function A_k for $k=3$ and $k=4$. Numerically, we find that the optimal choice for (λ,γ) is approximately $\lambda = 15.89$ and $\gamma = 4.4$ for $k=3$ and $\lambda = 14.97$ and $\gamma = 4.8$ for $k=4$. This leads to the numerical bounds*

$$A_k \le \begin{cases} 40.40, & k=3\\ 441.43, & k=4\end{cases}.$$

These constants are by no means optimal, but we are not aware of any better bound available in the literature. Inequalities of type $\mathbb{E}[\|\boldsymbol{X}\|_2^k] \le A_k\mathbb{E}[\|\boldsymbol{X}\|_2^2]^{k/2}$ are often referred to as the

[5]The R notebook for generating this figure can be found here `https://rpubs.com/adalalyan/Khintchine_constant`

Kintchine inequality [Khi23]. According to [CG18], [Bob99, Corollary 4.3] implies that $A_4 \leq 49$ *for one-dimensional* $\boldsymbol{X}$ *with log-concave density. Getting such a small constant in the multidimensional case would be of interest for applications to MCMC sampling.*

Chapter 4

Penalized Langevin Dynamics

Abstract

We study the problem of sampling from a probability distribution on $\mathbb{R}^p$ defined via a convex and smooth potential function. We first consider two continuous-time diffusion-type processes, termed Penalized Langevin dynamics (PLD) and Penalized Kinetic Langevin dynamics (PKLD), the drift of which is the negative gradient of the potential plus a linear penalty that vanishes when time goes to infinity. An upper bound on the Wasserstein-2 distance between the distribution of the PLD at time t and the target is established. This upper bound highlights the influence of the speed of decay of the penalty on the accuracy of approximation. As a consequence, considering the low-temperature limit we infer a new non-asymptotic guarantee of convergence of the penalized gradient flow for the optimization problem.
This chapter is based on a joint paper with Arnak Dalalyan called "Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets". It is accepted to *Advances in Neural Information Processing Systems (NeurIPS 2020)*.

4.1 Introduction

The problem of sampling from a probability distribution received a great deal of attention in machine learning literature. Gradient based MCMC methods such as the Langevin MC, the underdamped Langevin Monte Carlo, the Hamiltonian Monte Carlo and their Metropolis adjusted counterparts were shown to have attractive features both in practice and in theory. In particular, thanks to a large number of recent results, the case of smooth and strongly log-concave densities is now fairly well understood. In this case, non-asymptotic theoretical guarantees for various distances on probability distributions have been established, showing that the number of gradient evaluations necessary to achieve an error upper bounded by ε is a low order polynomial of the dimension, the condition number and the inverse precision $1/\varepsilon$. The dependence on the latter is even logarithmic for Metropolis adjusted methods.

The main focus of this paper is on the problem of sampling from densities[1]

$$\pi(\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta})), \qquad \boldsymbol{\theta} \in \mathbb{R}^p,$$

corresponding to a (weakly) convex potential function $f : \mathbb{R}^p \to \mathbb{R}$. In the sequel, a twice differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is said to satisfy (m, M)-SCGL condition (m-strongly convex and M-gradient Lipschitz), for some $M \geq m \geq 0$, if the following inequality is satisfied:

$$m\mathbf{I}_p \preceq \nabla^2 f(\boldsymbol{\theta}) \preceq M\mathbf{I}_p, \qquad \forall \boldsymbol{\theta} \in \mathbb{R}^p.$$

Here, for two square matrices $\mathbf{A}$ and $\mathbf{B}$, the relation $\mathbf{A} \preceq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semidefinite.

In this work, we wish to define a class of continuous-time processes, such that for every element $\{\boldsymbol{L}_t : t \geq 0\}$ of the class, the distribution of $\boldsymbol{L}$ at time t is close to the target distribution π. When the potential function f satisfies (m, M)-SCGL condition with $m \geq 0$, it is well-known that the vanilla Langevin dynamics $\boldsymbol{L}^{\mathsf{LD}}$, defined as the solution of

$$d\boldsymbol{L}_t^{\mathsf{LD}} = -\nabla f(\boldsymbol{L}_t^{\mathsf{LD}})dt + \sqrt{2}\, d\boldsymbol{W}_t, \tag{LD}$$

where $\boldsymbol{W}_t$ is a standard Wiener process independent of $\boldsymbol{L}_0^{\mathsf{LD}}$, has π as invariant distribution [Bha78]. Furthermore, when $m > 0$, the distribution ν_t^{LD} of $\boldsymbol{L}_t^{\mathsf{LD}}$ converges in Wasserstein distance (see below for a definition) exponentially fast to π [Vil08], that is

$$W_2(\nu_t^{\mathsf{LD}}, \pi) \leq e^{-mt} W_2(\nu_0^{\mathsf{LD}}, \pi). \tag{4.1}$$

A remarkable feature of this result is that it is dimension free. Another class of continuous time processes that has similar properties is the kinetic Langevin dynamics. It is defined as a system of two SDEs:

$$\begin{aligned} d\boldsymbol{L}_t^{\mathsf{KLD}} &= \boldsymbol{V}_t^{\mathsf{KLD}} dt; \\ d\boldsymbol{V}_t^{\mathsf{KLD}} &= -\Big(\eta \boldsymbol{V}_t^{\mathsf{KLD}} + \nabla f(\boldsymbol{L}_t^{\mathsf{KLD}})\Big)\, dt + \sqrt{2\eta}\boldsymbol{W}_t, \end{aligned} \tag{KLD}$$

where η is called the coefficient of friction. Important feature of this diffusion process is that it has an invariant distribution $\mathsf{P}(\boldsymbol{\theta}, \boldsymbol{v}) \propto \exp(-f(\boldsymbol{\theta}) - \|\boldsymbol{v}\|_2^2/2)$ (see [Pav14]). In addition, the ergodicity of the process is also well studied by many authors (see, e.g.

[1]We will use the same notation for the probability density functions and corresponding distributions.

[CCBJ18, EGZ19]). Recently [DRD20] have proved the following result.

Proposition 16 (Theorem 1 [DRD20]). *Suppose* $(\boldsymbol{L}, \boldsymbol{V}) \sim \nu_t$ *and* $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{V}}) \sim \widehat{\nu}_t$ *are two solutions of* (KLD)*, with initial distributions* $\nu_0 = \nu_0^{\mathsf{L}} \otimes \nu_0^{\mathsf{V}}$ *and* $\widehat{\nu}_0 = \widehat{\nu}_0^{\mathsf{L}} \otimes \nu_0^{\mathsf{V}}$*, respectively. The latter means that at time moment* 0 *the vectors* $\boldsymbol{V}_0$ *and* $\widehat{\boldsymbol{V}}_0$ *have the same distribution. Then, if* f *satisfies* (m, M)-SCGL*, then we have an exponential contraction in Wasserstein distance:*

$$W_2\left(\nu_t, \widehat{\nu}_t\right) \leq \alpha e^{-\beta t} W_2\left(\nu_0, \widehat{\nu}_0\right).$$

Here α *and* β *are time-independent constants depending on the parameters* m, M, η*.*

In the strongly convex case β is positive thus the result provides us with exponential convergence of the continuous-time diffusion. However, in the case when $m = 0$ the constant β that appears in the exponential, can be negative. Thus, we cannot make use of this inequality, when is non-strongly convex. Summing up, we have seen two methods that perform rapid convergence in Wasserstein distance when the potential function is strongly convex and that for both of them the convergence in the non-strongly convex case is left open.

It was established by [Bob99] that log-concave distributions satisfy the Poincaré inequality with the Poincaré constant $\mathcal{C}_{\mathsf{P}}$ that might depend on the dimension. According to [CLGL$^+$20], the exponentially fast convergence of (LD) to zero holds true with m replaced by $1/\mathcal{C}_{\mathsf{P}}$, when $m = 0$. In [KLS95], the authors conjectured that there is a universal constant $\mathcal{C}_{\mathsf{KLS}} > 0$ such that for any log-concave distribution π on $\mathbb{R}^p$, $\mathcal{C}_{\mathsf{P}} \leq \mathcal{C}_{\mathsf{KLS}} \|\mathbf{E}_{\boldsymbol{X} \sim \pi}[\boldsymbol{X}\boldsymbol{X}^\top]\|_{\text{op}}$, where $\|\mathbf{A}\|_{\text{op}}$ stands for the operator norm of the matrix $\mathbf{A}$. Despite important efforts made in recent years (see [AGB15, CG18, Che20]), this conjecture is still unproved. Note also that the Poincaré constant is, in general, hard to approximate and to estimate [PVBL$^+$19]. One approach to getting more tractable convergence bounds could be to find a tractable upper bound on the Poincaré constant of a distribution defined by a potential satisfying (m, M)-SCGL condition with $m = 0$. We develop here another approach, consisting in modifying the Langevin dynamics, so that the new dynamics has still a limiting distribution equal to π but for which we can get a tractable upper bound. A natural way of defining this new dynamics is to add to the potential f a strongly-convex penalty with a strong-convexity constant that vanishes when time goes to infinity. For a quadratic penalty function, this leads to the process $\boldsymbol{L}^{\mathsf{PLD}}$ termed penalized Langevin dynamics and defined by[2]

$$d\boldsymbol{L}_t^{\mathsf{PLD}} = -\big(\nabla f(\boldsymbol{L}_t^{\mathsf{PLD}}) + \alpha(t)\boldsymbol{L}_t^{\mathsf{PLD}}\big)\, dt + \sqrt{2}\, d\boldsymbol{W}_t, \qquad \text{(PLD)}$$

where $\alpha : [0, \infty) \to [0, \infty)$ is a time-dependent penalty factor tending to zero as $t \to \infty$.

[2]If a good initial guess $\boldsymbol{\theta}_0$ of the minimum point of f is available, it might be better to replace the penalty term by $\alpha(t)(\cdot - \boldsymbol{\theta}_0)$.

The main result of this work is an upper bound on $W_2(\nu_t^{\mathsf{PLD}}, \pi)$ that is valid for every continuously differentiable and decreasing penalty factor α. Optimizing over α, we show that the choice $\alpha(t) \sim 1/(2t)$, when $t \to \infty$, leads to a simple upper bound of the order $1/\sqrt{t}$.

Similarly for (KLD) we modify the drift term $-\nabla f(\boldsymbol{x})$ by adding a linear time-dependent penalty term $-\alpha(t)\boldsymbol{x}$. The Penalized Kinetic Langevin Dynamics (PKLD) is a stochastic process which is defined as the solution of the following system of stochastic differential equations:

$$\begin{aligned} d\boldsymbol{L}_t^{\mathsf{PKLD}} &= \boldsymbol{V}_t^{\mathsf{PKLD}} dt; \\ d\boldsymbol{V}_t^{\mathsf{PKLD}} &= -(\eta \boldsymbol{V}_t^{\mathsf{PKLD}} + \nabla f(\boldsymbol{L}_t^{\mathsf{PKLD}}) + \alpha(t)\boldsymbol{L}_t^{\mathsf{PKLD}}) + \sqrt{2\eta}\boldsymbol{W}_t. \end{aligned} \tag{PKLD}$$

Interestingly, using a suitably parametrized temperature-dependent potential function $f_\tau(\cdot) = f(\cdot)/\tau$ and a penalty factor, one can get an upper bound for the penalized gradient flow

$$\dot{\boldsymbol{X}}_t^{\mathsf{PGF}} = -\big(\nabla f(\boldsymbol{X}_t^{\mathsf{PGF}}) + \alpha_0(t)\boldsymbol{X}_t^{\mathsf{PGF}}\big), \qquad t \geq 0, \tag{PGF}$$

by passing to the low-temperature limit. This bound implies that $\|\boldsymbol{X}_t^{\mathsf{PGF}} - \boldsymbol{x}^*\|_2$ can be of the order $O(1/t^{1-\mathsf{q}})$, where $\boldsymbol{x}^*$ is a minimizer of the potential f and $\mathsf{q} \in [0, 1]$ is a parameter appearing in an additional condition imposed on f. To the best of our knowledge, the obtained bound is new, most previous results being valid for the objective function itself, not for the minimum point.

The rest of the paper is organized as follows. We start by stating the bound on the error of the PLD in Section 4.2. We also instantiate the bound to the penalty factors that are inversely proportional to time. In Section 4.3, we discuss the connections with the optimization problem, assessing the error of the PGF. Section 4.4 is devoted to relation to prior work. The proof of the main result, up to some technical lemmas, is presented in Section 4.A. Missing proofs are deferred to the supplementary material.

To complete this introduction, we introduce some notations. We consider the Wasserstein-2 distance

$$W_2(\nu, \nu') = \inf\Big\{\mathbf{E}[\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2^2]^{1/2} : \boldsymbol{\vartheta} \sim \nu \text{ and } \boldsymbol{\vartheta}' \sim \nu'\Big\},$$

where the minimum is over all joint distributions having ν and ν' as the first and the second marginal distributions. For any $\gamma > 0$, we define the probability density function $\pi_\gamma(\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta}) - \gamma\|\boldsymbol{\theta}\|_2^2)$, where $\|\boldsymbol{\theta}\|_2$ is the Euclidean norm. We also define $\mu_k(\pi) = \mathbb{E}_{\boldsymbol{X}\sim\pi}[\|\boldsymbol{X}\|_2^k]$, the moment of order k of π.

4.2 Convergence of penalized Langevin dynamics

In this section we explain our approach and state the main result. Without loss of generality, we will assume that the initial point for the PLD is the origin, $\boldsymbol{L}_0^{\mathsf{PLD}} = 0$. Note that if a good guess $\boldsymbol{\theta}_0$ of a minimizer of f is available, it is recommended to initialize PLD at $\boldsymbol{\theta}_0$. Our framework covers this case, since it suffices to apply our results to the translated function $\widetilde{f}(\cdot) = f(\boldsymbol{\theta}_0 + \cdot)$. Under the condition $\boldsymbol{L}_0^{\mathsf{PLD}} = 0$, the Wasserstein-2 distance at the starting point coincides with the second-order moment, $W_2(\nu_0^{\mathsf{PLD}}, \pi) = \sqrt{\mu_2(\pi)}$.

When $\alpha(t) = \alpha$ is a strictly positive constant, the distribution ν_T^{PLD}, for a large value of T, is close to the biased target π_α. Furthermore, in view of (4.1), the distance between these two distributions is smaller than a prescribed error level $\varepsilon > 0$ as soon as $T \geq (1/\alpha)\log(\sqrt{\mu_2(\pi)}/\varepsilon)$. On the other hand, one can choose α small enough such that the bias $W_2(\pi_\alpha, \pi)$ is smaller than ε. The discrete counterpart of this approach has been used in many recent works [Dal17b, DRDK19, CDWY18]. The approach we develop here extends these work to the case of time-dependent α and has the advantage of being asymptotically unbiased, when $t \to \infty$. In other words, it allows to choose α independently of ε and make the error smaller than ε by running PLD over a sufficiently large time period.

Theorem 14. *Suppose that π is a probability distribution with a potential function f that satisfies (m, M)-SCGL condition, where $m \geq 0$ and $M > 0$. Let $\alpha : [0, +\infty) \to \mathbb{R}$ be a non-increasing differentiable function, such that $m + \alpha(t) > 0$ for every $t \geq 0$. Then, for every positive number t and for $\beta(t) = \int_0^t (m + \alpha(u))\,du$, we have*

$$W_2(\nu_t^{\mathsf{PLD}}, \pi) \leq \sqrt{\mu_2(\pi)}\, e^{-\beta(t)} + 11\mu_2(\pi)\bigg\{e^{-\beta(t)}\int_0^t \frac{|\alpha'(s)|e^{\beta(s)}}{\sqrt{m+\alpha(s)}}\,ds + \frac{\alpha(t)}{\sqrt{m+\alpha(t)}}\bigg\}. \quad (4.2)$$

The proof being postponed to Section 4.A, the rest of this section is devoted to discussing the stated theorem and its consequences for some specific choices of the penalty factor. One can notice right away that in the case of a positive m, we can choose α to be zero, thereby obtaining the classical exponential convergence rate [Vil08]. In the rest of the discussion, we assume that $m = 0$.

The numerical constant 11 can certainly be improved. It is closely related to the fact that for a log-concave distribution ν, we have $\mu_4(\nu) \leq C_4\mu_2^2(\nu)$ for a universal constant C_4. proved to satisfy $C_4 \leq 442$ [DRDK19, Remark 3]. Improved bounds on C_4 will automatically yield improved numerical constant in Theorem 14. We also note that the Lipschitz constant M does not appear in inequality (4.2). Our proof, however, requires the finiteness of M. We believe that it is possible to relax the gradient-Lipschitz assumption by requiring from ∇f to be only locally Lipschitz-continuous. One can check that if we replace the penalty factor α by a larger function, the first term of the upper bound in

(4.2), proportional to $\exp\{-\int_0^t \alpha(s)\,ds\}$, becomes smaller. On the other hand, the second term increases when α increases [3] One can thus use inequality (4.2) for choosing the penalty factor α that offers a trade-off between the error of approximating the biased density $\pi_{\alpha(t)}$ by the PLD and the error of approximating the target π by the biased density $\pi_{\alpha(t)}$.

To "optimize" the upper bound with respect to α, let us for the moment ignore the first term in the curly parentheses in (4.2). In that case our functional of interest has two components, where one of them is increasing with respect to α, while the other is decreasing. (Here the monotony must be understood with a certain precaution, as in our case the mathematical concept is not well-defined.) These considerations suggest to choose the "optimal" α by balancing these two terms:

$$\sqrt{\mu_2(\pi)}\, e^{-\beta(t)} = 11\sqrt{\alpha(t)}\,\mu_2(\pi). \tag{4.3}$$

Taking the square of both sides, cancelling out some terms and using that $\alpha(t) = \beta'(t)$, we check that (4.3) is equivalent to

$$121\mu_2(\pi)\beta'(t)e^{2\beta(t)} = 1.$$

Solving this differential equation we get the following expression for $\beta(\cdot)$ and the corresponding expression for $\alpha(\cdot)$:

$$\beta^*(t) = \frac{1}{2}\log\left(\frac{2t}{121\mu_2(\pi)} + 1\right) \quad \text{and} \quad \alpha^*(t) = \frac{1}{2t + 121\mu_2(\pi)}.$$

It is easy to check that this choice of α ensures that the first and the last terms in the right hand side of (4.2) are of the order $1/\sqrt{t}$. Interestingly, the middle term in (4.2) turns out to be of the same order, up to a logarithmic factor. The precise statement of the consequence of Theorem 14 obtained by choosing $\alpha(t) = 1/(2t + A)$ for some $A > 0$ reads as follows.

Proposition 17. *If the potential function satisfies the (m, M)-SCGL condition with $m = 0$, then the error of PLD with $\alpha(t) = 1/(A + 2t)$, measured by the Wasserstein-2 distance, satisfies*

$$W_2(\nu_t^{\mathsf{PLD}}, \pi) \leq \frac{\sqrt{A\mu_2(\pi)} + 11\mu_2(\pi)\big(1 + \log\left(1 + (2/A)t\right)\big)}{\sqrt{A + 2t}}.$$

[3]This is clear for the last term, proportional to $\sqrt{\alpha(t)}$, whereas the corresponding claim for the first term in the curly parentheses less trivial.

