Theoretical contributions to Monte Carlo methods, and applications to Statistics

Lionel Riou-Durand

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

ENSAE ParisTech — École nationale de la statistique et de l'administration économique — université PARIS-SACLAY

NNT : 2019SACLG006

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École Doctorale de Mathématiques Hadamard (EDMH), ED 574

Établissement d'inscription : École Nationale de la Statistique et de l'Administration Économique (ENSAE)

Laboratoire d'accueil : Center for Research in Economics and Statistics (CREST), UMR 9194 CNRS

Spécialité de doctorat : Mathématiques appliquées

Lionel Riou-Durand

Contributions théoriques aux méthodes de Monte Carlo, et applications à la statistique

Date de soutenance : 5 juillet 2019

Après avis des rapporteurs :
Arnaud Guillin (Université Clermont-II)
Pierre Jacob (Harvard University)

Jury de soutenance :

Stéphanie Allassonnière	(Univ. Paris-V) Examinateur
Nicolas Chopin	(ENSAE) Directeur de thèse
Arnak Dalalyan	(ENSAE) Examinateur
Arnaud Guillin	(Univ. Clermont-II) Rapporteur
Pierre Jacob	(Harvard University) Rapporteur
Christian Robert	(Univ. PSL) Président du jury

Fondation mathématique FMJH Jacques Hadamard

CREST CENTER FOR RESEARCH IN ECONOMICS AND STATISTICS

Remerciements

Je remercie premièrement les chercheurs qui m'ont encadré durant ma thèse. Je pense bien sûr à Nicolas Chopin et Arnak Dalalyan, envers lesquels je suis infiniment reconnaissant, aussi bien pour leurs conseils que pour leur aide et leur soutien.

Je remercie ensuite mes collègues de travail, pour tous les moments que nous avons partagés, d'abord à Malakoff, puis sur le campus du plateau de Saclay. En voici une liste non exhautive désordonnée, merci à : Gautier, Simo, Léna, Pierre, Badr-Eddine, Geoffrey, Lucie, Boris, Solenne, Jérémy, Christophe, Alexis, Yannick, Alexander, The Tien, Vincent, Mehdi, Avo, Amir, Arya, Jules, Morgane, Jérôme, Wasfé, Edith, Pascale, Philip, Martin, Guillaume, Cristina, Marco, Sacha, Victor-Emmanuel.

Cette thèse marque la fin de mes études, qui se seraient sans doute arrêtées bien plus tôt sans l'aide que j'ai reçue à l'Université de Rennes 1. Je tiens à remercier ces enseignants qui ont beaucoup compté à mes yeux. Je parle de ceux qui m'ont redonné confiance, et qui m'ont transmis le goût des études et des mathématiques, merci à : Chantal Garcia, Chantal Guéguen, Sophie Larribeau, Nathalie Colombier, Isabelle Cadoret, Pascal Bouyaux, Jean-Christophe Breton, Bernard Delyon.

Enfin, un grand merci à mes proches pour leur soutien inconditionnel, je leur dédie ce manuscrit en contrepartie.

Contents

Chapter 1

An introduction to Monte Carlo methods

This chapter begins with a presentation of Monte Carlo methods, with a particular emphasis on their origin, their relevance in numerical integration and optimization, and some of their applications to Statistics. Section 1.1 introduces those notions as a motivation for a more general purpose, which is the sampling from a given probability distribution. Section 1.2 presents some applications of sampling-based methods to statistical inference for partially known models, in particular intractable likelihoods. Section 1.3 discusses the importance of obtaining explicit non-asymptotic guaranties for sampling-based procedures, especially for high dimensional probability distributions. Section 1.4 summarizes the contributions of this thesis.

1.1 Monte Carlo methods and applications

From a computational perspective, Monte Carlo methods are algorithms based on repeated random sampling, performed to approximate unknown numerical values. Monte Carlo methods, or at least those related to Markov Chains and their properties, were invented during the Second World War. They were developed by Jon von Neumann, Stanislaw Ulam and Nicholas Metropolis. The latter suggested the name *Monte Carlo*, referring to the Monte Carlo Casino in Monaco, as a code name for their works within nuclear weapons projects in Los Alamos. Monte Carlo methods are mainly used to solve numerical integration, or optimization problems where the other methods require prohibitively large computational resources. In Statistics nowadays, with the rise of computational techniques and power, Monte Carlo methods have become widely used, especially for Bayesian inference, model selection, and testing. Those statistical procedures require generating random samples from a given probability distribution, which is often non-trivial especially for high dimensional problems. We often encounter the word *sampling*, as a designation for generating those random samples from a computer, which is completely different from the usual sampling terminology that refers to selecting a random subset from a statistical population. From a mathematical perspective, sampling can be viewed as the art of transforming a uniform distribution into a given probability

distribution.

Throughout this thesis, we will assume implicitly that we have access to a random number generator, designed to perform exact uniform sampling between zero and one. One could argue that computers are not able to generate perfect randomness. Sad but true, a pseudo-random uniform sequence (u_j) produced by a computer is usually periodic, and there is a deterministic map D such that $u_{j+1} = D(u_j)$. Another could reply that existence of perfect randomness is highly questionable, and that randomness is mainly an attractive way of modelling the unknown. Both statements may appear rather convincing.

In any case, the scientific community agrees on saying that pseudo-random uniform sequences generated by computers are now so carefully designed that they share almost the same properties as perfect randomness. The most suspicious readers should remark that the reliability of some widely used pseudo-random generators is now such that periodicity should not pose a problem until the $(2^{19927} - 1)^{th}$ generated number[1]. On any computer nowadays, such an issue should not arise before a few billions of years. Therefore, we assume in the sequel that we are able to generate random variables identically and independently distributed (IID) with respect to the uniform distribution between zero and one.

Let us define a random vector uniformly distributed over the k-dimensional hypercube

$$(U_1, ..., U_k) \sim \mathcal{U}_{[0,1]^k}.$$

Let us define the measurable space $\boldsymbol{\Theta} \triangleq \mathbb{R}^p$ equipped with its Borel σ-field, that we will denote by $\mathscr{B}(\mathbb{R}^p)$. Our goal is now to sample from a given probability distribution Π defined over $\boldsymbol{\Theta}$.

Why such a goal? This is a legitimate question. From a general perspective, sampling methods are a standard basis for any Monte Carlo estimate. In Section 1.1.1 we give a brief introduction to Monte Carlo estimation, in the context of numerical integration. In Sections 1.1.2 to 1.1.4 we present some applications of Monte Carlo methods to statistics and machine learning. Sampling methods in statistics find most of their applications in the Bayesian framework. We give therefore a short introduction to the Bayesian paradigm, and emphasize the relevance of Bayesian estimators regardless of everyone's convictions from a decision theory perspective. We also show that Monte Carlo methods are not restricted to Bayesian inference, and present some applications in other frameworks such as maximum likelihood approximation and PAC-Bayesian estimation.

1.1.1 Numerical integration

We assume in this section that we are able to sample from a given probability distribution Π, defined over $\boldsymbol{\Theta} = \mathbb{R}^p$. We turn to the problem of approximating an expectation with respect to Π. For a given map $\varphi : \boldsymbol{\Theta} \mapsto \mathbb{R}$, we use the following notation for the expectation of φ with respect to Π if it exists:

$$\Pi(\varphi) = \int_{\boldsymbol{\Theta}} \varphi(\boldsymbol{\theta}) \Pi(d\boldsymbol{\theta}).$$

[1] Based on the periodicity of the Mersenne Twister pseudo-random number generator.

In such a setting, Monte Carlo estimation is a natural solution to approximating $\Pi(\varphi)$. Let $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m$ be m random variables identically and independently distributed from Π. Then for any measurable map $\varphi : \boldsymbol{\Theta} \mapsto \mathbb{R}$ such that $\Pi(|\varphi|) < +\infty$, the law of large numbers yields

$$\widehat{\Pi}_m(\varphi) \triangleq \frac{1}{m} \sum_{j=1}^{m} \varphi(\boldsymbol{\theta}_j) \overset{\text{a.s.}}{\to} \Pi(\varphi)$$

as m goes to infinity. In other words, the approximation $\widehat{\Pi}_m(\varphi)$ is a consistent estimator of $\Pi(\varphi)$. Essentially, the sample $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m)$ is to be seen as a random generation from a m-times repeated sampling procedure executed by a computer. The estimator $\widehat{\Pi}_m(\varphi)$ being a measurable function of this randomly generated sample, it is often referred to a Monte Carlo estimator, or as a Monte Carlo approximation of $\Pi(\varphi)$. If, in addition, the map φ is such that $\Pi(\varphi^2) < +\infty$, then the central limit theorem yields

$$\sqrt{m} \left(\widehat{\Pi}_m(\varphi) - \Pi(\varphi) \right) \overset{\mathscr{D}}{\to} \mathcal{N}(0, \Pi(\varphi^2) - \Pi(\varphi)^2)$$

as m goes to infinity. For such a function φ, this shows that the Monte Carlo approximation $\widehat{\Pi}_m(\varphi)$ will converge to $\Pi(\varphi)$ at the rate $\sqrt{m}$. Therefore, getting an estimation ten times more accurate would require sampling a hundred times more. It is not wrong to claim that such a speed is slow, thus the problem of controlling the Monte Carlo estimation error will often be limited, or at least coupled to controlling the computational complexity of such a task. Of course, Monte Carlo estimators are only useful when the distribution Π is partially known, in the sense that the expectation $\Pi(\varphi)$ does not admit a closed form for some function φ. In such a setting however, and when the dimension p is larger than one, Monte Carlo estimators are among the best known methods for estimating $\Pi(\varphi)$.

An important point to emphasize is the following. The problem of estimating an expectation should not be seen as a restriction to probability spaces. It is actually closely related to the problem of approximating any given integral with respect to some measure μ defined over $\boldsymbol{\Theta}$. Indeed, if we are able to sample from a distribution Π that admits a density $\pi(.)$ with respect to a measure μ, then for any map of the form $h \triangleq \pi \cdot \varphi$ such that $\Pi(|\varphi|) < +\infty$, we are also able to approximate the integral

$$\int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}) \mu(d\boldsymbol{\theta}) = \Pi(h/\pi)$$

by choosing the consistent Monte Carlo estimator $\widehat{\Pi}_m(h/\pi)$. Such a method is known as importance sampling, see Tokdar and Kass (2010) for a review of its mathematical foundations and properties. A straightforward remark is that we can expect such an approximation to be efficient only if $\Pi(h^2/\pi^2) < +\infty$ so that the central limit theorem holds. That condition will be satisfied essentially if the density π has large enough tails compared to the function of interest h. This emphasizes the motivation behind providing a wide choice of sampling distributions to the user.

1.1.2 The Bayesian paradigm

Statistical models have nowadays a common pattern. A certain amount of data $n \geq 1$ is observed by the user. The dataset itself, noted $\boldsymbol{Y}^{(n)}$, is assumed to be a random

realization from an unknown probability distribution, defined over a sample space $\mathcal{X}^{(n)}$. Choosing a statistical model is assuming that the unknown distribution is part of a particular family of distributions $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ indexed by a parameter $\boldsymbol{\theta}$ that belongs to the parameter space $\boldsymbol{\Theta}$. Inferring such a model is trying to recover which distribution, i.e. which parameter $\boldsymbol{\theta}$ was in charge of generating the random data observed by the user.

Bayesian inference relies on the following methodology. Define a probability distribution Π_0 over the measurable space $\boldsymbol{\Theta} = \mathbb{R}^p$, called *prior* distribution. The parameter $\boldsymbol{\theta}$ is now viewed as a random variable with distribution Π_0, and the distribution $\mathbb{P}_{\boldsymbol{\theta}}$ is viewed as the conditional distribution of $\boldsymbol{Y}^{(n)}$ given $\boldsymbol{\theta}$. In such a setting, the joint distribution of $(\boldsymbol{Y}^{(n)}, \boldsymbol{\theta})$ is fully specified, and in particular the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{Y}^{(n)}$, called *posterior* distribution, that we will denote $\Pi(.|\boldsymbol{Y}^{(n)})$. The posterior distribution is the heart of Bayesian inference.

From a Bayesian perspective, a distribution over $\boldsymbol{\Theta}$ is viewed as an amount of knowledge, or a measure of uncertainty, with respect to the unknown parameter. In that sense, the prior distribution refers to a knowledge with respect to $\boldsymbol{\theta}$ prior to any information provided by the data. Similarly, the posterior distribution refers to an amount of knowledge with respect to $\boldsymbol{\theta}$, updated by the information provided by the data. From a frequentist viewpoint, once a prior distribution Π_0 is chosen, the data-dependent distribution $\Pi(.|\boldsymbol{Y}^{(n)})$ is simply an estimator of the unknown and deterministic parameter $\boldsymbol{\theta}$. In that sense, such an estimator should be analyzed as any other estimator, and would require usual guarantees of convergence. Although those viewpoints may seem conflicting, not to choose between those two is perfectly allowed. A defender of the Bayesian viewpoint could be very interested in the fact that $\Pi(.|\boldsymbol{Y}^{(n)})$ converges at a certain speed when n grows, or that it is robust to misspecification, as well as a defender of the frequentist viewpoint could find very interesting to have another kind of inference available, especially if it makes the inferential framework more convenient.

Above all, it is important to understand that Bayesian estimators are relevant to Statistics, regardless of everyone's convictions from a decision theory perspective, and that in any cases, the posterior distribution is of central interest. It is also noteworthy that its use is not restricted to inference. Indeed, the posterior distribution is also widely used for testing and model selection. Most of the time, the distribution Π is not completely known, in the sense that no simple analytic form is available for computing its moments, or some quantiles of its marginal distributions, etc. And as a matter of fact, best known methods for approximating those quantities involve Monte Carlo methods and require to sample from Π.

1.1.3 Maximum Likelihood approximation

If sampling methods are particularly useful in the Bayesian framework, they find many other applications in statistics, in the standard frequentist framework as well. One particular framework that has been extensively studied is the framework of iid statistical models, i.e. statistical models that relies on a dataset composed by independently and identically distributed random variables. From a mathematical perspective, this frame-

work boils down to assuming that the dataset is composed by independent random variables $Y_1, ..., Y_n$ defined on the same state space $\mathcal{X}$ with the same probability distribution $\mathbb{P}_{\boldsymbol{\theta}}$. The latter distribution is unknown, and as in the previous section, we assume that it is part of a particular family of distributions $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. In such a setting, the main goal is to recover the parameter $\boldsymbol{\theta}$ which generated the iid dataset observed by the user.

A generic method of estimation is applicable when the underlying statistical model admits a *likelihood function*, that is when for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the distribution $\mathbb{P}_{\boldsymbol{\theta}}$ admits a density $f_{\boldsymbol{\theta}}(.)$ with respect to a fixed measure μ (that does not depend on $\boldsymbol{\theta}$). The likelihood function is a data dependent function of $\boldsymbol{\theta}$, defined in the iid framework by

$$\mathscr{L}(\boldsymbol{\theta}; Y_1, ..., Y_n) \triangleq \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(Y_i).$$

The corresponding estimation method, known as maximum likelihood estimation, was pioneered by Ronald Fisher in the beginning of the twentieth century, and later by Wilks (1938). A Maximum Likelihood Estimator (MLE), if it exists, is a maximizer of the map $\boldsymbol{\theta} \mapsto \mathscr{L}(\boldsymbol{\theta}; Y_1, ..., Y_n)$. Provided that the MLE is unique, we will denote it by $\widehat{\boldsymbol{\theta}}_n$. Maximum likelihood estimation became widely used, mainly due to its asymptotic properties. Under mild hypotheses on the model (Wald, 1949), it is shown that the MLE is strongly consistent, that is

$$\widehat{\boldsymbol{\theta}}_n \overset{\text{a.s.}}{\to} \boldsymbol{\theta}$$

when n goes to infinity. Assume that the likelihood function is twice differentiable with respect to $\boldsymbol{\theta}$, and define the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{\theta}}[-\nabla^2_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(Y_1)]$. Under further regularity assumptions on the likelihood function, including the non singularity of $\mathbf{I}(\boldsymbol{\theta})$, it is also shown that the MLE is asymptotically normal in the following sense:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \overset{\mathscr{D}}{\to} \mathcal{N}\left(\mathbf{0}_p, \mathbf{I}(\boldsymbol{\theta})^{-1}\right) \tag{1.1}$$

as n goes to infinity. A particularity of the Fisher information matrix is that its inverse is a lower bound for asymptotic variances of estimators, therefore the MLE is asymptotically optimal. Note that this property is not a claim of supremacy of the MLE from a statistical viewpoint. As a matter of fact, several Bayesian estimators, such as the mean of the posterior distribution, essentially share the same properties if the prior distribution is chosen carefully. Just as for the posterior mean, the interest of the MLE lies in the fact that it is a very generic method of estimation that benefits from guaranties of efficiency when n is large. The main difference from the Bayesian estimators comes from their computation. Just as posterior distributions usually do not have closed form for their moments, likelihoods usually do not have closed form for their maximizers, and numerical approximations of the MLE rely on optimization procedures while the approximation of the posterior moments rely on sampling procedures.

In appearance thus, MLE has nothing to do with sampling methods. However, when the statistical model admits a likelihood function which is partially computable, the MLE may not be approximated by standard optimization procedures. This is the case for instance, when the maps $f_{\boldsymbol{\theta}}(.)$ are only computable up to a normalizing factor that depends on $\boldsymbol{\theta}$. Stated in another form, the MLE is intractable when $f_{\boldsymbol{\theta}}(x) = h_{\boldsymbol{\theta}}(x)/\mathcal{Z}(\boldsymbol{\theta})$ for computationally tractable maps $h_{\boldsymbol{\theta}}(.)$, but a partition function $\mathcal{Z}(\boldsymbol{\theta}) = \int_{\mathcal{X}} h_{\boldsymbol{\theta}}(x)\mu(dx)$

that cannot be computed in a reasonable time. In this setting, referred as un-normalized model, some approximations of the MLE by sampling methods have proven their effectiveness, e.g. Geyer and Thompson (1992); Geyer (1994); Gutmann and Hyvärinen (2010); Gutmann and Hyvärinen (2012). Those methods rely on approximating the partition function by Monte Carlo. This requires to sample from a distribution on the sample space $\mathcal{X}$, chosen by the user. Obviously, the precision of the approximation will depend on the choice of the sampling distribution. Therefore, providing a wide choice of sampling distributions to the user allows him to approximate the MLE efficiently. This framework will be more detailed in Section 1.2 and Chapter 2.

1.1.4 The PAC-Bayesian framework

If the problem of approximating a partially known, or intractable distribution is a main focus in statistics, it is also of direct interest in some fields of machine learning. The machine learning framework, from a theoretical perspective, is often presented as follows. Let us define a sample space $\mathcal{X} \times \mathcal{Y}$. We assume here that $\mathcal{X} \triangleq \mathbb{R}^d$ and $\mathcal{Y} \triangleq \mathbb{R}$, and we consider a dataset $\mathcal{D}_n \triangleq (X_i, Y_i)_{i=1,\dots,n}$ identically and independently distributed from a distribution P defined over the measurable space $\mathcal{X} \times \mathcal{Y}$, equipped with its Borel σ-field. We also choose a set of predictors $\{f_{\boldsymbol{\theta}} : \mathcal{X} \mapsto \mathcal{Y}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ indexed over the parameter space $\boldsymbol{\Theta} = \mathbb{R}^p$. The choice of such a set will depend on the nature of the prediction problem.

The main question now is the following. Suppose that $(X, Y) \sim P$ is a new random vector independent from $\mathcal{D}_n$, such that only X is observed by the user. How can we predict Y? To assess the quality of a given predictor, we define a loss function $\ell : \mathcal{Y}^2 \mapsto \mathbb{R}_+$ and the corresponding risk function $R(\boldsymbol{\theta}) \triangleq \mathbb{E}[\ell(Y, f_{\boldsymbol{\theta}}(X))]$. A predictor is identified as good, if the error of prediction, measured by the loss function, is small in expectation. In other words, a good predictor is a map $f_{\boldsymbol{\theta}} : \mathcal{X} \mapsto \mathcal{Y}$ such that $R(\boldsymbol{\theta})$ is small.

If we assume that the map R has a unique minimizer $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, then finding the best predictor is equivalent to finding this minimizer, also known as *oracle*. Unfortunately, the risk function is unknown from the user, and the only information available concerning the distribution P is the dataset $\mathcal{D}_n$. Therefore, a natural idea is to minimize the empirical counterpart of the risk, defined as

$$r_n(\boldsymbol{\theta}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_{\boldsymbol{\theta}}(X_i)).$$

Empirical Risk Minimization (ERM) was pioneered by Vapnik (1992), who introduced what is now an important field of research in machine learning theory. Essentially, such a method relies on the fact that r_n converges pointwise to R by the law of large numbers. Therefore the minimizer of r_n is expected to a consistent estimator of the oracle.

Another idea was introduced in McAllester (1999), known as the PAC-Bayesian approach. Let us choose a distribution Π_0 over $\boldsymbol{\Theta}$. For some scale parameter $\lambda > 0$, we define the probability measure Π, such that

$$\Pi(d\boldsymbol{\theta}) \triangleq \frac{\exp\{-\lambda r_n(\boldsymbol{\theta})\}}{\int_{\boldsymbol{\Theta}} \exp\{-\lambda r_n\} d\Pi_0} \Pi_0(d\boldsymbol{\theta}).$$

This distribution Π is known in the literature as a Gibbs distribution (Catoni, 2007), or exponentially weighted aggregate (Dalalyan and Tsybakov, 2008). The scale parameter λ induces a certain spread in the distribution: a large λ will make the distribution Π more peaked around its mode. To solve the oracle learning problem, the parameter λ is usually tuned to be an increasing and divergent function of the sample size n, such that the distribution Π converges to a Dirac mass on $\boldsymbol{\theta}^*$ when the sample size goes to infinity. Therefore, the distribution Π is also a consistent estimator of the oracle, as well as the ERM. When $\lambda = n$, it may happen that $\exp\{-\lambda r_n(\boldsymbol{\theta})\}$ has a statistical interpretation, i.e. it may correspond to the likelihood of some IID statistical model. In such a setting, the distribution Π would correspond to the posterior distribution in the Bayesian framework, with prior distribution Π_0. Apart from this thin connection, the distribution Π should simply be seen as a data-dependent distribution that yields consistent estimators.

Estimators based on the Gibbs distribution Π are known as PAC-Bayesian estimators. While ERM usually requires optimization algorithms to minimize the function r_n, PAC-Bayesian estimators are usually approached using Monte Carlo methods. Indeed, as well as for the posterior distribution in the Bayesian framework, in most of applications the Gibbs distribution Π is partially known, in the sense that its moments do not have explicit forms. Therefore, their approximations often require sampling from Π. It is noteworthy that some recent methods, that rely on finding tractable approximations of the Gibbs distribution, has proven their effectiveness (Alquier et al., 2016). Those methods, known as Variational Bayes methods, form now a relevant alternative to Monte Carlo approximations of the Gibbs distribution. Nevertheless, the PAC-Bayesian framework presents another application of sampling methods in the field of statistics and machine learning.

1.1.5 From exact to approximate sampling

All those applications motivate our main goal, which is sampling from a given distribution Π, defined on $\boldsymbol{\Theta} = \mathbb{R}^p$. Fortunately, we usually have some information over the distribution Π. In most cases, some knowledge is available in an analytic sense: for instance, Π admits a density function $\pi(.)$ that is computationally tractable pointwise up to a normalizing constant, or differentiable in a closed form. To achieve the aforementioned, we are essentially allowed to perform any transformation of the vector $(U_1, ..., U_k)$. Stated in another way, our task boils down to finding a measurable map $T : [0,1]^k \mapsto \mathbb{R}^p$ such that

$$T(U_1, ..., U_k) \sim \Pi.$$

This task is usually referred to exact sampling. Such transformations exist, e.g. the inverse cumulative distribution function method, the rejection sampling algorithm[2].

Regrettably, in many situations, exact sampling is not possible in a feasible computational time. As a matter of fact, exact sampling is mostly restricted to very low dimension. One first issue is that T needs to be a computationally tractable map. Another problem is that exact sampling may require a large number k of uniform random variables.

[2]In that case, k is a geometrically distributed random variable.

In most applications of sampling methods to Statistics and Machine Learning, the target distribution Π cannot be sampled exactly. The objective is thus relaxed, and the challenge becomes to sample approximately from Π. Stated in an informal way, our new goal is to design a sequence of tractable maps (T_k) such that the following holds. For every $k \geq 1$, define the corresponding iterate of the sampling procedure

$$\boldsymbol{\theta}_k \triangleq T_k(U_1, ..., U_k)$$

and note $\mathscr{D}(\boldsymbol{\theta}_k)$ the distribution of the random vector $\boldsymbol{\theta}_k$, then for some k not too large we have

$$\mathscr{D}(\boldsymbol{\theta}_k) \approx \Pi.$$

How to measure the quality of such an approximation is a central question, and theoretical guarantees are required, some of those will be discussed shortly. We often refer to the generation of $\boldsymbol{\theta}_k$ as a sampling procedure, or algorithm, as it is usually done recursively by a computer. Among best known approximate sampling methods appear the so called Markov Chain Monte Carlo (MCMC) algorithms. The general principle of MCMC is the following. We define a Markov Chain in such a way that the distribution Π is its *stationary distribution*. A stationary distribution Π is a probability distribution that remains invariant after an iteration of the sampling procedure, i.e. a distribution that satisfies the following implication for any $k \in \mathbb{N}$:

$$\boldsymbol{\theta}_k \sim \Pi \;\Rightarrow\; \boldsymbol{\theta}_{k+1} \sim \Pi.$$

This property does not seem helpful, since we are still not able to sample from Π at a given step. However, under some assumptions, the stationary distribution is unique, and defines the long term distribution of the Markov Chain. Therefore, running the Markov Chain for a sufficiently long time allows us to sample approximately from Π. A particularly good review of MCMC algorithms and their theoretical guarantees was written by Roberts and Rosenthal (2004).

From a computational viewpoint, the complexity of a sampling task is directly related to the number k of sampled uniform variables. Computing the maps (T_k) may also be challenging, but MCMC samplers are designed to compute those maps sequentially, so that the overall complexity of the sampling procedure remains linear in k. Nevertheless, this often raises a trade-off between sampling accuracy and computational complexity.

From an asymptotic viewpoint, one of the first guarantees demanded for an approximate sampling procedure is the convergence in distribution, that is

$$\boldsymbol{\theta}_k \xrightarrow{\mathscr{D}} \Pi$$

as k goes to infinity. This guarantee may be interpreted as follows. Running the sampling algorithm as long as possible allows us to get the best possible sampling approximation of the target distribution Π. Another guarantee, particularly relevant concerning Monte Carlo estimation, is the strong law of large numbers. We say that a sequence $(\boldsymbol{\theta}_k)$ satisfies the strong law of large numbers if for any measurable $\varphi : \boldsymbol{\Theta} \mapsto \mathbb{R}$ such that $\Pi(|\varphi|) < +\infty$ we have

$$\widehat{\Pi}_m(\varphi) \triangleq \frac{1}{m} \sum_{k=1}^{m} \varphi(\boldsymbol{\theta}_k) \xrightarrow{\text{a.s.}} \Pi(\varphi) \tag{1.2}$$

as m goes to infinity.

Under mild conditions (ϕ-irreducibility, Harris recurrence, and aperiodicity, see Roberts and Rosenthal (2004) for definitions), a MCMC sampler will satisfy both guarantees. Convergence in distribution and a strong law of large numbers are often claimed as a proof of validity of the sampling procedure, at least in the limit. However, such a claim should be tempered. Those guarantees do not give any information on the speed of convergence. In particular, those are unhelpful when it comes down to comparing two sampling procedures on a computational basis, or to finding a good trade-off between computational complexity and sampling precision for a given algorithm.

If one hopes to get stronger guarantees, a finer analysis is required, which requires considering a metric over the space of probability distributions on $\mathbb{R}^p$. A well known distance is the Total Variation (TV) distance, which is also a norm over the set of signed measures on $\mathbb{R}^p$. If μ and ν are two probability measures on $\mathbb{R}^p$, then the TV distance between μ and ν is defined as

$$\|\mu - \nu\|_{\mathrm{TV}} \triangleq \sup_{A \in \mathscr{B}(\mathbb{R}^p)} |\mu(A) - \nu(A)|. \tag{1.3}$$

Naturally, other metrics exist, some of them will be introduced in Section 1.3 and Chapters 2 and 3. In this section however, for the sake of clarity, we will consider only the TV metric. For approximate sampling algorithms, a stronger but common guarantee is referred to geometric ergodicity. A sequence $(\boldsymbol{\theta}_k)$ generated by a sampling procedure will be called geometrically ergodic, with respect to the TV distance, if there exist some constants $C > 0$ and $\rho \in]0, 1[$ such that for every $k \geq 1$ we have

$$\|\mathscr{D}(\boldsymbol{\theta}_k) - \Pi\|_{\mathrm{TV}} \leq C\rho^k.$$

Such a result gives further information over the speed of convergence. It guarantees that the sampling approximation will converge to the target distribution exponentially fast with the number of iterates k. For a given MCMC sampler, there exist theoretical conditions (essentially a drift condition, defined in Roberts and Rosenthal (2004)) that ensures geometric ergodicity. For the random-walk Metropolis Hastings algorithm, probably the most generic MCMC sampler, Mengersen and Tweedie (1996) proved that geometric ergodicity holds essentially if and only if the distribution Π has sub-exponential tails.

Another property, particularly appealing for Monte Carlo estimation, is the existence of a Central Limit Theorem (CLT). We say that a sequence $(\boldsymbol{\theta}_k)$ satisfies a $\sqrt{m}$-CLT with respect to a measurable map $\varphi : \boldsymbol{\Theta} \mapsto \mathbb{R}$ if there exists a positive constant $\sigma^2(\varphi)$ such that

$$\sqrt{m}\left(\widehat{\Pi}_m(\varphi) - \Pi(\varphi)\right) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2(\varphi)),$$

where $\widehat{\Pi}_m(\varphi)$ is the sample average defined in (1.2). For MCMC samplers, geometric ergodicity is directly related to the existence of a central limit theorem, in the following sense. Theorem 24 of Roberts and Rosenthal (2004) states that if a sequence $(\boldsymbol{\theta}_k)$ is generated by a geometrically ergodic MCMC sampler, then it satisfies a $\sqrt{m}$-CLT for any measurable φ such that for some $\delta > 0$ we have $\Pi(|\varphi|^{2+\delta}) < +\infty$. In other words, a geometrically ergodic MCMC sampler satisfies a CLT for almost every possible function. In

such a case, the asymptotic variance $\sigma^2(\varphi)$ will be directly related to the autocorrelation function of the Markov Chain. The existence of a central limit theorem is useful in the sense that it allows the construction of asymptotic confidence intervals for Monte Carlo estimates.

From a sampling perspective, geometric ergodicity seems particularly appealing, as it ensures that the sampling error will decrease exponentially fast with the number of iterates k. However, the constants C and ρ are usually not explicit. Without additional information, geometric ergodicity is still unhelpful for determining which k to choose to reach a given precision, and how the computational complexity will scale with the dimension.

To answer those questions, one needs to introduce an explicit measure of the computational complexity. Such a measure can be defined through the *mixing time* of a sampling procedure. For a given precision level $\varepsilon > 0$, and a sequence $(\boldsymbol{\theta}_k)$ generated by a sampling procedure, the mixing time, with respect to the TV distance, is defined as

$$\mathcal{K}_{\varepsilon}^{\text{sam}} \triangleq \min\{k \in \mathbb{N} : \|\mathscr{D}(\boldsymbol{\theta}_k) - \Pi\|_{\text{TV}} \leq \varepsilon\}.$$

It is the minimum number of iterations required to reach a certain precision on the sampling approximation error. In other words, if the user knows a bound on the mixing time then he knows when to stop the sampling algorithm. Although such a measure has a simple interpretation, deriving bounds on the mixing time of a given sampling procedure can actually be quite hard, or not even feasible for complex sampling procedures.

As a matter of fact, it is known that the performance of most sampling schemes deteriorates very fast when the dimension increases. Approximate sampling for high dimensional probability distributions appears to be so challenging that it is often referred to a curse of dimensionality. Unsurprisingly, it is a very active field of research. Several approximate sampling procedures have proven their effectiveness into tackling this issue. Well known methods include: Langevin Monte Carlo, Hamiltonian Monte Carlo and Sequential Monte Carlo. To assess their performance, one branch of research, more theoretical, tries to establish non-asymptotic guarantees over the approximation error. These works are devoted to bounding the mixing times of approximate sampling procedures, and to study in particular their dependence on the dimension and the precision level. Such strong theoretical guarantees are by nature difficult to establish, and therefore require to restrict the study to simple algorithms within a particular family of sampling distributions. To date, explicit guarantees for mixing times are mostly restricted to unimodal and sub-exponential distributions. Another research branch, more methodological, is devoted to propose and assess new samplers for high dimensional distributions, usually through computationally intensive numerical experiments. Such a task is very valuable as well. It implies giving user-friendly advices on how to tune those samplers. Those two research branches are different by nature, nevertheless they are both very useful in finding approximate sampling solutions for high dimensional probability distributions.

A significant part of this thesis concerns high dimensional sampling, and is conducted within the first branch. Quantitative results for high dimensional sampling will be the focus of Section 1.3, but also Chapters 3 and 4. Prior to this, we will present some statistical problems posed in the framework of intractable likelihoods. We make a particular

emphasis on the inference of un-normalized models. This framework, briefly evoked in Section 1.1.4, will be the main focus of Section 1.2, but also Chapter 2.

1.2 Inference of un-normalized statistical models

In this section, we focus on sampling-based solutions for the inference of partially known statistical models. The lack of knowledge of a statistical model is to be understood here from a computational point of view. We consider the problem of inferring statistical models that involve intractable likelihood functions. More precisely, we study models whose likelihoods are only computable up to a normalizing factor. Several examples of un-normalized models are presented in the sequel. Those models poses problems for both Bayesian and frequentist inference. The problem of their inference is considered here in a unifying framework. We firstly show how MCMC samplers can be helpful in such a setting, before focusing on a generic method of estimation proposed by Geyer (1994), referred as Monte Carlo Maximum Likelihood Estimation (MC-MLE). We discuss the precision of the underlying approximation depending on the choice of the sampling distribution. Finally, we present a recent method of approximating the MLE proposed by Gutmann and Hyvärinen (2010); Gutmann and Hyvärinen (2012) referred as Noise Contrastive Estimation, which was observed to be more stable. Chapter 2 provides theoretical support to this observation.

1.2.1 A general framework

The framework of un-normalized models can be formalized as follows. We suppose in this section that $\boldsymbol{\Theta}$ can be any subset of $\mathbb{R}^p$. Consider a set of non-negative functions $\{h_{\boldsymbol{\theta}} : \mathcal{X} \mapsto \mathbb{R}_+, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ that are integrable with respect to a fixed measure μ, and define the function $\mathcal{Z}(\boldsymbol{\theta}) \triangleq \int_{\mathcal{X}} h_{\boldsymbol{\theta}}(x)\mu(dx)$. Therefore, for any fixed $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the function

$$f_{\boldsymbol{\theta}}(x) \triangleq \frac{h_{\boldsymbol{\theta}}(x)}{\mathcal{Z}(\boldsymbol{\theta})}, \qquad \forall x \in \mathcal{X}, \tag{1.4}$$

defines a probability density with respect to μ. The set of densities $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ defines a set of corresponding probability distributions $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, and therefore defines a statistical model. From a computational perspective, we assume that the maps $h_{\boldsymbol{\theta}}$ have closed form. However, this claim does not guarantee that the partition function $\mathcal{Z}(.)$ will be tractable. When the partition function is intractable, we say that the statistical model is un-normalized. Intractable partition functions precludes standard estimation procedures, and therefore un-normalized models are usually avoided by practitioners. Although, it is remarkable that exponential models, that form a widely studied family in statistics, have no guarantee of being normalized, as their canonical formulation boils down to assuming that $h_{\boldsymbol{\theta}}(x) = \exp\{\boldsymbol{\theta}^\top S(x)\}$ for some measurable map $S : \mathcal{X} \mapsto \mathbb{R}^p$. As a matter of fact, several exponential models used in practice are un-normalized, some of them are introduced hereafter.

Example 1: Exponential Random Graphical Models

A first example is the so called Exponential Random Graphical Model (ERGM), used in social network modelling (Robins et al., 2007). Assume that X is a random network, formalized by a random set of edges between n fixed nodes. Such a network has $n(n-1)/2$ possible edges. From a mathematical perspective, X can be defined as a random vector with binary components, that lives in the set $\mathcal{X} = \{0,1\}^{n(n-1)/2}$. The exponential modelization of this random phenomenon boils down to assuming that there is a map $S : \mathcal{X} \mapsto \mathbb{R}^p$ such that the probability of observing a network $x \in \mathcal{X}$ is proportional to $h_{\boldsymbol{\theta}}(x) = \exp\{\boldsymbol{\theta}^\top S(x)\}$. In the latter case, μ is the counting measure over $\mathcal{X}$ and, in the setting of networks, $S(x)$ is a summary statistic of the structure of the network (e.g. the number of edges, triangles, isolated nodes), that determines its likelihood. In this example, the partition function $\mathcal{Z}(\boldsymbol{\theta})$ is a sum over the set $\mathcal{X}$, with cardinal $2^{n(n-1)/2}$. Obviously, when the number of edges n is large, the partition function becomes computationally intractable, which places ERGM into the framework of un-normalized models.

Example 2: Ising Model

A second example of un-normalized model is the so-called Ising Model, named after the physicist Ernst Ising. In a particular field of Physics called statistical mechanics, the Ising model refers to a physical model of ferromagnetism, that formalizes mathematically the interaction between magnetic dipole moments of atomic spins. Fortunately, a deep understanding of what is a magnetic dipole moment or an atomic spin is not required in the sequel. We may limit our understanding to a modelization of very small magnetic entities called *spins*, that interact with their neighbors, and that can be either positive or negative. From a mathematical perspective, we assume that X is a random set of spins on a square lattice. If the lattice is composed by n^2 spins, then X forms a random vector with binary components, that lives in the set $\mathcal{X} = \{-1,1\}^{n^2}$. From a statistical viewpoint, the Ising model corresponds to an exponential modelization of this random phenomenon where the probability of observing a given lattice $x \in \mathcal{X}$ is assumed proportional to $h_{\boldsymbol{\theta}}(x) = \exp\{\boldsymbol{\theta}^\top S(x)\}$, where the statistic $S(x)$ summarizes the interactions between the spins of x (e.g. the number of pairs of neighbours with common spin). The computational problem faced in the Ising model is very close to the one faced for ERGM since the partition function is a sum over 2^{n^2} possible lattices, computationally intractable when n is large.

Example 3: Truncated Gaussian Model

If non computable partition functions arise in statistical modelling of interactions, or dependencies, those can arise in iid statistical models as well. A simple example is the Truncated Gaussian Model. It is common knowledge that Gaussian distributions belong to the exponential family, through a suitable reparametrization. A Gaussian distribution on $\mathbb{R}^d$ with mean $\mathbf{m}$ and non singular covariance matrix $\boldsymbol{\Sigma}$ admits a density with respect to Lebesgue's measure proportional to

$$h_{\mathbf{m},\boldsymbol{\Sigma}}(x) = \exp\left\{-\frac{1}{2}(x-\mathbf{m})^\top \boldsymbol{\Sigma}^{-1}(x-\mathbf{m})\right\}.$$

The partition function $\mathcal{Z}(\mathbf{m},\boldsymbol{\Sigma}) = \int_{\mathcal{X}} h_{\mathbf{m},\boldsymbol{\Sigma}}(x)dx$ is tractable when the Gaussian measure charges the whole space $\mathcal{X} = \mathbb{R}^d$. However, when it is truncated to a subspace of $\mathbb{R}^d$, for instance $\mathcal{X} =]0,+\infty[^d$, the partition function is multiplied by the Gaussian probability

measure of $\mathcal{X}$, which is intractable. The computational problem here lies in the fact that numerical approximations of the Gaussian measure of an arbitrary subset of $\mathbb{R}^d$ become inefficient when d increases. An interesting question is whether one can infer the parameters $(\mathbf{m}, \mathbf{\Sigma})$ from iid observations $Y_1, ..., Y_n$ with Gaussian distribution truncated to $\mathcal{X}$. In this setting, Truncated Gaussian Models also find their place among un-normalized models.

1.2.2 MCMC sampling for un-normalized models

A very well known MCMC method that is relevant here is the Metropolis Hastings (MH) algorithm. Let Π be a probability distribution on $\mathbb{R}^d$ with density $\pi(.)$ with respect to a measure ν. The MH algorithm is a generic MCMC sampler that is useful for sampling approximately from Π, provided that its density is computable pointwise up to a normalizing constant. We assume therefore that, for some integrable function $p(.)$ computable pointwise we have $\pi(x) = p(x)/C$ where $C = \int_{\mathbb{R}^d} p(x)\nu(dx)$. The use of the MH algorithm presupposes that there is a family of conditional densities $\{q(x|y) : y \in \mathbb{R}^d\}$ with respect to ν, such that we can sample from the corresponding probability distribution $Q(dx|y)$ for any $y \in \mathbb{R}^d$. This distribution is called *proposal distribution*. The Metropolis Hastings algorithm is defined as follows:

Algorithm 1 Metropolis Hastings algorithm

Require: An initial distribution ν_0 on $\mathbb{R}^d$, and a number of steps $m \geq 1$.
  $X_0 \sim \nu_0$
  for $j = 0, \ldots, m-1$ **do**
    $Z \sim Q(dx|X_j)$
    $U \sim \mathcal{U}_{]0,1[}$
    $\alpha \leftarrow \dfrac{p(Z)}{p(X_j)} \times \dfrac{q(X_j|Z)}{q(Z|X_j)}$
    $X_{j+1} \leftarrow Z\mathbb{1}_{U<\alpha} + X_j\mathbb{1}_{U>\alpha}$
  end for

The MH algorithm defines a Markov Chain that admits Π as an invariant distribution. This property follows from the definition of the variable α in Algorithm 1, referred as *acceptance ratio*. The sample $X_1, \ldots, X_m$ obtained is then particularly helpful in estimating moments of Π, as mentioned in Section 1.1.5.

In connection to un-normalized models, the MH algorithm can be used to sample approximately from any distribution $\mathbb{P}_{\boldsymbol{\psi}}$ with density $f_{\boldsymbol{\psi}}(x) = h_{\boldsymbol{\psi}}(x)/\mathcal{Z}(\boldsymbol{\psi})$ without the knowledge of the normalizing constant $\mathcal{Z}(\boldsymbol{\psi})$. In other words, one may use the MH algorithm to generate random data from any distribution that belongs to the statistical model. At first sight, this property does not seem helpful in solving the problem of inference caused by the intractable partition function. However, we will see in Sections 1.2.3 and 1.2.4 that this property is actually very helpful when it comes down to approximate the likelihood by Monte Carlo methods.

Prior to this, we briefly discuss the nature of the inferential issue inherent in an

un-normalized model, when it comes down to computing the MLE, or to performing Bayesian inference. If $x \in \mathcal{X}$ is some data observed by the user, the likelihood function $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(x)$ is said to be intractable, in the sense that $f_{\boldsymbol{\theta}}(x) = h_{\boldsymbol{\theta}}(x)/\mathcal{Z}(\boldsymbol{\theta})$ cannot be computed in practice for a given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. It is well known that the computing of the MLE often requires a numerical approximation by an optimization procedure, e.g. the gradient descent algorithm. Unfortunately, if the partition function $\mathcal{Z}(.)$ is intractable, the computing of its gradient at each step of the algorithm will not be feasible either. Let us now consider the Bayesian framework. The user chooses a prior distribution Π_0 on $\boldsymbol{\Theta}$ with density $\pi_0(.)$ with respect to a measure ν. The posterior distribution $\Pi(.|x)$ is then defined by the conditional density

$$\pi(\boldsymbol{\theta}|x) \triangleq \frac{\pi_0(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(x)}{\int_{\boldsymbol{\Theta}} \pi_0(\mathbf{u}) f_{\mathbf{u}}(x) \nu(d\mathbf{u})}, \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{1.5}$$

The approximation of the moments of the posterior distribution often relies on sampling methods, e.g. the Metropolis Hastings algorithm. In order to sample approximately from the posterior distribution $\Pi(.|x)$ with Algorithm 1, the computation of the denominator in (1.5) is not necessary, but the computation of $\mathcal{Z}(.)$ is required in the numerator, which is not feasible. Actually, its computation is needed twice at each step of the algorithm, because the probability transition kernel of the underlying Markov Chain, from a state $\boldsymbol{\theta}$ to a state $\boldsymbol{\psi}$ on $\boldsymbol{\Theta}$ depends on the ratio $\mathcal{Z}(\boldsymbol{\psi})/\mathcal{Z}(\boldsymbol{\theta})$. For that reason, this inferential problem is also known as *doubly intractable* distributions, in the Bayesian literature. Ironically, despite appearing as twice as unfeasible, the problem of sampling approximately from the posterior distribution received much attention, see for instance Møller et al. (2006); Murray et al. (2012); Lyne et al. (2015). Among the methods proposed by those authors, we present here the Exchange algorithm, introduced by Murray et al. (2012) as an improved version of the Auxiliary Variable algorithm proposed by Møller et al. (2006). The main assumption to allow its use is the following. Assume we are able to sample exactly from $\mathbb{P}_{\boldsymbol{\psi}}$ for any $\boldsymbol{\psi} \in \boldsymbol{\Theta}$. This assumption will be discussed shortly. The Exchange algorithm is defined as follows.

Algorithm 2 Exchange algorithm

Require: An initial distribution ν_0 on $\boldsymbol{\Theta}$, and a number of steps $m \geq 1$.
  $\boldsymbol{\theta}_0 \sim \nu_0$
  for $j = 0, \ldots, m-1$ **do**
    $\boldsymbol{\psi} \sim Q(d\mathbf{u}|\boldsymbol{\theta}_j)$
    $w \sim \mathbb{P}_{\boldsymbol{\psi}}$
    $U \sim \mathcal{U}_{]0,1[}$
    $$\alpha \leftarrow \frac{\pi_0(\boldsymbol{\psi}) f_{\boldsymbol{\psi}}(x)}{\pi_0(\boldsymbol{\theta}_j) f_{\boldsymbol{\theta}_j}(x)} \times \frac{f_{\boldsymbol{\theta}_j}(w)}{f_{\boldsymbol{\psi}}(w)} \times \frac{q(\boldsymbol{\theta}_j|\boldsymbol{\psi})}{q(\boldsymbol{\psi}|\boldsymbol{\theta}_j)}$$
    $\boldsymbol{\theta}_{j+1} \leftarrow \boldsymbol{\psi} \mathbb{1}_{U<\alpha} + \boldsymbol{\theta}_j \mathbb{1}_{U>\alpha}$
  end for

Algorithm 2 is a modified version of the Metropolis Hastings algorithm. Compared to the MH algorithm, at each step: an artificial random data w is drawn from a distribution $\mathbb{P}_{\boldsymbol{\psi}}$, and the acceptance ratio is multiplied by $f_{\boldsymbol{\theta}_j}(w)/f_{\boldsymbol{\psi}}(w)$. As for the MH algorithm,

the Exchange algorithm defines a Markov Chain with $\Pi(.|x)$ as a stationary distribution. Contrary to the MH algorithm, the Exchange algorithm does not require the computation of the partition function, because the intractable term $\mathcal{Z}(\boldsymbol{\theta}_j)/\mathcal{Z}(\boldsymbol{\psi})$ cancels out in the acceptance ratio α. In other words, replacing the maps $f_{\boldsymbol{\theta}_j}(.)$ and $f_{\boldsymbol{\psi}}(.)$ by $h_{\boldsymbol{\theta}_j}(.)$ and $h_{\boldsymbol{\psi}}(.)$ in Algorithm 2 does not change its definition.

Therefore, the Exchange algorithm can be used to generate a sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ helpful for estimating the moments of the posterior distribution $\Pi(.|x)$ without any knowledge of the partition function. However, running the Exchange algorithm requires being able to sample exactly from $\mathbb{P}_{\boldsymbol{\psi}}$ for any $\boldsymbol{\psi} \in \boldsymbol{\Theta}$. Unfortunately, this is a rather strong assumption. As emphasized in the beginning of this section, for a fixed $\boldsymbol{\psi} \in \boldsymbol{\Theta}$, the MH algorithm can be used to sample generate a Markov Chain with stationary distribution $\mathbb{P}_{\boldsymbol{\psi}}$, but there are few settings for which exact sampling is feasible in a reasonable computational time.

1.2.3 Monte Carlo Maximum Likelihood Estimation

We now turn to an approximation method of the MLE, called Monte Carlo Maximum Likelihood estimation. This method was introduced by Geyer and Thompson (1992). Its main theoretical justification was established a few years later by Geyer (1994). The general principle of this method is explained is the sequel. Suppose that for some $\boldsymbol{\psi} \in \boldsymbol{\Theta}$, a sample $X_1, \ldots, X_m$ is drawn by the Metropolis Hastings algorithm with target distribution $\mathbb{P}_{\boldsymbol{\psi}}$. Suppose also that the Markov Chain $(X_j)_{j \geq 0}$ satisfies a strong law of large numbers (see Section 1.1.5 for conditions). In this case, for any fixed $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have in particular

$$\frac{1}{m} \sum_{j=1}^{m} \frac{h_{\boldsymbol{\theta}}(X_j)}{h_{\boldsymbol{\psi}}(X_j)} \overset{\text{a.s.}}{\to} \frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\psi})}, \tag{1.6}$$

as m goes to infinity. In connection to Section 1.1.1, the empirical average in (1.6) can be seen as an importance sampling estimate of the intractable ratio $\mathcal{Z}(\boldsymbol{\theta})/\mathcal{Z}(\boldsymbol{\psi})$, where the random variables $X_1, \ldots, X_m$ are generated by a Markov Chain with long term distribution $\mathbb{P}_{\boldsymbol{\psi}}$.

Suppose now that some data $x \in \mathcal{X}$ is observed by the user, and suppose that the likelihood $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(x)$ has a unique maximizer, i.e. suppose that the MLE exists and is unique. Whatever is the value of $\boldsymbol{\psi}$, the MLE is also the maximizer of the log likelihood ratio function defined for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ by

$$\ell(\boldsymbol{\theta}) \triangleq \log \frac{h_{\boldsymbol{\theta}}(x)}{h_{\boldsymbol{\psi}}(x)} - \log \left\{ \frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\psi})} \right\}. \tag{1.7}$$

From this equivalence, Geyer (1994) proposed to approximate the MLE by maximizing the function

$$\ell_m^{\text{IS}}(\boldsymbol{\theta}) \triangleq \log \frac{h_{\boldsymbol{\theta}}(x)}{h_{\boldsymbol{\psi}}(x)} - \log \left\{ \frac{1}{m} \sum_{j=1}^{m} \frac{h_{\boldsymbol{\theta}}(X_j)}{h_{\boldsymbol{\psi}}(X_j)} \right\}. \tag{1.8}$$

The maximizer of $\ell_m^{\text{IS}}(.)$, if it exists, is called Monte Carlo Maximum Likelihood Estimator (MC-MLE). From (1.6), we deduce that for every fixed point $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, then $\ell_m^{\text{IS}}(\boldsymbol{\theta})$ is a consistent estimator of $\ell(\boldsymbol{\theta})$ as m goes to infinity. In other words, the map $\ell_m^{\text{IS}}(.)$ is an

estimator of the map $\ell(.)$, that is consistent pointwise. Moreover, Geyer (1994) showed, under mild assumptions, that the MC-MLE is also a consistent approximation of the MLE, when m goes to infinity.

Although, a legitimate question is whether this approximation is accurate for every sampling distribution $\mathbb{P}_{\boldsymbol{\psi}}$. The answer is no, this is illustrated in the sequel. Suppose that the Markov Chain $(X_j)_{j\geq 0}$ generated by the Metropolis Algorithm is geometrically ergodic. Then it follows from Roberts and Rosenthal (2004, Theorem 25), that $\ell_m^{\mathrm{IS}}(\boldsymbol{\theta})$ is an asymptotically normal estimator of $\ell(\boldsymbol{\theta})$ if and only if $\mathbb{P}_{\boldsymbol{\psi}}(h_{\boldsymbol{\theta}}^2/h_{\boldsymbol{\psi}}^2) < +\infty$. In other words, for a fixed sampling distribution $\mathbb{P}_{\boldsymbol{\psi}}$, the map $\ell_m^{\mathrm{IS}}(.)$ may be an acceptable approximation of the map $\ell(.)$ only on the set $\boldsymbol{\Theta}_\psi \triangleq \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathbb{P}_{\boldsymbol{\psi}}(h_{\boldsymbol{\theta}}^2/h_{\boldsymbol{\psi}}^2) < +\infty\}$. On a wide region of the parameter space, the likelihood function may not be approximated correctly. Therefore, we can expect the MC-MLE to be accurate only if the likelihood is well approximated around its maximizer.

Geyer (1994) emphasizes this fact, by showing that when the model is exponential, MC-MLE is asymptotically normal essentially if and only if the MLE belongs to $\boldsymbol{\Theta}_{\boldsymbol{\psi}}$. In exponential models, $\boldsymbol{\Theta}_{\boldsymbol{\psi}}$ is a convex subset of $\boldsymbol{\Theta}$ that contains $\boldsymbol{\psi}$. This supports the idea that MC-MLE is accurate essentially if $\boldsymbol{\psi}$ is close to the MLE. Chapter 2 provides further insight on this intuition.

In practice, the sampling distribution $\mathbb{P}_{\boldsymbol{\psi}}$ is chosen by the user, and the MLE is intractable. Therefore, the choice of the sampling distribution is an important issue. A heuristic strategy is the following. For an initial parameter $\boldsymbol{\psi}_0 \in \boldsymbol{\Theta}$ generate a sample $X_1, \ldots, X_m$ from the MH algorithm with respect to the target distribution $\mathbb{P}_{\boldsymbol{\psi}_0}$. Then compute the MC-MLE estimator, and plug its value to $\boldsymbol{\psi}_1$. Repeat the estimation process with a new sample drawn from $\mathbb{P}_{\boldsymbol{\psi}_1}$, and so on. The main hope when choosing this strategy is that the MC-MLE estimates may stabilize when those estimates eventually get close to the MLE. This methodology has no theoretical support from a general perspective. In specific models however, providing further guidance on the choice of the sampling distribution may be possible.

