Electron. J. Probab. **0** (2023), article no. 0, 1–39.
ISSN: 1083-6489 https://doi.org/10.1214/YY-TN

Stochastic Gradient Langevin Dynamics for (weakly) log-concave posterior distributions*

Marelys Crespo Navas† Sébastien Gadat‡ Xavier Gendre§¶

Abstract

In this paper, we investigate a continuous time version of the Stochastic Gradient Langevin Dynamics, introduced in [53], that incorporates a stochastic sampling step inside the traditional over-damped Langevin diffusion. This method is popular in machine learning for sampling a posterior distribution. We will pay specific attention to the computational cost in terms of n (the number of observations that produces the posterior distribution), and d (the dimension of the ambient space where the parameter of interest is living). We derive our analysis in the weakly convex framework, which is parameterized with the help of the Kurdyka-Łojasiewicz (KL) inequality, that permits to handle a vanishing curvature settings, which is far less restrictive when compared to the simple strongly convex case. We establish that the final horizon of simulation to obtain an ε approximation (in terms of entropy) is of the order $(d\log^2(n))^{(1+r)^2}[\log^2(\varepsilon^{-1}) + n^2 d^{2(1+r)}\log^{4(1+r)}(n)]$ with a Poissonian sub-sampling of parameter $(d\log^2(n))^{-(1+r)^2}$, where the parameter r is involved in the KL inequality and varies between 0 (strongly log-concave case) and 1 (limiting Laplace situation).

Keywords: Stochastic Gradient Langevin Dynamics; Log-concave models; Weak convexity.
MSC2020 subject classifications: 65C05; 62C10; 65C30; 60H35.
Submitted to EJP on XXX, final version accepted on XXX.

*This work is partially supported by the French Agence Nationale de la Recherche (ANR), project under reference ANR-PRC-CE23 MASDOL. S. Gadat acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future program (Investissements d'Avenir, grant ANR-17-EURE-0010). S. Gadat also gratefully acknowledges the Centre Lagrange for its support.
M. Crespo Navas received support from the University Research School EUR-MINT (State support managed by the National Research Agency for Future Investments program bearing the reference ANR-18-EURE-0023)

†Toulouse School of Economics (CNRS UMR 5314), Université Toulouse I Capitole.
E-mail: marelys.crespo-navas@ut-capitole.fr
‡Toulouse School of Economics (CNRS UMR 5314), Université Toulouse I Capitole, Intitut Universitaire de France. France.
E-mail: sebastien.gadat@tse-fr.eu
§Institut de Mathématiques de Toulouse (CNRS UMR 5219), Université de Toulouse Toulouse, France.
E-mail: xavier.gendre@math.univ-toulouse.fr
¶Pathway.com, Paris, France

1 Markovian Stochastic Langevin Dynamics and main results

1.1 Introduction

Motivations: In the recent past years, a huge amount of methods have been developed in machine learning to handle large scale massive datasets with a large number n of observations $(X_1, \dots, X_n)$ embedded in a high dimensional space $\mathbb{R}^d$. These methods generally involve either optimization of a data-dependent function (for frequentist learning) or sampling a data-dependent measure (for Bayesian learning with posterior distributions). In both approaches, a bottleneck lies on the size of n and d that usually generates numerical difficulties for the use of standard algorithms. In this paper, we are interested in the approximation of a posterior distribution following a Bayesian point of view with a statistical model described by a collection of densities $(p_\theta)_{\theta\in\mathbb{R}^d}$ on $\mathcal{X}$, where the parameter of interest θ belongs to $\mathbb{R}^d$ and where the $(X_i)_{1\leq i\leq n}$ are assumed to be i.i.d. observations in $\mathcal{X}$ distributed according to $p_{\theta^\star}$. A standard Bayesian approach consists of defining a prior distribution π_0 on $\mathbb{R}^d$ and then sample the posterior distribution denoted by μ_n (which will be proportional to $\exp(-U_{\nu_n})$ below) using a numerical probabilistic approximation with the help of a Langevin Dynamics (LD for short and also known as over-damped Langevin diffusion):

$$\mathrm{d}\theta_t = -\nabla_\theta U_{\nu_n}(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

In this work, we manage to deal with an adaptation of the Stochastic Gradient Langevin Dynamics (SGLD) algorithm proposed in [53], that exploits some old ideas of stochastic algorithms introduced in [47]: instead of using the previous equation, the authors propose a modification of the diffusion that generates a noisy drift in the SGLD due to a sampling strategy among the set of observations $(X_i)_{1\leq i\leq n}$. Before we provide some details on the precise objects and algorithm necessary to properly define this method, we first give some literature insights related to it.

State of the art: Approximating measures has a long-standing history and relies on Markov dynamics. We briefly introduce and describe below some well-known and recent results around this issue, and then motivate our work. For our purpose, we assume that U is any coercive function $\mathbb{R}^d \longrightarrow \mathbb{R}$.

- **(Over-damped) Langevin Dynamics (LD)**

$$\mathrm{d}\theta_t = -\nabla_\theta U(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \tag{1.1}$$

Under mild conditions on the potential U, [14, 48] proved that the LD defined in Equation (1.1) converges to the unique stationary distribution $\mu(\theta) \propto \exp(-U(\theta))$. Ergodicity and quantitative mixing properties of LD and many other sampling algorithms is a popular subject of research initiated in the probabilistic works around, roughly speaking, two strategies.

Coupling approaches The first one relies on pathwise considerations and dynamical properties of random dynamical systems and is built with some coupling argument and Lyapunov controls. We refer to the seminal contributions [42, 37], that exploits the approach of the Doeblin coupling and total variation (TV) bounds. Many extensions may be derived from this Lyapunov approach and may lead to Wasserstein or $\mathbb{L}^2$ upper bounds, we refer to [10] and the references therein of the same authors for a description of the link between Lyapunov conditions and ergodicity.

Functional inequality approaches The second strategy derives from spectral properties of Markov operators and is related to famous functional inequalities (Poincaré and log-Sobolev among others). The general idea is to differentiate the distance along the

time-evolution and apply a Gronwall Lemma to obtain a quantitative estimate of the long-time evolution of the semi-group. We refer to the seminal contributions of [35, 3], and to [4] for an almost exhaustive survey of all possible inequalities and consequences on the ergodicity of the Markov semi-groups. Finally, let us emphasize that some strong links exist between the spectral and the Lyapunov approaches, as pointed out by [2]. If functional inequalities are then strongly related to mixing properties and especially from a quantitative point of view, it is therefore necessary to develop a machinery that is able to assess these inequalities carefully, especially with a specific attention to our statistical setting of large n and d in the completely non-trivial situation where the target measure is *log-concave* but *not strongly log-concave*, which is a common feature of Bayesian posterior distributions.

• **Langevin Monte Carlo (LMC):** usually refers to the discretization of LD and allows for a concrete algorithm to approximate $\exp(-U(\theta))$. LMC stands for an approximation algorithm of LD, which is the Euler with a step-size $\eta > 0$ of Equation (1.1):

$$\vartheta_{k+1} = \vartheta_k - \eta \nabla_\vartheta U(\vartheta_k) + \sqrt{2\eta}\xi_{k+1}, \quad k \geq 0, \tag{1.2}$$

where $(\xi_k)_{k\geq 1}$ are standard Gaussian random variables in $\mathbb{R}^d$, mutually independent and independent of ϑ_0.

The mixing properties of LMC have been largely investigated during the past decade among the statistical and machine learning communities, strongly motivated by learning methods such as Exponentially Weighted Aggregation introduced by [19], which involves sampling a non log-concave and heavy tailed posterior distribution. Then, several works derive some quantitative estimates in simple or sophisticated frameworks.

Strongly convex situations: A first paper of Dalalyan [15] establishes the cost of LMC to obtain an ε TV bound in terms of d and ρ when the target measure is ρ strongly log-concave and proposes a penalized version of LMC to circumvent the lack of strong log-concavity when the target distribution is only log-concave. Since this pioneering paper, a huge impressive literature expanded. Among others, we refer to [23] that gives a careful study of discretized LMC.

Convex situations: Other papers relax the strongly convex assumption using a modification of the numerical scheme (1.2): we refer to [18] for a kinetic version of LMC and [17] where the penalized LMC in non strongly-concave situation is studied in depth. Among all these papers, first, the lack of strong log-concavity is dealt with a modification of the initial LMC using a surrogate and asymptotically vanishing penalty. Second, these papers assume that a noiseless gradient of the log-posterior is available at each iteration of the algorithm, which may not be realistic, especially when $U = U_{\nu_n}$ with large n. Finally, [1] provides some upper-bounds on the mixing time of LMC in both *constrained* convex and strongly convex cases using some explicit coupling and projections.

Functional inequality: On the machine learning side, mixing of LMC has also received an impressive recent amount of interest: [50] proves the convergence in Kullback-Leibler divergence assuming that the posterior distribution, defined as e^{-U}, satisfies a log-Sobolev inequality and the Hessian of U is bounded. Later on, [5] develops a close analysis using the Fisher information distance and the Poincaré inequality. Metropolis-Adjusted Langevin Algorithm, which is close to LMC (with the addition of a Metropolis correction), was studied in [13] using the chi-squared divergence for the class of smooth and strongly convex potentials. More recently, in a weakly log-concave setting, [24] studied the convergence in Kullback-Leibler divergence of LMC using a weak version of log-Sobolev inequality and obtain a link between a tail assumption on e^{-U} and log-Sobolev inequality. As emphasized below, the assumptions they made are closely related to ours and we will provide further details later on. Using similar ideas, in

[12], they assume that the invariant measure satisfies a Latala-Oleszkiewicz or modified log-Sobolev inequality and guarantee convergence on the Rényi divergence for LMC.

• **Stochastic Gradient Langevin Dynamic (SGLD):** This last framework studies the behaviour of LMC when an additional noise term is incorporated in (1.2), which generates a perturbed discretization:

$$\vartheta_{k+1} = \vartheta_k - \eta\widehat{\nabla_\vartheta U}(\vartheta_k) + \sqrt{2\eta}\xi_{k+1}, \tag{1.3}$$

where $\widehat{\nabla_\vartheta U}(\vartheta_k)$ is an unbiased estimation of $\nabla_\vartheta U(\vartheta_k)$ at iteration k. For example, if $U = \frac{1}{n}\sum_{i=1}^n U_i$, then $\widehat{\nabla_\vartheta U}(\vartheta_k) = \frac{1}{B}\sum_{i=1}^B \nabla_\vartheta U_{I_i}(\vartheta_k)$ will appear in Equation 1.3, where $1 \leq B \leq n$ and $I_1, \ldots, I_B \overset{\text{i.i.d.}}{\sim} \text{Uniform}([n])$. The computation of the full gradient over the entire dataset is replaced by computing the gradient over a subset.

Stochastic Gradient Langevin Dynamics (SGLD below) has attracted the interest of several works: [53] introduced this method and described its efficiency from a numerical point of view in the particular case of Bayesian learning, which is exactly our framework. Some recent advances and related contributions may be also cited: in [46], the authors derive some non-asymptotic bounds (in terms of 2-Wasserstein distance) assuming a dissipative and smooth potential for optimization purpose.

The contribution of [55] is also related to our work: the authors develop a machinery for the study of SGLD essentially based on the Poincaré inequality but the way the lower bound on the spectral gap involved in the LMC is dealt with appears to be inappropriate. In particular, the diffusion involved in SGLD is used at a very low-temperature, proportional to $1/n$, which generates some important troubles in the size of the spectral gap in non strongly log-concave framework.

Under similar assumptions to [46], in [57], the authors prove the convergence of SGLD in total variation distance using an isoperimetric inequality linked to the Cheeger constant. On the other hand, in [52], the authors propose a Laplacian smoothing version of SGLD and prove the convergence of the algorithm in 2-Wasserstein distance when the potential is dissipative, smooth and satisfies a bound variance property. They study the convex and non-convex cases, however, the dissipative property together with the smoothness guarantee that the target measure verifies a log-Sobolev inequality.

The most recent works are probably [20, 54], where the convergence of SGLD is studied in Kullback-Leibler divergence and in total variation. In both situations, [20] assume smoothness of the potential, in addition to a log-Sobolev inequality and a 4th moment growth condition for the convergence in Kullback-Leibler divergence and Poincaré inequality and a 6th moment growth condition in total variation.

• **Role of functional inequalities and convexity:** From the previous discussions on the related literature on LD, LMC and SGLD, it appears that functional inequalities play a key role, even in the log-concave situation for the approximation of the measure e^{-U} (or $e^{-U_{\nu_n}}$). Coming back to [46], the authors identify the important dependency of the spectral gap denoted by λ_* in their paper with the temperature level $1/\beta$ they introduced. They obtain some very highly pessimistic bounds in some general situations (see their discussion in Section 4 of [46]), they conclude their discussion by the urgent need to find some non-trivial situations where some better lower bounds of λ_* may be derived.

Metastability and spectral gap asymptotic: Indeed, the final remark of Section 4 of [46] is related to the well known meta-stability phenomenon: at a low temperature, the mixing rates of a lot of reversible and irreversible Markov semi-groups are strongly deteriorated by the low temperature settings, which is implicitly induced by a Bayesian posterior sampling problem with a large number n of observations. In a regime of variance noise of the order $O(\beta^{-1})$, the first study of large deviation principle of invariant measures

traces back to [26] where the authors establish the asymptotic of the spectral gap of the over-damped Langevin diffusion as $\exp(-I\beta)$ (Chapter 6 of [26]) where I is an explicit constant that depends on the potential of the Gibbs field. This result has been extended in depth by [35], which leads to the first precise analyses of the so-called simulated annealing method (see for example [33, 43]). These works, and more recent contributions with irreversible dynamical systems in a stochastic settings ([30, 27]) show that there is almost nothing to expect in meta-stable situations in terms of asymptotic behaviour of the spectral gap, and indirectly in terms of mixing rate.
Weakly convex case and Kurdyka-Łojasiewicz inequality: Hence, the only situation that may lead to reasonable results is an intermediary situation between the (almost) trivial strongly log-concave case and the meta-stable multi-welled case. This is the purpose of the weakly log-concave situation that is described by the family of Kurdyka-Łojasiewicz inequalities [38, 40] used in optimization theory [7] that have shown to be efficient for stochastic optimization [28] or for sampling [29]. Below, we will intensively use this way to parameterize the "weak convexity" setting and will also relate this assumption to the recent one introduced in [24]. We also refer to the recent contributions [8] that derives some functional inequalities within an intermediary framework in which the curvature ρ is related to their keystone function α that controls the constants involved in the functional inequalities they are studying.