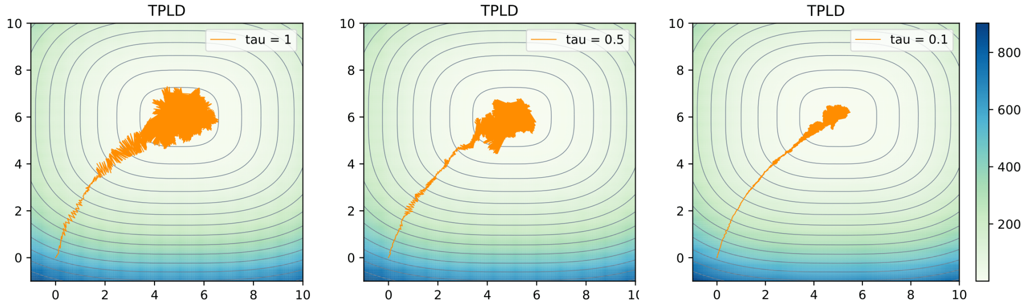


Figure 4.1: One random path of the TPLD, for $\tau = 1$ (left), $\tau = 0.5$ (middle) and $\tau = 0.1$ (right).

In particular, for $A = 2\mu_2(\pi)$, we get

$$W_2(\nu_t^{\mathsf{PLD}}, \pi) \le \frac{10\mu_2(\pi)\big\{1 + \log\big(1 + (t/\mu_2(\pi))\big)\big\}}{\sqrt{\mu_2(\pi) + t}}.$$

To complete this section, we present a quick argument showing that the right hand side of (4.2) cannot converge to zero faster than the rate $1/\sqrt{t}$. Suppose that for a specific choice of α, the right hand side of (4.2) is $o(1/t^{1/2})$. Then the last term is necessarily $o(1/t^{1/2})$, which implies that $\alpha(t) = o(1/t)$ when $t \to +\infty$. Thus, for some $c > 0$, $\alpha(t) \le 1/(4t)$ for every $t \ge c$. This means that $\beta(t) \le c\alpha(0) + ({}^1\!/\!{}_4)\log(t/c)$. Hence,

$$\exp(-\beta(t)) \ge (c/t)^{1/4}\exp(-c\alpha(0)).$$

This proves that the upper bound on the error of the PLD established in Theorem 14 may not tend to zero at a faster rate than $1/\sqrt{t}$, as t goes to infinity. This argument also shows that the optimizer $\alpha(\cdot)$ of the upper bound is asymptotically equivalent to $1/(2t)$ when $t \to \infty$.

4.2.1 Penalized Kinetic Langevin Dynamics

We define by ν_t^{PKLD} the distribution of the couple $(\boldsymbol{L}_t^{\mathsf{PKLD}}, \boldsymbol{V}_t^{\mathsf{PKLD}})$. Also by P_γ we denote the probability distribution on $\mathbb{R}^{2p}$, which has a density

$$\mathsf{P}_\gamma : (\boldsymbol{\theta}, \boldsymbol{v}) \to \exp(-f(\boldsymbol{\theta}) - \|\boldsymbol{v}\|_2^2/2 - \gamma\|\boldsymbol{\theta}\|_2^2/2).$$

We notice that by definition $\mathsf{P} = \mathsf{P}_0$.

Theorem 15. *Suppose that π is a probability distribution with a potential function f that satisfies (m, M)-SCGL condition, where $m \ge 0$ and $M > 0$. Let $\alpha : [0, +\infty) \to \mathbb{R}$ be a non-increasing differentiable function, such that $m + \alpha(t) > 0$ for every $t \ge 0$. We define the* *function β as below:*

$$\beta(t) := \int_0^t 2\eta\alpha(u)du.$$

Then, for every $t > 0$, we have

$$W_2(\nu_t^{\mathsf{PKLD}}, \mathsf{P}) \le \sqrt{\mu_2(\mathsf{P})}\, e^{-\beta(t)} + 11\mu_2(\mathsf{P})\left\{e^{-\beta(t)}\int_0^t \frac{|\alpha'(s)|e^{\beta(s)}}{\sqrt{m+\alpha(s)}}\,ds + \frac{\alpha(t)}{\sqrt{m+\alpha(t)}}\right\}.$$

Remark. The proof is postponed to the Section 4.B. It is based on a linear transformation of the vector $(\boldsymbol{L}_t^{\mathsf{PKLD}}, \boldsymbol{V}_t^{\mathsf{PKLD}})$, which was introduced in the proof of [DRD20][Theorem 1]. However, Theorem 15 is valid for any initial distribution ν_0^{PKLD}, while as to have a similar bound using the results from [DRD20] one needs to take $\boldsymbol{V}_0$ to be a standard normal vector independent from $\boldsymbol{L}_0^{\mathsf{PKLD}}$.

At the first glance it may seem surprising that we get almost the same upper bound on Wasserstein error for (PKLD) and (PLD). However, this is the case also for strongly convex potentials. As a consequence of [DRD20][Theorem 1], the authors deduce the following inequality:

$$W_2\left(\nu_t^{\boldsymbol{L}}, \widehat{\nu}_t^{\boldsymbol{L}}\right) \le \left(\frac{2M-m}{M-m}\right)^{1/2} \exp\left\{-\left(\sqrt{M}-\sqrt{M-m}\right)t\right\} W_2\left(\nu_0, \widehat{\nu}_0\right).$$

Here ν_t and $\widehat{\nu}_t$ are the distributions of two solutions of (PKLD) that (as mentioned in the remark above) have the same initial marginal distribution of the velocity component. Assuming that $M = O(1)$ and ignoring the multiplier in front, we have that

$$W_2\left(\nu_t^{\boldsymbol{L}}, \widehat{\nu}_t^{\boldsymbol{L}}\right) \lesssim \exp\left\{-mt\right\} W_2\left(\nu_0, \widehat{\nu}_0\right).$$

The right-hand side coincides with the upper bound described in (4.1). In view of the analysis done for Theorem 14, we deduce that the optimal choice of α is of order $O(1/t)$. The latter yields $O(1/t^{1/2})$ convergence in Wasserstein distance for (PKLD).

4.3 The counterpart in optimization: penalized gradient flow

In this section we draw the parallel between PLD and the penalized gradient flows, henceforth referred to as PGF, for a non-strongly convex function f. In the case of strongly convex functions, the gradient flow $\boldsymbol{X}_t^{\mathsf{GF}}$ defined by the differential equation $\dot{\boldsymbol{X}}_t^{\mathsf{GF}} = -\nabla f(\boldsymbol{X}_t^{\mathsf{GF}})$, converges exponentially fast to the minimum $\boldsymbol{x}_*$ of f, without the need to add a quadratic penalty. In contrast with this, for general non-strongly convex case functions f, only the convergence of the function $f(\boldsymbol{X}_t^{\mathsf{GF}})$ to $f(\boldsymbol{x}_*)$ at the rate $1/t$ can be established. The goal of this sections is to understand the convergence of the flow to the minimum point when a vanishing quadratic penalty is added to the cost function f; when does this flow converge, what is the impact of the penalty factor and what kind of rate can be achieved. To answers these questions, we assume in this section that $(0, M)$-SCGL holds true. We also assume that f has a unique minimum point denoted by $\boldsymbol{x}_*$.

In the analysis performed in the previous section, we can replace the function $f(\cdot)$ by the function $f_\tau(\cdot) = f(\cdot)/\tau$. The function f_τ has $\boldsymbol{x}_*$ as its point of minimum, whatever the real number $\tau > 0$. Moreover, if we define the tempered density function $\pi^\tau(\boldsymbol{\theta}) \propto \exp\left(-f_\tau(\boldsymbol{\theta})\right)$, the distribution π^τ tends to $\delta_{\boldsymbol{x}_*}$, the Dirac mass at $\boldsymbol{x}_*$. Clearly, f_τ satisfies $(0, M/\tau)$-SCGL condition. Thus, according to Theorem 14, the process $\boldsymbol{L}_t^\tau$, defined as

$$\boldsymbol{L}_t^\tau = \boldsymbol{L}_0^\tau - \frac{1}{\tau}\int_0^t \big(\nabla f(\boldsymbol{L}_s^\tau) + \alpha(s/\tau)\boldsymbol{L}_s^\tau\big) ds + \sqrt{2}\,\boldsymbol{W}_t,$$

converges to π^τ in Wasserstein distance, if $\alpha(\cdot)$ is a continuously differentiable and non-increasing function. We now introduce the tempered penalized Langevin dynamics (TPLD), as a time-scaled version of $\boldsymbol{L}^\tau$: $\boldsymbol{X}_t^{\mathsf{TPLD}} = \boldsymbol{L}_{t\tau}^\tau$ for every $\tau > 0$. One can check that this process satisfies the stochastic differential equation

$$d\boldsymbol{X}_t^{\mathsf{TPLD}} = -\big(\nabla f(\boldsymbol{X}_t^{\mathsf{TPLD}}) + \alpha(t)\boldsymbol{X}_t^{\mathsf{TPLD}}\big)\, dt + \sqrt{2\tau}\, d\bar{\boldsymbol{W}}_t, \tag{TPLD}$$

where $\bar{\boldsymbol{W}}_t = \tau^{-1/2}\boldsymbol{W}_{\tau t}$ is a standard Wiener process. To illustrate the behaviour of this process, Figure 4.1 shows one realization of TPLD for different values of τ, with the left plot corresponding to PLD. All the results of the previous section continue to hold for this tempered dynamics. In particular, the analog of the second claim of Proposition 17 in the case of the tempered diffusion takes the following form.

Proposition 18. *If the potential function satisfies the $(0, M)$-SCGL condition, then the* *error of* TPLD *with* $\alpha(t) = 1/(2\mu_2(\pi^\tau) + 2t)$ *satisfies*

$$W_2(\nu_t^{\mathsf{TPLD}}, \pi^\tau) \le \frac{10\mu_2(\pi^\tau)\big\{1 + \log\big(1 + (t/\mu_2(\pi^\tau))\big)\big\}}{\sqrt{\tau(\mu_2(\pi^\tau) + t)}}, \qquad \forall t \ge 0.$$

As mentioned above, for small τ, π^τ is close to the Dirac mass at the minimum point $\boldsymbol{x}_*$. The last result tells us that we can approximate π^τ arbitrarily well, by running the TPLD over a large time-period. But we can not replace τ by zero in this result, since the denominator of the right hand side vanishes and the result becomes vacuous. We show below that this can be repaired if an additional assumption is introduced.

Taking $\tau = 0$ in (TPLD), we get the penalized gradient flow (PGF):

$$d\boldsymbol{X}_t^{\mathsf{PGF}} = -\big(\nabla f(\boldsymbol{X}_u^{\mathsf{PGF}}) + \alpha(u)\boldsymbol{X}_u^{\mathsf{PGF}}\big)\, du, \qquad t \ge 0, \qquad \boldsymbol{X}_0^{\mathsf{PGF}} = \mathbf{0}.$$

Here we recognize the analog of PLD in the setting of the gradient flows. On the other hand, the Euclidean distance on $\mathbb{R}^p$ is equal to the Wasserstein distance between Dirac measures. This leads us to think that our approach for obtaining non-asymptotic error bounds for PLD is applicable to the penalized gradient flow. This turns out to be true, modulo the introduction of the following assumption.

Assumption $\mathbf{A}(\mathsf{D}, \mathsf{q})$: The minimum point $\boldsymbol{x}_\gamma$ of the (strongly convex and coercive) function $f_\gamma(\cdot) = f(\cdot) + \gamma\|\cdot\|_2^2/2$ satisfies

$$\|\boldsymbol{x}_\gamma - \boldsymbol{x}_{\tilde\gamma}\|_2 \le \frac{\mathsf{D}}{\tilde\gamma^{\mathsf{q}}}(\tilde\gamma - \gamma)\|\boldsymbol{x}_*\|_2, \qquad \forall \gamma < \tilde\gamma,$$

for some $\mathsf{D} > 0$ and $\mathsf{q} \in [0, 1]$.

Since $\boldsymbol{x}_\gamma$ a stationary point of f_γ, we have $\nabla f(\boldsymbol{x}_\gamma) = -\gamma\boldsymbol{x}_\gamma$. From this relation and [Nes04, Theorem 2.1.12], one can deduce that (a) if f satisfies (m, M)-SCGL condition with $m > 0$ then $\mathbf{A}(1/m, 0)$ holds and (b) if f satisfies $(0, M)$-SCGL condition then $\mathbf{A}(1, 1)$ holds.

Theorem 16. *Assume that* $\alpha : [0, \infty) \to [0, \infty)$ *is a continuously differentiable and non-increasing function. Let* $\beta(t) = \int_0^t \alpha(s)\, ds$ *be the antiderivative of* α*. If* f *satisfies* $\mathbf{A}(\mathsf{D}, \mathsf{q})$ *and* $(0, M)$-SCGL*, then*

$$\|\boldsymbol{X}_t^{\mathsf{PGF}} - \boldsymbol{x}_*\|_2 \le \|\boldsymbol{x}_*\|_2\Big(e^{-\beta(t)} + \mathsf{D}\int_0^t \frac{|\alpha'(s)|}{\alpha^{\mathsf{q}}(s)} e^{\beta(s)-\beta(t)} ds + \mathsf{D}\alpha^{1-\mathsf{q}}(t)\Big). \tag{4.4}$$

The proof can be found in Section 4.G of the supplementary material. When $\mathsf{q} = 1/2$, this result is the optimization counterpart of the inequality shown in Theorem 14. Once again, it is appealing to optimize the right hand side of (4.4) in order to choose the

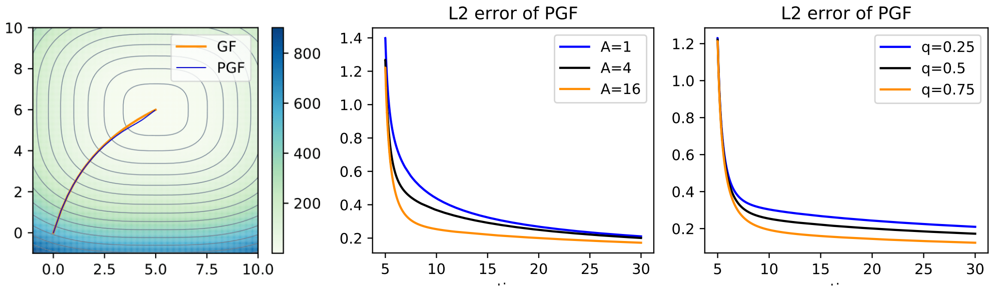


Figure 4.2: Left: The gradient flow (orange) and the penalized gradient flow (blue). Middle and Right: the mapping $t \mapsto \|\boldsymbol{X}_t^{\mathsf{PGF}} - \boldsymbol{x}_*\|_2$. In both plots we used the function $f(x, y) = |x-5|^3 + 2|y-6|^3$.

"best" penalty factor. Using arguments similar to those of previous section, *i.e.*, balancing the first and the last terms on the right hand side of (4.4), we get that the "optimal" convergence of PGF is achieved when

$$\alpha^*(t) = \frac{(1-\mathsf{q})}{t + \mathsf{D}^{1/(1-\mathsf{q})}(1-\mathsf{q})} \qquad \text{and} \qquad \beta^*(t) = (1-\mathsf{q}) \log\left(\frac{t}{\mathsf{D}^{1/(1-\mathsf{q})}(1-\mathsf{q})} + 1\right).$$

This leads us to make the recommendation of choosing $\alpha(t) = (1-\mathsf{q})/(t + A)$, for some positive A. If $\mathsf{q} = 1$, this amounts to considering the non-penalized gradient flow and (4.4) boils down to the fact that the distance from the gradient flow of a convex function to its minimum decreases. While for $\mathsf{q} < 1$, for the foregoing choice of the penalty factor, we get the error bound

$$\|\boldsymbol{X}_t^{\mathsf{PGF}} - \boldsymbol{x}_*\|_2 \leq \frac{A^{1-\mathsf{q}} + \mathsf{D} + \mathsf{D}\log(1 + (t/A))}{(t+A)^{1-\mathsf{q}}} \|\boldsymbol{x}_*\|_2. \tag{4.5}$$

To complete this section, we make some remarks on assumption $\mathbf{A}(\mathsf{D}, \mathsf{q})$. First, one can relax this assumption by requiring that the desired inequality holds for sufficiently small values of $\widetilde{\gamma}$ only. In this form, it can be easily seen that larger values of q correspond to weaker assumption. Second, even if the function f is not strongly convex, it may satisfy $\mathbf{A}(\mathsf{D}, 0)$. An example is the function $f(x) = \sqrt{(x - x_*)^2 + b^2}$. The second derivative of this function is equal to $b^2/((x-x_*)^2 + b^2)^{3/2}$. This implies that f satisfies $(0, 1/b)$-SCGL. We show in the supplementary material that it satisfies $\mathbf{A}(\mathsf{D}, 0)$ for some finite value of $\mathsf{D} > 0$. This is not really surprising, given that this function is strongly convex on any compact set. Another instructive example is the function $f(x) = |x - x_*|^a$ with $a \geq 2$. On compact sets, this function satisfies[4] $\mathbf{A}(\mathsf{D}, (a-2)/(a-1))$. Therefore, the error bound (4.5) implies that PGF converges to the minimum of f at the rate $1/t^{1/(a-1)}$, which is faster than the rate derived from the standard $O(1/t)$ bound for the non-penalized gradient flow. This

[4]See supplementary material.

behaviour is depicted in Figure 4.2.

4.4 Prior work and outlook

The relation of our results to some prior work has already been highlighted in previous sections. This section provides some complementary bibliographic remarks on recent advances on Langevin diffusions, gradient flows and their discrete counterparts.

Convergence of Langevin dynamics in continuous time has received a lot of attention in probability, see [CG09, CGR10, BGG12] and the references therein. An interesting known fact, for instance, is that the Langevin dynamics satisfies[5] $W_2(\nu^{\mathsf{LD}}_{t,\boldsymbol{x}}, \nu^{\mathsf{LD}}_{t,\boldsymbol{y}}) \le e^{-mt}\|\boldsymbol{x}-\boldsymbol{y}\|_2$ if and only if f is m-strongly convex. More recently, many papers in statistics and machine learning literature established non-asymptotic error bounds for discretized algorithms, mainly focusing on the convex case, see [**?** DM17, HKRC18, BEL18, SL19] in addition to previously cited papers. The non-convex case was emphasized in [CCA+18, MMS18, EMS18, MV19b].

The kinetic version of Langevin dynamics was proposed by [Kra40] in order to describe the motion of a particle with position $\boldsymbol{L}^{\mathsf{KLD}}_t$ and velocity $\boldsymbol{V}^{\mathsf{KLD}}$ in a chemical reaction. It was shown by [Nel67] that (LD) is a limit version of the rescaled kinetic diffusion $\bar{\boldsymbol{L}}_t := \boldsymbol{L}^{\mathsf{KLD}}_{\eta t}$, when $\eta \to \infty$. The process, in particular its behavior when the time t tends to infinity, has been studied by many (see e.g. [Bro97, Vil09, MM16, EGZ19, Nel67, MSH02]). Recently, with the rise of interest towards the Langevin sampling schemes, various discretized versions of (PKLD) have been studied, see [CCBJ18, SZTG20, EGZ19, Mon20, SL19].

In the optimization setting, the results on the convergence of the gradient flow for convex objectives have been known for a long time. More recently, [SBC16] derived a continuous-time second-order differential equation characterizing the Nesterov acceleration. Continuous-time Accelerated Mirror Descent was studied in [KBB15]. An approach based on Bregman-Lagrangian functional for continuous-time momentum and other methods was developed in [WWJ16, WRJ16]. Further results on related topics, relevant to machine learning, were obtained in [ZMSJ18, SRBd17, FRV18]. An overview of results on gradient flow beyond the Euclidean space setting can be found in [AGS08, San17].

On a related note, several studies took advantage of the fact that the distribution of the Langevin dynamics is a gradient flow in the space of measures [CB18, Ber18, DMM19, Wib18], in order to establish error bounds for sampling algorithms. The relation with MMD was studied by [AKSG19].

[5]Here, $\nu^{\mathsf{LD}}_{t,\boldsymbol{x}}$ is the distribution at time t of the LD starting at $\boldsymbol{x}$, and the inequality is assumed to hold for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ and any $t \in \mathbb{R}$.