1.2.4 Noise Contrastive Estimation

Section 1.2.3 was devoted to present an approximation of the MLE, and to assess its accuracy with respect to the Monte Carlo error. However, the definition of un-normalized models proposed in (1.4) presupposes that the observed data $x \in \mathcal{X}$ is fixed. This viewpoint does not take into account the dimension, nor the structure of the observed data. In particular, a point that should be emphasized is the computational burden of sampling on $\mathcal{X}$.

One underlying question is whether MC-MLE can be improved if a certain structure is assumed on the dataset. The most simple case one can consider is the framework of n iid observations. In this framework, the computational complexity of a sampling procedure on $\mathcal{X}$ is independent of the amount of observed data n. It will be mostly related to the dimension of $\mathcal{X}$.

Suppose that the dataset is constituted of n iid random variables $Y_1, \ldots, Y_n$ over the space $\mathcal{X}$. Suppose also that we are only able to generate m random variables $X_1, \ldots, X_m$

in a reasonable computational time. An interesting question is: how accurate will be MC-MLE, when the size of the dataset n is large? In particular, if the amount of data observed n is non-negligeable with respect to m, the inferential error of the MLE is expected to be also non-negligeable with respect to the Monte Carlo error of the MC-MLE. When the latter happens, both errors should be studied in a common framework. This point is discussed further in Chapter 2.

In the framework of un-normalized models for iid data, Gutmann and Hyvärinen (2010) proposed an estimation method known as Noise Contrastive Estimation (NCE). This method will be described in Chapter 2. As for MC-MLE, the principle of NCE relies on approximating a likelihood by Monte Carlo methods (Barthelmé and Chopin, 2015). A common aspect between NCE and MC-MLE is that both methods require sampling on the space $\mathcal{X}$. In Section 1.2.3, we emphasized the fact that the accuracy of MC-MLE may be sensitive to the choice the sampling distribution $\mathbb{P}_\psi$. A remarkable point is that the accuracy of NCE is much less sensitive to this choice. Gutmann and Hyvärinen (2012) provided numerical evidence that NCE is more stable than MC-MLE, especially when m/n is low. Chapter 2 provides a formal justification of this fact, and shows in the same time that the two methods have a lot of connections from a theoretical perspective.

1.3 Quantitative results for high dimensional sampling

In this section, we focus on the problem of controlling the approximation error of a sampling procedure. This error may be measured by several distances or divergences, some of them will be introduced in the sequel. The choice of a distance or divergence is necessary to quantify the error of an approximate sampling procedure after a certain number of iterations. We are interested here in providing explicit non asymptotic guarantees over the sampling error, in order to get a reliable measure of the computational complexity of the sampling procedure. A convenient framework to establish such results is the set of smooth and log concave distributions. In this setting, we focus on the Langevin Monte Carlo algorithm, whose non asymptotic study, pioneered by Dalalyan (2017b), inspired many authors in the last few years. After an overview of those results, we end up this introduction by presenting the Kinetic Langevin Monte Carlo algorithm, a special case of Hamiltonian Monte Carlo introduced by Cheng et al. (2018), which was shown to have a better scaling with the dimension. Chapters 3 and 4 provide further insight on its quantitative guarantees.

1.3.1 Measuring distances between probability distributions

Establishing quantitative guarantees on the error of an approximate sampling procedure requires choosing a metric over the space of probability distributions on $\mathbb{R}^p$. We will consider in this thesis mostly two well known distances. The first one is the total variation distance, already defined in (1.3). Another distance we will consider is the so-called Wasserstein distance. For any two probability measures μ and ν on $\mathbb{R}^p$, we first define

$\mathscr{C}(\mu, \nu)$ the corresponding set of *couplings*, that is the set of probability distributions on $\mathbb{R}^p \times \mathbb{R}^p$ with marginal distributions μ and ν. Then for any real $q \geq 1$, we define the q-Wasserstein distance between μ and ν, with respect to the Euclidean norm noted $\|.\|_2$, that is

$$W_q(\mu, \nu) \triangleq \left(\inf_{\varrho \in \mathscr{C}(\mu,\nu)} \int_{\mathbb{R}^p \times \mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^q \, d\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') \right)^{1/q}.$$

The Wasserstein distance, also known as the Wasserstein-Monge-Kantorovich distance, comes from optimal transport theory. For a complete introduction to its origin and properties see the book of Villani (2008). In many applications of Statistics and Machine Learning, the use of the Wasserstein distance has become popular, essentially because it often simplifies the analysis compared to the TV metric, and because it is directly linked to the moments of the distributions which are of particular interest within statistical estimation frameworks.

When the choice of a metric between probability distributions is too constraining, it is common to consider so called statistical *divergences*. Those are pseudo-metrics in the sense that they do not satisfy symmetry nor a triangular inequality. Let μ and ν be two probability distributions on $\mathbb{R}^p$ such that μ is absolutely continuous with respect to ν. Some well known divergences between μ and ν are the Kullback-Leibler divergence

$$\mathrm{KL}(\mu \| \nu) \triangleq \int_{\mathbb{R}^p} \log \left(\frac{d\mu}{d\nu}(\boldsymbol{\theta}) \right) \nu(d\boldsymbol{\theta})$$

and the χ^2 divergence

$$\chi^2(\mu \| \nu) \triangleq \int_{\mathbb{R}^p} \left(\frac{d\mu}{d\nu}(\boldsymbol{\theta}) - 1 \right)^2 \nu(d\boldsymbol{\theta}).$$

1.3.2 Comparing the complexity of two sampling procedures

Since our main goal is to sample approximately from Π, an interesting question is how to assess the quality of such an approximation. One convincing solution is to compute quantitative rates of convergence to the target distribution. Let $(\boldsymbol{\theta}_k)$ be a sequence of random vectors on $\mathbb{R}^p$, generated by an approximate sampling procedure. Once a metric, or a divergence, is chosen, then for a given precision level $\varepsilon > 0$, assessing the quality of the sampling approximation essentially boils down to controlling the mixing time of the sampling procedure. Recall that the mixing time of the sampling procedure, with respect to the TV distance, is

$$\mathcal{K}_\varepsilon^{\mathrm{sam}} = \min\{k \in \mathbb{N} : \|\mathscr{D}(\boldsymbol{\theta}_k) - \Pi\|_{\mathrm{TV}} \leq \varepsilon\}.$$

In this section, we may refer to the mixing time with respect to another metric or divergence. In this case, the TV distance should be replaced by the corresponding metric or divergence in the definition above.

Our new task is, for a given sampling procedure, to bound the number of iterations required to reach a certain precision, measured by a distance between probability distributions. The purpose of such a task is twofold. First, it is very convenient for the user,

because it provides a stopping rule to the sampling algorithm. Second, it is quite valuable from the theoretician's point of view, especially for comparing two sampling algorithms on the computational complexity basis. In particular, determining which algorithms have the best mixing times in terms of their scaling with the dimension is of tremendous importance. Moreover, it is essentially the only way of providing guidance towards efficient sampling procedures for (very) high dimensional distributions, simply because numerical experiments become too costly when the dimension increases.

Before stating precise results on this topic, let us remark that the mixing time of a sampling procedure can be related to a similar notion in optimization. Indeed, suppose we are interested in minimizing a certain function $f : \mathbb{R}^p \mapsto \mathbb{R}$, which is continuously differentiable, and has a unique minimizer $\boldsymbol{\theta}^*$. Suppose moreover that $(\boldsymbol{\vartheta}_k)$ is a determistic sequence, induced by an optimization algorithm, that converges to $\boldsymbol{\theta}^*$.

In such a setting, the computational complexity will be directly related to the convergence time of the optimization algorithm. For a given precision level $\varepsilon > 0$, we define the convergence time with respect to the Euclidean norm, that is

$$\mathcal{K}_\varepsilon^{\text{opt}} \triangleq \min\{k \in \mathbb{N} : \|\boldsymbol{\vartheta}_k - \boldsymbol{\theta}^*\|_2 \leq \varepsilon\}.$$

Just as the mixing time is for the sampling problem, the convergence time is the minimum number of iterations to reach a given accuracy. The only difference is that the approximation error is measured here by a distance to the minimizer, which is the Euclidean norm.

A priori, the optimization problem is different from the problem of sampling for a given distribution Π. Although, when it comes down to comparing two sampling algorithms on the computational complexity basis, the two problems should be related. The similarity between the mixing time and the convergence time motivates the idea of comparing two sampling algorithms in the same way as we do for comparing two optimization algorithms. We will see that other similarities between the sampling problem and the optimization problem appear when considering a more precise framework.

1.3.3 Focus on Langevin Monte Carlo

We now restrict our attention to the unimodal distributions Π that admit a positive and continously differentiable density with respect to Lebesgue measure on $\mathbb{R}^p$. Stated in another way, we focus on the distributions Π that admit a density noted π with respect to Lebesgue measure, of the form

$$\pi(\boldsymbol{\theta}) \triangleq \frac{e^{-f(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f(\mathbf{u})} d\mathbf{u}}, \qquad \forall \boldsymbol{\theta} \in \mathbb{R}^p,$$

where $f : \mathbb{R}^p \mapsto \mathbb{R}$ is a continuously differentiable function, which has a unique minimizer $\boldsymbol{\theta}^*$. Therefore, the map f needs to be coercive in such a way that π remains a probability density. In this context, the problem of minimizing f and the problem of sampling from a distribution with density $\pi(\boldsymbol{\theta}) \propto e^{-f(\boldsymbol{\theta})}$ actually have a particular connection. This connection lies in several similarities between the so called gradient descent algorithm and Langevin Monte Carlo algorithm.

Concerning the optimization problem, a very common strategy relies on running the Gradient Descent (GD) algorithm, defined as follows. For some deterministic $\boldsymbol{\vartheta}_0 \in \mathbb{R}^p$ and for some step size $h > 0$, define the sequence

$$\boldsymbol{\vartheta}_{k+1} = \boldsymbol{\vartheta}_k - h\nabla f(\boldsymbol{\vartheta}_k), \qquad k = 0, 1, 2, ... \tag{1.9}$$

Essentially, it is known that under some assumptions on f and h, the deterministic sequence $(\boldsymbol{\vartheta}_k)$ will converge to the minimizer $\boldsymbol{\theta}^*$.

Concerning the sampling problem, the so called Langevin Monte Carlo (LMC) algorithm (also known as the unadjusted Langevin algorithm) is defined as follows. Assume that $(\boldsymbol{\xi}_k)$ is a sequence of IID standard Gaussian vectors on $\mathbb{R}^p$. For some $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, that may be either deterministic or random, and for some step size $h > 0$, define

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - h\nabla f(\boldsymbol{\theta}_k) + \sqrt{2h}\boldsymbol{\xi}_{k+1}, \qquad k = 0, 1, 2, ... \tag{1.10}$$

As its definition essentially boils down to adding a noise term to a gradient descent step, the LMC is sometimes referred to the sampling counterpart of the GD algorithm. It is common knowledge that under some assumptions on f and h, the sequence $(\boldsymbol{\theta}_k)$ will be a Markov Chain that converges in distribution to a probability measure Π_h, that is close to Π when h is small enough. This property is a consequence of the fact that (1.10) is a discretization of the continuous-time Langevin process, defined through the following stochastic differential equation

$$d\boldsymbol{L}_t = -\nabla f(\boldsymbol{L}_t)dt + \sqrt{2}d\boldsymbol{W}_t, \qquad t \geq 0 \tag{1.11}$$

where $\{\boldsymbol{W}_t : t \geq 0\}$ stands for the standard p-dimensional Brownian motion. Let $\boldsymbol{L}_0 \in \mathbb{R}^p$ be either deterministic or random. Then, under some assumptions on f, (1.11) has a unique solution, which is a Markov process $\{\boldsymbol{L}_t : t \geq 0\}$ converging in distribution to Π when t goes to infinity. Unfortunately, there is no efficient method to sample exactly from the continuous-time Langevin process.

Built upon the Euler discretization of (1.11), LMC appeared as a natural candidate for sampling approximately from Π. From an asymptotic viewpoint, there is no convergence to the target distribution for a fixed step size $h > 0$. However, one may hope reaching a given precision by choosing first a step size h small enough such that Π_h is close to Π, then running the LMC algorithm for a sufficient number of steps so that $\mathscr{D}(\boldsymbol{\theta}_k)$ gets close to Π_h, and therefore close to Π. A first important contribution to the analysis of the probabilistic properties of the LMC was done by Roberts and Tweedie (1996). The authors show in particular that the chain may not be ergodic, or may even be transient, if the time step h is not carefully chosen. The sensitivity of the LMC to a bad choice of h, combined with the fact that there is no exact convergence to Π, influenced many authors in working on a modified version of the Langevin algorithm, that ensures the convergence in distribution to Π. This algorithm is known as the Metropolis Adjusted Langevin Algorithm (MALA). Unsurprisingly, MALA received a lot of interest during the following years. Several contributions to the analysis of its probabilistic properties were made by Roberts and Rosenthal (1998); Stramer and Tweedie (1999a,b); Roberts and Stramer (2002); Pillai et al. (2012); Eberle (2014); Dwivedi et al. (2018); Eberle and Majka (2019). Several Langevin type algorithms exist, based on other discretizations for

instance. Essentially, it is known that approximate sampling algorithms based on the properties of the continuous-time Langevin process perform well for high dimensional distributions. A proof of this fact will be given in the sequel.

Concerning the LMC, the non-asymptotic study of its theoretical guarantees was left out for a while, until very recently. A founding contribution to quantitative rates for LMC was made by Dalalyan (2017b), who established that under some simple assumptions on f, the mixing time (with respect to the TV distance) of the LMC depends polynomially on the dimension p and the precision level ε. This result was the first of its kind for LMC. Beside solving the problems of transience and non-ergodicity of the chain $(\boldsymbol{\theta}_k)$ by a simple choice of step size $h > 0$, it had the particular merit of providing non-asymptotic convergence rates with explicit dependence in the dimension together with small constants. This contribution aroused a new interest in LMC. Those results were significantly improved and extended to the Wasserstein distance by Durmus and Moulines (2016); Durmus and Moulines (2017); Dalalyan (2017a); Durmus et al. (2018); Dalalyan and Karagulyan (2019). It is noteworthy that those results do not conclude that LMC is more efficient than MALA, but they provide user-friendly bounds on the mixing time of LMC that have no equivalents in MALA's literature.

1.3.4 A convenient framework for quantitative rates

An interesting point from Dalalyan (2017b) was the connection made between the optimization and the sampling task, and the choice of assumptions on f induced by this comparison. This is explained in the sequel. A convenient framework to get explicit rates of convergence, for optimization problems, is to consider maps f that are both smooth and convex. Let us note ∇f for the gradient of f. In what follows, we assume that f is strongly convex and has a Lipschitz gradient, i.e. that for some positive constants m and M we have

$$\boldsymbol{A}_1 : \begin{cases} f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') - \nabla f(\boldsymbol{\theta}')^\top(\boldsymbol{\theta} - \boldsymbol{\theta}') \geq (m/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2, \\ \|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \leq M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \end{cases} \qquad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p.$$

Under assumption $\boldsymbol{A}_1$, the constants m and M are always such that $m \leq M$. Moreover, if the map f is assumed to be twice differentiable with Hessian $\nabla^2 f$, then assumption $\boldsymbol{A}_1$ is equivalent to requiring that $m\mathbf{I}_p \preceq \nabla^2 f(\boldsymbol{\theta}) \preceq M\mathbf{I}_p$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$. This remark brings us to consider a third parameter to measure the difficulty of the problem, for both optimization and sampling. This parameter is defined as $\varkappa \triangleq M/m$ and is called condition number, as it relates to the conditionning of the Hessian matrix $\nabla^2 f$. Essentially, it measures how much the curvature of f may vary when visiting the space $\boldsymbol{\Theta} = \mathbb{R}^p$.

Assuming that $\boldsymbol{A}_1$ holds, it is well known that an explicit convergence rate of the GD algorithm can be derived. This is the point of the following result (Boyd and Vandenberghe, 2004, Eq. 9.18). Assume that f is a continuously differentiable function such that for some positive constants m and M, assumption $\boldsymbol{A}_1$ holds. Let $(\boldsymbol{\vartheta}_k)$ be the deterministic sequence defined in (1.9) with $h = 1/(2M)$, then

$$\|\boldsymbol{\vartheta}_k - \boldsymbol{\theta}^*\|_2^2 \leq \frac{2\left(f(\boldsymbol{\vartheta}_0) - f(\boldsymbol{\theta}^*)\right)}{m}\left(1 - \frac{m}{2M}\right)^k, \qquad \forall k \in \mathbb{N}. \tag{1.12}$$

This shows that the approximation error to the minimizer of f is dimension free, and decreases exponentially fast with the number of iterates k. The initial term $(f(\boldsymbol{\vartheta}_0) - f(\boldsymbol{\theta}^*))$ is bounded by $(M/2)\|\boldsymbol{\vartheta}_0 - \boldsymbol{\theta}^*\|_2^2$. Therefore, if the initial distance to the minimizer is considered as a constant, then for a given precision level $\varepsilon > 0$, the convergence time $\mathcal{K}_\varepsilon^{\text{opt}}$ of the GD algorithm is dimension free and depends logarithmically on ε, but also on $\varkappa$.

Concerning the problem of sampling from Π, the first inequality of $\boldsymbol{A}_1$ requires the density $\pi(\boldsymbol{\theta}) \propto e^{-f(\boldsymbol{\theta})}$ to be strongly log-concave. For a detailed review on log-concave probability distributions and their properties, see Saumard and Wellner (2014). A main property is that a strongly log-concave distribution is necessarily unimodal and sub-Gaussian. A legitimate claim is that it is a major restriction in the space of probability distributions. Another legitimate claim is that, if one considers the optimization problem, the strong convexity of f is also a major restriction in the space of continuously differentiable functions. Just as most of geometric rates of convergence in optimization hold for convex functions, most results of geometric ergodicity in approximate sampling hold for sub-exponential distributions. Therefore, if one hopes to get explicit rates of convergence to Π, some assumption close to $\boldsymbol{A}_1$ will be required, essentially.

Concerning the Langevin diffusion process, we will see that assumption $\boldsymbol{A}_1$ is very convenient in many ways. A first remark is that if the second inequality of $\boldsymbol{A}_1$ holds, the diffusion process (1.11) has a unique strong solution $\{\boldsymbol{L}_t : t \geq 0\}$ which is a Markov process. Moreover, if assumption $\boldsymbol{A}_1$ holds and if $\mathscr{D}(\boldsymbol{L}_0)$ is absolutely continuous with respect to Lebesgue measure, then this diffusion Markov process is also geometrically ergodic, in the following sense (Dalalyan, 2017b, Lemma1):

$$\|\mathscr{D}(\boldsymbol{L}_t) - \Pi\|_{\text{TV}} \leq \frac{1}{2}\chi^2(\mathscr{D}(\boldsymbol{L}_0)\|\Pi)^{1/2} e^{-mt/2}, \qquad \forall t \geq 0. \tag{1.13}$$

In particular, for the initial distribution $\mathscr{D}(\boldsymbol{L}_0) = \mathcal{N}(\boldsymbol{\theta}^*, M^{-1}\mathbf{I}_p)$ then it can be shown that $\chi^2(\mathscr{D}(\boldsymbol{L}_0)\|\Pi) \leq \exp\{(p/2)\log(M/m)\}$. Therefore, in such a case for any $\varepsilon > 0$, the Langevin diffusion process after a time

$$T \geq \frac{2}{m}\left\{\frac{p}{4}\log\left(\frac{M}{m}\right) + \log\left(\frac{1}{2\varepsilon}\right)\right\},$$

will be such that $\|\mathscr{D}(\boldsymbol{L}_T) - \Pi\|_{\text{TV}} \leq \varepsilon$. In other words, the mixing time of the Langevin diffusion process for the TV distance depends linearly on the dimension p and logarithmically on ε and $\varkappa$. This result was improved by Durmus and Moulines (2016); Durmus and Moulines (2017), who showed that the dependence in the dimension p is actually logarithmic, for both the TV distance and the 2-Wasserstein distance, and for any deterministic starting point $\boldsymbol{L}_0 \in \mathbb{R}^p$. Obviously, the mixing time of the Langevin diffusion process is of little interest from the user's perspective since it cannot be sampled exactly. However, those results highlight the fact that an approximate sampling algorithm based on a discretization of (1.11) is expected to perform well for high dimensional distributions, provided that the discretization error has also polynomial dependence in the dimension.

For LMC, the discretization error was controlled by choosing a time step h small enough that has polynomial dependence on p and ε. Up to a logarithmic factor, the mixing time then scales as the inverse of h. We present hereafter a summary of the

best known rates for the mixing time of LMC, for the TV distance, Kullback-Leibler divergence and the Wasserstein-2 distance. Those rates are the results of several works already quoted. We emphasize here the contribution of Durmus et al. (2018), who recently improved the dependence on the condition number $\varkappa$. It is noteworthy that this result was made possible for TV and Kullback-Leibler by considering a slightly different sampling estimate, which relies on picking at random an element of the LMC chain after a burn-in period of the same order as the mixing time. As the corresponding distribution is an average of measures on the LMC chain, we refer to this sampling estimate as LMC with averaging.

Distance	Total variation	Kullback-Leibler	Wasserstein-2
Mixing time	$\varkappa p/\varepsilon^2$ △	$\varkappa p/\varepsilon$ △	$\varkappa p/\varepsilon^2$

Table 1.1: Scaling of LMC mixing time with $p, \varepsilon, \varkappa$, up to logarithmic factors. The results indicated by △ describe the behavior of LMC with averaging.

Several extensions of those results were studied; for instance Durmus and Moulines (2016) improved the dependence on the precision level for the Wasserstein-2 distance under further smoothness assumption, i.e. when the Hessian $\nabla^2 f$ is Lipschitz for the spectral norm. Dalalyan and Karagulyan (2019) proposed an extension of LMC to a case where the gradient ∇f is not known accurately. Both works also studied a varying step size approach of the LMC, proved optimal for a step size h of order $1/k$ at the k-th iterate, that essentially allows to get rid of a logarithmic factor in the mixing time. Several works also tried to establish similar guarantees when the potential function f is convex, but not strongly convex, that is when $m = 0$ in the first equation of $\boldsymbol{A}_1$. Among them, several isolated results can be found for instance in Dalalyan (2017b); Durmus and Moulines (2017); Durmus et al. (2018). Chapter 4 is an attempt to extend and unify those results into a common framework. It emphasizes the fact that the mixing time will scale polynomially with the dimension p if the moments of Π also scale polynomially with p, therefore additional assumptions on f are required. The literature for LMC is wide and this is by no means an exhaustive review of the works dedicated to this topic. In a different but related direction, a recent contribution by Cheng et al. (2018) established explicit non asymptotic guarantees for another algorithm, called Kinetic Langevin Monte Carlo. This is the main topic of the following section.

1.3.5 Kinetic Langevin Monte Carlo

The so called Kinetic Langevin Monte Carlo (KLMC) algorithm (a.k.a underdamped Langevin MCMC) is an approximate sampling algorithm, which belongs to a wide family of sampling algorithms known as Hamiltonian Monte Carlo (HMC). For a detailed explanation on HMC algorithms and their origin, see Duane et al. (1987); Neal (2011). As emphasized by Eberle et al. (2017), HMC algorithms are known to perform better than other MCMC algorithms for high dimensional distributions, although those observations were lacking of theoretical justification. Among HMC algorithms, KLMC has the particularity of being the first for which computations made possible to derive similar results as for LMC, i.e. explicit non asymptotic guarantees with small constants over the mixing

time. The KLMC algorithm itself was introduced by Cheng et al. (2018), which relies on a smart discretization of the so called kinetic Langevin diffusion process, defined by

$$d\begin{bmatrix}\boldsymbol{V}_t\\ \boldsymbol{L}_t\end{bmatrix} = \begin{bmatrix}-(\gamma \boldsymbol{V}_t + u\nabla f(\boldsymbol{L}_t))\\ \boldsymbol{V}_t\end{bmatrix} dt + \sqrt{2\gamma u}\begin{bmatrix}\mathbf{I}_p\\ \mathbf{0}_{p\times p}\end{bmatrix} d\boldsymbol{W}_t, \qquad t \geq 0, \tag{1.14}$$

where $\gamma > 0$ is the friction coefficient and $u > 0$ is the inverse mass. Under the same assumption $\boldsymbol{A}_1$ as for the Langevin diffusion, the continuous-time Markov process $(\boldsymbol{L}_t, \boldsymbol{V}_t)$ is also positive recurrent, while its invariant distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{2p}$. The corresponding invariant density is given by

$$p_*(\boldsymbol{\theta}, \boldsymbol{v}) \propto \exp\Big\{ - f(\boldsymbol{\theta}) - \frac{1}{2u}\|\boldsymbol{v}\|_2^2\Big\}, \qquad \boldsymbol{\theta} \in \mathbb{R}^p,\ \boldsymbol{v} \in \mathbb{R}^p. \tag{1.15}$$

In other words, under the invariant distribution, the components $\boldsymbol{L}$ and $\boldsymbol{V}$ are independent, $\boldsymbol{L}$ is distributed according to the target $\pi(\boldsymbol{\theta}) \propto e^{-f(\boldsymbol{\theta})}$, whereas $\boldsymbol{V}/\sqrt{u}$ is a standard Gaussian vector. To target a probability distribution through its product density with a Gaussian distribution on the extended space $\mathbb{R}^{2p}$ is actually a common feature of HMC algorithms. The diffusion process (1.14) cannot be sampled exactly, but a sampler can be derived from its discretization on $\mathbb{R}^{2p}$. The discretized version of the first component $(\boldsymbol{L}_t)$ can therefore be used for solving the problem of sampling from Π.

The algorithm of Cheng et al. (2018) will not be described here, but it relies essentially on approximating the term $\nabla f(\boldsymbol{L}_t)$ by $\nabla f(\boldsymbol{L}_0)$ in the stochastic differential equation (1.14). The key property of the new process induced by this approximation is that it is a Gaussian process with an explicit solution, that allows the computation of the conditional distribution of $(\boldsymbol{L}_t, \boldsymbol{V}_t)$ given $(\boldsymbol{L}_0, \boldsymbol{V}_0)$. By choosing a step size $h > 0$, and repeating this discretization principle at each step, one gets a tractable sequence of conditional Gaussian distributions and therefore a sampling algorithm.

As for p-dimensional Langevin type algorithms, the quality of the resulting sampler will depend on two key properties: the rate of mixing of (1.14) and the error induced by its discretization. A main contribution to the rate of mixing of kinetic diffusions was made by Eberle et al. (2017), under conditions that are more general than strong convexity of f. Cheng et al. (2018) took care of the second part by proving that their algorithm leads to a sampler that achieves a mixing time's scaling of $\varkappa^2 p^{1/2}/\varepsilon$ with respect to the Wasserstein-2 distance.

In other words, KLMC based on (1.14) converges faster than the standard LMC based on (1.11), at least with respect to the dimension p and the precision level ε. Emphasized in Chapter 3, this improved rate of convergence is mainly due to the higher smoothness of sample paths of the process $\{\boldsymbol{L}_t : t \geq 0\}$, that essentially leads to a smaller discretization error. However, the quadratic dependence on the condition number $\varkappa$ for KLMC is apparently worse than the linear dependence proven by Durmus et al. (2018) for LMC. In other words, from the results we have to this day, it is not impossible that LMC may behave better for very badly conditioned problems. Though, such a claim should be placed in the context of a new field of study. Indeed, the linear dependence proven by Durmus et al. (2018) is very recent, and therefore led Cheng et al. (2018) to formulate as an open question whether the dependence on $\varkappa$ could be improved to linear for KLMC. Chapter 3 answers partially to this question by proving that the dependence is improvable

to $\varkappa^{3/2}$. Nevertheless, a lack of lower bounds and optimality results in this field give the intuition that it is a bit early to draw conclusions on this fact.

A significant part of this thesis is actually devoted to the study of Kinetic Langevin Monte Carlo properties and extensions. In particular, Chapter 3 improves the results from Cheng et al. (2018) and proposes a second order algorithm whose mixing time's scaling is improved by a factor ε under further smoothness assumption, i.e. when the Hessian $\nabla^2 f$ is Lipschitz for the spectral norm. In another direction, an interesting question is whether quantitative results can be established for KLMC when the potential function f that is convex but not strongly convex, this topic is studied in Chapter 4.

1.4 Summary of the contributions

Chapter 2 presents a joint work with Nicolas Chopin, devoted to the inference of un-normalized statistical models. Chapter 3 is a joint work with Arnak Dalalyan, while Chapter 4 is a joint work with both Arnak Dalalyan and Avetik Karagulyan. Those two works relates to approximate sampling for high dimensional distributions, and in particular their non asymptotic guarantees.

1.4.1 Inference of un-normalized statistical models

In statistics and machine learning, dependency modelling is an important field of research. Statistical models can now be designed for handling very general dependence structures (e.g. graphical models, networks, spatial point processes). Inference in many of these models appears to be challenging, because they involve intractable likelihoods. Among others, these include problems caused by non computable "normalizing constants". Partition functions in graphical models are an example. Those "un-normalized" statistical models have posed a serious challenge for both bayesian and frequentist inference. This field raises the following central question: when a statistical model involves an intractable partition function, how can we infer this model, and how well? This problem was studied by several authors, e.g. Geyer (1994), Gutmann and Hyvärinen (2010), Barthelmé and Chopin (2015). The main solutions involve sampling methods. Un-normalized models are usually avoided by practitioners due to their difficulty for estimation. With the rise of computational techniques and power, research on intractable normalizing constants is useful in proposing and assessing new inferential methods.

Several inferential methods have been proposed to tackle this issue, for the most based on Monte-Carlo methods. Among them, we study some approximations of the Maximum Likelihood Estimator by sampling methods, especially the following two: a standard method called Monte-Carlo MLE (Geyer, 1994), and a more recent procedure, known as Noise Contrastive Estimation (Gutmann and Hyvärinen, 2010). Several works (Gutmann and Hyvärinen, 2010; Gutmann and Hyvärinen, 2012), based on simulations, observed that this new method is more stable, but no theoretical results had yet been proven. We manage to turn this observation into a formal result: we prove that this new estimation method is more robust to a bad choice of sampling distribution especially for

large observed datasets. We complete those results by a numerical comparison, assessing the gain of performance depending: on the distance between sampling and data distributions, and on the computational budget. Chapter 2 was published in Electronic Journal of Statistics in 2018 (Riou-Durand and Chopin, 2018). From a mathematical perspective, Chapter 2 employs several concepts such as: asymptotic properties of M-estimators, geometric ergodicity of MCMC samplers, Central Limit Theorems for Markov Chains.

1.4.2 Quantitative results for high dimensional sampling

High-dimensional models are an important focus of statistical research, emphasized in the present era of "big-data". In both frequentist and Bayesian settings, high dimension remains an important inferential issue. It appears so challenging that it is often referred as a "curse of dimensionality". For instance, Bayesian inference often requires sampling from a high dimensional posterior distribution. But it is known that the performance of most sampling schemes deteriorates exponentially fast when the dimension increases. Several approximate sampling procedures have proven their effectiveness into tackling this issue (e.g. Sequential Monte Carlo, Hamiltonian Monte Carlo, Langevin Monte Carlo). Among them, the (unadjusted) Langevin's algorithm was the first for which researchers were able to provide true non-asymptotic guarantees over the approximation error. Some recent results made a particular focus on this research area (Dalalyan, 2017b; Durmus and Moulines, 2016), as they showed that under some strong conditions over the distribution, Langevin's "mixing time" (number of iterations needed to reach a given accuracy) grows polynomially fast (only), when the dimension increases. Researchers are currently trying to extend these results to other algorithms, under milder hypotheses. These non-asymptotic studies are very useful in providing guaranties and guidance to efficient high dimensional sampling procedures.

Following several articles on the mixing time of Langevin Monte Carlo for strongly log concave distributions (Dalalyan, 2017b; Durmus and Moulines, 2016), a new sampling algorithm was very recently proposed by Cheng et al. (2018). The authors showed that under the same assumptions, the mixing time of this algorithm with respect to the Wasserstein-2 distance, improves the one of Langevin Monte Carlo, in terms of dependence to both dimension and precision level. This sampling method, referred as Kinetic Langevin Monte Carlo (a.k.a. underdamped Langevin MCMC), can be viewed as a special case of Hamiltonian Monte Carlo. It was to our knowledge the first non-asymptotic result of this type for HMC. Chapter 3 is devoted to provide a better understanding of the properties of Kinetic Langevin Monte Carlo. In addition to that, we show that the mixing time dependence on regularity parameters is improvable, compared to Cheng et al. (2018), together with numerical constants. Finally, we design a second-order algorithm which, under stronger regularity assumptions, leads to an improved mixing time, in terms of dependence to the precision level. Chapter 3 was submitted to Bernoulli in 2018 (Dalalyan and Riou-Durand, 2018). From a mathematical perspective, Chapter 3 employs several concepts such as: Langevin diffusions, Couplings, Wasserstein distance.

Another interesting direction in the literature of approximate sampling is whether quantitative results can be established under milder hypotheses than adopted by Dalalyan (2017b); Durmus and Moulines (2016); Durmus and Moulines (2017). More precisely, a

main question is whether polynomial rates of convergence can still be obtained when the sampling distribution is log concave, but not strongly. For LMC, several contributions answered partially to this question (e.g. Dalalyan (2017b); Durmus and Moulines (2017); Durmus et al. (2018)). Chapter 4 is devoted to extend those results, but also to establish their counterpart for KLMC. We establish non asymptotic bounds on the mixing time of LMC and KLMC (first and second order), with respect to Wasserstein-q distances and the bounded-Lipschitz distance. Those bounds depend on the knowledge of moments of the sampling distribution. We provide several conditions that allows the user to bound those moments, and obtain at the same time polynomial rates with the dimension. To this date, Chapter 4 is still a working paper. From a mathematical perspective, Chapter 4 employs several concepts such as: Langevin diffusions, Moments of log concave distributions, Wasserstein and bounded-Lipschitz distances.

1.5 Résumé substantiel des contributions

Le Chapitre 2 présente un travail réalisé conjointement avec Nicolas Chopin, consacré à l'inférence de modèles statistiques non-normalisés. Le Chapitre 3 est un travail réalisé conjointement avec Arnak Dalalyan, tandis que le Chapitre 4 est un travail réalisé conjointement avec Arnak Dalalyan et Avetik Karagulyan. Ces deux travaux portent sur les méthodes d'échantillonnage aléatoire approché pour des distributions de grande dimension, et en particulier leurs garanties non-asymptotiques.

1.5.1 Inférence des modèles statistiques non-normalisés

En statistique et en apprentissage, la modélisation de la dépendance est un domaine de recherche important. Des modèles statistiques sont désormais conçus pour traiter des structures complexes de dépendance (par exemple des modèles graphiques, des réseaux, des processus ponctuels spatiaux). L'inférence dans la plupart de ces modèles est difficile à réaliser, car ils impliquent des fonctions de vraisemblance incalculables. Ces problèmes sont parfois causés par des "constantes de normalisation" non calculables. Les fonctions de partition dans les modèles graphiques en sont un exemple. Ces modèles statistiques "non-normalisés" sont un défi pour l'inférence Bayésienne comme fréquentiste. Ce champ de recherche tente de répondre à la question suivante : lorsqu'un modèle statistique implique une fonction de partition incalculable, comment peut-on estimer ce modèle, et à quel coût ? Ce problème a été étudié par plusieurs auteurs, e.g. Geyer (1994), Gutmann and Hyvärinen (2010), Barthelmé and Chopin (2015). Les modèles non-normalisés sont généralement évités par les praticiens en raison de leur difficulté d'estimation. Avec l'essor de la puissance de calcul et des méthodes computationnelles en statistique, leur inférence devient envisageable. La recherche sur les problèmes de constantes de normalisation incalculables est utile pour proposer et évaluer de nouvelles méthodes d'inférence. Plusieurs méthodes ont été proposées pour pallier ces difficultés, la plupart basées sur les méthodes de Monte-Carlo. Parmi celles-ci, nous étudions certaines approximations de l'estimateur du Maximum de Vraisemblance, en particulier les deux suivantes : une méthode standard appelée *Monte-Carlo MLE* (Geyer, 1994), et une méthode plus récente, appelée *Noise Contrastive Estimation* (Gutmann and Hyvärinen, 2010). Plusieurs travaux (Gutmann and Hyvärinen, 2010; Gutmann and Hyvärinen, 2012), basés sur des expérimentations numériques, ont observé que cette nouvelle méthode est plus stable, mais aucun résultat théorique n'avait encore été prouvé. Le Chapitre 2 répond à cette question : nous prouvons que cette nouvelle méthode d'estimation est plus robuste à un mauvais choix de distribution d'échantillonnage, surtout lorsque le nombre de données observées est grand. Nous complétons ces résultats par une étude numérique, évaluant le gain de performance en fonction de la distance entre la distribution d'échantillonnage et la distribution des données, et du budget computationnel. Le Chapitre 2 a été publié dans *Electronic Journal of Statistics* en 2018 (Riou-Durand and Chopin, 2018). D'un point de vue mathématique, le Chapitre 2 utilise plusieurs concepts comme : les propriétés asymptotiques des M-estimateurs, l'ergodicité géométrique des algorithmes MCMC, les théorèmes de la limite centrale pour les chaînes de Markov.

1.5.2 Résultats quantitatifs pour l'échantillonnage en grande dimension

Les modèles statistiques de grande dimension forment un centre d'intérêt majeur pour la recherche en statistique, particulièrement dans l'ère actuelle du "big data". En statistique fréquentiste comme Bayésienne, la grande dimension demeure un important problème pour l'inférence. Les problèmes posés par la grande dimension sont si difficile qu'on les associent souvent à une "malédiction de la dimension". Par exemple, l'inférence Bayésienne passe souvent par l'échantillonnage aléatoire d'une loi a posteriori de grande dimension. Plusieurs procédures d'échantillonnage approché ont prouvé leur efficacité dans la résolution de ce problème (e.g. *Sequential Monte Carlo*, *Hamiltonian Monte Carlo*, *Langevin Monte Carlo*). Parmi eux, l'algorithme *Langevin Monte Carlo* (LMC) fut le premier pour lequel les chercheurs ont été en mesure de fournir de véritables garanties non-asymptotiques sur l'erreur d'approximation. Certains résultats récents ont particulièrement développé cette ligne de recherche (Dalalyan, 2017b; Durmus and Moulines, 2016) en montrant, sous certaines conditions sur la distribution d'échantillonnage, que le *mixing time* de LMC (nombre d'itérations nécessaires pour atteindre une précision donnée) a une dépendance polynomiale (seulement) en la dimension. Les chercheurs tentent actuellement d'étendre ces résultats à d'autres algorithmes, sous des hypothèses plus faibles.

Suite à plusieurs articles sur le *mixing time* de LMC pour les distributions fortement log-concave (Dalalyan, 2017b; Durmus and Moulines, 2016), un nouvel algorithme d'échantillonnage a été proposé très récemment par Cheng et al. (2018). Les auteurs ont montré que, sous les mêmes hypothèses, le *mixing time* de cet algorithme par rapport à la distance de Wasserstein-2, améliore le *mixing time* de LMC, en termes de sa dépendance en la dimension et le niveau de précision. Cette méthode d'échantillonnage, appelée *Kinetic Langevin Monte Carlo* (KLMC) (alias *underdamped Langevin MCMC*), peut être considérée comme un cas particulier de *Hamiltonian Monte Carlo* (HMC). A notre connaissance, c'est le premier résultat non-asymptotique de ce type pour HMC. Le chapitre 3 est consacré à une meilleure compréhension des propriétés de KLMC. En outre, nous montrons qu'il est possible d'améliorer: la dépendance du *mixing time* en les paramètres de régularité de la loi cible, ainsi que les constantes numériques. Enfin, nous proposons un algorithme de second ordre qui, dans le cadre d'hypothèses de régularité plus fortes, conduit à un meilleur *mixing time*, en termes de dépendance par rapport au niveau de précision. Le Chapitre 3 a été soumis à Bernoulli en 2018 (Dalalyan and Riou-Durand, 2018). D'un point de vue mathématique, le Chapitre 3 utilise plusieurs concepts tels que : les processus de diffusions de Langevin, les couplages, la distance de Wasserstein.

Une autre ligne de recherche intéressante dans la littérature sur l'échantillonnage approché concerne les résultats quantitatifs sous des hypothèses plus faibles que celles adoptées par Dalalyan (2017b); Durmus and Moulines (2016); Durmus and Moulines (2017). Plus précisément, une question principale est de savoir si une vitesse de convergence polynomiale en la dimension peut encore être obtenue lorsque la distribution d'échantillonnage est log concave, mais pas fortement. Pour le LMC, plusieurs contributions ont répondu partiellement à cette question e.g. Dalalyan (2017b); Durmus and Moulines (2017); Durmus et al. (2018). Le Chapitre 4 est consacré à l'extension de ces

résultats, mais aussi au développement de leurs analogues pour KLMC. Nous établissons des bornes non-asymptotiques sur le *mixing time* de LMC et KLMC (premier et deuxième ordre), pour les distances Wasserstein-q et bounded-Lipschitz. Ces bornes dépendent des moments de la distribution d'échantillonnage. Nous proposons plusieurs conditions qui permettent à l'utilisateur de borner ces moments, et d'obtenir ainsi une vitesse polynomiale en la dimension. A ce jour, le Chapitre 4 est encore un travail en cours. D'un point de vue mathématique, le Chapitre 4 utilise plusieurs concepts tels que : les processus de diffusions de Langevin, les moments de distributions log-concave, les distances de Wasserstein et bounded-Lipschitz.

Part I

Inference of un-normalized statistical models

Chapter 2

Noise contrastive estimation: asymptotic properties, formal comparison with MC-MLE

A statistical model is said to be un-normalised when its likelihood function involves an intractable normalising constant. Two popular methods for parameter inference for these models are MC-MLE (Monte Carlo maximum likelihood estimation), and NCE (noise contrastive estimation); both methods rely on simulating artificial data-points to approximate the normalising constant. While the asymptotics of MC-MLE have been established under general hypotheses (Geyer, 1994), this is not so for NCE. We establish consistency and asymptotic normality of NCE estimators under mild assumptions. We compare NCE and MC-MLE under several asymptotic regimes. In particular, we show that, when $m \to \infty$ while n is fixed (m and n being respectively the number of artificial data-points, and actual data-points), the two estimators are asymptotically equivalent. Conversely, we prove that, when the artificial data-points are IID, and when $n \to \infty$ while m/n converges to a positive constant, the asymptotic variance of a NCE estimator is always smaller than the asymptotic variance of the corresponding MC-MLE estimator. We illustrate the variance reduction brought by NCE through a numerical study.

2.1 Introduction

Consider a set of probability densities $\{f_\theta : \theta \in \Theta\}$ with respect to some measure μ, defined on a space $\mathcal{X}$, such that:

$$f_\theta(x) = \frac{h_\theta(x)}{\mathcal{Z}(\theta)}$$

where h_θ is non-negative, and $\mathcal{Z}(\theta)$ is a normalising constant, $\mathcal{Z}(\theta) = \int_{\mathcal{X}} h_\theta(x)\mu(\mathrm{d}x)$. A model based on such a family of densities is said to be un-normalised if function h_θ may be computed point-wise, but $\mathcal{Z}(\theta)$ is not available (i.e. it may not be computed in a reasonable CPU time).

Un-normalised models arise in several areas of machine learning and Statistics, such

as deep learning (Salakhutdinov and Hinton, 2009), computer vision (Wang et al., 2013), image segmentation (Gu and Zhu, 2001), social network modelling (Caimo and Friel, 2011), directional data modelling (Walker, 2011), among others. In most applications, data-points are assumed to be IID (independent and identically distributed); see however e.g. Mnih and Teh (2012) or Barthelmé and Chopin (2015) for applications of non-IID un-normalised models. In that spirit, we consider an un-normalised model of IID variables $Y_1, \ldots, Y_n$, with log-likelihood (divided by n):

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log h_\theta(y_i) - \log \mathcal{Z}(\theta). \tag{2.1}$$

The fact that $\mathcal{Z}(\theta)$ is intractable precludes standard maximum likelihood estimation.

Geyer (1994) wrote a seminal paper on un-normalised models, in which he proposed to estimate θ by maximising function

$$\ell^{\mathrm{IS}}_{n,m}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{h_\theta(y_i)}{h_\psi(y_i)} - \log\left\{\frac{1}{m}\sum_{j=1}^{m} \frac{h_\theta(x_j)}{h_\psi(x_j)}\right\} \tag{2.2}$$

where the x_j's are m artificial data-points generated from a user-chosen distribution $\mathbb{P}_\psi$ with density $f_\psi(x) = h_\psi(x)/\mathcal{Z}(\psi)$. Although notation $\mathbb{P}_\psi$ suggests that the distribution of the artificial data-points belongs to the considered parametric model, this is not compulsory. The only required assumption is that the model is dominated by $\mathbb{P}_\psi$ (i.e. $\mathbb{P}_\theta \ll \mathbb{P}_\psi$ for every $\theta \in \Theta$). The empirical average inside the second log is a consistent (as $m \to \infty$) importance sampling estimate of $\mathcal{Z}(\theta)/\mathcal{Z}(\psi)$. Function $\ell^{\mathrm{IS}}_{n,m}$ is thus an approximation of the log-likelihood ratio $\ell_n(\theta) - \ell_n(\psi)$, whose maximiser is the MLE.

In many applications, the easiest way to sample from $\mathbb{P}_\psi$ is to use MCMC (Markov chain Monte Carlo). Geyer (1994) established the asymptotic properties of the MC-MLE estimates under general conditions; in particular that the x_j's are realisations of an ergodic process. This is remarkable, given that most of the theory on M-estimation (i.e. estimation obtained by maximising functions) is restricted to IID data.

More recently, Gutmann and Hyvärinen (2012) proposed an alternative approach to parameter estimation of un-normalised models, called noise contrastive estimation (NCE). It also relies on simulating artificial data-points $x_1, \ldots, x_m$ from distribution $\mathbb{P}_\psi$. The method consists in maximising the likelihood of a logistic classifier, where actual (resp. artificial) data-points are assigned label 1 (resp. 0). With symbols, the log-likelihood divided by n rewrites:

$$\ell^{\mathrm{NCE}}_{n,m}(\theta, \nu) = \frac{1}{n}\sum_{i=1}^{n} \log q_{\theta,\nu}(y_i) + \frac{1}{m}\sum_{i=1}^{m} \log\left\{(1 - q_{\theta,\nu}(x_i))^{m/n}\right\} \tag{2.3}$$

where $q_{\theta,\nu}(x)$, the probability of label 1 for a value x, is defined through odd-ratio function:

$$\log\left\{\frac{q_{\theta,\nu}(x)}{1 - q_{\theta,\nu}(x)}\right\} = \log\left\{\frac{h_\theta(x)}{h_\psi(x)}\right\} + \nu + \log\left(\frac{n}{m}\right).$$

The NCE estimator of θ is obtained by maximising function $\ell^{\mathrm{NCE}}_{n,m}(\theta, \nu)$ with respect to both $\theta \in \Theta$ and $\nu \in \mathbb{R}$. In particular, when the considered model is exponential, i.e.

when $h_\theta(x) = \exp\left\{\theta^T S(x)\right\}$, for some statistic S, $\ell_{n,m}^{\text{NCE}}$ is the log-likelihood of a standard logistic regression, with covariate $S(x)$. In that case, implementing NCE is particularly straightforward.

This paper has two objectives: first, to establish the asymptotic properties of NCE when the artificial data-points are generated from an ergodic process (typically a MCMC sampler) in order to show that NCE is as widely applicable as MC-MLE; second, to compare the statistical efficiency of both methods.

As a preliminary step, we replace the original log-likelihood by a function defined on the extended space $\Theta \times \mathbb{R}$, called Poisson transform:

$$\ell_n(\theta, \nu) = \frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{h_\theta(y_i)}{h_\psi(y_i)}\right\} + \nu - e^\nu \times \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)}. \tag{2.4}$$

This function is so called as it corresponds to the log-likelihood (up to a linear transformation) of a Poisson process with intensity $h_\theta(y) + \nu$, see Barthelmé and Chopin (2015) for details. The main property of this transformation is that it produces exactly the same MLE as the original likelihood: $(\widehat{\theta}_n, \widehat{\nu}_n)$ maximises (2.4) if and only if $\widehat{\theta}_n$ maximises (2.1) and $\widehat{\nu}_n = \log\left\{\mathcal{Z}(\psi)/\mathcal{Z}(\widehat{\theta}_n)\right\}$.

In the same way, we replace the MC-MLE log-likelihood by function

$$\ell_{n,m}^{\text{IS}}(\theta, \nu) = \frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{h_\theta(y_i)}{h_\psi(y_i)}\right\} + \nu - \frac{e^\nu}{m}\sum_{j=1}^{m} \frac{h_\theta(x_j)}{h_\psi(x_j)} \tag{2.5}$$

which has the same maximiser (with respect to θ) as function (2.2).

We thus obtain three objective functions defined with respect to the same parameter space, $\Theta \times \mathbb{R}$. This will greatly facilitate our analysis. The paper is organised as follows. In Section 2.2, we introduce the set up and notations. In Section 2.3, we study the behaviour of the NCE estimator as $m \to \infty$ (while n is kept fixed). We prove that the NCE estimator converges to the MLE at the same $m^{-1/2}$ rate as the MC-MLE estimator, and the difference between the two estimators converges faster, at rate m^{-1}. In Section 2.4, we let both m and n go to infinity while $m/n \to \tau > 0$. We obtain asymptotic variances for both estimators, which admit a simple and interpretable decomposition. Using this decomposition, we are able to establish that when the artificial data-points are IID, the asymptotic variance of NCE is always smaller than the asymptotic variance of MC-MLE (for the same computational budget). Section 2.5 assesses this variance reduction in a numerical example. Section 2.6 discusses the practical implications of our results. All proofs are delegated to the appendix.

2.2 Set-up and notations

Unless explicitly stated, we will consider Θ to be an open subset of $\mathbb{R}^d$, with natural topology associated to the Euclidian norm. We consider a parametric statistical model $\{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\}$, corresponding to n IID data-points lying in space $\mathcal{X} \subset \mathbb{R}^k$, associated with the corresponding Borel σ-Field. We assume that the model is identifiable, and equipped

with some dominating measure μ, inducing the log-likelihood (2.1). From now on, we work directly with the "extended" version of approximate and exact log-likelihoods, i.e. functions (2.3), (2.4) and (2.5), which are functions of extended parameter $\xi = (\theta, \nu)$, with $\xi \in \Xi = \Theta \times \mathbb{R}$. When convenient, we also write $\ell_n(\xi)$ for $\ell_n(\theta, \nu)$ and so on. An open ball in Ξ, centered on ξ and of radius ϵ, is denoted $B(\xi, \epsilon)$. We may also use this notation for balls in Θ.

The point of this paper is to study and compare point estimates $\widehat{\xi}^{\mathrm{IS}}_{n,m}$ and $\widehat{\xi}^{\mathrm{NCE}}_{n,m}$, which maximise functions (2.5) and (2.3). For the sake of generality, we allow these estimators to be approximate maximisers; i.e. we will refer to $\widehat{\xi}^{\mathrm{IS}}_{n,m}$ as an approximate MC-MLE if

$$\ell^{\mathrm{IS}}_{n,m}(\widehat{\xi}^{\mathrm{IS}}_{n,m}) \geq \sup_{\xi \in \Xi} \ell^{\mathrm{IS}}_{n,m}(\xi) - o(1) \quad \text{a.s.} \tag{2.6}$$

and with a similar definition for $\widehat{\xi}^{\mathrm{NCE}}_{n,m}$. The meaning of symbol $o(1)$ in (2.6) depends on the asymptotic regime: in Section 2.3, n is kept fixed, while $m \to \infty$, hence $o(1)$ means "*converges to zero as* $m \to \infty$". In Section 2.4, both m and n go to infinity, and the meaning of $o(1)$ must be adapted accordingly.

In both asymptotic regimes, the main assumption regarding the sampling process is as follows.

(X1) The artificial data-points are realisations of a $\mathbb{P}_\psi-$ergodic process $(X_j)_{j \geq 1}$.

By $\mathbb{P}_\psi-$ergodicity, we mean that the following law of large number holds:

$$\frac{1}{m} \sum_{j=1}^{m} \varphi(X_j) \underset{m \to \infty}{\to} \mathbb{E}_\psi \left[\varphi(X)\right] = \int_{\mathcal{X}} \varphi(x) f_\psi(x) \mu(\mathrm{d}x) \quad \text{a.s.}$$

for any measurable, real-valued function φ such that $\mathbb{E}_\psi \left[|\varphi(X)|\right] < +\infty$.

Assumption (X1) is mild. For instance, if the X_j's are generated by a MCMC algorithm, this is equivalent to assuming that the simulated chain is aperiodic and ϕ-irreducible, which is true for all practical MCMC samplers; see e.g. Roberts and Rosenthal (2004).

2.3 Asymptotics of the Monte Carlo error

In this section, the analysis is conditional on the observed data: n and $y_1, ..., y_n$ are fixed. The only source of randomness is the Monte Carlo error, and the quantity we seek to estimate is the (intractable) MLE. This regime was first studied for MC-MLE by Geyer (1994). For convenience, we suppose that the MLE exists and is unique; or equivalently that $\widehat{\xi}_n = (\widehat{\theta}_n, \widehat{\nu}_n)$ is the unique maximiser of ℓ_n.

2.3.1 Consistency

We are able to prove NCE consistency (towards the MLE) using the same approach as Geyer (1994) for MC-MLE. Our consistency result relies on the following assumptions:

(C1) The random sequence $\left(\widehat{\xi}^{\mathrm{NCE}}_{n,m}\right)_{m\geq 1}$ is an approximate NCE estimator, which belongs to a compact set almost surely.

(H1) The maps $\theta \mapsto h_\theta(x)$ are:

1. lower semi-continuous at each $\theta \in \Theta$, except for x in a $\mathbb{P}_\psi$-null set that may depend on θ;
2. upper semi-continuous, for any x not in a $\mathbb{P}_\psi$-null set (that does not depend on θ), and for all $x = y_i$, $i = 1, \ldots, n$.

Theorem 1. *Under assumptions (X1), (C1) and (H1), almost surely:* $\widehat{\xi}^{\mathrm{NCE}}_{n,m} \underset{m\to\infty}{\to} \widehat{\xi}_n$.

This result is strongly linked to Theorems 1 and 4 of Geyer (1994), which state that $\widehat{\theta}^{\mathrm{IS}}_{n,m} \to \widehat{\theta}_n$ as $m \to \infty$ under essentially the same assumptions. These assumptions are very mild: they basically require continuity of the maps $\theta \mapsto h_\theta(x)$, without any integrability condition.