Sketch of our contributions: Taking together the statistical considerations and limitations, we are motivated in this paper in the study of a *continuous time* stochastic Gradient Langevin Dynamic. This process will be described precisely in the next paragraph as well as the Kurdyka-Łojasiewicz setup parameterized by a real value r, which varies between 0 (strongly convex case) and 1 (limiting Laplace asymptotic tail). We will show that the final horizon of simulation to obtain an ε approximation is of the order:

$$(d\log(n)^2)^{(1+r)^2}[\log^2(\varepsilon^{-1}) + n^2 d^{2(1+r)}\log^{4(1+r)}(n)]$$

with a Poissonian sub-sampling of parameter $(d\log^2(n))^{-(1+r)^2}$.

An important advantage of considering SGLD is that it is not necessary to have all the observations available at the initial instant, but only the total number of observations that will be used and modify the potential every exponential time.

Structure of the paper: The rest of the introduction consists of the definitions of the algorithm in Subsection 1.2, the way we assess the quality of our result with an entropy criterion in Subsection 1.3, as well as the quantitative weakly log-concave assumption in Subsection 1.4. We finally state our main result in Subsection 1.5 and provide two examples in Subsection 1.6.

In order to prove the main result, we first present in Section 2 the classical tools related to Markov semi-groups and we establish an inequality for the entropy that depends on the Dirichlet form. Section 3 is dedicated to the study of some functional inequalities that links the entropy and the Dirichlet form, allowing us, in Section 4, to establish a differential inequality for the entropy and proving the main result. Section 5 finally presents our technical results.

1.2 Continuous time evolution

Below, we briefly define the continuous time SGLD algorithm for Bayesian learning, for which a discretized form has been introduced in [53]. For this purpose, we consider a statistical model that is built with the help of a function $(x,\theta)\longmapsto p_\theta(x)$ where $\theta\in\mathbb{R}^d$ encodes the parameter of the statistical model and x the observation in a space denoted

by $\mathcal{X}$. We then assume that we have n i.i.d. observations denoted by $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ distributed according to p_{θ^*}, with $n \geq 2$. Given a prior distribution π_0 on $\mathbb{R}^d$, the posterior distribution μ_n is then defined as:

$$\mu_n(\theta) \propto \pi_0(\theta) \times \prod_{i=1}^{n} p_\theta(\mathbf{X}_i).$$

We introduce the log-parameterization that leads to the Gibbs form:

$$U_x(\theta) = -[\log \pi_0(\theta) + n \log p_\theta(x)],$$

and we then observe that:

$$\mu_n(\theta) \propto \exp\left(-\frac{1}{n}\sum_{i=1}^{n} U_{\mathbf{X}_i}(\theta)\right) = \exp\left(-U_{\nu_n}(\theta)\right),$$

where ν_n refers to the empirical distribution and U_{ν_n}, the average value of $U_X(\theta)$ when $X \sim \nu_n$:

$$\nu_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{X}_i}(x) \quad \text{and} \quad U_{\nu_n}(\theta) = \mathbb{E}_{X\sim\nu_n}[U_X(\theta)].$$

The standard Langevin Dynamics approach relies on the ergodic behaviour of the stochastic differential equation:

$$\mathrm{d}\theta_t = -\nabla U_{\nu_n}(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

that possesses, under some mild assumptions, a unique invariant distribution μ_n.

The SGLD algorithm takes benefit of both, sampling with a stochastic differential equation and homogenization of the drift that may be written as an expectation on X that is sampled uniformly over the set of observations according to ν_n. The leading idea is to replace the expectation in U_{ν_n} that depends on the overall set of observations $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ by a single unique observation that is randomized uniformly all along the evolution of the stochastic differential equation, and modified according to a Markov exponential clock. That being said, we can write an explicit formal definition of the algorithm as follows. We define $\left(\xi_j^{(n)}\right)_{j\geq 1}$ an infinite sequence of exponential random variables of mean α_n^{-1} that will be fixed later on and $\xi_0^{(n)} = 0$.

We also consider a sequence $\left(V_j^{(n)}\right)_{j\geq 1}$ of i.i.d. random variables uniformly distributed in $[n] = \{1, 2, \dots, n\}$. We then define the process $(X_t)_{t\geq 0}$ as a jump process that takes its values in $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ such that:

$$X_t = \mathbf{X}_{V_j^{(n)}}, \qquad \text{if } \sum_{k=0}^{j-1} \xi_k^{(n)} \leq t < \sum_{k=0}^{j} \xi_k^{(n)}, \quad j \geq 1.$$

Informally, $(X_t)_{t\geq 0}$ should be understood as follows: the process takes the value of one observation uniformly chosen from the n observations $\mathbf{X}_1, \dots, \mathbf{X}_n$ during exponential times with intensity α_n. The stochastic Langevin over-damped diffusion we consider is then given by the joint evolution $(\theta_t, X_t)_{t\geq 0}$ and that is defined by:

$$\mathrm{d}\theta_t = -\nabla_\theta U_{X_t}(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad t > 0, \tag{1.4}$$

where $(B_t)_{t\geq 0}$ is a multivariate standard Brownian Motion.

For the completeness, we present the continuous time SGLD algorithm in Algorithm 1, which is built using the Langevin over-damped diffusion (1.4).

Algorithm 1: Continuous time SGLD

Data: $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ i.i.d. observations distributed according to p_{θ^*}, n_0 initial distribution, π_0 prior distribution

$t_0 = 0$
Generate θ_0 according to n_0
for $k = 0, 1, \dots$ **do**
  Pick X_k uniformly in $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$
  Generate ξ_k according to an Exponential distribution with mean α_n^{-1}
  $t_{k+1} = t_k + \xi_k$
  $\theta_{t_{k+1}} = \theta_{t_k} - \int_{t_k}^{t_{k+1}} \nabla_\theta U_{X_k}(\theta_s)\mathrm{d}s + \sqrt{2}B_{\xi_k}$
end
return $\lim_{k\to\infty} \theta_{t_k}$

1.3 Entropic divergence

To assess the long-time behaviour of the continuous time SGLD, we introduce several notations related to the pair $(\theta_t, X_t)_{t\geq 0}$. Below, we denote by λ_d the Lebesgue measure over $\mathbb{R}^d$ and by ν_c the counting measure over $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$. The semi-group induced by $\mathcal{L}$, defined in Equation (2.1) and being elliptic on the θ coordinate, trivially irreducible and finitely supported on the x coordinate, makes the law of (θ_t, X_t) absolutely continuous with respect to the measure $\lambda_d \times \nu_c$ as soon as $t > 0$.

We introduce the notation of m_t to refer to the joint density of (θ_t, X_t) at time t with respect to $\lambda_d \times \nu_c$. In the meantime, n_t denotes the marginal distribution of θ_t and $m_t(\cdot|\theta)$ the conditional distribution of X_t given $\theta_t = \theta$. That is:

$$\mathrm{Law}(\theta_t, X_t) = m_t, \quad n_t(\theta) = \sum_{i=1}^{n} m_t(\theta, \mathbf{X}_i), \quad m_t(x|\theta) = \frac{m_t(\theta, x)}{n_t(\theta)},$$

for $\theta \in \mathbb{R}^d$ and $x \in \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$.

To show that the algorithm recovers the correct asymptotic behaviour, *i.e.* that $n_t(\theta) \longrightarrow \mu_n$ when $t \longrightarrow \infty$, we consider the relative entropy (or Kullback-Leibler divergence) of n_t with respect to μ_n that is well defined thanks to the ellipticity, and given by:

$$J_t = Ent_{\mu_n}\left(\frac{n_t}{\mu_n}\right) = \int_{\mathbb{R}^d} \log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right) \mathrm{d}n_t(\theta).$$

J_t measures at any time $t > 0$ a divergence between the instantaneous law of the process at time t and the (presumably) invariant distribution μ_n of the process $(\theta_t, X_t)_{t\geq 0}$. It would also be possible to measure this difference between the two distributions in terms of the $\mathbb{L}^2$ or the χ-square distance and to produce a theoretical analysis with the help of functional analysis but it would rely on stronger assumptions on the function U_{ν_n}.

1.4 Main assumptions

1.4.1 Weak convexity:

We will study a continuous time version of SGLD into a weakly convex framework, *i.e.* when U_{ν_n} is assumed to be convex but not necessarily strongly convex. SGLD has recently received an important interest in the machine learning community and has been studied in various situations where functional inequalities are involved. We refer to [55, 20] (uniform Poincaré inequality) and to [46, 52, 20] (Log-Sobolev inequality), where the functional inequalities play a crucial role to analyze the behavior of the process. In

[46, 55, 52], the authors develop a 2-Wasserstein analysis of the algorithm assuming restrictive assumptions like dissipativity. On the other hand, in [20], they study the convergence in Kullback-Leibler divergence and in total variation. In both analysis, they assume smoothness of the potential, in addition to a Log-Sobolev inequality and a 4th moment growth condition for the convergence in Kullback-Leibler divergence and Poincaré inequality and a 6th moment growth condition in total variation.

Importantly, Poincaré or Log-Sobolev inequalities are not so innocent since they generally require convexity (see e.g. [6, 4]) to be reasonably dimension-dependent, and even strong convexity to be dimension free. Otherwise, the constant involved in these functional inequalities are exponentially degraded by the "temperature" (n^{-1} in our case) and the dimension (d for us) as indicated in [35].

In our work, we have chosen to parameterize this lack of strong convexity with the help of the Kurdyka-Łojasiewicz inequality [38, 40], which is a standard tool in optimization to describe the transition between convexity and strong convexity and makes the bounds more explicit. This assumption allows to observe how the entropy evolves according to the key exponent involved in the KL inequality. In particular, it makes possible to understand the influence of the lack of strong convexity that is more or less hidden in the uniform Poincaré or Log-Sobolev inequalities. We introduce a parametric form of the KL inequalities following [28].

For this purpose, we denote by $\mathcal{C}^2$ the set of twice continuously differentiable functions and for any $\mathcal{C}^2$-function V, we denote the spectrum of the Hessian matrix of V as $Sp(\nabla^2 V(\theta))$. Furthermore, if V is convex, we denote

$$\underline{\lambda}_{\nabla^2 V}(\theta) = \inf Sp(\nabla^2 V(\theta)), \quad \theta \in \mathbb{R}^d.$$

Hypothesis 1.1 ($\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$)**.** *We say that a function $V : \mathbb{R}^d \to \mathbb{R}$ satisfies a $\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$-condition if:*

a) V is a $\mathcal{C}^2$-function,
b) V is convex and $\min_{\theta\in\mathbb{R}^d} V(\theta) = V(\theta^*) > 0$,
c) ∇V is L-Lipschitz and
d) there exist some constants $0 \le r < 1$ and $\mathfrak{c} > 0$ such that

$$\mathfrak{c} V^{-r}(\theta) \le \underline{\lambda}_{\nabla^2 V}(\theta), \quad \forall \theta \in \mathbb{R}^d. \tag{1.5}$$

Let us briefly comment this assumption.

- In [29], a slightly different parameterization is used with the introduction of another exponent $0 \le q \le r$ related to $\overline{\lambda}_{\nabla^2 V}(\theta) = \sup Sp(\nabla^2 V(\theta))$, $\theta \in \mathbb{R}^d$. The authors also assume the upper bound $\overline{\lambda}_{\nabla^2 V}(\theta) \le \tilde{\mathfrak{c}} V^{-q}(\theta)$. When $r = q$, they recover a global standard KL inequality (see [28, 7]). Here, we have chosen to simplify this assumption and use a rough upper bound on the eigenvalues of the Hessian matrix given by the Lipschitz constant L, *i.e.* in the last inequality we simply use $\tilde{\mathfrak{c}} = L$ and $q = 0$.
- The case $r = 0$ is of course associated to the strongly convex situation where the curvature of the function is uniformly lower bounded by $\mathfrak{c}$, and the case $r = 1$ would correspond to the limiting Laplace situation.
- We shall observe that if $V(\theta) = (1 + \|\theta\|_2^2)^p$ with $p \in [1/2, 1]$, then V satisfies a $\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$-condition with $r = \frac{1-p}{p}$ and $\mathfrak{c} = 2p(1 - 2(1-p))$, see Remark 7 of [29] for further details. In particular, the larger p, the smaller r, which translates into a better curvature of the potential function. It is expected that the complexity of the algorithm increases with the lack of curvature, *i.e.* is an increasing function of r.

- The $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ assumption in our work is tightly related to that one of [24]. They assume that the potential function U is degenerately convex at infinity, which means that there exists a function $\widetilde{U}$ such that for a constant $\epsilon \geq 0$, $\|U - \widetilde{U}\|_\infty \leq \epsilon$ where $\underline{\lambda}_{\nabla^2 \widetilde{U}}(\theta) \geq \kappa(1 + \frac{1}{4}\|\theta\|_2^2)^{-\tau/2}$ for some $\kappa > 0$ and $\tau \geq 0$. This parameterization is similar to the one considered in (1.5), however, we bound the eigenvalues by a power of the same function U, while they use a power of $\|\theta\|_2^2$. The example above which satisfies a $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$-condition with $r = \frac{1-p}{p}$ is also degenerated at infinity taking $\widetilde{U}(\theta) = U(\theta) = (1 + \|\theta\|^2)^p$ and $\tau = 2(1-p) = \frac{2r}{1+r}$.