The results presented in the present work can be generalized in various directions. In particular, it would be interesting to relax the smoothness assumption, following an argument from, *e.g.*, [DMP18, CDJB19, MFWB19, SKR19], to develop a similar theory for the kinetic Langevin dynamics [EGZ19, CCBJ18, DRD20, MCJ$^+$19] or to consider the case of a Lévy process driven stochastic differential equation in the spirit of [SZTG20, LMW19].

4.5 Conclusion

We put forward a family of time-inhomogeneous diffusion processes that converge to a pre-specified target distribution and, therefore, can be used for approximate sampling. These processes are defined as penalized Langevin dynamics with a penalty that vanishes when time goes to infinity. The penalty allows to ensure strong convexity, which helps to handle situations where the original log-density is not strongly convex. We established a simple non-asymptotic error bound showing that the rate of convergence in the Wasserstein-2 distance is $o(1/\sqrt{t})$. We have also discussed analogous results for the penalized gradient flow. The important next step to investigate in future works is the analysis of discretized versions of the penalized Langevin dynamics.

Appendix to Chapter 4

4.A Proof of Theorem 14

Recall that for every $\gamma \in \mathbb{R}$, π_γ is the probability distribution with density proportional to $\exp(-f(\boldsymbol{\theta}) - \gamma\|\boldsymbol{\theta}\|_2^2/2)$. The triangle inequality for the Wasserstein distance yields

$$W_2(\nu_t^{\mathsf{PLD}}, \pi) \leq W_2(\nu_t^{\mathsf{PLD}}, \pi_{\alpha(t)}) + W_2(\pi_{\alpha(t)}, \pi), \tag{4.6}$$

for every $t > 0$. We will bound these two terms separately, but let us start by stating two technical lemmas. The first one is a consequence of the well-known transportation cost inequality (see [GL10, Corollary 7.2]), whereas the second one establishes the smoothness and the monotony with respect to γ of the second-order moment of π_γ. The proofs of these lemmas are postponed to Section 4.C.

Lemma 15. *Let π be a probability density function such that the potential $f = -\log(\pi)$ satisfies the $(m, +\infty)$-SCGL condition. Let $\widetilde{\gamma} \geq \gamma$ be real numbers, such that $m + \gamma \geq 0$. Then*

$$W_2(\pi_{\widetilde{\gamma}}, \pi_\gamma) \leq \frac{11(\widetilde{\gamma} - \gamma)}{\sqrt{m + \widetilde{\gamma}}} \mu_2(\pi_\gamma).$$

Lemma 16. *Suppose that π has a finite fourth-order moment. Then $\gamma \mapsto \mu_2(\pi_\gamma)$ is continuously differentiable and non-increasing, when $\gamma \in [0, +\infty)$.*

If we apply Lemma 15 with $\gamma = 0$ and $\widetilde{\gamma} = \alpha(t)$, then we obtain

$$W_2(\pi_{\alpha(t)}, \pi) \leq \frac{11\alpha(t)}{\sqrt{m + \alpha(t)}} \mu_2(\pi). \tag{4.7}$$

This provides the desired upper bound on the second term of the right hand side of (4.6). To bound the first term, $W_2(\nu_t^{\mathsf{PLD}}, \pi_{\alpha(t)})$, we aim at obtaining a Gronwall-type inequality for the function

$$\phi(t) := W_2(\nu_t^{\mathsf{PLD}}, \pi_{\alpha(t)}).$$

To this end, we consider an auxiliary stochastic process $\{\widetilde{\boldsymbol{L}}_u : u \geq t\}$, defined as a solution of the following stochastic differential equation

$$d\widetilde{\boldsymbol{L}}_u = -\big(\nabla f(\widetilde{\boldsymbol{L}}_u) + \alpha(t)\widetilde{\boldsymbol{L}}_u\big)du + \sqrt{2}d\boldsymbol{W}_u,$$

with the starting point $\widetilde{\boldsymbol{L}}_t = \boldsymbol{L}_t^{\mathsf{PLD}}$. This is in fact the Langevin diffusion corresponding to the potential $f(\cdot) + \alpha(t)\|\cdot\|_2^2/2$. Therefore, $\pi_{\alpha(t)}$ is the invariant distribution of $\widetilde{\boldsymbol{L}}$, and it is $(m+\alpha(t))$-strongly log-concave. Let $\boldsymbol{Q}_{t,\delta}$ be the distribution of the random vector $\widetilde{\boldsymbol{L}}_{t+\delta}$. The triangle inequality yields

$$\phi(t+\delta) \leq W_2\left(\nu_{t+\delta}^{\mathsf{PLD}}, \boldsymbol{Q}_{t,\delta}\right) + W_2\left(\boldsymbol{Q}_{t,\delta}, \pi_{\alpha(t)}\right) + W_2\left(\pi_{\alpha(t)}, \pi_{\alpha(t+\delta)}\right).$$

Recalling the definition of $\pi_{\alpha(t)}$ and $\boldsymbol{Q}_{t,\delta}$, we therefore find ourselves in the case of classical Langevin diffusion. Hence, one can apply (4.1) to get the bound

$$W_2(\boldsymbol{Q}_{t,\delta}, \pi_{\alpha(t)}) \leq \exp\big(-\delta(m+\alpha(t))\big)W_2(\nu_t^{\mathsf{PLD}}, \pi_{\alpha(t)}) = \exp\left(-\delta(m+\alpha(t))\right)\phi(t).$$

Applying Lemma 15 to $\pi_{\alpha(t)}$ and $\pi_{\alpha(t+\delta)}$, we get

$$W_2\left(\pi_{\alpha(t)}, \pi_{\alpha(t+\delta)}\right) \leq \frac{11(\alpha(t) - \alpha(t+\delta))}{\sqrt{m+\alpha(t)}}\mu_2(\pi_{\alpha(t+\delta)}).$$

Thus we obtain a bound for $\phi(t+\delta)$, that depends linearly on $\phi(t)$:

$$\phi(t+\delta) \leq W_2\left(\nu_{t+\delta}^{\mathsf{PLD}}, \boldsymbol{Q}_{t,\delta}\right) + e^{-\delta(m+\alpha(t))}\phi(t) + \frac{11(\alpha(t)-\alpha(t+\delta))}{\sqrt{m+\alpha(t)}}\mu_2(\pi_{\alpha(t+\delta)}). \tag{4.8}$$

Let us subtract $\phi(t)$ from both sides of (4.8) and divide by δ:

$$\begin{aligned}\frac{\phi(t+\delta)-\phi(t)}{\delta} \leq \frac{1}{\delta}\cdot W_2\left(\nu_{t+\delta}^{\mathsf{PLD}}, \boldsymbol{Q}_{t,\delta}\right) &+ \frac{\exp\left(-\delta(m+\alpha(t))\right)-1}{\delta}\cdot\phi(t)\\ &+ \frac{11(\alpha(t)-\alpha(t+\delta))}{\delta\sqrt{m+\alpha(t)}}\mu_2(\pi_{\alpha(t+\delta)}).\end{aligned} \tag{4.9}$$

The next lemma provides an upper bound on $W_2\left(\nu_{t+\delta}^{\mathsf{PLD}}, \boldsymbol{Q}_{t,\delta}\right)$ showing that it is $o(\delta)$, when $\delta \to 0$.

Lemma 17. *For every $t, \delta > 0$, and for every integrable function $\alpha : [t, t+\delta] \to [0,\infty)$,*

$$W_2\left(\nu_{t+\delta}^{\mathsf{PLD}}, \boldsymbol{Q}_{t,\delta}\right) \leq \left(\phi(t) + \mu_2^{1/2}(\pi)\right)\exp\left\{M\delta + \int_0^\delta \alpha(t+u)\,du\right\}\int_0^\delta \big|\alpha(t+s)-\alpha(t)\big|\,ds.$$

When δ tends to 0, according to Lemma 17, the first term of the right-hand side of (4.9) vanishes. Thus, after passing to the limit, we are left with the following Gronwall-type inequality:

$$\phi'(t) \le -(m+\alpha(t))\phi(t) - \frac{11\alpha'(t)}{\sqrt{m+\alpha(t)}} \cdot \mu_2(\pi_{\alpha(t)}). \tag{4.10}$$

Here we tacitly used the fact that $\mu_2(\pi_{\alpha(t+\delta)}) \to \mu_2(\pi_{\alpha(t)})$, whenever $\delta \to 0$, which is due to the continuity of $\alpha(t)$ and Lemma 16. Recalling that the function $\beta(t)$ is given by $\beta(t) = \int_0^t (m+\alpha(s))\, ds$, one can rewrite (4.10) as

$$\left(\phi(t)e^{\beta(t)}\right)' \le -\frac{11\alpha'(t)e^{\beta(t)}}{\sqrt{m+\alpha(t)}}\mu_2(\pi_{\alpha(t)}) \le -\frac{11\alpha'(t)e^{\beta(t)}}{\sqrt{m+\alpha(t)}}\mu_2(\pi).$$

Therefore we infer the following bound on $\phi(t)$:

$$\phi(t) \le \phi(0)e^{-\beta(t)} - 11\mu_2(\pi)\int_0^t \frac{\alpha'(s)}{\sqrt{m+\alpha(s)}} e^{\beta(s)-\beta(t)} ds.$$

Combining this bound with (4.6) and (4.7), we obtain the inequality

$$W_2(\nu_t^{\mathsf{PLD}}, \pi) \le W_2(\nu_0, \pi_{\alpha(0)})e^{-\beta(t)} - 11\mu_2(\pi)\int_0^t \frac{\alpha'(s)e^{\beta(s)-\beta(t)}}{\sqrt{m+\alpha(s)}} ds + \frac{11\alpha(t)\mu_2(\pi)}{\sqrt{m+\alpha(t)}}.$$

Lemma 16 yields $W_2(\nu_0, \pi_{\alpha(0)}) = \sqrt{\mu_2(\pi_{\alpha(0)})} \le \sqrt{\mu_2(\pi)}$. This completes the proof of Theorem 14, since the derivative of α is negative.

4.B Proof of Theorem 15

The triangle inequality for the Wasserstein distance yields

$$W_2(\nu_t^{\mathsf{PKLD}}, \mathsf{P}) \le W_2(\nu_t^{\mathsf{PKLD}}, \mathsf{P}_{\alpha(t)}) + W_2(\mathsf{P}_{\alpha(t)}, \mathsf{P}), \tag{4.11}$$

for every $t > 0$. Lemma 15 yields the following bound on the second term:

$$W_2(\mathsf{P}_{\alpha(t)}, \mathsf{P}) = W_2(\pi_{\alpha(t)}, \pi) \le \frac{11\alpha(t)}{\sqrt{m+\alpha(t)}}\mu_2(\pi). \tag{4.12}$$

To bound the term $W_2(\nu_t^{\mathsf{PKLD}}, \mathsf{P}_{\alpha(t)})$ we will consider it as a function of time:

$$\phi(t) := W_2(\nu_t^{\mathsf{PKLD}}, \mathsf{P}_{\alpha(t)}).$$

The method to bound $\phi(t)$ is similar to the method introduced for the proof of Theorem 14. We first bound $\phi(t+\delta)$ and then obtain a Gronwall-type bound on the derivative of ϕ. In the end we apply Gronwall inequality and deduce the proof of the theorem.

For every t we fix the drift penalty in PKLD and introduce the following the SDE:

$$\begin{aligned} d\boldsymbol{L}_u &= \boldsymbol{V}_u du; \\ d\boldsymbol{V}_u &= -\big(\eta \boldsymbol{V}_u + \nabla f(\boldsymbol{L}_u) + \alpha(t)\boldsymbol{L}_u\big)du + \sqrt{2\eta}d\boldsymbol{W}_u, \end{aligned} \tag{†-KLD}$$

where $u \geq t$ and $\boldsymbol{W}$ is the same Brownian motion as in PKLD. Let $(\widetilde{\boldsymbol{L}}_u, \widetilde{\boldsymbol{V}}_u)_{u\geq 0}$ and $(\boldsymbol{L}'_u, \boldsymbol{V}'_u)_{u\geq 0}$ be the solutions of (†-KLD) with following initial conditions:

$$\big(\widetilde{\boldsymbol{L}}_0, \widetilde{\boldsymbol{V}}_0\big) = \big(\boldsymbol{L}_t^{\mathsf{PKLD}}, \boldsymbol{V}_t^{\mathsf{PKLD}}\big) \qquad \text{and} \qquad (\boldsymbol{L}'_0, \boldsymbol{V}'_0) \sim \mathsf{P}_{\alpha(t)}.$$

In addition we assume that the Wasserstein distance of $\mathsf{P}_{\alpha(t)}$ and $\boldsymbol{Q}_{t,\delta}$ is attained for the vectors $\big(\widetilde{\boldsymbol{L}}_0, \widetilde{\boldsymbol{V}}_0\big)$ and $(\boldsymbol{L}'_0, \boldsymbol{V}'_0)$:

$$\left\| \begin{bmatrix} \widetilde{\boldsymbol{L}}_0 - \boldsymbol{L}'_0 \\ \widetilde{\boldsymbol{V}}_0 - \boldsymbol{L}'_0 \end{bmatrix} \right\|_{\mathbb{L}_2} = W_2(\mathsf{P}_{\alpha(t)}, \boldsymbol{Q}_{t,\delta}).$$

Here we denote by $\boldsymbol{Q}_{t,\delta}$, the probability distribution $\mathcal{L}(\widetilde{\boldsymbol{L}}_\delta, \widetilde{\boldsymbol{V}}_\delta)$. The triangle inequality for Wasserstein distance yields

$$\phi(t+\delta) \leq W_2\left(\nu_{t+\delta}^{\mathsf{PKLD}}, \boldsymbol{Q}_{t,\delta}\right) + W_2\left(\boldsymbol{Q}_{t,\delta}, \mathsf{P}_{\alpha(t)}\right) + W_2\left(\mathsf{P}_{\alpha(t)}, \mathsf{P}_{\alpha(t+\delta)}\right). \tag{4.13}$$

Equation †-KLD is in fact the kinetic Langevin diffusion, which has $\mathsf{P}_{\alpha(t)}$ as its invariant distribution. Therefore $(\boldsymbol{L}'_u, \boldsymbol{V}'_u) \sim \mathsf{P}_{\alpha(t)}$ for all u, which implies the following inequality:

$$W_2\left(\boldsymbol{Q}_{t,\delta}, \mathsf{P}_{\alpha(t)}\right) \leq \left\| \begin{bmatrix} \widetilde{\boldsymbol{L}}_{t+\delta} - \boldsymbol{L}'_{t+\delta} \\ \widetilde{\boldsymbol{V}}_{t+\delta} - \boldsymbol{V}'_{t+\delta} \end{bmatrix} \right\|_{\mathbb{L}_2}.$$

To bound the right-hand side we introduce the following lemma.

Lemma 18. *Let $(\boldsymbol{L}^1, \boldsymbol{V}^1)$ and $(\boldsymbol{L}^2, \boldsymbol{V}^2)$ be two solutions of (†-KLD). Then for every $v \in (0, \eta/2)$*

$$\left\| \begin{bmatrix} \boldsymbol{L}_u^1 - \boldsymbol{L}_u^2 \\ \boldsymbol{V}_u^1 - \boldsymbol{V}_u^2 \end{bmatrix} \right\|_{\mathbb{L}_2} \leq 2\eta \exp\left\{(-\alpha(t)) \vee \big(M + \alpha(t) - \eta^2\big) \cdot \frac{u}{\eta}\right\} \left\| \begin{bmatrix} \boldsymbol{L}_0^1 - \boldsymbol{L}_0^2 \\ \boldsymbol{V}_0^1 - \boldsymbol{V}_0^2 \end{bmatrix} \right\|_{\mathbb{L}_2}.$$

The proof of the lemma can be found in the Section 4.F.1. Thus

$$W_2(\boldsymbol{Q}_{t,\delta}, \mathsf{P}_{\alpha(t)}) \leq 2\eta \exp\left\{(-\alpha(t)) \vee \big(M + \alpha(t) - \eta^2\big)\frac{\delta}{\eta}\right\} W_2(\boldsymbol{Q}_{t,0}, \mathsf{P}_{\alpha(t)}).$$

Since $\widetilde{\boldsymbol{L}}_0 = \boldsymbol{L}_t^{\mathsf{PKLD}}$, we get $W_2(\boldsymbol{Q}_{t,0}, \pi_{\alpha(t)}) = \phi(t)$. Due to the assumption of the theorem

$\eta > \sqrt{M}$, which means that the argument in the exponential function is negative for all t. In particular, since $\alpha(t) \to 0$ for large values of t, we get that $2\alpha(t) < \eta^2 - M$. Summing up we get the following inequality:

$$\limsup_{\delta \to 0} \frac{1}{\delta} \big(W_2(\boldsymbol{Q}^{\boldsymbol{L}}_{t,\delta}, \pi_{\alpha(t)}) - \phi(t) \big) \le -2\eta\alpha(t)\phi(t). \tag{4.14}$$

The next lemma provides an upper bound on $W_2\left(\nu^{\mathsf{PKLD}}_{t+\delta}, \boldsymbol{Q}_{t,\delta}\right)$ showing that it is $o(\delta)$, when $\delta \to 0$.

Lemma 19. *For every $t, \delta > 0$, and for every integrable function $\alpha : [t, t+\delta] \to [0, \infty)$,*

$$W_2\left(\nu^{\mathsf{PKLD}}_{t+\delta}, \boldsymbol{Q}_{t,\delta}\right) \le \left(c\phi(t) + \mu_2^{1/2}(\mathsf{P}_{\alpha(t)})\right) \exp\left\{ \int_0^\delta G(t+u)\, du \right\} \int_0^\delta \big|\alpha(t+s) - \alpha(t)\big|\, ds,$$

where $G(t+u) := \max(\eta + 1, M + \alpha(t+s))$ and c is a constant that does not depend on t and δ.

Finally, the last term of (4.13) is bounded using Lemma 15:

$$W_2(\mathsf{P}_{\alpha(t+\delta)}, \mathsf{P}_{\alpha(t)}) = W_2(\pi_{\alpha(t+\delta)}, \pi_{\alpha(t)}) \le \frac{11(\alpha(t) - \alpha(t+\delta))}{\sqrt{m + \alpha(t)}} \mu_2(\pi_{\alpha(t+\delta)}).$$

Thus

$$\limsup_{\delta \to 0} \frac{1}{\delta} W_2(\mathsf{P}_{\alpha(t+\delta)}, \mathsf{P}_{\alpha(t)}) \le \frac{11|\alpha'(t)|}{\sqrt{m + \alpha(t)}} \mu_2(\pi_{\alpha(t)}). \tag{4.15}$$

Combining (4.13), (4.14), Lemma 19 and (4.15) we get the following equality:

$$\begin{aligned} \phi'(t) &= \lim_{\delta \to 0} \frac{1}{\delta} \big(\phi(t+\delta) - \phi(t) \big) \\ &\le -2\eta\alpha(t)\phi(t) - \frac{11\alpha'(t)}{\sqrt{m + \alpha(t)}} \cdot \mu_2(\pi_{\alpha(t)}). \end{aligned}$$

Recalling that the function $\beta(t)$ is the antiderivative of $2\eta\alpha(t)$,

$$\beta(t) = \int_0^t (\sqrt{M + \alpha(u)} - \sqrt{M - m}) \left(\frac{2M - m + \alpha(u)}{M - m} \right)^{1/2} du,$$

we deduce

$$\big(\phi(t) e^{\beta(t)}\big)' \le -\frac{11\alpha'(t) e^{\beta(t)}}{\sqrt{m + \alpha(t)}} \mu_2(\pi_{\alpha(t)}) \le \frac{11|\alpha'(t)| e^{\beta(t)}}{\sqrt{m + \alpha(t)}} \mu_2(\pi).$$

The latter is true due to Lemma 16. Therefore we infer the following bound on $\phi(t)$:

$$\phi(t) \leq \phi(0)e^{-\beta(t)} + 11\mu_2(\pi) \int_0^t \frac{|\alpha'(s)|}{\sqrt{m+\alpha(s)}} e^{\beta(s)-\beta(t)} ds.$$

Combining this bound with (4.11) and (4.12), we obtain the inequality

$$W_2(\nu_t^{\mathsf{PKLD}}, \mathsf{P}) \leq W_2(\nu_0^{\mathsf{PKLD}}, \mathsf{P}_{\alpha(0)})e^{-\beta(t)} + 11\mu_2(\mathsf{P}) \int_0^t \frac{|\alpha'(s)|e^{\beta(s)-\beta(t)}}{\sqrt{m+\alpha(s)}} ds + \frac{11\alpha(t)\mu_2(\mathsf{P})}{\sqrt{m+\alpha(t)}}.$$

This completes the proof.