Remark 1. *As noticed by Geyer (1994), the proof does not require Θ to be a subset of $\mathbb{R}^d$, consistency of MC-MLE as well as Theorem 1 hold more generally as soon as Θ is a separable metric space.*

2.3.2 Asymptotic normality, comparison with MCMC-MLE

In order to compare the Monte Carlo error of MC-MLE and NCE estimators, we make the following extra assumptions:

(H2) The maps $\theta \mapsto h_\theta(x)$ are twice continuously differentiable in a neighborhood of $\widehat{\theta}_n$ for $\mathbb{P}_\psi-$almost every x, and for $x = y_i$, $i = 1, \ldots, n$. The Hessian matrix $\mathbf{H} = \nabla^2 \ell_n(\widehat{\theta}_n)$ is invertible. Moreover, for some $\varepsilon > 0$

$$\int_{\mathcal{X}} a_\varepsilon(x) \sup_{\theta \in B(\widehat{\theta}_n, \varepsilon)} h_\theta(x) \mu(\mathrm{d}x) < +\infty$$

where $a_\varepsilon(x) = 1 + \sup_{\theta \in B(\widehat{\theta}_n, \varepsilon)} \|\nabla_\theta \log h_\theta(x)\|^2 + \sup_{\theta \in B(\widehat{\theta}_n, \varepsilon)} \|\nabla^2_\theta \log h_\theta(x)\|$.

(G1) Estimators $\widehat{\xi}^{\mathrm{IS}}_{n,m}$ and $\widehat{\xi}^{\mathrm{NCE}}_{n,m}$ converge to $\widehat{\xi}_n$ almost surely, and are such that

$$\nabla \ell^{\mathrm{IS}}_{n,m}(\widehat{\xi}^{\mathrm{IS}}_{n,m}) = o\left(m^{-1}\right), \qquad \nabla \ell^{\mathrm{NCE}}_{n,m}(\widehat{\xi}^{\mathrm{NCE}}_{n,m}) = o\left(m^{-1}\right).$$

(I1) For some $\varepsilon > 0$ the following integrability condition holds:

$$\mathbb{E}_\psi\left[b_\varepsilon(X) \sup_{\theta \in B(\widehat{\theta}_n, \varepsilon)} \left(\frac{h_\theta(X)}{h_\psi(X)}\right)^2\right] < +\infty$$

where $b_\varepsilon(x) = 1 + \sup_{\theta \in B(\widehat{\theta}_n, \varepsilon)} \|\nabla_\theta \log h_\theta(x)\|$.

Measurability of the suprema in (H2) and (I1) is ensured by the lower semi-continuity of the two first differentials in a neighbourhood of $\widehat{\theta}_n$. Assumption (H2) is a regularity condition that ensures in particular that the partition function $\theta \mapsto \mathcal{Z}(\theta) = \int_{\mathcal{X}} h_\theta(x)\mu(\mathrm{d}x)$ is twice differentiable under the integral sign, in a neighbourhood of $\widehat{\theta}_n$. Following Theorem 1, Assumption (G1) is trivial as soon as Assumptions (C1) and (H1) hold and $\widehat{\xi}^{\mathrm{IS}}_{n,m}$ and $\widehat{\xi}^{\mathrm{NCE}}_{n,m}$ are exact maximisers; in that case the gradients are zero. Integrability Assumption (I1) is the critical assumption. It is essentially a (locally uniform) second moment condition on the importance weights, with $\mathbb{P}_{\widehat{\theta}_n}$ as the target distribution.

Theorem 2. *Under assumptions (X1), (H2), (G1) and (I1):*

$$m\left(\widehat{\xi}^{\mathrm{NCE}}_{n,m} - \widehat{\xi}^{\mathrm{IS}}_{n,m}\right) \underset{m\to\infty}{\to} n\left(-\mathcal{H}(\widehat{\xi}_n)\right)^{-1} v(\widehat{\xi}_n) \quad a.s. \tag{2.7}$$

where $\mathcal{H}(\xi) = \nabla^2_\xi \ell_n(\xi)$*, and* $v(\xi)$ *is defined as follows: let* $g_\xi(x) = \log h_\theta(x) + \nu$*, then*

$$v(\xi) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\xi g_\xi(y_i)\left(\frac{\exp\{g_\xi(y_i)\}}{h_\psi(y_i)}\right) - \mathbb{E}_\psi\left[\nabla_\xi g_\xi(X)\left(\frac{\exp\{g_\xi(X)\}}{h_\psi(X)}\right)^2\right].$$

Before discussing the implications of Theorem 2, it is important to consider Geyer (1994)'s result about asymptotic normality of MC-MLE, which relies on the following assumption:

(N) For some covariance matrix $\mathbf{A}$ we have:

$$\sqrt{m}\nabla\ell^{\mathrm{IS}}_{n,m}(\widehat{\theta}_n) \underset{m\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}_d\left(\mathbf{0}_d, \mathbf{A}\right)$$

As noticed by Geyer (1994), asymptotics of MC-MLE are quite similar to the asymptotics of maximum likelihood, and it can be shown that under assumptions (X1), (H2), (G1) and (N),

$$\sqrt{m}\left(\widehat{\theta}^{\mathrm{IS}}_{n,m} - \widehat{\theta}_n\right) \underset{m\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}_d\left(\mathbf{0}_d, \mathbf{H}^{-1}\mathbf{A}\mathbf{H}^{-1}\right).$$

Theorem 2 shows that the difference between the two point estimates is $\mathcal{O}(m^{-1})$, which is negligible relative to the $\mathcal{O}_{\mathbb{P}}(m^{-1/2})$ rate of convergence to $\widehat{\theta}_n$. This proves that, when n is fixed, both approaches are asymptotically equivalent when it comes to approximate the MLE. In particular, Slutsky's lemma implies asymptotic normality of the NCE estimator with the same asymptotic variance as for MC-MLE.

Assumptions (H2) and (I1) admit a much simpler formulation when the model belongs to an exponential family. This is the point of the following Proposition.

Proposition 1. *If the parametric model is exponential, i.e. if* $h_\theta(x) = \exp\left\{\theta^T S(x)\right\}$ *for some statistic* S*, then assumptions (H2) and (I1) are equivalent to the following assumptions (H2-exp) and (I1-exp):*

(H2-exp) *The Hessian matrix of the log-likelihood* $\mathbf{H} = \nabla^2\ell_n(\widehat{\theta}_n)$ *is invertible.*

(I1-exp) *The MLE* $\widehat{\theta}_n$ *lies in the interior of* $\Theta_\psi = \left\{\theta \in \Theta : \mathbb{E}_\psi\left[\left(\frac{h_\theta(X)}{h_\psi(X)}\right)^2\right] < +\infty\right\}$.

The set Θ_ψ is convex whenever Θ is. In particular, this is true when Θ coincides with the natural space of parameters, defined as $\widetilde{\Theta} = \{\theta \in \mathbb{R}^d : \int_{\mathcal{X}} \exp\left\{\theta^T S(x)\right\}\mu(\mathrm{d}x) < +\infty\}$. If $\mathbb{P}_\psi \in \{\mathbb{P}_\theta : \theta \in \Theta\}$, then (I1-exp) holds as soon as $2\widehat{\theta}_n - \psi$ lies in the interior of $\widetilde{\Theta}$.

Remark 2. *Condition (N) requires a* $\sqrt{m}$*-CLT (central limit theorem) for the function* $\varphi : x \mapsto \left(\nabla_\theta \log h_\theta\right)(h_\theta/h_\psi)(x)$ *at* $\theta = \widehat{\theta}_n$*. There has been an extensive literature on CLT's for Markov Chains, see e.g. Roberts and Rosenthal (2004) for a review. In particular, if* $(X_j)_{j\geq 1}$ *is a geometrically ergodic Markov Chain with stationary distribution* $\mathbb{P}_\psi$*, then assumption (N) holds if for some* $\delta > 0$*,* $\varphi \in \mathbb{L}_{2+\delta}(\mathbb{P}_\psi)$*. This assumption is very similar to assumption (I1), especially when the model is exponential.*

In practice, implications of Theorem 2 must be considered cautiously, as the Euclidian norm of the limit in (2.7) will typically increase with n. For several well-known un-normalised models (e.g. Ising models, Exponential Random Graph Models), n is equal to one, in which case NCE and MC-MLE will always produce very close estimates. For other models however, it is known that the two estimators may behave differently, especially when the number of actual data-points is big and when simulations have a high computational cost (see Gutmann and Hyvärinen (2012)).

To investigate to which extent both approaches provide a good approximation of the true parameter value in these models, we will require both m and n to go to infinity. As it turns out, this will also make it possible to do finer comparison between $\widehat{\xi}^{\mathrm{IS}}_{n,m}$ and $\widehat{\xi}^{\mathrm{NCE}}_{n,m}$ (and thus between $\widehat{\theta}^{\mathrm{NCE}}_{n,m}$ and $\widehat{\theta}^{\mathrm{IS}}_{n,m}$). This is the point of the next section.

2.4 Asymptotics of the overall error

We now assume that observations y_i are realisations of IID random variables Y_i, with probability density $f_{\theta^\star}$, for some true parameter $\theta^\star \in \Theta$, while the artificial data-points $(X_j)_{j\geq 1}$ remain generated from a $\mathbb{P}_\psi$-ergodic process. We also assume that $(Y_i)_{i\geq 1}$ and $(X_j)_{j\geq 1}$ are independent sequences; this regime was first studied for NCE by Gutmann and Hyvärinen (2012), although the X_j's were assumed IID in that paper.

This asymptotic regime has some drawbacks: it assumes that the model is well specified, and that $\mathbb{P}_\psi$ is chosen independently from the data. This is rarely true in practice, as the user will generally try to choose $\mathbb{P}_\psi$ as close as possible to the data distribution to reduce the mean square error (see Section 2.5). Nevertheless, allowing both m and n to go to infinity turns out to provide a better understanding of the asymptotic behaviours of NCE and MC-MLE, at least for situations where the number of actual data-points may be large.

We assume implicitly that $m = m_n$ is a non-decreasing sequence of positive integers going to infinity when n does, while $m_n/n \to \tau \in (0, +\infty)$. Every limit when n goes to infinity should be understood accordingly. Finally, $\xi^\star = (\theta^\star, \nu^\star)$ stands for the true extended parameter, where $\nu^\star = \log\{\mathcal{Z}(\psi)/\mathcal{Z}(\theta^\star)\}$.

2.4.1 Consistency

Our results concerning the overall consistency (to $\xi^\star$, as both m and $n \to \infty$) of MC-MLE and NCE rely on the following assumptions:

(C2) The random sequences $\left(\widehat{\xi}^{\mathrm{IS}}_{n,m}\right)_{n\geq 1}$ and $\left(\widehat{\xi}^{\mathrm{NCE}}_{n,m}\right)_{n\geq 1}$ are approximate MC-MLE and NCE estimators, and belong to a compact set almost surely.

(H3) The maps $\theta \mapsto h_\theta(x)$ are continuous for $\mathbb{P}_\psi$-almost every x, and for any $\theta \in \Theta$ there is some $\varepsilon > 0$ such that

$$\int_{\mathcal{X}} \sup_{\phi \in B(\theta,\varepsilon)} \left(\log \frac{h_\phi(x)}{h_{\theta^\star}(x)}\right)_+ h_{\theta^\star}(x)\mu(\mathrm{d}x) < +\infty.$$

Theorem 3. *Under assumptions (X1), (C2) and (H3), both estimators $\widehat{\xi}^{\mathrm{IS}}_{n,m}$ and $\widehat{\xi}^{\mathrm{NCE}}_{n,m}$ converge almost surely to $\xi^\star$ as $n, m \to \infty$, while $m/n \to \tau$.*

Our proofs of NCE and MC-MLE consistency are mainly inspired from Wald (1949)'s famous proof of MLE consistency, for which the same integrability condition (H3) is required. It is noteworthy that, under this regime, MC-MLE and NCE consistency essentially rely on the same assumptions as MLE consistency.

Remark 3. *As noticed by Wald (1949), the proof does not require Θ to be a subset of $\mathbb{R}^d$. Theorem 3 holds as soon as Θ is a metric space.*

Proposition 2. *If the parametric model is exponential, i.e. if $h_\theta(x) = \exp\left\{\theta^T S(x)\right\}$ for some measurable statistic S, then assumption (H3) always holds.*

2.4.2 Asymptotic normality

To ensure the asymptotic normality of both NCE and MC-MLE estimates, we make the following assumption.

(X2) The sequence $(X_j)_{j\geq 1}$ is a Harris ergodic Markov chain (that is, aperiodic, ϕ-irreducible and positive Harris recurrent; for definitions see Meyn and Tweedie (2012)), with stationary distribution $\mathbb{P}_\psi$.

The Markov kernel associated with the chain $(X_j)_{j\geq 1}$, noted $P(x, \mathrm{d}y)$, is reversible (satisfies detailed balance) with respect to $\mathbb{P}_\psi$, that is

$$\mathbb{P}_\psi(\mathrm{d}x)P(x, \mathrm{d}y) = \mathbb{P}_\psi(\mathrm{d}y)P(y, \mathrm{d}x). \tag{2.8}$$

Moreover, the chain $(X_j)_{j\geq 1}$ is *geometrically ergodic*, i.e. there is some $\rho \in [0, 1)$ and a positive measurable function M such that for $\mathbb{P}_\psi$-almost every x

$$\|P^n(x, .) - \mathbb{P}_\psi(.)\|_{\mathrm{TV}} \leq M(x)\rho^n \tag{2.9}$$

where $P^n(x; dy)$ denote the n-step Markov transition kernel corresponding to P, and $\|.\|_{\mathrm{TV}}$ stands for the total variation norm.

Under (X2), for any measurable, real-valued function φ such that $\mathbb{E}_\psi[\varphi^2] < \infty$, then a $\sqrt{m}$-CLT holds, i.e.

$$\sqrt{m}\left(\frac{1}{m}\sum_{j=1}^{m}\varphi(X_j) - \mathbb{E}_\psi\left[\varphi(X)\right]\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_\varphi^2) \tag{2.10}$$

where

$$\sigma_\varphi^2 = \mathbb{V}_\psi(\varphi(X)) + 2\sum_{i=1}^{\infty}\mathrm{Cov}(\varphi(X_0), \varphi(X_i)).$$

In the equation above, $\mathrm{Cov}(\varphi(X_0), \varphi(X_i))$ stands for the i-th lag autocovariance of the chain at stationarity; that is with respect to the distribution defined by $X_0 \sim \mathbb{P}_\psi$ and $X_{i+1}|X_i \sim P(X_i, .)$. The sequence of artificial data-points $(X_j)_{j\geq 1}$ is not assumed stationary. Since the chain is Harris recurrent, (2.10) holds whenever $X_1 = x$ for any $x \in \mathcal{X}$ (see e.g. Roberts and Rosenthal (2004), especially Theorem 4 and Proposition 29).

For convenience, we choose to assume that the kernel is reversible (which is true for any Metropolis-Hastings algorithm), but the reversibility assumption (2.8) is not compulsory, and may be replaced by slightly stronger integrability assumptions (see e.g. Roberts and Rosenthal (2004)); in particular, if reversibility is not assumed then (2.10) holds whenever $\varphi \in \mathbb{L}_{2+\delta}(\mathbb{P}_\psi)$. The critical assumption is geometric ergodicity.

Geometric ergodicity is obviously stronger than assumption (X1) which only requires a law of large numbers to hold. Nevertheless, geometric ergodicity remains a state of the art condition to ensure CLT's for Markov chains (see e.g. Roberts and Rosenthal (2004) and Bradley et al. (2005)), while it can often be checked for practical MCMC samplers. We present assumption (X2) as a practical condition for ensuring CLT's when the artificial data-points are generated from a MCMC sampler, while it also covers the IID case without loss of generality. Note though that CLT's can hold under sharper conditions, when the Markov chain satisfies polynomial ergodicity for instance (see Jones (2004)).

Our asymptotic normality results rely on the following assumptions:

(H4) The maps $\theta \mapsto h_\theta(x)$ are twice continuously differentiable in a neighborhood of $\theta^\star$ for $\mathbb{P}_\psi$-almost every x; the Fisher Information $\mathbf{I}(\theta) = \mathbb{V}_\theta\Big(\nabla_\theta \log h_\theta(Y)\Big)$ is invertible at $\theta = \theta^\star$; and for some $\varepsilon > 0$

$$\int_{\mathcal{X}} c_\varepsilon(x) \sup_{\theta \in B(\theta^\star, \varepsilon)} h_\theta(x)\mu(\mathrm{d}x) < \infty$$

where $c_\varepsilon(x) = 1 + \sup_{\theta \in B(\theta^\star, \varepsilon)} \|\nabla_\theta \log h_\theta(x)\|^2 + \sup_{\theta \in B(\theta^\star, \varepsilon)} \|\nabla^2_\theta \log h_\theta(x)\|.$

(G2) Estimators $\widehat{\xi}^{\mathrm{IS}}_{n,m}$ and $\widehat{\xi}^{\mathrm{NCE}}_{n,m}$ converge in probability to $\xi^\star$, and are such that

$$\nabla \ell^{\mathrm{IS}}_{n,m}(\widehat{\xi}^{\mathrm{IS}}_{n,m}) = o_{\mathbb{P}}\left(n^{-1/2}\right), \qquad \nabla \ell^{\mathrm{NCE}}_{n,m}(\widehat{\xi}^{\mathrm{NCE}}_{n,m}) = o_{\mathbb{P}}\left(n^{-1/2}\right).$$

(I3) At $\theta = \theta^\star$, the following integrability condition holds:

$$\mathbb{E}_\psi\left[d_\theta(X)\left(\frac{h_\theta(X)}{h_\psi(X)}\right)^2\right] < \infty$$

where $d_\theta(x) = 1 + \|\nabla_\theta \log h_\theta(x)\|^2$.

Theorem 4. *Under assumptions (X2), (H4) and (G2), we have*

$$\sqrt{n}\left(\widehat{\xi}_{n,m}^{\mathrm{NCE}} - \xi^\star\right) \xrightarrow{\mathcal{D}} \mathcal{N}_{d+1}\left(0, \mathbf{V}_\tau^{\mathrm{NCE}}(\xi^\star)\right)$$

where

$$\begin{aligned}
\mathbf{V}_\tau^{\mathrm{NCE}}(\xi) &= \mathbf{J}_\tau(\xi)^{-1}\left\{\mathbf{\Sigma}_\tau(\xi) + \tau^{-1}\mathbf{\Gamma}_\tau(\xi)\right\}\mathbf{J}_\tau(\xi)^{-1},\\
\mathbf{J}_\tau(\xi) &= \mathbb{E}_\theta\left[(\nabla_\xi \nabla_\xi^T g_\xi)\left(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\right)(Y)\right],\\
\mathbf{\Sigma}_\tau(\xi) &= \mathbb{V}_\theta\left((\nabla_\xi g_\xi)\left(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\right)(Y)\right),\\
\mathbf{\Gamma}_\tau(\xi) &= \mathbb{V}_\psi\left(\varphi_\xi^{\mathrm{NCE}}(X)\right) + 2\sum_{i=1}^{+\infty}\mathrm{Cov}\left(\varphi_\xi^{\mathrm{NCE}}(X_0), \varphi_\xi^{\mathrm{NCE}}(X_i)\right),\\
\varphi_\xi^{\mathrm{NCE}}(x) &= (\nabla_\xi g_\xi)\frac{f_\theta}{f_\psi}\left(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\right)(x).
\end{aligned}$$

Moreover, under assumptions (X2), (H4), (G2) and (I3), we have

$$\sqrt{n}\left(\widehat{\xi}_{n,m}^{\mathrm{IS}} - \xi^\star\right) \xrightarrow{\mathcal{D}} \mathcal{N}_{d+1}\left(0, \mathbf{V}_\tau^{\mathrm{IS}}(\xi^\star)\right)$$

where

$$\begin{aligned}
\mathbf{V}_\tau^{\mathrm{IS}}(\xi) &= \mathbf{J}(\xi)^{-1}\left\{\mathbf{\Sigma}(\xi) + \tau^{-1}\mathbf{\Gamma}(\xi)\right\}\mathbf{J}(\xi)^{-1},\\
\mathbf{J}(\xi) &= \mathbb{E}_\theta\left[\nabla_\xi \nabla_\xi^T g_\xi(Y)\right],\\
\mathbf{\Sigma}(\xi) &= \mathbb{V}_\theta\left(\nabla_\xi g_\xi(Y)\right),\\
\mathbf{\Gamma}(\xi) &= \mathbb{V}_\psi\left(\varphi_\xi^{\mathrm{IS}}(X)\right) + 2\sum_{i=1}^{+\infty}\mathrm{Cov}\left(\varphi_\xi^{\mathrm{IS}}(X_0), \varphi_\xi^{\mathrm{IS}}(X_i)\right),\\
\varphi_\xi^{\mathrm{IS}}(x) &= (\nabla_\xi g_\xi)\frac{f_\theta}{f_\psi}(x).
\end{aligned}$$

Remark 4. *Second moment condition (I3) is critical. It basically forbids $\mathbb{P}_\psi$ to be chosen as a too thin tail distribution compared to $\mathbb{P}_{\theta^*}$. Assumption (I3) is needed for establishing MC-MLE asymptotic normality, but not for NCE (inequality 2.21 shows that condition (H4) is enough). This already shows that, under the considered regime, NCE is more robust (to $\mathbb{P}_\psi$) than MC-MLE.*

Assumptions (H4) and (I3) admit a simpler formulation when the model is exponential, as shown by the following proposition.

Proposition 3. *If the parametric model is exponential, i.e. if $h_\theta(x) = \exp\left\{\theta^T S(x)\right\}$ for some statistic S, then assumptions (H4) and (I3) are equivalent to the following assumptions (H4-exp) and (I3-exp):*

(H4-exp) *Fisher Information* $\mathbf{I}(\theta) = \mathbb{V}_\theta\Big(\nabla_\theta \log h_\theta(Y)\Big)$ *is invertible at* $\theta = \theta^\star$.

(I3-exp) *Parameter* $\theta^\star$ *belongs to the interior of* $\Theta_\psi = \Big\{\theta : \mathbb{E}_\psi\Big[\Big(\frac{h_\theta(X)}{h_\psi(X)}\Big)^2\Big] < \infty\Big\}$.

In particular, if $\mathbb{P}_\psi \in \{\mathbb{P}_\theta\}_{\theta\in\Theta}$, *then (I3-exp) holds as soon as* $2\theta^\star - \psi$ *belongs to the interior of* $\widetilde{\Theta} = \Big\{\theta \in \mathbb{R}^d : \int_{\mathcal{X}} \exp\Big\{\theta^T S(x)\Big\}\mu(\mathrm{d}x) < \infty\Big\}$.

2.4.3 Comparison of asymptotic variances

Theorem 5. *If the artificial data-points* $(X_j)_{j\geq 1}$ *are IID, then under assumptions (H4) and (I3),* $\mathbf{V}_\tau^{\mathrm{IS}}(\xi^\star) \succcurlyeq \mathbf{V}_\tau^{\mathrm{NCE}}(\xi^\star)$, *i.e.* $\mathbf{V}_\tau^{\mathrm{IS}}(\xi^\star) - \mathbf{V}_\tau^{\mathrm{NCE}}(\xi^\star)$ *is a positive semi-definite matrix.*

Theorem 5 shows that, asymptotically, when $m/n \to \tau > 0$, and when the artificial data-points are IID, the variance of a NCE estimator is always lower than the variance of the corresponding MC-MLE estimator. This inequality is with respect to the Loewner partial order on symmetric matrices. To our knowledge, this is the first theoretical result proving that NCE dominates MC-MLE in terms of mean square error. We failed however to extend this result to correlated Markov chains.

This inequality holds for any fixed ratio $\tau \in (0, +\infty)$, and any given sampling distribution $\mathbb{P}_\psi$, but the sharpness of the bound remains unknown. Typically, the bigger is τ, the closer the two variances will be, as the ratio $\tau f_\psi / \tau f_\psi + f_{\theta^\star}$ gets closer to one. It is also the case when the sampling distribution $\mathbb{P}_\psi$ is close to the true data distribution $\mathbb{P}_{\theta^\star}$. Geyer (1994) noticed that MC-MLE performs better when $\mathbb{P}_\psi$ is close to $\mathbb{P}_{\theta^\star}$. Next proposition shows that when $\mathbb{P}_\psi = \mathbb{P}_{\theta^\star}$, both variances can be related to the variance of the MLE.

Proposition 4. *If the artificial data-points are IID sampled from* $\mathbb{P}_\psi = \mathbb{P}_{\theta^\star}$, *then under assumptions (H4) and (I3) we have*

$$\mathbf{V}_\tau^{\mathrm{NCE}}(\xi^\star) = \mathbf{V}_\tau^{\mathrm{IS}}(\xi^\star) = (1 + \tau^{-1})\mathbf{V}^{\mathrm{MLE}}(\xi^\star)$$

where $\mathbf{V}^{\mathrm{MLE}}(\xi) = \mathbf{J}(\xi)^{-1}\mathbf{\Sigma}(\xi)\mathbf{J}(\xi)^{-1}$.

It is straightforward to check that, under the usual conditions ensuring asymptotic normality of the MLE, the extended maximiser of the Poisson Transform ℓ_n is also asymptotically normal with variance $\mathbf{V}^{\mathrm{MLE}}(\xi^\star)$. This proposition shows what we can expect from NCE and MC-MLE in a ideal scenario where the sampling distribution is the same as the true data distribution.

2.5 Numerical example

This section presents a numerical example that illustrates how the variance reduction brought by NCE may vary according to the sampling distribution $\mathbb{P}_\psi$ and the ratio τ.

We consider observations IID distributed from the multivariate Gaussian distribution $\mathcal{N}_p(\mu, \Sigma)$ truncated to $(0, +\infty)^p$; that is $Y_1, ..., Y_n$ are IID with the following probability density with respect to Lebesgue's measure:

$$f_{\mu,\Sigma}(x) = \frac{1}{\mathcal{Z}(\mu, \Sigma)} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \mathbb{1}_{(0,+\infty)^p}(x)$$

where

$$\mathcal{Z}(\mu, \Sigma) = (2\pi)^{p/2}|\Sigma|^{1/2}\mathbb{P}\left(W \in (0, +\infty)^p\right), \qquad W \sim \mathcal{N}_p(\mu, \Sigma).$$

The probability $\mathbb{P}\left(W \in (0, +\infty)^p\right)$ is intractable for almost every (μ, Σ). Numerical approximations of such probabilities quickly become inefficient when p increases.

It is well known that (truncated) Gaussian densities form an exponential family under the following parametrisation: for a given $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{S}_p^{++}$ (the set of positive definite matrices of size p), define $\theta = (\Sigma^{-1}\mu, \text{triu}(-(1/2)\Sigma^{-1}))$, and $S(x) = (x, \text{triu}(xx^T))$, where triu(.) is the upper triangular part. This parametrisation is minimal and the natural parameter space is a convex open subset of $\mathbb{R}^q$ where $q = p + p(p+1)/2$. Indeed, under the exponential formulation, we have $\Theta = \Theta_1 \times \Theta_2$ where $\Theta_1 = \mathbb{R}^p$ and Θ_2 is on open cone of $\mathbb{R}^{p(p+1)/2}$, in bijection with $\mathbb{S}_p^{++}$ through the function triu(.).

The observations are sampled IID from $\mathbb{P}_\theta$ (by reject method) for some true parameter $\theta = \theta^\star$, corresponding to

$$\mu^\star = \begin{pmatrix} 1 \\ -1 \\ 0.5 \end{pmatrix}, \qquad \Sigma^\star = \begin{pmatrix} 1 & 0.5 & 1 \\ 0.5 & 1.5 & 0.3 \\ 1 & 0.3 & 2 \end{pmatrix},$$

in the usual Gaussian parametrisation. The artificial data-points are sampled IID from $\mathbb{P}_\psi$, corresponding to the density $f_{\mu,\Sigma}$ with $\mu = \mathbf{0}_p$ and $\Sigma = \lambda \mathbf{I}_p$ for some $\lambda > 0$. The sample size is fixed to $n = 1000$, while m is chosen such that the ratio m/n is equal to $\tau \in \{1, 5, 20, 100\}$. The distribution $\mathbb{P}_\psi$ is chosen as stated above for $\lambda \in [1.5, 20]$.

Figure 2.1 plots estimates and confidence intervals of the mean square error ratio (mean euclidian norm square error of the estimator divided by the asymptotic variance of the MLE) of both estimators (NCE and MC-MLE), based on 1000 independent replications. (Regarding the denominator of this ratio, note that the variance of the MLE may be estimated by performing noise contrastive estimation with $\mathbb{P}_\psi = \mathbb{P}_{\theta^*}$, see Proposition 4.)

To facilitate the direct comparison between NCE and MC-MLE, we also plot in Figure 2.2 estimates and confidence intervals of the MSE ratio of MC-MCLE over NCE. As expected from Theorem 5, this ratio is always higher than one; it becomes larger and larger as τ decreases, or as λ moves away from its optimal value (around 4). This suggests that NCE is more robust than MC-MLE to a poor choice for the reference distribution (especially thin tails distributions, i.e. when λ goes to zero).

Finally, we discuss a technical difficulty related to the constrained nature of the parameter space Θ. In principle, both the NCE and the MC-MLE estimators should be obtained through constrained optimisation (i.e. as maximisers of their respective objective functions over Θ). However, it is much easier (here, and in many cases) to perform

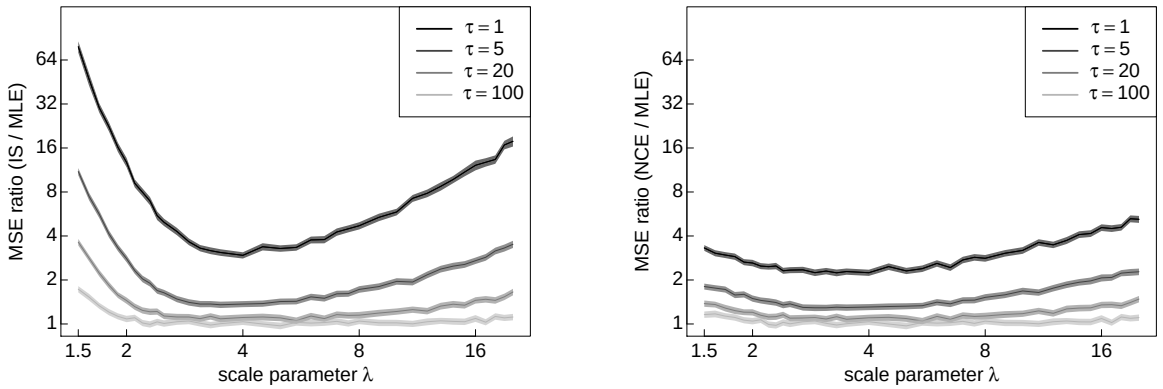

Figure 2.1: Estimates and confidence intervals of the Mean Square Error ratios of MC-MLE (left) and NCE (right), compared to the MLE. The MSE ratio depends both on the variance of the proposal distribution λ and the number of artificial data-points $m = \tau \times n$ ($n = 1000$). A log-scale is used for both axes.

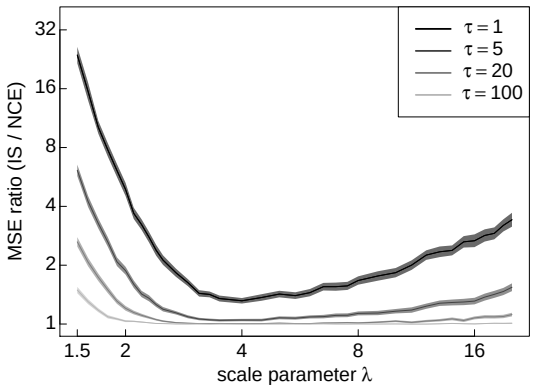

Figure 2.2: Estimates and confidence intervals of the Mean Square Error ratios of MC-MLE, compared to the NCE. The MSE ratio depends both on the variance of the proposal distribution λ and the number of artificial data-points $m = \tau \times n$ ($n = 1000$). A log-scale is used for both axes.

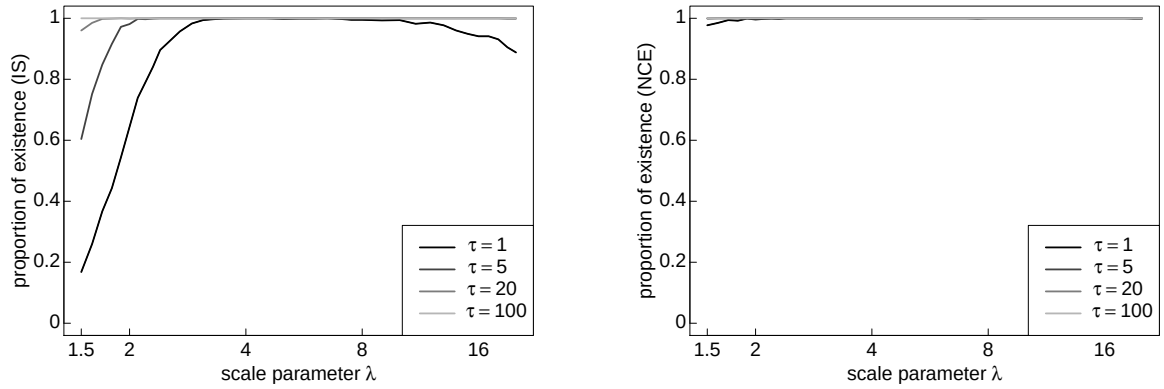

Figure 2.3: Estimates and confidence intervals of the probability of existence of MC-MLE (left) and NCE (right) estimators. For a fixed $n = 1000$, the probability of belonging to Θ is lower for MC-MLE, especially for small values of the variance of the proposal distribution λ and the number of artificial data-points $m = \tau \times n$. A log-scale is used for both axes.

an unconstrained optimisation (over $\mathbb{R}^q$). We must check then that the so obtained solution fulfils the constraint that defines Θ (here, that the solution corresponds to a matrix Σ which is definite positive). Figure 2.3 plots estimates and confidence intervals of the probability that both estimators belong to Θ. We see that NCE (when implemented without constraints) is much more likely to produce estimates that belong to Θ.

Note also that when the considered model is an exponential family (as in this case), both functions $\ell^{\mathrm{IS}}_{n,m}$ and $\ell^{\mathrm{NCE}}_{n,m}$ are convex. This implies that, when the unconstrained maximiser of these functions do not fulfil the constraint that defines Θ, then the constrained maximiser does not exist. (Any solution of the constrained optimisation program lies on the boundary of the constrained set.)

2.6 Conclusion

The three practical conclusions we draw from our results are that: (a) NCE is as widely applicable as MCMC-MLE (including when the X'_js are generated using MCMC); (b) NCE and MC-MLE are asymptotically equivalent (as $m \to \infty$) when n is fixed; (c) NCE may provide lower-variance estimates than MC-MLE when n is large (provided that $m = \mathcal{O}(n)$). The variance reduction seems to be more important when the ratio $\tau = m/n$ is small, or when the reference distribution (for generating the X_j's) is poorly chosen. Note that we proved (c) under the assumption that the X_j's are IID, but we conjecture it also holds when they are generated using MCMC. Proving this conjecture may be an interesting avenue for future research.

As mentioned in the introduction, another advantage of NCE is its ease of implementation. In particular, when the considered model is exponential, NCE boils down to

performing a standard logistic regression. For all these reasons, it seems reasonable to recommend NCE instead of MC-MLE to perform inference for un-normalised models.

2.7 Main Proofs

2.7.1 Technical lemmas

The following lemmas are prerequisites for the proofs of our main theorems. Most of them are 'classical' results, but for the sake of completeness, we provide the proofs of these lemmas in the attached supplement.

All these lemma apply to a $\mathbb{P}_\psi$-ergodic sequence of random variables, $(X_j)_{j\geq 1}$.

First lemma is a slightly disguised version of the law of large numbers, combined with the monotone convergence of a sequence of test functions.

Lemma 1. *Let $(f_m)_{m\geq 1}$ be a non-decreasing sequence of measurable, non negative real-valued functions converging pointwise towards f. Then we have:*

$$\frac{1}{m}\sum_{j=1}^{m} f_m(X_j) \xrightarrow[m\to+\infty]{a.s.} \mathbb{E}_\psi[f(X)].$$

This result holds whether the expectation is finite or infinite.

Second lemma is a natural generalisation of Lemma 1 to dominated convergence.

Lemma 2. *Let $(f_m)_{m\geq 1}$, f and g be measurable, real-valued functions, such that $(f_m)_{m\geq 1}$ converges pointwise towards f; for any $m \geq 1$, $|f_m| \leq g$; and $\mathbb{E}_\psi[g(X)] < +\infty$. Then we have:*

$$\frac{1}{m}\sum_{j=1}^{m} f_m(X_j) \xrightarrow[m\to+\infty]{a.s.} \mathbb{E}_\psi[f(X)].$$

Third lemma is a generalisation of Lemma 1 to the degenerate case where the expectation is infinite. In that case, Lemma 3 shows that the monotonicity assumption is unnecessary.

Lemma 3. *Let $(f_m)_{m\geq 1}$, f and g be measurable, real-valued functions, such that $(f_m)_{m\geq 1}$ converges pointwise towards f; g is non negative, $\mathbb{E}_\psi[g(X)] < +\infty$; for any $m \geq 1$, $f_m \leq g$; and $\mathbb{E}_\psi[f(X)_-] = +\infty$ where f_- stands for the negative part of f. Then we have:*

$$\frac{1}{m}\sum_{j=1}^{m} f_m(X_j) \xrightarrow[m\to+\infty]{a.s.} -\infty.$$

Fourth lemma is a uniform law of large numbers. It is well known in the IID case. This result does not actually require the independence assumption. We present a generalisation of this result to ergodic processes. The proof is due to Bernard Delyon, who made it available in an unpublished course in French (Delyon (2018)). We present an English translation of the proof in the supplement.

Lemma 4. *Let K a compact subset of $\mathbb{R}^d$; $(\theta, x) \mapsto \varphi(\theta, x)$ a measurable function defined on $K \times \mathcal{X}$ whose values lie on $\mathbb{R}^p$; and suppose that the maps $\theta \mapsto \varphi(\theta, x)$ are continuous for $\mathbb{P}_\psi$-almost every x. Moreover, suppose that*

$$\mathbb{E}_\psi\left[\sup_{\theta \in K} \|\varphi(\theta, X)\|\right] < +\infty.$$

Then the function $\theta \mapsto \mathbb{E}_\psi\Big[\varphi(\theta, X)\Big]$ defined on K is continuous, and we have

$$\sup_{\theta \in K} \left\| \frac{1}{m} \sum_{j=1}^{m} \varphi(\theta, X_j) - \mathbb{E}_\psi\left[\varphi(\theta, X)\right] \right\| \xrightarrow[m \to +\infty]{a.s.} 0.$$

Consequently, if there is a random sequence $(\widetilde{\theta}_m)_{m \geq 1}$ converging almost surely to some parameter $\widetilde{\theta} \in \Theta$. Then we have

$$\left\| \frac{1}{m} \sum_{j=1}^{m} \varphi(\widetilde{\theta}_m, X_j) - \mathbb{E}_\psi\left[\varphi(\widetilde{\theta}, X)\right] \right\| \underset{m \to \infty}{\to} 0 \quad a.s.$$

Fifth lemma is also a well known result. It is often used to prove the weak convergence (usually asymptotic normality) of Z-estimators.

Lemma 5. *Define any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $(\ell_n(\theta, \omega))_{n \geq 1}$ be measurable real-valued functions defined on $\mathbb{R}^d \times \Omega$. Let $\theta^\star \in \mathbb{R}^d$ and $\varepsilon > 0$ such that for any $n \geq 1$ and for $\mathbb{P}$-almost every $\omega \in \Omega$ the map $\theta \mapsto \ell_n(\theta, \omega)$ is C^2 on $B(\theta^\star, \varepsilon)$. Let $(\widehat{\theta}_n)_{n \geq 1}$ be a random sequence converging in probability to $\theta^\star$. Suppose also that:*

(a) $\{\nabla_\theta \ell_n(\theta)\}_{|_{\theta = \widehat{\theta}_n}} = o_{\mathbb{P}}(n^{-1/2})$,

(b) $\sup_{\theta \in B(\theta^\star, \varepsilon)} \|\nabla^2_\theta \ell_n(\theta) - \mathcal{H}(\theta)\| \xrightarrow{\mathbb{P}} 0$, *for some $\mathbb{R}^{d \times d}$ valued function $\mathcal{H}$ continuous at $\theta^\star$, such that $\mathcal{H}(\theta^\star)$ is full rank,*

(c) $\sqrt{n}\{\nabla_\theta \ell_n(\theta)\}_{|_{\theta = \theta^\star}} \xrightarrow{\mathcal{D}} Z$, *for some random vector Z.*

Then

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) + \mathcal{H}(\theta^\star)^{-1} \sqrt{n}\{\nabla_\theta \ell_n(\theta)\}_{|_{\theta = \theta^\star}} \xrightarrow{\mathbb{P}} 0_{\mathbb{R}^d},$$

and, consequently

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{D}} -\mathcal{H}(\theta^\star)^{-1} Z.$$

Sixth lemma is a technical tool required for proving asymptotic normality of NCE. It is particularly straightforward to prove in the IID case. We present a generalisation of this result to reversible, geometrically ergodic Markov chains.

Lemma 6. *Assume that (X2) holds. Let $(f_n)_{n\geq 1}$, f and g be measurable, real-valued functions, such that $(f_n)_{n\geq 1}$ converges pointwise towards f; for any $n \geq 1$, $|f_n| \leq g$; and $\mathbb{E}_\psi[g(X)^2] < \infty$. Then we have*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\Big\{f_n(X_i) - f(X_i)\Big\} - \mathbb{E}\Big[f_n(X) - f(X)\Big]\right) \xrightarrow{\mathbb{P}} 0,$$

and, consequently

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f_n(X_i) - \mathbb{E}[f_n(X)]\right) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \sigma_f^2\Big),$$

where $\sigma_f^2 = \mathbb{V}_\psi(f(X)) + 2\sum_{i=1}^{+\infty}\mathrm{Cov}(f(X_0), f(X_i)) < +\infty$.

2.7.2 Proof of Theorem 1

A standard approach to establish consistency of M-estimators is to prove some Glivenko-Cantelli result (uniform convergence), but, to the best of our knowledge, no such result exists under the general assumption that the underlying random variables (the X_j's in our case) are generated from an ergodic process. Instead, we follow Geyer (1994)'s approach, which relies on establishing that function $-\ell_{n,m}^{\mathrm{NCE}}$ epiconverges to $-\ell_n$. Epiconvergence is essentially a one sided locally uniform convergence, that ensures the convergence of minimisers; for a succint introduction to epiconvergence, see Appendix A of Geyer (1994) and Chapter 7 of Rockafellar and Wets (2009).

We follow closely Geyer (1994). In particular, Theorem 4 of Geyer (1994) shows that: if a sequence of functions $\ell_{n,m}$ hypoconverges to some function ℓ_n which has a unique maximiser $\widehat{\theta}_n$ and if a random sequence $(\widehat{\theta}_{n,m})_{m\geq 1}$ is an approximate maximiser of $\ell_{n,m}$ which belongs to a compact set almost surely, then $\widehat{\theta}_{n,m}$ converges to $\widehat{\theta}_n$ almost surely. Consequently, to prove Theorem 1, we only have to prove that $\ell_{n,m}^{\mathrm{NCE}}$ hypoconverges to ℓ_n (i.e. that $-\ell_{n,m}^{\mathrm{NCE}}$ epiconverges to $-\ell_n$); that is

$$\ell_n(\theta,\nu) \leq \inf_{B\in\mathcal{N}(\theta,\nu)} \liminf_{m\to+\infty} \sup_{(\phi,\mu)\in B} \Big\{\ell_{n,m}^{\mathrm{NCE}}(\phi,\mu)\Big\} \tag{2.11}$$

$$\ell_n(\theta,\nu) \geq \inf_{B\in\mathcal{N}(\theta,\nu)} \limsup_{m\to+\infty} \sup_{(\phi,\mu)\in B} \Big\{\ell_{n,m}^{\mathrm{NCE}}(\phi,\mu)\Big\} \tag{2.12}$$

where $\mathcal{N}(\theta,\nu)$ denotes the set of neighborhoods of the point (θ,ν).

Since $\Xi = \Theta \times \mathbb{R}$ is a separable metric space, there exists a countable base $\mathcal{B} = \{B_1, B_2, ...\}$ for the considered topology. For any point (θ,ν) define the countable base of neighborhoods $\mathcal{N}_c(\theta,\nu) = \mathcal{B} \cap \mathcal{N}(\theta,\nu)$ which can replace $\mathcal{N}(\theta,\nu)$ in the infima of the preceding inequalities. Choose a countable dense subset $\Gamma_c = \{(\theta_1,\nu_1),(\theta_2,\nu_2),...\}$ as follows. For each k let (θ_k,ν_k) be a point of B_k such that:

$$\ell_n(\theta_k,\nu_k) \geq \sup_{(\phi,\mu)\in B_k}\{\ell_n(\phi,\mu)\} - \frac{1}{k}.$$

The proof is very similar to Theorem 1 of Geyer (1994). However, in this slightly different proof, we will need

$$\lim_{m\to+\infty} \left[\frac{1}{m}\sum_{j=1}^{m} \log\left\{\left(1+e^{\nu}\frac{nh_\theta(X_j)}{mh_\psi(X_j)}\right)^{\frac{m}{n}}\right\}\right] = \mathbb{E}_\psi\left[e^{\nu}\frac{h_\theta(X)}{h_\psi(X)}\right] = e^{\nu}\frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)} \tag{2.13}$$

and

$$\lim_{m\to+\infty} \frac{1}{m}\sum_{j=1}^{m} \log\left(1+\frac{n}{m}\inf_{(\phi,\mu)\in B}\left[e^{\mu}\frac{h_\phi(X_j)}{h_\psi(X_j)}\right]\right)^{\frac{m}{n}} = \mathbb{E}_\psi\left[\inf_{(\phi,\mu)\in B}\left\{e^{\mu}\frac{h_\phi(X)}{h_\psi(X)}\right\}\right] \tag{2.14}$$

to hold simultaneously with probability one for any $(\theta,\nu)\in\Gamma_c$ and any $B\in\mathcal{B}$. For any fixed (θ,ν), Lemma 1 applies to the maps $x\mapsto(1+\frac{x}{m})^m$, and since any countable union of null sets is still a null set, convergence holds simultaneously for every element of Γ_c and $\mathcal{B}$ with probability one. One may note that infima in the last equation are measurable under (H1) (in that case, an infima over any set $B\in\mathcal{B}$ can be replaced by an infima over the countable dense subset $B\cap\Gamma_c$).

Proving inequality (2.11) is straightforward:

$$\forall B\in\mathcal{B},\quad \forall(\theta,\nu)\in B\cap\Gamma_c,\qquad \ell_n(\theta,\nu)=\lim_{m\to+\infty}\ell^{\mathrm{NCE}}_{n,m}(\theta,\nu)\leq\liminf_{m\to+\infty}\sup_{(\phi,\mu)\in B}\left\{\ell^{\mathrm{NCE}}_{n,m}(\phi,\mu)\right\}$$

and thus

$$\inf_{B\in\mathcal{N}_c(\theta,\nu)}\sup_{(\phi,\mu)\in B\cap\Gamma_c}\{\ell_n(\phi,\mu)\}\leq\inf_{B\in\mathcal{N}_c(\theta,\nu)}\liminf_{m\to+\infty}\sup_{(\phi,\mu)\in B}\left\{\ell^{\mathrm{NCE}}_{n,m}(\phi,\mu)\right\}.$$

(Geyer, 1994) proved that $\theta\mapsto\mathcal{Z}(\theta)$ is lower semi-continuous (cf Theorem 1). This result directly implies that $(\theta,\nu)\mapsto\ell_n(\theta,\nu)$ is upper semi-continuous as a sum of upper semi-continuous functions. Thus the left hand side is equal to $l(\theta,\nu)$ by construction of Γ_c.

The proof of the second inequality also follows closely Geyer (1994):

$$\begin{aligned}
&\inf_{B\in\mathcal{N}(\theta,\nu)}\limsup_{m\to+\infty}\sup_{(\phi,\mu)\in B}\left\{\ell^{\mathrm{NCE}}_{n,m}(\phi,\mu)\right\}\\
&\qquad\leq\inf_{B\in\mathcal{N}(\theta,\nu)}\left\{\sup_{(\phi,\mu)\in B}\left[\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{h_\phi(y_i)}{h_\psi(y_i)}\right\}+\mu\right]\right.\\
&\qquad\quad-\liminf_{m\to+\infty}\inf_{(\phi,\mu)\in B}\left[\frac{1}{n}\sum_{i=1}^{n}\log\left\{1+\frac{n}{m}e^{\mu}\frac{h_\phi(y_i)}{h_\psi(y_i)}\right\}\right]\\
&\qquad\quad\left.-\liminf_{m\to+\infty}\frac{1}{m}\sum_{j=1}^{m}\log\left(1+\frac{n}{m}\inf_{(\phi,\mu)\in B}\left[e^{\mu}\frac{h_\phi(X_j)}{h_\psi(X_j)}\right]\right)^{\frac{m}{n}}\right\}\\
&\qquad=\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{h_\theta(y_i)}{h_\psi(y_i)}\right\}+\nu-\sup_{B\in\mathcal{N}(\theta,\nu)}\mathbb{E}_\psi\left[\inf_{(\phi,\mu)\in B}\left\{e^{\mu}\frac{h_\phi(X)}{h_\psi(X)}\right\}\right].
\end{aligned}$$

The inequality follows directly from superadditivity of the supremum (and subadditivity of the infimum) and the continuity and monotonicity of the maps $x\mapsto\log(1+\frac{n}{m}x)^{\frac{m}{n}}$.

The last equality holds because the infimum over $\mathcal{N}(\theta,\nu)$ can be replaced by the infimum over the countable set $\mathcal{A}_c(\theta,\nu)$: the set of open balls centered on (θ,ν) of radius k^{-1}, $k \geq 1$, which means the infimum is also the limit of a decreasing sequence, which can be splitted into three terms. The second term converges deterministically to zero, while convergences (2.13) and (2.14) apply for the first and third terms.

To conclude, apply the monotone convergence theorem to the remaining term:

$$\begin{aligned}\sup_{B\in\mathcal{A}_c(\theta,\nu)} \mathbb{E}_\psi\left[\inf_{(\phi,\mu)\in B}\left\{e^\mu \frac{h_\phi(X)}{h_\psi(X)}\right\}\right] &= \mathbb{E}_\psi\left[\sup_{B\in\mathcal{A}_c(\theta,\nu)}\inf_{(\phi,\mu)\in B}\left\{e^\mu \frac{h_\phi(X)}{h_\psi(X)}\right\}\right] \\ &= \mathbb{E}_\psi\left[e^\nu \frac{h_\theta(X)}{h_\psi(X)}\right] = e^\nu \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)}.\end{aligned}$$

2.7.3 Proof of Theorem 2

Define $g_\xi(x) = \log h_\theta(x) + \nu$, and the following gradients (dropping n and m in the notation for convenience):

$$\begin{aligned}\Psi^{\mathrm{NCE}}(\xi) = \nabla \ell^{\mathrm{NCE}}_{n,m}(\xi) = \frac{1}{n}\sum_{i=1}^n \nabla_\xi g_\xi(y_i)\left(\frac{m h_\psi(y_i)}{m h_\psi(X_j) + n\exp\{g_\xi(y_i)\}}\right) \\ - \frac{1}{m}\sum_{j=1}^m \nabla_\xi g_\xi(X_j)\left(\frac{m\exp\{g_\xi(X_j)\}}{m h_\psi(X_j) + n\exp\{g_\xi(X_j)\}}\right),\end{aligned}$$

$$\Psi^{\mathrm{IS}}(\xi) = \nabla \ell^{\mathrm{IS}}_{n,m}(\xi) = \frac{1}{n}\sum_{i=1}^n \nabla_\xi g_\xi(y_i) - \frac{1}{m}\sum_{j=1}^m \nabla_\xi g_\xi(X_j)\left(\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\right).$$

By Taylor-Lagrange, for any component k, $1 \leq k \leq d+1$, there exists (a random variable) $\xi_m^{(k)} \in [\widehat{\xi}^{\mathrm{IS}}_{n,m}; \widehat{\xi}^{\mathrm{NCE}}_{n,m}]$ such that

$$\Psi^{\mathrm{IS}}_k(\widehat{\xi}^{\mathrm{IS}}_{n,m}) = \Psi^{\mathrm{IS}}_k(\widehat{\xi}^{\mathrm{NCE}}_{n,m}) + \left\{\nabla\Psi^{\mathrm{IS}}_k(\xi_m^{(k)})\right\}^T\left(\widehat{\xi}^{\mathrm{IS}}_{n,m} - \widehat{\xi}^{\mathrm{NCE}}_{n,m}\right)$$

where $\Psi^{\mathrm{IS}}_k(\xi)$ denotes the $k-$th component of $\Psi^{\mathrm{IS}}(\xi)$, and $[\widehat{\xi}^{\mathrm{IS}}_{n,m}; \widehat{\xi}^{\mathrm{NCE}}_{n,m}]$ denotes the line segment in $\mathbb{R}^{d+1}$ which joins $\widehat{\xi}^{\mathrm{IS}}_{n,m}$ and $\widehat{\xi}^{\mathrm{NCE}}_{n,m}$.