  Below, we will establish that an important consequence of $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ is that the invariant distribution verifies a *weak* log-Sobolev inequality, that will be useful to assess the ergodic behaviour of our procedure. In [24], even though not exactly equivalent, the strategy is rather close and the authors establish a *modified* log-Sobolev inequality with the help of a perturbation strategy (see in particular Appendix A of [24]).

In Section 5.1 we recall some important consequences of the KL inequality obtained in Lemma 15 of [29]. In particular, the growth of any function that satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ is lower and upper bounded by a positive power of the distance to its minimizer.

Remark 1.2. If inequality (1.5) holds for a constant $\mathfrak{c}$, then it holds for all positive values less than $\mathfrak{c}$. For that reason, we assume that $\mathfrak{c} \leq (8L/(1+r))^{1+r}$, which will be used in Proposition 5.10.

1.4.2 Others assumptions

We state below an important consequence of a "population" satisfying the $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ assumption, but before, let us state some mild assumptions on π_0.

Hypothesis 1.3 ($\mathcal{H}_{\pi_0}(\ell_0)$)**.** *$\pi_0$ is a log-concave $\mathcal{C}^2$-function such that $\min_{\theta \in \mathbb{R}^d} -\log \pi_0(\theta) > 0$ and $\theta \mapsto \nabla \log \pi_0(\theta)$ is ℓ_0-Lipschitz.*

Since the prior distribution is chosen by the user, our $\mathcal{H}_{\pi_0}(\ell_0)$ hypothesis is not restrictive and some typical examples satisfy these conditions, such as Gaussian, Weibull and Gamma, both with shape parameter larger than 1, Gumbel, among others.

The following proposition shows the consequence on the function U_{ν_n} of assuming the hypothesis $\mathcal{H}_{\pi_0}(\ell_0)$ and $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ on $\theta \mapsto -\log p_\theta(x)$, for any x. The proof of Proposition 1.4 may be found in Section 5.2.

Proposition 1.4. *We assume $\mathcal{H}_{\pi_0}(\ell_0)$ and that there exist $(\mathfrak{c}, r)$ such that for any x: $\theta \longmapsto -\log p_\theta(x)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$. Then U_{ν_n} satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c} n^{1+r}, nL + \ell_0)$, and in particular, for any $\mathbf{X}_i$, $U_{\mathbf{X}_i}$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c} n^{1+r}, nL + \ell_0)$.*

We introduce the notation $a \lesssim_{uc} b$ ($a \gtrsim_{uc} b$) which means $a \leq cb$ ($a \geq cb$) where c is a universal constant *i.e.* a positive constant independent of n and d.

We assume that the minimizers and the minimums of the functions $U_{\mathbf{X}_i}$ are contained in balls of radius depending of n and d.

Hypothesis 1.5 ($\mathcal{H}_{\min}$)**.** *There exists a constant $\beta \geq 0$ such that*

$$\max_i \|\arg\min U_{\mathbf{X}_i}\|_2 \lesssim_{uc} \sqrt{d}\log^\beta(n) \quad \textit{and} \quad \max_i \min_{\theta \in \mathbb{R}^d} U_{\mathbf{X}_i}(\theta) \lesssim_{uc} d\log^{2\beta}(n).$$

Assumption $\mathcal{H}_{\min}$ is not restrictive. In dimension $d = 1$, consider a concentrated i.i.d. sample $(X_i)_{1 \leq i \leq n}$ with a suitable sub-Gaussian like behaviour for which the Laplace transform of $|\arg\min U_{X_i}|$, for any i, is upper bounded as:

$$\mathbb{E}[\exp(\lambda |\arg\min U_{X_i}|)] \leq \exp(\sigma^2 \lambda^k), \quad \forall \lambda > 0,$$

where $k \geq 1$ and $\sigma > 0$ are fixed constants. The Markov inequality and the upper bound above imply that, in this case for any $\lambda > 0$ and $s > 0$:

$$\begin{aligned}\mathbb{P}\left(\max_i |\arg\min U_{X_i}|) > s\right) &\leq e^{-s\lambda}\mathbb{E}\left(\max_i \exp(\lambda|\arg\min U_{X_i}|)\right)\\ &\leq e^{-s\lambda}\mathbb{E}\left(\sum_{i=1}^n \exp(\lambda|\arg\min U_{X_i}|)\right)\\ &\leq n\exp(-s\lambda+\sigma^2\lambda^k).\end{aligned}$$

If we choose $\lambda = \sigma^{-\frac{2}{k}}\log^{\frac{1}{k}}(n/\delta)$ and $s = 2\sigma^{\frac{2}{k}}\log^{\frac{k-1}{k}}(n/\delta)$ where δ is a small positive constant, then:

$$\mathbb{P}\left(\max_i |\arg\min U_{\mathbf{X}_i}| > 2\sigma^{\frac{2}{k}}\log^{\frac{k-1}{k}}(n/\delta)\right) \leq \delta.$$

Therefore the value of β involved in $\mathcal{H}_{\min}$ is given by $\beta = \frac{k-1}{k}$. We recover in particular the situation where $\beta = 1/2$ when $k = 2$. For larger dimensions, the result may be extended using that $\|x\|_2^2 \leq d\,\max_{1\leq j\leq d}(x^j)^2$, where x^j is the j-th component of x. We should keep in mind from this last discussion that even if $\mathcal{H}_{\min}$ is stated (and makes sense) for any value of $\beta \geq 0$, it holds in general for $0 \leq \beta \leq 1$. If we replace $\|\arg\min U_{\mathbf{X}_i}\|_2^2$ by $\min U_{\mathbf{X}_i}$ in the previous example, we obtain an analogous behavior of $\max_i \min U_{\mathbf{X}_i}$.

Hypothesis $\mathcal{H}_{\min}$ and $\mathcal{H}_{\pi_0}(\ell_0)$ lead to an almost similar behaviour of the minimizer and the minimum of U_{ν_n}. Details appear in Proposition 5.3.

1.5 Long-time entropy convergence

We introduce for any time $t \geq 0$, the density function of θ_t with respect to μ_n, which is given by:

$$h_t(\theta) = \frac{n_t(\theta)}{\mu_n(\theta)}.$$

We will choose the initial distribution n_0 such that h_0 is bounded, then h_t will also be bounded, for any $t > 0$. This property plays an important role in proving the main theorem of the paper and it will be studied in Section 3. Let us denote as 0_d and I_d the null vector of $\mathbb{R}^d$ and the identity matrix of size d respectively .

Hypothesis 1.6 ($\mathcal{H}_{\mathbf{n_0}}(L,\ell_0)$)**.** *A positive constant σ^2 exists such that n_0 is the density function of a $\mathcal{N}(0_d,\sigma^2 I_d)$ random variable. Moreover, there exist two universal constants c_1 and c_2 such that $0 < c_1 \leq c_2 < 1$ and:*

$$\frac{c_1}{nL+\ell_0} \leq \sigma^2 \leq \frac{c_2}{nL+\ell_0}.$$

This hypothesis guarantees the boundedness of h_0 and the initial entropy:

$$J_0 = \int_{\mathbb{R}^d} \log\left(h_0(\theta)\right) \mathrm{d}n_0(\theta) \lesssim_{uc} n(d\log^{2\beta}(n))^{1+r} + rd\log\left(d/n\right).$$

We refer to Proposition 3.9 in Section 3 for further details.

The next result assesses a mixing property in terms of decrease of the entropy and therefore states the convergence of n_t towards the correct measure μ_n.

Theorem 1.7. *Assume $\mathcal{H}_{\pi_0}(\ell_0)$, $\mathcal{H}_{\min}$, $\mathcal{H}_{\mathbf{n_0}}(L,\ell_0)$ and that each $\theta \mapsto -\log p_\theta(\mathbf{X}_i)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c},L)$. Then:*

1. *If $r = 0$, then μ_n satisfies a log-Sobolev inequality with constant $C_{\mathrm{LSI}}(\mu_n) = \frac{2}{\mathfrak{c}n}$. Moreover, for any $\varepsilon > 0$, if $\alpha_n = n$ and*

$$t \gtrsim_{uc} n^{-1}\left[\log(\varepsilon^{-1}) + \log(d) + \log(n)\right],$$

then $J_t \leq \varepsilon$.

2. *If $0 < r < 1$, then μ_n satisfies a Poincaré inequality with constant $C_{\mathrm{PI}}(\mu_n)$, indistinctly denoted as C_{PI}. Let us define $c_{n,d} = n^4 \left(d \log^{2\beta}(n)\right)^{1+r}$ and $O_{n,d} = \left(\frac{C_1 d}{n}\right)^{\frac{dr}{2}} \exp\left(C_2 n \left(d \log^{2\beta}(n)\right)^{1+r}\right)$, where C_1 and C_2 are positive universal constants. For any $t > 0$,*

$$J_t \lesssim_{uc} \left(J_0 + O_{n,d} + c_{n,d} C_{\mathrm{PI}}\right) e^{-\sqrt{\frac{3t}{16 \mathfrak{a} C_{\mathrm{PI}}}}},$$

where $\mathfrak{a}$ is a universal constant. For any $\varepsilon > 0$, if $\alpha_n = \left(d \log^{2\beta}(n)\right)^{-(1+r)^2}$, then $J_t \leq \varepsilon$ if

$$t \gtrsim_{uc} \left(d \log^{2\beta}(n)\right)^{(1+r)^2} \left[\log^2(\varepsilon^{-1}) + n^2 \left(d \log^{2\beta}(n)\right)^{2(1+r)}\right].$$

If we denote t_ε the smallest value such that $J_{t_\varepsilon} \leq \varepsilon$, then the choice of α_n guarantees that the mean number of jumps $\alpha_n t_\varepsilon$ of the process $(X_t)_{0 \leq t \leq t_\varepsilon}$ is the minimum possible.

It is important to point out that in the weakly log concave case, the bound used for the Poincaré constant C_{PI} adds noise in terms of the dimension d and possibly in n. If the constant C_{PI} were explicitly calculable, then $J_t \leq \varepsilon$ if $\alpha_n = \frac{1}{C_{\mathrm{PI}}}$ and:

$$t \gtrsim_{uc} C_{\mathrm{PI}} \left[\log^2(\varepsilon^{-1}) + n^2 \left(d \log^{2\beta}(n)\right)^{2(1+r)}\right].$$

We observe that the choice of α_n is similar in both cases, ignoring proportional constants. In terms of ε, moving from the strongly log-concave case to the weakly log-concave case requires squaring the dependence on ε. Additionally, the dependence on n and d deteriorates as well.

1.6 Examples

We present two examples. First, we apply the convergence result stated in Theorem 1.7 in a synthetic situation and we compare the result with the one that would be obtained for the LD algorithm. As a second example, we study an application to the Bayesian logistic regression.

Synthetic example. Consider $0 < r < 1$ and the potential function $U^{(r)}(\theta, x) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ as:

$$U^{(r)}(\theta, x) = (1 + x^2)(1 + \|\theta\|_2^2)^{\frac{1}{1+r}}.$$

We then observe $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. random variables in $\mathbb{R}$, and we define:

$$U^{(r)}_{\nu_n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} U^{(r)}(\theta, \mathbf{X}_i)$$

which verifies a $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c} n^{1+r}, L)$-condition with constants $\mathfrak{c} = \frac{2(1-r)}{(1+r)^2}(1 + \min \mathbf{X}_i^2)^{1+r}$ and $L = \frac{2}{1+r}(1 + \max \mathbf{X}_i^2)$. This choice of potential leads to the weakly log-concave density function:

$$\mu_n^{(r)}(\theta) = \frac{e^{-U^{(r)}_{\nu_n}(\theta)}}{\mathcal{Z}_n}.$$

We consider two stochastic processes to sample the distribution $\mu_n^{(r)}$. Let $\left(\theta_t^{(1)}\right)_{t \geq 0}$ be the solution of the over-damped Langevin Dynamics:

$$\mathrm{d}\theta_t^{(1)} = -\nabla_\theta U^{(r)}_{\nu_n}(\theta_t^{(1)})\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad t > 0,$$

while $\left(\theta_t^{(2)}\right)_{t\geq 0}$ is the solution of the continuous time SGLD defined by:

$$\mathrm{d}\theta_t^{(2)} = -\nabla_\theta U_{X_t}^{(r)}(\theta_t^{(2)})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad t > 0,$$

where $(X_t)_{t\geq 0}$ is taking the value of one observation uniformly chosen from the n observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ during exponential times with intensity α_n. In both cases $(W_t)_{t\geq 0}$ and $(B_t)_{t\geq 0}$ are d-dimensional standard Brownian Motions and we assume that $\theta_0^{(1)}$ and $\theta_0^{(2)}$ are centered Gaussian random variables with a tuned covariance matrix as is specified in hypothesis $\mathcal{H}_{\mathbf{n_0}}(L, \ell_0)$.

To find the order of convergence (in terms of entropy) for $\theta_t^{(1)}$, we proceed in a similar way as is done in the proof of Theorem 1.7. Let $\widetilde{n}_t$ be the law of $\theta_t^{(1)}$, we denote by $\widetilde{J}_t$ and $\widetilde{\mathcal{E}}_t$ the entropy and the Dirichlet form of $\widetilde{n}_t$ with respect to μ_n respectively. Then $\widetilde{J}_t$ satisfies the equality:

$$\partial_t \left\{\widetilde{J}_t\right\} = -\widetilde{\mathcal{E}}_t.$$

We could use weak log-Sobolev inequality and prove that a sufficient condition to obtain $\widetilde{J}_t \leq \varepsilon$ is:

$$t \gtrsim_{uc} \left(d\log^{2\beta}(n)\right)^{(1+r)^2} \left[\log^2(\varepsilon^{-1}) + n^2\left(d\log^{2\beta}(n)\right)^{2(1+r)}\right]. \tag{1.6}$$

This is indeed the same result that we will obtain for $\theta_t^{(2)}$. Theorem 1.7, guarantees that if $\alpha_n = \left(d\log^{2\beta}(n)\right)^{-(1+r)^2}$ then the time to obtain an ε-error is given by (1.6). Although the results coincide in order of convergence, it does not mean that they have the same proportionality constants.