4.C Proofs of the lemmas used in Theorem 14

4.C.1 Proof of Lemma 15

We denote by $D_{\mathrm{KL}}(\pi_\gamma || \pi_{\widetilde{\gamma}})$ the Kullback-Leibler divergence between the distributions π_γ and $\pi_{\widetilde{\gamma}}$. Since $\pi_{\widetilde{\gamma}}$ is $(m+\widetilde{\gamma})$-strongly log-concave, the transportation cost inequality [GL10, Corollary 7.2] yields

$$W_2^2(\pi_{\widetilde{\gamma}}, \pi_\gamma) \leq \frac{2}{m+\widetilde{\gamma}} D_{\mathrm{KL}}(\pi_\gamma || \pi_{\widetilde{\gamma}}). \tag{4.16}$$

Let us denote by c_γ the logarithm of the normalizing constant for π_γ so that $\pi_\gamma(\boldsymbol{\theta}) = \exp(-f(\boldsymbol{\theta}) - (1/2)\gamma\|\boldsymbol{\theta}\|_2^2 + c_\gamma)$. Similarly, we denote by $c_{\widetilde{\gamma}}$ the logarithm of the normalizing constant of $\pi_{\widetilde{\gamma}}$. This readily yields

$$\begin{aligned}
D_{\mathrm{KL}}(\pi_\gamma || \pi_{\widetilde{\gamma}}) &= \int_{\mathbb{R}^p} \pi_\gamma(\boldsymbol{\theta}) \log\left(\frac{\pi_\gamma(\boldsymbol{\theta})}{\pi_{\widetilde{\gamma}}(\boldsymbol{\theta})}\right) d\boldsymbol{\theta} \\
&= \int_{\mathbb{R}^p} \pi_\gamma(\boldsymbol{\theta}) \left(1/2(\widetilde{\gamma} - \gamma)\|\boldsymbol{\theta}\|_2^2 + c_\gamma - c_{\widetilde{\gamma}}\right) d\boldsymbol{\theta} \\
&= 1/2(\widetilde{\gamma} - \gamma)\mu_2(\pi_\gamma) + c_\gamma - c_{\widetilde{\gamma}}.
\end{aligned}$$

Using the inequality $e^{-x} \leq 1 - x + (1/2)x^2$ for all $x > 0$ implies the following upper bound on $c_\gamma - c_{\widetilde{\gamma}}$:

$$\begin{aligned}
c_\gamma - c_{\widetilde{\gamma}} &= \log\left(\int_{\mathbb{R}^p} \pi_\gamma(\boldsymbol{\theta}) \exp\left(1/2(\gamma - \widetilde{\gamma})\|\boldsymbol{\theta}\|_2^2\right) d\boldsymbol{\theta}\right) \\
&\leq \log\left(1 + 1/2(\gamma - \widetilde{\gamma})\mu_2(\pi_\gamma) + 1/8(\gamma - \widetilde{\gamma})^2\mu_4(\pi_\gamma)\right).
\end{aligned}$$

Since $\log(1+x) \le x$ for $x > -1$ we get

$$D_{\mathrm{KL}}(\pi_\gamma || \pi_{\widetilde{\gamma}}) \le {}^1\!/_8(\gamma - \widetilde{\gamma})^2 \mu_4(\pi_\gamma).$$

Since $m + \gamma \ge 0$, the distribution π_γ is log-concave. Thus, in view of [DRDK19, Remark 3], we have the inequality $\mu_4(\pi_\gamma) \le 442\mu_2^2(\pi_\gamma)$. Finally, combining these bounds with (4.16), we get

$$W_2(\pi_{\widetilde{\gamma}}, \pi_\gamma) \le \sqrt{\frac{2}{m+\widetilde{\gamma}}} \times \frac{(\widetilde{\gamma} - \gamma)\mu_4^{1/2}(\pi_\gamma)}{\sqrt{8}} \le \frac{11\mu_2(\pi_\gamma)}{\sqrt{m+\widetilde{\gamma}}}(\widetilde{\gamma} - \gamma).$$

This completes the proof of the lemma.

4.C.2 Proof of Lemma 16

For $k \in \mathbb{N} \cup \{0\}$, define

$$h_k(\gamma) = \int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^k \exp\left(-f(\boldsymbol{\theta}) - \gamma\|\boldsymbol{\theta}\|_2^2/2\right) d\boldsymbol{\theta}.$$

If $\pi \in \mathcal{P}_k(\mathbb{R}^p)$ then the function h_k is continuous on $[0; +\infty)$. Indeed, if the sequence $\{\gamma_n\}_n$ converges γ_0, when $n \to +\infty$, then the function $\|\boldsymbol{\theta}\|_2^k \exp\left(-f(\boldsymbol{\theta}) - ({}^1\!/_2)\gamma_n\|\boldsymbol{\theta}\|_2^2\right)$ is upper-bounded by $\|\boldsymbol{\theta}\|_2^k \exp\left(-f(\boldsymbol{\theta})\right)$. Thus in view of the dominated convergence theorem, we can interchange the limit and the integral. Since, by definition,

$$\mu_k(\pi_\gamma) = \frac{h_k(\gamma)}{h_0(\gamma)},$$

we get the continuity of $\mu_2(\pi_\gamma)$ and $\mu_4(\pi_\gamma)$. Let us now prove that $h_k(t)$ is continuously differentiable, when $\pi \in \mathcal{P}_{k+2}(\mathbb{R}^p)$. The integrand function in the definition of h_k is a continuously differentiable function with respect to t. In addition, its derivative is continuous and is as well integrable on $\mathbb{R}^p$, as we supposed that π has the $(k+2)$-th moment. Therefore, Leibniz integral rule yields the following

$$h_k'(\gamma) = -\frac{1}{2}\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^{k+2} \exp\left(-f(\boldsymbol{\theta}) - \gamma\|\boldsymbol{\theta}\|_2^2/2\right) d\boldsymbol{\theta} = -\frac{1}{2}h_{k+2}(t).$$

The latter yields the smoothness of h_k. Finally, in order to prove the monotony of $\mu_2(\pi_\gamma)$, we will simply calculate its derivative

$$\begin{aligned}(\mu_2(\pi_\gamma))' &= -\frac{1}{2h_0(\gamma)}h_4(\gamma) - \frac{h_0'(\gamma)}{h_0(\gamma)^2}h_2(\gamma) \\ &= \frac{1}{2}\left(\mu_2^2(\pi_\gamma) - \mu_4(\pi_\gamma)\right).\end{aligned}$$

Since the latter is always negative, this completes the proof of the lemma.

4.C.3 Proof of Lemma 17

From the definition of Wasserstein distance, we have

$$W_2\left(\nu^{\mathsf{PLD}}_{t+\delta}, \boldsymbol{Q}_{t,\delta}\right) \leq \|\widetilde{\boldsymbol{L}}_{t+\delta} - \boldsymbol{L}^{\mathsf{PLD}}_{t+\delta}\|_{\mathbb{L}_2}.$$

In view of the definition of the process $\widetilde{\boldsymbol{L}}$, we can write

$$\widetilde{\boldsymbol{L}}_{t+\delta} - \boldsymbol{L}^{\mathsf{PLD}}_{t+\delta} = \int_t^{t+\delta} \left(\nabla f(\boldsymbol{L}^{\mathsf{PLD}}_s) - \nabla f(\widetilde{\boldsymbol{L}}_s) + \alpha(s)\boldsymbol{L}^{\mathsf{PLD}}_s - \alpha(t)\widetilde{\boldsymbol{L}}_s\right) ds.$$

Therefore we have

$$\|\widetilde{\boldsymbol{L}}_{t+\delta} - \boldsymbol{L}^{\mathsf{PLD}}_{t+\delta}\|_{\mathbb{L}_2} \leq \Big\| \underbrace{\int_t^{t+\delta} \left(\nabla f(\boldsymbol{L}^{\mathsf{PLD}}_s) - \nabla f(\widetilde{\boldsymbol{L}}_s)\right) ds}_{:=T_1} \Big\|_{\mathbb{L}_2} + \Big\| \underbrace{\int_t^{t+\delta} \left(\alpha(s)\boldsymbol{L}^{\mathsf{PLD}}_s - \alpha(t)\widetilde{\boldsymbol{L}}_s\right) ds}_{:=T_2} \Big\|_{\mathbb{L}_2}.$$

Now let us analyze these two terms separately. We start with T_1:

$$\begin{aligned}
\|T_1\|_{\mathbb{L}_2} &= \left\| \int_t^{t+\delta} \left(\nabla f(\boldsymbol{L}^{\mathsf{PLD}}_s) - \nabla f(\widetilde{\boldsymbol{L}}_s)\right) ds \right\|_{\mathbb{L}_2} \\
&\leq \int_t^{t+\delta} \left\|\nabla f(\boldsymbol{L}^{\mathsf{PLD}}_s) - \nabla f(\widetilde{\boldsymbol{L}}_s)\right\|_{\mathbb{L}_2} ds \\
&\leq M \int_t^{t+\delta} \|\boldsymbol{L}^{\mathsf{PLD}}_s - \widetilde{\boldsymbol{L}}_s\|_{\mathbb{L}_2} ds.
\end{aligned}$$

These are due to the Minkowskii inequality and the Lipschitz continuity of the gradient. In order to bound the second term T_2, we will add and subtract the term $\alpha(t+s)\widetilde{\boldsymbol{L}}_{t+s}$. Similar to the case above, we get the following upper bound:

$$\begin{aligned}
\|T_2\|_{\mathbb{L}_2} &\leq \int_t^{t+\delta} \alpha(s) \left\|\boldsymbol{L}^{\mathsf{PLD}}_s - \widetilde{\boldsymbol{L}}_s\right\|_{\mathbb{L}_2} ds + \int_t^{t+\delta} |\alpha(s) - \alpha(t)| \left\|\widetilde{\boldsymbol{L}}_s\right\|_{\mathbb{L}_2} ds \\
&= \int_0^{\delta} \alpha(t+s) \left\|\boldsymbol{L}^{\mathsf{PLD}}_{t+s} - \widetilde{\boldsymbol{L}}_{t+s}\right\|_{\mathbb{L}_2} ds + \int_0^{\delta} |\alpha(t+s) - \alpha(t)| \left\|\widetilde{\boldsymbol{L}}_{t+s}\right\|_{\mathbb{L}_2} ds.
\end{aligned}$$

Recall that $\widetilde{\boldsymbol{L}}_{t+s}$ is the solution of Langevin SDE with an $(m + \alpha(t))$-strongly convex potential function, and $Q_{t,s}$ is its distribution on $\mathbb{R}^p$. Thus, the triangle inequality yields

$$\left\|\widetilde{\boldsymbol{L}}_{t+s}\right\|_{\mathbb{L}_2} = W_2(\boldsymbol{Q}_{t,s}, \delta_0) \leq W_2(\boldsymbol{Q}_{t,s}, \pi_{\alpha(t)}) + W_2(\pi_{\alpha(t)}, \delta_0)$$

and applying the strong-convexity of $f_{\alpha(t)}$ we have

$$\big\|\widetilde{\boldsymbol{L}}_{t+s}\big\|_{\mathbb{L}_2} \leq W_2(\nu_t^{\mathsf{PLD}}, \pi_{\alpha(t)})\exp(-ms - \alpha(t)s) + \sqrt{\mu_2(\pi_{\alpha(t)})} \leq V_t$$

Summing up, we have

$$\big\|\boldsymbol{L}_{t+\delta}^{\mathsf{PLD}} - \widetilde{\boldsymbol{L}}_{t+\delta}\big\|_{\mathbb{L}_2} \leq \int_0^\delta \big(M + \alpha(t+s)\big)\|\boldsymbol{L}_{t+s}^{\mathsf{PLD}} - \widetilde{\boldsymbol{L}}_{t+s}\|_{\mathbb{L}_2} ds + \widetilde{\alpha}_t(\delta)\, V_t,$$

where $\widetilde{\alpha}_t(\delta)$ is an auxiliary function defined as

$$\widetilde{\alpha}_t(\delta) := \int_0^\delta |\alpha(t+s) - \alpha(t)|\, ds.$$

Now let us define $\Phi(s) = \|\boldsymbol{L}_{t+s}^{\mathsf{PLD}} - \widetilde{\boldsymbol{L}}_{t+s}\|_{\mathbb{L}_2}$. The last inequality can be rewritten as

$$\Phi(\delta) \leq \int_0^\delta \big(M + \alpha(t+s)\big)\Phi(s)\, ds + \widetilde{\alpha}_t(\delta)\, V_t.$$

The (integral form of the) Gronwall inequality, lemma 21, implies that

$$\begin{aligned}
\Phi(\delta) &\leq V_t \int_0^\delta \widetilde{\alpha}_t(s)\big(M + \alpha(t+s)\big)e^{\int_s^\delta (M+\alpha(t+u))\, du}\, ds + \widetilde{\alpha}_t(\delta)V_t \\
&= V_t \int_0^\delta \widetilde{\alpha}_t'(s)\, e^{\int_s^\delta (M+\alpha(t+u))\, du}\, ds \\
&\leq V_t\, \widetilde{\alpha}_t(\delta)\, \exp\left\{M\delta + \int_0^\delta \alpha(t+u)\, du\right\}.
\end{aligned}$$

This completes the proof.

4.C.4 Different forms of the Gronwall inequality

In this section, we provide two forms of the Gronwall inequality that are used in the present work. For the sake of the self-containedness, the proofs of these inequalities are also provided.

Lemma 20 (Differential form)**.** *Let $A : [a, b] \to \mathbb{R}$ and $B : [a, b] \to \mathbb{R}$ be two functions. If the function $\Phi : [a, b] \to \mathbb{R}$ satisfies the recursive differential inequality*

$$\Phi'(x) \leq A(x)\Phi(x) + B(x), \qquad \forall x \in [a, b], \tag{4.17}$$

then it also satisfies the inequality

$$\Phi(x) \leq \Phi(a) \exp\left\{ \int_a^x A(z)\, dz \right\} + \int_a^x B(s) \exp\left\{ \int_s^x A(z)\, dz \right\} ds, \qquad \forall x \in [a, b].$$

Proof. To ease notation, we set $E(x) = \exp\{-\int_a^x A(z)\, dz\}$. By multiplying both sides of (4.17) by $E(x)$, we get

$$\Big(\Phi(x)E(x)\Big)' \leq B(x)E(x), \qquad \forall x \in [a, b].$$

Integrating this inequality, we arrive at

$$\Phi(x)E(x) \leq \Phi(a)E(a) + \int_a^x B(s)E(s)\, ds.$$

Dividing both sides of this inequality by $E(x) > 0$ and taking into account that $E(a) = 1$, we get the claim of the lemma. □

Lemma 21 (Integral form). *Let $A : [a, b] \to [0, +\infty)$ and $B : [a, b] \to \mathbb{R}$ be two functions. If the function $\Phi : [a, b] \to \mathbb{R}$ satisfies the recursive integral inequality*

$$\Phi(x) \leq \int_a^x A(s)\Phi(s)\, ds + B(x), \qquad \forall x \in [a, b],$$

then it also satisfies the inequality

$$\Phi(x) \leq \int_a^x A(s)B(s) \exp\left\{ \int_s^x A(z)\, dz \right\} ds + B(x), \qquad \forall x \in [a, b]. \tag{4.18}$$

Proof. We set

$$\Psi(x) = \exp\left\{ -\int_a^x A(z)\, dz \right\} \int_a^x A(s)\Phi(s)\, ds.$$

We have

$$\begin{aligned}
\Psi'(x) &= -A(x)\Psi(x) + \exp\left\{ -\int_a^x A(z)\, dz \right\} A(x)\Phi(x) \\
&\leq -A(x)\Psi(x) + \exp\left\{ -\int_a^x A(z)\, dz \right\} A(x)\Big(\int_a^x A(s)\Phi(s)\, ds + B(x) \Big) \\
&= -A(x)\Psi(x) + A(x)\Psi(x) + A(x)B(x) \exp\left\{ -\int_a^x A(z)\, dz \right\}.
\end{aligned}$$

Therefore,

$$\Psi(x) \leq \Psi(a) + \int_a^x A(s)B(s) \exp\left\{ -\int_a^s A(z)\, dz \right\} ds.$$

Replacing Ψ by its expression and using the fact that $\Psi(a) = 0$, we get

$$\exp\left\{ -\int_a^x A(z)\,dz \right\} \int_a^x A(s)\Phi(s)\,ds \le \int_a^x A(s)B(s)\exp\left\{ -\int_a^s A(z)\,dz \right\} ds.$$

This implies that

$$\int_a^x A(s)\Phi(s)\,ds \le \int_a^x A(s)B(s)\exp\left\{ \int_s^x A(z)\,dz \right\} ds.$$

Combining this inequality with (4.18), we get the claim of the lemma. □

4.D Proof of Proposition 17

For the penalty factor $\alpha(t) = 1/(A+2t)$, we get $\beta(t) = \int_0^t \alpha(s)\,ds = (1/2)\log\big(1 + (2/A)t\big)$. This implies that

$$\sqrt{\mu_2(\pi)}\,e^{-\beta(t)} + 11\mu_2(\pi)\sqrt{\alpha(t)} = \frac{\sqrt{A\mu_2(\pi)} + 11\mu_2(\pi)}{\sqrt{A+2t}}.$$

Finally, the middle term in the right hand side of (4.2) takes the form

$$\begin{aligned} 11\mu_2(\pi)\int_0^t \frac{|\alpha'(s)|}{\sqrt{\alpha(s)}} e^{\beta(s)-\beta(t)} ds &= \frac{11\mu_2(\pi)}{\sqrt{A+2t}} \int_0^t \frac{2}{A+2s} ds \\ &= \frac{11\mu_2(\pi)}{\sqrt{A+2t}} \log\left(1 + (2/A)t\right). \end{aligned}$$

Combining these relations, we get the claim of the proposition.

4.E (Weakly) convex potentials: what is known and what we can hope for

Many recent papers investigated the case of strongly convex potential; this case is now rather well understood. Let us briefly summarize here some facts and conjectures that can shed some light on the broader case of weakly convex potential. This might help to understand what can be expected to be proved in the framework studied in this work.

The ergodicity properties of the Langevin process are closely related to such notions of functional analysis as the spectral gap, the Poincaré and the log-Sobolev inequalities. Thus, the generator of a Markov semi-group associated with an m-strongly convex potential has a spectral-gap $\mathcal{C}_{\text{SG}}$ at least equal to m. This property was exploited by [Dal17b] to derive guarantees on the LMC algorithm. It is known that the spectral gap exists if and only if the invariant density satisfies the Poincaré inequality. Furthermore, the spectral gap is equal to the inverse of the Poincaré constant $\mathcal{C}_{\text{P}}$. Furthermore, distributions associated to m-strongly convex potentials satisfy the log-Sobolev inequality with the constant $\mathcal{C}_{\text{LS}} \leq 1/m$. This property was used by [**?**] to extend the guarantees to the Wasserstein-2 distance.