By assumption (G1), the left hand side is $o_{\mathbb{P}}(m^{-1})$. The matrix form yields:

$$o_{\mathbb{P}}\left(m^{-1}\right) = \Psi^{\mathrm{IS}}(\widehat{\xi}^{\mathrm{NCE}}_{n,m}) + \mathbf{H}^{\mathrm{IS}}_m\left(\widehat{\xi}^{\mathrm{IS}}_{n,m} - \widehat{\xi}^{\mathrm{NCE}}_{n,m}\right), \qquad \mathbf{H}^{\mathrm{IS}}_m = \begin{pmatrix} \left\{\nabla\Psi^{\mathrm{IS}}_1(\xi_m^{(1)})\right\}^T \\ \vdots \\ \left\{\nabla\Psi^{\mathrm{IS}}_{d+1}(\xi_m^{(d+1)})\right\}^T \end{pmatrix}.$$

Let us prove first the convergence of the Hessian matrix. Lemma 4 can be applied to each row component of the following matrix-valued function, the uniform norm of which is $\mathbb{P}_\psi$-integrable under (H2):

$$\varphi_h : (\xi, x) \mapsto \left(\frac{1}{n}\sum_{i=1}^n \nabla^2_\xi g_\xi(y_i)\right) - \left(\nabla^2_\xi g_\xi(x) + \nabla_\xi g_\xi(x)\left\{\nabla_\xi g_\xi(x)\right\}^T\right)\left(\frac{\exp\{g_\xi(x)\}}{h_\psi(x)}\right).$$

Convergences of the $d+1$ rows of $\mathbf{H}_m^{\mathrm{IS}}$ can be combined to get the following result:

$$\left\|\mathbf{H}_m^{\mathrm{IS}} - \mathcal{H}(\widehat{\xi}_n)\right\| \underset{m\to\infty}{\to} 0 \qquad \text{a.s.}$$

where

$$\mathcal{H}(\xi) = \mathbb{E}_\psi\Big[\varphi_h(\xi, X)\Big] = \nabla_\xi^2 \ell_n(\xi).$$

It turns out that $\mathcal{H}(\widehat{\xi}_n)$ is invertible as soon as (H2) holds. This is the point of the following lemma. This implies in particular that $\mathbf{H}_m^{\mathrm{IS}}$ is eventually invertible with probability one.

Lemma 7. *Assume (H2) holds. At the point $\xi = \widehat{\xi}_n$, the Hessian matrix of the Poisson Transform $\nabla_\xi^2 l_n(\xi)$ is negative definite if and only if the Hessian of the log-likelihood $\nabla_\theta^2 l_n(\theta)$ is definite negative.*

The proof of Lemma 7 follows from a direct block matrix computation (using Schur's complement). For the sake of completeness, we present a proof in the supplement.

Now, let us prove the convergence of the gradient. By assumption (G1), we can write $\Psi^{\mathrm{IS}}(\widehat{\xi}_{n,m}^{\mathrm{NCE}}) = \Delta_m + o\Big(m^{-1}\Big)$, where:

$$\begin{aligned}
\Delta_m &= \Psi^{\mathrm{IS}}(\widehat{\xi}_{n,m}^{\mathrm{NCE}}) - \Psi^{\mathrm{NCE}}(\widehat{\xi}_{n,m}^{\mathrm{NCE}}) \\
&= \left\{ \frac{1}{n}\sum_{i=1}^n \nabla_\xi g_\xi(y_i)\Big(\frac{n\exp\{g_\xi(y_i)\}}{mh_\psi(y_i) + n\exp\{g_\xi(y_i)\}}\Big) \right. \\
&\quad \left. - \frac{1}{m}\sum_{j=1}^m \nabla_\xi g_\xi(X_j)\Big(\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\Big)\Big(\frac{n\exp\{g_\xi(X_j)\}}{mh_\psi(X_j) + n\exp\{g_\xi(X_j)\}}\Big) \right\}\Big|_{\xi=\widehat{\xi}_{n,m}^{\mathrm{NCE}}}
\end{aligned}$$

hence

$$\begin{aligned}
\frac{m}{n}\Delta_m &= \left\{ \frac{1}{n}\sum_{i=1}^n \nabla_\xi g_\xi(y_i)\Big(\frac{\exp\{g_\xi(y_i)\}}{h_\psi(y_i)}\Big)\Big(1 - \frac{n\exp\{g_\xi(y_i)\}}{mh_\psi(y_i) + n\exp\{g_\xi(y_i)\}}\Big) \right. \\
&\quad \left. - \frac{1}{m}\sum_{j=1}^m \nabla_\xi g_\xi(X_j)\Big(\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\Big)^2\Big(1 - \frac{n\exp\{g_\xi(X_j)\}}{mh_\psi(X_j) + n\exp\{g_\xi(X_j)\}}\Big) \right\}\Big|_{\xi=\widehat{\xi}_{n,m}^{\mathrm{NCE}}} \\
&= \left\{ \frac{1}{n}\sum_{i=1}^n \nabla_\xi g_\xi(y_i)\Big(\frac{\exp\{g_\xi(y_i)\}}{h_\psi(y_i)}\Big) - \frac{1}{m}\sum_{j=1}^m \nabla_\xi g_\xi(X_j)\Big(\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\Big)^2 \right. \\
&\quad - \frac{1}{n}\sum_{i=1}^n \nabla_\xi g_\xi(y_i)\Big(\frac{\exp\{g_\xi(y_i)\}}{h_\psi(y_i)}\Big)\Big(\frac{n\exp\{g_\xi(y_i)\}}{mh_\psi(y_i) + n\exp\{g_\xi(y_i)\}}\Big) \\
&\quad \left. + \frac{1}{m}\sum_{j=1}^m \nabla_\xi g_\xi(X_j)\Big(\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\Big)^2\Big(\frac{n\exp\{g_\xi(X_j)\}}{mh_\psi(X_j) + n\exp\{g_\xi(X_j)\}}\Big) \right\}\Big|_{\xi=\widehat{\xi}_{n,m}^{\mathrm{NCE}}}.
\end{aligned}$$

The two last terms of the right hand side are residuals for which we want to bound the uniform norm over the ball $B(\widehat{\theta}_n, \varepsilon)$. The sup norm of the second term is eventually

bounded by:

$$\frac{1}{m} \sup_{\xi \in B(\widehat{\xi}_n, \varepsilon)} \sum_{i=1}^{n} \|\nabla_\xi g_\xi(y_i)\| \left(\frac{\exp\{g_\xi(y_i)\}}{h_\psi(y_i)} \right)^2 \underset{m \to \infty}{\to} 0.$$

The sup norm of the third term is eventually bounded by $\frac{1}{m} \sum_{j=1}^{m} f_m(X_j)$ where

$$f_m(x) = \sup_{\xi \in B(\widehat{\xi}_n, \varepsilon)} \|\nabla_\xi g_\xi(x)\| \left(\frac{\exp\{g_\xi(x)\}}{h_\psi(x)} \right)^2 \left(\frac{n \exp\{g_\xi(x)\}}{m h_\psi(x) + n \exp\{g_\xi(x)\}} \right)$$

and Lemma 2 applies under (I1) to the sequence $(f_m)_{m \geq 1}$ converging pointwise towards 0, and dominated by the integrable function $g(x) = \sup_{\xi \in B(\widehat{\xi}_n, \varepsilon)} \|\nabla_\xi g_\xi(x)\| \left(\frac{\exp\{g_\xi(x)\}}{h_\psi(x)} \right)^2$.

The limit of $(m/n)\Delta_m$ is thus dictated by the behaviour of the first term. We apply Lemma 4 to the following vector-valued function, whose uniform norm is integrable under (I1) and under the continuity of the deterministic part assumed in (H2):

$$\varphi_g : (\xi, x) \mapsto \left(\frac{1}{n} \sum_{i=1}^{n} \nabla_\xi g_\xi(y_i) \frac{\exp\{g_\xi(y_i)\}}{h_\psi(y_i)} \right) - \nabla_\xi g_\xi(x) \left(\frac{\exp\{g_\xi(x)\}}{h_\psi(x)} \right)^2.$$

Lemma 4 yields $(m/n)\Delta_m \underset{m \to +\infty}{\longrightarrow} v(\widehat{\xi}_n)$ a.s. where

$$v(\xi) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\xi g_\xi(y_i) \left(\frac{\exp\{g_\xi(y_i)\}}{h_\psi(y_i)} \right) - \mathbb{E}_\psi \left[\nabla_\xi g_\xi(X) \left(\frac{\exp\{g_\xi(X)\}}{h_\psi(X)} \right)^2 \right].$$

Combination of these facts ensure that on a set of probability one, we have eventually:

$$\frac{m}{n} \left(\widehat{\xi}^{\mathrm{IS}}_{n,m} - \widehat{\xi}^{\mathrm{NCE}}_{n,m} \right) = o(1) + \left(-\mathbf{H}^{\mathrm{IS}}_m \right)^{-1} \left(\frac{m}{n} \Delta_m + o(1) \right) \underset{m \to \infty}{\to} \left(-\mathcal{H}(\widehat{\xi}_n) \right)^{-1} v(\widehat{\xi}_n).$$

2.7.4 Proof of Theorem 3

The proof of MC-MLE consistency under the considered regime is a very straightforward adaptation of Wald's proof of consistency for the MLE. We thus choose to present in appendix only the proof of NCE consistency, which is slightly more technical, although the sketch is similar. For the sake of completeness, a proof of MC-MLE consistency is presented in the supplement.

NCE consistency

For convenience, we choose to analyse a slightly different objective function (sharing the same maximiser with $\ell^{\mathrm{NCE}}_{n,m}$), defined as:

$$M^{\mathrm{NCE}}_n(\theta, \nu) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \varphi_{(\theta,\nu)}(Y_i) - \zeta^{(n)}_{(\theta,\nu)}(Y_i) \right\} - \left(\frac{m_n}{n} \right) \times \frac{1}{m_n} \sum_{j=1}^{m_n} \zeta^{(n)}_{(\theta,\nu)}(X_j) \tag{2.15}$$

where $\varphi_{(\theta,\nu)}(x) = \log\left\{\frac{e^\nu h_\theta(x)}{e^{\nu^\star} h_{\theta^\star}(x)}\right\}$ and $\zeta^{(n)}_{(\theta,\nu)}(x) = \log\left\{\frac{\frac{m_n}{n} h_\psi(x) + e^\nu h_\theta(x)}{\frac{m_n}{n} h_\psi(x) + e^{\nu^\star} h_{\theta^\star}(x)}\right\}$.

We begin our proof with the following lemma.

Lemma 8. *For any fixed (θ,ν), almost surely, $M_n^{\mathrm{NCE}}(\theta,\nu) \underset{n\to\infty}{\to} \mathcal{M}_\tau^{\mathrm{NCE}}(\theta,\nu)$, where:*

$$\mathcal{M}_\tau^{\mathrm{NCE}}(\theta,\nu) = \mathbb{E}_{\theta^\star}\left[\log\left\{\frac{e^\nu h_\theta}{e^{\nu^\star} h_{\theta^\star}}\right\} - \log\left\{\frac{\tau h_\psi + e^\nu h_\theta}{\tau h_\psi + e^{\nu^\star} h_{\theta^\star}}\right\}\right] - \tau \mathbb{E}_\psi\left[\log\left\{\frac{\tau h_\psi + e^\nu h_\theta}{\tau h_\psi + e^{\nu^\star} h_{\theta^\star}}\right\}\right]$$

Moreover, $(\theta^\star,\nu^\star)$ is the unique maximiser of $\mathcal{M}_\tau^{\mathrm{NCE}}(\theta,\nu)$.

Proof. For any fixed (θ,ν), the sequence $\zeta^{(n)}_{(\theta,\nu)}$ is eventually dominated (by a $\mathbb{P}_\psi$-integrable function), since for any $c > 0$ (in particular for $c = \tau \pm \varepsilon$) we have by Jensen's inequality:

$$\mathbb{E}_\psi\left[\log\left\{\frac{c h_\psi + e^\nu h_\theta}{c h_\psi + e^{\nu^\star} h_{\theta^\star}}\right\}\right] \geq \mathbb{E}_\psi\left[\log\left\{\frac{f_\psi}{f_\psi + \frac{1}{c} f_{\theta^\star}}\right\}\right] \geq -\log\left(1 + \frac{1}{c}\right) \tag{2.16}$$

$$\mathbb{E}_\psi\left[\log\left\{\frac{c h_\psi + e^\nu h_\theta}{c h_\psi + e^{\nu^\star} h_{\theta^\star}}\right\}\right] \leq \mathbb{E}_\psi\left[\log\left\{\frac{f_\psi + \frac{e^\nu \mathcal{Z}(\theta)}{c\mathcal{Z}(\psi)} f_\theta}{f_\psi}\right\}\right] \leq \log\left(1 + \frac{e^\nu \mathcal{Z}(\theta)}{c\mathcal{Z}(\psi)}\right) \tag{2.17}$$

Moreover, $\zeta^{(n)}_{(\theta,\nu)}$ converges pointwise to $\zeta^{\infty}_{(\theta,\nu)}(x) = \log\left\{\frac{\tau h_\psi(x) + e^\nu h_\theta(x)}{\tau h_\psi(x) + e^{\nu^\star} h_{\theta^\star}(x)}\right\}$, thus Lemma 2 applies: the second empirical average in (2.15) converges almost surely to $\mathbb{E}_\psi\left[\zeta^{\infty}_{(\theta,\nu)}(X)\right]$.

Now, the sequence $\left\{\varphi_{(\theta,\nu)} - \zeta^{(n)}_{(\theta,\nu)}\right\}$ is upper bounded by the positive part of $\varphi_{(\theta,\nu)}$ which is $\mathbb{P}_{\theta^\star}$-integrable. In particular, if $\mathbb{E}_{\theta^\star}\left[\left(\varphi_{(\theta,\nu)} - \zeta^{\infty}_{(\theta,\nu)}\right)_-\right] = +\infty$, then Lemma 3 applies and the first empirical average in (2.15) converges towards $-\infty$.

Conversely, suppose that $\mathbb{E}_{\theta^\star}\left[\left(\varphi_{(\theta,\nu)} - \zeta^{\infty}_{(\theta,\nu)}\right)_-\right] < \infty$. The law of large numbers would apply directly if the sequence m_n/n was exactly equal to τ. To handle this technical issue, we can consider the two following inequalities. Note that for any $a \geq b > 0$:

$$\log\left\{\frac{a h_\psi(x) + e^\nu h_\theta(x)}{a h_\psi(x) + e^{\nu^\star} h_{\theta^\star}(x)}\right\} \leq \log\left\{\frac{\frac{a}{b} b h_\psi(x) + \frac{a}{b} e^\nu h_\theta(x)}{b h_\psi(x) + e^{\nu^\star} h_{\theta^\star}(x)}\right\}$$

$$\log\left\{\frac{b h_\psi(x) + e^\nu h_\theta(x)}{b h_\psi(x) + e^{\nu^\star} h_{\theta^\star}(x)}\right\} \leq \log\left\{\frac{a h_\psi(x) + e^\nu h_\theta(x)}{\frac{b}{a} a h_\psi(x) + \frac{b}{a} e^{\nu^\star} h_{\theta^\star}(x)}\right\}$$

This yields a useful uniform bound for any $a, b > 0$:

$$\left|\log\left\{\frac{a h_\psi(x) + e^\nu h_\theta(x)}{a h_\psi(x) + e^{\nu^\star} h_{\theta^\star}(x)}\right\} - \log\left\{\frac{b h_\psi(x) + e^\nu h_\theta(x)}{b h_\psi(x) + e^{\nu^\star} h_{\theta^\star}(x)}\right\}\right| \leq \left|\log a - \log b\right| \tag{2.18}$$

Thus, if $\mathbb{E}_{\theta^\star}\left[\left(\varphi_{(\theta,\nu)} - \zeta^{\infty}_{(\theta,\nu)}\right)_-\right] < +\infty$, then the uniform bound (2.18) also ensures that:

$$\mathbb{E}_{\theta^\star}\left[\left(\varphi_{(\theta,\nu)} - \log\left\{\frac{c h_\psi + e^\nu h_\theta}{c h_\psi + e^{\nu^\star} h_{\theta^\star}}\right\}\right)_-\right] < +\infty$$

for any positive $c > 0$. The sequence can now be easily dominated and Lemma 2 applies; the first empirical average in (2.15) converges to $\mathbb{E}_{\theta^\star}\Big[\varphi_{(\theta,\nu)}(Y) - \zeta^\infty_{(\theta,\nu)}(Y)\Big]$.

Finally, let us prove that $(\theta^\star, \nu^\star)$ is the unique maximiser of $\mathcal{M}_\tau^{\mathrm{NCE}}$. We have:

$$\begin{aligned}
\mathcal{M}_\tau^{\mathrm{NCE}}(\theta,\nu) &= \frac{1}{\mathcal{Z}(\psi)}\Bigg[\int_{\mathcal{X}} -\log\left\{\frac{e^{\nu^\star}h_{\theta^\star}(x)}{e^{\nu}h_\theta(x)}\right\}e^{\nu^\star}h_{\theta^\star}(x) \\
&\qquad + \log\left\{\frac{\tau h_\psi(x) + e^{\nu^\star}h_{\theta^\star}(x)}{\tau h_\psi(x) + e^{\nu}h_\theta(x)}\right\}\Big(\tau h_\psi(x) + e^{\nu^\star}h_{\theta^\star}(x)\Big)\lambda(\mathrm{d}x)\Bigg] \\
&\leq \frac{1}{\mathcal{Z}(\psi)}\Bigg[\int_{\mathcal{X}} -\log\left\{\frac{e^{\nu^\star}h_{\theta^\star}(x)}{e^{\nu}h_\theta(x)}\right\}e^{\nu^\star}h_{\theta^\star}(x) \\
&\qquad + \log\left\{\frac{\tau h_\psi(x)}{\tau h_\psi(x)}\right\}\tau h_\psi(x) + \log\left\{\frac{e^{\nu^\star}h_{\theta^\star}(x)}{e^{\nu}h_\theta(x)}\right\}e^{\nu^\star}h_{\theta^\star}(x)\lambda(\mathrm{d}x)\Bigg] \\
&= 0
\end{aligned}$$

by the log-sum inequality, which applies with equality if and only if $e^{\nu}h_\theta(x) = e^{\nu^\star}h_{\theta^\star}(x)$ for $\mathbb{P}_{\theta^\star}$ almost every x. This occurs if and only if ν and θ are chosen such that $f_{\theta^\star}(x) = \frac{e^\nu}{\mathcal{Z}(\psi)}h_\theta(x)$. The model being identifiable, there is only one choice for both the unnormalized density and the normalizing constant; $\theta = \theta^\star$ and $\nu = \nu^\star$.

□

We now prove that the NCE estimator converges almost surely to this unique maximiser. Let $\eta > 0$, and define $K_\eta = \{\xi \in K : d(\xi, \xi^\star) \geq \eta\}$ where K is the compact set defined in (C2).

Under (H3), monotone convergence ensures that for any $\xi \in K_\eta$:

$$\lim_{\varepsilon \downarrow 0}\ \mathbb{E}_{\theta^\star}\left[\sup_{\beta \in B(\xi,\varepsilon)}\Big(\varphi_\beta(Y) - \zeta^\infty_\beta(Y)\Big)\right] = \mathbb{E}_{\theta^\star}\Big[\varphi_\xi(Y) - \zeta^\infty_\xi(Y)\Big]$$

and

$$\lim_{\varepsilon \downarrow 0}\ \mathbb{E}_{\psi}\left[\inf_{\beta \in B(\xi,\varepsilon)}\zeta^\infty_\beta(X)\right] = \mathbb{E}_{\psi}\Big[\zeta^\infty_\xi(X)\Big].$$

Indeed, since maps $\theta \mapsto h_\theta(x)$ are continuous, the two previous expectations (on the left hand side) are respectively bounded from above for ε small enough, and bounded from below for any ε.

Thus, for any $\xi \in K_\eta$ and any $\gamma > 0$ we can find $\varepsilon_\xi > 0$ such that simultaneously:

$$\mathbb{E}_{\theta^\star}\left[\sup_{\beta \in B(\xi,\varepsilon_\xi)}\Big(\varphi_\beta(Y) - \zeta^\infty_\beta(Y)\Big)\right] \leq \mathbb{E}_{\theta^\star}\Big[\varphi_\xi(Y) - \zeta^\infty_\xi(Y)\Big] + \frac{\gamma}{2}$$

$$\mathbb{E}_{\psi}\left[\inf_{\beta \in B(\xi,\varepsilon_\xi)}\zeta^\infty_\beta(X)\right] \geq \mathbb{E}_{\psi}\Big[\zeta^\infty_\xi(X)\Big] - \frac{\gamma}{2\tau}.$$

The compactness assumption ensures that there is a finite set $\{\xi_1, ..., \xi_p\} \subset K_\eta$ such that $K_\eta \subset \bigcup_{k=1}^p B(\xi_k, \varepsilon_{\xi_k})$. This yields the following inequality:

$$\sup_{\xi \in K_\eta} M_n^{\mathrm{NCE}}(\xi) \leq \max_{k=1,\dots,p} \left\{ \frac{1}{n} \sum_{i=1}^n \sup_{q \geq n} \sup_{\beta \in B(\xi_k, \varepsilon_{\xi_k})} \left(\varphi_\beta(Y_i) - \zeta_\beta^{(q)}(Y_i) \right) \right.$$
$$\left. - \left(\frac{m_n}{n} \right) \times \frac{1}{m_n} \sum_{j=1}^{m_n} \inf_{\beta \in B(\xi_k, \varepsilon_{\xi_k})} \zeta_\beta^{(n)}(X_j) \right\}$$

Choose any x for which the map $\theta \mapsto h_\theta(x)$ is continuous, and any $\xi \in K_\eta$. From the definition of $\zeta_\beta^{(n)}$, the following convergence is trivial:

$$\inf_{\beta \in B(\xi, \varepsilon_\xi)} \left(\zeta_\beta^{(n)}(x) \right) \underset{n \to +\infty}{\longrightarrow} \inf_{\beta \in B(\xi, \varepsilon_\xi)} \left(\zeta_\beta^{\infty}(x) \right).$$

Moreover, using inequalities (2.16) et (2.17), one can easily show that the sequence $\left\{ \inf_{\beta \in B(\xi, \varepsilon_\xi)} \zeta_\beta^{(n)} \right\}$ is dominated (by a $\mathbb{P}_\psi$-integrable function). Lemma 2 applies:

$$\frac{1}{m_n} \sum_{j=1}^{m_n} \inf_{\beta \in B(\xi_k, \varepsilon_{\xi_k})} \zeta_\beta^{(n)}(X_j) \underset{n \to +\infty}{\longrightarrow} \mathbb{E}_\psi \left[\inf_{\beta \in B(\xi, \varepsilon_\xi)} \zeta_\beta^{\infty}(X) \right] \qquad \text{a.s.}$$

Now, subadditivity of the supremum and inequality (2.18) yield

$$\left| \sup_{\beta \in B(\xi, \varepsilon_\xi)} \left(\varphi_\beta(x) - \zeta_\beta^{(n)}(x) \right) - \sup_{\beta \in B(\xi, \varepsilon_\xi)} \left(\varphi_\beta(x) - \zeta_\beta^{(\infty)}(x) \right) \right|$$
$$\leq \sup_{\beta \in B(\xi, \varepsilon_\xi)} \left| \zeta_\beta^{(n)}(x) - \zeta_\beta^{\infty}(x) \right| \leq \left| \log \frac{m_n}{n} - \log \tau \right| \underset{n \to +\infty}{\longrightarrow} 0$$

while monotonicity ensures that

$$\sup_{\beta \in B(\xi, \varepsilon_\xi)} \left(\varphi_\beta - \zeta_\beta^{\infty} \right) \leq \sup_{q \geq n} \sup_{\beta \in B(\xi, \varepsilon_\xi)} \left(\varphi_\beta - \zeta_\beta^{(q)} \right) \leq \sup_{\beta \in B(\xi, \varepsilon_\xi)} \left(\varphi_\beta \right)_+ .$$

In the last inequality, the right hand side is $\mathbb{P}_{\theta^\star}$-integrable under (H3), and the sequence (in the middle) converges pointwise towards its lower bound whose negative part has either finite or infinite expectation. In both cases, either Lemma 2 or Lemma 3 can be applied and ensures that, almost surely:

$$\frac{1}{n} \sum_{i=1}^n \sup_{q \geq n} \sup_{\beta \in B(\xi_k, \varepsilon_{\xi_k})} \left(\varphi_\beta(Y_i) - \zeta_\beta^{(q)}(Y_i) \right) \underset{n \to +\infty}{\longrightarrow} \mathbb{E}_{\theta^\star} \left[\sup_{\beta \in B(\xi, \varepsilon_\xi)} \left(\varphi_\beta(Y) - \zeta_\beta^{\infty}(Y) \right) \right]$$

Combining these convergences simultaneously on a finite set, we get almost surely:

$$\limsup_{n \to +\infty} \sup_{\xi \in K_\eta} M_n^{\mathrm{NCE}}(\xi) \leq \max_{k=1,\dots,p} \left\{ \mathbb{E}_{\theta^\star} \left[\sup_{\beta \in B(\xi_k, \varepsilon_{\xi_k})} \left(\varphi_\beta(Y) - \zeta_\beta^{\infty}(Y) \right) \right] \right.$$
$$\left. - \tau \mathbb{E}_\psi \left[\inf_{\beta \in B(\xi_k, \varepsilon_{\xi_k})} \zeta_\beta^{\infty}(X) \right] \right\}$$
$$\leq \sup_{\xi \in K_\eta} \mathcal{M}_\tau^{\mathrm{NCE}}(\xi) + \gamma$$

This leads to the following inequality since γ is arbitrary small:

$$\limsup_{n\to+\infty} \sup_{\xi\in K_\eta} M_n^{\mathrm{NCE}}(\xi) \leq \sup_{\xi\in K_\eta} \mathcal{M}_\tau^{\mathrm{NCE}}(\xi) \qquad \text{a.s.} \tag{2.19}$$

This last inequality is the heart of the proof. To conclude, we need only to show that the right hand side is negative, this is the aim of the following lemma.

Lemma 9. *Under (H3), the map $\xi \mapsto \mathcal{M}_\tau^{\mathrm{NCE}}(\xi)$ is upper semi continuous.*

The proof of Lemma 9 is straightforward. For the sake of completeness, we present a proof in the supplement.

Since an upper semi continuous function achieves its maximum on any compact set, this lemma proves in particular that $\sup_{\xi\in K_\eta} \mathcal{M}_\tau^{\mathrm{NCE}}(\xi) < 0$.

Thus inequality (2.19) implies that we can always find some $\alpha < 0$ such that eventually $\sup_{\xi\in K_\eta} M_n^{\mathrm{NCE}}(\xi) < \alpha$, while (C2) implies that $M_n^{\mathrm{NCE}}(\widehat{\xi}_{n,m}^{\mathrm{IS}}) \geq \sup_{\xi\in\Xi} M_n^{\mathrm{NCE}}(\xi) - \delta_n$ where $\delta_n \to 0$, and where

$$\sup_{\xi\in\Xi} M_n^{\mathrm{NCE}}(\xi) \geq M_n^{\mathrm{NCE}}(\xi^\star) \xrightarrow[n\to+\infty]{a.s.} \mathcal{M}^{\mathrm{NCE}}(\xi^\star) = 0.$$

Combination of these facts show that with probability one we have eventually:

$$M_n^{\mathrm{NCE}}(\widehat{\xi}_{n,m}^{\mathrm{IS}}) > \alpha > \sup_{\xi\in K_\eta} M_n^{\mathrm{NCE}}(\xi).$$

This is enough to prove strong consistency. Indeed, with probability one, $\widehat{\xi}_{n,m}^{\mathrm{NCE}}$ eventually escapes from K_η (otherwise there would be a contradiction with the inequality above). Since the sequence belongs to K by assumption, the sequence has no choice but to stay eventually in the ball of radius η. Thus with probability one, for any $\eta > 0$, we have eventually $d(\widehat{\xi}_{n,m}^{\mathrm{NCE}}, \xi^\star) < \eta$. This is the definition of almost sure convergence.

2.7.5 Proof of Theorem 4

The proof of MC-MLE asymptotic normality is entirely classical. We choose to present in appendix only the proof of NCE asymptotic normality, which follows the same sketch but is slightly more technical. For the sake of completeness, a proof of MC-MLE asymptotic normality is presented in the supplement.

NCE asymptotic normality

Let $G_n^{\mathrm{NCE}}(\xi) = \nabla_\xi \ell_{n,m}^{\mathrm{NCE}}(\xi)$ and $\mathbf{H}_n^{\mathrm{NCE}}(\xi) = \nabla_\xi^2 \ell_{n,m}^{\mathrm{NCE}}(\xi)$. We have:

$$\begin{aligned} G_n^{\mathrm{NCE}}(\xi) = &\frac{1}{n}\sum_{i=1}^{n} \nabla_\xi g_\xi(Y_i) \left(\frac{m_n h_\psi}{m_n h_\psi + n\exp\{g_\xi\}}\right)(Y_i) \\ &- \frac{1}{m_n}\sum_{j=1}^{m_n} \nabla_\xi g_\xi(X_j) \frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)} \left(\frac{m_n h_\psi}{m_n h_\psi + n\exp\{g_\xi\}}\right)(X_j) \end{aligned}$$

$$
\begin{aligned}
\mathbf{H}_n^{\mathrm{NCE}}(\xi) = & \frac{1}{n}\sum_{i=1}^{n} \nabla_\xi^2 g_\xi(Y_i)\left(\frac{m_n h_\psi}{m_n h_\psi + n\exp\{g_\xi\}}\right)(Y_i) \\
& - \frac{1}{n}\sum_{i=1}^{n} \nabla_\xi \nabla_\xi^T g_\xi(Y_i)\left(\frac{m_n h_\psi n\exp\{g_\xi\}}{(m_n h_\psi + n\exp\{g_\xi\})^2}\right)(Y_i) \\
& - \frac{1}{m_n}\sum_{j=1}^{m_n}\left\{(\nabla_\xi^2 + \nabla_\xi\nabla_\xi^T) g_\xi(X_j)\right\}\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\left(\frac{m_n h_\psi}{m_n h_\psi + n\exp\{g_\xi\}}\right)(X_j) \\
& + \frac{1}{m_n}\sum_{j=1}^{m_n} \nabla_\xi \nabla_\xi^T g_\xi(X_j)\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\left(\frac{m_n h_\psi n\exp\{g_\xi\}}{(m_n h_\psi + n\exp\{g_\xi\})^2}\right)(X_j)
\end{aligned}
$$

We firstly show that the study can be reduced to the following random sequences:

$$
\begin{aligned}
G_n^\tau(\xi) = & \frac{1}{n}\sum_{i=1}^{n} \nabla_\xi g_\xi(Y_i)\left(\frac{\tau h_\psi}{\tau h_\psi + \exp\{g_\xi\}}\right)(Y_i) \\
& - \frac{1}{m_n}\sum_{j=1}^{m_n} \nabla_\xi g_\xi(X_j)\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\left(\frac{\tau h_\psi}{\tau h_\psi + \exp\{g_\xi\}}\right)(X_j)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{H}_n^{\tau}(\xi) = & \frac{1}{n}\sum_{i=1}^{n} \nabla_\xi^2 g_\xi(Y_i)\left(\frac{\tau h_\psi}{\tau h_\psi + \exp\{g_\xi\}}\right)(Y_i) \\
& - \frac{1}{n}\sum_{i=1}^{n} \nabla_\xi \nabla_\xi^T g_\xi(Y_i)\left(\frac{\tau h_\psi \exp\{g_\xi\}}{(\tau h_\psi + \exp\{g_\xi\})^2}\right)(Y_i) \\
& - \frac{1}{m_n}\sum_{j=1}^{m_n}\left\{(\nabla_\xi^2 + \nabla_\xi\nabla_\xi^T) g_\xi(X_j)\right\}\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\left(\frac{\tau h_\psi}{\tau h_\psi + \exp\{g_\xi\}}\right)(X_j) \\
& + \frac{1}{m_n}\sum_{j=1}^{m_n} \nabla_\xi \nabla_\xi^T g_\xi(X_j)\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\left(\frac{\tau h_\psi \exp\{g_\xi\}}{(\tau h_\psi + \exp\{g_\xi\})^2}\right)(X_j)
\end{aligned}
$$

To do so, we show that almost surely $\sup_{\xi\in B(\xi^\star,\varepsilon)} \|\mathbf{H}_n^{\mathrm{NCE}}(\xi) - \mathbf{H}_n^{\tau}(\xi)\| \underset{n\to\infty}{\to} 0$.

Splitting the uniform norm into four parts yields:

$$
\begin{aligned}
\sup_{\xi\in B(\xi^\star,\varepsilon)}\left\|\mathbf{H}_n^{\mathrm{NCE}}(\xi) - \mathbf{H}_n^{\tau}(\xi)\right\| \leq & \frac{1}{n}\sum_{i=1}^{n} \sup_{\xi\in B(\xi^\star,\varepsilon)}\left\|\nabla_\xi^2 g_\xi(Y_i)\right\|\eta_n^\tau(Y_i) \\
& + \frac{1}{n}\sum_{i=1}^{n} \sup_{\xi\in B(\xi^\star,\varepsilon)}\left\|\nabla_\xi \nabla_\xi^T g_\xi(Y_i)\right\|\mathbf{\Gamma}_n^\tau(Y_i) \\
& + \frac{1}{m_n}\sum_{j=1}^{m_n} \sup_{\xi\in B(\xi^\star,\varepsilon)}\left\|(\nabla_\xi^2 + \nabla_\xi\nabla_\xi^T) g_\xi(X_j)\right\|\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\eta_n^\tau(X_j) \\
& + \frac{1}{m_n}\sum_{j=1}^{m_n} \sup_{\xi\in B(\xi^\star,\varepsilon)}\left\|\nabla_\xi \nabla_\xi^T g_\xi(X_j)\right\|\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}\gamma_n^\tau(X_j) \qquad (2.20)
\end{aligned}
$$

where the sequences of functions

$$
\eta_n^\tau = \sup_{\xi\in B(\xi^\star,\varepsilon)}\left|\frac{m_n h_\psi}{m_n h_\psi + n\exp\{g_\xi\}} - \frac{\tau h_\psi}{\tau h_\psi + \exp\{g_\xi\}}\right|
$$

and

$$\gamma_n^\tau = \sup_{\xi \in B(\xi^\star, \varepsilon)} \left| \frac{m_n h_\psi n \exp\{g_\xi\}}{(m_n h_\psi + n \exp\{g_\xi\})^2} - \frac{\tau h_\psi \exp\{g_\xi\}}{(\tau h_\psi + \exp\{g_\xi\})^2} \right|$$

are both upper bounded by 1 and converge pointwise (for any $x \in \mathcal{X}$) to 0 (use the continuity of $\xi \mapsto g_\xi(x)$).

Lemma 2 applies to each empirical average in (2.20) (every integrability condition holds under (H4)). The sum converges to 0 almost surely.

Now, we prove that $\forall a \in \mathbb{R}^{d+1} \quad a^T \sqrt{n}\Big(G_n^{\mathrm{NCE}}(\xi^\star) - G_n^\tau(\xi^\star)\Big) \xrightarrow{\mathbb{P}} 0$.

Define $\eta_{\theta,\tau}^{(n)} = \frac{m_n f_\psi}{m_n f_\psi + n f_\theta} - \frac{\tau f_\psi}{\tau f_\psi + f_\theta}$. At the point $\xi = \xi^\star$ we have:

$$\begin{aligned}\sqrt{n}\Big(G_n^{\mathrm{NCE}}(\xi^\star) - G_n^\tau(\xi^\star)\Big) &= \sqrt{n}\left\{ \frac{1}{n}\sum_{i=1}^n \Big(\nabla_\xi g_\xi\Big)\eta_{\theta,\tau}^{(n)}(Y_i) - \mathbb{E}_\theta\left[\Big(\nabla_\xi g_\xi\Big)\eta_{\theta,\tau}^{(n)}(Y)\right]\right\}_{|\xi=\xi^\star} \\ &\quad - \sqrt{\frac{n}{m_n}} \times \sqrt{m_n}\left\{ \frac{1}{m_n}\sum_{j=1}^{m_n} \Big(\nabla_\xi g_\xi\Big)\frac{f_\theta}{f_\psi}\eta_{\theta,\tau}^{(n)}(X_j) - \mathbb{E}_\theta\left[\Big(\nabla_\xi g_\xi\Big)\eta_{\theta,\tau}^{(n)}(Y)\right]\right\}_{|\xi=\xi^\star}\end{aligned}$$

The sequence $\left|\eta_{\theta,\tau}^{(n)}\right|$ is upper bounded by 1 and converges pointwise towards 0. Moreover, for any $c > \tau$, the sequence $\left|\eta_{\theta,\tau}^{(n)}\right|$ is also eventually upper bounded by $2\frac{cf_\psi}{cf_\psi + f_\theta}$. This ensures that both second moment conditions required holds under (H4) since:

$$\begin{aligned}\int_{\mathcal{X}} \|\nabla_\xi g_\xi\|^2 \left(\frac{f_\theta}{f_\psi}\right)^2 \left(\frac{cf_\psi}{cf_\psi + f_\theta}\right)^2 f_\psi \mathrm{d}\mu &= c\int_{\mathcal{X}} \|\nabla_\xi g_\xi\|^2 \left(\frac{cf_\psi}{cf_\psi + f_\theta}\right)\left(\frac{f_\theta}{cf_\psi + f_\theta}\right) f_\theta \mathrm{d}\mu \\ &\leq c \times \mathbb{E}_\theta\left[\|\nabla_\xi g_\xi\|^2\right] < +\infty \qquad (2.21)\end{aligned}$$

We can thus apply Lemma 6:

$$\begin{aligned}\sqrt{n}\left\{ \frac{1}{n}\sum_{i=1}^n \Big(a^T\nabla_\xi g_\xi\Big)\eta_{\theta,\tau}^{(n)}(Y_i) - \mathbb{E}_\theta\left[\Big(a^T\nabla_\xi g_\xi\Big)\eta_{\theta,\tau}^{(n)}(Y)\right]\right\}_{|\xi=\xi^\star} &\xrightarrow{\mathbb{P}} 0 \\ \sqrt{m_n}\left\{ \frac{1}{m_n}\sum_{j=1}^{m_n} \Big(a^T\nabla_\xi g_\xi\Big)\frac{f_\theta}{f_\psi}\eta_{\theta,\tau}^{(n)}(X_j) - \mathbb{E}_\theta\left[\Big(a^T\nabla_\xi g_\xi\Big)\eta_{\theta,\tau}^{(n)}(Y)\right]\right\}_{|\xi=\xi^\star} &\xrightarrow{\mathbb{P}} 0\end{aligned}$$

Finally, Cramér-Wold's device applies: $\sqrt{n}\Big(G_n^{\mathrm{NCE}}(\xi^\star) - G_n^\tau(\xi^\star)\Big) \xrightarrow{\mathbb{P}} 0_{\mathbb{R}^{d+1}}$.

Now, we can work directly with G_n^τ and $\mathbf{H}_n^\tau$, which is much easier. Indeed, Lemma 4

yields $\sup_{\xi \in B(\xi^\star, \varepsilon)} \|\mathbf{H}_n^\tau(\xi) - \mathbf{H}_\tau(\xi)\| \underset{n\to\infty}{\to} 0$ almost surely, where:

$$\begin{aligned}
\mathbf{H}_\tau(\xi) &= \mathbb{E}_{\theta^\star}\left[\nabla_\xi^2 g_\xi(Y)\left(\frac{\tau h_\psi}{\tau h_\psi + \exp\{g_\xi\}}\right)(Y)\right] \\
&- \mathbb{E}_{\theta^\star}\left[\nabla_\xi \nabla_\xi^T g_\xi(Y)\left(\frac{\tau h_\psi \exp\{g_\xi\}}{(\tau h_\psi + \exp\{g_\xi\})^2}\right)(Y)\right] \\
&- \mathbb{E}_{\psi}\left[\left\{(\nabla_\xi^2 + \nabla_\xi \nabla_\xi^T) g_\xi(X)\right\}\frac{\exp\{g_\xi(X)\}}{h_\psi(X)}\left(\frac{\tau h_\psi}{\tau h_\psi + \exp\{g_\xi\}}\right)(X)\right] \\
&+ \mathbb{E}_{\psi}\left[\nabla_\xi \nabla_\xi^T g_\xi(X)\frac{\exp\{g_\xi(X)\}}{h_\psi(X)}\left(\frac{\tau h_\psi \exp\{g_\xi\}}{(\tau h_\psi + \exp\{g_\xi\})^2}\right)(X)\right]
\end{aligned}$$

The only condition required is that the supremum norm of each integrand is integrable, which is satisfied under (H4) (bound the ratios by one).

Note also that, at the point $\xi = \xi^\star$, functions $\mathbf{H}_\tau$ and $-\mathbf{J}_\tau$ coincide, where:

$$\mathbf{J}_\tau(\xi) = \mathbb{E}_\theta\left[(\nabla_\xi \nabla_\xi^T g_\xi)\left(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\right)(Y)\right]$$

A quick block matrix calculation shows that Schur's complement in $-\mathbf{J}_\tau(\xi)$ is proportional to:

$$\mathbf{I}_\tau(\theta) = \mathbb{V}_{X\sim\mathbb{Q}_\tau}\Big(\nabla_\theta \log h_\theta(X)\Big)$$

where $\mathbb{Q}_\tau$ refers to the probability measure whose density with respect to μ is defined as $q_\tau(x) \propto \frac{\tau f_\psi(x) f_\theta(x)}{\tau f_\psi(x) + f_\theta(x)}$. Note that $\mathbb{P}_\theta \ll \mathbb{Q}_\tau$ since the model is dominated by $\mathbb{P}_\psi$.

In particular, $\mathbf{J}_\tau(\xi^\star)$ is invertible if and only if $\mathbf{I}_\tau(\theta^\star)$ is invertible. Since $\mathbf{I}_\tau(\theta)$ is a covariance matrix, if it is not full rank, then $\nabla_\theta \log h_\theta(X)$ belongs to a hyperplane $\mathbb{Q}_\tau$-almost surely (and thus $\mathbb{P}_\theta$-almost surely). This contradicts assumption (H4) since the Fisher Information is full rank. Thus $\mathbf{I}_\tau(\theta^\star)$ and $\mathbf{J}_\tau(\xi^\star)$ are both invertible.

Now, we prove the weak convergence of the gradient:

$$\begin{aligned}
\sqrt{n} G_n^\tau(\xi^\star) = \sqrt{n}\Bigg(&\frac{1}{n}\sum_{i=1}^n (\nabla_\xi g_\xi)\left(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\right)(Y_i) - \mathbb{E}_\theta\left[(\nabla_\xi g_\xi)\left(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\right)(Y)\right]\Bigg)_{|\xi=\xi^\star} \\
-\sqrt{\frac{n}{m_n}}\sqrt{m_n}\Bigg(&\frac{1}{m_n}\sum_{j=1}^{m_n} (\nabla_\xi g_\xi)\frac{f_\theta}{f_\psi}\left(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\right)(X_j) - \mathbb{E}_\theta\left[(\nabla_\xi g_\xi)\left(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\right)(Y)\right]\Bigg)_{|\xi=\xi^\star}
\end{aligned}$$

Slutsky's lemma applies as follows. It is noteworthy that second moment conditions hold under (H4) only (use inequality (2.21)), whereas assumption (I3) is necessary for proving MC-MLE asymptotic normality (see 2.8.3).

$$\sqrt{n} G_n^\tau(\xi^\star) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0_{\mathbb{R}^{d+1}}, \mathbf{\Sigma}_\tau(\xi^\star) + \tau^{-1}\mathbf{\Gamma}_\tau(\xi^\star)\Big)$$

where

$$\mathbf{\Sigma}_\tau(\xi) = \mathbb{V}_\theta\left((\nabla_\xi g_\xi)\Big(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\Big)(Y)\right),$$
$$\mathbf{\Gamma}_\tau(\xi) = \mathbb{V}_\psi\left(\varphi_\xi^{\mathrm{NCE}}(X)\right) + 2\sum_{i=1}^{+\infty}\mathrm{Cov}\Big(\varphi_\xi^{\mathrm{NCE}}(X_0), \varphi_\xi^{\mathrm{NCE}}(X_i)\Big),$$
$$\varphi_\xi^{\mathrm{NCE}} = (\nabla_\xi g_\xi)\frac{f_\theta}{f_\psi}\Big(\frac{\tau f_\psi}{\tau f_\psi + f_\theta}\Big).$$

Finally, Lemma 5 applies:

$$\sqrt{n}\left(\widehat{\xi}_{n,m}^{\mathrm{NCE}} - \xi^\star\right) \xrightarrow{\mathcal{D}} \mathcal{N}_{d+1}\Big(0, \mathbf{V}_\tau^{\mathrm{NCE}}(\xi^\star)\Big)$$

where $\mathbf{V}_\tau^{\mathrm{NCE}}(\xi) = \mathbf{J}_\tau(\xi)^{-1}\left\{\mathbf{\Sigma}_\tau(\xi) + \tau^{-1}\mathbf{\Gamma}_\tau(\xi)\right\}\mathbf{J}_\tau(\xi)^{-1}$.

2.7.6 Proof of Theorem 5

For convenience, we will use some shorthand notations. Define the real-valued measurable functions $Q = f_\theta/f_\psi$ and $R = \tau f_\psi/(\tau f_\psi + f_\theta)$. Note that we have the relationship $QR = \tau(1-R)$. In the following, assume that $\xi = \xi^\star$, and for any measurable function φ, note that $\mathbb{E}_\theta[\varphi]$ stands for the expectation of $\varphi(X)$ where $X \sim \mathbb{P}_\theta$, and that $\nabla\nabla^T g_\xi$ stands for the measurable matrix-valued function $x \mapsto \nabla_\xi g_\xi(x)(\nabla_\xi g_\xi(x))^T$. We begin with the following computations:

$$\begin{aligned}
\mathbf{J}(\xi) &= \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi\Big],\\
\mathbf{\Sigma}(\xi) &= \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi\Big] - \mathbb{E}_\theta\Big[\nabla g_\xi\Big]\mathbb{E}_\theta\Big[\nabla^T g_\xi\Big],\\
\mathbf{\Gamma}(\xi) &= \mathbb{E}_\psi\Big[\nabla\nabla^T g_\xi Q^2\Big] - \mathbb{E}_\psi\Big[\nabla g_\xi Q\Big]\mathbb{E}_\psi\Big[\nabla^T g_\xi Q\Big]\\
&= \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi (R^{-1}-1)\Big] \times \tau - \mathbb{E}_\theta\Big[\nabla g_\xi\Big]\mathbb{E}_\theta\Big[\nabla^T g_\xi\Big],\\
\mathbf{J}_\tau(\xi) &= \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R\Big],\\
\mathbf{\Sigma}_\tau(\xi) &= \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R^2\Big] - \mathbb{E}_\theta\Big[\nabla g_\xi R\Big]\mathbb{E}_\theta\Big[\nabla^T g_\xi R\Big],\\
\mathbf{\Gamma}_\tau(\xi) &= \mathbb{E}_\psi\Big[\nabla\nabla^T g_\xi Q^2 R^2\Big] - \mathbb{E}_\psi\Big[\nabla g_\xi QR\Big]\mathbb{E}_\psi\Big[\nabla^T g_\xi QR\Big]\\
&= \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R(1-R)\Big] \times \tau - \mathbb{E}_\theta\Big[\nabla g_\xi R\Big]\mathbb{E}_\theta\Big[\nabla^T g_\xi R\Big].
\end{aligned}$$

Fortunately, the expression of the asymptotic variances simplify, as shown by the following lemma.

Lemma 10. *Let Z be any real-valued, non-negative measurable function such that $\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big]$ is finite and invertible. Then:*

$$\mathbf{M} := \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big]^{-1}\mathbb{E}_\theta\Big[\nabla g_\xi Z\Big]\mathbb{E}_\theta\Big[\nabla^T g_\xi Z\Big]\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big]^{-1} = \begin{pmatrix} 0_{\mathbb{R}^{d\times d}} & 0_{\mathbb{R}^d} \\ 0_{\mathbb{R}^d}^T & 1 \end{pmatrix}.$$

The proof of Lemma 10 follows from a direct block matrix computation. For the sake of completeness, we present a proof in the supplement.

Let $\mathbf{M}$ be defined as in Lemma 10, matrix calculations yield

$$\begin{aligned}
\mathbf{J}(\xi)^{-1}\mathbf{\Sigma}(\xi)\mathbf{J}(\xi)^{-1} &= \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi\Big]^{-1} - \mathbf{M},\\
\mathbf{J}_\tau(\xi)^{-1}\mathbf{\Sigma}_\tau(\xi)\mathbf{J}_\tau(\xi)^{-1} &= \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R\Big]^{-1}\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R^2\Big]\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R\Big]^{-1} - \mathbf{M},\\
\mathbf{J}(\xi)^{-1}\mathbf{\Gamma}(\xi)\mathbf{J}(\xi)^{-1} &= \tau\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi\Big]^{-1}\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi(R^{-1}-1)\Big]\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi\Big]^{-1} - \mathbf{M},\\
\mathbf{J}_\tau(\xi)^{-1}\mathbf{\Gamma}_\tau(\xi)\mathbf{J}_\tau(\xi)^{-1} &= \tau\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R\Big]^{-1}\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R(1-R)\Big]\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R\Big]^{-1} - \mathbf{M}.
\end{aligned}$$

Summing up these expressions we finally get

$$\mathbf{V}_\tau^{\mathrm{IS}}(\xi) = \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi\Big]^{-1}\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R^{-1}\Big]\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi\Big]^{-1} - (1+\tau^{-1})\mathbf{M},$$

$$\mathbf{V}_\tau^{\mathrm{NCE}}(\xi) = \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R\Big]^{-1} - (1+\tau^{-1})\mathbf{M}.$$

Now, to compare these variances, the idea is the following: $(x,y)\mapsto x^2/y$ is a convex function on $\mathbb{R}^2$, which means Jensen's inequality ensures that for any random variables X, Y such that the following expectations exist we have $\mathbb{E}[X^2/Y] \geq \mathbb{E}[X]^2/\mathbb{E}[Y]$. Here the variances are matrices, but it turns out that it is possible to use a generalization of Jensen's inequality to the Loewner partial order on matrices. We introduce the following notations:

$\mathbb{M}_{n,m}$ is the set of $n\times m$ matrices,
$\mathbb{S}_n$ is the set of $n\times n$ symmetric matrices,
$\mathbb{S}_n^+$ is the set of ($n\times n$ symmetric) positive semi-definite matrices,
$\mathbb{S}_n^{++}$ is the set of ($n\times n$ symmetric) positive definite matrices,
$\mathcal{R}(A)$ is the range of A,
$\Delta_{n,m} = \Big\{(A,B)\in\mathbb{S}_n^+\times\mathbb{M}_{n,m} \;:\; \mathcal{R}(B)\subset\mathcal{R}(A)\Big\}$,
$A^\dagger$ denotes the Moore-Penrose pseudo-inverse of A,
$\succcurlyeq$ denotes the Loewner partial order ($A_1\succcurlyeq A_2$ iff $A_1-A_2\in\mathbb{S}_n^+$).

Lemma 11. *Let A,B be random matrices such that $(A,B)\in\Delta_{n,m}$ with probability one for some positive integers n,m. Let $\varphi : (A,B)\mapsto B^TA^\dagger B$ defined on $\Delta_{n,m}$. Then $\mathbb{E}[\varphi(A,B)]\succcurlyeq\varphi(\mathbb{E}[A],\mathbb{E}[B])$ provided that the three expectations exist.*

Proof. We just have to prove that f is convex with respect to $\succcurlyeq$, i.e. that for any $\lambda\in[0,1]$, and any $(A_1,B_1),(A_2,B_2)\in\Delta_{n,m}$ we have $\lambda\varphi(A_1,B_1)+(1-\lambda)\varphi(A_2,B_2)\succcurlyeq\varphi(\lambda(A_1,B_1)+(1-\lambda)(A_2,B_2))$. Indeed, if this convex relationship on matrices is satisfied then for any $x\in\mathbb{R}^m$, the real-valued map $q:(A,B)\mapsto x^T\varphi(A,B)x$ is necessarily convex

on $\Delta_{n,m}$. Consequently, Jensen's inequality applies, i.e. for any random $(A,B) \in \Delta_{n,m}$ a.s. and any $x \in \mathbb{R}^m$ we have

$$x^T \mathbb{E}[\varphi(A,B)]x = \mathbb{E}[q(A,B)] \geq q(\mathbb{E}[A], \mathbb{E}[B]) = x^T \varphi(\mathbb{E}[A], \mathbb{E}[B])x$$

which is the claim of the lemma.

Now, to prove that φ is convex with respect to $\succcurlyeq$, we use a property of the generalized Schur's complement in positive semi-definite matrices (see Boyd and Vandenberghe (2004) p.651): let $A \in \mathbb{S}_n, B \in \mathbb{M}_{n,m}, C \in \mathbb{S}_m$, and consider the block symmetric matrix

$$D = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}.$$

Then we have

$$D \succcurlyeq 0 \quad \Leftrightarrow \quad A \succcurlyeq 0 \ , \ \mathcal{R}(B) \subset \mathcal{R}(A) \ , \ C - B^T A^\dagger B \succcurlyeq 0.$$

This leads to a straightforward proof of the convexity of φ. To our knowledge, the following trick is due to Ando (1979), whose original proof was restricted to positive definite matrices. We use the generalized Schur's complement to extend this result to any $(A,B) \in \Delta_{n,m}$: let $\lambda \in [0,1]$, and $(A_1,B_1),(A_2,B_2) \in \Delta_{n,m}$. The sum of two positive semi definite matrices is positive semi-definite thus we have

$$\lambda \begin{pmatrix} A_1 & B_1 \\ B_1^T & B_1^T A_1^\dagger B_1 \end{pmatrix} + (1-\lambda) \begin{pmatrix} A_2 & B_2 \\ B_2^T & B_2^T A_2^\dagger B_2 \end{pmatrix} \succcurlyeq 0$$

which is the same as

$$\begin{pmatrix} \lambda A_1 + (1-\lambda)A_2 & \lambda B_1 + (1-\lambda)B_2 \\ \lambda B_1^T + (1-\lambda)B_2^T & \lambda B_1^T A_1^\dagger B_1 + (1-\lambda) B_2^T A_2^\dagger B_2 \end{pmatrix} \succcurlyeq 0.$$

Consequently, the generalised Schur's complement in this last block matrix is also positive semi-definite, i.e.

$$\lambda B_1^T A_1^\dagger B_1 + (1-\lambda) B_2^T A_2^\dagger B_2 \succcurlyeq \Big(\lambda B_1 + (1-\lambda)B_2\Big)^T \Big[\lambda A_1 + (1-\lambda)A_2\Big]^\dagger \Big(\lambda B_1 + (1-\lambda)B_2\Big)$$

which proves the convexity of φ with respect to $\succcurlyeq$, and thus the claim of the lemma. □

Finally, we compare the asymptotic variances of the two estimators. Note that for any $(A,B) \in \mathbb{S}_n \times \mathbb{S}_n^{++}$, and for every $x \in \mathbb{R}^n$, we have

$$x^T A x \geq 0 \quad \Leftrightarrow \quad x^T BABx \geq 0.$$

Indeed, if A is semi definite positive then for some integer k we can find $P \in \mathbb{M}_{k,n}$ such that $A = P^T P$, moreover, B being symmetric we have $x^T BABx = \|PBx\|^2 \geq 0$. The direct implication is enough since $B^{-1} \in \mathbb{S}_n^{++}$.