Formally, we can conclude that both processes would have the same order of convergence, however, the continuous time SGLD process does not need to compute the average of the functions $U^{(r)}(\theta, \mathbf{X}_i)$, only change the observation $\mathbf{X}_i$ every exponential clock.

Bayesian logistic regression. This example is inspired by the Bayesian logistic regression problem studied in [22] (see also [34, 31, 45]).

Consider $n \geq 1$ i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_1, \ldots, X_n$ are d-dimensional input variables and $Y_1, \ldots, Y_n$ are binary output responses. The output responses are distributed as Bernoulli random variables such that:

$$Y_i \sim \text{Ber}\left(\phi(\theta^\top X_i)\right), \quad i \in \{1, \ldots, n\},$$

where ϕ is the logit function defined by $\phi(x) = (1 + e^{-x})^{-1}$, $x \in \mathbb{R}$ and $\theta \in \mathbb{R}^d$ is the parameter of interest. We consider a prior distribution given by the following density with respect to Lebesgue measure on $\mathbb{R}^d$:

$$\pi_0^{(r)}(\theta) \propto \exp\left\{-a\left(1 + \|\theta\|_2^2\right)^{\frac{1}{1+r}}\right\}, \quad \theta \in \mathbb{R}^d,$$

where a is a positive constant and the parameter $r \in (0, 1)$ is related to weakly log-concavity. The log-concave posterior distribution of θ is given by the density:

$$\mu_n^{(r)}\left(\theta|(X_1, Y_1), \ldots, (X_n, Y_n)\right) \propto \exp\left\{-\sum_{i=1}^n \ell_i(\theta) - a\left(1 + \|\theta\|_2^2\right)^{\frac{1}{1+r}}\right\},$$

where the log-likelihoods are $\ell_i(\theta) = \log\left(1 + \exp\left((1 - 2Y_i)\theta^\top X_i\right)\right)$. We now introduce the potential:

$$U_{\nu_n}^{(r)}(\theta) = \sum_{i=1}^n \ell_i(\theta) + a\left(1 + \|\theta\|_2^2\right)^{\frac{1}{1+r}},$$

which verifies a $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, nL)$-condition with parameters $\mathfrak{c} = \frac{2(1-r)}{(1+r)^2}a^{1+r}$ and $L = \frac{2}{1+r}\max_{i,j} X_{ij}^2$, where X_{ij} represents the j-th component of X_i.

The proof of the main result is based on the fact that U_{ν_n} satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL+\ell_0)$, so if now $U^{(r)}_{\nu_n}$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, nL)$, we obtain that the bound of the Poincaré constant is worse:

$$C_{\mathrm{PI}} \lesssim_{uc} n^{(1+r)(2+r)} \left(d\log^{2\beta}(n)\right)^{(1+r)^2},$$

and as a consequence we obtain an ε-error if $\alpha_n = n^{-(1+r)(2+r)} \left(d\log^{2\beta}(n)\right)^{-(1+r)^2}$ and

$$t \gtrsim_{uc} n^{(1+r)(2+r)} \left(d\log^{2\beta}(n)\right)^{(1+r)^2} \left[\log^2(\varepsilon^{-1}) + n^2 \left(d\log^{2\beta}(n)\right)^{2(1+r)}\right].$$

Once again, we recall that in the weakly log-concave case the bound of the Poincaré constant adds an important dependence in terms of n and d. One way to improve this result would be to consider, for example $a = n$ in the definition of $\pi_0^{(r)}$, so we would obtain the original result of Theorem 1.7.

2 Markov tools and evolution of the entropy J_t

It is straightforward to verify that the joint evolution of $(\theta_t, X_t)_{t\geq 0}$ exists and is weakly unique (in law) with the help of the Martingale Problem (MP below). For this purpose, we preliminary define the operator $\mathcal{L}$ that acts on any function $f \in \mathcal{C}^2(\mathbb{R}^d \times \mathcal{X})$ as:

$$\mathcal{L}f(\theta, x) = \underbrace{-\langle \nabla_\theta U_x(\theta), \nabla_\theta f(\theta, x)\rangle + \Delta_\theta f(\theta, x)}_{\mathcal{L}_1 f(\theta,x)} + \underbrace{\frac{\alpha_n}{n}\sum_{i=1}^{n}[f(\theta, \mathbf{X}_i) - f(\theta, x)]}_{\mathcal{L}_2 f(\theta,x)}, \tag{2.1}$$

for all $(\theta, x) \in \mathbb{R}^d \times \mathcal{X}$.

The operator $\mathcal{L}$ is divided into two terms, $\mathcal{L}_1$ acts on the component θ and is associated to the diffusion part, while $\mathcal{L}_2$ is the jump operator that acts on the x component. Thanks to the finite number of observations $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, we can apply the results of Sections 4 and 5 of chapter 4 of [25] and deduce the following result:

Proposition 2.1. *Assume that for any $x \in \mathcal{X}$, U_x is $\mathcal{C}^2(\mathbb{R}^d)$ and $\nabla_\theta U_x$ is L_x-Lipschitz, then for any initial distribution ν on $\mathbb{R}^d \times \mathcal{X}$, the martingale problem $(\mathcal{L}, \nu)$ is well-posed. The associated (weakly) unique process $(\theta_t, X_t)_{t\geq 0}$ is a Feller Markov process associated to the semi-group $\mathcal{L}$. In particular, the θ component verifies the S.D.E.* (1.4).

If we denote by $\mathcal{L}^\star$ the adjoint operator of $\mathcal{L}$ in $\mathbb{L}^2(\mathbb{R}^d \times \mathcal{X})$, the Kolmogorov forward equation (or Fokker-Plank equation) yields:

$$\partial_t\{m_t(\theta, x)\} = \mathcal{L}^\star m_t(\theta, x), \tag{2.2}$$

in the weak sense. However, using Remark 3.19 of [39], we could verify that the first and second derivatives of m_t with respect to θ are bounded, therefore, m_t is differentiable in time and Equation (2.2) is satisfied in the strong sense.

We now introduce two operators, which will be the keystone of our work. The first one describes the infinitesimal action on the θ coordinate under the average effect of X_t at time t that applies for any function $f \in \mathcal{C}^2(\mathbb{R}^d)$ as:

$$\mathcal{G}_t f(\theta) = -\sum_{i=1}^{n}\langle \nabla_\theta f(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle m_t(\mathbf{X}_i|\theta) + \Delta_\theta f(\theta). \tag{2.3}$$

The second one is very close to the first one except that the average effect of X_t is replaced by the targeted ideal distribution ν_n. It leads to the definition: for any function $f \in \mathcal{C}^2(\mathbb{R}^d)$,

$$\begin{aligned}\mathcal{G}f(\theta) &= -\sum_{i=1}^{n}\langle \nabla_\theta f(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle \nu_n(\mathbf{X}_i) + \Delta_\theta f(\theta) \\ &= -\langle \nabla_\theta f(\theta), \nabla_\theta U_{\nu_n}(\theta)\rangle + \Delta_\theta f(\theta). \end{aligned} \tag{2.4}$$

2.1 Study of J_0 and $\partial_t\{J_t\}$

Our starting point is to establish a differential inequality for J_t. But first, we state the following proposition in which it is proved that the initial entropy J_0 is bounded.

Proposition 2.2. *Assume $\mathcal{H}_{\min}$,$\mathcal{H}_{\mathbf{n_0}}(L,\ell_0)$, $\mathcal{H}_{\pi_\mathbf{0}}(\ell_0)$ and that, for any x, $\theta \mapsto -\log p_\theta(x)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$, then:*

i) Two positive constants C_1 and C_2 exist, which are independent from n and d and such that:

$$\|h_0\|_\infty \lesssim_{uc} (C_1 d/n)^{\frac{dr}{2}} \exp\left(C_2 n (d\log^{2\beta}(n))^{1+r}\right).$$

ii) As a consequence:

$$J_0 = \int_{\mathbb{R}^d} \log\left(h_0(\theta)\right) \mathrm{d}n_0(\theta) \lesssim_{uc} n(d\log^{2\beta}(n))^{1+r} + rd\log\left(d/n\right).$$

The proof of Proposition 2.2 may be found in Section 5.3.

Let us now define the Dirichlet form of $\sqrt{n_t(\theta)/\mu_n(\theta)}$ (proportional to the relative Fisher information of n_t with respect to μ_n) as:

$$\mathcal{E}_t = \int_{\mathbb{R}^d} \left\| \nabla\left(\sqrt{\frac{n_t(\theta)}{\mu_n(\theta)}}\right)\right\|_2^2 \mathrm{d}\mu_n(\theta) = \frac{1}{4}\int_{\mathbb{R}^d} \left\|\nabla \log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\right\|_2^2 \mathrm{d}n_t(\theta).$$

The following proposition shows the link between $\partial_t\{J_t\}$ and $\mathcal{E}_t$.

Proposition 2.3. *Assume $\mathcal{H}_{\min}$, $\mathcal{H}_{\pi_\mathbf{0}}(\ell_0)$ and for each $\mathbf{X}_i$, $\theta \to -\log p_\theta(\mathbf{X}_i)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$. Then, for any $t > 0$,*

$$\partial_t\{J_t\} \le -3\mathcal{E}_t + c_{n,d} e^{-\frac{2\alpha_n}{3}t},$$

where $c_{n,d} \lesssim_{uc} n^4\left(d\log^{2\beta}(n)\right)^{1+r}$.

Proof of Proposition 2.3. The existence of $\partial_t\{J_t\}$ and the equalities:

$$\partial_t\{J_t\} = \int_{\mathbb{R}^d}\left(1+\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\right)\partial_t\{n_t(\theta)\}\mathrm{d}\theta = \int_{\mathbb{R}^d}\mathcal{G}_t\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\mathrm{d}n_t(\theta).$$

are justified in Section 5.4. Then, we are led to split $\partial_t\{J_t\}$ into two terms :

$$\begin{aligned}\partial_t\{J_t\} &= \int_{\mathbb{R}^d}\mathcal{G}_t\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\mathrm{d}n_t(\theta), \\ &= \underbrace{\int_{\mathbb{R}^d}\mathcal{G}\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\mathrm{d}n_t(\theta)}_{J_{1,t}} + \underbrace{\int_{\mathbb{R}^d}(\mathcal{G}_t-\mathcal{G})\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\mathrm{d}n_t(\theta)}_{J_{2,t}}. \end{aligned} \tag{2.5}$$

We study each term separately.

- Study of $J_{1,t}$. Since $\mathcal{G}$ is a diffusion operator and μ_n is the invariant measure associated to $\mathcal{G}$, then we can use the classical link between $J_{1,t}$ and the Dirichlet form $\mathcal{E}_t$ (see [4]):

$$\int_{\mathbb{R}^d} \mathcal{G}\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\mathrm{d}n_t(\theta) = \int_{\mathbb{R}^d}\frac{n_t(\theta)}{\mu_n(\theta)}\,\mathcal{G}\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\mathrm{d}\mu_n(\theta) = -4\mathcal{E}_t. \tag{2.6}$$

- Study of $J_{2,t}$. We use the difference between $\mathcal{G}$ and $\mathcal{G}_t$, for any twice differentiable function f:

$$\begin{aligned}(\mathcal{G}_t-\mathcal{G})\,f(\theta) &= -\sum_{i=1}^n \langle\nabla_\theta f(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle\left[m_t(\mathbf{X}_i|\theta)-\nu_n(\mathbf{X}_i)\right]\\ &= -\frac{1}{n}\sum_{i=1}^n\langle\nabla_\theta f(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right],\end{aligned}$$

where we used that $\nu_n(\mathbf{X}_i)=\frac{1}{n}$, for any i. Then, the term $J_{2,t}$ may be computed as:

$$\begin{aligned}|J_{2,t}| &= \left|\int_{\mathbb{R}^d}(\mathcal{G}_t-\mathcal{G})\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\mathrm{d}n_t(\theta)\right|\\ &= \frac{1}{n}\left|\int_{\mathbb{R}^d}\sum_{i=1}^n\langle\nabla_\theta\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]\mathrm{d}n_t(\theta)\right|.\end{aligned}$$

Using the Cauchy-Schwarz inequality with respect to the measure $\nu_n(\mathbf{X}_i)\times\mathrm{d}n_t(\theta)$ in the first line and $2ab\le a^2+b^2$, for any a and b, in the second line, we obtain that:

$$\begin{aligned}|J_{2,t}| &\le 2\mathcal{E}_t^{\frac{1}{2}}\left(\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i=1}^n\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2^2\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\mathrm{d}n_t(\theta)\right)^{\frac{1}{2}}\\ &\le \mathcal{E}_t+\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i=1}^n\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2^2\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\mathrm{d}n_t(\theta).\end{aligned}$$

Replacing Equation (2.6) and the inequality above in Equation (2.5), we get:

$$\partial_t\{J_t\} \le -3\mathcal{E}_t+\underbrace{\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i=1}^n\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2^2\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\mathrm{d}n_t(\theta)}_{\Delta_t}. \tag{2.7}$$

We then focus on the second term of the right hand side. For this purpose, we consider a strictly positive function $g(t)$, which will be fixed later and we split Δ_t into two terms as:

$$\begin{aligned}\Delta_t =& \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i=1}^n\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2^2\,\mathbb{1}_{\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2\le g(t)}\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\mathrm{d}n_t(\theta)\\ &+\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i=1}^n\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2^2\,\mathbb{1}_{\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2> g(t)}\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\mathrm{d}n_t(\theta)\\ =&\Delta_{1,t}+\Delta_{2,t}.\end{aligned} \tag{2.8}$$

We study each term separately.