Note that the log-Sobolev inequality is stronger than the Poincaré inequality and $\mathcal{C}_{\text{P}} \leq \mathcal{C}_{\text{LS}}$. For m-strongly convex potentials, we have $\mathcal{C}_{\text{SG}}^{-1} = \mathcal{C}_{\text{P}} \leq \mathcal{C}_{\text{LS}} \leq 1/m$. Results in [Dal17b**?**] imply that in order to get a Wasserstein distance smaller than $\varepsilon\sqrt{p/m}$, it suffices to perform a number of LMC iterations proportional to $(M/m)^2\varepsilon^{-2}$, up to logarithmic factors. A formal proof of the fact that the same result holds for the densities satisfying the log-Sobolev inequality with constant $1/m$ (but which are not necessarily m-strongly log-concave) was given in [VW19].

On the other hand, it was established by [Bob99] that any log-concave distribution satisfies the Poincaré inequality. However, the Poincaré constant might depend on the dimension. In [KLS95], the authors conjectured that there is a universal constant $\mathcal{C}_{\text{KLS}} > 0$ such that for any log-concave distribution π on $\mathbb{R}^p$,

$$\mathcal{C}_{\text{P}} \leq \mathcal{C}_{\text{KLS}} \|\mathbf{E}_\pi[\boldsymbol{X}\boldsymbol{X}^\top]\|_{\text{op}} := \mathcal{C}_{\text{KLS}}\mu_{\text{op}}(\pi). \tag{KLS}$$

Despite important efforts made in recent years (see [AGB15, CG18]), this conjecture is still unproved. Finally, in the recent paper [CLGL+20], Corollary 4 establishes that $W_2(\mu_t^{\mathsf{LD}}, \pi) \leq \sqrt{2\mathcal{C}_{\text{P}}\chi^2(\nu_0\|\pi)}\, e^{-t/\mathcal{C}_{\text{P}}}$. While the exponential in t convergence to zero is a very appealing property of this result, it comes with two shortcomings. To the previously mentioned difficulty of assessing the Poincaré constant, one has to add the challenging problem of finding a meaningful upper bound on the χ^2-divergence between the initial distribution and the target.

What can we hope for in the light of the previous discussion? As shown in [Dal17b, Lemma 5], for f satisfying (m, M)-SCGL, choosing $\nu_0 = \mathcal{N}(\boldsymbol{x}_*, M^{-1}\mathbf{I}_p)$ yields $\chi^2(\nu_0\|\pi) \leq (M/m)^{p/2}$. In the case $m = 0$, it might be possible to replace m by $1/\mathcal{C}_{\text{P}}$ in this result. If in addition, we admit inequality KLS, then we get

$$\begin{aligned}
W_2^2(\mu_t^{\mathsf{LD}}, \pi) &\leq 2\mathcal{C}_{\text{P}}\,(M\mathcal{C}_{\text{P}})^{p/2} e^{-2t/\mathcal{C}_{\text{P}}} \\
&\leq 2\mathcal{C}_{\text{KLS}}\mu_{\text{op}}(\pi)\,\big(M\mathcal{C}_{\text{KLS}}\mu_{\text{op}}(\pi)\big)^{p/2} e^{-2t/\mathcal{C}_{\text{KLS}}\mu_{\text{op}}(\pi)}.
\end{aligned}$$

This is, probably, the best upper bound one could hope for in the general log-concave setting by Langevin diffusion based algorithms. We see that it has three drawbacks as compared to our result stated in Proposition 17. First, it requires the knowledge of a minimizer $\boldsymbol{x}^*$. Second, it involves the Lipschitz constant M of the gradient. Third, it is heavily based on $\mathcal{C}_{\text{KLS}}$, which might be very large.

4.F Proofs of the lemmas used in Theorem 15

4.F.1 Proof of Lemma 18

We define an auxiliary random vector (ψ_t, z_t) in $\mathbb{R}^{2p}$ in the following manner:

$$\begin{pmatrix} \psi_t \\ z_t \end{pmatrix} = \underbrace{\begin{pmatrix} \lambda_+ I_p & I_p \\ -\lambda_- I_p & -I_p \end{pmatrix}}_{:=A} \cdot \begin{pmatrix} \boldsymbol{L}_t^1 - \boldsymbol{L}_t^2 \\ \boldsymbol{V}_t^1 - \boldsymbol{V}_t^2 \end{pmatrix},$$

where λ_+ and λ_- are two positive numbers such that $\lambda_+ + \lambda_- = \eta$ and $\lambda_+ > \lambda_-$. Taylor's theorem in its integral form yields

$$\nabla f\left(\boldsymbol{L}_t^1\right) - \nabla f\left(\boldsymbol{L}_t^2\right) = \mathbf{H}_t\left(\boldsymbol{L}_t^1 - \boldsymbol{L}_t^2\right)$$

with $\mathbf{H}_t \triangleq \int_0^1 \nabla^2 f\left(\boldsymbol{L}_t^1 - x\left(\boldsymbol{L}_t^1 - \boldsymbol{L}_t^2\right)\right) dx$. Since $(\boldsymbol{L}^1, \boldsymbol{V}^1)$ and $\left(\boldsymbol{L}^2, \boldsymbol{V}^2\right)$ satisfy the (†-KLD), combining with the equality above we obtain

$$\begin{aligned}
\frac{d}{dt}\psi_t &= -\eta\left(\boldsymbol{V}_t^1 - \boldsymbol{V}_t^2\right) - \left(\nabla f\left(\boldsymbol{L}_t^1\right) - \nabla f\left(\boldsymbol{L}_t^2\right)\right) + \lambda_+\left(\boldsymbol{V}_t^1 - \boldsymbol{V}_t^2\right) \\
&= \frac{(\lambda_+ - \eta)\left(\lambda_-\psi_t + \lambda_+ z_t\right)}{\lambda_- - \lambda_+} - \frac{\mathbf{H}_t\left(\psi_t + z_t\right)}{\lambda_+ - \lambda_-} \\
&= \frac{\left(\lambda_-^2\mathbf{I} - \mathbf{H}_t\right)\psi_t + \left(\lambda_-\lambda_+\mathbf{I} - \mathbf{H}_t\right)z_t}{\lambda_+ - \lambda_-}
\end{aligned}$$

In the above inequalities, we have used that $\lambda_+ - \eta = -\lambda_-$. Similar computations yield

$$\begin{aligned}
\frac{d}{dt}z_t &= \eta\left(\boldsymbol{V}_t^1 - \boldsymbol{V}_t^2\right) + \left(\nabla f\left(\boldsymbol{L}_t^1\right) - \nabla f\left(\boldsymbol{L}_t^2\right)\right) - \lambda_-\left(\boldsymbol{V}_t^1 - \boldsymbol{V}_t^2\right) \\
&= \frac{(\eta - \lambda_-)\left(\lambda_-\psi_t + \lambda_+ z_t\right)}{\lambda_- - \lambda_+} + \frac{\mathbf{H}_t\left(\psi_t + z_t\right)}{\lambda_+ - \lambda_-} \\
&= \frac{\left(\mathbf{H}_t - \lambda_-\lambda_+\mathbf{I}\right)\psi_t + \left(\mathbf{H}_t - \lambda_+^2\mathbf{I}\right)z_t}{\lambda_+ - \lambda_-}.
\end{aligned}$$

From these equations, we deduce that

$$
\begin{aligned}
\frac{d}{dt}\left\|\begin{bmatrix}\psi_t\\ z_t\end{bmatrix}\right\|_2^2 &= 2\psi_t^\top\frac{d\psi_t}{dt}+2z_t^\top\frac{dz_t}{dt}\\
&=\frac{2}{\lambda_+-\lambda_-}\left\{\psi_t^\top\left(\lambda_-^2\mathbf{I}-\mathbf{H}_t\right)\psi_t+z_t^\top\left(\mathbf{H}_t-\lambda_+^2\mathbf{I}\right)z_t\right\}\\
&\le\frac{2}{\lambda_+-\lambda_-}\left\{\left(\lambda_-^2-m-\alpha(t)\right)\|\psi_t\|_2^2+\left(M+\alpha(t)-\lambda_+^2\right)\|z_t\|_2^2\right\}\\
&\le\frac{2\left\{\left(\lambda_-^2-m-\alpha(t)\right)\vee\left(M+\alpha(t)-\lambda_+^2\right)\right\}}{\lambda_+-\lambda_-}\left\|\begin{bmatrix}\psi_t\\ z_t\end{bmatrix}\right\|_2^2
\end{aligned}
$$

Thus applying Grownwall inequality, we have the following:

$$
\left\|\begin{bmatrix}\psi_t\\ z_t\end{bmatrix}\right\|_2\le\exp\left\{\frac{\left(\lambda_-^2-m-\alpha(t)\right)\vee\left(M+\alpha(t)-\lambda_+^2\right)}{\lambda_+-\lambda_-}t\right\}\left\|\begin{bmatrix}\psi_0\\ z_0\end{bmatrix}\right\|_2,\quad\forall t\ge 0.
$$

On the other hand

$$
\left\|\begin{bmatrix}\boldsymbol{L}_t^1-\boldsymbol{L}_t^2\\ \boldsymbol{V}_t^1-\boldsymbol{V}_t^2\end{bmatrix}\right\|_{\mathbb{L}_2}\le\|A\|_2\left\|\begin{bmatrix}\psi_t\\ z_t\end{bmatrix}\right\|_2,
$$

which yields the following inequality:

$$
\left\|\begin{bmatrix}\boldsymbol{L}_t^1-\boldsymbol{L}_t^2\\ \boldsymbol{V}_t^1-\boldsymbol{V}_t^2\end{bmatrix}\right\|_{\mathbb{L}_2}\le\|A\|_2\|A^{-1}\|_2\exp\left\{\frac{\left(\lambda_-^2-\alpha(t)\right)\vee\left(M+\alpha(t)-\lambda_+^2\right)}{\lambda_+-\lambda_-}t\right\}\left\|\begin{bmatrix}\boldsymbol{L}_0^1-\boldsymbol{L}_0^2\\ \boldsymbol{V}_0^1-\boldsymbol{V}_0^2\end{bmatrix}\right\|_{\mathbb{L}_2}.
$$

The next technical lemmas gives an upper bound for $\|A\|_2\|A^{-1}\|_2$.

Lemma 22. *For every $\lambda_+>\lambda_->0$, such that $\lambda_++\lambda_->2$ the following inequality is true*

$$
\left\|\begin{pmatrix}\lambda_+I_p & I_p\\ -\lambda_-I_p & -I_p\end{pmatrix}\right\|_2\times\left\|\begin{pmatrix}\lambda_+I_p & I_p\\ -\lambda_-I_p & -I_p\end{pmatrix}^{-1}\right\|_2\le\frac{2(\lambda_++\lambda_-)^2}{\lambda_+-\lambda_-}.
$$

This lemma yields the inequality below:

$$
\left\|\begin{bmatrix}\boldsymbol{L}_t^1-\boldsymbol{L}_t^2\\ \boldsymbol{V}_t^1-\boldsymbol{V}_t^2\end{bmatrix}\right\|_{\mathbb{L}_2}\le\frac{2\eta^2}{\eta-2\lambda_-}\exp\left\{\frac{\left(\lambda_-^2-m-\alpha(t)\right)\vee\left(M+\alpha(t)-(\eta-\lambda_-)^2\right)}{\eta-2\lambda_-}t\right\}\left\|\begin{bmatrix}\boldsymbol{L}_0^1-\boldsymbol{L}_0^2\\ \boldsymbol{V}_0^1-\boldsymbol{V}_0^2\end{bmatrix}\right\|_{\mathbb{L}_2}.
$$

The right-hand side of the inequality is an increasing function of λ_-, as $\lambda_-\in[0,\eta/2)$.

Therefore

$$\left\|\begin{bmatrix} \boldsymbol{L}_t^1 - \boldsymbol{L}_t^2 \\ \boldsymbol{V}_t^1 - \boldsymbol{V}_t^2 \end{bmatrix}\right\|_{\mathbb{L}_2} \leq 2\eta \exp\left\{(-m - \alpha(t)) \vee \left(M + \alpha(t) - \eta^2\right)\frac{t}{\eta}\right\} \left\|\begin{bmatrix} \boldsymbol{L}_0^1 - \boldsymbol{L}_0^2 \\ \boldsymbol{V}_0^1 - \boldsymbol{V}_0^2 \end{bmatrix}\right\|_{\mathbb{L}_2}.$$

4.F.2 Proof of Lemma 19

Proof. From the definition of Wassertein distance and triangle inequality, we have

$$W_2\left(\nu_{t+\delta}^{\mathsf{PKLD}}, \boldsymbol{Q}_{t,\delta}\right) \leq \|\boldsymbol{L}_{t+\delta} - \widetilde{\boldsymbol{L}}_{t+\delta}\|_{\mathbb{L}_2} + \|\boldsymbol{V}_{t+\delta} - \widetilde{\boldsymbol{V}}_{t+\delta}\|_{\mathbb{L}_2} := \Phi(\delta).$$

Since the processes $(\boldsymbol{L}, \boldsymbol{V})$ and $(\widetilde{\boldsymbol{L}}, \widetilde{\boldsymbol{V}})$ have the same Gaussian noise, then using the formulas (PKLD) and (†-KLD), we obtain

$$\begin{aligned}\Phi(\delta) \leq (\eta+1) &\left\|\int_0^\delta \left[\boldsymbol{V}_{t+s} - \widetilde{\boldsymbol{V}}_{t+s}\right] ds\right\|_{\mathbb{L}_2} \\ &+ \left\|\int_0^\delta \left[\nabla f(\boldsymbol{L}_{t+s}) - \nabla f(\widetilde{\boldsymbol{L}}_{t+s}) + \alpha(t+s)\boldsymbol{L}_{t+s} - \alpha(t)\widetilde{\boldsymbol{L}}_{t+s}\right] ds\right\|_{\mathbb{L}_2}.\end{aligned}$$

Since the gradient of f is M-Lipschitz continuous, adding and substracting $\alpha(t+s)\widetilde{\boldsymbol{L}}_{t+s}$ in the integral we get the following inequality:

$$\begin{aligned}\Phi(\delta) \leq (\eta+1) &\int_0^\delta \left\|\boldsymbol{V}_{t+s} - \widetilde{\boldsymbol{V}}_{t+s}\right\|_{\mathbb{L}_2} ds + \int_0^\delta (M + \alpha(t+s)) \left\|\boldsymbol{L}_{t+s} - \widetilde{\boldsymbol{L}}_{t+s}\right\|_{\mathbb{L}_2} ds \\ &+ \int_0^\delta |\alpha(t+s) - \alpha(t)| \left\|\widetilde{\boldsymbol{L}}_{t+s}\right\|_{\mathbb{L}_2} ds.\end{aligned}$$

Thus,

$$\Phi(\delta) \leq \int_0^\delta \max\left(\eta+1, M + \alpha(t+s)\right)\Phi(s) ds + \int_0^\delta |\alpha(t+s) - \alpha(t)| \left\|\widetilde{\boldsymbol{L}}_{t+s}\right\|_{\mathbb{L}_2} ds.$$

Let us show now, that $\|\boldsymbol{L}_{t+s}\|_{\mathbb{L}_2}$ is bounded by a constant that does not depend on s. In order to do that we express the $\mathbb{L}_2$-norm of the random vector as the Wasserstein distance between Dirac measure and its distribution:

$$\left\|\widetilde{\boldsymbol{L}}_{t+s}\right\|_{\mathbb{L}_2} \leq \left\|(\widetilde{\boldsymbol{L}}_{t+s}, \widetilde{\boldsymbol{V}}_{t+s})\right\|_{\mathbb{L}_2} = W_2(\boldsymbol{Q}_{t,s}, \delta_0) \leq W_2(\boldsymbol{Q}_{t,s}, \mathsf{P}_{\alpha(t)}) + W_2(\mathsf{P}_{\alpha(t)}, \delta_0).$$

As we know from [DRD20][Theorem 1], the Wasserstein error at the moment s is bounded by the initial error at moment 0: $W_2(\boldsymbol{Q}_{t,s}, \mathsf{P}_{\alpha(t)}) \leq c \times W_2(\boldsymbol{Q}_{t,0}, \mathsf{P}_{\alpha(t)})$. Here c is a constant that depends neither on t nor on s. Therefore

$$\begin{aligned}\left\|\widetilde{\boldsymbol{L}}_{t+s}\right\|_{\mathbb{L}_2} &\le cW_2(\boldsymbol{Q}_{t,0}, \mathsf{P}_{\alpha(t)}) + W_2(\mathsf{P}_{\alpha(t)}, \delta_0)\\ &= cW_2(\nu_t^{\mathsf{PKLD}}, \mathsf{P}_{\alpha(t)}) + \big(\mu_2(\mathsf{P}_{\alpha(t)})\big)^{1/2}.\end{aligned}$$

We define the right hand side of the last inequality as V_t. Summing up, we get the following recurrent inequality for the function $\Phi(\delta)$:

$$\Phi(\delta) \le \int_0^\delta \max\big(\eta+1, M+\alpha(t+s)\big)\Phi(s)ds + V_t\int_0^\delta |\alpha(t+s)-\alpha(t)|ds.$$

Finally, applying Gronwall's inverse inequality we conclude the proof. □

4.F.3 Proof of Lemma 22

We need to prove that

$$\|A\|\,\|A^{-1}\| \le \frac{2(\lambda_+ + \lambda_-)^2}{\lambda_+ - \lambda_-}, \qquad \text{where} \qquad A = \begin{pmatrix} \lambda_+ I_p & I_p \\ -\lambda_- I_p & -I_p \end{pmatrix}.$$

Let us first notice that applying a simple permutation of rows and columns the matrix A becomes a block matrix. Indeed,

$$\|A\| = \left\|\begin{pmatrix} B & 0_2 & \dots & 0_2 \\ 0_2 & B & & \\ \vdots & & \ddots & \vdots \\ 0_2 & & \dots & B \end{pmatrix}\right\|, \qquad \text{where} \qquad B = \begin{pmatrix} \lambda_+ & 1 \\ -\lambda_- & -1 \end{pmatrix}.$$

Therefore $\|A\| \le \|B\|$. Similarly one can check that $\|A^{-1}\| \le \|B^{-1}\|$. Let us bound these two norms separately. In order to find the eigenvalues of $B^\top B$ we have to solve the following quadratic equation:

$$(x-2)\left(x - (\lambda_+^2 + \lambda_-^2)\right)x = (\lambda_+ + \lambda_-)^2.$$

We see that for the value of $x_0 = 2(\lambda_+ + \lambda_-)^2$ the quadratic function is larger than the right-hand side. In addition we notice that x_0 is larger than both roots of the left-hand side, thus the solution x bounded by x_0. The latter yields that $\|B\| \le \sqrt{2}(\lambda_+ + \lambda_-)$.

Similarly one can verify that $\|B^{-1}\| \leq \sqrt{2}(\lambda_+ + \lambda_-)/(\lambda_+ - \lambda_-)$. Therefore

$$\|A\|\|A^{-1}\| \leq \|B\|\|B^{-1}\| \leq \frac{2(\lambda_+ + \lambda_-)^2}{(\lambda_+ - \lambda_-)}.$$

This concludes the proof.