Consequently, the relation $\mathbf{V}_\tau^{\mathrm{IS}}(\xi) \succcurlyeq \mathbf{V}_\tau^{\mathrm{NCE}}(\xi)$ is equivalent to the relation

$$\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R^{-1}\Big] \succcurlyeq \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi\Big]\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi R\Big]^{-1}\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi\Big]. \tag{2.22}$$

Inequality (2.22) is a direct application of Lemma 11 (let $B = \nabla\nabla^T g_\xi$, $A = BR$; note that $(A,B) \in \Delta_{d+1,d+1}$ almost surely; and use basic properties of the pseudo-inverse).

2.8 Supplementary Proofs

2.8.1 Proofs of technical lemmas

Proof of Lemma 1

For all $k \in \mathbb{N}$, eventually (for any $m \geq k$) we have

$$\frac{1}{m}\sum_{j=1}^{m} f_k(X_j) \leq \frac{1}{m}\sum_{j=1}^{m} f_m(X_j) \leq \frac{1}{m}\sum_{j=1}^{m} f(X_j).$$

Moreover, since $(X_j)_{j\geq 1}$ is $\mathbb{P}_\psi$-ergodic, the law of large numbers applies (even if the expectations are infinite, since f_k and f are non-negative):

$$\frac{1}{m}\sum_{j=1}^{m} f_k(X_j) \xrightarrow[m\to+\infty]{a.s.} \mathbb{E}_\psi[f_k(X)] \quad \text{and} \quad \frac{1}{m}\sum_{j=1}^{m} f(X_j) \xrightarrow[m\to+\infty]{a.s.} \mathbb{E}_\psi[f(X)].$$

Thus, there is a set of probability one on which for every $k \in \mathbb{N}$,

$$\mathbb{E}_\psi[f_k(X)] \leq \liminf_{m\to+\infty} \frac{1}{m}\sum_{j=1}^{m} f_m(X_j) \leq \limsup_{m\to+\infty} \frac{1}{m}\sum_{j=1}^{m} f_m(X_j) \leq \mathbb{E}_\psi[f(X)].$$

Since the inequality holds for any $k \in \mathbb{N}$, it also holds for the supremum over k:

$$\sup_{k\in\mathbb{N}} \mathbb{E}_\psi[f_k(X)] \leq \liminf_{m\to+\infty} \frac{1}{m}\sum_{j=1}^{m} f_m(X_j) \leq \limsup_{m\to+\infty} \frac{1}{m}\sum_{j=1}^{m} f_m(X_j) \leq \mathbb{E}_\psi[f(X)].$$

Finally, the monotone convergence theorem yields

$$\sup_{k\in\mathbb{N}} \mathbb{E}_\psi[f_k(X)] = \lim_{k\to+\infty} \mathbb{E}_\psi[f_k(X)] = \mathbb{E}_\psi\Big[\lim_{k\to+\infty} f_k(X)\Big] = \mathbb{E}_\psi[f(X)].$$

Consequently, the lower and upper limits are both equal to $\mathbb{E}_\psi[f(X)]$ almost surely.

Proof of Lemma 2

Since $(X_j)_{j\geq 1}$ is $\mathbb{P}_\psi$-ergodic and f is dominated by the integrable function g, the law of large numbers applies to function f. Thus, we just need to prove that

$$\left|\frac{1}{m}\sum_{j=1}^{m} \{f_m(X_j) - f(X_j)\}\right| \xrightarrow[m\to+\infty]{a.s.} 0.$$

To do so, use the fact that

$$\begin{aligned}\left|\frac{1}{m}\sum_{j=1}^{m} \{f_m(X_j) - f(X_j)\}\right| &\leq \frac{1}{m}\sum_{j=1}^{m} |f_m(X_j) - f(X_j)| \\ &\leq \frac{1}{m}\sum_{j=1}^{m} \sup_{k\geq m} \Big|f_k(X_j) - f(X_j)\Big|.\end{aligned}$$

Define $h_m = 2g - \sup_{k \geq m} |f_k - f|$ and note that $(h_m)_{m \geq 1}$ is a non-decreasing sequence of non negative functions converging pointwise towards $2g$. Lemma 1 yields

$$\frac{1}{m} \sum_{j=1}^{m} h_m(X_j) \xrightarrow[m \to +\infty]{a.s.} \mathbb{E}_\psi[2g(X)].$$

Finally g is integrable, thus the remainder converges almost surely towards zero:

$$\frac{1}{m} \sum_{j=1}^{m} \sup_{k \geq m} \left| f_k(X_j) - f(X_j) \right| = \frac{2}{m} \sum_{j=1}^{m} g(X_j) - \frac{1}{m} \sum_{j=1}^{m} h_m(X_j) \xrightarrow[m \to +\infty]{a.s.} 0.$$

Proof of Lemma 3

Since g is integrable and $(X_j)_{j \geq 1}$ is $\mathbb{P}_\psi$-ergodic we have

$$\frac{1}{m} \sum_{j=1}^{m} g(X_j) \xrightarrow[m \to +\infty]{a.s.} \mathbb{E}_\psi\Big[g(X)\Big] < +\infty.$$

Thus we only need to show that

$$\frac{1}{m} \sum_{j=1}^{m} \{g(X_j) - f_m(X_j)\} \xrightarrow[m \to +\infty]{a.s.} +\infty.$$

Define $h_m = g - \sup_{k \geq m} f_k$, an increasing sequence of non negative functions converging pointwise to $g - f$. Lemma 1 applies whether $g - f$ is integrable or not:

$$\frac{1}{m} \sum_{j=1}^{m} (g(X_j) - f_m(X_j)) \geq \frac{1}{m} \sum_{j=1}^{m} h_m(X_j) \xrightarrow[m \to +\infty]{a.s.} \mathbb{E}_\psi\Big[g(X) - f(X)\Big].$$

The following inequality shows that the expectation is indeed infinite:

$$\mathbb{E}_\psi\Big[g(X) - f(X)\Big] = \mathbb{E}_\psi\Big[\Big(g(X) - f(X)_+\Big) + f(X)_-\Big] \geq \mathbb{E}_\psi\Big[f(X)_-\Big] = +\infty.$$

Proof of Lemma 4

To begin, note that measurability of the supremum is ensured by the lower semi-continuity of the maps $\theta \mapsto \varphi(\theta, x)$ on a set of probability one that does not depend on θ.

For every $\theta \in K$, consider the following function:

$$f_\theta(\eta) = \mathbb{E}_\psi\left[\sup_{\phi \in B(\theta, \eta)} \|\varphi(\phi, X) - \varphi(\theta, X)\|\right].$$

Dominated convergence implies that $f_\theta(\eta)$ converges to zero when η goes to zero. This is enough to ensure the continuity of the map $\theta \mapsto \mathbb{E}_\psi\Big[\varphi(\theta, X)\Big]$ because of the following inequality:

$$\sup_{\phi \in B(\theta, \eta)} \|\mathbb{E}_\psi\Big[\varphi(\phi, X) - \varphi(\theta, X)\Big]\| \leq f_\theta(\eta).$$

Let $\varepsilon > 0$. For every $\theta \in K$, we can always find $\eta_{(\theta,\varepsilon)} > 0$ small enough such that $f_\theta(\eta_{(\theta,\varepsilon)}) < \varepsilon$. Note that the open balls centered on $\theta \in K$ of radius $\eta_{(\theta,\varepsilon)}$, form an open cover of K, from which we can extract a finite subcover thanks to the compactness assumption. Thus we can build a finite set $\{\phi_1, ..., \phi_I\} \subset K$ (centers of the balls) such that

$$K \subset \bigcup_{i=1}^{I} B_i, \qquad B_i = B(\phi_i, \eta_{(\phi_i,\varepsilon)}).$$

Now, for any $\theta \in K$ define i_θ as the smallest integer $i \in \{1, ..., I\}$ such that $\theta \in B_i$, and consider the following equality:

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m} \varphi(\theta, X_j) - \mathbb{E}_\psi\Big[\varphi(\theta, X)\Big] = \frac{1}{m}\sum_{j=1}^{m} \Big\{\varphi(\theta, X_j) - \varphi(\phi_{i_\theta}, X_j)\Big\} \\ + \frac{1}{m}\sum_{j=1}^{m} \varphi(\phi_{i_\theta}, X_j) - \mathbb{E}_\psi\Big[\varphi(\phi_{i_\theta}, X)\Big] \\ + \mathbb{E}_\psi\Big[\varphi(\phi_{i_\theta}, X)\Big] - \mathbb{E}_\psi\Big[\varphi(\theta, X)\Big]\end{aligned}$$

The three last terms are functions of θ for which we want to bound the uniform norm.

The uniform norm of the third term is lower than ε since $\forall \theta \in K$, $d(\theta, \phi_{i_\theta}) \leq \eta_{(\phi_{i_\theta},\varepsilon)}$. The second term converges to zero by the law of large number since $\{\phi_1, ..., \phi_J\}$ is finite. Finally, the uniform norm of the second term can be bounded by

$$U_m = \max_{1\leq i\leq I} \left\{\frac{1}{m}\sum_{j=1}^{m} \sup_{\theta\in B_i} \|\varphi(\theta, X_j) - \varphi(\phi_i, X_j)\|\right\}.$$

The supremum are integrable by assumption, thus the law of large numbers applies:

$$U_m \xrightarrow[m\to+\infty]{a.s.} \max_{1\leq i\leq I} f_{\phi_i}(\eta_{(\phi_i,\varepsilon)}) < \varepsilon.$$

To sum up, we have just proven that for any $\varepsilon > 0$, almost surely,

$$\limsup_{m\to+\infty} \left\{\sup_{\theta\in K} \left\|\frac{1}{m}\sum_{j=1}^{m} \varphi(\theta, X_j) - \mathbb{E}_\psi\Big[\varphi(\theta, X)\Big]\right\|\right\} < 2\varepsilon.$$

Since ε is arbitrary small, we get the first claim of the lemma.

Now, if $\widetilde{\theta}_m \to \widetilde{\theta}$, we have eventually $\|\widetilde{\theta}_m - \widetilde{\theta}\| \leq \varepsilon$ with probability one. This yields the following inequality for m large enough:

$$\begin{aligned}\left\|\frac{1}{m}\sum_{j=1}^{m} \varphi(\widetilde{\theta}_m, X_j) - \mathbb{E}_\psi\Big[\varphi(\widetilde{\theta}, X)\Big]\right\| \leq \sup_{\theta\in B(\widetilde{\theta}_m,\varepsilon)} \left\|\frac{1}{m}\sum_{j=1}^{m} \varphi(\theta, X_j) - \mathbb{E}_\psi\Big[\varphi(\theta, X)\Big]\right\| \\ + \left\|\mathbb{E}_\psi\Big[\varphi(\widetilde{\theta}_m, X)\Big] - \mathbb{E}_\psi\Big[\varphi(\widetilde{\theta}, X)\Big]\right\|.\end{aligned}$$

The first term converges to zero since the first claim of the lemma applies to the compact closure of $B(\widetilde{\theta}, \varepsilon)$. The continuity of the map $\theta \mapsto \mathbb{E}_\psi\Big[\varphi(\theta, X)\Big]$ ensures that the second term also goes to zero, proving the second claim of the Lemma.

Proof of Lemma 5

Let $\varepsilon > 0$, and $G_n(\theta, \omega) = \nabla_\theta \ell_n(\theta, \omega)$ defined on $B(\theta^\star, \varepsilon)$. Define also $g_k^{(n)}(\theta)$ as the k-th component of $G_n(\theta)$. By assumption, for any $\delta > 0$,

$$\left\{\omega \in \Omega : \max\left(\|\widehat{\theta}_n - \theta^\star\| \ , \ \|\sqrt{n} G_n(\widehat{\theta}_n)\| \ , \ \sup_{\theta \in B(\theta^\star, \varepsilon)} \|\nabla_\theta^2 \ell_n(\theta) - \mathcal{H}(\theta)\|\right) \leq \delta\right\}$$

defines a sequence of sets whose probability goes to one.

On any of these sets (for a fixed ω), Taylor Lagrange's theorem ensures that we can find $(\widetilde{\theta}_j^{(n)})_{j=1,\dots,d}$ on the segment line $[\theta^\star, \widehat{\theta}_n]$ such that

$$G_n(\widehat{\theta}_n) = G_n(\theta^\star) + \mathbf{H}_n(\widehat{\theta}_n - \theta^\star), \qquad \mathbf{H}_n = \begin{pmatrix} \left(\nabla_\theta g_1^{(n)}(\widetilde{\theta}_1^{(n)})\right)^T \\ \vdots \\ \left(\nabla_\theta g_d^{(n)}(\widetilde{\theta}_d^{(n)})\right)^T \end{pmatrix}.$$

In particular, for any $\delta \in]0, \varepsilon[$,

$$\|\mathbf{H}_n - \mathcal{H}(\theta^\star)\| \leq d \sup_{\theta \in B(\theta^\star, \varepsilon)} \|\nabla_\theta^2 \ell_n(\theta) - \mathcal{H}(\theta)\| + \sum_{j=1}^{d} \|\mathcal{H}(\widetilde{\theta}_j^{(n)}) - \mathcal{H}(\theta^\star)\|.$$

For any $j = 1, ..., d$, the distance between $\widetilde{\theta}_j^{(n)}$ and $\theta^\star$ is at most δ, and $\mathcal{H}$ is continuous, thus δ can always be chosen small enough such that $\mathbf{H}_n$ is invertible. We thus have:

$$\widehat{\theta}_n - \theta^\star = \mathbf{H}_n^{-1}\Big\{G_n(\widehat{\theta}_n) - G_n(\theta^\star)\Big\}$$

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) + \mathcal{H}(\theta^\star)^{-1}\sqrt{n} G_n(\theta^\star) = \mathbf{H}_n^{-1}\sqrt{n} G_n(\widehat{\theta}_n) - \Big\{\mathbf{H}_n^{-1} - \mathcal{H}(\theta^\star)^{-1}\Big\}\sqrt{n} G_n(\theta^\star)$$

The right hand side converges in probability to zero because $G_n(\widehat{\theta}_n) = o_{\mathbb{P}}(n^{-1/2})$ by assumption, and because $\sqrt{n} G_n(\theta^\star)$ converges in distribution and is thus bounded in probability. The last equality being true on a sequence of sets whose probability goes to one, this implies that the left hand side must also converge to zero in probability.

The last conclusion follows from Slutsky's lemma.

Proof of Lemma 6

Before proving the lemma, we recall a powerful result from Jones (2004). Under (X2), the chain $(X_j)_{j \geq 1}$ is asymptotically uncorrelated with exponential decay, i.e. there is some $\gamma > 0$ such that

$$\rho(n) = \sup\Big\{\text{Corr}(U, V) \ , \ U \in L^2(\mathcal{F}_1^k) \ , \ V \in L^2(\mathcal{F}_{k+n}^\infty) \ , \ k \geq 1\Big\} = \mathcal{O}(e^{-\gamma n})$$

where $\mathcal{F}_k^m$ is the sigma-algebra generated by $X_k, ..., X_m$.

Let $h_n = f_n - f$, and note that $\mathbb{V}_\psi(h_n(X_0)) \leq \mathbb{E}_\psi\Big[(h_n(X_0))^2\Big] \underset{n\to\infty}{\to} 0$ by dominated convergence. Combined with the previous result, this implies that

$$\frac{1}{n}\mathbb{V}\left(\sum_{i=1}^n h_n(X_i)\right) = \mathbb{V}_\psi(h_n(X_0)) \times \left\{1 + 2\sum_{i=1}^n \frac{n-i}{n}\mathrm{Corr}(h_n(X_0), h_n(X_i))\right\} \underset{n\to\infty}{\to} 0$$

since

$$\left|\sum_{i=1}^n \frac{n-i}{n}\mathrm{Corr}(h_n(X_0), h_n(X_i))\right| \leq \sum_{i=1}^{+\infty} \rho(i) < +\infty.$$

The first claim of the lemma follows from Chebyshev's inequality, since for any $\varepsilon > 0$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n h_n(X_i) - \mathbb{E}\Big[h_n(X)\Big] \geq \frac{\varepsilon}{\sqrt{n}}\right) \leq \frac{1}{n\varepsilon^2}\mathbb{V}\left(\sum_{i=1}^n h_n(X_i)\right) \underset{n\to\infty}{\to} 0.$$

Finally, under (X2) a $\sqrt{n}$-CLT holds for f dominated by g

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \sigma_f^2\Big).$$

An application of Slutsky's lemma yields the second claim of the lemma.

2.8.2 Proofs of the remaining lemmas

Proof of Lemma 7

Assumption (H2) ensures in particular that the partition function $\theta \mapsto \mathcal{Z}(\theta)$ is differentiable in a neighborhood of $\widehat{\theta}_n$. Write the Hessian of the Poisson Transform as the following block matrix:

$$\nabla^2_{(\theta,\nu)}\ell_n(\theta,\nu) = \begin{pmatrix} A & b \\ b^T & c \end{pmatrix}$$

where

$$\begin{aligned}
A &= \nabla^2_\theta \ell_n(\theta,\nu) = \frac{1}{n}\sum_{i=1}^n \nabla^2_\theta \log h_\theta(y_i) - e^\nu \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)}\frac{\nabla^2_\theta \mathcal{Z}(\theta)}{\mathcal{Z}(\theta)}, \\
b &= \nabla_\theta \frac{\partial}{\partial\nu}\ell_n(\theta,\nu) = -e^\nu \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)}\frac{\nabla_\theta \mathcal{Z}(\theta)}{\mathcal{Z}(\theta)}, \\
c &= \frac{\partial^2}{\partial\nu^2}\ell_n(\theta,\nu) = -e^\nu \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)} < 0.
\end{aligned}$$

The Hessian of the Poisson transform is negative definite if and only if Schur's complement of c in the Hessian also is. Use the following equality to compute it:

$$\nabla^2_\theta \log \mathcal{Z}(\theta) = \frac{\nabla^2_\theta \mathcal{Z}(\theta)}{\mathcal{Z}(\theta)} - \frac{\nabla_\theta \mathcal{Z}(\theta)\Big(\nabla_\theta \mathcal{Z}(\theta)\Big)^T}{\mathcal{Z}(\theta)^2},$$

$$A - bc^{-1}b^T = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 \log h_\theta(y_i) - e^\nu \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)} \nabla_\theta^2 \log \mathcal{Z}(\theta).$$

At the point $\xi = \widehat{\xi}_n$, Schur's complement of c is also the Hessian of the log likelihood:

$$\nabla_\theta^2 \ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 \log h_\theta(y_i) - \nabla_\theta^2 \log \mathcal{Z}(\theta), \qquad \left\{ e^\nu \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)} \right\}\Big|_{\xi=\widehat{\xi}_n} = 1.$$

Proof of Lemma 9

Let $\xi_n \to \xi$, we have

$$\lim_{n\to+\infty} \sup_{k\geq n} \mathcal{M}_\tau^{\mathrm{NCE}}(\xi_k) \leq \frac{1}{\mathcal{Z}(\psi)} \lim_{n\to+\infty} \left\{ \int_{\mathcal{X}} \sup_{k\geq n} \varphi_k(x) \mu(\mathrm{d}x) \right\},$$

where

$$\varphi_k(x) = \log\left\{ \frac{e^{\nu_k} h_{\theta_k}(x)}{e^{\nu^\star} h_{\theta^\star}(x)} \right\} e^{\nu^\star} h_{\theta^\star}(x) + \log\left\{ \frac{\tau h_\psi(x) + e^{\nu^\star} h_{\theta^\star}(x)}{\tau h_\psi(x) + e^{\nu_k} h_{\theta_k}(x)} \right\} \Big(\tau h_\psi(x) + e^{\nu^\star} h_{\theta^\star}(x) \Big).$$

The sequence $\left\{ \sup_{k\geq n} \ \varphi_k \right\}$ is a decreasing sequence converging pointwise. It may be bounded from above thanks to the log-sum inequality, since for any k we have

$$\varphi_k \leq \log\left\{ \frac{e^{\nu_k} h_{\theta_k}}{e^{\nu^\star} h_{\theta^\star}} \right\} e^{\nu^\star} h_{\theta^\star} + \log\left\{ \frac{\tau h_\psi}{\tau h_\psi} \right\} \tau h_\psi + \log\left\{ \frac{e^{\nu^\star} h_{\theta^\star}}{e^{\nu_k} h_{\theta_k}} \right\} e^{\nu^\star} h_{\theta^\star} = 0.$$

Monotone convergence theorem applies:

$$\lim_{n\to+\infty} \sup_{k\geq n} \mathcal{M}_\tau^{\mathrm{NCE}}(\xi_k) \leq \frac{1}{\mathcal{Z}(\psi)} \int_{\mathcal{X}} \lim_{n\to+\infty} \varphi_n(x) \mu(\mathrm{d}x) = \mathcal{M}_\tau^{\mathrm{NCE}}(\xi).$$

Proof of Lemma 10

Without loss of generality, we may suppose that $\mathbb{E}_\theta[Z] = 1$. Recall the following expressions:

$$\nabla g_\xi = \begin{pmatrix} \nabla g_\theta \\ 1 \end{pmatrix} \qquad \nabla\nabla^T g_\xi = \begin{pmatrix} \nabla\nabla^T g_\theta & \nabla g_\theta \\ \nabla^T g_\theta & 1 \end{pmatrix}.$$

We thus have

$$\mathbb{E}_\theta\Big[\nabla g_\xi Z\Big] \mathbb{E}_\theta\Big[\nabla^T g_\xi Z\Big] = \begin{pmatrix} \mathbb{E}_\theta[\nabla g_\theta Z] \mathbb{E}_\theta[\nabla^T g_\theta Z] & \mathbb{E}_\theta[\nabla g_\theta Z] \\ \mathbb{E}_\theta[\nabla^T g_\theta Z] & 1 \end{pmatrix},$$

$$\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big] = \begin{pmatrix} \mathbb{E}_\theta[\nabla\nabla^T g_\theta Z] & \mathbb{E}_\theta[\nabla g_\theta Z] \\ \mathbb{E}_\theta[\nabla^T g_\theta Z] & 1 \end{pmatrix}.$$

We use the following decomposition

$$\mathbb{E}_\theta\Big[\nabla g_\xi Z\Big] \mathbb{E}_\theta\Big[\nabla^T g_\xi Z\Big] = \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big] - \begin{pmatrix} A_Z & 0 \\ 0 & 0 \end{pmatrix}$$

where Schur's complement $A_Z = \mathbb{E}_\theta[\nabla\nabla^T g_\theta Z] - \mathbb{E}_\theta[\nabla g_\theta Z]\mathbb{E}_\theta[\nabla^T g_\theta Z]$ is definite positive.

So we can re-write the matrix $\mathbf{M}$ as:

$$\mathbf{M} = \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big]^{-1} - \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big]^{-1} \begin{pmatrix} A_Z & 0 \\ 0 & 0 \end{pmatrix} \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big]^{-1}.$$

Now, on the one hand, an inverse block matrix calculation yields

$$\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big]^{-1} = \begin{pmatrix} A_Z^{-1} & -A_Z^{-1}\mathbb{E}_\theta[\nabla g_\theta Z] \\ -\mathbb{E}_\theta[\nabla^T g_\theta Z]A_Z^{-1} & 1 + \mathbb{E}_\theta[\nabla^T g_\theta Z]A_Z^{-1}\mathbb{E}_\theta[\nabla g_\theta Z] \end{pmatrix},$$

while, on the other hand, a direct computation yields

$$\mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big]^{-1} \begin{pmatrix} A_Z & 0 \\ 0 & 0 \end{pmatrix} \mathbb{E}_\theta\Big[\nabla\nabla^T g_\xi Z\Big]^{-1} = \begin{pmatrix} A_Z^{-1} & -A_Z^{-1}\mathbb{E}_\theta[\nabla g_\theta Z] \\ -\mathbb{E}_\theta[\nabla^T g_\theta Z]A_Z^{-1} & \mathbb{E}_\theta[\nabla^T g_\theta Z]A_Z^{-1}\mathbb{E}_\theta[\nabla g_\theta Z] \end{pmatrix}.$$

The matrix $\mathbf{M}$ being the difference between these two quantities, we get the claim of the lemma.

2.8.3 Proofs of MC-MLE consistency and asymptotic normality

MC-MLE consistency

The following proof is a straightforward adaptation of Wald's proof of consistency for the MLE (Wald (1949)). The sketch of proof is mainly inspired from Geyer (2012), which has the merit of giving a very accessible presentation of this technical proof.

To begin, define the opposite of the Kullback-Leibler divergence:

$$\lambda(\theta) = \mathbb{E}_{\theta^\star}\left[\log \frac{f_\theta(Y)}{f_{\theta^\star}(Y)}\right] \leq 0.$$

Since the model is identifiable, λ has a unique maximizer achieved at $\theta^\star$. It may be $-\infty$ for some values of θ, but this does not pose problems in the following proof.

For convenience, we choose to analyse the MC-MLE objective function through the following translational motion (sharing the same maximiser with $\ell^{\mathrm{IS}}_{n,m}$):

$$M^{\mathrm{IS}}_n(\theta,\nu) = \frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{e^\nu h_\theta(Y_i)}{e^{\nu^\star} h_{\theta^\star}(Y_i)}\right\} + 1 - e^\nu \frac{1}{r_n}\sum_{j=1}^{r_n} \frac{h_\theta(X_j)}{h_\psi(X_j)}.$$

For any $\xi \in \Xi = \Theta \times \mathbb{R}$, the law of large numbers yields $M^{\mathrm{IS}}_n(\xi) \xrightarrow[n\to+\infty]{a.s.} \mathcal{M}^{\mathrm{IS}}(\xi)$ where

$$\mathcal{M}^{\mathrm{IS}}(\theta,\nu) = \lambda(\theta) + \nu + \log\frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)} + 1 - e^\nu \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)} \leq 0.$$

Note that by construction $\mathcal{M}^{\mathrm{IS}}$ also has a unique maximiser at $\xi^\star = (\theta^\star, \nu^\star)$.

Let $\eta > 0$. Define $K_\eta = \{\xi \in K : d(\xi, \xi^\star) \geq \eta\}$ where K is the compact set defined in (C2). Under (H3), continuity of the maps $\theta \mapsto h_\theta(x)$ and monotone convergence ensure that for any $\xi \in K_\eta$,

$$\lim_{\varepsilon \downarrow 0} \mathbb{E}_{\theta^\star}\left[\sup_{(\phi,\mu)\in B(\xi,\varepsilon)} \log \frac{e^\mu h_\phi(Y)}{e^{\nu^\star} h_{\theta^\star}(Y)}\right] = \mathbb{E}_{\theta^\star}\left[\log \frac{e^\nu h_\theta(Y)}{e^{\nu^\star} h_{\theta^\star}(Y)}\right],$$

while dominated convergence ensures that

$$\lim_{\varepsilon \downarrow 0} \mathbb{E}_{\psi}\left[\inf_{(\phi,\mu)\in B(\xi,\varepsilon)} e^\mu \frac{h_\phi(X)}{h_\psi(X)}\right] = e^\nu \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)}.$$

Thus for any $\xi \in K_\eta$ and $\gamma > 0$, we can always find $\varepsilon_\xi > 0$ such that simultaneously:

$$\mathbb{E}_{\theta^\star}\left[\sup_{(\phi,\mu)\in B(\xi,\varepsilon_\xi)} \log \frac{e^\mu h_\phi(Y)}{e^{\nu^\star} h_{\theta^\star}(Y)}\right] \leq \mathbb{E}_{\theta^\star}\left[\log \frac{e^\nu h_\theta(Y)}{e^{\nu^\star} h_{\theta^\star}(Y)}\right] + \frac{\gamma}{2},$$

and

$$\mathbb{E}_{\psi}\left[\inf_{(\phi,\mu)\in B(\xi,\varepsilon_\xi)} e^\mu \frac{h_\phi(X)}{h_\psi(X)}\right] \geq e^\nu \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)} - \frac{\gamma}{2}.$$

The set of open balls $\{B(\xi, \varepsilon_\xi) \ : \ \xi \in K\}$ form an open cover of K_η from which we can extract a finite subcover by compactness, i.e. we can build a finite set $\{\xi_1, ..., \xi_p\} \subset K_\eta$ such that $K_\eta \subset \bigcup_{k=1}^p B(\xi_k, \varepsilon_{\xi_k})$. This yields the following inequality:

$$\sup_{\xi\in K_\eta} M_n^{\mathrm{IS}}(\xi) \leq \max_{k=1,\ldots,p}\left\{\frac{1}{n}\sum_{i=1}^n \sup_{(\phi,\mu)\in B(\xi_k,\varepsilon_{\xi_k})}\left(\log \frac{e^\mu h_\phi(Y_i)}{e^{\nu^\star} h_{\theta^\star}(Y_i)}\right) + 1 - \frac{1}{r_n}\sum_{j=1}^{r_n} \inf_{(\phi,\mu)\in B(\xi_k,\varepsilon_{\xi_k})}\left(e^\mu \frac{h_\phi(X_j)}{h_\psi(X_j)}\right)\right\}.$$

The right hand side converges almost surely as the law of large numbers applies simultaneously on a finite set. We can thus bound the upper limit:

$$\limsup_{n\to+\infty}\sup_{\xi\in K_\eta} M_n^{\mathrm{IS}}(\xi) \leq \max_{k=1,\ldots,p}\left\{\mathbb{E}_{\theta^\star}\left[\sup_{(\phi,\mu)\in B(\xi_k,\varepsilon_{\xi_k})}\left(\log \frac{e^\mu h_\phi(Y)}{e^{\nu^\star} h_{\theta^\star}(Y)}\right)\right] + 1 - \mathbb{E}_{\psi}\left[\inf_{(\phi,\mu)\in B(\xi_k,\varepsilon_{\xi_k})}\left(e^\mu \frac{h_\phi(X)}{h_\psi(X)}\right)\right]\right\},$$

$$\limsup_{n\to+\infty}\sup_{\xi\in K_\eta} M_n^{\mathrm{IS}}(\xi) \leq \max_{k=1,\ldots,p} \mathcal{M}^{\mathrm{IS}}(\xi_k) + \gamma \leq \sup_{\xi\in K_\eta} \mathcal{M}^{\mathrm{IS}}(\xi) + \gamma.$$

Moreover γ is arbitrary small, thus the inequality still holds when γ is zero:

$$\limsup_{n\to+\infty}\sup_{\xi\in K_\eta} M_n^{\mathrm{IS}}(\xi) \leq \sup_{\xi\in K_\eta} \mathcal{M}^{\mathrm{IS}}(\xi) \quad \text{a.s.} \tag{2.23}$$

To conclude, let us prove that the right hand side is negative. Indeed, subadditivity of the supremum yields

$$\sup_{\xi\in K_\eta}\mathcal{M}^{\mathrm{IS}}(\theta,\nu)\leq\sup_{\xi\in K_\eta}\lambda(\theta)+\sup_{\xi\in K_\eta}\left(\nu+\log\frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)}+1-e^{\nu}\frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\psi)}\right)$$

where the second term is non positive by construction. Under (H3), it is easy to check that λ is upper semi continuous, which implies in particular that λ achieves its maximum on any compact set. Consequently: $\sup_{\xi\in K_\eta}\mathcal{M}^{\mathrm{IS}}(\xi)\leq\sup_{\xi\in K_\eta}\lambda(\theta)<0$.

The last part of the proof is the same as for NCE consistency (see the appendix).

MC-MLE asymptotic normality

For convenience, for any $\xi=(\theta,\nu)$, we note $g_\xi(x)=\nu+\log h_\theta(x)$.

Let $G_n^{\mathrm{IS}}(\xi)=\nabla_\xi\ell_{n,m}^{\mathrm{IS}}(\xi)$ and $\mathbf{H}_n^{\mathrm{IS}}(\xi)=\nabla_\xi^2\ell_{n,m}^{\mathrm{IS}}(\xi)$. We have

$$G_n^{\mathrm{IS}}(\xi)=\frac{1}{n}\sum_{i=1}^n\nabla_\xi g_\xi(Y_i)-\frac{1}{m_n}\sum_{j=1}^{m_n}\nabla_\xi g_\xi(X_j)\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)},$$

$$\mathbf{H}_n^{\mathrm{IS}}(\xi)=\frac{1}{n}\sum_{i=1}^n\nabla_\xi^2 g_\xi(Y_i)-\frac{1}{m_n}\sum_{j=1}^{m_n}\left\{(\nabla_\xi^2+\nabla_\xi\nabla_\xi^T)g_\xi(X_j)\right\}\frac{\exp\{g_\xi(X_j)\}}{h_\psi(X_j)}. \tag{2.24}$$

We start by proving that, almost surely,

$$\sup_{\xi\in B(\xi^\star,\varepsilon)}\|\mathbf{H}_n^{\mathrm{IS}}(\xi)-\mathbf{H}(\xi)\|\underset{n\to\infty}{\to}0, \tag{2.25}$$

where

$$\mathbf{H}(\xi)=\mathbb{E}_{\theta^\star}\left[\nabla_\xi^2 g_\xi(Y)\right]-\mathbb{E}_\psi\left[\left\{(\nabla_\xi^2+\nabla_\xi\nabla_\xi^T)g_\xi(X)\right\}\frac{\exp\{g_\xi(X)\}}{h_\psi(X)}\right].$$

To prove (2.25), split the supremum norm in two and apply Lemma 4 to both empirical averages in definition (2.24). Both supremum norms are integrable under (H4), this is proven in the following.

$$\nabla_\xi^2 g_\xi(x)=\begin{pmatrix}\nabla_\theta^2\log h_\theta(x) & 0\\ 0 & 0\end{pmatrix}\qquad\nabla_\xi\nabla_\xi^T g_\xi(x)=\begin{pmatrix}\nabla_\theta\nabla_\theta^T\log h_\theta(x) & \nabla_\theta\log h_\theta(x)\\ \nabla_\theta^T\log h_\theta(x) & 1\end{pmatrix}$$

First supremum norm is integrable under (H4), since

$$\int_{\mathcal{X}}\sup_{\xi\in B(\xi^\star,\varepsilon)}\|\nabla_\xi^2 g_\xi(x)\|h_{\theta^\star}\mu(\mathrm{d}x)\leq\int_{\mathcal{X}}\sup_{\theta\in B(\theta^\star,\varepsilon)}\|\nabla_\theta^2\log h_\theta(Y)\|\sup_{\theta\in B(\theta^\star,\varepsilon)}h_\theta(x)\mu(\mathrm{d}x)<+\infty.$$

For the second one, use the following decomposition:

$$\|(\nabla_\xi^2+\nabla_\xi\nabla_\xi^T)g_\xi(x)\|_1=\|(\nabla_\theta^2+\nabla_\theta\nabla_\theta^T)\log h_\theta(x)\|_1+2\|\nabla_\theta\log h_\theta(x)\|_1+1,$$

$$\|(\nabla_\theta^2+\nabla_\theta\nabla_\theta^T)\log h_\theta(x)\|_1\leq\|\nabla_\theta^2\log h_\theta(x)\|_1+\|\nabla_\theta\log h_\theta(x)\|_1^2,$$

$$\|\nabla_\theta \log h_\theta(x)\|_1 \leq 1 + \|\nabla_\theta \log h_\theta(x)\|_1^2.$$

This yields a finite upper bound under (H4), since

$$\int_{\mathcal{X}} \sup_{\xi \in B(\xi^\star,\varepsilon)} \|(\nabla_\xi^2 + \nabla_\xi \nabla_\xi^T) g_\xi(x)\|_1 \exp\{g_\xi(x)\} \mu(\mathrm{d}x) \leq$$
$$e^{\nu^\star + \varepsilon} \int_{\mathcal{X}} \sup_{\theta \in B(\theta^\star,\varepsilon)} \Big(\|\nabla_\theta^2 \log h_\theta(x)\|_1 + 3\|\nabla_\theta \log h_\theta(x)\|_1^2 + 3 \Big) \sup_{\theta \in B(\theta^\star,\varepsilon)} h_\theta(x) \mu(\mathrm{d}x) < +\infty.$$

Note also that, at the point $\xi = \xi^\star$, functions $\mathbf{H}$ and $-\mathbf{J}$ coincide, where

$$\mathbf{J}(\xi) = \mathbb{E}_\theta \left[\nabla_\xi \nabla_\xi^T g_\xi(Y) \right] = \begin{pmatrix} \mathbb{E}_\theta \Big[\nabla_\theta \nabla_\theta^T \log h_\theta(Y) \Big] & \mathbb{E}_\theta \Big[\nabla_\theta \log h_\theta(Y) \Big] \\ \mathbb{E}_\theta \Big[\nabla_\theta^T \log h_\theta(Y) \Big] & 1 \end{pmatrix}.$$

In particular, the matrix $\mathbf{J}(\xi^\star)$ is definite positive, since Schur's complement is also the Fisher Information, definite positive by assumption:

$$\mathbb{E}_\theta \Big[\nabla_\theta \nabla_\theta^T g_\theta(Y) \Big] - \mathbb{E}_\theta \Big[\nabla_\theta g_\theta(Y) \Big] \mathbb{E}_\theta \Big[\nabla_\theta^T g_\theta(Y) \Big] = \mathbb{V}_\theta \Big(\nabla_\theta \log h_\theta(Y) \Big) = \mathbf{I}(\theta).$$

Now we establish the weak convergence of the gradient. We have

$$\sqrt{n} G_n^{\mathrm{IS}}(\xi^\star) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \nabla_\xi g_\xi(Y_i) - \mathbb{E}_\theta \Big[\nabla_\xi g_\xi(Y) \Big] \right)_{|\xi = \xi^\star}$$
$$- \sqrt{\frac{n}{m_n}} \sqrt{m_n} \left(\frac{1}{m_n} \sum_{j=1}^{m_n} \nabla_\xi g_\xi(X_j) \frac{f_\theta(X_j)}{f_\psi(X_j)} - \mathbb{E}_\theta \Big[\nabla_\xi g_\xi(Y) \Big] \right)_{|\xi = \xi^\star}.$$

Simulations and observations are assumed independent, thus Slutsky's lemma yields the following. Second moment conditions hold under (I3).

$$\sqrt{n} G_n^{\mathrm{IS}}(\xi^\star) \xrightarrow{\mathcal{D}} \mathcal{N}_{d+1} \Big(0, \mathbf{\Sigma}(\xi^\star) + \tau^{-1} \mathbf{\Gamma}(\xi^\star) \Big),$$

where

$$\mathbf{\Sigma}(\xi) = \mathbb{V}_\theta \Big(\nabla_\xi g_\xi(Y) \Big) = \begin{pmatrix} \mathbf{I}(\theta) & 0 \\ 0 & 0 \end{pmatrix},$$

and

$$\mathbf{\Gamma}(\xi) = \mathbb{V}_\psi \left(\varphi_\xi^{\mathrm{IS}}(X) \right) + 2 \sum_{i=1}^{+\infty} \mathrm{Cov} \Big(\varphi_\xi^{\mathrm{IS}}(X_0), \varphi_\xi^{\mathrm{IS}}(X_i) \Big), \qquad \varphi_\xi^{\mathrm{IS}} = (\nabla_\xi g_\xi) \frac{f_\theta}{f_\psi}.$$

Finally, Lemma 5 applies:

$$\sqrt{n} \left(\widehat{\xi}_{n,m}^{\mathrm{IS}} - \xi^\star \right) \xrightarrow{\mathcal{D}} \mathcal{N} \Big(0_{\mathbb{R}^{d+1}}, \mathbf{V}_\tau^{\mathrm{IS}}(\xi^\star) \Big)$$

where $\mathbf{V}_\tau^{\mathrm{IS}}(\xi) = \mathbf{J}(\xi)^{-1} \left\{ \mathbf{\Sigma}(\xi) + \tau^{-1} \mathbf{\Gamma}(\xi) \right\} \mathbf{J}(\xi)^{-1}$.

2.8.4 Proofs related to exponential families

The following calculations are entirely classical. For the sake of completeness, we present the few tricks required for proving Propositions 1, 2 and 3.

To begin, define $b(x) = \operatorname{sgn}(S(x))$, the vector composed by the signs of each component of $S(x)$. Note that for any $\theta \in \Theta$, the following supremum is necessarily achieved on the boundary of the 1-ball, in the direction of the sign vector:

$$\sup_{\|\phi-\theta\|_1 \leq \varepsilon} \exp\left\{\phi^T S(x)\right\} = \exp\left\{(\theta + \varepsilon b(x))^T S(x)\right\}. \tag{2.26}$$

Since $\|S(x)\|_1 = b(x)^T S(x)$, we have (for the 1-norm for instance):

$$\sup_{\phi \in B(\theta,\varepsilon)} \left(\log \frac{h_\phi(x)}{h_{\theta^\star}(x)}\right) = (\theta - \theta^\star)^T S(x) + \varepsilon \|S(x)\| \leq (\|\theta - \theta^\star\| + \varepsilon)\|S(x)\|,$$

which proves the claim of Proposition 2, since

$$\int_{\mathcal{X}} \sup_{\phi \in B(\theta,\varepsilon)} \left(\log \frac{h_\phi(x)}{h_{\theta^\star}(x)}\right)_+ h_{\theta^\star}(x)\mu(\mathrm{d}x) \leq (\|\theta - \theta^\star\| + \varepsilon) \int_{\mathcal{X}} \|S(x)\| h_{\theta^\star}(x)\mu(\mathrm{d}x) < +\infty.$$

For Propositions 1 and 3, use also the fact that $\|S(x)\|_1 = b(x)^T S(x)$ and that $y \leq e^y$ for any $y \in \mathbb{R}$. We have

$$\|S(x)\|_1^2 \leq \varepsilon^{-2} \exp\left\{2\varepsilon b(x)^T S(x)\right\}. \tag{2.27}$$

Equations (2.26) and (2.27) can be combined as follows:

$$\begin{aligned}\int_{\mathcal{X}} (1 + \|S(x)\|^2) \sup_{\phi \in B(\theta,\varepsilon)} h_\phi(x)\mu(\mathrm{d}x) \leq & \sum_{b \in \{-1,1\}^d} \int_{\mathcal{X}} \exp\left\{(\theta + b\varepsilon)^T S(x)\right\} \mu(\mathrm{d}x) \\ & + \varepsilon^{-2} \sum_{b \in \{-1,1\}^d} \int_{\mathcal{X}} \exp\left\{(\theta + 3b\varepsilon)^T S(x)\right\} \mu(\mathrm{d}x),\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_\psi\left[(1 + \|S(X)\|^2) \sup_{\phi \in B(\theta,\varepsilon)} \left(\frac{h_\phi(X)}{h_\psi(X)}\right)^2\right] \leq & \sum_{b \in \{-1,1\}^d} \mathbb{E}_\psi\left[\left(\frac{\exp\left\{(\theta + b\varepsilon)^T S(X)\right\}}{h_\psi(X)}\right)^2\right] \\ & + \varepsilon^{-2} \sum_{b \in \{-1,1\}^d} \mathbb{E}_\psi\left[\left(\frac{\exp\left\{(\theta + 2b\varepsilon)^T S(X)\right\}}{h_\psi(X)}\right)^2\right].\end{aligned}$$

Choosing $\theta = \widehat{\theta}_n$ in the preceding inequalities yields Proposition 1, while choosing $\theta = \theta^\star$ yields Proposition 3.

Part II

Quantitative results for high dimensional sampling

Chapter 3

On sampling from a log-concave density using kinetic Langevin diffusions

Langevin diffusion processes and their discretizations are often used for sampling from a target density. The most convenient framework for assessing the quality of such a sampling scheme corresponds to smooth and strongly log-concave densities defined on $\mathbb{R}^p$. The present work focuses on this framework and studies the behavior of the Monte Carlo algorithm based on discretizations of the kinetic Langevin diffusion. We first prove the geometric mixing property of the kinetic Langevin diffusion with a mixing rate that is optimal in terms of its dependence on the condition number. We then use this result for obtaining improved guarantees of sampling using the kinetic Langevin Monte Carlo method, when the quality of sampling is measured by the Wasserstein distance. We also consider the situation where the Hessian of the log-density of the target distribution is Lipschitz-continuous. In this case, we introduce a new discretization of the kinetic Langevin diffusion and prove that this leads to a substantial improvement of the upper bound on the sampling error measured in Wasserstein distance.

3.1 Introduction

Markov processes and, more particularly, diffusion processes are often used in order to solve the problem of sampling from a given density π. This problem can be formulated as follows. Assume that we are able to generate an arbitrary number of independent standard Gaussian random variables $\xi_1, \dots, \xi_K$. For a given precision level $\varepsilon > 0$ and a given metric d on the space of probability measures, the goal is to devise a function F_ε such that the distribution ν_K of the random variable $\vartheta_K = F_\varepsilon(\xi_1, \dots, \xi_K)$ satisfies $d(\mu_K, \pi) \leq \varepsilon$. For solving this task, it is often assumed that we can have access to the evaluations of the probability density function of π as well as its derivatives. Among different functions F_ε having the aforementioned property, the most interesting are those that require the smallest number of computations.

Markov Chain Monte Carlo methods hinge on random variables ϑ_K and associated functions F_ε defined by recursion $\vartheta_k = G_\varepsilon(\vartheta_{k-1}, \xi_k)$, $k = 1, \ldots, K$, where G_ε is some function of two arguments. For a given target distribution π, if one succeeds to design a function G_ε such that the Markov process $\{\vartheta_k; k \in \mathbb{N}\}$ is ergodic with invariant density π then, for large K, the distribution of ϑ_K will be close to π. Therefore, if the evaluation of G_ε involves only simple operations, we get a solution of the task of approximate sampling from π. Of course, it is important to address the problem of the choice of the number of iterations K ensuring that the sampling error is smaller than ε. However, it is even more important to be able to design functions G_ε, often referred to as the update rule, with desired properties presented above.

Discretization of continuous-time Markov processes is a successful generic method for defining update rules. The idea is to start by specifying a continuous-time Markov process, $\{L_t : t \geq 0\}$, which is provably positive recurrent and has the target π as invariant distribution[1]. The second step is to set-up a suitable time-discretization of the continuous-time process. More precisely, since $\{L_t\}$ is a Markov process, for any step-size $h > 0$, there is a mapping G such that $L_{kh} \overset{\mathscr{D}}{=} G(L_{(k-1)h}, \xi_k)$, $k = 1, \ldots, K$, where ξ_k is a standard Gaussian random variable independent of $L_{(k-1)h}$. This mapping G might not be available in a closed form. Therefore, the last step is to approximate G by a tractable mapping G_ε. Langevin diffusions are a class of continuous-time Markov processes for which the invariant density is available in closed-form. For this reason, they are suitable candidates for applying the generic approach of the previous paragraph.

Let m and M be two positive constants such that $m \leq M$. Throughout this work, we will assume that the target distribution π has a density with respect to the Lebesgue measure on $\mathbb{R}^p$, which is of the form $\pi(\boldsymbol{\theta}) = Ce^{-f(\boldsymbol{\theta})}$ for a function f that is m-strongly convex and with an M-Lipschitz gradient. The (highly overdamped) Langevin diffusion having π as invariant distribution is defined as a strong solution to the stochastic differential equation

$$d\boldsymbol{L}_t = -\nabla f(\boldsymbol{L}_t)\, dt + \sqrt{2}\, d\boldsymbol{W}_t, \qquad t \geq 0, \tag{3.1}$$

where $\boldsymbol{W}$ is a p-dimensional standard Brownian motion. The update rule associated to this process, obtained by using the Euler discretization, is given by the equation $G_\varepsilon(\boldsymbol{L}_{(k-1)h}, \boldsymbol{\xi}_k) = -h\nabla f(\boldsymbol{L}_{(k-1)h}) + \sqrt{2h}\, \boldsymbol{\xi}_k$ with $\boldsymbol{\xi}_k \overset{\mathscr{D}}{=} h^{-1/2}(\boldsymbol{W}_{kh} - \boldsymbol{W}_{(k-1)h})$ being a p-dimension standard Gaussian vector. The resulting approximate sampling method is often called Langevin Monte Carlo (LMC) or Unadjusted Langevin Algorithm (ULA). Its update rule follows from (3.1) by replacing the function $t \mapsto \nabla f(\boldsymbol{L}_t)$ by its piecewise constant approximation. Therefore, the behavior of the LMC is governed by the following two characteristics of the continuous-time process: the mixing rate and the smoothness of the sample paths. A quantitative bound on the mixing rate allows us to choose a time horizon T such that the distribution of the random vector $\boldsymbol{L}_T$ is within a distance $\varepsilon/2$ of the target distribution, whereas the smoothness of sample paths helps us to design a step-size h so that the distribution of the discretized process at $K = T/h$ is within a distance $\varepsilon/2$ of the distribution of $\boldsymbol{L}_T$. For the LMC, we know that the Langevin

[1] More generally, one can consider a Markov process having an invariant distribution that is close to π.

diffusion mixes exponentially fast with the precise rate e^{-mt}. In addition, almost all sample paths of $\boldsymbol{L}$ are Hölder continuous of degree α, for every $\alpha < 1/2$. Combining these properties, it has been shown that it suffices $K_\varepsilon = O((p/\varepsilon^2)\log(p/\varepsilon^2))$ iterations for the LMC algorithm to achieve an error smaller than ε (both in total-variation and Wasserstein distances); see (Dalalyan, 2017b) for the first nonasymptotic result of this type and (Durmus and Moulines, 2016; Durmus and Moulines, 2017; Dalalyan and Karagulyan, 2019) for improved versions of it.

Under the same assumptions on the log-target f, one can consider the kinetic Langevin diffusion, also known as the second-order Langevin process, defined by

$$d\begin{bmatrix}\boldsymbol{V}_t\\ \boldsymbol{L}_t\end{bmatrix} = \begin{bmatrix}-(\gamma \boldsymbol{V}_t + u\nabla f(\boldsymbol{L}_t))\\ \boldsymbol{V}_t\end{bmatrix} dt + \sqrt{2\gamma u}\begin{bmatrix}\mathbf{I}_p\\ \mathbf{0}_{p\times p}\end{bmatrix} d\boldsymbol{W}_t, \qquad t \geq 0, \tag{3.2}$$

where $\gamma > 0$ is the friction coefficient and $u > 0$ is the inverse mass. As proved in (Nelson, 1967, Theorem 10.1), the highly overdamped Langevin diffusion (3.1) is obtained as a limit of the rescaled kinetic diffusion $\bar{\boldsymbol{L}}_t = \boldsymbol{L}_{\gamma t}$, where $\boldsymbol{L}$ is defined as in (3.2) with $u = 1$, when the friction coefficient γ tends to infinity.

The continuous-time Markov process $(\boldsymbol{L}_t, \boldsymbol{V}_t)$ is positive recurrent and its invariant distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{2p}$. The corresponding invariant density is given by

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

This means that under the invariant distribution, the components $\boldsymbol{L}$ and $\boldsymbol{V}$ are independent, $\boldsymbol{L}$ is distributed according to the target π, whereas $\boldsymbol{V}/\sqrt{u}$ is a standard Gaussian vector. Therefore, one can use this process for solving the problem of sampling from π. As discussed above, the quality of the resulting sampler will depend on two key properties of the process: rate of mixing and smoothness of sample paths. The rate of mixing of kinetic diffusions has been recently studied by Eberle et al. (2017) under conditions that are more general than strong convexity of f. In strongly convex case, a more tractable result has been obtained by Cheng et al. (2018). It establishes that for $\gamma = 2$ and $u = 1/M$, the mixing rate in the Wasserstein distance is $e^{-(m/2M)t}$; see Theorem 5 in (Cheng et al., 2018). On the other hand, sample paths of the process $\{\boldsymbol{L}\}$ defined in (3.2) are smooth of order $1 + \alpha$, for every $\alpha \in [0, 1/2[$. Combining these two properties, (Cheng et al., 2018) prove that a suitable discretization of (3.2) leads to a sampler that achieves an error smaller than ε in a number of iterations K satisfying $K = O((p/\varepsilon^2)^{1/2}\log(p/\varepsilon))$.

It follows from the discussion of previous paragraphs that the kinetic LMC based on (3.2) converges faster than the standard LMC based on (3.1). Furthermore, this improved rate of convergence is mainly due to the higher smoothness of sample paths of the underlying Markov process. The main purpose of the present work is to pursue the investigation of the kinetic Langevin Monte Carlo (KLMC) initiated in (Cheng et al., 2018) by addressing the following questions:

Q1. What is the rate of mixing of the continuous-time kinetic Langevin diffusion for general values of the parameters u and γ?

Q2. Is it possible to improve the rate of convergence of the KLMC by optimizing it over the choice of u, γ and the step-size ?

Q3. If the function f happens to have a Lipschitz-continuous Hessian, is it possible to devise a discretization that takes advantage of this additional smoothness and leads to improved rates of convergence?

The rest of the paper is devoted to answering these questions. The rate of mixing for the continuous-time process is discussed in Section 3.2. In a nutshell, we show that if $\gamma \geq \sqrt{(M+m)u}$, then the rate of mixing is of order $e^{-(um/\gamma)t}$. Non-asymptotic guarantees for the KLMC algorithm are stated and discussed in Section 3.3. They are in the same spirit as those established in (Cheng et al., 2018), but have an improved dependence on the condition number, the ratio of the Lipschitz constant M and the strong convexity constant m. Our result has also improved constants and is much less sensitive to the choice of the initial distribution. These improvements are achieved thanks to a more careful analysis of the discretization error of the Langevin process. Finally, we present in Section 3.4 a new discretization, termed second-order KLMC, of the kinetic Langevin diffusion that exploits the knowledge of the Hessian of f. Its error, measured in the Wasserstein distance W_2 is shown to be bounded by ε for a number of iterations that scales as $(p/\varepsilon)^{1/2}$. Thus, we get an improvement of order $(1/\varepsilon)^{1/2}$ over the first-order KLMC algorithm.

3.2 Mixing rate of the kinetic Langevin diffusion

Let us denote by $\mathbf{P}_t^{\boldsymbol{L}}$ the transition probability at time t of the kinetic diffusion $\boldsymbol{L}$ defined by (3.2). This means that $\mathbf{P}_t^{\boldsymbol{L}}$ is a Markov kernel given by $\mathbf{P}_t^{\boldsymbol{L}}((\boldsymbol{x},\boldsymbol{v}),B) = \mathbf{P}(\boldsymbol{L}_t \in B | \boldsymbol{V}_0 = \boldsymbol{v}, \boldsymbol{L}_0 = \boldsymbol{x})$, for every $\boldsymbol{v},\boldsymbol{x} \in \mathbb{R}^p$ and any Borel set $B \subset \mathbb{R}^p$. For any probability distribution μ on $\mathbb{R}^p \times \mathbb{R}^p$, we denote $\mu\mathbf{P}_t^{\boldsymbol{L}}$ the (unconditional) distribution of the random variable $\boldsymbol{L}_t$ when the starting distribution of the process $(\boldsymbol{V},\boldsymbol{L})$ is μ (*i.e.*, when $(\boldsymbol{V},\boldsymbol{L}_0) \sim \mu$).