- Study of $\Delta_{1,t}$. We introduce a weighted $\mathbb{L}^2$ distance between the conditional distribution of X_t given $\theta_t=\theta$ and the measure ν_n. This distance is denoted by I_t and is defined as:

$$I_t = \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i=1}^n\left(\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right)^2\mathrm{d}n_t(\theta). \tag{2.9}$$

This quantity measures the average closeness (with respect to θ) of the conditional law of x given θ at time t to ν_n. Using the definition of I_t, the first term of (2.8) is bounded by:

$$\Delta_{1,t} \leq g^2(t) I_t.$$

In Section 5.5, Proposition 5.7, we show that $I_t \leq (n-1)e^{-2\alpha_n t}$. Then, we get that:

$$\Delta_{1,t} \leq n g^2(t) e^{-2\alpha_n t}.$$

- Study of $\Delta_{2,t}$. We recall that for any i, $\nu_n(X_i) = 1/n$ and $m_t(\mathbf{X}_i|\theta) \in [0,1]$, then $\left|\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)} - 1\right| \leq n$, which implies that:

$$\Delta_{2,t} \leq n \int_{\mathbb{R}^d} \sum_{i=1}^n \|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2^2 \mathbb{1}_{\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2 > g(t)} \mathrm{d}n_t(\theta).$$

The Cauchy-Schwarz inequality leads to:

$$\begin{aligned}\Delta_{2,t} &\leq n \left(\int_{\mathbb{R}^d} \sum_{i=1}^n \|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2^4 \, \mathrm{d}n_t(\theta) \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \sum_{i=1}^n \mathbb{1}_{\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2 > g(t)} \mathrm{d}n_t(\theta) \right)^{\frac{1}{2}} \\ &= n \left(\sum_{i=1}^n \mathbb{E}_{n_t} \left(\|\nabla_\theta U_{\mathbf{X}_i}(\theta_t)\|_2^4 \right) \right)^{\frac{1}{2}} \left(\sum_{i=1}^n \mathbb{P}_{n_t} \left(\|\nabla_\theta U_{\mathbf{X}_i}(\theta_t)\|_2 > g(t) \right) \right)^{\frac{1}{2}}. \quad (2.10)\end{aligned}$$

Thanks to Propositions 1.4 and 5.1, we deduce that for any $\theta \in \mathbb{R}^d$:

$$\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2^2 \leq 2(nL + \ell_0) U_{\mathbf{X}_i}(\theta),$$

then:

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{n_t} \left(\|\nabla_\theta U_{\mathbf{X}_i}(\theta_t)\|_2^4 \right) &\leq \frac{4(nL+\ell_0)^2}{n} \sum_{i=1}^n \mathbb{E}_{n_t} \left(U_{\mathbf{X}_i}^2(\theta_t) \right) \\ &\leq \frac{4(nL+\ell_0)^2}{n} \mathbb{E}_{n_t} \left[\left(\sum_{i=1}^n U_{\mathbf{X}_i}(\theta_t) \right)^2 \right] \\ &\leq 4n(nL+\ell_0)^2 \mathbb{E}_{n_t} \left(U_{\nu_n}^2(\theta_t) \right), \quad (2.11)\end{aligned}$$

where we used the relation $\|.\|_2 \leq \|.\|_1$ in $\mathbb{R}^n$ in the second step, that is, for any $\theta \in \mathbb{R}^d$ and any i, $U_{\mathbf{X}_i}(\theta) \geq 0$, then $U_{\mathbf{X}_1}^2(\theta) + \ldots + U_{\mathbf{X}_n}^2(\theta) \leq (U_{\mathbf{X}_1}(\theta) + \ldots + U_{\mathbf{X}_n}(\theta))^2$. Taking expectation in θ, we obtain the inequality.

Furthermore,

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \mathbb{P}_{n_t} \left(\|\nabla_\theta U_{\mathbf{X}_i}(\theta_t)\|_2 > g(t) \right) &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{P}_{n_t} \left(2(nL+\ell_0) U_{\mathbf{X}_i}(\theta_t) > g^2(t) \right) \\ &\leq \frac{2(nL+\ell_0)}{n g^2(t)} \sum_{i=1}^n \mathbb{E}_{n_t} \left(U_{\mathbf{X}_i}(\theta_t) \right) \\ &\leq \frac{2(nL+\ell_0)}{g^2(t)} \mathbb{E}_{n_t} \left(U_{\nu_n}(\theta_t) \right), \quad (2.12)\end{aligned}$$

where we used Markov inequality in the second line.

Replacing (2.11) and (2.12) in (2.10), we get that:

$$\begin{aligned}\Delta_{2,t} &\leq \frac{n^{5/2}[2(nL+\ell_0)]^{3/2}}{g(t)}\left[\mathbb{E}_{n_t}(U^2_{\nu_n}(\theta_t))\mathbb{E}_{n_t}\left(U_{\nu_n}(\theta_t)\right)\right]^{1/2}\\ &\leq Cn^{\frac{11}{2}}\left(d\log^{2\beta}(n)\right)^{\frac{3(1+r)}{2}} g^{-1}(t),\end{aligned}$$

where C is a positive constant obtained when we apply Proposition 5.10 with $\alpha = 2$ and $\alpha = 1$ to upper the moments of $U_{\nu_n}(\theta_t)$.

Going back to Δ_t in (2.8), we have proven that:

$$\Delta_t \leq ng^2(t)e^{-2\alpha_n t} + Cn^{\frac{11}{2}}\left(d\log^{2\beta}(n)\right)^{\frac{3(1+r)}{2}} g^{-1}(t).$$

Optimizing with respect to $g(t)$, we deduce that:

$$\Delta_t \leq Cn^4\left(d\log^{2\beta}(n)\right)^{1+r} e^{-\frac{2\alpha_n t}{3}}, \quad \forall t \geq 0.$$

We conclude the proof by replacing the bound of Δ_t in (2.7). □

3 Functional inequalities and $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$

This section studies some functional inequalities that links the Dirichlet form and the relative entropy. When the probability measure is strongly log-concave (*i.e.* $r = 0$ under a $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ condition) a standard approach is to apply the log-Sobolev inequality (LSI for short). This idea relies on the initial works of [32] where LSI were introduced. The consequences of LSI to exponential ergodicity has then been an extensive field of research and we refer to [4] for an overview on this topic. A popular sufficient condition that ensures LSI is the log strong-concavity of the targeted measure (see among other [3]) and an impressive amount of literature has been focused on the existing links between these functional inequalities, ergodicity of the semi-group, transport inequalities and Lyapunov conditions. We refer to [10, 2] (these two works are far from being exhaustive). The great interest of LSI has then been observed in machine learning and statistics more recently as testified by the recent works in Monte Carlo samplings of [41, 44].

A popular way to extend LSI from the strongly concave situation to a more general case relies on the "strong convexity outside a ball" hypothesis using the perturbation argument of the seminal contributions of [35]. If this method proves to be suitable for the study of the simulated annealing process in [43], [35], it appears to be doubtful for the study of sampling problems with convex potentials V that satisfies $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ as this settings do not imply an asymptotic strong convexity of $\theta \longmapsto V(\theta)$ for large values of $\|\theta\|_2$. That being said, and maybe an even worst consequence of such approach, is the unavoidable dependency on the dimension for the LSI constant when using a perturbation approach, which leads to a serious exponential degradation of the convergence rates with the dimension of the ambient space.

To overcome these difficulties, we have chosen to use a slightly different functional inequality that may be considered as an innocent modification of LSI, but that indeed appears to be well suited to weakly log-concave setting described through an $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ assumption (*i.e.* $0 < r < 1$ under $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$) . For this purpose, we shall use weak log-Sobolev inequalities (WLSI for short below) that have been introduced in [9] and whose interest has been extensively studied in many works to obtain exponentially sub-linear rates of mixing. To derive such inequalities, our starting point will be the contribution of [11] that makes the link between Lyapunov conditions and WLSI. Our

approach based on $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ certainly shares some similarities with the recent work of [8] where some functional inequalities (Poincaré and transport inequalities) are obtained within a framework of variable curvature bound.

It is worth to mention that we pay special attention to the dependence of all constants involved in the functional inequalities in terms of n, d and r.

3.1 Poincaré, log-Sobolev and weak log-Sobolev inequalities

In this section we will focus on describing Poincaré, log-Sobolev and weak log-Sobolev inequalities in a general setting. These inequalities link the norm of a function or a related quantity to the norm of its derivative. Hence, we start by presenting some definitions and well-known facts.

Let be $\mathcal{B}(\mathbb{R}^d)$ the Borel algebra on $\mathbb{R}^d$ and m a probability measure defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Consider a function $f : \mathbb{R}^d \to \mathbb{R}$. For any $p \geq 1$, we define $\mathbb{L}^p_m(\mathbb{R}^d)$: the space of functions with finite p-norm

$$\|f\|_{\mathbb{L}^p_m(\mathbb{R}^d)} = \left[\int_{\mathbb{R}^d} |f|^p \mathrm{d}m\right]^{1/p},$$

whenever this quantity is finite. Moreover,

$$\mathcal{H}^1_m(\mathbb{R}^d) = \left\{ f : \mathbb{R}^d \to \mathbb{R}; f \in \mathbb{L}^2_m(\mathbb{R}^d), \nabla f \in \left(\mathbb{L}^2_m(\mathbb{R}^d)\right)^d \right\},$$

where ∇f is defined in the weak sense. $\mathcal{C}^1_b(\mathbb{R}^d)$ is the set of bounded and once differentiable functions on $\mathbb{R}^d$.

For any function $f \in \mathbb{L}^1_m(\mathbb{R}^d)$, we define its variance as:

$$Var_m(f) = \int_{\mathbb{R}^d} (f - m[f])^2 \mathrm{d}m,$$

where $m[f] = \int_{\mathbb{R}^d} f \mathrm{d}m$ and when $f \in \mathcal{H}^1_m(\mathbb{R}^d)$, the Dirichlet form of f is defined as:

$$\mathcal{E}_m(f) = \int_{\mathbb{R}^d} \|\nabla f\|_2^2 \mathrm{d}m.$$

The Poincaré inequality links the variance of f to its Dirichlet form.

Definition 3.1 (Poincaré inequality)**.** *The probability measure m satisfies a Poincaré inequality if there exists a constant $C_{\mathrm{PI}}(m)$ such that for any $f \in \mathcal{H}^1_m(\mathbb{R}^d)$,*

$$Var_m(f) \leq C_{\mathrm{PI}}(m) \mathcal{E}_m(f).$$

The optimal constant $C_{\mathrm{PI}}(m)$ is referred to as the Poincaré constant.

Remark 3.2. An important property of log-concave measures is that they satisfy a Poincaré inequality and in this situation a bound on the Poincaré constant may be found in Theorem 1.2 of [6].

We briefly introduce the log-Sobolev inequality (LSI). For any function $f : \mathbb{R}^d \to \mathbb{R}$ such that $\int_{\mathbb{R}^d} f^2 |\log(f)| \mathrm{d}m < \infty$, we define the entropy of f^2 as:

$$Ent_m(f^2) = \int_{\mathbb{R}^d} f^2 \log(f^2) \mathrm{d}m - \int_{\mathbb{R}^d} f^2 \mathrm{d}m \log\left(\int_{\mathbb{R}^d} f^2 \mathrm{d}m\right),$$

in this definition $0 \log 0$ is interpreted as 0.

Definition 3.3 (Log-Sobolev inequality)**.** *The probability measure m satisfies a LSI if there exists a positive constant $C_{\mathrm{LSI}}(m)$ such that for any $f \in \mathcal{H}^1(\mathbb{R}^d)$,*

$$Ent_m(f^2) \leq C_{\mathrm{LSI}}(m) \mathcal{E}_m(f).$$

Remark 3.4. In the particular case when m is a c-strongly log-concave measure, a LSI is verified, see [3], and the log-Sobolev constant is $C_{\mathrm{LSI}}(m) = \frac{2}{c}$, which is independent of the dimension d.

However, as mentioned before, in a weakly log-concave setting such as $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ when $0 < r < 1$, a weak log-Sobolev inequality (WLSI) would seem to be suitable to derive a good functional inequality which links the entropy and the Dirichlet form.

Definition 3.5 (Weak log-Sobolev inequality)**.** *The probability measure m satisfies a WLSI if a non-increasing function $\varphi : (0, +\infty) \mapsto \mathbb{R}_+$ exists such that for any $f \in \mathcal{C}^1_b(\mathbb{R}^d)$ and any $s > 0$,*

$$Ent_m(f^2) \leq \varphi(s)\mathcal{E}_m(f) + s\, Osc^2(f),$$

where $Osc(f) = \sup f - \inf f$.

This functional inequality was introduced in [9] and in said study they prove that the above definition is only necessary for small values of s, since it always holds that $Ent_m(f^2) \leq \left(\frac{1}{e} + \frac{1}{2}\right) Osc^2(f)$.

The following proposition establishes an important link between the Poincaré inequality and the WLSI, which can be obtained as a particular case of Proposition 3.1 in [9]. As discussed above, the function φ is described only for small values of s, in particular, we only describe it for $0 < s \leq e^{-1}$.

Proposition 3.6. *Assume that m satisfies a Poincaré inequality of constant $C_{\mathrm{PI}}(m)$, then m satisfies a WLSI. Moreover the function φ could be defined as follows:*

$$\varphi(s) = \mathfrak{a} C_{\mathrm{PI}}(m) \log\left(1/s\right), \quad 0 < s \leq e^{-1},$$

where $\mathfrak{a} > 0$ is a universal constant.

The proof may be found in Section 5.6.

3.2 Functional inequalities under $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$

In order to find a good functional inequality that links J_t and $\mathcal{E}_t$, we specify that the measure m is actually μ_n and the function is

$$f(\theta) = \sqrt{h_t(\theta)} = \sqrt{n_t(\theta)/\mu_n(\theta)},$$

where $t > 0$. Then we deduce that $J_t = Ent_{\mu_n}(h_t)$ and $\mathcal{E}_t = \mathcal{E}_{\mu_n}(\sqrt{h_t})$.

Under the $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ hypothesis, we are able to describe two situations: when $r = 0$ the probability measure μ_n is strongly log-concave and we will apply LSI, while when $0 < r < 1$, μ_n is weakly log-concave and we will use WLSI. We study each situation separately.

3.2.1 Strongly log-concave case

From Proposition 1.4, we observe that U_{ν_n} satisfies a $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL + \ell_0)$-condition, and then, $r = 0$ implies that U_{ν_n} is a $\mathfrak{c}n$-strongly convex function. The following proposition is an immediate consequence of Remark 3.4, therefore we omit the proof.