4.G Penalized Gradient Flow

4.G.1 Proof of Theorem 16

We recall that for every $\gamma \in \mathbb{R}$, is given by $f_\gamma(\cdot) := f(\cdot) + \gamma \| \cdot \|_2^2/2$. We define $\boldsymbol{x}(\gamma)$ the minimum point of f_γ. In particular, $\boldsymbol{x}_0 = \boldsymbol{x}_*$. The triangle inequality yields

$$\|\boldsymbol{X}_t^{\mathsf{PGF}} - \boldsymbol{x}_*\|_2 \leq \|\boldsymbol{X}_t^{\mathsf{PGF}} - \boldsymbol{x}_{\alpha(t)}\|_2 + \|\boldsymbol{x}_{\alpha(t)} - \boldsymbol{x}_0\|_2 \tag{4.19}$$

for every $t > 0$. We will bound these two terms separately. **A**(A, q) for $\gamma = 0$ and $\widetilde{\gamma} = \alpha(t)$ yields the following bound on the second term:

$$\|\boldsymbol{x}_{\alpha(t)} - \boldsymbol{x}_0\|_2 \leq \mathsf{D}\alpha(t)^{1-\mathsf{q}}.$$

To bound the first term of (4.19), we aim at obtaining a Gronwall-type inequality for the function

$$\phi(t) := \|\boldsymbol{X}_t^{\mathsf{PGF}} - \boldsymbol{x}_{\alpha(t)}\|_2.$$

To this end, we consider an auxiliary stochastic process $\{\widetilde{\boldsymbol{X}}_u : u \geq t\}$, defined as a solution of the following differential equation

$$d\widetilde{\boldsymbol{X}}_u = -\big(\nabla f(\widetilde{\boldsymbol{X}}_u) + \alpha(t)\widetilde{\boldsymbol{X}}_u\big)du,$$

with the starting point $\widetilde{\boldsymbol{X}}_t = \boldsymbol{X}_t$. This is in fact the gradient flow corresponding to the strongly-convex potential $f_{\alpha(t)}$. The triangle inequality yields

$$\phi(t+\delta) \leq \left\|\boldsymbol{X}_{t+\delta}^{\mathsf{PGF}} - \widetilde{\boldsymbol{X}}_{t+\delta}\right\|_2 + \left\|\widetilde{\boldsymbol{X}}_{t+\delta} - \boldsymbol{x}_{\alpha(t)}\right\|_2 + \left\|\boldsymbol{x}_{\alpha(t)} - \boldsymbol{x}_{\alpha(t+\delta)}\right\|_2.$$

From the linear convergence of the gradient flow of an $\alpha(t)$-strongly convex function, we get the following:

$$\left\|\boldsymbol{X}_{t+\delta}^{\mathsf{PGF}} - \boldsymbol{x}_{\alpha(t)}\right\|_2 \leq \exp\big(-\delta\alpha(t)\big)\left\|\widetilde{\boldsymbol{X}}_t - \boldsymbol{x}_{\alpha(t)}\right\|_2 = \exp\left(-\delta\alpha(t)\right)\phi(t).$$

In order to bound the distance between $\boldsymbol{x}_{\alpha(t)}$ and $\boldsymbol{x}_{\alpha(t+\delta)}$, we use again $\mathbf{A}(\mathsf{A},\mathsf{q})$ condition, thus

$$\left\|\boldsymbol{x}_{\alpha(t)} - \boldsymbol{x}_{\alpha(t+\delta)}\right\|_2 \leq \frac{\mathsf{D}}{\alpha^{\mathsf{q}}(t)}(\alpha(t) - \alpha(t+\delta))\|\boldsymbol{x}_*\|_2.$$

Thus we obtain a bound for $\phi(t+\delta)$, that depends linearly on $\phi(t)$:

$$\phi(t+\delta) \leq \left\|\boldsymbol{X}_{t+\delta}^{\mathsf{PGF}} - \widetilde{\boldsymbol{X}}_{t+\delta}\right\|_2 + e^{-\delta\alpha(t)}\phi(t) + \frac{\mathsf{D}}{\alpha^{\mathsf{q}}(t)}(\alpha(t) - \alpha(t+\delta))\|\boldsymbol{x}_*\|_2. \tag{4.20}$$

Let us subtract $\phi(t)$ from both sides of (4.20) and divide by δ:

$$\begin{aligned}\frac{\phi(t+\delta) - \phi(t)}{\delta} \leq \frac{1}{\delta}\cdot\left\|\boldsymbol{X}_{t+\delta}^{\mathsf{PGF}} - \widetilde{\boldsymbol{X}}_{t+\delta}\right\|_2 &+ \frac{\exp\left(-\delta\alpha(t)\right) - 1}{\delta}\cdot\phi(t) \\ &+ \frac{\mathsf{D}(\alpha(t) - \alpha(t+\delta))}{\delta\alpha^{\mathsf{q}}(t)}\|\boldsymbol{x}_*\|_2.\end{aligned} \tag{4.21}$$

The next lemma provides an upper bound on $\left\|\boldsymbol{X}_{t+\delta}^{\mathsf{PGF}} - \widetilde{\boldsymbol{X}}_{t+\delta}\right\|_2$ showing that it is $o(\delta)$, when $\delta \to 0$.

Lemma 23. *Suppose f satisfies (m, M)-SCGL with $m = 0$. Then for every $t, \delta > 0$, and for every integrable function $\alpha : [t, t+\delta] \to [0, \infty)$,*

$$\|\widetilde{\boldsymbol{X}}_{t+\delta} - \boldsymbol{X}_{t+\delta}^{\mathsf{PGF}}\|_2 \leq \left(\phi(t) + \|\boldsymbol{x}_{\alpha(t)}\|_2\right)\exp\left\{M\delta + \int_0^\delta \alpha(t+u)\,du\right\}\int_0^\delta \left|\alpha(t+s) - \alpha(t)\right| ds.$$

The proof can be found in the Section 4.G.2. When δ tends to 0, according to Lemma 23, the first term of the right-hand side of (4.21) vanishes. Thus, after passing to the limit, we are left with the following Gronwall-type inequality:

$$\phi'(t) \leq -\alpha(t)\phi(t) - \frac{\mathsf{D}\alpha'(t)}{\alpha^{\mathsf{q}}(t)}\cdot\|\boldsymbol{x}_*\|_2. \tag{4.22}$$

Here we tacitly used the fact that $\|\boldsymbol{x}_{\alpha(t+\delta)}\|_2 \leq \|\boldsymbol{x}_0\|_2$. Recalling that the function $\beta(t)$ is given by $\beta(t) = \int_0^t \alpha(s)ds$, one can rewrite (4.22) as

$$\left(\phi(t)e^{\beta(t)}\right)' \leq -\frac{\mathsf{D}\alpha'(t)e^{\beta(t)}}{\alpha^{\mathsf{q}}(t)}\|\boldsymbol{x}_*\|_2.$$

Therefore we infer the following bound on $\phi(t)$:

$$\phi(t) \leq \phi(0)e^{-\beta(t)} - \mathsf{D}\|\boldsymbol{x}_*\|_2\int_0^t \frac{\alpha'(s)}{\alpha^{\mathsf{q}}(s)}e^{\beta(s)-\beta(t)}ds.$$

Combining this bound with (4.7), we obtain the inequality

$$\|\boldsymbol{X}_t^{\mathsf{PGF}} - \boldsymbol{x}_0\|_2 \leq \|\boldsymbol{X}_0^{\mathsf{PGF}} - \boldsymbol{x}(\alpha(0))\|_2 e^{-\beta(t)} - \mathsf{D}\|\boldsymbol{x}_*\|_2 \int_0^t \frac{\alpha'(s)}{\alpha^{\mathsf{q}}(s)} e^{\beta(s)-\beta(t)} ds + \mathsf{D}\|\boldsymbol{x}_*\|_2 \alpha(t)^{1-\mathsf{q}}.$$

Since the process starts at point 0, $\|\boldsymbol{X}_0^{\mathsf{PGF}} - \boldsymbol{x}(\alpha(0))\|_2 = \|\boldsymbol{x}_{\alpha(0)}\|_2$. The next lemma bounds $\|\boldsymbol{x}_{\alpha(0)}\|_2$.

Lemma 24. *The function* $\gamma \mapsto \|\boldsymbol{x}(\gamma)\|_2$ *is a non-increasing continuous function on the interval* $[0, \infty)$.

Therefore, $\|\boldsymbol{x}_{\alpha(0)}\|_2 \leq \|\boldsymbol{x}_0\|_2 = \|\boldsymbol{x}_*\|_2$, which completes the proof of Theorem 16.

4.G.2 Proof of Lemma 23

From the definition of $\widetilde{\boldsymbol{X}}$, we can write

$$\widetilde{\boldsymbol{X}}_{t+\delta} - \boldsymbol{X}_{t+\delta}^{\mathsf{PGF}} = \int_t^{t+\delta} \left(\nabla f(\boldsymbol{X}_s^{\mathsf{PGF}}) - \nabla f(\widetilde{\boldsymbol{X}}_s) + \alpha(s)\boldsymbol{X}_s^{\mathsf{PGF}} - \alpha(t)\widetilde{\boldsymbol{X}}_s\right) ds.$$

Therefore we have

$$\|\widetilde{\boldsymbol{X}}_{t+\delta} - \boldsymbol{X}_{t+\delta}^{\mathsf{PGF}}\|_2 \leq \Bigg\| \underbrace{\int_t^{t+\delta} \left(\nabla f(\boldsymbol{X}_s^{\mathsf{PGF}}) - \nabla f(\widetilde{\boldsymbol{X}}_s)\right) ds}_{:=T_1} \Bigg\|_2 + \Bigg\| \underbrace{\int_t^{t+\delta} \left(\alpha(s)\boldsymbol{X}_s^{\mathsf{PGF}} - \alpha(t)\widetilde{\boldsymbol{X}}_s\right) ds}_{:=T_2} \Bigg\|_2.$$

Now let us analyze these two terms separately. We start with T_1:

$$\begin{aligned}
\|T_1\|_2 &= \left\| \int_t^{t+\delta} \left(\nabla f(\boldsymbol{X}_s^{\mathsf{PGF}}) - \nabla f(\widetilde{\boldsymbol{X}}_s)\right) ds \right\|_2 \\
&\leq \int_t^{t+\delta} \left\| \nabla f(\boldsymbol{X}_s^{\mathsf{PGF}}) - \nabla f(\widetilde{\boldsymbol{X}}_s) \right\|_2 ds \\
&\leq M \int_t^{t+\delta} \|\boldsymbol{X}_s^{\mathsf{PGF}} - \widetilde{\boldsymbol{X}}_s\|_2 ds.
\end{aligned}$$

These are due to the Minkowskii inequality and the Lipschitz continuity of the gradient. In order to bound the second term T_2, we will add and subtract the term $\alpha(t+s)\widetilde{\boldsymbol{X}}_{t+s}$. Similar to the case above, we get the following upper bound:

$$\begin{aligned}
\|T_2\|_2 &\leq \int_t^{t+\delta} \alpha(s) \left\|\boldsymbol{X}_s^{\mathsf{PGF}} - \widetilde{\boldsymbol{X}}_s\right\|_2 ds + \int_t^{t+\delta} |\alpha(s) - \alpha(t)| \left\|\widetilde{\boldsymbol{X}}_s\right\|_2 ds \\
&= \int_0^{\delta} \alpha(t+s) \left\|\boldsymbol{X}_{t+s}^{\mathsf{PGF}} - \widetilde{\boldsymbol{X}}_{t+s}\right\|_2 ds + \int_0^{\delta} |\alpha(t+s) - \alpha(t)| \left\|\widetilde{\boldsymbol{X}}_{t+s}\right\|_2 ds.
\end{aligned}$$

Recall that $\widetilde{\boldsymbol{X}}_{t+s}$ is the gradient flow of an $(m+\alpha(t))$-strongly convex potential function. Thus, the triangle inequality yields

$$\begin{aligned}
\big\|\widetilde{\boldsymbol{X}}_{t+s}\big\|_2 &\le \big\|\widetilde{\boldsymbol{X}}_{t+s} - \boldsymbol{x}(t)\big\|_2 + \|\boldsymbol{x}(t)\|_2 \\
&\le \big\|\widetilde{\boldsymbol{X}}_{t} - \boldsymbol{x}(t)\big\|_2 \exp(-ms - \alpha(t)s) + \|\boldsymbol{x}(t)\|_2 \\
&\le \big\|\boldsymbol{X}^{\mathsf{PGF}}_{t} - \boldsymbol{x}(t)\big\|_2 + \|\boldsymbol{x}(t)\|_2 := V_t.
\end{aligned}$$

Summing up, we have

$$\big\|\boldsymbol{X}^{\mathsf{PGF}}_{t+\delta} - \widetilde{\boldsymbol{X}}_{t+\delta}\big\|_2 \le \int_0^\delta \big(M + \alpha(t+s)\big)\|\boldsymbol{X}^{\mathsf{PGF}}_{t+s} - \widetilde{\boldsymbol{X}}_{t+s}\|_2 ds + \widetilde{\alpha}_t(\delta)\, V_t,$$

where $\widetilde{\alpha}_t(\delta)$ is an auxiliary function defined as

$$\widetilde{\alpha}_t(\delta) := \int_0^\delta |\alpha(t+s) - \alpha(t)|\, ds.$$

Now let us define $\Phi(s) = \|\boldsymbol{X}^{\mathsf{PGF}}_{t+s} - \widetilde{\boldsymbol{X}}_{t+s}\|_{\mathbb{L}_2}$. The last inequality can be rewritten as

$$\Phi(\delta) \le \int_0^\delta \big(M + \alpha(t+s)\big)\Phi(s)\, ds + \widetilde{\alpha}_t(\delta)\, V_t.$$

The (integral form of the) Gronwall inequality implies that

$$\begin{aligned}
\Phi(\delta) &\le V_t \int_0^\delta \widetilde{\alpha}_t(s)\big(M + \alpha(t+s)\big) e^{\int_s^\delta (M+\alpha(t+u))\, du}\, ds + \widetilde{\alpha}_t(\delta) V_t \\
&= V_t \int_0^\delta \widetilde{\alpha}_t'(s)\, e^{\int_s^\delta (M+\alpha(t+u))\, du}\, ds \\
&\le V_t\, \widetilde{\alpha}_t(\delta)\, \exp\left\{M\delta + \int_0^\delta \alpha(t+u)\, du\right\}.
\end{aligned}$$

This completes the proof.

4.G.3 Proof of Lemma 24

Suppose that $\gamma_1 < \gamma_2$. Let us show that $\|\boldsymbol{x}(\gamma_1)\|_2 > \|\boldsymbol{x}(\gamma_2)\|_2$. Let us consider the function f_{γ_2}. We have that

$$f_{\gamma_2}(\boldsymbol{x}(\gamma_2)) \le f_{\gamma_2}(\boldsymbol{x}(\gamma_1)) = f(\boldsymbol{x}(\gamma_1)) + \gamma_2\|\boldsymbol{x}(\gamma_1)\|_2/2.$$

The definition of f_{γ_1} yields

$$\begin{aligned} f_{\gamma_2}(\boldsymbol{x}(\gamma_2)) &\leq f_{\gamma_1}(\boldsymbol{x}(\gamma_1)) + (\gamma_2 - \gamma_1)\|\boldsymbol{x}(\gamma_1)\|_2/2 \\ &\leq f_{\gamma_1}(\boldsymbol{x}(\gamma_2)) + (\gamma_2 - \gamma_1)\|\boldsymbol{x}(\gamma_1)\|_2/2 \\ &\leq f_{\gamma_2}(\boldsymbol{x}(\gamma_2)) + (\gamma_2 - \gamma_1)\big(\|\boldsymbol{x}(\gamma_1)\|_2 - \|\boldsymbol{x}(\gamma_2)\|_2\big)/2. \end{aligned}$$

Here the second passage is valid, as $\boldsymbol{x}(\gamma_1)$ is the minimum point of f_{γ_1}. Since $\gamma_2 > \gamma_1$, the difference $\|\boldsymbol{x}(\gamma_1)\|_2 - \|\boldsymbol{x}(\gamma_2)\|_2$ is positive. Thus the monotony is proved.
To prove the continuity of the function we take a sequence γ_n that tends to γ_0 and show that $\boldsymbol{x}_{\gamma_n} \to \boldsymbol{x}_{\gamma_0}$. Assumption $\mathbf{A}(\mathsf{D},\mathsf{q})$ yields

$$\|\boldsymbol{x}_{\gamma_n} - \boldsymbol{x}_{\gamma_0}\|_2 \leq \frac{\mathsf{D}}{\max(\gamma_n,\gamma_0)^{\mathsf{q}}}|\gamma_n - \gamma_0|\|\boldsymbol{x}_*\|_2, \qquad \forall n \in \mathbb{N}.$$

Since $\mathsf{q} < 1$, the ratio of $|\gamma_n - \gamma_0|$ and $\max(\gamma_n,\gamma_0)^{\mathsf{q}}$ tends to zero, when $n \to 0$. This concludes the proof.

4.H Examples of functions satisfying condition $\mathbf{A}(\mathsf{D},\mathsf{q})$

In this section we consider several functions that are convex but not strongly convex and satisfy $\mathbf{A}(\mathsf{D},\mathsf{q})$ condition presented in Section 4.3.

4.H.1 Locally strongly convex functions

We prove that locally strongly convex functions satisfy $\mathbf{A}(\mathsf{D},0)$. Recalling Lemma 24 we get that $\|\boldsymbol{x}_\gamma\|_2 \leq \|\boldsymbol{x}_*\|_2$. Thus the we can consider the function only on $\mathcal{B}(0,\|x_*\|_2)$. Since f is locally strongly convex, there exists m_* such that it is m_*-strongly convex in the ball $\mathcal{B}(0,\|x_*\|_2)$. The latter means, that $f_{\widetilde{\gamma}}$ is $(m_* + \widetilde{\gamma})$-strongly convex. Therefore [Nes04][Theorem 2.1.9] yields the following:

$$\|\boldsymbol{x}_\gamma - \boldsymbol{x}_{\widetilde{\gamma}}\|_2 \leq \frac{1}{m_* + \widetilde{\gamma}}\|\nabla f_{\widetilde{\gamma}}(\boldsymbol{x}_\gamma) - \nabla f_{\widetilde{\gamma}}(\boldsymbol{x}_{\widetilde{\gamma}})\|_2.$$

Using the optimality condition for differentiable functions one gets $\nabla f_{\widetilde{\gamma}}(\boldsymbol{x}_\gamma) = (\widetilde{\gamma} - \gamma)\boldsymbol{x}_\gamma$ for all $\gamma \geq 0$. Therefore, for every $0 \leq \gamma < \widetilde{\gamma}$, we obtain

$$\|\boldsymbol{x}_\gamma - \boldsymbol{x}_{\widetilde{\gamma}}\|_2 \leq \frac{1}{m_* + \widetilde{\gamma}}\|(\widetilde{\gamma} - \gamma)\boldsymbol{x}_\gamma\|_2 \leq \frac{\widetilde{\gamma} - \gamma}{m_*}\|\boldsymbol{x}_\gamma\|_2.$$

The latter is true due to Lemma 24. Thus f satisfies $\mathbf{A}(1/m_*,0)$.