Since the process $(\boldsymbol{V},\boldsymbol{L})$ is ergodic, whatever the initial distribution, for large values of t the distribution of $\boldsymbol{L}_t$ is close to the invariant distribution. We want to quantify how fast does this convergence occur. Furthermore, we are interested in a nonasymptotic result in the Wasserstein-Kantorovich distance W_2, valid for a large set of possible values (γ, u).

A first observation is that, without loss of generality, we can focus our attention to the case $u = 1$. This is made formal in the next lemma.

Lemma 1. *Let $(\boldsymbol{V},\boldsymbol{L})$ be the kinetic Langevin diffusion defined by (3.2). The modified process $(\bar{\boldsymbol{V}}_t, \bar{\boldsymbol{L}}_t) = (u^{-1/2}\boldsymbol{V}_{t/\sqrt{u}}, \boldsymbol{L}_{t/\sqrt{u}})$ is an kinetic Langevin diffusion as well with associated parameters $\bar{\gamma} = \gamma/\sqrt{u}$ and $\bar{u} = 1$.*

The proof of this result is straightforward and therefore is omitted. Note that it shows that the parameter u merely represents a time scale (the speed of running over the path of the process $\boldsymbol{L}$). Therefore, in the rest of this paper, we will consider the parameter u to be equal to 1.

Theorem 1. *Assume that the function f is twice differentiable with a Hessian matrix $\nabla^2 f$ satisfying $m\mathbf{I}_p \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\mathbf{I}_p$ for every $\boldsymbol{x} \in \mathbb{R}^p$. Let μ_1, μ_2 and μ_2' be three probability*

measures on $\mathbb{R}^p$. *Let us define the product measures* $\mu = \mu_1 \otimes \mu_2$ *and* $\mu' = \mu_1 \otimes \mu_2'$. *For every* $\gamma, t > 0$, *there exist numbers* $\alpha \leq \sqrt{2}/\gamma$ *and* $\beta \geq \{m \wedge (\gamma^2 - M)\}/\gamma$ *such that*

$$W_2(\mu \mathbf{P}_t^{\boldsymbol{L}}, \mu' \mathbf{P}_t^{\boldsymbol{L}}) \leq \alpha e^{-\beta t} W_2(\mu, \mu'). \tag{3.3}$$

More precisely, for every $v \in [0, \gamma/2[$, *we have*[2]

$$W_2(\mu \mathbf{P}_t^{\boldsymbol{L}}, \mu' \mathbf{P}_t^{\boldsymbol{L}}) \leq \frac{\sqrt{2((\gamma - v)^2 + v^2)}}{\gamma - 2v} \exp\left\{ \frac{(v^2 - m) \vee (M - (\gamma - v)^2)}{\gamma - 2v}\, t \right\} W_2(\mu, \mu'). \tag{3.4}$$

The proof of this result is postponed to Section 3.7. Here, we will discuss some consequences of it and present the main ingredient of the proof. First of all, note that this result implies that for $\gamma^2 > 2 \vee M$, the operator $\mathbf{P}_t^{\boldsymbol{L}}$ is a contraction. The rate of this contraction is characterized by the parameter β. If we optimize the exponent in (3.4) with respect to v, we get the optimal rates of contraction reported in Table 3.1.

If we consider the case $\gamma = 2\sqrt{Mu} = 2\sqrt{M}$ previously studied in (Cheng et al., 2018), then the best rate of contraction provided by (3.4) corresponds to $v = \sqrt{M} - \sqrt{M - m}$, and the upper bound of Theorem 1 reads as

$$W_2(\mu \mathbf{P}_t^{\boldsymbol{L}}, \mu' \mathbf{P}_t^{\boldsymbol{L}}) \leq \left(\frac{2M - m}{M - m} \right)^{1/2} \exp\Big\{ - \Big(\sqrt{M} - \sqrt{M - m} \Big)\, t \Big\} W_2(\mu, \mu'). \tag{3.5}$$

One can check that the constant $\sqrt{M} - \sqrt{M - m}$ that we obtain within the exponential is optimal, in the sense that one gets exactly this constant in the case where f is the bivariate quadratic function $f(x_1, x_2) = (m/2)x_1^2 + (M/2)x_2^2$. This constant is slightly better than the one obtained in (Cheng et al., 2018, Lemma 8) for the particular choice of the time scale $u = 1/M$. Indeed, if we rewrite the two results in the common time-scale $u = 1$, (Cheng et al., 2018, Lemma 8) provides the contraction rate $\beta = m/(2\sqrt{M})$, which is smaller than (but asymptotically equivalent to) $\sqrt{M} - \sqrt{M - m}$.

Another relevant consequence is obtained by instantiating (3.3) to the case $\gamma \geq \sqrt{M + m}$. This leads to the bound

$$\gamma \geq \sqrt{M + m} \quad \Longrightarrow \quad W_2(\mu \mathbf{P}_t^{\boldsymbol{L}}, \mu' \mathbf{P}_t^{\boldsymbol{L}}) \leq \sqrt{2} \exp\Big\{ - (m/\gamma)\, t \Big\} W_2(\mu, \mu'). \tag{3.6}$$

This result is interesting since it allows to optimize the argument of the exponent with respect to γ for fixed t. The corresponding optimized constant is $m/\sqrt{M + m}$, which improves on the constant obtained in (3.5) for $\gamma = 2\sqrt{M}$. When M/m becomes large, the improvement factor gets close to 2.

We now describe the main steps of the proof of Theorem 1. The main idea is to consider along with the process $(\boldsymbol{V}, \boldsymbol{L})$, another process $(\boldsymbol{V}', \boldsymbol{L}')$ that satisfies the same SDE (3.2) as $(\boldsymbol{V}, \boldsymbol{L})$, with the same Brownian motion but with different initial conditions. One easily checks that

$$d \begin{bmatrix} \boldsymbol{V}_t - \boldsymbol{V}_t' \\ \boldsymbol{L}_t - \boldsymbol{L}_t' \end{bmatrix} = \begin{bmatrix} -(\gamma(\boldsymbol{V}_t - \boldsymbol{V}_t') + \nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_t')) \\ \boldsymbol{V}_t - \boldsymbol{V}_t' \end{bmatrix} dt \qquad t \geq 0. \tag{3.7}$$

[2] One can observe that (3.3) can be deduced from (3.4) by taking $v = 0$.

$\gamma^2 \in$	$]0, M]$	$]M, m+M]$	$[m+M, 3m+M[$	$[3m+M, +\infty[$
rate of contraction, β	NA	$\frac{\gamma^2 - M}{\gamma}$	$\frac{\gamma}{2} - \frac{M-m}{2\sqrt{2(m+M)-\gamma^2}}$	$\frac{\gamma - \sqrt{\gamma^2 - 4m}}{2}$
Obtained by Thm. 1 with	-	$v = 0$	$v = \frac{\gamma - \sqrt{2(m+M)-\gamma^2}}{2}$	$v = \frac{\gamma - \sqrt{\gamma^2-4m}}{2}$

Table 3.1: The rates of contraction of the distribution of the kinetic Langevin diffusion $\boldsymbol{L}_t$ for $u = 1$ and varying γ. The reported values are obtained by optimizing the bound in Theorem 1 with respect to v. In the overdamped case $\gamma^2 \geq 3m + M$, the obtained rates coincide with those that can be directly computed for quadratic functions f and, therefore, are optimal.

Using the mean value theorem, we infer that for a suitable symmetric matrix $\mathbf{H}_t$, we have $\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}'_t) = \mathbf{H}_t(\boldsymbol{L}_t - \boldsymbol{L}'_t)$. Furthermore, $\mathbf{H}_t$ being the Hessian of a strongly convex function satisfies $\mathbf{H}_t \succeq m\mathbf{I}_p$. Then, (3.7) can be rewritten as

$$\frac{d}{dt}\begin{bmatrix} \boldsymbol{V}_t - \boldsymbol{V}'_t \\ \boldsymbol{L}_t - \boldsymbol{L}'_t \end{bmatrix} = \begin{bmatrix} -\gamma \mathbf{I}_p & -\mathbf{H}_t \\ \mathbf{I}_p & \mathbf{0}_{p\times p} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_t - \boldsymbol{V}'_t \\ \boldsymbol{L}_t - \boldsymbol{L}'_t \end{bmatrix} \qquad t \geq 0. \tag{3.8}$$

In a small neighborhood of any fixed time instance t_0, (3.8) is close to a linear differential equation with the associated matrix

$$\mathbf{M}(t_0) = \begin{bmatrix} -\gamma \mathbf{I}_p & -\mathbf{H}_{t_0} \\ \mathbf{I}_p & \mathbf{0}_{p\times p} \end{bmatrix}.$$

It is well-known that the solution of such a differential equation will tend to zero if and only if the real parts of all the eigenvalues of $\mathbf{M}(t_0)$ are negative. The matrix $\mathbf{M}(t_0)$ is not symmetric; it is in most cases diagonalizable but its eigenvectors generally depend on t_0. To circumvent this difficulty, we determine the transformations diagonalizing the surrogate matrix

$$\mathbf{M} = \begin{bmatrix} -\gamma \mathbf{I}_p & -v^2 \mathbf{I}_p \\ \mathbf{I}_p & \mathbf{0}_{p\times p} \end{bmatrix}, \qquad \text{for some } v \in [0, \gamma/2[.$$

This yields an invertible matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{M}\mathbf{P}$ is diagonal. We can thus rewrite (3.8) in the form

$$\frac{d}{dt}\mathbf{P}^{-1}\begin{bmatrix} \boldsymbol{V}_t - \boldsymbol{V}'_t \\ \boldsymbol{L}_t - \boldsymbol{L}'_t \end{bmatrix} = \{\mathbf{P}^{-1}\mathbf{M}(t)\mathbf{P}\}\mathbf{P}^{-1}\begin{bmatrix} \boldsymbol{V}_t - \boldsymbol{V}'_t \\ \boldsymbol{L}_t - \boldsymbol{L}'_t \end{bmatrix} \qquad t \geq 0. \tag{3.9}$$

Interestingly, we prove that the quadratic form associated with the matrix $\mathbf{P}^{-1}\mathbf{M}(t)\mathbf{P}$ is negative definite and this provides the desired result. Furthermore, we use this same matrix $\mathbf{P}$ for analyzing the discretized version of the kinetic Langevin diffusion and proving the main result of the next section.

3.3 Error bound for the KLMC in Wasserstein distance

Let us start this section by recalling the KLMC algorithm, the sampler derived from a suitable time-discretization of the kinetic diffusion, introduced by Cheng et al. (2018). Let us define the sequence of functions ψ_k by $\psi_0(t) = e^{-\gamma t}$ and $\psi_{k+1}(t) = \int_0^t \psi_k(s)\, ds$. Recall that f is assumed twice differentiable and, without loss of generality, the parameter u is assumed to be equal to one. The discretization involves a step-size $h > 0$ and is defined by the following recursion:

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\vartheta}_k) \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\vartheta}_k) \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \boldsymbol{\xi}_{k+1} \\ \boldsymbol{\xi}'_{k+1} \end{bmatrix}, \tag{3.10}$$

where $(\boldsymbol{\xi}_{k+1}, \boldsymbol{\xi}'_{k+1})$ is a $2p$-dimensional centered Gaussian vector satisfying the following conditions:

- $(\boldsymbol{\xi}_j, \boldsymbol{\xi}'_j)$'s are iid and independent of the initial condition $(\boldsymbol{v}_0, \boldsymbol{\vartheta}_0)$,
- for any fixed j, the random vectors $\big((\boldsymbol{\xi}_j)_1, (\boldsymbol{\xi}'_j)_1\big)$, $\big((\boldsymbol{\xi}_j)_2, (\boldsymbol{\xi}'_j)_2\big)$, …, $\big((\boldsymbol{\xi}_j)_p, (\boldsymbol{\xi}'_j)_p\big)$ are iid with the covariance matrix

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

This recursion may appear surprising, but one can check that it is obtained by first replacing in (3.2), on each time interval $t \in [kh, (k+1)h]$, the gradient $\nabla f(\boldsymbol{L}_t)$ by $\nabla f(\boldsymbol{L}_{kh})$, by renaming $(\boldsymbol{V}_{kh}, \boldsymbol{L}_{kh})$ into $(\boldsymbol{v}_k, \boldsymbol{\vartheta}_k)$ and by explicitly solving the obtained linear SDE (which leads to an Ornstein-Uhlenbeck process). To the best of our knowledge, the algorithm (3.10), that we will refer to as KLMC, has been first proposed by Cheng et al. (2018). The next result characterizes its approximation properties.

Theorem 2. *Assume that the function f is twice differentiable with a Hessian matrix $\nabla^2 f$ satisfying $m\mathbf{I}_p \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\mathbf{I}_p$ for every $\boldsymbol{x} \in \mathbb{R}^p$. In addition, let the initial condition of the KLMC algorithm be drawn from the product distribution $\mu = \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) \otimes \nu_0$. For every $\gamma \geq \sqrt{m+M}$ and $h \leq m/(4\gamma M)$, the distribution ν_k of the kth iterate $\boldsymbol{\vartheta}_k$ of the KLMC algorithm (3.10) satisfies*

$$W_2(\nu_k, \pi) \leq \sqrt{2}\Big(1 - \frac{0.75mh}{\gamma}\Big)^k W_2(\nu_0, \pi) + \frac{Mh\sqrt{2p}}{m}.$$

The proof of this theorem, postponed to Section 3.8, is inspired by the proof in (Cheng et al., 2018), but with a better control of the discretization error. This allows us to achieve the following improvements as compared to aforementioned paper:

- The second term in the upper bound provided by Theorem 2 scales linearly as a function of the condition number $\varkappa \triangleq M/m$, whereas the corresponding term in (Cheng et al., 2018) scales as $\varkappa^{3/2}$.

- The impact of the initial distribution ν_0 on the overall error of sampling appears only in the first term, which is multiplied by a sequence that has an exponential decay in k. As a consequence, if we denote by K the number of iterations sufficient for the error to be smaller than a prescribed level ε, our result leads to an expression of K in which $W_2(\nu_0, \pi)$ is within a logarithm. Recall that the expression of K in (Cheng et al., 2018, Theorem 1) scales linearly in $W_2(\nu_0, \pi)$.
- The numerical constants of Theorem 2 are much smaller than those of the corresponding result in (Cheng et al., 2018).

In order to ease the comparison of our result to (Cheng et al., 2018, Theorem 1), let us apply Theorem 2 to

$$h = \frac{m}{4M\sqrt{m+M}} \bigwedge \frac{0.94\varepsilon}{\varkappa\sqrt{2p}} \tag{3.11}$$

and $\gamma = \sqrt{m+M}$, which corresponds to the tightest upper bound furnished by our theorem. Note that in (Cheng et al., 2018) it is implicitly assumed that p/ε^2 is large enough so that the second term in the minimum appearing in (3.11) is smaller than the first term. From (3.11) we obtain that[3]

$$K_{\mathrm{KLMC}} \geq \frac{\sqrt{m+M}}{0.75m}\left(\frac{4M\sqrt{m+M}}{m} \bigvee \frac{\varkappa\sqrt{2p}}{0.94\varepsilon}\right) \log\left(\frac{24W_2(\nu_0,\pi)}{\varepsilon}\right) \tag{3.12}$$

iterations are sufficient for having $W_2(\nu_K, \pi) \leq \varepsilon$. After some simplifications, we get

$$K_{\mathrm{KLMC}} \geq 3\varkappa^{3/2}\left\{(16\varkappa) \bigvee \frac{p}{m\varepsilon^2}\right\}^{1/2} \log\left(\frac{24W_2(\nu_0,\pi)}{\varepsilon}\right) \tag{3.13}$$

Remind that the corresponding result in Cheng et al. (2018) requires K to satisfy[4]

$$K \geq 52\varkappa^2\left\{\frac{p}{m\varepsilon^2}\right\}^{1/2} \log\left(\frac{24W_2(\nu_0,\pi)}{\varepsilon}\right).$$

Thus, the improvement in terms of the number of iterations we obtain is at least by a factor $17\sqrt{\varkappa}$, whenever $\kappa \leq p/(16m\varepsilon^2)$.

It is also helpful to compare the obtained result (3.13) to the analogous result for the highly overdamped Langevin diffusion (Durmus and Moulines, 2016). Using (Durmus et al., 2018, Eq. (22)), one can check that this is enough to choose an integer

$$K_{\mathrm{LMC}} \geq 2\varkappa\left\{1 \bigvee \frac{2.18p}{m\varepsilon^2}\right\} \log\left(\frac{24W_2(\nu_0,\pi)}{\varepsilon}\right), \tag{3.14}$$

such that K_{LMC} iterations of the LMC algorithm are sufficient to arrive at an error bounded by ε. Comparing (3.13) and (3.14), we see that the KLMC is preferable to the

[3]This value of K is obtained by choosing h and K so that the second term in the upper bound of Theorem 2 is equal to $(1-\sqrt{2}/24)\epsilon$ whereas the first term is smaller than $(\sqrt{2}/24)\epsilon$.

[4]This lower bound on K is obtained by replacing $\mathcal{D}^2 \triangleq \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2$ by 0 in (Cheng et al., 2018, Theorem 1).

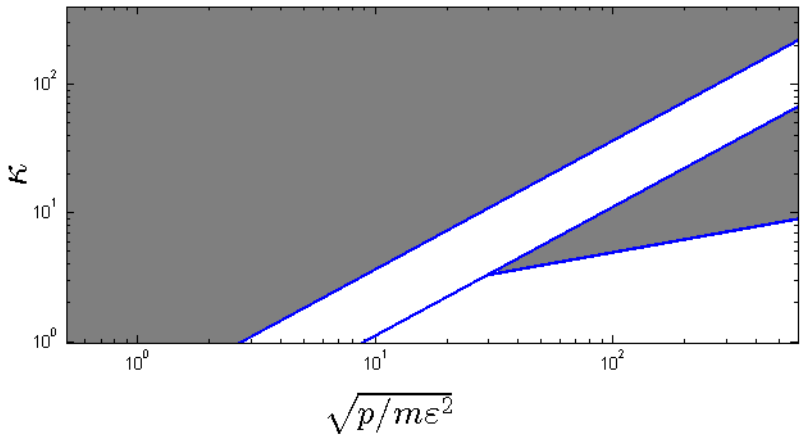

Figure 3.1: This plot represents in the plane defined by coordinates $(\sqrt{p/m\varepsilon^2}, \varkappa)$ the regions where LMC leads to smaller error than the KLMC (in gray). Please note that the axes are in logarithmic scale.

LMC when $p/(m\varepsilon^2)$ is large as compared to the condition number $\varkappa$. This is typically the case when the dimensionality is high or a high precision approximation is required. The order of preference is reversed when the condition number $\varkappa$ is large as compared to $p/(m\varepsilon^2)$. Such a situation corresponds to settings where the target log-density f is nearly flat (m is small) or has a gradient that may increase very fast (M is large). As an important conclusion, we can note that none of these two methods is superior to the other in general. The plot in Figure 3.1 illustrates this fact by showing in gray the regions where LMC outperforms KLMC.

3.4 Second-order KLMC and a bound on its error

In this section, we propose another discretization of the kinetic Langevin process, which is applicable when the function f is twice differentiable. We show below that this new discretization leads to a provably better sampling error under the condition that the Hessian matrix of f is Lipschitz-continuous with respect to the spectral norm. At any iteration $k \in \mathbb{N}$, we define $\mathbf{H}_k = \nabla^2 f(\boldsymbol{\vartheta}_k)$ and

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\vartheta}_k) - \varphi_2(h)\mathbf{H}_k\boldsymbol{v}_k \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\vartheta}_k) - \varphi_3(h)\mathbf{H}_k\boldsymbol{v}_k \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \boldsymbol{\xi}_{k+1}^{(1)} - \mathbf{H}_k\boldsymbol{\xi}_{k+1}^{(3)} \\ \boldsymbol{\xi}_{k+1}^{(2)} - \mathbf{H}_k\boldsymbol{\xi}_{k+1}^{(4)} \end{bmatrix}, \tag{3.15}$$

where

- ψ_0, ψ_1, ψ_2 are defined as in the beginning of the previous section,
- $\varphi_{k+1}(t) = \int_0^t e^{-\gamma(t-s)}\psi_k(s)\,ds$ for every $t > 0$,
- the $4p$ dimensional random vectors $(\boldsymbol{\xi}_{k+1}^{(1)}, \boldsymbol{\xi}_{k+1}^{(2)}, \boldsymbol{\xi}_{k+1}^{(3)}, \boldsymbol{\xi}_{k+1}^{(4)})$ are iid Gaussian with zero mean,

- for any fixed j, the 4-dimensional random vectors $\Big([(\boldsymbol{\xi}_j^{(1)})_1, (\boldsymbol{\xi}_j^{(2)})_1, (\boldsymbol{\xi}_j^{(3)})_1, (\boldsymbol{\xi}_j^{(4)})_1], \ldots, [(\boldsymbol{\xi}_j^{(1)})_p, (\boldsymbol{\xi}_j^{(2)})_p, (\boldsymbol{\xi}_j^{(3)})_p, (\boldsymbol{\xi}_j^{(4)})_p]\Big)$ are iid with the covariance matrix

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

This definition is somewhat complicated, but it follows from an application of the second-order Taylor approximation to the drift term of the kinetic Langevin diffusion[5]. At this stage, one can note that if the Hessian $\mathbf{H}_k$ is zero, then the update rule (3.15) boils down to the update rule of the KLMC algorithm in (3.10). Iterating the update rule (3.15) we get a random variable that will be henceforth called KLMC2 or second-order kinetic Langevin Monte-Carlo algorithm.

Theorem 3. *Assume that, for some constants $m, M, M_2 > 0$, the function f is m-strongly convex, its gradient is M-Lipschitz, and its Hessian is M_2-Lipschitz for the spectral norm. In addition, let the initial condition of the second-order KLMC algorithm be drawn from the product distribution $\mu = \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) \otimes \nu_0$. For every*

$$\gamma \geq \sqrt{m+M} \quad \text{and} \quad h \leq \frac{m}{5\gamma M} \wedge \frac{m}{4\sqrt{5p}\, M_2},$$

the distribution ν_k^{KLMC2} of the kth iterate $\boldsymbol{\vartheta}_k^{\mathrm{KLMC2}}$ of the second-order KLMC algorithm (3.15) satisfies[6]

$$W_2(\nu_k^{\mathrm{KLMC2}}, \pi) \leq \sqrt{2}\Big(1 - \frac{mh}{4\gamma}\Big)^k W_2(\nu_0, \pi) + \frac{2h^2 M_2 p}{m} + \frac{h^2 M\sqrt{2Mp}}{m} + \frac{8M}{m}\, h e^{-p/2}.$$

Several important consequences can be drawn from this result. First, the value of the parameter γ minimizing the right hand side is its smallest possible value $\gamma = \sqrt{m+M}$. Second, one can note that the last term of the obtained upper bound is independent of dimension p and decreases exponentially fast in $1/h$. This term is in most cases negligible with respect to the other terms involved in the upper bound. In particular, we deduce from this result that if the Lipschitz constants M and M_2 are bounded and the strong convexity constant m is bounded away from zero, then the KLMC2 algorithm achieves the precision level ε after K_ε iterations, with K_ε being of order $\sqrt{p/\varepsilon}$, up to a logarithmic factor. Finally, if we neglect the last term in the upper bound of Theorem 3, and choose the parameters h and k so that the other terms are equal to $\varepsilon/\sqrt{4m}$, we get that the number of iteration K_ε to achieve an error $\varepsilon/\sqrt{m}$ scales, up to a logarithmic factor, as $\sqrt{M}/(mh_\varepsilon) = \sqrt{p}\,\varkappa_2^2 + \sqrt{p/\varepsilon}\,\varkappa_2^{5/4}$, where $\varkappa_2 = (M_2^{2/3} + Mp^{-1/3})/m$ is a version of the condition number taking into account the Hessian-Lipschitz assumption.

It is interesting to compare this result to the convergence result for the LMCO algorithm established in (Dalalyan and Karagulyan, 2019). We can note that the number of

[5] For more detailed explanations, see Section 3.9.1

[6] One can see from the proof that $e^{-p/2}$ in this inequality can be replaced by the smaller quantity $e^{-\frac{m^2}{160 M_2^2 h^2}}$.

iterations that are sufficient for the KLMC2 to achieve the error ε is much smaller than the corresponding number for the LMCO: $\sqrt{p/\varepsilon}$ versus p/ε. In addition, the KLMC2 algorithm does not need to compute matrix exponentials neither to do matrix inversion. The most costly operations are that of computing the products of the $p \times p$ Hessian and the vectors $\boldsymbol{v}_k$, $\boldsymbol{\xi}_{k+1}^3$ and $\boldsymbol{\xi}_{k+1}^3$. In most cases, the complexity of these computations scales linearly in p.

As a conclusion, to the best of our knowledge, the second-order KLMC algorithm provides the best known convergence rate $\sqrt{p/\varepsilon}$ for a target density π having a log-density that is concave and Hessian-Lipschitz.

3.5 Related work

The idea of using the Langevin diffusion (see (Pavliotis, 2014) for an introduction to this topic) for approximating a random variable drawn from its invariant distribution is quite old and can be traced back at least to (Roberts and Tweedie, 1996). Since then, many papers focused on analyzing the asymptotic behavior of the Langevin-based methods under various assumptions, see (Lamberton and Pagès, 2003, 2002; Stramer and Tweedie, 1999a,b; Douc et al., 2004; Pillai et al., 2012; Xifara et al., 2014; Roberts and Stramer, 2002; Roberts and Rosenthal, 1998; Bou-Rabee and Hairer, 2013) and the references therein. Convergence to the invariant distribution for Langevin processes is studied in (Desvillettes and Villani, 2001; Helffer and Nier, 2005; Dolbeault et al., 2015).

Non-asymptotic and computable bounds on the convergence to equilibrium of the kinetic Langevin diffusion have been recently obtained in (Eberle et al., 2017; Cheng et al., 2018; Cheng et al., 2018). While (Cheng et al., 2018) considers only the convex case, (Eberle et al., 2017; Cheng et al., 2018) deal also with nonconvexity. On the one hand, (Cheng et al., 2018) provide results only for a fixed value of parameters $(\gamma, u) = (2, 1/M)$. On the other hand, if we instantiate results of (Eberle et al., 2017) to the case of convex functions f, convergence to the invariant density is proved under the condition $\gamma^2 \geq 30Mu$. This is to be compared to the conditions of Theorem 1 that establishes exponential convergence as soon as $\gamma^2 > Mu$.

Nonasymptotic bounds on the precision of the Langevin Monte Carlo under strong convexity have been established in (Dalalyan, 2017b) and then extended and refined in a series of papers (Durmus and Moulines, 2016; Bubeck, 2015; Dalalyan, 2017a; Cheng and Bartlett, 2018; Durmus and Moulines, 2017; Brosse et al., 2017; Durmus et al., 2018; Luu et al., 2017; Bernton, 2018). Very recently, it was proved in (Dwivedi et al., 2018) that applying a Metropolis-Hastings correction to the LMC leads to improved dependence on the target precision ϵ of the number of gradient evaluations. The fact that the discretized version of the kinetic Langevin diffusion may outperform its highly overdamped counterpart was observed and quantified in (Cheng et al., 2018).

Previous work has also studied the precision of Langevin algorithms in the case when the gradient evaluations are contaminated by some noise (Dalalyan, 2017a; Dalalyan and Karagulyan, 2019; Cheng et al., 2018; Baker et al., 2018; Chatterji et al., 2018) and the relation with stochastic optimization (Raginsky et al., 2017; Zhang et al., 2017; Xu et al.,

2017; Dieuleveut et al., 2017). There are certainly many other papers related to the present work that are not mentioned in this section. There is a vast literature on this topic and it will be impossible to quote all the papers. We believe that the papers cited here and the references therein provide a good overview of the state of the art.

3.6 Conclusion

In order to summarize the content of the previous sections, let us return, on by one, to the questions raised in the introduction. First, concerning the mixing properties of the kinetic Langevin diffusion for general values of u and γ, we have established that as soon as $\gamma^2 > Mu$, the process mixes exponentially fast with a rate at least equal to $\{mu \wedge (\gamma^2 - Mu)\}/\gamma$. Therefore, for fixed values of m, M and u, the nearly fastest rate of mixing is obtained for $\gamma^2 = (m+M)u$ and is equal to $m/\sqrt{m+M}$.

To answer the second question, we have seen that optimization with respect to γ and u leads to improved constants but does not improve the rate. Indeed, if we use the values of γ and u used in (Cheng et al., 2018) (that is $\gamma = 2$ and $u = 1/M$, which in view of Lemma 1 are equivalent to $\gamma = 2\sqrt{M}$ and $u = 1$) lead to a bound on the number of iterates sufficient to achieve a precision ε that is of the same order as the optimized one given in (3.12). Interestingly, our analysis revealed that not only the numerical constants of the result in (Cheng et al., 2018) can be improved, but also the dependence on the condition number $\varkappa = M/m$ can be made better. Indeed, we have managed to replace the factor $\varkappa^2$ by $\varkappa^{3/2}$. Such an improvement might have important consequences in generalizing the results to the case of a convex function which is not strongly convex. This line of research will be explored in a future work. Our bound exhibits also a better dependence on the error of the first step: it is logarithmic in our result while it was linear in (Cheng et al., 2018).

Finally, we have given an affirmative answer to the third question. We have shown that leveraging second-order information may reduce the number of steps of the algorithm by a factor proportional to $1/\sqrt{\varepsilon}$, where ε is the target precision. In order to better situate this improvement in the context of prior work, the table below reports the order of magnitude of the number of steps[7] of Langevin related algorithms in the strongly convex case:

1st-order LMC	1st-order KLMC	2nd-order KLMC
(Durmus and Moulines, 2016) (Dalalyan and Karagulyan, 2019)	(Cheng et al., 2018) and Theorem 2	Theorem 3
p/ε	$\sqrt{p}/\varepsilon$	$\sqrt{p/\varepsilon}$

[7] To ease the comparison, we consider $\varkappa$ as a fixed constant and do not report the dependence on $\varkappa$ in this table.

3.7 Proof of the mixing rate

This section is devoted to proofs of the results stated in Section 3.2. Let $\boldsymbol{L}_0, \boldsymbol{L}_0'$ and $\boldsymbol{V}_0$ be three p-dimensional random vectors defined on the same probability space such that

- $\boldsymbol{V}_0$ is independent of $(\boldsymbol{L}_0, \boldsymbol{L}_0')$,
- $\boldsymbol{V}_0 \sim \mu_1$, whereas $\boldsymbol{L}_0 \sim \mu_2$ and $\boldsymbol{L}_0' \sim \mu_2'$,
- $W_2^2(\mu_2, \mu_2') = \mathbf{E}[\|\boldsymbol{L}_0 - \boldsymbol{L}_0'\|_2^2]$.

Let $\boldsymbol{W}$ be a Brownian motion on the same probability space. We define $(\boldsymbol{V}, \boldsymbol{L})$ and $(\boldsymbol{V}', \boldsymbol{L}')$ as kinetic Langevin diffusion processes driven by the same Brownian motion $\boldsymbol{W}$ and satisfying the initial condition $\boldsymbol{V}_0' = \boldsymbol{V}_0$. From the definition of the Wasserstein distance, it follows that

$$W_2^2(\mu \mathbf{P}_t^{\boldsymbol{L}}, \mu' \mathbf{P}_t^{\boldsymbol{L}}) \leq \mathbf{E}[\|\boldsymbol{L}_t - \boldsymbol{L}_t'\|_2^2].$$

In view of this inequality, it suffices to find an appropriate upper bound on the right hand side of the last display, in order to prove Theorem 1. This upper bound is provided below in Proposition 1.

Proposition 1. *Let $\boldsymbol{V}_0, \boldsymbol{L}_0$ and $\boldsymbol{L}_0'$ be random vectors in $\mathbb{R}^p$. Let $(\boldsymbol{V}_t, \boldsymbol{L}_t)$ and $(\boldsymbol{V}_t', \boldsymbol{L}_t')$ be kinetic Langevin diffusions driven by the same Brownian motion and starting from $(\boldsymbol{V}_0, \boldsymbol{L}_0)$ and $(\boldsymbol{V}_0, \boldsymbol{L}_0')$, respectively. Let v be an arbitrary real number from $[0, \gamma/2)$. We have*

$$\|\boldsymbol{L}_t - \boldsymbol{L}_t'\|_2 \leq \frac{\sqrt{2((\gamma - v)^2 + v^2)}}{\gamma - 2v} \exp\left\{ \frac{(v^2 - m) \vee (M - (\gamma - v)^2)}{\gamma - 2v}\, t \right\} \|\boldsymbol{L}_0 - \boldsymbol{L}_0'\|_2, \quad \forall t \geq 0.$$

Remark 1. *As a consequence, we can see that for $\gamma^2 \geq 2(M + m)$ by setting*

$$v = \frac{\gamma - \sqrt{\gamma^2 - 4m}}{2} \geq \frac{m}{\gamma}.$$

we arrive at

$$\|\boldsymbol{L}_t - \boldsymbol{L}_t'\|_2 \leq \left(\frac{2\gamma^2 - 4m}{\gamma^2 - 4m} \right)^{1/2} e^{-vt} \|\boldsymbol{L}_0 - \boldsymbol{L}_0'\|_2, \qquad \forall t \geq 0.$$

Proof. We will use the following short hand notations $\psi_t \triangleq (\boldsymbol{V}_t + \lambda_+ \boldsymbol{L}_t) - (\boldsymbol{V}_t' + \lambda_+ \boldsymbol{L}_t')$ and $z_t \triangleq (-\boldsymbol{V}_t - \lambda_- \boldsymbol{L}_t) + \boldsymbol{V}_t' + \lambda_- \boldsymbol{L}_t'$, where λ_+ and λ_- are two positive numbers such that $\lambda_+ + \lambda_- = \gamma$ and $\lambda_+ > \lambda_-$. First note that using Taylor's theorem with the remainder term in integral form, we get

$$\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_t') = \mathbf{H}_t (\boldsymbol{L}_t - \boldsymbol{L}_t')$$

with $\mathbf{H}_t \triangleq \int_0^1 \nabla^2 f(\boldsymbol{L}_t - x(\boldsymbol{L}_t - \boldsymbol{L}_t'))dx$. In view of this formula and the fact that $(\boldsymbol{V}, \boldsymbol{L})$ and $(\boldsymbol{V}', \boldsymbol{L}')$ satisfy the SDE (3.2), we obtain

$$\begin{aligned}
\frac{d}{dt}\psi_t &= -\gamma(\boldsymbol{V}_t - \boldsymbol{V}_t') - (\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_t')) + \lambda_+(\boldsymbol{V}_t - \boldsymbol{V}_t') \\
&= \frac{(\lambda_+ - \gamma)(\lambda_- \psi_t + \lambda_+ z_t)}{\lambda_- - \lambda_+} - \frac{\mathbf{H}_t(\psi_t + z_t)}{\lambda_+ - \lambda_-} \\
&= \frac{(\lambda_-^2 \mathbf{I} - \mathbf{H}_t)\psi_t + (\lambda_- \lambda_+ \mathbf{I} - \mathbf{H}_t) z_t}{\lambda_+ - \lambda_-}.
\end{aligned}$$

In the above inequalities, we have used that $\lambda_+ - \gamma = -\lambda_-$. Similar computations yield

$$\begin{aligned}\frac{d}{dt} z_t &= \gamma(\boldsymbol{V}_t - \boldsymbol{V}'_t) + (\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}'_t)) - \lambda_-(\boldsymbol{V}_t - \boldsymbol{V}'_t) \\ &= \frac{(\gamma - \lambda_-)(\lambda_- \psi_t + \lambda_+ z_t)}{\lambda_- - \lambda_+} + \frac{\mathbf{H}_t(\psi_t + z_t)}{\lambda_+ - \lambda_-} \\ &= \frac{(\mathbf{H}_t - \lambda_-\lambda_+\mathbf{I})\psi_t + (\mathbf{H}_t - \lambda_+^2\mathbf{I})z_t}{\lambda_+ - \lambda_-}.\end{aligned}$$

From these equations, we deduce that

$$\begin{aligned}\frac{d}{dt}\left\|\begin{bmatrix}\psi_t \\ z_t\end{bmatrix}\right\|_2^2 &= 2\psi_t^\top \frac{d\psi_t}{dt} + 2z_t^\top \frac{dz_t}{dt} \\ &= \frac{2}{\lambda_+ - \lambda_-}\Big\{\psi_t^\top(\lambda_-^2\mathbf{I} - \mathbf{H}_t)\psi_t + z_t^\top(\mathbf{H}_t - \lambda_+^2\mathbf{I})z_t\Big\} \\ &\le \frac{2}{\lambda_+ - \lambda_-}\Big\{(\lambda_-^2 - m)\|\psi_t\|_2^2 + (M - \lambda_+^2)\|z_t\|_2^2\Big\} \\ &\le \frac{2\{(\lambda_-^2 - m) \vee (M - \lambda_+^2)\}}{\lambda_+ - \lambda_-}\left\|\begin{bmatrix}\psi_t \\ z_t\end{bmatrix}\right\|_2^2.\end{aligned}$$

An application of Gronwall's inequality yields

$$\left\|\begin{bmatrix}\psi_t \\ z_t\end{bmatrix}\right\|_2 \le \exp\left\{\frac{(\lambda_-^2 - m) \vee (M - \lambda_+^2)}{\lambda_+ - \lambda_-}\, t\right\}\left\|\begin{bmatrix}\psi_0 \\ z_0\end{bmatrix}\right\|_2, \qquad \forall t \ge 0.$$

Since $\boldsymbol{V}_0 = \boldsymbol{V}'_0$ and $\boldsymbol{L}_t - \boldsymbol{L}'_t = (\psi_t + z_t)/(\lambda_+ - \lambda_-)$, we get

$$\begin{aligned}\|\boldsymbol{L}_t - \boldsymbol{L}'_t\|_2 &\le \frac{\sqrt{2}}{\lambda_+ - \lambda_-}\left\|\begin{bmatrix}\psi_t \\ z_t\end{bmatrix}\right\|_2 \\ &\le \frac{\sqrt{2(\lambda_+^2 + \lambda_-^2)}}{\lambda_+ - \lambda_-}\exp\left\{\frac{(\lambda_-^2 - m) \vee (M - \lambda_+^2)}{\lambda_+ - \lambda_-}\, t\right\}\|\boldsymbol{L}_0 - \boldsymbol{L}'_0\|_2, \qquad \forall t \ge 0,\end{aligned}$$

and the claim of the proposition follows. □

3.8 Proof of the convergence of the first-order KLMC

This section contains the complete proof of Theorem 2. We first write

$$W_2(\nu_k, \pi) = W_2(\nu_k, \mu^* \mathbf{P}_{kh}^{\boldsymbol{L}}), \tag{3.16}$$

where $\mu^* = \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) \otimes \pi$ and $\mu^* \mathbf{P}_{kh}^{\boldsymbol{L}}$ is the distribution[8] of the kinetic Langevin process $\boldsymbol{L}$ at time instant kh when the initial condition of this process is drawn from μ^*. In order to upper bound the term in the right hand side of the last display, we introduce

[8] In other words, $\mu^* \mathbf{P}_{kh}^{\boldsymbol{L}}$ is the first marginal of the distribution $\mu^* \mathbf{P}_{kh}^{(\boldsymbol{L},\boldsymbol{V})}$, the last notation being standard in the theory of Markov processes.

the discretized version of the kinetic Langevin diffusion: $(\widetilde{\boldsymbol{V}}_0, \widetilde{\boldsymbol{L}}_0) \sim \mu$ and for every $j = 0, 1, \ldots, k$ and for every $t \in]jh, (j+1)h]$,

$$\begin{aligned}\widetilde{\boldsymbol{V}}_t &= \widetilde{\boldsymbol{V}}_{jh} e^{-\gamma(t-jh)} - \int_{jh}^{t} e^{-\gamma(t-s)} ds \nabla f(\widetilde{\boldsymbol{L}}_{jh}) + \sqrt{2\gamma} \int_{jh}^{t} e^{-\gamma(t-s)}\, d\boldsymbol{W}_{jh+s} \\ \widetilde{\boldsymbol{L}}_t &= \widetilde{\boldsymbol{L}}_{jh} + \int_{jh}^{t} \widetilde{\boldsymbol{V}}_{jh+s}\, ds. \end{aligned} \tag{3.17}$$

We stress that $\boldsymbol{W}$ in the above formula is the same Brownian motion as the one used for defining the process $(\boldsymbol{V}, \boldsymbol{L})$. Furthermore, we choose $\widetilde{\boldsymbol{V}}_0 = \boldsymbol{V}_0$ and $(\boldsymbol{L}_0, \widetilde{\boldsymbol{L}}_0)$ so that

$$W_2^2(\nu_0, \pi) = \mathbf{E}[\|\boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\|_2^2]. \tag{3.18}$$

The process $(\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{L}})$ realizes the synchronous coupling between the sequences $\{(\boldsymbol{v}_j, \boldsymbol{\vartheta}_j); j = 0, \ldots, k\}$ and $\{(\boldsymbol{V}_{jh}, \boldsymbol{L}_{jh}); j = 0, \ldots, k\}$. Indeed, one easily checks by mathematical induction that $(\widetilde{\boldsymbol{V}}_{jh}, \widetilde{\boldsymbol{L}}_{jh})$ has exactly the same distribution as the vector $(\boldsymbol{v}_j, \boldsymbol{\vartheta}_j)$. Therefore, we have

$$W_2(\nu_k, \mu^* \mathbf{P}_{kh}^{\boldsymbol{L}}) \leq \left(\mathbf{E}[\|\widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh}\|_2^2]\right)^{1/2} \triangleq \|\widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh}\|_{\mathbb{L}_2}.$$

Let $\mathbf{P}$ be the matrix used in the proof of the contraction in continuous time for $v = 0$, that is

$$\mathbf{P} = \frac{1}{\gamma} \begin{bmatrix} \mathbf{0}_{p\times p} & -\gamma \mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p \end{bmatrix}, \qquad \mathbf{P}^{-1} = \begin{bmatrix} \mathbf{I}_p & \gamma \mathbf{I}_p \\ -\mathbf{I}_p & \mathbf{0}_{p\times p} \end{bmatrix}.$$

We will now evaluate the sequence

$$A_k \triangleq \left\| \mathbf{P}^{-1} \begin{bmatrix} \widetilde{\boldsymbol{V}}_{kh} - \boldsymbol{V}_{kh} \\ \widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh} \end{bmatrix} \right\|_{\mathbb{L}_2}.$$

The rest of the proof, devoted to upper bounding the last $\mathbb{L}_2$-norm, is done by mathematical induction. On each time interval $[jh, (j+1)h]$, we introduce an auxiliary continuous-time kinetic Langevin process $(\boldsymbol{V}', \boldsymbol{L}')$ such that $(\boldsymbol{V}'_{jh}, \boldsymbol{L}'_{jh}) = (\widetilde{\boldsymbol{V}}_{jh}, \widetilde{\boldsymbol{L}}_{jh})$ and

$$d \begin{bmatrix} \boldsymbol{V}'_t \\ \boldsymbol{L}'_t \end{bmatrix} = \begin{bmatrix} -(\gamma \boldsymbol{V}'_t + \nabla f(\boldsymbol{L}'_t)) \\ \boldsymbol{V}'_t \end{bmatrix} dt + \sqrt{2\gamma u} \begin{bmatrix} \mathbf{I}_p \\ \mathbf{0}_{p\times p} \end{bmatrix} d\boldsymbol{W}_t, \qquad t \in [jh, (j+1)h]. \tag{3.19}$$

By the triangle inequality, we have

$$\begin{aligned} A_{j+1} &\leq \left\| \mathbf{P}^{-1} \begin{bmatrix} \widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h} \\ \widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h} \end{bmatrix} \right\|_{\mathbb{L}_2} + \left\| \mathbf{P}^{-1} \begin{bmatrix} \boldsymbol{V}'_{(j+1)h} - \boldsymbol{V}_{(j+1)h} \\ \boldsymbol{L}'_{(j+1)h} - \boldsymbol{L}_{(j+1)h} \end{bmatrix} \right\|_{\mathbb{L}_2} \\ &\leq \left\| \mathbf{P}^{-1} \begin{bmatrix} \widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h} \\ \widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h} \end{bmatrix} \right\|_{\mathbb{L}_2} + e^{-mh/\gamma} A_j, \end{aligned} \tag{3.20}$$

where in the last inequality we have used the contraction established in continuous time. For the first norm in the right hand side of the last display, we use the fact that the

considered processes $(\boldsymbol{V}', \boldsymbol{L}')$ and $(\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{L}})$ have the same value at the time instant jh. Therefore,

$$\begin{aligned}
\|\widetilde{\boldsymbol{V}}_t - \boldsymbol{V}'_t\|_{\mathbb{L}_2} &= \left\| \int_{jh}^{t} e^{-\gamma(t-s)} \Big(\nabla f(\boldsymbol{L}'_s) - \nabla f(\boldsymbol{L}'_{jh}) \Big)\, ds \right\|_{\mathbb{L}_2} \\
&\le \int_{jh}^{t} \left\| \nabla f(\boldsymbol{L}'_s) - \nabla f(\boldsymbol{L}'_{jh}) \right\|_{\mathbb{L}_2} ds \\
&\le M \int_{jh}^{t} \left\| \boldsymbol{L}'_s - \boldsymbol{L}'_{jh} \right\|_{\mathbb{L}_2} ds \\
&\le M \int_{jh}^{t} \int_{jh}^{s} \left\| \boldsymbol{V}'_u \right\|_{\mathbb{L}_2} du\, ds \\
&= M \int_{jh}^{t} (t-u) \left\| \boldsymbol{V}'_u \right\|_{\mathbb{L}_2} du \\
&\le M \int_{jh}^{t} (t-u)\, du \max_{u \in [jh, (j+1)h]} \left\| \boldsymbol{V}'_u \right\|_{\mathbb{L}_2} \\
&= \frac{M(t-jh)^2}{2} \max_{u \in [jh, (j+1)h]} \left\| \boldsymbol{V}'_u \right\|_{\mathbb{L}_2}
\end{aligned}$$

and

$$\begin{aligned}
\|\widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h}\|_2 &= \left\| \int_{jh}^{(j+1)h} (\widetilde{\boldsymbol{V}}_t - \boldsymbol{V}'_t)\, dt \right\|_2 \\
&\le \int_{jh}^{(j+1)h} \|\widetilde{\boldsymbol{V}}_t - \boldsymbol{V}'_t\|_2\, dt \\
&\le \frac{M}{2} \int_{jh}^{(j+1)h} (t-jh)^2\, dt \max_{u \in [jh, (j+1)h]} \left\| \boldsymbol{V}'_u \right\|_{\mathbb{L}_2} \\
&\le \frac{Mh^3}{6} \max_{u \in [jh, (j+1)h]} \left\| \boldsymbol{V}'_u \right\|_{\mathbb{L}_2}.
\end{aligned}$$

Lemma 2. *For every $u \in [jh, (j+1)h]$, we have*

$$\|\boldsymbol{V}'_u\|_{\mathbb{L}_2} \le \sqrt{p} + A_j.$$

Proof. We have

$$\begin{aligned}
\|\boldsymbol{V}'_u\|_{\mathbb{L}_2} &= \|\boldsymbol{V}_u\|_{\mathbb{L}_2} + \|\boldsymbol{V}'_u - \boldsymbol{V}_u\|_{\mathbb{L}_2} \\
&= \sqrt{p} + \|[\mathbf{I}_p,\ \mathbf{0}_p]\mathbf{P}\mathbf{P}^{-1}[(\boldsymbol{V}'_u - \boldsymbol{V}_u)^\top, (\boldsymbol{L}'_u - \boldsymbol{L}_u)^\top]\|_{\mathbb{L}_2} \\
&\le \sqrt{p} + \|[\mathbf{I}_p,\ \mathbf{0}_p]\mathbf{P}\| \cdot \|\mathbf{P}^{-1}[(\boldsymbol{V}'_u - \boldsymbol{V}_u)^\top, (\boldsymbol{L}'_u - \boldsymbol{L}_u)^\top] \\
&\le \sqrt{p} + \|[\mathbf{I}_p,\ \mathbf{0}_p]\mathbf{P}\| \cdot \|\mathbf{P}^{-1}[(\boldsymbol{V}'_{jh} - \boldsymbol{V}_{jh})^\top, (\boldsymbol{L}'_{jh} - \boldsymbol{L}_{jh})^\top]\|_{\mathbb{L}_2} \\
&= \sqrt{p} + \|[\mathbf{I}_p,\ \mathbf{0}_p]\mathbf{P}\| \cdot A_j.
\end{aligned}$$

Recall that

$$\mathbf{P} = \frac{1}{\gamma} \begin{bmatrix} \mathbf{0}_{p\times p} & -\gamma \mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p \end{bmatrix},$$

which implies that $\|[\mathbf{I}_p,\ \mathbf{0}_p]\mathbf{P}\| = 1$. This completes the proof of the lemma. □

From this lemma and previous inequalities, we infer that

$$\begin{aligned}
&\left\| \mathbf{P}^{-1} \begin{bmatrix} \widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h} \\ \widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h} \end{bmatrix} \right\|_{\mathbb{L}_2} \\
&\quad \le \left\{ \left(\|\widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h}\|_{\mathbb{L}_2} + \gamma \, \|\widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h}\|_{\mathbb{L}_2} \right)^2 + \|\widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h}\|^2_{\mathbb{L}_2} \right\}^{1/2} \\
&\quad \le \left\{ \left(1 + \frac{\gamma h}{3} \right)^2 + 1 \right\}^{1/2} \frac{Mh^2}{2} \Big(\sqrt{p} + A_j \Big).
\end{aligned}$$

Choosing $h \le 1/(4\gamma)$, we arrive at

$$\left\| \mathbf{P}^{-1} \begin{bmatrix} \widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h} \\ \widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h} \end{bmatrix} \right\|_{\mathbb{L}_2} \le 0.75 \, Mh^2 \Big(\sqrt{p} + A_j \Big).$$

Combining this inequality and (3.20), for every $h \le m/(4\gamma M)$, we get

$$\begin{aligned}
A_{j+1} &\le 0.75 \, Mh^2 \Big(\sqrt{p} + A_j \Big) + e^{-hm/\gamma} A_j && (3.21) \\
&= 0.75 Mh^2 \sqrt{p} + (e^{-hm/\gamma} + 0.75 \, Mh^2) A_j. && (3.22)
\end{aligned}$$

Using the inequality $e^{-x} \le 1 - x + \frac{1}{2}x^2$, we can derive from (3.22) that

$$\begin{aligned}
A_{j+1} &\le 0.75 Mh^2 \sqrt{p} + \left(1 - \frac{hm}{\gamma} + \frac{h^2 m^2}{2\gamma^2} + 0.75 \, Mh^2 \right) A_j \\
&\le 0.75 Mh^2 \sqrt{p} + \left(1 - \frac{0.75 mh}{\gamma} \right) A_j.
\end{aligned}$$

Unfolding this recursive inequality, we arrive at

$$A_k \le \frac{Mh\gamma\sqrt{p}}{m} + \left(1 - \frac{0.75 mh}{\gamma} \right)^k A_0.$$

Finally, one easily checks that $A_0 = \gamma W_2(\nu_0, \pi)$ and

$$\|\widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh}\|_{\mathbb{L}_2} \le \|[\mathbf{0}_{p\times p} \; \mathbf{I}_p]\mathbf{P}\| A_k = \gamma^{-1}\sqrt{2} A_k.$$

Putting all these pieces together, we arrive at

$$\begin{aligned}
W_2(\nu_k, \pi) &\le \|\widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh}\|_{\mathbb{L}_2} \\
&\le \gamma^{-1}\sqrt{2} A_k \\
&\le \frac{Mh\sqrt{2p}}{m} + \sqrt{2} \Big(1 - \frac{0.75 mh}{\gamma} \Big)^k (A_0/\gamma) \\
&= \frac{Mh\sqrt{2p}}{m} + \sqrt{2} \Big(1 - \frac{0.75 mh}{\gamma} \Big)^k W_2(\nu_0, \pi),
\end{aligned}$$

and the claim of Theorem 2 follows.

3.9 Proofs for the second-order discretization of the kinetic Langevin diffusion

We start this section by providing some explanations on the definition of the KLMC2 algorithm. We turn then to the proof of Theorem 3.

3.9.1 Explanations on the origin of the KLMC2 algorithm

Recall that the kinetic diffusion is given by the equation

$$d\begin{bmatrix}\boldsymbol{V}_t\\ \boldsymbol{L}_t\end{bmatrix} = \begin{bmatrix}-(\gamma \boldsymbol{V}_t + \nabla f(\boldsymbol{L}_t))\\ \boldsymbol{V}_t\end{bmatrix} dt + \sqrt{2\gamma}\begin{bmatrix}\mathbf{I}_p\\ \mathbf{0}_{p\times p}\end{bmatrix} d\boldsymbol{W}_t. \tag{3.23}$$

From (3.23), by integration by parts, we can deduce that

$$\begin{aligned} e^{\gamma t}\boldsymbol{V}_t &= \boldsymbol{V}_0 + \int_0^t e^{\gamma s}\, d\boldsymbol{V}_s + \gamma \int_0^t e^{\gamma s}\boldsymbol{V}_s\, ds \\ &= \boldsymbol{V}_0 - \int_0^t e^{\gamma s}\nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2\gamma}\int_0^t e^{\gamma s}\, d\boldsymbol{W}_s. \end{aligned}$$

Therefore, we have

$$\boldsymbol{V}_t = e^{-\gamma t}\boldsymbol{V}_0 - \int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2\gamma}\int_0^t e^{-\gamma(t-s)}\, d\boldsymbol{W}_s, \tag{3.24}$$

$$\boldsymbol{L}_t = \boldsymbol{L}_0 + \int_0^t \boldsymbol{V}_s\, ds. \tag{3.25}$$

Lemma 3. *For every $\gamma > 0$ and $t > 0$, we have for any $k, j \in \mathbb{N}$*

$$\varphi_{k+1}(t) = \int_0^t \varphi_k(s)ds, \qquad \varphi_{k+j+1}(t) = \int_0^t \psi_k(s)\psi_j(t-s)ds$$

Proof. Fubini's Theorem and a change of variables yield

$$\begin{aligned} \int_0^t \varphi_k(s)ds &= \int_0^t \int_0^s e^{-\gamma(s-r)}\psi_{k-1}(r)drds \\ &= \int_0^t \int_0^{t-r} e^{-\gamma s}\psi_{k-1}(r)dsdr \\ &= \int_0^t \int_0^{t-s} e^{-\gamma s}\psi_{k-1}(r)drds \\ &= \int_0^t e^{-\gamma s}\psi_k(t-s)ds = \varphi_{k+1}(t). \end{aligned}$$

This is the first claim of the lemma.