Proposition 3.7. *Assume $\mathcal{H}_{\pi_0}(\ell_0)$ and that each $\theta \mapsto -\log p_\theta(\mathbf{X}_i)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ with $r = 0$. Then μ_n verifies a LSI with constant $C_{\mathrm{LSI}}(\mu_n) = \frac{2}{\mathfrak{c}n}$, which is independent of d and indistinctly denoted as C_{LSI}. In particular:*

$$J_t \leq C_{\mathrm{LSI}} \mathcal{E}_t.$$

3.2.2 Weakly log-concave case

Since μ_n is a log-concave measure, then a Poincaré inequality is verified with constant $C_{\mathrm{PI}}(\mu_n)$, from now on denoted as C_{PI}. In addition, using Proposition 3.6, we observe that μ_n also satisfies a WLSI with function

$$\varphi(s) = \mathfrak{a} C_{\mathrm{PI}} \log(1/s), \quad 0 < s \leq e^{-1},$$

where $\mathfrak{a}$ is a universal constant. Then, in order to apply WLSI to $f = \sqrt{h_t}$, we need to find bounds for C_{PI} and for $Osc(\sqrt{h_t})$.

The next proposition states two lower bounds on the Poincaré constant within the $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ framework when $0 < r < 1$. The first one always holds, regardless the value of $(X_1, \ldots, X_n)$ that may be been randomly sampled. The second one has to be considered with high probability, with respect to the sampling process $(X_1, \ldots, X_n)$. The proof of Proposition 3.8 is deferred to Section 5.3.

Proposition 3.8. *Assume $\mathcal{H}_{\min}$,$\mathcal{H}_{\mathbf{n_0}}(L, \ell_0)$, $\mathcal{H}_{\pi_0}(\ell_0)$ and for any x, $\theta \mapsto -\log p_\theta(x)$ satisfies $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ where $0 < r < 1$, then:*

i) For any sample $(X_1, \ldots, X_n)$, it holds:

$$C_{\mathrm{PI}} \lesssim_{uc} \left(d \log^{2\beta}(n) \right)^{(1+r)^2}.$$

ii) Assume that $\theta \mapsto \mathbb{P}_\theta$ is injective and θ_0 exists such that $(X_1, \ldots, X_n) \sim \mathbb{P}_{\theta_0}$. If locally around θ_0, $\theta \mapsto \|\theta - \theta_0\|_2^{-\alpha} W_1(\mathbb{P}_\theta, \mathbb{P}_{\theta_0})$ does not vanish, then:

$$\mathbb{E}_{(X_1,\ldots,X_n)\sim\mathbb{P}_{\theta_0}}[C_{\mathrm{PI}}] \lesssim_{uc} \left(\frac{d \log n}{n} \right)^{\alpha}.$$

We are finally led to show that h_t is a bounded function and then compute an upper bound of the oscillation $Osc\left(\sqrt{h_t}\right)$, for any time $t > 0$. For this purpose, we observe that the Markov semi-group induces that $h_t = n_t/\mu_n = P_t h_0$ where $h_0 = n_0/\mu_n$. The next proposition implies the boundedness of h_t over $\mathbb{R}^d$ when n_0 is chosen as a Gaussian distribution with a carefully tuned covariance matrix.

Proposition 3.9. *Assume $\mathcal{H}_{\min}$,$\mathcal{H}_{\mathbf{n_0}}(L, \ell_0)$, $\mathcal{H}_{\pi_0}(\ell_0)$ and that, for any x, $\theta \mapsto -\log p_\theta(x)$ satisfies $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$. Then h_t is bounded for any $t > 0$ and there exist two universal constants $C_1 > 0$ and $C_2 > 0$ such that:*

$$Osc^2(\sqrt{h_t}) \leq Osc(h_t) \leq Osc(h_0) \lesssim_{uc} (C_1 d/n)^{\frac{dr}{2}} \exp\left(C_2 n (d \log^{2\beta}(n))^{1+r} \right).$$

The proof of Proposition 3.9 may be found in Section 5.3.

As a consequence of Remark 3.2 and Propositions 3.6 and 3.9, we state the following proposition, which we will not prove.

Proposition 3.10. *Assume $\mathcal{H}_{\min}$,$\mathcal{H}_{\mathbf{n_0}}(L, \ell_0)$, $\mathcal{H}_{\pi_0}(\ell_0)$ and that, for any x, $\theta \mapsto -\log p_\theta(x)$ satisfies $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$. Then:*

i) μ_n satisfies a Poincaré inequality with constant C_{PI}.

ii) For any $t > 0$ and any $s > 0$,

$$J_t \leq \varphi(s) \mathcal{E}_t + s O_{n,d},$$

where $O_{n,d} = (C_1 d/n)^{\frac{dr}{2}} \exp\left(C_2 n (d \log^{2\beta}(n))^{1+r} \right)$ and φ could be defined as:

$$\varphi(s) = \mathfrak{a} C_{\mathrm{PI}} \log(1/s), \quad 0 < s \leq e^{-1},$$

where $\mathfrak{a} > 0$ is a universal constant.

4 Proof of the main result

In Proposition 2.3, we proved that for any $t > 0$:

$$\partial_t\{J_t\} \leq -3\mathcal{E}_t + c_{n,d}e^{-\frac{2\alpha_n}{3}t}, \tag{4.1}$$

where $c_{n,d} \lesssim_{uc} n^4 \left(d\log^{2\beta}(n)\right)^{1+r}$. Once again, we study the cases strongly log-concave case ($r = 0$) and weakly log-concave case ($0 < r < 1$) separately.

Strongly log-concave case ($r = 0$). From Proposition 3.7, μ_n verifies a LSI with constant $C_{\mathrm{LSI}} = \frac{2}{\mathfrak{c}n}$. Then, in particular:

$$J_t \leq C_{\mathrm{LSI}}\mathcal{E}_t.$$

We combine this inequality with (4.1) to get that for any $t > 0$,

$$\partial_t\{J_t\} \leq -\frac{3J_t}{C_{\mathrm{LSI}}} + c_{n,d}e^{-\frac{2\alpha_n}{3}t}.$$

Applying Gronwall's lemma, we deduce that:

$$J_t \leq J_0e^{-\frac{3t}{C_{\mathrm{LSI}}}} + c_{n,d}e^{-\frac{3t}{C_{\mathrm{LSI}}}}\int_0^t e^{\left(\frac{3}{C_{\mathrm{LSI}}}-\frac{2\alpha_n}{3}\right)u}\mathrm{d}u.$$

Now, we upper bound the second term assuming that $\alpha_n \neq \frac{9}{2C_{\mathrm{LSI}}}$, the same bound is obtained if $\alpha_n = \frac{9}{2C_{\mathrm{LSI}}}$. We apply the mean value theorem to the function $y \mapsto -e^{-yt}$ in the interval $\left(\min\left\{\frac{3}{C_{\mathrm{LSI}}}, \frac{2\alpha_n}{3}\right\}, \max\left\{\frac{3}{C_{\mathrm{LSI}}}, \frac{2\alpha_n}{3}\right\}\right)$ to get that:

$$e^{-\frac{3t}{C_{\mathrm{LSI}}}}\int_0^t e^{\left(\frac{3}{C_{\mathrm{LSI}}}-\frac{2\alpha_n}{3}\right)u}\mathrm{d}u = \frac{e^{-\frac{2\alpha_n t}{3}} - e^{-\frac{3t}{C_{\mathrm{LSI}}}}}{\frac{3}{C_{\mathrm{LSI}}} - \frac{2\alpha_n}{3}} = te^{-at} \leq te^{-\min\left\{\frac{3}{C_{\mathrm{LSI}}}, \frac{2\alpha_n}{3}\right\}t},$$

where $a \in \left(\min\left\{\frac{3}{C_{\mathrm{LSI}}}, \frac{2\alpha_n}{3}\right\}, \max\left\{\frac{3}{C_{\mathrm{LSI}}}, \frac{2\alpha_n}{3}\right\}\right)$.

Using that for any $x \geq 0$, $xe^{-x} \leq e^{-x/2}$, with $x = \min\left\{\frac{3\mathfrak{c}n}{2}, \frac{2\alpha_n}{3}\right\}t$, we obtain that:

$$e^{-\frac{3t}{C_{\mathrm{LSI}}}}\int_0^t e^{\left(\frac{3}{C_{\mathrm{LSI}}}-\frac{2\alpha_n}{3}\right)u}\mathrm{d}u \leq \max\left\{\frac{C_{\mathrm{LSI}}}{3}, \frac{3}{2\alpha_n}\right\}e^{-\min\left\{\frac{3}{2C_{\mathrm{LSI}}}, \frac{\alpha_n}{3}\right\}t}.$$

We proved that:

$$J_t \leq \left(J_0 + c_{n,d}\max\left\{\frac{C_{\mathrm{LSI}}}{3}, \frac{3}{2\alpha_n}\right\}\right)e^{-\min\left\{\frac{3}{2C_{\mathrm{LSI}}}, \frac{\alpha_n}{3}\right\}t}, \quad t > 0.$$

A sufficient condition to obtain $J_t \leq \varepsilon$ is that

$$\left(J_0 + c_{n,d}\max\left\{\frac{C_{\mathrm{LSI}}}{3}, \frac{3}{2\alpha_n}\right\}\right)e^{-\min\left\{\frac{3}{2C_{\mathrm{LSI}}}, \frac{\alpha_n}{3}\right\}t} \leq \varepsilon,$$

or equivalently:

$$t \geq \max\left\{\frac{2C_{\mathrm{LSI}}}{3}, \frac{3}{\alpha_n}\right\}\left[\log(\varepsilon^{-1}) + \log\left(J_0 + c_{n,d}\max\left\{\frac{C_{\mathrm{LSI}}}{3}, \frac{3}{2\alpha_n}\right\}\right)\right].$$

Using that $C_{\mathrm{LSI}} = \frac{2}{\mathfrak{c}n}$, $J_0 \lesssim_{uc} d$, $c_{n,d} \lesssim_{uc} n^4 d\log^{2\beta}(n)$ and taking $\alpha_n = n$, we deduce that:

$$t \gtrsim_{uc} n^{-1}\left[\log(\varepsilon^{-1}) + \log(d) + \log(n)\right] \implies J_t \leq \varepsilon.$$

The choice of $\alpha_n = n$ guarantees that the expected number of jumps until obtaining an ε-error is minimal.

Weakly log-concave case ($0 < r < 1$). From Proposition 3.10, we deduce a differential inequality for J_t. For any $t > 0$ and $0 < s \leq e^{-1}$:

$$\partial_t \{J_t\} \leq -\frac{3J_t}{\varphi(s)} + \frac{3s\, O_{n,d}}{\varphi(s)} + c_{n,d} e^{-\frac{2\alpha_n}{3}t}.$$

Using Gronwall's lemma, we get that:

$$J_t \leq \inf_{0<s\leq e^{-1}} \left\{ J_0 e^{-\frac{3t}{\varphi(s)}} + sO_{n,d} + c_{n,d} e^{-\frac{3t}{\varphi(s)}} \int_0^t e^{\left(\frac{3}{\varphi(s)} - \frac{2\alpha_n}{3}\right)u} \mathrm{d}u \right\}.$$

The mean value theorem, as in the strongly log-concave case, implies that:

$$J_t \leq \inf_{0<s\leq e^{-1}} \left\{ J_0 e^{-\frac{3t}{\varphi(s)}} + sO_{n,d} + c_{n,d} \max\left\{ \frac{\varphi(s)}{3}, \frac{3}{2\alpha_n} \right\} e^{-\min\left\{\frac{3}{2\varphi(s)}, \frac{\alpha_n}{3}\right\}t} \right\}.$$

We then choose $s = s_t = e^{-A\sqrt{t}}$, with $A > 0$. We observe that $0 < s_t \leq e^{-1}$ if $t \geq A^{-2}$, and then:

$$\varphi(s_t) = \mathfrak{a} C_{\mathrm{PI}} \log\left(1/s_t\right) = \mathfrak{a} A C_{\mathrm{PI}} \sqrt{t},$$

which leads to

$$J_t \leq J_0 e^{-\frac{3\sqrt{t}}{\mathfrak{a} A C_{\mathrm{PI}}}} + O_{n,d} e^{-A\sqrt{t}} + c_{n,d} \max\left\{ \frac{\mathfrak{a} A C_{\mathrm{PI}} \sqrt{t}}{3}, \frac{3}{2\alpha_n} \right\} e^{-\min\left\{\frac{3}{2\mathfrak{a} A C_{\mathrm{PI}}\sqrt{t}}, \frac{\alpha_n}{3}\right\}t}.$$

The choice $A = \sqrt{\frac{3}{\mathfrak{a} C_{\mathrm{PI}}}}$ guarantees that the first two terms are of the same order with respect to t. Moreover, the condition $t \geq A^{-2}$ becomes $t \geq \frac{\mathfrak{a} C_{\mathrm{PI}}}{3}$.

Then,

$$J_t \leq (J_0 + O_{n,d}) e^{-\sqrt{\frac{3t}{\mathfrak{a} C_{\mathrm{PI}}}}} + c_{n,d} \max\left\{ \sqrt{\frac{\mathfrak{a} C_{\mathrm{PI}} t}{3}}, \frac{3}{2\alpha_n} \right\} e^{-\min\left\{\sqrt{\frac{3}{4\mathfrak{a} C_{\mathrm{PI}} t}}, \frac{\alpha_n}{3}\right\}t}.$$

We assume that $\sqrt{\frac{3}{4\mathfrak{a} C_{\mathrm{PI}} t}} \leq \frac{\alpha_n}{3}$, that is, $t \gtrsim_{uc} \frac{1}{\alpha_n^2 C_{\mathrm{PI}}}$, then

$$\begin{aligned} J_t &\leq \left(J_0 + O_{n,d} + c_{n,d} \sqrt{\frac{\mathfrak{a} C_{\mathrm{PI}} t}{3}} \right) e^{-\sqrt{\frac{3t}{4\mathfrak{a} C_{\mathrm{PI}}}}} \\ &\lesssim_{uc} \left(J_0 + O_{n,d} + c_{n,d} C_{\mathrm{PI}} \right) e^{-\sqrt{\frac{3t}{16\mathfrak{a} C_{\mathrm{PI}}}}}, \end{aligned}$$

where we used the inequality: for any $x \geq 0$, $xe^{-x} \leq e^{-x/2}$), with $x = \sqrt{\frac{3t}{4\mathfrak{a} C_{\mathrm{PI}}}}$.