4.H.2 Cubic function $f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_*\|_2^3$

In this section we show that the cubic function satisfies **A**$(1/\sqrt{3\|\boldsymbol{x}_*\|_2}, 1/2)$. It is straightforward to verify that the function f is convex. f_γ is strongly convex and the optimality condition for $\boldsymbol{x}_\gamma$ yields the following equality:

$$\nabla f(\boldsymbol{x}_\gamma) + \gamma \boldsymbol{x}_\gamma = 3\|\boldsymbol{x}_\gamma - \boldsymbol{x}_*\|_2(\boldsymbol{x}_\gamma - \boldsymbol{x}_*) + \gamma \boldsymbol{x}_\gamma = 0. \tag{4.23}$$

In the case when $\boldsymbol{x}_* = 0$, the penalized minimum point $\boldsymbol{x}_\gamma$ equals 0, for every γ, thus we suppose in the following that $\boldsymbol{x}_* \neq 0$. Since the norm is scalar, (4.23) yields that the vectors $\boldsymbol{x}_\gamma - \boldsymbol{x}_*$ and $\boldsymbol{x}_\gamma$ are co-linear. Therefore there exists a real number λ_γ such that $\boldsymbol{x}_\gamma = \lambda_\gamma \boldsymbol{x}_*$. Lemma 24 implies that $|\lambda_\gamma| \leq 1$, thus the following quadratic equality is true:

$$-3\|\boldsymbol{x}_*\|_2(\lambda_\gamma - 1)^2 \boldsymbol{x}_* + \gamma\lambda_\gamma \boldsymbol{x}_* = 0. \tag{4.24}$$

As said in the beginning, $\boldsymbol{x}_* \neq 0$, therefore it its coefficient that is equal to zero. Solving the quadratic equation with respect to λ_γ, we get the following formula:

$$\lambda_\gamma = 1 - \frac{\gamma}{\gamma/2 + \sqrt{3\gamma\|\boldsymbol{x}_*\|_2 + \gamma^2/4}}.$$

According to Lemma 24, for every $\widetilde{\gamma} > \gamma$, we have $|\lambda_\gamma| > |\lambda_{\widetilde{\gamma}}|$. On the other hand, from (4.24) one deduces that $\lambda_\gamma > 0$, for every $\gamma > 0$. Thus, inserting the found value for λ_γ, we obtain the following inequality:

$$\begin{aligned}
\|\boldsymbol{x}_\gamma - \boldsymbol{x}_{\widetilde{\gamma}}\|_2 &= \|\boldsymbol{x}_*\|_2 \left(\frac{\widetilde{\gamma}}{\widetilde{\gamma}/2 + \sqrt{3\widetilde{\gamma}\|\boldsymbol{x}_*\|_2 + \widetilde{\gamma}^2/4}} - \frac{\gamma}{\gamma/2 + \sqrt{3\gamma\|\boldsymbol{x}_*\|_2 + \gamma^2/4}} \right) \\
&\leq \frac{(\widetilde{\gamma} - \gamma)\|\boldsymbol{x}_*\|_2}{\widetilde{\gamma}/2 + \sqrt{3\widetilde{\gamma}\|\boldsymbol{x}_*\|_2 + \widetilde{\gamma}^2/4}} \leq \frac{(\widetilde{\gamma} - \gamma)\|\boldsymbol{x}_*\|_2}{\sqrt{3\widetilde{\gamma}\|\boldsymbol{x}_*\|_2}}.
\end{aligned}$$

Therefore f satisfies **A**$(1/\sqrt{3\|\boldsymbol{x}_*\|_2}, 1/2)$.

4.H.3 Power function $f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_*\|_2^a$

For $a \geq 2$, we consider the function $f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_*\|_2^a$. We show here that f satisfies **A**$((1/a\|\boldsymbol{x}_*\|_2^{a-2})^{1/(a-1)}, (a-2)/(a-1))$. Since f_γ is a differentiable strongly-convex function, we get the following equation for $\boldsymbol{x}_\gamma$:

$$a\|\boldsymbol{x}_\gamma - \boldsymbol{x}_*\|_2^{a-2}(\boldsymbol{x}_\gamma - \boldsymbol{x}_*) + \gamma \boldsymbol{x}_\gamma = 0. \tag{4.25}$$

Similar to the previous case, we notice that $\boldsymbol{x}_\gamma - \boldsymbol{x}_*$ and $\boldsymbol{x}_\gamma$ are co-linear. Thus, there exists λ_γ such that $\boldsymbol{x}_\gamma = (1-\lambda_\gamma)\boldsymbol{x}_*$. Since $\boldsymbol{x}_*$ is assumed to be non-zero, in order to calculate $\boldsymbol{x}_\gamma$, one needs to solve the following equation:

$$|\lambda_\gamma|^{a-2}\lambda_\gamma = \frac{\gamma(1-\lambda_\gamma)}{a\|\boldsymbol{x}_*\|^{a-2}}. \tag{4.26}$$

Thus the p-dimensional equation (4.25) reduces to equation (4.26) involving a one-dimensional unknown. Lemma 24 yields $\lambda_{\widetilde{\gamma}} > \lambda_\gamma > 0$ for every $\widetilde{\gamma} > \gamma \geq 0$. In addition, from (4.26), we have that $\lambda_\gamma \leq 1$ for every $\gamma > 0$. It is straightforward to verify that for every $\gamma \geq 0$, (4.26) has exactly one solution satisfying these conditions.

Lemma 25. *Let $\alpha \geq 1$. If $(\lambda_s : s \in (0,1))$ satisfies $\lambda_s^\alpha = s(1-\lambda_s)$ for every $s \in (0,1)$, then*

$$|\lambda_s - \lambda_{s'}| \leq \frac{|s-s'|}{(s \vee s')^{(\alpha-1)/\alpha}}, \qquad \forall s', s \in (0,1).$$

Proof. Without loss of generality, we assume that $s' \leq s$. Computing the derivative of both sides of the identity $\lambda_s^\alpha = s(1-\lambda_s)$, we get

$$\lambda_s' = \frac{1-\lambda_s}{\alpha\lambda_s^{\alpha-1} + s} \geq 0.$$

This implies that $\lambda_{s'} \leq \lambda_s$. In addition,

$$\begin{aligned}
\lambda_s - \lambda_{s'} &\leq \frac{\lambda_s^\alpha - \lambda_{s'}^\alpha}{\lambda_s^{\alpha-1}} \\
&= \frac{s(1-\lambda_s) - s'(1-\lambda_{s'})}{\lambda_s^{\alpha-1}} \\
&= \frac{(s-s')(1-\lambda_{s'})}{\lambda_s^{\alpha-1}} - \frac{s(\lambda_s - \lambda_{s'})}{\lambda_s^{\alpha-1}}.
\end{aligned}$$

Rearranging the terms, we arrive at

$$\begin{aligned}
\lambda_s - \lambda_{s'} &\leq \frac{(s-s')(1-\lambda_{s'})}{\lambda_s^{\alpha-1}\big(1 + \frac{s}{\lambda_s^{\alpha-1}}\big)} \\
&= \frac{(s-s')(1-\lambda_{s'})}{\lambda_s^{\alpha-1} + s}
\end{aligned}$$

In the last fraction, the numerator is bounded by $s - s'$, while the denominator satisfies

$$\begin{aligned}
\lambda_s^{\alpha-1} + s &= \big(s(1-\lambda_s)\big)^{(\alpha-1)/\alpha} + s \\
&\geq \big(s(1-s^{1/\alpha})\big)^{(\alpha-1)/\alpha} + s \\
&\geq s^{(\alpha-1)/\alpha}(1-s^{1/\alpha}) + s = s^{(\alpha-1)/\alpha}.
\end{aligned}$$

This completes the proof of the lemma. □

Applying this lemma to (4.26), we get

$$\lambda_{\widetilde{\gamma}} - \lambda_{\gamma} \leq \frac{\widetilde{\gamma} - \gamma}{a\|\boldsymbol{x}_*\|_2^{a-2}(\widetilde{\gamma}/a\|\boldsymbol{x}_*\|_2^{a-2})^{(a-2)/(a-1)}} = \frac{\widetilde{\gamma} - \gamma}{a^{1/(a-1)}\|\boldsymbol{x}_*\|_2^{(a-2)/(a-1)}\widetilde{\gamma}^{(a-2)/(a-1)}},$$

for all $\gamma, \widetilde{\gamma}$ satisfying $0 \leq \gamma \leq \widetilde{\gamma} \leq a\|\boldsymbol{x}_*\|_2^{a-2}$. In conclusion, we get

$$\begin{aligned} \|\boldsymbol{x}_{\widetilde{\gamma}} - \boldsymbol{x}_{\gamma}\|_2 &\leq \frac{\widetilde{\gamma} - \gamma}{a^{1/(a-1)}\|\boldsymbol{x}_*\|_2^{(a-2)/(a-1)}\widetilde{\gamma}^{(a-2)/(a-1)}}\|\boldsymbol{x}_*\|_2 \\ &\leq \frac{\widetilde{\gamma} - \gamma}{\widetilde{\gamma}^{(a-2)/(a-1)}}(\|\boldsymbol{x}_*\|_2/a)^{1/(a-1)}. \end{aligned}$$

This concludes the proof.

Chapter 5

Summary and Perspectives

5.1 Summary of the Thesis

In this thesis, we studied various versions of Langevin sampling from log-concave and strongly log-concave targets.

Our first contribution consists of three main parts. First, we proposed a framework of analysis of convergence of the LMC with inexact gradients. The framework includes both the deterministic and stochastic biases of the gradient evaluations. We proved explicit non-asymptotic bounds on the Wasserstein error for the noisy LMC with strongly convex and gradient-Lipschitz potentials, as well as the case of second-order smooth potentials. Then, we proved that the LMC with a varying step-size can achieve faster convergence rate when compared to the constant step-sized LMC. In particular, it allows to get rid of the logarithmic multiplier in the rate. In the end, we proposed a second-order method LMCO$'$, which performs as many iterations as LMCO, while the computational complexity of each of its iterations is of order p.

In our second contribution, we focused on the non-strongly log-concave targets. We introduced a fixed-time quadratic penalization to the potential function. The penalization is then applied to LMC, KLMC and KLMC-2 algorithms. Non-asymptotic bounds are proved for W_1, W_2 and total-variation errors, with explicit dependence on the penalization magnitude. In this chapter we also justified the importance of scaling the error depending on the choice of the probability measure distance. In the end, we proved several bounds on the second-order moment for log-concave distributions in two particular cases. First, the distribution was assumed to be strongly log-concave inside some Euclidean ball, then it was assumed to be so outside some Euclidean ball.

In the third contribution, we study again the general log-concave case. This time we focused on the continuous-time sampling schemes. In particular, we proposed a linear penalization to the drift term in the Langevin diffusion, and we called the resulting SDE "Penalized Langevin Dynamics" (PLD). We proved a bound on the mixing-time of PLD in W_2 distance with explicit dependence on the penalization term. We then showed that optimizing over the penalization yields polynomial convergence. Similarly, we introduced Penalized Kinetic Langevin Dynamics (PKLD), by performing the same trick on KLD. Polynomial convergence rates are proved also for PKLD. In the end, we extend our results to the problem of convex optimization, exploiting its connection to the sampling problem.

5.2 Perspectives

This work has left several open questions and it can be extended. Here is a non-exhaustive list of possible future work.

- In Chapter 4 we have studied the continuous time sampling scheme. However, it is not implementable on a computer; thus, a discrete method is required. As mentioned previously, there are several discretization methods for Langevin-type processes. Of particular interest in this case is the mid-point discretization scheme, proposed by [SL19]. The latter method is proved to have a better dependence on the condition number κ, when compared to the classical Euler-Maruyama discretization scheme (see [DRD20] and [CCBJ18]). Therefore, applying the mid-point method to the PLD may yield to an iterative algorithm with tractable convergence bounds.
- As we have seen, the LMC is a biased sampling method. It does not converge to the target distribution π. Metropolis Adjusted Langevin Algorithm (MALA) solves this issue by correcting each iteration using a rejection-acceptance procedure (see [CDWY20]). For the kinetic method, however, this method is not well studied. It would be interesting to study how the Metropolis-Hastings correction step can be adapted to the case of kinetic Langevin algorithm.
- In [DMM19], the LMC is analyzed from the perspective of convex optimization on the space $(\mathcal{P}_2(\mathbb{R}^p), W_2)$. It is based on the seminal paper by [JKO98], which essentially interprets the Fokker-Planck equation as a gradient flux of a functional. A clever application of this theory by [DMM19] has given the best known convergence for the standard LMC. Despite the efficiency of the technique, it has not been used yet in other settings of the sampling problem. An extension of this result for the mid-point discretization method or kinetic Langevin algorithm could potentially improve the state-of-the-art.

Bibliography

[AFEB16] P. Alquier, N. Friel, R. Everitt, and A. Boland. Noisy Monte-Carlo: convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26(1):29–47, Jan 2016.

[AFMP11] Y. Atchadé, G. Fort, E. Moulines, and P. Priouret. Adaptive Markov chain Monte-Carlo: theory and methods. In *Bayesian time series models*, pages 32–51. Cambridge Univ. Press, Cambridge, 2011.

[AGB15] David Alonso-Gutiérrez and Jesús Bastero. *Approaching the Kannan-Lovász-Simonovits and variance conjectures*, volume 2131 of *Lecture Notes in Mathematics*. Springer, Cham, 2015.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[AH12] P. Alquier and M. Hebiri. Transductive versions of the LASSO and the Dantzig Selector. *J. Statist. Plann. Inference,* 142(9):2485–2500, 2012.

[AJKH14] A. Alfonsi, B. Jourdain, and A. Kohatsu-Higa. Pathwise optimal transport bounds between a one-dimensional diffusion and its Euler scheme. *Ann. Appl. Probab.*, 24(3):1049–1080, 06 2014.

[AJKH15] Aurelien Alfonsi, Benjamin Jourdain, and Arturo Kohatsu-Higa. Optimal transport bounds between the time-marginals of a multidimensional diffusion and its Euler scheme. *Electron. J. Probab.*, 20:31 pp., 2015.

[AKSG19] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems 32*, pages 6484–6494. Curran Associates, Inc., 2019.

[ARW16] C. Andrieu, J. Ridgway, and N. Whiteley. Sampling normalizing constants in high dimensions using inhomogeneous diffusions. *ArXiv e-prints*, December 2016.

[BB11] Léon Bottou and Olivier Bousquet. 13 the tradeoffs of large-scale learning. *Optimization for machine learning*, page 351, 2011.

[BBC+08] Dominique Bakry, Franck Barthe, Patrick Cattiaux, Arnaud Guillin, et al. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.

[BBN+03] Keith Ball, Franck Barthe, Assaf Naor, et al. Entropy jumps in the presence of a spectral gap. *Duke Mathematical Journal*, 119(1):41–63, 2003.

[BC+09] Jonathan M Borwein, O Chan, et al. Uniform bounds for the complementary incomplete gamma function. *Mathematical Inequalities and Applications*, 12:115–121, 2009.

[BDM18] N. Brosse, A. Durmus, and É. Moulines. Normalizing constants of log-concave densities. *Electronic journal of statistics*, 12(1):851–889, 2018.

[BDMP17] N. Brosse, A. Durmus, É. Moulines, and M. Pereyra. Sampling from a log-concave distribution with compact support with proximal Langevin Monte-Carlo. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 319–342, 07–10 Jul 2017.

[BEL18] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected Langevin Monte-Carlo. *Discrete & Computational Geometry*, 59(4):757–783, Jun 2018.

[Ben16] Dobriyan M Benov. The manhattan project, the first electronic computer and the Monte-Carlo method. *Monte Carlo Methods and Applications*, 22(1):73–79, 2016.

[Ber18] Espen Bernton. Langevin Monte-Carlo and JKO splitting. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1777–1798. PMLR, 2018.

[BFFN19] Jack Baker, Paul Fearnhead, Emily B. Fox, and Christopher Nemeth. Control variates for stochastic gradient mcmc. *Statistics and Computing*, 29(3):599–615, May 2019.

[BGG12] François Bolley, Ivan Gentil, and Arnaud Guillin. Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations. *J. Funct. Anal.*, 263(8):2430–2457, 2012.

[Bha78] R. N. Bhattacharya. Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *Ann. Probab.*, 6(4):541–553, 08 1978.

[BL76] Herm Jan Brascamp and Elliott H Lieb. On extensions of the brunn-minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366 – 389, 1976.

[BM13] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Neural Information Processing Systems (NIPS)*, 2013.

[BMP+94] James O Berger, Elías Moreno, Luis Raul Pericchi, M Jesús Bayarri, José M Bernardo, Juan A Cano, Julián De la Horra, Jacinto Martín, David Ríos-Insúa, Bruno Betrò, et al. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

[Bob99] S. G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, 27(4):1903–1921, 1999.

[Bor83] Christer Borell. Convexity of measures in certain convex cones in vector space σ-algebras. *Mathematica Scandinavica*, 53(1):125–144, 1983.

[Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[Bro80] L Brown. Examples of berger's phenomenon in the estimation of independent normal means. *The Annals of Statistics*, pages 572–585, 1980.

[Bro97] Roger Brockett. Oscillatory descent for function minimization. In *Current and future directions in applied mathematics*, pages 65–82. Springer, 1997.

[BRT+09] Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of LASSO and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

[BTW07] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[C+47] Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[CB18] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of ALT2018*, 2018.

[CCA+18] Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *CoRR*, abs/1805.01648, 2018.

[CCBJ18] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323. PMLR, 2018.

[CD98] Ming-Hui Chen and Dipak K Dey. Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 322–343, 1998.

[CDC15] Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pages 2278–2286, 2015.

[CDJB19] Niladri S Chatterji, Jelena Diakonikolas, Michael I Jordan, and Peter L Bartlett. Langevin Monte-Carlo without smoothness. *arXiv preprint arXiv:1905.13285*, 2019.

[CDWY18] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mcmc sampling algorithms on polytopes. *The Journal of Machine Learning Research*, 19(1):2146–2231, 2018.

[CDWY20] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte-Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 2020.

[CFM+18] Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, and Michael Jordan. On the theory of variance reduction for stochastic gradient Monte-Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR, 2018.

[CG09] Patrick Cattiaux and Arnaud Guillin. Trends to equilibrium in total variation distance. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):117–145, 2009.

[CG18] Patrick Cattiaux and Arnaud Guillin. On the Poincaré constant of log-concave measures. *arXiv preprint arXiv:1810.08369*, 2018.

[CGR10] Patrick Cattiaux, Arnaud Guillin, and Cyril Roberto. Poincaré inequality and the L^p convergence of semi-groups. *Electron. Commun. Probab.*, 15:270–280, 2010.

[Cha77] Gregory J Chaitin. Algorithmic information theory. *IBM journal of research and development*, 21(4):350–359, 1977.

[Cha82] Gregory J Chaitin. Gödel's theorem and information. *International Journal of Theoretical Physics*, 21(12):941–954, 1982.

[Cha88] Gregory J Chaitin. Randomness in arithmetic. *Scientific American*, 259(1):80–85, 1988.

[Che20] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *arXiv preprint arXiv:2011.13661*, 2020.

[CLGL+20] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. Exponential ergodicity of mirror-langevin diffusions. *Advances in Neural Information Processing Systems*, 33, 2020.

[CMMR12] Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, and Christian P Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

[CN91] Andrew Caplin and Barry Nalebuff. Aggregation and imperfect competition: On the existence of equilibrium. *Econometrica: Journal of the Econometric Society*, pages 25–59, 1991.

[CR98] George Casella and Christian P Robert. Post-processing accept-reject samples: recycling and rescaling. *Journal of Computational and Graphical Statistics*, 7(2):139–157, 1998.

[Cro72] Lawrence Crone. The singular value decomposition of matrices and cheap numerical filtering of systems of linear equations. *Journal of the Franklin Institute*, 294(2):133–136, 1972.

[CT07] Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[CV19] Zongchen Chen and Santosh S. Vempala. Optimal convergence rate of Hamiltonian Monte-Carlo for strongly logconcave distributions. *CoRR*, abs/1905.02313, 2019.

[CZ13] E.K.P. Chong and S.H. Zak. *An Introduction to Optimization*. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2013.

[Dal17a] Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte-Carlo and gradient descent. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689, 07–10 Jul 2017.

[Dal17b] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. *J. R. Stat. Soc. B*, 79:651 – 676, 2017.

[Dan61] HE Daniels. The asymptotic efficiency of a maximum likelihood estimator. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 151–163. University of California Press Berkeley, 1961.

[DBLJ14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *arXiv preprint arXiv:1407.0202*, 2014.

[DCWY18] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference on Learning Theory*, pages 793–797. PMLR, 2018.

[DDB+20] Aymeric Dieuleveut, Alain Durmus, Francis Bach, et al. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Annals of Statistics*, 48(3):1348–1382, 2020.

[Dev86] Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265, 1986.