The second claim of the lemma is true for $j = 0$ and any $k \in \mathbb{N}$ by definition. By

induction we get

$$\begin{aligned}\int_0^t \psi_k(s)\psi_j(t-s)ds &= \int_0^t \psi_k(s)\int_0^{t-s}\psi_{j-1}(r)drds\\ &= \int_0^t\int_0^{t-r}\psi_k(s)\psi_{j-1}(r)dsdr\\ &= \int_0^t \psi_{k+1}(t-r)\psi_{j-1}(r)dr\\ &= \int_0^t \psi_{k+j}(r)\psi_0(t-r)dr = \varphi_{k+j+1}(t).\end{aligned}$$

This completes the proof of the lemma. □

If the function f is twice continuously differentiable, then, for small values of s, the value $\nabla f(\boldsymbol{L}_s)$ appearing in (3.24) can be approximated by an affine function of $\boldsymbol{L}_s$:

$$\begin{aligned}\nabla f(\boldsymbol{L}_s) &\approx \nabla f(\boldsymbol{L}_0) + \nabla^2 f(\boldsymbol{L}_0)(\boldsymbol{L}_s - \boldsymbol{L}_0)\\ &= \nabla f(\boldsymbol{L}_0) + \nabla^2 f(\boldsymbol{L}_0)\int_0^s \boldsymbol{V}_w\, dw\\ &\approx \nabla f(\boldsymbol{L}_0) + \psi_1(s)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0 + \sqrt{2\gamma}\,\nabla^2 f(\boldsymbol{L}_0)\int_0^s \psi_1(s-w)\, d\boldsymbol{W}_w. \qquad (3.26)\end{aligned}$$

From the above approximation, we can infer that

$$\begin{aligned}\int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{L}_s)\, ds &\approx \psi_1(t)\nabla f(\boldsymbol{L}_0) + \varphi_2(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0\\ &\quad + \sqrt{2\gamma}\,\nabla^2 f(\boldsymbol{L}_0)\int_0^t e^{-\gamma(t-s)}\int_0^s \psi_1(s-w)\, d\boldsymbol{W}_w\, ds\\ &= \psi_1(t)\nabla f(\boldsymbol{L}_0) + \varphi_2(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0 + \sqrt{2\gamma}\,\nabla^2 f(\boldsymbol{L}_0)\int_0^t \varphi_2(t-w)\, d\boldsymbol{W}_w. \qquad (3.27)\end{aligned}$$

In the last step of the above equation, we have used that

$$\begin{aligned}\int_0^t e^{-\gamma(t-s)}\int_0^s \psi_1(s-w)\, d\boldsymbol{W}_w\, ds &= \int_0^t\int_w^t e^{-\gamma(t-s)}\psi_1(s-w)\, ds\, d\boldsymbol{W}_w\\ &= \int_0^t\int_0^{t-w} e^{-\gamma(t-w-u)}\psi_1(u)\, du\, d\boldsymbol{W}_w\\ &= \int_0^t \varphi_2(t-w)\, d\boldsymbol{W}_w.\end{aligned}$$

Combining the last approximation and the diffusion equation (3.24), we arrive at

$$\begin{aligned}\boldsymbol{V}_t &\approx e^{-\gamma t}\boldsymbol{V}_0 - \psi_1(t)\nabla f(\boldsymbol{L}_0) - \varphi_2(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0\\ &\quad - \sqrt{2\gamma}\,\nabla^2 f(\boldsymbol{L}_0)\int_0^t \varphi_2(t-s)\, d\boldsymbol{W}_s + \sqrt{2\gamma}\int_0^t e^{-\gamma(t-s)}\, d\boldsymbol{W}_s.\end{aligned}$$

This approximation will be used for defining the discretized version of the process $\boldsymbol{V}$. In order to define the discretized version of $\boldsymbol{L}$, we will simply use the plug-in approximation

of $\boldsymbol{V}$, and then integrate. This leads to

$$\begin{aligned}\boldsymbol{L}_t &= \boldsymbol{L}_0 + \int_0^t \boldsymbol{V}_s\, ds \\ &\approx \boldsymbol{L}_0 + \psi_1(t)\boldsymbol{V}_0 - \psi_2(t)\nabla f(\boldsymbol{L}_0) - \varphi_3(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0 \\ &\quad - \sqrt{2\gamma}\,\nabla^2 f(\boldsymbol{L}_0)\int_0^t \varphi_3(t-w)\, d\boldsymbol{W}_w + \sqrt{2\gamma}\int_0^t \psi_1(t-w)\, d\boldsymbol{W}_w.\end{aligned}$$

3.9.2 Proof of Theorem 3

Recall that we have defined in Section 3.4 the following functions

$$\varphi_{k+1}(t) = \int_0^t e^{-\gamma(t-s)}\psi_k(s)ds, \qquad k \geq 1.$$

We first evaluate the error of one iteration of the KLMC2 algorithm. To this end, we introduce the processes

$$\begin{aligned}\widetilde{\boldsymbol{V}}_t = e^{-\gamma t}\widetilde{\boldsymbol{V}}_0 - \Big(\psi_1(t)\nabla f(\widetilde{\boldsymbol{L}}_0) + \varphi_2(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\widetilde{\boldsymbol{V}}_0\Big) \\ + \sqrt{2\gamma}\left(\int_0^t e^{-\gamma(t-s)}d\boldsymbol{W}_s - \nabla^2 f(\widetilde{\boldsymbol{L}}_0)\int_0^t \varphi_2(t-s)d\boldsymbol{W}_s\right)\end{aligned}$$

and

$$\begin{aligned}\widetilde{\boldsymbol{L}}_t = \widetilde{\boldsymbol{L}}_0 + \psi_1(t)\widetilde{\boldsymbol{V}}_0 - \Big(\psi_2(t)\nabla f(\widetilde{\boldsymbol{L}}_0) + \varphi_3(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\widetilde{\boldsymbol{V}}_0\Big) \\ + \sqrt{2\gamma}\left(\int_0^t \psi_1(t-s)d\boldsymbol{W}_s - \nabla^2 f(\widetilde{\boldsymbol{L}}_0)\int_0^t \varphi_3(t-s)d\boldsymbol{W}_s\right).\end{aligned}$$

In what follows, we will use the following matrices to perform a linear transformation of the space $\mathbb{R}^{2p}$:

$$\mathbf{P} = \gamma^{-1}\cdot\begin{bmatrix}\mathbf{0}_{p\times p} & -\gamma\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p\end{bmatrix}, \qquad \mathbf{P}^{-1} = \begin{bmatrix}\mathbf{I}_p & \gamma\mathbf{I}_p \\ -\mathbf{I}_p & \mathbf{0}_{p\times p}\end{bmatrix}. \tag{3.28}$$

We need an auxiliary process, denoted by $(\widehat{\boldsymbol{V}}, \widehat{\boldsymbol{L}})$, which at time 0 coincides with $(\boldsymbol{V}, \boldsymbol{L})$ but evolves according to exactly the same dynamics as $(\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{L}})$.

Proposition 2. *Assume that, for some constants $m, M, M_2 > 0$, the function f is m-strongly convex, its gradient is M-Lipschitz, and its Hessian is M_2-Lipschitz for the spectral norm. If the parameter γ and the step size t of the kinetic Langevin diffusion are such that*

$$t \leq \frac{1}{5\gamma},$$

then

$$\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_t - \widehat{\boldsymbol{V}}_t \\ \boldsymbol{L}_t - \widehat{\boldsymbol{L}}_t\end{bmatrix}\right\|_{\mathbb{L}_2} \leq 0.25 \times t^3(M_2\sqrt{p^2+2p} + M^{3/2}\sqrt{p}).$$

Proof. From the definition of $\mathbf{P}^{-1}$, we compute

$$\begin{aligned}\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_t-\widehat{\boldsymbol{V}}_t\\ \boldsymbol{L}_t-\widehat{\boldsymbol{L}}_t\end{bmatrix}\right\|_{\mathbb{L}_2} &= \left\{\|\boldsymbol{V}_t-\widehat{\boldsymbol{V}}_t+\gamma(\boldsymbol{L}_t-\widehat{\boldsymbol{L}}_t)\|_{\mathbb{L}_2}^2+\|\boldsymbol{V}_t-\widehat{\boldsymbol{V}}_t\|_{\mathbb{L}_2}^2\right\}^{1/2}\\ &\le \left\{\left(\|\boldsymbol{V}_t-\widehat{\boldsymbol{V}}_t\|_{\mathbb{L}_2}+\gamma\|\boldsymbol{L}_t-\widehat{\boldsymbol{L}}_t\|_{\mathbb{L}_2}\right)^2+\|\boldsymbol{V}_t-\widehat{\boldsymbol{V}}_t\|_{\mathbb{L}_2}^2\right\}^{1/2}\end{aligned}$$

where the upper bound follows from Minkowski's inequality. We now give upper bounds for the $\mathbb{L}_2$-norm of processes $\boldsymbol{V}-\widehat{\boldsymbol{V}}$ and $\boldsymbol{L}-\widehat{\boldsymbol{L}}$.

Lemma 4. *For any time step $t>0$ we have*

$$\|\widehat{\boldsymbol{V}}_t-\boldsymbol{V}_t\|_{\mathbb{L}_2}\le\frac{t^3(M_2\sqrt{p^2+2p}+M^{3/2}\sqrt{p})}{6},$$

$$\|\widehat{\boldsymbol{L}}_t-\boldsymbol{L}_t\|_{\mathbb{L}_2}\le\frac{t^4(M_2\sqrt{p^2+2p}+M^{3/2}\sqrt{p})}{24}.$$

Proof. Recall that $\psi_1(t)=\int_0^t e^{-\gamma(t-s)}ds$, $\psi_2(t)=\int_0^t se^{-\gamma(t-s)}ds$ and

$$\boldsymbol{V}_t=e^{-\gamma t}\boldsymbol{V}_0-\int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{L}_s)ds+\sqrt{2\gamma}\int_0^t e^{-\gamma(t-s)}d\boldsymbol{W}_s.$$

We compute

$$\begin{aligned}\widehat{\boldsymbol{V}}_t-\boldsymbol{V}_t=\int_0^t e^{-\gamma(t-s)}(\nabla f(\boldsymbol{L}_s)-\nabla f(\boldsymbol{L}_0))ds-\varphi_2(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0\\ -\sqrt{2\gamma}\nabla^2 f(\boldsymbol{L}_0)\int_0^t\varphi_2(t-s)d\boldsymbol{W}_s.\end{aligned}$$

By Taylor's theorem, we have

$$\nabla f(\boldsymbol{L}_s)-\nabla f(\boldsymbol{L}_0)=\mathbf{H}_s\cdot(\boldsymbol{L}_s-\boldsymbol{L}_0),\qquad \mathbf{H}_s\triangleq\int_0^1\nabla^2 f(\boldsymbol{L}_s+h(\boldsymbol{L}_0-\boldsymbol{L}_s))dh.$$

This yields the following convenient re-writing of the first integral

$$\begin{aligned}&\int_0^t e^{-\gamma(t-s)}(\nabla f(\boldsymbol{L}_s)-\nabla f(\boldsymbol{L}_0))ds\\ &\quad=\underbrace{\int_0^t e^{-\gamma(t-s)}(\mathbf{H}_s-\nabla^2 f(\boldsymbol{L}_0))(\boldsymbol{L}_s-\boldsymbol{L}_0)ds}_{\triangleq\boldsymbol{A}_t}+\underbrace{\nabla^2 f(\boldsymbol{L}_0)\int_0^t\int_0^s e^{-\gamma(t-s)}\boldsymbol{V}_r drds}_{\triangleq\boldsymbol{C}_t}.\end{aligned}$$

Now, we replace $\boldsymbol{V}_r$ by its explicit expression

$$\boldsymbol{V}_r=e^{-\gamma r}\boldsymbol{V}_0-\int_0^r e^{-\gamma(r-w)}\nabla f(\boldsymbol{L}_w)\,dw+\sqrt{2\gamma}\int_0^r e^{-\gamma(r-w)}\,d\boldsymbol{W}_w.$$

By integrating twice, we compute

$$\begin{aligned}\boldsymbol{C}_t=\varphi_2(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0+\sqrt{2\gamma}\nabla^2 f(\boldsymbol{L}_0)\int_0^t\varphi_2(t-s)d\boldsymbol{W}_s\\ \underbrace{-\nabla^2 f(\boldsymbol{L}_0)\int_0^t\int_0^s\int_0^r e^{-\gamma(t-s)}e^{-\gamma(r-w)}\nabla f(\boldsymbol{L}_w)dwdrds}_{\triangleq\boldsymbol{B}_t}\end{aligned}$$

Summing the two expressions allows some terms to cancel out leading to

$$\widehat{\boldsymbol{V}}_t - \boldsymbol{V}_t = \boldsymbol{A}_t - \boldsymbol{B}_t,$$

where

$$\begin{aligned}
\boldsymbol{A}_t &= \int_0^t \int_0^1 e^{-\gamma(t-s)} \left(\nabla^2 f(\boldsymbol{L}_s + h(\boldsymbol{L}_0 - \boldsymbol{L}_s)) - \nabla^2 f(\boldsymbol{L}_0)\right) \cdot (\boldsymbol{L}_s - \boldsymbol{L}_0) dh ds, \\
\boldsymbol{B}_t &= \nabla^2 f(\boldsymbol{L}_0) \int_0^t \int_0^s \int_0^r e^{-\gamma(t-s)} e^{-\gamma(r-w)} \nabla f(\boldsymbol{L}_w) dw dr ds.
\end{aligned}$$

We now control $\mathbb{L}_2$-norm of processes $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$. Bounding $e^{-\gamma(t-s)}$ by one, Minkowski's inequality in its integral version and the Lipschitz assumption on the Hessian yield

$$\begin{aligned}
\|\boldsymbol{A}_t\|_{\mathbb{L}_2} &\leq \int_0^t \int_0^1 \mathbf{E}\left[\| \left(\nabla^2 f(\boldsymbol{L}_s + h(\boldsymbol{L}_0 - \boldsymbol{L}_s)) - \nabla^2 f(\boldsymbol{L}_0)\right) \cdot (\boldsymbol{L}_s - \boldsymbol{L}_0)\|_2^2\right]^{1/2} dh ds \\
&\leq M_2 \int_0^t \int_0^1 \mathbf{E}\left[(1-h)^2 \|\boldsymbol{L}_s - \boldsymbol{L}_0\|_2^4\right]^{1/2} dh ds \\
&= \frac{M_2}{2} \int_0^t \left\{ \mathbf{E}\left[\left\| \int_0^s \boldsymbol{V}_r dr \right\|_2^4\right]^{1/4} \right\}^2 ds \\
&\leq \frac{M_2}{2} \int_0^t \left\{ \int_0^s \mathbf{E}\left[\|\boldsymbol{V}_r\|_2^4\right]^{1/4} dr \right\}^2 ds \\
&= \frac{M_2}{2} \int_0^t \left\{ \int_0^s \mathbf{E}\left[\|\boldsymbol{V}_0\|_2^4\right]^{1/4} dr \right\}^2 ds \\
&= \frac{M_2 t^3}{6} \mathbf{E}\left[\|\boldsymbol{V}_0\|_2^4\right]^{1/2},
\end{aligned}$$

where we have used the stationarity of the process $\boldsymbol{V}_r$. Since $\boldsymbol{V}_0$ is standard Gaussian, we get $\mathbf{E}\left[\|\boldsymbol{V}_0\|_2^4\right] = p^2 + 2p$.

In the same way, Minkowski's inequality in its integral version yields

$$\begin{aligned}
\|\boldsymbol{B}_t\|_{\mathbb{L}_2} &\leq \int_0^t \int_0^s \int_0^r \|\nabla^2 f(\boldsymbol{L}_0) \nabla f(\boldsymbol{L}_w)\|_{\mathbb{L}_2} dw dr ds \\
&\leq \int_0^t \int_0^s \int_0^r M \|\nabla f(\boldsymbol{L}_w)\|_{\mathbb{L}_2} dw dr ds \\
&= M \|\nabla f(\boldsymbol{L}_0)\|_{\mathbb{L}_2} \int_0^t \int_0^s \int_0^r dw dr ds \\
&= \frac{t^3 M}{6} \|\nabla f(\boldsymbol{L}_0)\|_{\mathbb{L}_2},
\end{aligned}$$

where last equalities follow from the stationarity of $\boldsymbol{L}_w$. Since $\boldsymbol{L}_0 \sim \pi$ (Dalalyan, 2017a, Lemma 2) ensures that $\|\nabla f(\boldsymbol{L}_0)\|_{\mathbb{L}_2} \leq \sqrt{Mp}$, and the first claim of the lemma follows.

The bound for process $\boldsymbol{L} - \widehat{\boldsymbol{L}}$ follows from Minkowski's inequality combined with the

bound just proven:

$$\begin{aligned}\|\widehat{\boldsymbol{L}}_t - \boldsymbol{L}_t\|_{\mathbb{L}_2} &\le \int_0^t \|\widehat{\boldsymbol{V}}_s - \boldsymbol{V}_s\|_{\mathbb{L}_2} ds \\ &\le \int_0^t \left(\frac{t^3(M_2\sqrt{p^2+2p} + M^{3/2}\sqrt{p})}{6} \right) ds \\ &= \frac{t^4(M_2\sqrt{p^2+2p} + M^{3/2}\sqrt{p})}{24}.\end{aligned}$$

This completes the proof of the lemma. □

The claim of the proposition follows from the assumption $\gamma t \le 1/5$ and that

$$\sqrt{\left(\frac{1}{6} + \frac{1}{5\times 24}\right)^2 + \left(\frac{1}{6}\right)^2} \le 0.25$$

□

The next, perhaps the most important, step of the proof is to assess the distance between the random vectors $(\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{L}}_t)$ and $(\widetilde{\boldsymbol{V}}_t, \widetilde{\boldsymbol{L}}_t)$.

Proposition 3. *Assume that, for some constants $m, M, M_2 > 0$, the function f is m-strongly convex, its gradient is M-Lipschitz, and its Hessian is M_2-Lipschitz for the spectral norm. If the parameter γ and the step size t of the kinetic Langevin diffusion satisfy the inequalities*

$$\gamma^2 \ge m + M, \qquad t \le \frac{1}{5\gamma\varkappa},$$

then, for the $(2p)\times(2p)$ matrix $\mathbf{P}$ defined in (3.28), and for every $a \ge 5p$, it holds

$$\begin{aligned}\left\| \mathbf{P}^{-1} \begin{bmatrix} \widehat{\boldsymbol{V}}_t - \widetilde{\boldsymbol{V}}_t \\ \widehat{\boldsymbol{L}}_t - \widetilde{\boldsymbol{L}}_t \end{bmatrix} \right\|_{\mathbb{L}_2} &\le \left(1 - \frac{mt}{2\gamma} + \frac{M_2\sqrt{a}\,t^2}{\gamma}\right) \left\| \mathbf{P}^{-1} \begin{bmatrix} \boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_0 \\ \boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0 \end{bmatrix} \right\|_{\mathbb{L}_2} \\ &\quad + \sqrt{2}\,t^2(M-m)e^{-(a-p)/8}.\end{aligned}$$

Proof. **Step 1:** After change of basis, the new discretized process rewrites:

$$\begin{aligned}\mathbf{P}^{-1} \begin{bmatrix} \widehat{\boldsymbol{V}}_t - \widetilde{\boldsymbol{V}}_t \\ \widehat{\boldsymbol{L}}_t - \widetilde{\boldsymbol{L}}_t \end{bmatrix} &= \Big\{ \mathbf{I}_{2p} - \psi_1(t) \underbrace{\mathbf{P}^{-1}\mathbf{R}_0\mathbf{P}}_{\triangleq \mathbf{Q}_0} - \underbrace{\mathbf{P}^{-1}\mathbf{E}_0(t)\mathbf{P}}_{\triangleq \mathbf{N}_0(t)} \Big\} \cdot \mathbf{P}^{-1} \begin{bmatrix} \boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_0 \\ \boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0 \end{bmatrix} \\ &\quad + \mathbf{P}^{-1} \begin{bmatrix} \varphi_2(t)(\nabla^2 f(\boldsymbol{L}_0) - \nabla^2 f(\widetilde{\boldsymbol{L}}_0))\boldsymbol{V}_0 \\ \varphi_3(t)(\nabla^2 f(\boldsymbol{L}_0) - \nabla^2 f(\widetilde{\boldsymbol{L}}_0))\boldsymbol{V}_0 \end{bmatrix},\end{aligned}$$

where

$$\mathbf{R}_0 = \begin{bmatrix} \gamma\mathbf{I}_p & \mathbf{H}_0 \\ -\mathbf{I}_p & \mathbf{0}_{p\times p} \end{bmatrix}, \qquad \mathbf{E}_0(t) \triangleq \begin{bmatrix} \varphi_2(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & \mathbf{0}_{p\times p} \\ \varphi_3(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & -\psi_2(t)\mathbf{H}_0 \end{bmatrix}.$$

By Minkowski's inequality and the definition of $\mathbf{P}^{-1}$, we get

$$\begin{aligned}\left\| \mathbf{P}^{-1} \begin{bmatrix} \widehat{\boldsymbol{V}}_t - \widetilde{\boldsymbol{V}}_t \\ \widehat{\boldsymbol{L}}_t - \widetilde{\boldsymbol{L}}_t \end{bmatrix} \right\|_{\mathbb{L}_2} &\le \left\| \Big\{ \mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0 - \mathbf{N}_0(t) \Big\} \cdot \mathbf{P}^{-1} \begin{bmatrix} \boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_0 \\ \boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0 \end{bmatrix} \right\|_{\mathbb{L}_2} \\ &\quad + \xi_2(t) \left\| \Big(\nabla^2 f(\boldsymbol{L}_0) - \nabla^2 f(\widetilde{\boldsymbol{L}}_0) \Big) \cdot \boldsymbol{V}_0 \right\|_{\mathbb{L}_2}\end{aligned}$$

where

$$\xi_2(t) \triangleq \sqrt{\Big(\varphi_2(t) + \gamma\varphi_3(t)\Big)^2 + \varphi_2(t)^2}.$$

We have

$$\begin{aligned}\varphi_2(t) + \gamma\varphi_3(t) &= \int_0^t e^{-\gamma(t-s)}(\psi_1(s) + \gamma\psi_2(s))\, ds \\ &= \frac{1}{\gamma}\int_0^t e^{-\gamma(t-s)}(1 - e^{-\gamma s} + s\gamma - 1 + e^{-\gamma s})\, ds \leq t^2/2.\end{aligned}$$

Therefore, $\xi_2(t) \leq t^2/\sqrt{2}$.

Step 2: We give an upper bound for the following spectral norm

$$\|\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0 - \mathbf{N}_0(t)\| \leq \|\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0\| + \|\mathbf{N}_0(t)\|.$$

We will start by proving that that

$$\|\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0\| \leq 1 - \psi_1(t)(m/\gamma) + 0.5\psi_1(t)^2 M(\alpha + m^2/(M\gamma^2))$$

where $\alpha \triangleq \max(1 - M/\gamma^2, 3M/\gamma^2 - 1)$.

First, we control the eigenvalues of

$$(\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0)(\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0)^\top = \mathbf{I}_{2p} - 2\psi_1(t)\left(\frac{\mathbf{Q}_0 + \mathbf{Q}_0^\top}{2}\right) + \psi_1(t)^2\mathbf{Q}_0\mathbf{Q}_0^\top.$$

For convenience, we use the notation $\mathbf{\Sigma}_0 \triangleq \gamma^{-1}\mathbf{H}_0$ in the following. Direct computations yield

$$\begin{aligned}\mathbf{Q}_0 &= \begin{bmatrix} \mathbf{\Sigma}_0 & \mathbf{\Sigma}_0 \\ -\mathbf{\Sigma}_0 & \gamma\mathbf{I}_p - \mathbf{\Sigma}_0 \end{bmatrix}, \\ \mathbf{S}_0 \triangleq \frac{\mathbf{Q}_0 + \mathbf{Q}_0^\top}{2} &= \begin{bmatrix} \mathbf{\Sigma}_0 & 0_{p\times p} \\ 0_{p\times p} & \gamma\mathbf{I}_p - \mathbf{\Sigma}_0 \end{bmatrix}, \\ \mathbf{Q}_0\mathbf{Q}_0^\top &= \begin{bmatrix} 2\mathbf{\Sigma}_0^2 & \gamma\mathbf{\Sigma}_0 - 2\mathbf{\Sigma}_0^2 \\ \gamma\mathbf{\Sigma}_0 - 2\mathbf{\Sigma}_0^2 & (\gamma\mathbf{I}_p - \mathbf{\Sigma}_0)^2 + \mathbf{\Sigma}_0^2 \end{bmatrix}.\end{aligned}$$

Let us define the symmetric matrix

$$\mathbf{E}_0 \triangleq \begin{bmatrix} \mathbf{\Sigma}_0^2 & \mathbf{H}_0 - 2\mathbf{\Sigma}_0^2 \\ \mathbf{H}_0 - 2\mathbf{\Sigma}_0^2 & \mathbf{\Sigma}_0^2 \end{bmatrix}.$$

so that the following equality holds: $\mathbf{Q}_0\mathbf{Q}_0^\top = \mathbf{S}_0^2 + \mathbf{E}_0$.

Regrouping the quadratic form yields

$$(\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0)(\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0)^\top = (\mathbf{I}_{2p} - \psi_1(t)\mathbf{S}_0)^2 + \psi_1(t)^2\mathbf{E}_0.$$

Lemma 5. *Assume that $\gamma^2 \geq m + M$, then the following holds:*

$$(m/\gamma)\mathbf{I}_{2p} \preceq \mathbf{S}_0 \preceq (\gamma - m/\gamma)\mathbf{I}_{2p}, \qquad \|\mathbf{E}_0\| \leq M\alpha.$$

Proof. The condition $\gamma^2 \geq m + M$ implies that $(m/\gamma)\mathbf{I}_p \preceq \mathbf{\Sigma}_0 \preceq (\gamma - m/\gamma)\mathbf{I}_p$. The first claim of the lemma follows directly.

Now, let us compute the eigenvalues of the symmetric matrix $\mathbf{E}_0$. We diagonalize $\mathbf{H}_0$ and note $(\lambda_j^{\mathbf{H}_0})_{j=1,\dots,p}$ its eigenvalues. By solving $\det(\mathbf{E}_0 - \lambda \mathbf{I}_{2p}) = 0$ we get p equations, i.e. for every $j = 1, ..., p$ we need to solve:

$$\lambda^2 - 2a_j\lambda + a_j^2 - b_j^2 = 0, \qquad a_j = \left(\gamma^{-1}\lambda_j^{\mathbf{H}_0}\right)^2, \qquad b_j = \lambda_j^{\mathbf{H}_0} - 2a_j.$$

The solutions are $\lambda_j = a_j \pm |b_j|$. For every $j = 1, ..., p$, we get

$$|\lambda_j| \leq \max\left(\lambda_j^{\mathbf{H}_0} - \left(\gamma^{-1}\lambda_j^{\mathbf{H}_0}\right)^2, 3\left(\gamma^{-1}\lambda_j^{\mathbf{H}_0}\right)^2 - \lambda_j^{\mathbf{H}_0}\right).$$

The function $x \mapsto \max(x - (x/\gamma)^2, 3(x/\gamma)^2 - x)$ is increasing on $\mathbb{R}_+$. Since $\lambda_j^{\mathbf{H}_0}$ is upper bounded by M, the second claim of the lemma follows. □

Now, we apply Lemma 5. Since $\mathbf{S}_0$ and $\mathbf{E}_0$ are symmetric, we have

$$\begin{aligned}\|\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0\|^2 &\leq \|(\mathbf{I}_{2p} - \psi_1(t)\mathbf{S}_0)^2\| + \psi_1(t)^2\|\mathbf{E}_0\| \\ &\leq (1 - \psi_1(t)(m/\gamma))^2 + \psi_1(t)^2 M\alpha \\ &= 1 - 2\psi_1(t)(m/\gamma) + \psi_1(t)^2 M(\alpha + m^2/(M\gamma^2)).\end{aligned}$$

Finally, for any $x \leq 1$, $\sqrt{1-x} \leq (1 - x/2)$, we get

$$\|\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0\| \leq 1 - \psi_1(t)(m/\gamma) + 0.5\psi_1(t)^2 M(\alpha + m^2/(M\gamma^2)).$$

Now we turn to the bound of $\|\mathbf{N}_0(t)\|$. Direct calculation yields

$$\begin{aligned}\mathbf{N}_0(t) &= \mathbf{P}^{-1}\mathbf{E}_0(t)\mathbf{P} \\ &= \gamma^{-1}\begin{bmatrix}\mathbf{I}_p & \gamma\mathbf{I}_p \\ -\mathbf{I}_p & \mathbf{0}_{p\times p}\end{bmatrix}\begin{bmatrix}\varphi_2(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & \mathbf{0}_{p\times p} \\ \varphi_3(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & -\psi_2(t)\mathbf{H}_0\end{bmatrix}\begin{bmatrix}\mathbf{0}_{p\times p} & -\gamma\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p\end{bmatrix} \\ &= \gamma^{-1}\begin{bmatrix}(\varphi_2 + \gamma\varphi_3)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & \mathbf{0} \\ -\varphi_2\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{0} & -\gamma\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p\end{bmatrix} - \psi_2(t)\gamma^{-1}\begin{bmatrix}\mathbf{0} & \gamma\mathbf{H}_0 \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{0} & -\gamma\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{0}_{p\times p} & -(\varphi_2(t) + \gamma\varphi_3(t))\nabla^2 f(\widetilde{\boldsymbol{L}}_0) \\ \mathbf{0}_{p\times p} & \varphi_2(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\end{bmatrix} - \psi_2(t)\begin{bmatrix}\mathbf{H}_0 & \mathbf{H}_0 \\ \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p}\end{bmatrix}.\end{aligned}$$

Since $\nabla^2 f(\widetilde{\boldsymbol{L}}_0)$ and $\mathbf{H}_0$ are both upper bounded by $M\mathbf{I}_p$ and $0 \leq \psi_2(t) \leq \varphi_2(t) \leq \varphi_2(t) + \gamma\varphi_3(t) \leq t^2/2$, we get

$$\|\mathbf{N}_0(t)\| \leq \sqrt{2}Mt^2.$$

Summing the two upper bounds, we get

$$\|\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0 - \mathbf{N}_0(t)\| \leq \rho_t \triangleq \left\{1 - \frac{\psi_1(t)m}{\gamma} + \frac{\psi_1(t)^2 M}{2}\left(\alpha + \frac{m^2}{M\gamma^2}\right) + M\sqrt{2}\,t^2\right\}.$$

Taylor's expansion ensures that $t - \gamma t^2/2 \leq \psi_1(t) \leq t$ and, therefore,

$$\rho_t \leq 1 - \frac{mt}{\gamma} + \frac{Mt^2}{2}\underbrace{\left(\alpha + \frac{m^2}{M\gamma^2} + \frac{m}{M} + 2\sqrt{2}\right)}_{\leq 2 + 2\sqrt{2} \leq 5}.$$

Finally, we use the condition $t \le 1/(5\gamma\varkappa)$ to bound ρ_t by $1 - mt/(2\gamma)$.

Step 3: We control the $\mathbb{L}_2$-norm of $(\nabla^2 f(\boldsymbol{L}_0) - \nabla^2 f(\widetilde{\boldsymbol{L}}_0))\boldsymbol{V}_0$.

Since $m\mathbf{I}_p \preccurlyeq \nabla^2 f(x) \preccurlyeq M\mathbf{I}_p$, combined with the fact that the Hessian is M_2-Lipschitz, we get

$$\left\|\left(\nabla^2 f(\boldsymbol{L}_0) - \nabla^2 f(\widetilde{\boldsymbol{L}}_0)\right) \cdot \boldsymbol{V}_0\right\|_2 \le \min\left(M - m, M_2\|\boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\|_2\right) \|\boldsymbol{V}_0\|_2.$$

Using the obvious inequality $\|\boldsymbol{V}_0\|_2^2 \le a + (\|\boldsymbol{V}_0\|_2^2 - a)_+$, for every $a > 0$, this implies that

$$\begin{aligned}
\mathbf{E}\left[\left\|\left(\nabla^2 f(\boldsymbol{L}_0) - \nabla^2 f(\widetilde{\boldsymbol{L}}_0)\right)\, \boldsymbol{V}_0\right\|_2^2\right] &\le \mathbf{E}\left[\min\left((M-m)^2, M_2^2\|\boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\|_2^2\right) \|\boldsymbol{V}_0\|_2^2\right] \\
&\le M_2^2 a\, \mathbf{E}\left[\|\boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\|_2^2\right] + (M-m)^2 \mathbf{E}\left[(\|\boldsymbol{V}_0\|_2^2 - a)_+\right] \\
&\overset{(1)}{\le} M_2^2 a\, \|\boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\|_{\mathbb{L}_2}^2 + 4(M-m)^2 e^{-(a-p)/4},
\end{aligned}$$

where inequality (1) is valid for every $a \ge 5p$ according to well-known bounds on the χ^2 distribution; see for instance (Collier and Dalalyan, 2017, Lemmas 5-6). Finally, recall that

$$\|\boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\|_{\mathbb{L}_2} \le \gamma^{-1}\sqrt{2}\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_0 \\ \boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2}.$$

Taking square roots yields the claim of the proposition. □

The last piece of the proof is the following proposition.

Proposition 4. *Assume that, for some constants $m, M, M_2 > 0$, the function f is m-strongly convex, its gradient is M-Lipschitz, and its Hessian is M_2-Lipschitz for the spectral norm. If the parameter γ and the step size h of the kinetic Langevin diffusion satisfy the inequalities*

$$\gamma^2 \ge m + M, \qquad h \le \frac{1}{5\gamma\varkappa} \wedge \frac{m}{4\sqrt{5p}\, M_2}.$$

Then

$$\begin{aligned}
\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh} - \widetilde{\boldsymbol{V}}_{kh} \\ \boldsymbol{L}_{kh} - \widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2} &\le \left(1 - \frac{mh}{4\gamma}\right)^k \left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_0 \\ \boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2} \\
&\quad + \frac{4\sqrt{2}\,(M-m)}{m}\,\gamma h e^{-\frac{m^2}{160 M_2^2 h^2}} + \gamma h^2\left(\frac{M_2}{m}\sqrt{p^2+2p} + \frac{M^{3/2}}{m}\sqrt{p}\right).
\end{aligned}$$

Proof. Minkowski's inequality yields

$$\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh} - \widetilde{\boldsymbol{V}}_{kh} \\ \boldsymbol{L}_{kh} - \widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2} \le \left\|\mathbf{P}^{-1}\begin{bmatrix}\widehat{\boldsymbol{V}}_{kh} - \widetilde{\boldsymbol{V}}_{kh} \\ \widehat{\boldsymbol{L}}_{kh} - \widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2} + \left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh} - \widehat{\boldsymbol{V}}_{kh} \\ \boldsymbol{L}_{kh} - \widehat{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2}.$$

For $k \ge 0$, define

$$x_k = \left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh} - \widetilde{\boldsymbol{V}}_{kh} \\ \boldsymbol{L}_{kh} - \widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2}.$$

By Proposition 2 and Proposition 3, we thus have

$$x_{k+1} \le \left(1 - \frac{mh}{2\gamma} + \frac{M_2\sqrt{a}\,h^2}{\gamma}\right) x_k + \sqrt{2}\,h^2(M-m)e^{-(a-p)/8} + 0.25h^3\left(M_2\sqrt{p^2+2p} + M^{3/2}\sqrt{p}\right).$$

Assuming that $\sqrt{a} = m/(4M_2h) \ge \sqrt{5p}$ and unfolding the last recursion, we get

$$x_{k+1} \le \left(1 - \frac{mh}{4\gamma}\right)^{k+1} x_0 + \frac{4\sqrt{2}\,(M-m)}{m}\,\gamma h e^{-(a-p)/8} + \gamma h^2\left(\frac{M_2}{m}\sqrt{p^2+2p} + \frac{M^{3/2}}{m}\sqrt{p}\right).$$

Easy algebra shows that

$$\frac{a-p}{8} = \frac{a}{10} + \frac{a-5p}{40} \ge \frac{a}{10} = \frac{m^2}{160M_2^2h^2}.$$

This is exactly the claim of the proposition. □

To complete the proof of Theorem 3, we need to do some simple algebra. First of all, using the relations

$$W_2(\nu_k,\pi) \le \gamma^{-1}\sqrt{2}\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh} - \widetilde{\boldsymbol{V}}_{kh}\\ \boldsymbol{L}_{kh} - \widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2}, \qquad W_2(\nu_0,\pi) = \gamma^{-1}\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_0\\ \boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2}$$

as well as the inequality $p^2 + 2p \le 2p^2$ (since $p \ge 2$), we arrive at

$$\begin{aligned} W_2(\nu_k,\pi) \le \sqrt{2}\left(1 - \frac{mh}{4\gamma}\right)^k W_2(\nu_0,\pi) \\ + \frac{8\,(M-m)}{m}\,h e^{-\frac{m^2}{160M_2^2h^2}} + \sqrt{2}\,h^2\left(\frac{M_2p}{m}\sqrt{2} + \frac{M^{3/2}}{m}\sqrt{p}\right). \end{aligned} \tag{3.29}$$

This leads to the claim of the theorem.

Chapter 4

Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets

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 and the bounded-Lipschitz distance. 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.

4.1 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 (Bubeck, 2015), the case of non-convex smooth optimization being much more delicate (Jain and

Kar, 2017). 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 (Durmus et al., 2018) 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 $\mu_k = \mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\|\boldsymbol{\vartheta}\|_2^k]$, where $\|\boldsymbol{v}\|_q = (\sum_j |v_j|^q)^{1/q}$ is the usual ℓ_q-norm for any $q \geq 1$.

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

$$\boldsymbol{\vartheta}_{k+1,\boldsymbol{h}} = \boldsymbol{\vartheta}_{k,\boldsymbol{h}} - h_{k+1}\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}}) + \sqrt{2h_{k+1}}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{4.1}$$

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 Kth 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 (Durmus et al., 2018, 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

$$D_{\mathrm{KL}}(\bar{\nu}_K\|\pi) \leq \frac{\mu_2}{2Kh} + Mph, \qquad D_{\mathrm{KL}}(\bar{\nu}_K^{\mathrm{opt}}\|\pi) \leq \sqrt{\frac{2Mp\mu_2}{K}}. \tag{4.2}$$

Note that the second inequality above is obtained from the first one by using the step-size $h_{\mathrm{opt}} = (2KMp/\mu_2)^{-1/2}$ 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 $M\mu_2 \leq \kappa p^\beta$, for some dimension-free positive constants β and κ, we get

$$D_{\mathrm{KL}}(\bar{\nu}_K^{\mathrm{opt}}\|\pi) \leq \sqrt{\frac{2\kappa p^{1+\beta}}{K}}.$$

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,

$d_{\mathrm{TV}}(\bar{\nu}_K, \pi) \leq \sqrt{D_{\mathrm{KL}}(\bar{\nu}_K, \pi)/2}$ 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 $K_{\mathrm{LMCa,TV}}(p, \varepsilon)$ sufficient for the total-variation-error of the LMC with averaging (hereafter, LMCa) to be smaller than ε is

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

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 condition number κ. We will focus only on those measures of quality of sampling that can be directly used for evaluating the quality of approximating expectations.

4.2 Further precisions on the analyzed methods

Since our main motivation for considering the sampling problem comes from statistical applications, we will focus on the following distances between the probability distributions: Monge-Kantorovich-Wasserstein distances W_q and the bounded-Lipschitz distance (also called Fortet-Mourier distance) defined by

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

$$d_{\mathrm{bL}}(\nu, \nu') = \sup\Big\{\mathbf{E}_{\boldsymbol{\vartheta}\sim\nu}[\varphi(\boldsymbol{\vartheta})] - \mathbf{E}_{\boldsymbol{\vartheta}\sim\nu'}[\varphi(\boldsymbol{\vartheta})] : \sup_{\boldsymbol{\theta}\in\mathbb{R}^p} |\varphi(\boldsymbol{\theta})| + \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'} \frac{|\varphi(\boldsymbol{\theta}) - \varphi(\boldsymbol{\theta}')|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2} \leq 1\Big\}.$$

It is well known that the bounded-Lipschitz distance is stronger than both the total-variation and W_1 distances in the sense that $d_{\mathrm{bL}}(\nu, \nu') \leq d_{\mathrm{TV}}(\nu, \nu') \wedge W_1(\nu, \nu')$ for every pair of probability measures (ν, ν'). It is also clear, in view of the Hölder inequality, that the distances W_q are increasing functions of q.

Our main contributions are upper bounds on quantities of the form $W_q(\nu_K, \pi)$ and $d_{\mathrm{bL}}(\nu_K, \pi)$ where π is a log-concave target density function and ν_K is 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 (standard) Langevin diffusion. The latter is the highly overdamped version of the former, see (Nelson, 1967). The Langevin diffusion, having π as invariant distribution, is defined as a solution[1] to the stochastic differential equation

$$d\boldsymbol{L}_t = -\nabla f(\boldsymbol{L}_t)\, dt + \sqrt{2}\, d\boldsymbol{W}_t, \qquad t \geq 0, \tag{4.3}$$

where $\boldsymbol{W}$ is a p-dimensional standard Brownian motion independent of the initial value $\boldsymbol{L}_0$. The LMC algorithm presented in (4.1) is merely the Euler-Maruyama discretization

[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, (Pavliotis, 2014)) ensuring that π and p^* are invariant densities of, respectively, processes (4.3) and (4.4) are fulfilled.

of the process $\boldsymbol{L}$. The kinetic Langevin diffusion $\{\boldsymbol{L}_t : t \geq 0\}$, also known as the second-order Langevin process, is defined by

$$d\begin{bmatrix}\boldsymbol{V}_t \\ \boldsymbol{L}_t\end{bmatrix} = \begin{bmatrix}-(\gamma \boldsymbol{V}_t + \nabla f(\boldsymbol{L}_t)) \\ \boldsymbol{V}_t\end{bmatrix} dt + \sqrt{2\gamma}\begin{bmatrix}\mathbf{I}_p \\ \mathbf{0}_{p\times p}\end{bmatrix} d\boldsymbol{W}_t, \qquad t \geq 0, \tag{4.4}$$

where $\gamma > 0$ is the friction coefficient. The process $\boldsymbol{V}_t$ is often called the velocity process since the second row in (4.4) implies that $\boldsymbol{V}_t$ is the time derivative of $\boldsymbol{L}_t$. The continuous-time Markov process $(\boldsymbol{L}_t, \boldsymbol{V}_t)$ 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, in equilibrium, the random variable $\boldsymbol{L}_t$ has the target distribution π.

Time-discretized versions of Langevin diffusion processes (4.3) and (4.4) are used for (approximately) 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 $M > 0$: $\|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \leq M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$.*

Condition 2. *The function f is convex on $\mathbb{R}^p$. Furthermore, for some positive constants κ and β, we have $V_\pi = \mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\|\boldsymbol{\vartheta} - \mathbf{E}[\boldsymbol{\vartheta}]\|_2^2] \leq \kappa p^\beta / M$.*

For m-strongly convex functions f, 2 is satisfied with $\kappa = M/m$, according to Brascamp-Lieb inequality (Brascamp and Lieb, 1976). We will show that this condition is also satisfied for functions f that are convex everywhere and strongly convex inside a ball, as well as 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 $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$.*

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 (4.3) and (4.4), 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 $f_\alpha(\boldsymbol{\theta}) \triangleq f(\boldsymbol{\theta}) + \alpha\|\boldsymbol{\theta}\|_2^2/2$. Therefore, the corresponding surrogate distribution has the density

$$\pi_\alpha(\boldsymbol{\theta}) \triangleq \frac{e^{-f_\alpha(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f_\alpha(\boldsymbol{v})} d\boldsymbol{v}}. \tag{4.5}$$

The parameter α, together with the step-size $\boldsymbol{h}$, is considered as a tuning parameter of the algorithms to be calibrated. Too 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, too small values of α will lead to a small approximation error but slow convergence. The next result quantifies the approximation of π by π_α, for different distances.

Proposition 5. *We have the following bounds, for any $\alpha \geq 0$ and $q \in [1, +\infty[$*

$$d_{\mathrm{TV}}(\pi, \pi_\alpha) \leq \alpha\mu_2/2 + \exp\{\alpha\mu_2/2\} - 1$$
$$W_q^q(\pi, \pi_\alpha) \leq 2^{q-2}\alpha\mu_{q+2} + 2^{q-1}\mu_q(\exp\{\alpha\mu_2/2\} - 1).$$

If, in addition, $\alpha\mu_2 \leq 1/5$, then

$$d_{\mathrm{TV}}(\pi, \pi_\alpha) \leq 1.03\alpha\mu_2, \qquad W_q^q(\pi, \pi_\alpha) \leq 1.03 \cdot 2^{q-1}\alpha\mu_{q+2} \leq C_q\alpha\mu_2^{(q+2)/2},$$

where C_q is a numerical constant depending only on q. For instance, $C_1 \leq 70$ and $C_2 \leq 1197$.

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. Note that the values of the constants C_q are derived from the following version[2] of a consequence of Borell's lemma (Brazitikos et al., 2014, Theorem 2.4.6)

$$\mu_3^{1/3} \leq 4.1\mu_2^{1/2}, \qquad \mu_4^{1/4} \leq 5\mu_2^{1/2}.$$

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.

4.3 Measuring the complexity of a sampling scheme

We have already introduced the notation $K_{\mathsf{Meth},\mathsf{Crit}}(p, \varepsilon)$, the number of iterations that guarantee that method Meth 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 (4.1) and the kinetic Langevin process (4.4), 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 $K_{\mathrm{LMC},\mathsf{Crit}}(p, \varepsilon) \leq K_{\mathrm{KLMC},\mathsf{Crit}}(p, \varepsilon)$ 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

[2]The corresponding result, stated in Lemma 9, is (to the best of our knowledge) the first attempt to provide optimized constants.

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[3]. 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 $K_{\mathrm{LMC,TV}}$ (Dalalyan, 2017b; Durmus and Moulines, 2016) and K_{LMC,W_2} (Durmus and Moulines, 2016; Dalalyan and Tsybakov, 2009). 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[4] which depend on both dimension and the scale. For this reason, we suggest the following definition of the analogue of K in the case of Wasserstein distances:

$$K_{\mathsf{Meth},W_q}(p,\varepsilon) = \min\{k \in \mathbb{N} : W_q(\nu_k^{\mathsf{Meth}}, \pi) \le \varepsilon\sqrt{\mu_2(\pi)},\ \forall\, \pi \in \mathscr{P}\}, \tag{4.6}$$

where Meth 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. Definition (4.6), 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 (Tat Lee et al., 2018) as well. Indeed, in the context of m-strongly log-concave distributions, Tat Lee et al. (2018) propose to find the smallest k such that $W_2(\nu_k^{\mathsf{Meth}}, \pi) \le \varepsilon/\sqrt{m}$. This is close to our proposal, since in the case of m-strongly log-concave distributions, it follows from the Brascamp-Lieb inequality that $\sup_\pi \sqrt{\mu_2} = \sqrt{p/m}$.

4.4 Overview of main contributions

In this work, we analyze three methods, LMC, KLMC (Cheng et al., 2018) and KLMC2 (Dalalyan and Riou-Durand, 2018), applied to the strong-convexified potential $f_\alpha(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + (\alpha/2)\|\boldsymbol{\theta}\|_2^2$ in order to cope with the lack of strong convexity. We briefly recall these algorithms and present a summary of the main contributions of this work.

[3]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.

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

4.4.1 Considered Markov chain Monte-Carlo methods

We first recall the definition of the Langevin Monte Carlo algorithm. For some positive parameter h, referred to as the step-size and for some initial distribution ν_0 on $\mathbb{R}^p$ chosen by the user, LMC algorithm starts from $\boldsymbol{\vartheta}_0 \sim \nu_0$. The iterations of the LMC algorithm are defined by the update rule

$$\boldsymbol{\vartheta}_{k+1} = \boldsymbol{\vartheta}_k - h\nabla f(\boldsymbol{\vartheta}_k) + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{4.7}$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k, \ldots$ is a sequence of mutually independent, and independent of $\boldsymbol{\vartheta}_0$, centered Gaussian vectors with covariance matrices equal to identity.

We now recall the definition of Kinetic Langevin Monte Carlo of first and second order. 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 first specify the noise structure 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 from the initial condition $(\boldsymbol{v}_0, \boldsymbol{\vartheta}_0)$.

To specify the covariance structure of these Gaussian variables, we define at first 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 note $\xi_{k,j}$ for the scalar j-th component of the vector $\boldsymbol{\xi}_k$, 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 is a sampler derived from a suitable time-discretization of the kinetic diffusion, introduced by Cheng et al. (2018). Applied to the strong-convexified potential f_α, for a step-size $h > 0$, it is defined by the following recursion:

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)(\nabla f(\boldsymbol{\vartheta}_k) + \alpha\boldsymbol{\vartheta}_k) \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)(\nabla f(\boldsymbol{\vartheta}_k) + \alpha\boldsymbol{\vartheta}_k) \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \boldsymbol{\xi}_{k+1}^{(1)} \\ \boldsymbol{\xi}_{k+1}^{(2)} \end{bmatrix}. \tag{4.8}$$

The KLMC2 algorithm, introduced by Dalalyan and Riou-Durand (2018), is derived from a second order discretization, applicable when the function f is twice differentiable. Applied to the strong-convexified potential f_α, for a step-size $h > 0$, it is defined as follows. At any iteration $k \in \mathbb{N}$, define the gradient $\boldsymbol{g}_{k,\alpha} = \nabla f(\boldsymbol{\vartheta}_k) + \alpha\boldsymbol{\vartheta}_k$, the Hessian $\mathbf{H}_{k,\alpha} = \nabla^2 f(\boldsymbol{\vartheta}_k) + \alpha\mathbf{I}_p$, and the recursion

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\boldsymbol{g}_{k,\alpha} - \varphi_2(h)\mathbf{H}_{k,\alpha}\boldsymbol{v}_k \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\boldsymbol{g}_{k,\alpha} - \varphi_3(h)\mathbf{H}_{k,\alpha}\boldsymbol{v}_k \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \boldsymbol{\xi}_{k+1}^{(1)} - \mathbf{H}_{k,\alpha}\boldsymbol{\xi}_{k+1}^{(3)} \\ \boldsymbol{\xi}_{k+1}^{(2)} - \mathbf{H}_{k,\alpha}\boldsymbol{\xi}_{k+1}^{(4)} \end{bmatrix}. \tag{4.9}$$

4.4.2 Summary of the obtained complexity bounds

Without going into details here, we mention in the table 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 equal to 1.

	LMCa	LMC		KLMC	KLMC2
Cond.	1-2	1-2	1-3	1-2	1-3
W_2	–	p^2/ε^6	$p^{2.5}/\varepsilon^5$	p^2/ε^5	p^2/ε^4
W_1	–	p^2/ε^4	$p^{2.5}/\varepsilon^3$	p^2/ε^3	p^2/ε^2
d_{bL}	p^2/ε^4 △	p^3/ε^4 □	p^3/ε^3	$p^{2.5}/\varepsilon^3$	p^2/ε^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 (Durmus et al., 2018). The second to last row is obtained from the last row by using the fact that the bounded-Lipschitz distance is upper bounded by the total-variation distance. 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 (Dalalyan, 2017b). All the remaining cells of the table are filled in by the results obtained in the present work. It is worth mentioning here, that using Metropolis-Hastings adjustment of the LMC (termed MALA), Dwivedi et al. (2018) obtained the complexity

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

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

4.4.3 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 (Dalalyan, 2017b), consisting 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. Let dist be a general distance on the set of all probability measures and Meth^f be the instantiation of a sampling algorithm to the potential function f.

We will denote by $\nu_{k,\alpha}^{\mathsf{Meth}}$ the distribution of the random vector obtained after performing k iterations of the algorithm Meth with the surrogate potential $f_\alpha(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + (\alpha/2)\|\boldsymbol{\theta}\|_2^2$. Our goal is, in a first stage, to establish an upper bound on the distance between the sampling distribution $\nu_{k,\alpha}^{\mathsf{Meth}}$ and the target π. The methods we analyze here, being discretizations of continuous-time diffusion processes, depend on the step of discretization h. 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 4.3, as a function of p, ε and the condition number κ.

To achieve this goal, we first use the triangle inequality

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

Then, the second term of the right hand side of the last displayed equation is bounded using Proposition 5. Finally, the distance between the sampling density $\nu_{k,\alpha}^{\mathsf{Meth}}$ 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.

4.5 Precision and computational complexity of the LMC

In this section we present the non-asymptotic convergence rates for non-strongly convex 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.

4.5.1 The case of gradient-Lipschitz potential

First we give explicit conditions on the parameters α, h and K to have convergence error for LMC algorithm smaller than ε, in the case of Gradient-Lipschitz potential function.