A sufficient condition to get $J_t \leq \varepsilon$ is that the right hand of the previous equation is less than ε, which is verified if:

$$t \gtrsim_{uc} C_{\mathrm{PI}} \left[\log^2(\varepsilon^{-1}) + \log^2\left(J_0 + O_{n,d} + c_{n,d} C_{\mathrm{PI}}\right) \right]$$

and $t \gtrsim_{uc} \frac{1}{\alpha_n^2 C_{\mathrm{PI}}}$. To find α_n, we look for the value of α_n that minimizes $\alpha_n \cdot t_\varepsilon$, where t_ε is the minimum time required to achieve an ε-error. From this, we observe that $\alpha_n = \frac{1}{C_{\mathrm{PI}}}$. Of course, in the choice of α_n we ignore multiplicative constants that do not depend on n nor d.

To obtain explicit estimates that depend on n and d, we use Proposition 3.8, where we showed that $C_{\mathrm{PI}} \lesssim_{uc} \left(d \log^{2\beta}(n) \right)^{(1+r)^2}$. Instead of working with the constant C_{PI} in φ, we directly use the upper bound $\kappa \left(d \log^{2\beta}(n) \right)^{(1+r)^2}$. This allows us to keep all previous

computations unchanged, with the only difference being the substitution of C_{PI} by its upper bound.

Therefore, using the values of $O_{n,d}$, $c_{n,d}$ and the upper bound of J_0, we finally observe that if $\alpha_n = \left(d\log^{2\beta}(n)\right)^{-(1+r)^2}$ and

$$t \gtrsim_{uc} \left(d\log^{2\beta}(n)\right)^{(1+r)^2}\left[\log^2(\varepsilon^{-1}) + n^2\left(d\log^{2\beta}(n)\right)^{2(1+r)}\right],$$

then $J_t \leq \varepsilon$.

5 Technical results

5.1 Growth properties under the Kurdyka-Łojasiewicz inequality

We remind here some important consequences of the KL inequality that implies several relationships between the function and the norm of its gradient. The proof of these inequalities may be found in Lemma 15 of [29] (a small mistake appears and we correct the statement with a factor 2 in our work).

Proposition 5.1. *Assume that a function V satisfies $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$, then for all $\theta \in \mathbb{R}^d$,*

$$\frac{2\mathfrak{c}}{1-r}\left[V^{1-r}(\theta) - \min V^{1-r}\right] \leq \|\nabla V(\theta)\|^2 \leq 2L\left[V(\theta) - \min V\right] \leq 2LV(\theta).$$

It is furthermore possible to assess a minimal and maximal growth property of any function that satisfies $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$, which is necessarily lower and upper bounded by a positive power of the distance to its minimizer.

Proposition 5.2. *Assume that a function V satisfies $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$, then for all $\theta \in \mathbb{R}^d$,*

$$V^{1+r}(\theta) \geq V^{1+r}(\theta) - \min(V)^{1+r} \geq \frac{(1+r)\mathfrak{c}}{2}\|\theta - \arg\min V\|^2$$

and

$$V(\theta) - \min(V) \leq \frac{L}{2}\|\theta - \arg\min V\|^2.$$

5.2 Properties of U_{ν_n} and $\mathcal{Z}_n$

We star by proving Proposition 1.4, which shows how a $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ condition on $\theta \mapsto -\log p_\theta(x)$ is translated to U_{ν_n}.

Proof of Proposition 1.4. First, we observe that if each $\theta \mapsto \nabla \log p_\theta(\mathbf{X}_i)$ is L-Lipschitz and $\theta \mapsto \nabla \log \pi_0$ is ℓ_0-Lipschitz, then the triangle inequality implies that

$$\|\nabla U_{\nu_n}(\theta_1) - \nabla U_{\nu_n}(\theta_2)\|_2 \leq \|\nabla U_{\mathbf{X}_1}(\theta_1) - \nabla U_{\mathbf{X}_1}(\theta_2)\|_2 \leq (nL + \ell_0)\|\theta_1 - \theta_2\|_2,$$

for any $\theta_1, \theta_2 \in \mathbb{R}^d$. Then $\theta \mapsto \nabla U_{\nu_n}(\theta)$ is $(nL + \ell_0)$-Lipschitz.

Second, we consider the lower-bound property on the curvature and observe that for any $\theta \in \mathbb{R}^d$,

$$\underline{\lambda}_{\nabla^2 U_{\nu_n}}(\theta) = \inf_{e\in\mathbb{R}^d:\|e\|_2=1} e^T(\nabla^2 U_{\nu_n}(\theta))e \geq \frac{1}{n}\sum_{i=1}^n \inf_{e\in\mathbb{R}^d:\|e\|_2=1} e^T(\nabla^2 U_{\mathbf{X}_i}(\theta))e.$$

The log concavity of the prior yields

$$\underline{\lambda}_{\nabla^2 U_{\nu_n}}(\theta) \geq \frac{1}{n}\sum_{i=1}^n \underline{\lambda}_{\nabla^2(-n\log p_\theta(\mathbf{X}_i))} = \sum_{i=1}^n \underline{\lambda}_{\nabla^2(-\log p_\theta(\mathbf{X}_i))}.$$

Then, the $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ property applied to each term of the sum above and the condition $\min_{\theta\in\mathbb{R}^d} -\log\pi_0(\theta) > 0$ from $\mathcal{H}_{\pi_\mathbf{0}}(\ell_0)$ yield

$$\underline{\lambda}_{\nabla^2 U_{\nu_n}}(\theta) \geq \mathfrak{c}\sum_{i=1}^n [-\log p_\theta(\mathbf{X}_i)]^{-r} \geq \mathfrak{c}n^r \sum_{i=1}^n U^{-r}_{\mathbf{X}_i}(\theta) = \mathfrak{c}n^{1+r}\left(\frac{1}{n}\sum_{i=1}^n U^{-r}_{\mathbf{X}_i}(\theta)\right).$$

From the Jensen inequality, we finally deduce that:

$$\underline{\lambda}_{\nabla^2 U_{\nu_n}}(\theta) \geq \mathfrak{c}n^{1+r}\left(\frac{1}{n}\sum_{i=1}^n U^{-r}_{\mathbf{X}_i}(\theta)\right) \geq \mathfrak{c}n^{1+r} U^{-r}_{\nu_n}(\theta).$$

We conclude that U_{ν_n} satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL+\ell_0)$. The proof is similar when we consider $U_{\mathbf{X}_i}$ instead of U_{ν_n}. □

The following proposition shows the growth of the minimum and the minimizer of U_{ν_n} with respect to d and n.

Proposition 5.3. *Assume $\mathcal{H}_{\pi_\mathbf{0}}(\ell_0)$, $\mathcal{H}_{\min}$ and that for any x: $\theta \longmapsto -\log p_\theta(x)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$. Then:*

$$\|\arg\min U_{\nu_n}\|_2 \lesssim_{uc} (\sqrt{d}\log^\beta(n))^{1+r} \quad \textit{and} \quad \min_{\theta\in\mathbb{R}^d} U_{\nu_n}(\theta) \lesssim_{uc} nd\log^{2\beta}(n).$$

Proof. Proposition 1.4 shows that U_{ν_n} satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL+\ell_0)$. Hence, we can apply Proposition 5.2 with $\theta = 0$ and $\arg\min U_{\nu_n} = \theta^*_n$. We deduce that:

$$\|\theta^*_n\|_2^2 \leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}} U^{1+r}_{\nu_n}(0).$$

To obtain an upper bound of $U_{\nu_n}(0)$ we first find an upper bound of $U_{\mathbf{X}_i}(0)$ using $\mathcal{H}_{\min}$ and once again Proposition 5.2, for all i, as follows:

$$U_{\mathbf{X}_i}(0) \leq \min\ U_{\mathbf{X}_i} + \frac{nL+\ell_0}{2}\|\arg\min U_{\mathbf{X}_i}\|_2^2 \lesssim_{uc} nd\log^{2\beta}(n),$$

then $U_{\nu_n}(0) \lesssim_{uc} nd\log^{2\beta}(n)$ and:

$$\|\theta^*_n\|_2^2 \lesssim_{uc} \left(d\log^{2\beta}(n)\right)^{1+r}.$$

The second part comes from $U_{\nu_n}(\theta^*_n) \leq U_{\nu_n}(0)$. □

Proposition 5.4. *Assume $\mathcal{H}_{\pi_\mathbf{0}}(\ell_0)$, $\mathcal{H}_{\mathbf{n_0}}(L, \ell_0)$, $\mathcal{H}_{\min}$ and that for any x: $\theta \longmapsto -\log p_\theta(x)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$. Then the normalizing constant of μ_n verifies the following inequality:*

$$\mathcal{Z}_n \leq 2\left(\frac{2\pi}{\mathfrak{c}n^{1+r}}\right)^{d/2} d^{dr/2}.$$

Proof. In Proposition 1.4, we proved that U_{ν_n} satisfies a $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL+\ell_0)$-condition, then if we apply Proposition 5.2 to U_{ν_n} and denote $\theta^\star_n = \arg\min U_{\nu_n}$, we deduce that:

$$U_{\nu_n}(\theta) \geq a_{n,r}\|\theta - \theta^*_n\|_2^{\frac{2}{1+r}},$$

where $a_{n,r} = n\left(\frac{(1+r)\mathfrak{c}}{2}\right)^{\frac{1}{1+r}}$. We compute an upper bound of $\mathcal{Z}_n$ using the inequality above:

$$\mathcal{Z}_n = \int_{\mathbb{R}^d} e^{-U_{\nu_n}(\theta)}\mathrm{d}\theta \leq \int_{\mathbb{R}^d} e^{-a_{n,r}\|\theta-\theta^\star_n\|_2^{\frac{2}{1+r}}}\mathrm{d}\theta.$$

A change of variable and the well known equality

$$\int_{\mathbb{R}^d} e^{-a\|\theta\|_2^\ell} \mathrm{d}\theta = \frac{\pi^{d/2}\Gamma(d/\ell+1)}{a^{d/\ell}\Gamma(d/2+1)}, \quad \forall a>0, \quad \forall \ell>0,$$

imply that:

$$\mathcal{Z}_n \leq \int_{\mathbb{R}^d} e^{-a_{n,r}\|\theta\|_2^{\frac{2}{1+r}}} \mathrm{d}\theta \leq \left(\frac{\pi}{a_{n,r}^{1+r}}\right)^{d/2} \frac{\Gamma\left(\frac{d(1+r)}{2}+1\right)}{\Gamma\left(\frac{d}{2}+1\right)}.$$

Then, from standard relations on the Gamma function we conclude that:

$$\mathcal{Z}_n \leq 2\left(\frac{2\pi}{\mathfrak{c}n^{1+r}}\right)^{d/2} d^{dr/2}.$$

□

5.3 Smoothness and boundedness of the semi-group

Proof of Proposition 2.2. $\underline{i)}$ We start by proving that $h_0 = n_0(\theta)/\mu_n(\theta)$ is a bounded function. For all $\theta \in \mathbb{R}^d$,

$$h_0(\theta) = (2\pi\sigma^2)^{-d/2}\mathcal{Z}_n e^{-\frac{\|\theta\|_2^2}{2\sigma^2}+U_{\nu_n}(\theta)} \leq 2\left(\sigma^2\mathfrak{c}n^{1+r}\right)^{-d/2} d^{\frac{dr}{2}} e^{-\frac{\|\theta\|_2^2}{2\sigma^2}+U_{\nu_n}(\theta)}, \tag{5.1}$$

where we used that n_0 is the density function of a $\mathcal{N}(0_d, \sigma^2 I_d)$ random variable and the bound of $\mathcal{Z}_n$ obtained in Proposition 5.4. Let us focus on the exponent of Equation (5.1).

From Proposition 1.4, U_{ν_n} satisfies a $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL+\ell_0)$-condition. Then, we denote $\theta_n^\star = \arg\min U_{\nu_n}$ and apply Proposition 5.2 to U_{ν_n}:

$$U_{\nu_n}(\theta) \leq U_{\nu_n}(\theta_n^\star) + \frac{(nL+\ell_0)}{2}\|\theta-\theta_n^\star\|_2^2.$$

The exponent of (5.1) satisfies the following inequality:

$$-\frac{\|\theta\|_2^2}{2\sigma^2} + U_{\nu_n}(\theta) \leq -\frac{\|\theta\|_2^2}{2\sigma^2} + U_{\nu_n}(\theta_n^\star) + \frac{(nL+\ell_0)}{2}\|\theta-\theta_n^\star\|_2^2.$$

For all $\sigma^2 \leq \frac{c_2}{nL+\ell_0}$ where $0<c_2<1$, a straightforward optimization on θ yields:

$$-\frac{\|\theta\|_2^2}{2\sigma^2} + \frac{(nL+\ell_0)}{2}\|\theta-\theta_n^\star\|_2^2 \leq \frac{(nL+\ell_0)}{2(1-c_2)}\|\theta_n^\star\|_2^2,$$

then:

$$-\frac{\|\theta\|_2^2}{2\sigma^2} + U_{\nu_n}(\theta) \leq U_{\nu_n}(\theta_n^\star) + \frac{(nL+\ell_0)}{2(1-c_2)}\|\theta_n^\star\|_2^2 \leq C_2 n\left(d\log^{2\beta}(n)\right)^{1+r}, \tag{5.2}$$

where we used Proposition 5.3 in the last step and we define C_2 as a constant. Replacing (5.2) in (5.1), we get that:

$$\|h_0\|_\infty \leq 2\left(\sigma^2\mathfrak{c}n^{1+r}\right)^{-d/2} d^{dr/2} e^{C_2 n\left(d\log^{2\beta}(n)\right)^{1+r}} \leq \left(\frac{C_1 d}{n}\right)^{\frac{dr}{2}} e^{C_2 n\left(d\log^{2\beta}(n)\right)^{1+r}},$$

where we used $\frac{c_1}{nL+\ell_0} \leq \sigma^2$ from hypothesis $\mathcal{H}_{\mathbf{n_0}}(L,\ell_0)$ and one more time we define C_1 as a universal constant.