[Dev06] Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

[DFB17] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

[DHL17] Arnak S Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the LASSO. *Bernoulli*, 23(1):552–581, 2017.

[DJD88] Sudhakar Dharmadhikari and Kumar Joag-Dev. *Unimodality, convexity, and applications*. Elsevier, 1988.

[DJHS92] David L Donoho, Iain M Johnstone, Jeffrey C Hoch, and Alan S Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):41–67, 1992.

[DK19] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte-Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

[DM17] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 06 2017.

[DM19] Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 11 2019.

[DMM19] Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of Langevin Monte-Carlo via convex optimization. *J. Mach. Learn. Res.*, 20:73–1, 2019.

[DMP18] Alain Durmus, Éric Moulines, and Marcelo Pereyra. Efficient Bayesian Computation by Proximal Markov Chain Monte-Carlo: When Langevin Meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1), 2018.

[DMR04] R. Douc, E. Moulines, and Jeffrey S. Rosenthal. Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Ann. Appl. Probab.*, 14(4):1643–1665, 2004.

[Doo42] Joseph L Doob. The brownian movement and stochastic equations. *Annals of Mathematics*, pages 351–369, 1942.

[DRD20] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

[DRDK19] Arnak S. Dalalyan, Lionel Riou-Durand, and Avetik Karagulyan. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. 2019.

[DT09] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, June 18-21, 2009*, pages 1–10, 2009.

[EGZ19] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Ann. Probab.*, 47(4):1982–2010, 07 2019.

[EMS18] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems 31*, pages 9671–9680. 2018.

[FK85] Ludwig Fahrmeir and Heinz Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, pages 342–368, 1985.

[FRV18] Guilherme Franca, Daniel Robinson, and Rene Vidal. ADMM and accelerated ADMM as continuous dynamical systems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1559–1567. PMLR, 10–15 Jul 2018.

[GBVV14] Apostolos Giannopoulos, Silouanos Brazitikos, Petros Valettas, and Beatrice-Helen Vritsiou. *Geometry of Isotropic Convex Bodies*. 05 2014.

[GC11] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte-Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(2):123–214, 2011.

[Gew89] John Geweke. Bayesian inference in econometric models using Monte-Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

[Gew91] John Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing science and statistics: Proceedings of the 23rd symposium on the interface*, volume 571, page 578. Citeseer, 1991.

[GL10] N. Gozlan and C. Léonard. Transport inequalities. A survey. *Markov Process. Related Fields*, 16(4):635–736, 2010.

[GM94] Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

[GP20] Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the ruppert-polyak averaging stochastic algorithm. 2020.

[GPP20] Sébastien Gadat, Fabien Panloup, and Clément Pellegrini. On the cost of Bayesian posterior mean strategy for log-concave models. *arXiv preprint arXiv:2010.06420*, 2020.

[GR93] A Gelman and DB Rubin. Discussion on gibbs sampler and other mcmc methods. *Journal of the Royal Statistical Society*, pages 73–73, 1993.

[Gri93] A. Griewank. Some bounds on the complexity of gradients, Jacobians, and Hessians. In P.M. Pardalos, editor, *Complexity in Nonlinear Optimization*, pages 128–161. World Scientific publishers, 1993.

[GS91] Alan E Gelfand and Adrian FM Smith. Gibbs sampling for marginal posterior expectations. *Communications in Statistics-Theory and Methods*, 20(5-6):1747–1766, 1991.

[GS94] Alan E Gelfand and Sujit K Sahu. On Markov chain Monte-Carlo acceleration. *Journal of Computational and Graphical Statistics*, 3(3):261–276, 1994.

[GW92] Walter R Gilks and Pascal Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

[HA01] H Brian Hwarng and HT Ang. A simple neural network for arma (p, q) time series. *Omega*, 29(4):319–333, 2001.

[Har04] Gilles Hargé. A convex/log-concave correlation inequality for gaussian measure and an application to abstract Wiener spaces. *Probability theory and related fields*, 130(3):415–440, 2004.

[Has70] W Keith Hastings. Monte-Carlo sampling methods using Markov chains and their applications. 1970.

[Hes95] Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

[HH06] C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.*, 1(1):145–168, 2006.

[HJLS13] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[HK70] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[HKRC18] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin dynamics. *Advances In Neural Information Processing Systems 31 (Nips 2018)*, 31, 2018.

[HM54] John M Hammersley and K William Morton. Poor man's Monte-Carlo. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(1):23–38, 1954.

[Hor62] AE Horel. Applications of ridge analysis toregression problems. *Chem. Eng. Progress.*, 58:54–59, 1962.

[HSR19] Liam Hodgkinson, Robert Salomone, and Fred Roosta. Implicit Langevin algorithms for sampling from log-concave densities. *arXiv preprint arXiv:1903.12322*, 2019.

[HZ17] Jonathan Huggins and James Zou. Quantifying the accuracy of approximate diffusions and Markov chains. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 382–391, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[IW14] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014.

[JH00] S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361, 2000.

[JK17] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336, 2017.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[KBB15] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2845–2853. Curran Associates, Inc., 2015.

[KD20] Avetik Karagulyan and Arnak Dalalyan. Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets. *Advances in Neural Information Processing Systems*, 33, 2020.

[Khi23] A. Khintchine. Über dyadische Brüche. *Math. Z.*, 18(1):109–116, 1923.

[Kib03] BM Golam Kibria. Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation*, 32(2):419–435, 2003.

[KLS95] R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete Comput. Geom.*, 13(3-4):541–559, 1995.

[Kol33] Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

[Kol09] Vladimir Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

[Kra40] Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

[Lan08] Paul Langevin. Sur la théorie du mouvement brownien. *Compt. Rendus*, 146:530–533, 1908.

[LC70] Lucien Le Cam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3):802–828, 1970.

[LC90] Lucien Le Cam. Maximum likelihood: an introduction. *International Statistical Review/Revue Internationale de Statistique*, pages 153–171, 1990.

[LC06] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[LD75] Erich Leo Lehmann and Howard J D'Abrera. *Nonparametrics: statistical methods based on ranks.* Holden-day, 1975.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

[LFC17] Tung Duy Luu, Jalal Fadili, and Christophe Chesneau. Sampling from non-smooth distribution through Langevin diffusion. working paper or preprint, August 2017.

[Liu96] Jun S Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and computing*, 6(2):113–119, 1996.

[LMW19] Mingjie Liang, Mateusz B. Majka, and Jian Wang. Exponential ergodicity for SDEs and mckean-vlasov processes with lévy noise, 2019.

[Loj82] Stanislaw Lojasiewicz. Sur les trajectoires du gradient d'une fonction analytique. *Seminari di geometria*, 1983:115–117, 1982.

[LV06a] L. Lovasz and S. Vempala. Fast algorithms for logconcave functions: Sampling, Rounding, Integration and Optimization. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 57–68, 2006.

[LV06b] L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM J. Comput.*, 35(4):985–1005 (electronic), 2006.

[Mar54] Andrew W Marshall. The use of multi-stage sampling schemes in Monte-Carlo computations. Technical report, RAND CORP SANTA MONICA CALIF, 1954.

[MCC+19] Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I Jordan. Is there an analog of Nesterov Acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

[MCJ+19] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

[Met87] N Metropolis. The beginning. *Los Alamos Science*, 15:125–130, 1987.

[MFWB19] Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. An efficient sampling algorithm for non-smooth composite potentials. *CoRR*, abs/1910.00551, 2019.

[MJ51] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[MM16] Stéphane Mischler and Clément Mouhot. Exponential stability of slowly decaying solutions to the kinetic-Fokker-Planck equation. *Archive for rational mechanics and analysis*, 221(2):677–723, 2016.

[MMS18] Mateusz B. Majka, Aleksandar Mijatović, and Lukasz Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity, 2018.

[Mon20] Pierre Monmarché. High-dimensional mcmc with a standard splitting scheme for the underdamped Langevin diffusion, 2020.

[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.

[MS17] Oren Mangoubi and Aaron Smith. Rapid mixing of Hamiltonian Monte-Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

[MS19] Oren Mangoubi and Aaron Smith. Mixing of Hamiltonian Monte-Carlo on strongly log-concave distributions 2: Numerical integrators. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 586–595, 2019.

[MSH02] Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

[MT93] Sean P Meyn and Richard L Tweedie. Stability of Markovian processes ii: Continuous-time processes and sampled chains. *Advances in Applied Probability*, pages 487–517, 1993.

[MT$^+$96] Kerrie L Mengersen, Richard L Tweedie, et al. Rates of convergence of the hastings and metropolis algorithms. *Annals of Statistics*, 24(1):101–121, 1996.

[MT12] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[MU49] Nicholas Metropolis and Stanislaw Ulam. The Monte-Carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

[MV19a] Oren Mangoubi and Nisheeth K Vishnoi. Nonconvex sampling with the metropolis-adjusted Langevin algorithm. pages 2259–2293, 2019.

[MV19b] Oren Mangoubi and Nisheeth K. Vishnoi. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2259–2293. PMLR, 2019.

[MW00] Ian W McKeague and Wolfgang Wefelmeyer. Markov chain Monte-Carlo and Rao–Blackwellization. *Journal of statistical planning and inference*, 85(1-2):171–182, 2000.

[NDH$^+$17] T. Nagapetyan, A. B. Duncan, L. Hasenclever, S. J. Vollmer, L. Szpruch, and K. Zygalakis. The True Cost of Stochastic Gradient Langevin Dynamics. *ArXiv e-prints*, June 2017.

[Nel67] Edward Nelson. *Dynamical Theories of Brownian Motion*. Department of Mathematics. Princeton University, 1967.

[Nes04] Yu. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

[Nor82] Lennart Nordberg. A procedure for determination of a good ridge parameter in linear regression. *Communications in Statistics-Simulation and Computation*, 11(3):285–309, 1982.

[NP00] Pierpaolo Natalini and Biagio Palumbo. Inequalities for the incomplete gamma function. *Math. Inequal. Appl*, 3(1):69–77, 2000.

[OZ00] Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

[Par81] Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

[Pav14] Grigorios A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.

[PB14] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[Per14] M. Pereyra. Proximal Markov chain Monte-Carlo algorithms. Technical report, arXiv:1306.0187, 2014.

[Pré73] András Prékopa. On logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum*, 34:335–343, 1973.

[PVBL$^+$19] Loucas Pillaud-Vivien, Francis Bach, Tony Lelièvre, Alessandro Rudi, and Gabriel Stoltz. Statistical estimation of the Poincaré constant and application to sampling multimodal distributions, 2019.

[QL$^+$12] Feng Qi, Qiu-Ming Luo, et al. Bounds for the ratio of two gamma functions—from Wendel's and related inequalities to logarithmically completely monotonic functions. *Banach Journal of Mathematical Analysis*, 6(2):132–158, 2012.

[Rao] Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2.

[RC13] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[Rip09] Brian D Ripley. *Stochastic simulation*, volume 316. John Wiley & Sons, 2009.

[RK16] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte-Carlo method*, volume 10. John Wiley & Sons, 2016.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[Rob96] Christian P Robert. Intrinsic losses. *Theory and decision*, 40(2):191–214, 1996.

[Roc70] R Tyrrell Rockafellar. *Convex analysis*, volume 36. Princeton university press, 1970.

[RR55] Marshall N Rosenbluth and Arianna W Rosenbluth. Monte-Carlo calculation of the average extension of molecular chains. *The Journal of Chemical Physics*, 23(2):356–359, 1955.

[RR98] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268, 1998.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, 07–10 Jul 2017.

[RS02] G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, 4(4):337–357 (2003), 2002.

[RT96a] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[RT96b] Gareth O Roberts and Richard L Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[San17] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[SFCM13] A. Schreck, G. Fort, S. Le Corff, and E. Moulines. A shrinkage-thresholding metropolis adjusted Langevin algorithm for Bayesian variable selection. *preprint*, arXiv:1312.5658, 2013.

[SKR19] Adil Salim, Dmitry Koralev, and Peter Richtarik. Stochastic proximal Langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems 32*, pages 6653–6664. 2019.

[SL19] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2098–2109, 2019.

[SRBd17] Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d'Aspremont. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems 30*, pages 1109–1118. Curran Associates, Inc., 2017.

[ST+99a] Sarjinder Singh, Derrick Shannon Tracy, et al. Ridge-regression using scrambled responses. *Metrika*, 41(2):147–157, 1999.

[ST99b] O. Stramer and R. L. Tweedie. Langevin-type models. I. Diffusions with given stationary distributions and their discretizations. *Methodol. Comput. Appl. Probab.*, 1(3):283–306, 1999.

[ST99c] O. Stramer and R. L. Tweedie. Langevin-type models. II. Self-targeting candidates for MCMC algorithms. *Methodol. Comput. Appl. Probab.*, 1(3):307–328, 1999.

[SW14] A. Saumard and J. A. Wellner. Log-concavity and strong log-concavity: a review. *Stat. Surv.*, 8:45–114, 2014.

[SZTG20] Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gürbüzbalaban. Fractional underdamped Langevin dynamics: Retargeting SGD with momentum under heavy-tailed gradient noise. *CoRR*, abs/2002.05685, 2020.

[Tal02] Denis Talay. Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme. *Markov Process. Related Fields*, 8(2):163–198, 2002.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[Tie94] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[TSV18] Yin Tat Lee, Zhao Song, and Santosh S. Vempala. Algorithmic Theory of ODEs and Sampling from Well-conditioned Logconcave Densities. *arXiv e-prints*, page arXiv:1812.06243, Dec 2018.

[Tsy08] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[TTV16] Yee Whye Teh, Alexandre H. Thiery, and Sebastian J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(7):1–33, 2016.

[Tuk56] HF Tukey. Conditional Monte-Carlo for normal samples. In *Proc. Symp. on Monte-Carlo Methods*, pages 64–79. John Wiley and Sons, 1956.

[Ula52] Stanislaw Ulam. Random processes and transformations. In *Proceedings of the International Congress on Mathematics*, volume 2, pages 264–275. Citeseer, 1952.

[vdG07] Sara van de Geer. The deterministic LASSO. In *Proc. of Joint Statistical Meeting*, 2007.

[vdGL13] Sara van de Geer and Johannes Lederer. The Lasso, correlated design, and improved oracle inequalities. *IMS Collections*, 9:303–316, 2013.

[VDK83] Herman K Van Dijk and Teunis Kloek. Experiments with some alternatives for simple importance sampling in Monte-Carlo integration. Technical report, 1983.

[Vem05] Santosh Vempala. Geometric random walks: a survey. *Combinatorial and computational geometry*, 52(573-612):2, 2005.

[Vil08] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[Vil09] Cedric Villani. *Hypocoercivity*. Number 949-951. American Mathematical Soc., 2009.

[VN51] John Von Neumann. 13. various techniques used in connection with random digits. *Appl. Math Ser*, 12(36-38):3, 1951.

[VNR46] John Von Neumann and Robert D Richtmyer. Statistical methods in neutron diffusion. pages 17–36. University of California Press, 1946.

[VW19] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32*, pages 8094–8106. 2019.

[VZ15] S. J. Vollmer and K. C. Zygalakis. (Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics. *ArXiv e-prints*, January 2015.

[Wen00] Eshetu Wencheko. Estimation of the signal-to-noise in the linear regression model. *Statistical Papers*, 41(3):327, 2000.

[Wib18] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 06–09 Jul 2018.

[Wri95] Raymond E Wright. Logistic regression. 1995.

[WRJ16] Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A lyapunov analysis of momentum methods in optimization, 2016.

[WT11] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 681–688, 2011.

[WWJ16] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[XCZG18] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, pages 3126–3137, 2018.

[XSL+14] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statist. Probab. Lett.*, 91:14–19, 2014.

[ZMSJ18] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3900–3909. 2018.

[ZXG19] Difan Zou, Pan Xu, and Quanquan Gu. Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics. In *AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 2936–2945, 2019.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre: L'échantillonnage avec Langevin Monté-Carlo

Mots clés: échantillonbage, chaines de Markov, statistiques, Langevin Monte-Carlo, optimisation, probabilités

Résumé: L'échantillonnage des lois aléatoire est un problème de grande importance en statistiques et en machine learning. Les approches générales sur ce sujet sont souvent divisées en deux catégories: fréquentiste vs bayésienne. L'approche fréquentiste correspond à la minimisation du risque empirique, c'est-à-dire à l'estimation du maximum de la vraisemblance qui est un problème d'optimisation, tandis que l'approche bayésienne revient à intégrer la loi postérieure. Cette dernière approche nécessite souvent des méthodes approximatives car l'intégrale n'est généralement pas tractable. Dans ce manuscrit, nous allons étudier la méthode de Langevin, basée sur la discrétisation de l'EDS de Langevin. La première partie de l'introduction pose le cadre mathématique et l'intérêt d'étudier la question de l'échantillonnage. La suite de l'introduction s'attache à la présentation des méthodes d'échantillonnage.

Le premier article concerne les bornes non-asymptotiques sur la convergence en distance de Wasserstein de Langevin Monte-Carlo pour les fonctions potentielles régulières et fortement convexes. Nous établissons d'abord des bornes explicites pour LMC avec des step-sizes variantes. Puis nous étudions la convergence pour des fonctions potentielles avec des gradients stochastiques. Enfin, deux types de discrétisation sont présentés, pour les fonctions potentielles plus régulières.

Dans le deuxième article nous abordons le problème d'échantillonnage de loi log-concave (pas fortement) en utilisant LMC, KLMC et KLMC2. Nous proposons une pénalisation quadratique constante de la fonction potentielle. Puis nous prouvons des bornes non-asymptotiques sur l'erreur de Wasserstein de ces méthodes pour le choix de pénalisation optimale. Enfin, nous soulignons l'importance du choix de l'échelle pour le mesurage des complexités des différentes méthodes.

La troisième contribution principale est concentrée sur la convergence de la diffusion de Langevin dans le cas log-concave. Une pénalisation variable dans le temps est proposée pour la fonction de potentiel. Nous prouvons des bornes explicites pour cette méthode nommée Penalized Langevin Dynamics. À la fin, le lien entre les algorithmes de Langevin et l'optimisation convexe est établi, ce qui nous permet de prouver des bornes similaires pour le gradient flow.

Title: Sampling with the Langevin Monte-Carlo

Keywords: sampling, Langevin Monte-Carlo, Markov chains, optimization, probability, statistics

Abstract: Sampling from probability distributions is a problem of significant importance in Statistics and Machine Learning. The approaches for the latter can be roughly classified into two main categories, that is the frequentist and the Bayesian. The first is the MLE or ERM which boils down to optimization, while the other requires the integration of the posterior distribution. Approximate sampling methods are hence applied to estimate the integral. In this manuscript, we focus mainly on Langevin sampling which is based on discretizations of Langevin SDEs. The first half of the introductory part presents the general mathematical framework of statistics and optimization, while the rest aims at the historical background and mathematical development of sampling algorithms.

The first main contribution provides non-asymptotic bounds on convergence LMC in Wasserstein error. We first prove the bounds for LMC with the time-varying step. Then we establish bounds in the case when the gradient is available with a noise. In the end, we study the convergence of two versions of discretization, when the Hessian of the potential is regular.

In the second main contribution, we study the sampling from log-concave (non-strongly) distributions using LMC, KLMC, and KLMC with higher-order discretization. We propose a constant square penalty for the potential function. We then prove non-asymptotic bounds in Wasserstein distances and provide the optimal choice of the penalization parameter. In the end, we highlight the importance of scaling the error for different error measures.

The third main contribution focuses on the convergence properties of convex Langevin diffusions. We propose to penalize the drift with a linear term that vanishes over time. Explicit bounds on the convergence error in Wasserstein distance are proposed for the Penalized Langevin Dynamics and Penalized Kinetic Langevin Dynamics. Also, similar bounds are proved for the Gradient Flow of convex functions.

Institut Polytechnique de Paris

91120 Palaiseau, France