Theorem 4. *Suppose that the potential function f is Gradient-Lipschitz. If the following conditions are satisfied:*

$$\alpha \leq \frac{1}{5\mu_2} \bigwedge \frac{M}{20}, \qquad h \leq \frac{1}{M} \bigwedge \frac{\alpha\varepsilon^2}{18.9Mp}, \qquad K \geq \frac{2}{\alpha h} \log(3W_0(\nu_0, \pi_\alpha)/\varepsilon),$$

then for any $q \in [1, 2]$ the implications below are true:

$$\alpha \leq \frac{\varepsilon}{3.1\mu_2} \Rightarrow d_{\mathrm{bL}}(\nu_K, \pi) \leq \varepsilon; \tag{4.10}$$

$$\alpha \leq \frac{\varepsilon^q}{3.1 \cdot 6^{q-1}\mu_{q+2}} \Rightarrow W_q(\nu_K, \pi) \leq \varepsilon. \tag{4.11}$$

In order to give more insight on the rates that we obtain for our algorithm, let us derive explicit order for iteration numbers. 2 and Lemma 9 yield that in the case of W_q error one needs to perform

$$K = \widetilde{O}\left(\frac{\kappa^{q+2}p^{\beta(q+2)+1}}{\varepsilon^{2q+2}}\right)$$

gradient iterations. Similarly, for the bounded Lipschitz distance we obtain

$$K = \widetilde{O}\left(\frac{\kappa^2 p^{2\beta+1}}{\varepsilon^4}\right).$$

Proof. We are going to prove both implications parallelly. Recalling the properties of W_q and d_{bL} we have

$$W_q(\nu_K, \pi) \leq W_2(\nu_K, \pi_\alpha) + W_q(\pi_\alpha, \pi),$$
$$d_{\mathrm{bL}}(\nu_K, \pi) \leq W_2(\nu_K, \pi_\alpha) + d_{TV}(\pi_\alpha, \pi).$$

π_α is α-strongly log-concave with f_α as its potential function. By the definition f_α will also have a Lipschitz continuous gradient. As $h \leq 1/M$, then (Durmus et al., 2018, Theorem 9) implies the following bound on $W_2(\nu_K, \pi_\alpha)$.

$$\begin{aligned} W_2(\nu_K, \pi_\alpha) &\leq (1 - \alpha h)^{K/2} W(\nu_0, \pi_\alpha) + (2h(M + \alpha)p/\alpha)^{1/2} \\ &\leq (1 - \alpha h)^{K/2} W(\nu_0, \pi_\alpha) + (2.1hMp/\alpha)^{1/2}. \end{aligned}$$

The latter is true, due to $\alpha \leq M/20$. In order to bound the remaining terms we will use Proposition 5. Since $\alpha \leq 1/(5\mu_2)$ we obtain

$$W_q(\nu_K, \pi) \leq (1 - \alpha h)^{K/2} W(\nu_0, \pi_\alpha) + (2.1hMp/\alpha)^{1/2} + \left(1.03 \cdot 2^{q-1} \alpha \mu_{q+2}\right)^{1/q}$$
$$d_{\mathrm{bL}}(\nu_K, \pi) \leq (1 - \alpha h)^{K/2} W(\nu_0, \pi_\alpha) + (2.1hMp/\alpha)^{1/2} + 1.03\alpha\mu_2.$$

We will now prove that the obtained bounds are smaller than ε, by showing that each term on the right hand side is smaller than $\varepsilon/3$. Let us start with $(1 - \alpha h)^{K/2} W(\nu_0, \pi_\alpha)$:

$$\begin{aligned} (1 - \alpha h)^{K/2} W(\nu_0, \pi_\alpha) &\leq \exp(-\alpha h K/2) W(\nu_0, \pi_\alpha) \\ &\leq \exp\left(-\log(3W(\nu_0, \pi_\alpha)/\varepsilon)\right) W(\nu_0, \pi_\alpha) \\ &\leq \frac{\varepsilon}{3}. \end{aligned}$$

For the second summand we have

$$2.1hMp/\alpha \leq \frac{\alpha\varepsilon^2}{18.9Mp} \cdot \frac{2.1Mp}{\alpha} = \frac{\varepsilon^2}{9}.$$

Finally, (4.10) implies

$$1.03\alpha\mu_2 \leq \varepsilon/3 \quad \text{and} \quad \left(1.03 \cdot 2^{q-1} \alpha \mu_{q+2}\right)^{1/q} \leq \varepsilon/3,$$

respectively. This concludes the proof. □

4.5.2 The case of Hessian-Lipschitz potential

In what follows we are going to analyze the convergence of LMC algorithm, assuming that the potential function is Hessian-Lipschitz.

Theorem 5. *Suppose that the potential function f is Hessian-Lipschitz. If the following conditions are satisfied:*

$$\alpha \leq \frac{1}{5\mu_2} \bigwedge \frac{M}{20}, \qquad h \leq \frac{\alpha\varepsilon}{3M_2 p} \bigwedge \frac{\alpha\varepsilon}{17M^{3/2}p^{1/2}},$$

$$K \geq \frac{1}{\alpha h} \log(3W(\nu_0, \pi_\alpha)/\varepsilon),$$

then for any $q \in [1, 2]$ *the implications below are true:*

$$\alpha \leq \frac{\varepsilon}{3.1\mu_2} \Rightarrow d_{\mathrm{bL}}(\nu_K, \pi) \leq \varepsilon;$$
$$\alpha \leq \frac{\varepsilon^q}{3.1 \cdot 6^{q-1}\mu_{q+2}} \Rightarrow W_q(\nu_K, \pi) \leq \varepsilon.$$

In order to give more insight on the rates that we obtain for our algorithm, let us derive again explicit order for iteration numbers. 2 and Lemma 9 yield that in the case of W_q error one needs to perform

$$K = \widetilde{O}\left(\frac{\kappa^{q+2} p^{\beta(q+2)+1}}{\varepsilon^{2q+1}}\right)$$

gradient iterations. Similarly, for the bounded Lipschitz distance we obtain

$$K = \widetilde{O}\left(\frac{\kappa^2 p^{2\beta+1}}{\varepsilon^3}\right).$$

Proof. As in the previous case we are going to prove the implications parallelly. Again we exploit the inequalitites below:

$$W_q(\nu_K, \pi) \leq W_2(\nu_K, \pi_\alpha) + W_q(\pi_\alpha, \pi),$$
$$d_{\mathrm{bL}}(\nu_K, \pi) \leq W_2(\nu_K, \pi_\alpha) + d_{\mathrm{TV}}(\pi_\alpha, \pi).$$

π_α is α-strongly log-concave with f_α as its potential function. By definition, f_α is Hessian-Lipschitz and strongly-convex. As $h \leq 1/M$, then (Dalalyan and Karagulyan, 2019, Theorem 5) implies the following bound on $W_2(\nu_K, \pi_\alpha)$.

$$\begin{aligned} W_2(\nu_K, \pi_\alpha) &\leq (1 - \alpha h)^K W(\nu_0, \pi_\alpha) + \frac{M_2 hp}{2\alpha} + \frac{13(M+\alpha)^{3/2} hp^{1/2}}{5\alpha} \\ &\leq (1 - \alpha h)^K W(\nu_0, \pi_\alpha) + \frac{M_2 hp}{2\alpha} + \frac{2.8 M^{3/2} hp^{1/2}}{\alpha}. \end{aligned} \tag{4.12}$$

The latter is true, due to $\alpha \leq M/20$. In order to bound the remaining terms we will use Proposition 5. We have shown already in the proof of Theorem 4, that under these conditions on α, the errors $W_q(\nu_K, \pi)$ and $d_{\mathrm{bL}}(\nu_K, \pi)$ are both smaller than $\varepsilon/3$. Similar to the previous case, $(1 - \alpha h)^{K/2} W(\nu_0, \pi_\alpha)$ is shown to be smaller than ε. Now we will prove that the second and the third summands of (4.12) are smaller than $\varepsilon/6$.

$$\frac{M_2 hp}{2\alpha} \leq \frac{M_2 p}{2\alpha} \cdot \frac{\alpha\varepsilon}{3M_2 p} \leq \frac{\varepsilon}{6}.$$

Similarly,

$$\frac{2.8 M^{3/2} hp^{\frac{1}{2}}}{5\alpha} \leq \frac{2.8 M^{3/2} p^{1/2}}{\alpha} \cdot \frac{\alpha\varepsilon}{17 M^{3/2} p^{1/2}} \leq \frac{\varepsilon}{6}.$$

□

Summing up, we observe that second-order smoothness results faster convergence in both distances in terms of ε.

4.6 Precision and computational complexity of the KLMC and KLMC2

Theorem 6. *Suppose that the potential function f satisfies Conditions 1, 2. Set the parameters $\alpha, \gamma, h > 0$ such that*

$$\alpha \leq \frac{1}{5\mu_2} \bigwedge \frac{M}{20}, \qquad \gamma \geq \sqrt{1.1M}, \qquad h \leq \frac{\alpha}{4.2M} \times \left(\frac{1}{\gamma} \bigwedge \frac{\varepsilon}{\sqrt{2p}} \right)$$

and let ν_k be the distribution of the k-th iterate $\boldsymbol{\vartheta}_k$ of the KLMC algorithm, tuned by those parameters.

If the number of iterations K is such that

$$K \geq \frac{4\gamma}{3\alpha h} \log \left(\frac{6W_2(\nu_0, \pi_\alpha)}{\varepsilon} \right),$$

then following implications hold

$$\begin{aligned} \alpha \leq \varepsilon^2/(8.3\mu_4) &\quad \Rightarrow \quad W_2(\nu_K, \pi) \leq \varepsilon, \\ \alpha \leq \varepsilon/(2.1\mu_3) &\quad \Rightarrow \quad W_1(\nu_K, \pi) \leq \varepsilon, \\ \alpha \leq \varepsilon/(2.1\mu_2) &\quad \Rightarrow \quad d_{\mathrm{bL}}(\nu_K, \pi) \leq \varepsilon. \end{aligned}$$

Using Condition 2 and Lemma 9 to control the moments μ_2, μ_3 and μ_4, we compute the scaling of the mixing time of KLMC with respect to $p, \varepsilon, \kappa, \beta$. Up to logarithmic factors, the mixing time scales as $\kappa^4 p^{4\beta+1/2}/\varepsilon^5$ for the Wasserstein-2 distance, $\kappa^3 p^{3\beta+1/2}/\varepsilon^3$ for the Wasserstein-1 distance, and $\kappa^2 p^{2\beta+1/2}/\varepsilon^3$ for the bounded Lipschitz distance.

Proof. We make use of the following relationships between distances for any two probability measures μ and ν:

$$d_{\mathrm{bL}}(\mu, \nu) \leq W_1(\mu, \nu) \leq W_2(\mu, \nu), \qquad d_{\mathrm{bL}}(\mu, \nu) \leq d_{\mathrm{TV}}(\mu, \nu).$$

Combined with the triangular inequality, this yields

$$\begin{aligned} W_2(\nu_k, \pi) &\leq W_2(\nu_k, \pi_\alpha) + W_2(\pi, \pi_\alpha), \\ W_1(\nu_k, \pi) &\leq W_2(\nu_k, \pi_\alpha) + W_1(\pi, \pi_\alpha), \\ d_{\mathrm{bL}}(\nu_k, \pi) &\leq W_2(\nu_k, \pi_\alpha) + d_{\mathrm{TV}}(\pi, \pi_\alpha). \end{aligned}$$

We control the common term $W_2(\nu_k, \pi_\alpha)$ as follows. By construction, the convexified potential f_α is α-strongly convex and its gradient ∇f_α is $(M+\alpha)$-Lipschitz. Therefore, a direct application of Dalalyan and Riou-Durand (2018) (Theorem 2) ensures that, if the parameters $\alpha, \gamma, h > 0$ are such that

$$\alpha \leq \frac{M}{20}, \qquad \gamma \geq \sqrt{1.1M}, \qquad h \leq \frac{\alpha}{4.2\gamma M},$$

then the distribution of the KLMC sampler after k iterates satisfies

$$W_2(\nu_k, \pi_\alpha) \leq \sqrt{2} \left(1 - \frac{3\alpha h}{4\gamma} \right)^k W_2(\nu_0, \pi_\alpha) + 1.05 \frac{Mh\sqrt{2p}}{\alpha},$$

where $1.05M$ is an upper bound for the Lipschitz constant $(M + \alpha)$.

The right hand side of the resulting inequality is a sum of two terms which are both bounded by $\varepsilon/4$ if

$$h \leq \frac{\alpha\varepsilon}{4, 2M\sqrt{2p}}, \qquad k \geq \frac{4\gamma}{3\alpha h} \log\left(\frac{4\sqrt{2}W_2(\nu_0, \pi_\alpha)}{\varepsilon}\right).$$

Therefore, assumptions of the Theorem ensures that $W_2(\nu_k, \pi_\alpha) \leq \varepsilon/2$.

Concerning the distances between π and π_α, Proposition 5 applies since $\alpha \leq 1/(5\mu_2)$ by assumption. This yields the following implications, and the claim of the theorem follows.

$$\begin{aligned} \alpha \leq \varepsilon^2/(8.3\mu_4) &\Rightarrow W_2(\pi, \pi_\alpha) \leq \varepsilon/2 \\ \alpha \leq \varepsilon/(2.1\mu_3) &\Rightarrow W_1(\pi, \pi_\alpha) \leq \varepsilon/2 \\ \alpha \leq \varepsilon/(2.1\mu_2) &\Rightarrow d_{\mathrm{bL}}(\pi, \pi_\alpha) \leq \varepsilon/2 \end{aligned}$$

□

Theorem 7. *Suppose that the potential function f satisfies Conditions 1- 3. Set the parameters $\alpha, \gamma, h > 0$ such that*

$$\alpha \leq \frac{1}{5\mu_2} \bigwedge \frac{M}{20}, \qquad \gamma \geq \sqrt{1.1M}, \qquad h \leq \frac{\alpha}{5.25\gamma M} \bigwedge \frac{\alpha}{4M_2\sqrt{5p}}$$

and let ν_k be the distribution of the k-th iterate $\boldsymbol{\vartheta}_k$ of the KLMC2 algorithm, tuned by those parameters.

If the step size h is chosen small enough such that

$$h \leq \frac{\alpha}{13M_2 \log^{1/2}\left(\frac{10}{\varepsilon\sqrt{M}}\right)} \bigwedge \left(\frac{\alpha\varepsilon}{12(M_2 p + M^{3/2}p^{1/2})}\right)^{1/2}$$

and the number of iterations K is such that

$$K \geq \frac{4\gamma}{\alpha h} \log\left(\frac{12W_2(\nu_0, \pi_\alpha)}{\varepsilon}\right)$$

then following implications hold

$$\begin{aligned} \alpha \leq \varepsilon^2/(8.3\mu_4) &\Rightarrow W_2(\nu_K, \pi) \leq \varepsilon, \\ \alpha \leq \varepsilon/(2.1\mu_3) &\Rightarrow W_1(\nu_K, \pi) \leq \varepsilon, \\ \alpha \leq \varepsilon/(2.1\mu_2) &\Rightarrow d_{\mathrm{bL}}(\nu_K, \pi) \leq \varepsilon. \end{aligned}$$

Using Condition 2 and Lemma 9 to control the moments μ_2, μ_3 and μ_4, we compute the scaling of the mixing time of KLMC2 with respect to $p, \varepsilon, \kappa, \beta$. Up to logarithmic factors, the mixing time scales as $\kappa^4 p^{4\beta+1/2}/\varepsilon^4$ for the Wasserstein-2 distance, $\kappa^3 p^{3\beta+1/2}/\varepsilon^2$ for the Wasserstein-1 distance, and $\kappa^2 p^{2\beta+1/2}/\varepsilon^2$ for the bounded Lipschitz distance. This improves the mixing time of KLMC by a factor ε for all three distances.

Proof. As shown previously in the proof of Theorem 6, the following inequalities always hold

$$\begin{aligned} W_2(\nu_k, \pi) &\leq W_2(\nu_k, \pi_\alpha) + W_2(\pi, \pi_\alpha), \\ W_1(\nu_k, \pi) &\leq W_2(\nu_k, \pi_\alpha) + W_1(\pi, \pi_\alpha), \\ d_{\mathrm{bL}}(\nu_k, \pi) &\leq W_2(\nu_k, \pi_\alpha) + d_{\mathrm{TV}}(\pi, \pi_\alpha). \end{aligned}$$

The control of the three distances between π and π_α is already made in the proof of Theorem 6. Therefore, we need only to ensure that the common term $W_2(\nu_k, \pi_\alpha)$ is bounded by $\varepsilon/2$, this is proved in the sequel. By construction, the convexified potential f_α is α-strongly convex and its gradient ∇f_α is $(M+\alpha)$-Lipschitz. Moreover, the Hessian matrix $\nabla^2 f_\alpha$ is also M_2-Lipschitz for the spectral norm. Therefore, a direct application of Theorem 3 of (Dalalyan and Riou-Durand (2018)) ensures that, if the parameters $\alpha, \gamma, h > 0$ are such that

$$\alpha \leq \frac{M}{20}, \qquad \gamma \geq \sqrt{1.1M}, \qquad h \leq \frac{\alpha}{5.25\gamma M} \bigwedge \frac{\alpha}{4M_2\sqrt{5p}}$$

then the distribution of the KLMC2 sampler after k iterates satisfies

$$\begin{aligned} W_2(\nu_k, \pi_\alpha) \leq \sqrt{2}\left(1 - \frac{\alpha h}{4\gamma}\right)^k W_2(\nu_0, \pi_\alpha) + \frac{2h^2 M_2 p}{\alpha} + \frac{h^2(1.05M)^{3/2}\sqrt{2p}}{\alpha} \\ + \frac{8.4hM}{\alpha} \exp\left\{-\frac{\alpha^2}{160M_2^2h^2}\right\}. \end{aligned}$$

where $1.05M$ is an upper bound for the Lipschitz constant $(M+\alpha)$. To simplify the last expression, we use the fact that $1.05^{3/2}\sqrt{2} \leq 2$ and $h \leq \alpha/(5.25\gamma M)$ where $\gamma \geq \sqrt{M}$. This yields

$$W_2(\nu_k, \pi_\alpha) \leq \sqrt{2}\left(1 - \frac{\alpha h}{4\gamma}\right)^k W_2(\nu_0, \pi_\alpha) + \frac{2h^2(M_2 p + M^{3/2}p^{1/2})}{\alpha} + \frac{1.6}{\sqrt{M}} \exp\left\{-\frac{(\alpha/h)^2}{160M_2^2}\right\}.$$

In this inequality, the right hand side is a sum of three terms that are all bounded by $\varepsilon/6$ if the following two inequalities hold:

$$\begin{aligned} k &\geq \frac{4\gamma}{\alpha h} \log\left(\frac{12W_2(\nu_0, \pi_\alpha)}{\varepsilon}\right), \\ h &\leq \frac{\alpha}{13M_2 \log^{1/2}\left(\frac{10}{\varepsilon\sqrt{M}}\right)} \bigwedge \left(\frac{\alpha\varepsilon}{12(M_2 p + M^{3/2}p^{1/2})}\right)^{1/2}. \end{aligned}$$

This proves the claim of the Theorem. □

4.7 Bounding moments

From the user's perspective, computing the mixing time of LMC or KLMC for a convex potential f requires the computation of some moments of the distribution $\pi(d\boldsymbol{\theta})$. Those

moments usually involve intractable intergrals. In this section, we thus propose explicit bounds on the moments

$$\mu_a \triangleq \int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}, \qquad a > 0.$$

Only assuming that f is convex ensures that $\pi(d\boldsymbol{\theta})$ is sub-exponential, but it does not guarantee that the moment μ_a will scale polynomially with the dimension. For instance, the distribution $\pi(d\boldsymbol{\theta}) \propto \exp\{-2^{-p}\|\boldsymbol{\theta}\|_2\}d\boldsymbol{\theta}$ has moments $\mu_a = 2^{pa}\Gamma(p+a)/\Gamma(p)$ that scales exponentially with the dimension. Therefore, to provide polynomial bounds, one needs to make additional assumptions on the potential function f. We investigate the case where f is m-strongly convex, inside, respectively outside, a ball of radius R around the mode $\boldsymbol{\theta}^*$. We manage to provide user-friendly bounds on μ_a, with small constants. If m and R are dimension free, then we show that μ_a scales as $(p \log p)^a$, respectively $(p \log p)^{a/2}$. The dependence on the dimension is sharp within a poly-log factor.

Proposition 6. *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 $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq R$. Then for every $a > 0$ we have*

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq A \vee B + \frac{2^{a+2}}{(mR)^a \Gamma(p/2)}$$

where

$$A = \left\{ \frac{3}{mR} \left((p+a)\log(p+a) + p \log_+ \left(\frac{2M}{m^2 R^2} \right) \right) \right\}^a$$

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 2. *If the assumptions of Proposition 6 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 6, the dominant term is A when p is large, while the dominant term is B when R is large. The residual term $2^{a+2}/((mR)^a\Gamma(p/2))$ goes to zero whenever p or R goes 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, this is proven in Lemma 8. Note that when R goes to infinity we recover exactly the bound of the strongly convex case proven in Lemma 6.

Proposition 7. *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*

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \left(1 + \frac{2}{\Gamma(p/2)} \right) A^a$$

where

$$A = (2R) \bigvee \left(\frac{6(p+a)}{m} \log \left(\frac{16pM^2}{m^2} \right) \right)^{1/2}.$$

Proposition 8. *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*

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq e^{mR^2/2} \left(\frac{p}{m}\right)^{a/2} \left\{2^{a-1}\left(1 + (1 + a/p)^{a/2-1}\right)\right\}^{\mathbb{1}_{a>2}}.$$

Remark 3. *Under the assumptions of Proposition 7, then*

$$\mu_a = \widetilde{O}\left(R^a \bigvee \left(\frac{p}{m}\right)^{a/2}\right).$$

In the bound of Proposition 7, if m and R are assumed to be dimension free constants, then μ_a scales as $(p \log p)^{a/2}$. This rate is improved in Proposition 8 to $p^{a/2}$, which is optimal. Note that if R goes to zero, we recover exactly the bound of the strongly convex case proven in Lemma 6. However, the bound of Proposition 7 is sharper when R is large. Assuming that m is a dimension-free constant, the bound remains polynomial in p whenever R grows at most polynomially with the dimension.

4.8 Postponed proofs

4.8.1 Proof of Proposition 6

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 begin by computing the map

$$\begin{aligned}
\widetilde{m}(r) &\triangleq 2\int_0^1 m(ry)(1-y)dy \\
&= 2\int_0^1 m\mathbb{1}_{]0,R[}(ry)(1-y)dy \\
&= 2m\int_0^{1\wedge R/r} (1-y)dy \\
&= m\mathbb{1}_{r<R} + m\left(2\frac{R}{r} - \frac{R^2}{r^2}\right)\mathbb{1}_{r\geq R}.
\end{aligned}$$

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 computations into the two following parts:

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{B_A} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} + \int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Concerning the second term, for any $r > A$, we have $\widetilde{m}(r)r^2/2 = mRr - mR^2/2 \geq mRr/2$. Applying Lemma 7 yields

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq \left((2/M)^{p/2}\Gamma(p/2)/2\right)^{-1} \int_A^{+\infty} r^{p+a-1} e^{-mRr/2} dr.$$

We now use the following inequality on the incomplete Gamma function from Natalini and Palumbo (2000), (also available in Borwein et al. (2009)). Let $B > 1$, and $q \geq 1$, then if $x \geq (B/(B-1))(q-1)$ then

$$\int_x^{+\infty} y^{q-1} e^{-y} dy \leq B x^{q-1} e^{-x}.$$

We apply this inequality for $B = 2$ and $q = p + a$. If one assumes that $A \geq 2(p + a - 1)/(mR)$, then

$$\begin{aligned}\int_A^{+\infty} r^{p+a-1} e^{-mRr/2} dr &= \left(\frac{2}{mR}\right)^{a+p} \int_{mRA/2}^{+\infty} y^{p+a-1} e^{-y} dy \\ &\leq \left(\frac{2}{mR}\right)^{a+p} 2 \left(\frac{mRA}{2}\right)^{p+a-1} e^{-mRA/2}.\end{aligned}$$

This yields

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq \frac{2^{a+2}}{(mR)^a \Gamma(p/2)} \left(\frac{2M}{m^2 R^2}\right)^{p/2} \left(\frac{mRA}{2}\right)^{p+a-1} e^{-mRA/2}.$$

The last bound ensures that the inequality

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq \frac{2^{a+2}}{(mR)^a \Gamma(p/2)} \tag{4.13}$$

is fulfilled whenever $\varphi(x) \triangleq x - c \log(x) - b \geq 0$, where

$$x = \frac{mRA}{2}, \qquad c = p + a - 1, \qquad b = \frac{p}{2} \log\left(\frac{2M}{m^2 R^2}\right).$$

Taylor's expansion around $1.5(c+1)\log(c+1)$ yields

$$\varphi(1.5(c+1)\log(c+1) + 3(b \vee 0)) = \varphi(1.5(c+1)\log(c+1)) + \varphi'(y) \times 3(b \vee 0)$$

for some $y \geq 1.5(c+1)\log(c+1)$, which implies that $\varphi'(y) \geq 1 - c/(1.5(c+1)\log(c+1)) \geq 1/3$. We get

$$\varphi(1.5(c+1)\log(c+1)+3(b\vee 0)) \geq 1.5(c+1)\log(c+1) - c\log(1.5(c+1)\log(c+1)) + (b\vee 0) - b \geq 0.$$

Since the map φ is increasing on $[c, +\infty[$ and $1.5(c+1)\log(c+1) + 3(b \vee 0) \geq c$, we conclude that (4.13) is fulfilled for any

$$A \geq A^* \triangleq \frac{3}{mR} \left((p+a)\log(p+a) + p \log_+\left(\frac{2M}{m^2 R^2}\right)\right).$$

Recall that $A \geq R$ by assumption, this brings two cases to consider. Firstly, if $R < A^*$, then for $A = A^*$ we have

$$\int_{B_A} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq A^a.$$

Secondly, if $R \geq A^*$, then 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 6 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 (4.13) is fulfilled in both cases, the claim of the Proposition follows.

4.8.2 Proof of Proposition 7

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

$$m(r) = m\mathbb{1}_{]R,+\infty[}(r)$$

We begin by computing the map

$$\begin{aligned}
\widetilde{m}(r) &\triangleq 2\int_0^1 m(ry)(1-y)dy \\
&= 2\int_0^1 m\mathbb{1}_{]R,+\infty[}(ry)(1-y)dy \\
&= 2m\mathbb{1}_{r>R}\int_{R/r}^1 (1-y)dy \\
&= m\left(1-R/r\right)^2\mathbb{1}_{r>R}.
\end{aligned}$$

Let $A \geq 2R$ 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 7 yields

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta})d\boldsymbol{\theta} \leq \left((2/M)^{p/2}\Gamma(p/2)/2\right)^{-1} \int_A^{+\infty} r^{p+a-1}e^{-mr^2/8}dr$$

since $\widetilde{m}(r) \geq r^2/4$ for every $r \geq A \geq 2R$.

We now use the following inequality on the incomplete Gamma function from Natalini and Palumbo (2000), (also available in Borwein et al. (2009)). Let $B > 1$, and $q \geq 1$, then if $x \geq (B/(B-1))(q-1)$ then

$$\int_x^{+\infty} y^{q-1}e^{-y}dy \leq Bx^{q-1}e^{-x}.$$

We apply this inequality for $B = 2$ and $q = (p+a)/2$. If one assumes that $mA^2/8 \geq (p+a)/2 - 1$, then

$$\begin{aligned}
\int_A^{+\infty} r^{p+a-1}e^{-mr^2/8}dr &= 2^{-1}\left(\frac{8}{m}\right)^{(p+a)/2}\int_{mA^2/8}^{+\infty} y^{(p+a)/2-1}e^{-y}dy \\
&\leq \left(\frac{8}{m}\right)^{(p+a)/2}\left(\frac{mA^2}{8}\right)^{(p+a)/2-1} e^{-mA^2/8} \\
&= A^a\left(\frac{8}{m}\right)^{p/2}\left(\frac{mA^2}{8}\right)^{p/2-1} e^{-mA^2/8}.
\end{aligned}$$

This yields

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta})d\boldsymbol{\theta} \leq \frac{2A^a}{\Gamma(p/2)}\left(\frac{4M}{m}\right)^{p/2}\left(\frac{mA^2}{8}\right)^{p/2-1} e^{-mA^2/8}.$$

The last bound ensures that the inequality

$$\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq \frac{2A^a}{\Gamma(p/2)} \tag{4.14}$$

is fulfilled whenever $\varphi(x) \triangleq x - c\log(x) - b \geq 0$, where

$$x = \frac{mA^2}{8}, \qquad c = \frac{p}{2} - 1, \qquad b = \frac{p}{2}\log\left(\frac{4M}{m}\right) > 0.$$

Taylor's expansion around $1.5(c+1)\log(c+1)$ yields

$$\varphi(1.5(c+1)\log(c+1) + 3b) = \varphi(1.5(c+1)\log(c+1)) + \varphi^{'}(y) \times 3b$$

for some $y \geq 1.5(c+1)\log(c+1)$, which implies that $\varphi^{'}(y) \geq 1 - c/(1.5(c+1)\log(c+1)) \geq 1/3$. We get

$$\varphi(1.5(c+1)\log(c+1)+3b) \geq 1.5(c+1)\log(c+1) - c\log(1.5(c+1)\log(c+1)) + b - b \geq 0.$$

Since the map φ is increasing on $[c, +\infty[$ and $1.5(c+1)\log(c+1) + 3b \geq c$, we conclude that (4.14) is fulfilled for any

$$\begin{aligned} A^2 &\geq \frac{6}{m}\left(p\log(p/2) + 2p\log(4M/m)\right) \\ &= \frac{6p}{m}\log\left(\frac{16pM^2}{m^2}\right). \end{aligned}$$

Finally, we choose A such that this inequality and the two additional assumptions: $A \geq 2R$ and $mA^2/8 \geq (p+a)/2 - 1$ hold, that is

$$A = (2R) \bigvee \left(\frac{6(p+a)}{m}\log\left(\frac{16pM^2}{m^2}\right)\right)^{1/2}.$$

Such a choice yields the claim of the Proposition.

4.8.3 Proof of Proposition 8

Define $f = -\log\pi$ and for any $\boldsymbol{\theta} \in \mathbb{R}^p$:

$$\bar{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + \frac{m}{2}\left(\|\boldsymbol{\theta}\|_2 - R\right)^2 \mathbb{1}_{\|\boldsymbol{\theta}\|_2 \leq R}.$$

The map $\bar{f}$ is m-strongly convex, moreover $\bar{f}(\boldsymbol{\theta}) \geq f(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \mathbb{R}^p$, this yields

$$\begin{aligned} \int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} &\geq \int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a e^{-\bar{f}(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \bar{C} \int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \left(e^{-\bar{f}(\boldsymbol{\theta})}/\bar{C}\right) d\boldsymbol{\theta} \end{aligned}$$

where

$$\bar{C} \triangleq \int_{\mathbb{R}^p} e^{-\bar{f}(\boldsymbol{\theta})} d\boldsymbol{\theta} \geq e^{mR^2/2}.$$

Applying Lemma 6 to the probability density $e^{-\bar{f}(\boldsymbol{\theta})}/\bar{C}$ yields the claim of the Proposition.

4.8.4 Proof of Proposition 5

Without loss of generality we may assume that $\int_{\mathbb{R}^p} \exp(-f(\boldsymbol{\theta}))d\boldsymbol{\theta} = 1$. We first give upper and lower bounds to the normalizing constant of π_α, that is

$$c_\alpha \triangleq \int_{\mathbb{R}^p} \pi(\boldsymbol{\theta}) e^{-\alpha\|\boldsymbol{\theta}\|_2^2/2} d\boldsymbol{\theta}.$$

The constant c_α is an expectation with respect to the density π, it can be trivially upper bounded by 1, and lower bounded by Jensen's inequality applied to the convex map $x \mapsto e^{-x}$. These two facts yield

$$\exp\{-\alpha\mu_2/2\} \leq c_\alpha \leq 1.$$

Now we control the distance between densities π and π_α at any fixed $\boldsymbol{\theta} \in \mathbb{R}^p$:

$$\begin{aligned}
|\pi(\boldsymbol{\theta}) - \pi_\alpha(\boldsymbol{\theta})| &= \pi(\boldsymbol{\theta}) \left| 1 - \frac{e^{-\alpha\|\boldsymbol{\theta}\|_2^2/2}}{c_\alpha} \right| \\
&\leq \pi(\boldsymbol{\theta}) \left\{ \left(1 - e^{-\alpha\|\boldsymbol{\theta}\|_2^2/2}\right) + e^{-\alpha\|\boldsymbol{\theta}\|_2^2/2} \left(\frac{1}{c_\alpha} - 1\right) \right\} \\
&\leq \pi(\boldsymbol{\theta}) \left(\alpha\|\boldsymbol{\theta}\|_2^2/2 + \exp\{\alpha\mu_2/2\} - 1\right).
\end{aligned}$$

The Total Variation distance between densities π and π_α is easily bounded by integrating the previous inequality, that is

$$\begin{aligned}
d_{\mathrm{TV}}(\pi_\alpha, \pi) &= \int_{\mathbb{R}^p} |\pi(\boldsymbol{\theta}) - \pi_\alpha(\boldsymbol{\theta})| d\boldsymbol{\theta} \\
&\leq \int_{\mathbb{R}^p} \pi(\boldsymbol{\theta}) \left(\alpha\|\boldsymbol{\theta}\|_2^2/2 + \exp\{\alpha\mu_2/2\} - 1\right) d\boldsymbol{\theta} \\
&\leq \alpha\mu_2/2 + \exp\{\alpha\mu_2/2\} - 1
\end{aligned}$$

which is the first claim of the proposition.

To bound W_q for any $q \geq 1$, we use an inequality from Villani (2008) (Theorem 6.15, page 115):

$$W_q^q(\mu, \nu) \leq 2^{q-1} \int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^q |\mu(\boldsymbol{\theta}) - \nu(\boldsymbol{\theta})| d\boldsymbol{\theta}.$$

Combining this with the bound on $|\pi(\boldsymbol{\theta}) - \pi_\alpha(\boldsymbol{\theta})|$ shown above, we have

$$\begin{aligned}
W_q^q(\pi, \pi_\alpha) &\leq 2^{q-1} \int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^q \pi(\boldsymbol{\theta}) \left(\alpha\|\boldsymbol{\theta}\|_2^2/2 + \exp\{\alpha\mu_2/2\} - 1\right) \\
&\leq 2^{q-2}\alpha\mu_{q+2} + 2^{q-1}\mu_q(\exp\{\alpha\mu_2/2\} - 1).
\end{aligned}$$

which is the second claim of the proposition.

Finally, the monotonicity of the $\mathbb{L}_q$ norm ensures that $\mu_q\mu_2 \leq \mu_{q+2}$. Numerical constants follow from the inequality $e^x - 1 \leq 1.06x$ for $x \leq 1/10$, and from Lemma 9.

4.8.5 Technical lemmas

Lemma 6. *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. Durmus and Moulines (2016) 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 monotonicity 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 Hargé (2004). 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

$$\mathbb{E}[g(Y - \mathbb{E}[Y])] \leq \mathbb{E}[g(X - \mathbb{E}[X])].$$

Now $f = -\log \pi$ is m-strongly convex, thus for the particular choice $\mu = \mathbf{0}_p$ and $\Sigma = m\mathbf{I}_p$, then π/φ remains log-concave. Applied to the convex map $g : \boldsymbol{\theta} \mapsto \|\boldsymbol{\theta}\|_2^a$, the inequality of Hargé (2004) 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 Qi et al. (2012) (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 for $x = p/2$ and $y = a/2 > 1$, this yields

$$\mathbb{E}_\pi[\|\boldsymbol{\theta} - \mathbb{E}_\pi[\boldsymbol{\theta}]\|_2^a] \leq \left(\frac{p}{m}\right)^{a/2} \left(1 + (1 + a/p)^{a/2-1}\right). \tag{4.15}$$

We now bound the distance between the mean and the mode

$$\begin{aligned} \mathcal{D} &\triangleq \|\mathbb{E}_\pi[\boldsymbol{\theta}] - \boldsymbol{\theta}^*\|_2 \\ &\leq \mathbb{E}_\pi[\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2] \\ &\leq \left(\frac{p}{m}\right)^{1/2}. \end{aligned} \tag{4.16}$$

For any $x, y \geq 0$ we have $(x+y)^a \leq 2^{a-1}(x^a + y^a)$, this yields

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq 2^{a-1} \left(\mathbb{E}[\|\boldsymbol{\theta} - \mathbb{E}_\pi[\boldsymbol{\theta}]\|_2^a] + \mathcal{D}^a\right)$$

Using bounds (4.15) and (4.16) in the last expression yields the claim of the lemma for $a > 2$. □

Lemma 7. *Assume there exists a measurable map $m : [0, +\infty[\mapsto [0, M]$ such that 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 $B_A \triangleq \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq A\}$. We have*

$$\int_{(B_A)^c} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \left((2/M)^{p/2}\Gamma(p/2)/2\right)^{-1} \int_A^{+\infty} r^{p+a-1} e^{-\widetilde{m}(r)r^2/2} dr$$

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 $\nabla^2 f \preceq M\mathbf{I}_p$.

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

$$\begin{aligned}
\int_{(B_A)^c} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} &\leq C^{-1} \int_{\|\boldsymbol{\theta}\|_2 \geq A} \|\boldsymbol{\theta}\|_2^a e^{-f(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&\leq \left(\int_{\mathbb{R}^p} e^{-M\|\boldsymbol{\theta}\|_2^2/2} d\boldsymbol{\theta} \right)^{-1} \int_{\|\boldsymbol{\theta}\|_2 \geq A} \|\boldsymbol{\theta}\|_2^a e^{-\widetilde{m}(\|\boldsymbol{\theta}\|_2)\|\boldsymbol{\theta}\|_2^2/2} d\boldsymbol{\theta} \\
&= \left(\int_0^{+\infty} r^{p-1} e^{-Mr^2/2} dr \right)^{-1} \int_A^{+\infty} r^{a+p-1} e^{-\widetilde{m}(r)r^2/2} dr \\
&= \left((2/M)^{p/2}\Gamma(p/2)/2 \right)^{-1} \int_A^{+\infty} r^{a+p-1} e^{-\widetilde{m}(r)r^2/2} dr
\end{aligned}$$

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 8. *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 $p \geq 2 \vee (a-1)$

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \geq (0.1)\Gamma(p+a)/\Gamma(p) \underset{p\to+\infty}{\sim} 0.1p^a.$$

This proves that, under assumptions of Proposition 6 (here with $m = R = 1$), the dependence p^a is not improvable.

Proof. Remark first that $f(\boldsymbol{\theta}) = \varphi(\|\boldsymbol{\theta}\|_2)$ where

$$\varphi(r) \triangleq 0.5r^2 \mathbb{1}_{r\leq 1} + r\mathbb{1}_{r>1}.$$

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, and the last inequality follows from the fact that $\Gamma(.)$ is an increasing function on $[2, +\infty[$. This proves the claim of the Lemma. □

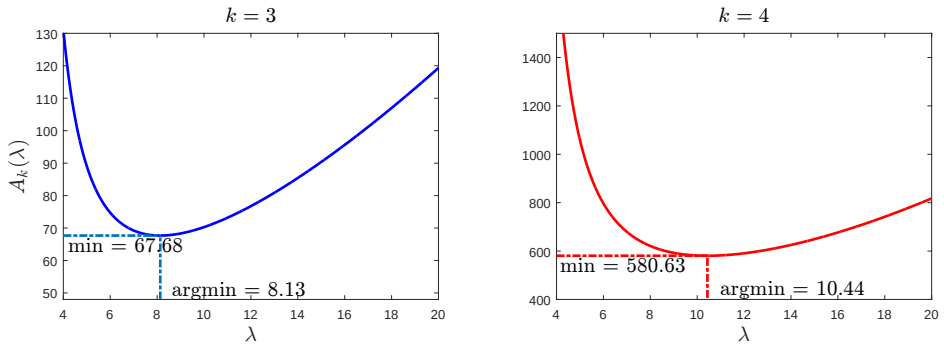

Figure 4.1: Plots of bounds from Lemma 9 for $k = 3$ and $k = 4$.

Lemma 9. *Let $\Gamma(k, x)$ be the upper incomplete Gamma function. If $k \geq 2$ is a positive integer, then $\mu_k \leq A_k \mu_2^{k/2}$ where $A_k = \min_{\lambda>2} A_k(\lambda)$ with*

$$A_k(\lambda) = \frac{\sqrt{\lambda-1}}{\lambda}\left[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\right]^k k\Gamma\left(k, \frac{\log(\lambda-1)}{2}\right) + \lambda^{k/2}. \tag{4.17}$$

Proof. Let us define A by $\{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2^2 \leq \lambda\mu_2\}$, for any $\lambda > 1$. From Tchebyshev's inequality we have

$$\pi(A) \geq 1 - \frac{\mathbb{E}_\pi[\|\boldsymbol{\theta}\|_2^2]}{\lambda\mu_2} = 1 - \frac{1}{\lambda}.$$

The A being symmetric, Proposition 2.14 from (Ledoux, 2001) implies the following inequality:

$$1 - \pi(tA) \leq \pi(A) \left(\frac{1 - \pi(A)}{\pi(A)}\right)^{(t+1)/2},$$

for every real number t larger than 1. Since the right-hand side is a decreasing function of $\pi(A)$, we obtain the following bound on $\pi(tA^C)$:

$$\pi(tA^C) \leq \frac{1}{\lambda \cdot (\lambda - 1)^{(t-1)/2}}.$$

Now let us introduce random variable η as $\|\boldsymbol{\theta}\|_2/\sqrt{\mu_2}$, where $\boldsymbol{\theta} \sim \pi$. It is clear that (4.17) is equivalent to

$$\mathbb{E}[\eta^k] \leq \frac{\sqrt{\lambda - 1}}{\lambda}\left[\frac{2\sqrt{\lambda}}{\log(\lambda - 1)}\right]^k \cdot k\Gamma\left(k, \frac{\log(\lambda - 1)}{2}\right) + \lambda^{k/2}.$$

Since $\eta > 0$ almost surely,

$$\mathbb{E}[\eta^k] = \int_{\mathbb{R}} P(\eta^k > t)dt = k\int_{\mathbb{R}} t^{k-1}P(\eta > t)dt.$$

Thus the proof lemma of reduces to bound the tail of η. The definition of η yields

$$P(\eta > t) = P(\|\boldsymbol{\theta}\|_2 > t\sqrt{\mu_2}) = \pi\left(\frac{t}{\sqrt{\lambda}} \cdot A^C\right) \leq \frac{1}{\lambda \cdot (\lambda - 1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}},$$

when $t > \sqrt{\lambda}$. Therefore we have

$$\begin{aligned}
\mathbb{E}[\eta^k] &\leq k\int_{\sqrt{\lambda}}^{\infty} \frac{t^{k-1}}{\lambda \cdot (\lambda - 1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}}dt + \int_0^{\sqrt{\lambda}} kt^{k-1}P(\eta > t)dt \\
&\leq k\int_{\sqrt{\lambda}}^{\infty} \frac{t^{k-1}}{\lambda \cdot (\lambda - 1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}}dt + \lambda^{k/2}.
\end{aligned}$$

One can notice that the first integral can be calculated using the upper incomplete gamma function $\Gamma(k, z)$. Indeed, the change of variable $z = t\log(\lambda - 1)/(2\sqrt{\lambda})$ yields

$$\begin{aligned}
\int_{\sqrt{\lambda}}^{\infty} \frac{t^{k-1}}{\lambda \cdot (\lambda - 1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}}\, dt &= \frac{\sqrt{\lambda - 1}}{\lambda}\int_{\sqrt{\lambda}}^{\infty} t^{k-1}\exp\left(-\ell_n(\lambda - 1)\frac{t}{2\sqrt{\lambda}}\right)dt \\
&= \frac{\sqrt{\lambda - 1}}{\lambda}\left[\frac{2\sqrt{\lambda}}{\log(\lambda - 1)}\right]^k \int_{\frac{\log(\lambda-1)}{2}}^{\infty} z^{k-1}e^{-z}\, dz \\
&= \frac{\sqrt{\lambda - 1}}{\lambda}\left[\frac{2\sqrt{\lambda}}{\log(\lambda - 1)}\right]^k \Gamma\left(k, \frac{\log(\lambda - 1)}{2}\right).
\end{aligned}$$

Finally, bounding the incomplete Gamma function by factorial we obtain

$$\mathbb{E}[\eta^k] \leq k \cdot \frac{\sqrt{\lambda-1}}{\lambda}\left[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\right]^k \Gamma\left(k, \frac{\log(\lambda-1)}{2}\right) + \lambda^{k/2}.$$

This concludes the proof. □

Remark 4. *Figure 4.1 shows the shape of the function* $\lambda \mapsto A_k(\lambda)$ *for* $k = 3$ *and* $k = 4$. *We see, in particular, that the optimal choice for* λ *is approximately* 8.13 *for* $k = 3$ *and* 10.44 *for* $k = 4$. *This leads to the numerical bounds*

$$A_k \leq \begin{cases} 67.7, & k = 3 \\ 580.7, & k = 4 \end{cases}.$$

These constants are by no means optimal, but we are not aware of any better bound available in the literature.

Bibliography

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414.

Ando, T. (1979). Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra and its Applications*, 26:203–241.

Baker, J., Fearnhead, P., Fox, E. B., and Nemeth, C. (2018). Control variates for stochastic gradient MCMC. *Statistics and Computing*.

Barthelmé, S. and Chopin, N. (2015). The Poisson transform for unnormalised statistical models. *Statistics and Computing*, 25(4):767–780.

Bernton, E. (2018). Langevin Monte Carlo and JKO splitting. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1777–1798.

Borwein, J. M., Chan, O., et al. (2009). Uniform bounds for the complementary incomplete gamma function. *Mathematical Inequalities and Applications*, 12:115–121.

Bou-Rabee, N. and Hairer, M. (2013). Nonasymptotic mixing of the MALA algorithm. *IMA J. Numer. Anal.*, 33(1):80–110.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Bradley, R. C. et al. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2:107–144.

Brascamp, H. J. and Lieb, E. H. (1976). 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.

Brazitikos, S., Giannopoulos, A., Valettas, P., and Vritsiou, B.-H. (2014). *Geometry of isotropic convex bodies*, volume 196. American Mathematical Soc.

Brosse, N., Durmus, A., Moulines, É., and Pereyra, M. (2017). Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 319–342.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55.

Catoni, O. (2007). PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*.

Chatterji, N., Flammarion, N., Ma, Y., Bartlett, P., and Jordan, M. (2018). On the theory of variance reduction for stochastic gradient Monte Carlo. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 764–773, Stockholmsmassan, Stockholm Sweden.

Cheng, X. and Bartlett, P. (2018). Convergence of Langevin MCMC in KL-divergence. In *Proceedings of ALT2018*.

Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. (2018). Sharp Convergence Rates for Langevin Dynamics in the Nonconvex Setting. *ArXiv e-prints*.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. (2018). Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the Conference on Learning Theory (COLT2018)*.

Collier, O. and Dalalyan, A. S. (2017). Minimax estimation of a p-dimensional linear functional in sparse Gaussian models and robust estimation of the mean. submitted 1712.05495, arXiv.

Dalalyan, A. (2017a). Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689.

Dalalyan, A. and Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61.

Dalalyan, A. S. (2017b). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676.

Dalalyan, A. S. and Karagulyan, A. (2019). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*.

Dalalyan, A. S. and Riou-Durand, L. (2018). On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv preprint arXiv:1807.09382*.

Dalalyan, A. S. and Tsybakov, A. B. (2009). 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.

Delyon, B. (2018). Estimation paramétrique. *Unpublished lecture notes*.

Desvillettes, L. and Villani, C. (2001). On the trend to global equilibrium in spatially inhomogeneous entropy-dissipating systems: the linear Fokker-Planck equation. *Comm. Pure Appl. Math.*, 54(1):1–42.

Dieuleveut, A., Durmus, A., and Bach, F. (2017). Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. *ArXiv e-prints*.

Dolbeault, J., Mouhot, C., and Schmeiser, C. (2015). Hypocoercivity for linear kinetic equations conserving mass. *Trans. Amer. Math. Soc.*, 367(6):3807–3828.

Douc, R., Moulines, E., and Rosenthal, J. S. (2004). Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Ann. Appl. Probab.*, 14(4):1643–1665.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222.

Durmus, A., Majewski, S., and Miasojedow, B. (2018). Analysis of Langevin Monte Carlo via convex optimization. *arXiv preprint arXiv:1802.09188*.

Durmus, A. and Moulines, E. (2016). High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *ArXiv e-prints*.

Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587.

Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. (2018). Log-concave sampling: Metropolis-hastings algorithms are fast! In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 793–797.

Eberle, A. (2014). Error bounds for Metropolis-Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337–377.

Eberle, A., Guillin, A., and Zimmer, R. (2017). Couplings and quantitative contraction rates for Langevin dynamics. *ArXiv e-prints*.

Eberle, A. and Majka, M. B. (2019). Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability*, 24.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 261–274.

Geyer, C. J. (2012). The Wald consistency theorem. *Unpublished lecture notes*.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):657–683.

Gu, M. G. and Zhu, H.-T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(2):339–355.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13:307–361.

Hargé, G. (2004). 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.

Helffer, B. and Nier, F. (2005). *Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians*, volume 1862 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

Jain, P. and Kar, P. (2017). Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336.

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probab. Surv.*, 1:299–320.

Lamberton, D. and Pagès, G. (2002). Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, 8(3):367–405.

Lamberton, D. and Pagès, G. (2003). Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift. *Stoch. Dyn.*, 3(4):435–451.

Ledoux, M. (2001). *The concentration of measure phenomenon*. American Mathematical Soc.

Luu, T. D., Fadili, J., and Chesneau, C. (2017). Sampling from non-smooth distribution through Langevin diffusion. hal-01492056v3.

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H., and Simpson, D. (2015). On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical science*, 30(4):443–467.

McAllester, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The annals of Statistics*, 24(1):101–121.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Mnih, A. and Teh, Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1751–1758.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov Chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Murray, I., Ghahramani, Z., and MacKay, D. (2012). MCMC for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

Natalini, P. and Palumbo, B. (2000). Inequalities for the incomplete gamma function. *Math. Inequal. Appl*, 3(1):69–77.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2.

Nelson, E. (1967). *Dynamical Theories of Brownian Motion*. Department of Mathematics. Princeton University.

Pavliotis, G. A. (2014). *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York. Diffusion processes, the Fokker-Planck and Langevin equations.

Pillai, N. S., Stuart, A. M., and Thiéry, A. H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Ann. Appl. Probab.*, 22(6):2320–2356.

Qi, F., Luo, Q.-M., et al. (2012). 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.

Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703.

Riou-Durand, L. and Chopin, N. (2018). Noise contrastive estimation: Asymptotic properties, formal comparison with mc-mle. *Electron. J. Statist.*, 12(2):3473–3518.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71.

Roberts, G. O. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, 4(4):337–357 (2003).

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007). Recent developments in exponential random graph (p*) models for social networks. *Social networks*, 29(2):192–215.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Salakhutdinov, R. and Hinton, G. (2009). Deep Boltzmann machines. In *Artificial Intelligence and Statistics*, pages 448–455.

Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45.

Stramer, O. and Tweedie, R. L. (1999a). Langevin-type models. I. Diffusions with given stationary distributions and their discretizations. *Methodol. Comput. Appl. Probab.*, 1(3):283–306.

Stramer, O. and Tweedie, R. L. (1999b). Langevin-type models. II. Self-targeting candidates for MCMC algorithms. *Methodol. Comput. Appl. Probab.*, 1(3):307–328.

Tat Lee, Y., Song, Z., and Vempala, S. S. (2018). Algorithmic Theory of ODEs and Sampling from Well-conditioned Logconcave Densities. *arXiv e-prints*, page arXiv:1812.06243.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Wald, A. (1949). Note on the consistency of the Maximum Likelihood Estimate. *Ann. Math. Statistics*, 20:595–601.

Walker, S. G. (2011). Posterior sampling when the normalizing constant is unknown. *Comm. Statist. Simulation Comput.*, 40(5):784–792.

Wang, C., Komodakis, N., and Paragios, N. (2013). Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610–1627.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S., and Girolami, M. (2014). Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statist. Probab. Lett.*, 91:14–19.

Xu, P., Chen, J., Zou, D., and Gu, Q. (2017). Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization. *ArXiv e-prints*.

Zhang, Y., Liang, P., and Charikar, M. (2017). A hitting time analysis of stochastic gradient Langevin dynamics. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1980–2022.

List of Figures

List of Tables

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Contributions théoriques aux méthodes de Monte Carlo, et applications à la statistique

Mots Clefs : échantillonnage MCMC, M-estimateurs, ergodicité géométrique, temps de mélange, couplages, distance de Wasserstein

Résumé :
La première partie de cette thèse concerne l'inférence de modèles statistiques non normalisés. Nous étudions deux méthodes d'inférence basées sur de l'échantillonnage aléatoire : Monte-Carlo MLE (Geyer, 1994), et Noise Contrastive Estimation (Gutmann et Hyvarinen, 2010). Cette dernière méthode fut soutenue par une justification numérique d'une meilleure stabilité, mais aucun résultat théorique n'avait encore été prouvé. Nous prouvons que Noise Contrastive Estimation est plus robuste au choix de la distribution d'échantillonnage. Nous évaluons le gain de précision en fonction du budget computationnel. La deuxième partie de cette thèse concerne l'échantillonnage aléatoire approché pour les distributions de grande dimension. La performance de la plupart des méthodes d'échantillonnage se détériore rapidement lorsque la dimension augmente, mais plusieurs méthodes ont prouvé leur efficacité (e.g. Hamiltonian Monte Carlo, Langevin Monte Carlo). Dans la continuité de certains travaux récents (Eberle et al., 2017 ; Cheng et al., 2018), nous étudions certaines discrétisations d'un processus connu sous le nom de kinetic Langevin diffusion. Nous établissons des vitesses de convergence explicites vers la distribution d'échantillonnage, qui ont une dépendance polynomiale en la dimension. Notre travail améliore et étend les résultats de Cheng et al. pour les densités log-concaves.

Title : Theoretical contributions to Monte Carlo methods, and applications to statistics

Keys words : MCMC sampling, M-estimators, geometric ergodicity, mixing times, couplings, Wasserstein distance

Abstract : The first part of this thesis concerns the inference of un-normalized statistical models. We study two methods of inference based on sampling, known as Monte-Carlo MLE (Geyer, 1994), and Noise Contrastive Estimation (Gutmann and Hyvarinen, 2010). The latter method was supported by numerical evidence of improved stability, but no theoretical results had yet been proven. We prove that Noise Contrastive Estimation is more robust to the choice of the sampling distribution. We assess the gain of accuracy depending on the computational budget. The second part of this thesis concerns approximate sampling for high dimensional distributions. The performance of most samplers deteriorates fast when the dimension increases, but several methods have proven their effectiveness (e.g. Hamiltonian Monte Carlo, Langevin Monte Carlo). In the continuity of some recent works (Eberle et al., 2017; Cheng et al., 2018), we study some discretizations of the kinetic Langevin diffusion process and establish explicit rates of convergence towards the sampling distribution, that scales polynomially fast when the dimension increases. Our work improves and extends the results established by Cheng et al. for log-concave densities.