$\underline{ii)}$ Using that $x \mapsto \log x$ is increasing and the part $i)$, we have that:

$$J_0 \leq \log\left(\|h_0\|_\infty\right) \lesssim_{uc} n(d\log^{2\beta}(n))^{1+r} + rd\log\left(d/n\right).$$

□

Proof of Proposition 3.8 . $\underline{i)}$. The proof relies on an argument set up with a "fixed" sample $(X_1, \dots, X_n)$. Our starting point is Proposition 1.4 where we proved that U_{ν_n} satisfies a $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL + \ell_0)$-condition. Then, we apply Proposition 5.2 with $\theta^*_n = \arg\min U_{\nu_n}$ and deduce that:

$$\|\theta - \theta^*_n\|_2^2 \leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}} U^{1+r}_{\nu_n}(\theta).$$

We use the fact that for any distribution μ and any $a \in \mathbb{R}^d$, we have $Var_\mu(\theta) \leq \mathbb{E}_\mu(\|\theta - a\|_2^2)$. Taking $\mu = \mu_n$ and $a = \theta^*_n$, then:

$$Var_{\mu_n}(\theta) \leq \mathbb{E}_{\mu_n}\left(\|\theta - \theta^*_n\|_2^2\right) \leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}} \mathbb{E}_{\mu_n}[U^{1+r}_{\nu_n}(\theta)].$$

We then use the ergodic behaviour of $(\theta_t)_{t\geq 0}$ and observe that there exists a constant C independent from n and d such that:

$$Var_{\mu_n}(\theta) \leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}} \limsup_{t\geq 0} \mathbb{E}_{n_t}[U^{1+r}_{\nu_n}(\theta_t)] \leq C\left(d\log^{2\beta}(n)\right)^{(1+r)^2},$$

where the last inequality comes from Proposition 5.10.

We now use the Bobkov bound on the Poincaré constant for a log-concave distribution, see Theorem 1.2 of [6], and deduce that a universal constant K exists such that:

$$C_{\mathrm{PI}}(\mu_n) \leq 4K^2 Var_{\mu_n}(\theta).$$

Using the upper bound of the variance, we deduce that a universal $\kappa > 0$ exists such that:

$$C_{\mathrm{PI}}(\mu_n) \leq \kappa\left(d\log^{2\beta}(n)\right)^{(1+r)^2}.$$

$\underline{ii)}$. For the second point, we consider a situation on average over the samples and the result uses the concentration of the posterior distribution around its mean. We know from Theorem 3 of [29] that a constant $c > 0$ exists such that:

$$\mathbb{E}_{(X_1,\dots,X_n)\sim \mathbb{P}_{\theta_0}}[Var_{\mu_n}(\theta)] \leq c\epsilon_{n,d},$$

with $\epsilon_{n,d} = \left(\frac{Ld\log n}{n}\right)^{\alpha^{-1}}$. The result follows using the Jensen inequality and the Bobkov bound. □

Proof of Proposition 3.9 . We proved in Proposition 2.2 that h_0 is bounded and

$$\|h_0\|_\infty \leq \left(\frac{C_1 d}{n}\right)^{\frac{dr}{2}} e^{C_2 n\left(d\log^{2\beta}(n)\right)^{1+r}},$$

where C_1 and C_2 are two positive constants independent of n and d. In addition, by definition, for any measurable function h_0:

$$h_t(\theta) = P_t h_0(\theta) = \mathbb{E}[h_0(\theta_t)|\theta_0 = \theta], \quad t > 0,$$

where P_t is a Markov Feller semi-group. If h_0 is bounded, then $\inf h_0 \leq h_t(\theta) \leq \sup h_0$, for any $\theta \in \mathbb{R}^d$ and any $t > 0$. In addition,

$$Osc\left(\sqrt{h_t}\right) \leq Osc\left(\sqrt{h_0}\right), \quad t > 0.$$

We get the statement taking into account that $Osc\left(\sqrt{h_0}\right) \leq \sqrt{\|h_0\|_\infty}$. □

5.4 Regularity of J_t

Let's start by proving the following result that will be used several times.

Proposition 5.5. *Let $\mathcal{G}_t$ be the diffusion operator under the average effect of X_t, defined in Equation* (2.3)*. If $\mathcal{G}_t^\star$ is the adjoint operator of $\mathcal{G}_t$ in $\mathbb{L}^2(\mathbb{R}^d)$, then for any $t > 0$:*

$$\partial_t\{n_t(\theta)\} = \mathcal{G}_t^\star n_t(\theta). \tag{5.3}$$

Proof. Let's prove (5.3) in the weak sense. Consider $g \in \mathcal{C}^2(\mathbb{R}^d)$ and define $\widetilde{g}(\theta, x) = g(\theta)$, for any $x \in \mathcal{X}$. We deduce that:

$$\begin{aligned}
\int_{\mathbb{R}^d} g(\theta)\partial_t\{n_t(\theta)\}\mathrm{d}\theta &= \int_{\mathbb{R}^d} g(\theta)\partial_t\left\{\sum_{i=1}^n m_t(\theta, \mathbf{X}_i)\right\}\mathrm{d}\theta \\
&= \int_{\mathbb{R}^d} \sum_{i=1}^n g(\theta)\partial_t\{m_t(\theta, \mathbf{X}_i)\}\mathrm{d}\theta \\
&= \int_{\mathbb{R}^d} \sum_{i=1}^n \widetilde{g}(\theta, \mathbf{X}_i)\partial_t\{m_t(\theta, \mathbf{X}_i)\}\mathrm{d}\theta \\
&= \int_{\mathbb{R}^d} \sum_{i=1}^n \mathcal{L}\widetilde{g}(\theta, \mathbf{X}_i)\ m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta,
\end{aligned}$$

where we used the definition of n_t in the first step and Kolmogorov forward Equation (2.2) in the last one. Since $\widetilde{g}$ does not depend on x, we observe that $\mathcal{L}_2\widetilde{g}(\theta, \mathbf{X}_i) = 0$ and we only need to compute the remaining term $\mathcal{L}_1\widetilde{g}(\theta, \mathbf{X}_i)$:

$$\begin{aligned}
\int_{\mathbb{R}^d} g(\theta)\partial_t\{n_t(\theta)\}\mathrm{d}\theta &= \int_{\mathbb{R}^d} \sum_{i=1}^n \mathcal{L}_1\widetilde{g}(\theta, \mathbf{X}_i)\ m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta \\
&= \int_{\mathbb{R}^d} \sum_{i=1}^n \left[-\langle\nabla_\theta g(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle + \Delta_\theta g(\theta)\right] m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta \\
&= -\int_{\mathbb{R}^d} \sum_{i=1}^n \langle\nabla_\theta g(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle m_t(\mathbf{X}_i|\theta)\mathrm{d}n_t(\theta) + \int_{\mathbb{R}^d} \Delta_\theta g(\theta)\mathrm{d}n_t(\theta) \\
&= \int_{\mathbb{R}^d} \mathcal{G}_t g(\theta)\mathrm{d}n_t(\theta),
\end{aligned}$$

where we used the fact that $m_t(\theta, \mathbf{X}_i) = m_t(\mathbf{X}_i|\theta)n_t(\theta)$. Using one more time Remark 3.19 of [39], we could verify that the first and second derivatives of n_t with respect to θ are bounded, therefore, n_t is differentiable in time and Equation (5.3) is satisfied in the strong sense. □

Now let's focus on proving that J_t is finite. Using the ellipticity of the semi-group generated by $\mathcal{G}_t$, we can use the result of [36] and have that for any $t > 0$, $n_t \in \mathcal{C}^\infty(\mathbb{R}^d)$ and the irreducibility yields $\forall t \geq 0$, $n_t > 0$. In Proposition 3.9 we will prove that $h_0 = n_0(\theta)/\mu_n(\theta)$ is bounded and a standard consequence is that:

$$\left\|\frac{n_t(\theta)}{\mu_n(\theta)}\right\|_\infty = \|h_t\|_\infty \leq \|h_0\|_\infty.$$

Therefore $J_t < \infty$, for any $t > 0$.

In the following proposition it is shown that J_t is differentiable and that it is possible to exchange derivative and the integral sign.

Proposition 5.6. *Assume $\mathcal{H}_{\pi_0}(\ell_0)$, $\mathcal{H}_{\min}$, $\mathcal{H}_{\mathbf{n_0}}(L,\ell_0)$ and that each $\theta \mapsto -\log p_\theta(\mathbf{X}_i)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c},L)$. Then:*

$$\partial_t\{J_t\} = \int_{\mathbb{R}^d}\left(1+\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\right)\partial_t\{n_t(\theta)\}\mathrm{d}\theta = \int_{\mathbb{R}^d}\mathcal{G}_t\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\mathrm{d}n_t(\theta).$$

Proof. We need to verify that the function of $t\mapsto \log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)n_t(\theta)$ is differentiable, which is an immediate consequence of the fact that $t\mapsto m_t(\theta,x)$ is differentiable for any $(\theta,x)\in\mathbb{R}^d\times\mathcal{X}$. Then so are $t\mapsto n_t(\theta)$ and $t\mapsto \log(n_t(\theta)/\mu_n(\theta))n_t(\theta)$.

The second part of the proof consists of finding an integrable function g such that $\left|\partial_t\left\{\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)n_t(\theta)\right\}\right|\leq g(\theta)$ for any $t>0$.

For any $\theta\in\mathbb{R}^d$ and any $t>0$, we observe that:

$$\begin{aligned}\left|\partial_t\left\{\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)n_t(\theta)\right\}\right| &= \left|\left(1+\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\right)\partial_t\{n_t(\theta)\}\right| \\ &\leq |1+\log(\|h_0\|_\infty)||\partial_t\{n_t(\theta)\}|.\end{aligned}$$

We then have to focus on $\partial_t\{n_t(\theta)\}$. Let us denote by $p_{0,t}(\vartheta,\cdot)$ the density function of $\theta_t|\theta_0=\vartheta$, then the density of θ_t could be rewritten as:

$$n_t(\theta) = \int_{\mathbb{R}^d} n_0(\vartheta)p_{0,t}(\vartheta,\theta)\mathrm{d}\vartheta$$

where n_0 is the density function of θ_0 and we assumed in hypothesis $\mathcal{H}_{\mathbf{n_0}}(L,\ell_0)$ that n_0 is a Gaussian distribution. We recall that $p_{0,t}(\vartheta,\cdot)$ satisfies the Fokker Planck equation:

$$\partial_t\{p_{0,t}(\vartheta,\theta)\} = \mathcal{G}_t^\star p_{0,t}(\vartheta,\theta),$$

where $\mathcal{G}_t^\star$ is the adjoint operator of $\mathcal{G}_t$ in $\mathbb{L}^2(\mathbb{R}^d)$. Using the Remark 3.19 of [39], for any multi-index $\alpha\in\mathbb{N}^d$, the functions $(t,\theta)\mapsto\frac{\partial^{|\alpha|}}{\partial\theta^\alpha}p_{0,t}(\vartheta,\theta)$ are continuous and bounded by a polynomial of $\|\vartheta\|_2$ for any $t\in[T^{-1},T]$ and $T>1$. Therefore, the previous equation is satisfied in the strong sense when $t\in[T^{-1},T]$. The following steps are true for $t\in[T^{-1},T]$.

Taking into account that we can exchange derivative with the integral sign to find $\partial_t\{n_t(\theta)\}$ since the spatial derivatives of $p_{0,t}$ are bounded by a polynomial function and n_0 is a Gaussian density, we then observe that $\partial_t\{n_t(\theta)\}$ could be written as follows:

$$\partial_t\{n_t(\theta)\} = \int_{\mathbb{R}^d} n_0(\vartheta)\partial_t\left\{p_{0,t}(\vartheta,\theta)\right\}\mathrm{d}\vartheta = \int_{\mathbb{R}^d}\mathcal{G}_t n_0(\vartheta)p_{0,t}(\vartheta,\theta)\mathrm{d}\vartheta.$$

Using the definition of $\mathcal{G}_t$ in Equation (2.3), we deduce that for any $\vartheta\in\mathbb{R}^d$:

$$\mathcal{G}_t n_0(\vartheta) = \frac{1}{\sigma^2}\left(\sum_{i=1}^n\langle\vartheta,\nabla_\vartheta U_{\mathbf{X}_i}(\vartheta)\rangle m_t(\mathbf{X}_i|\vartheta) + \frac{\|\vartheta\|_2^2}{\sigma^2} - d\right)n_0(\vartheta),$$

where $\theta_0\sim\mathcal{N}(0_d,\sigma^2 I_d)$ and $m_t(\cdot|\vartheta)$ is the conditional distribution of X_t when $\theta_0=\vartheta$. Note that $m_t(\cdot|\vartheta)$ is the probability distribution of a discrete random variable, then $m_t(\mathbf{X}_i|\vartheta)\leq 1$, for any i.

The Lipschitz regularity of $\nabla_\theta U_{\mathbf{X}_i}(\theta)$ guarantees that there exists a positive polynomial q_2 of degree 2 such that:

$$|\mathcal{G}_t n_0(\vartheta)| \leq q_2(\vartheta)n_0(\vartheta).$$

The bound above and the Cauchy-Schwarz inequality lead to:

$$\begin{aligned}|\partial_t\{n_t(\theta)\}| &\leq \int_{\mathbb{R}^d} q_2(\vartheta) n_0(\vartheta) p_{0,t}(\vartheta,\theta)\mathrm{d}\vartheta \\ &\leq \left(\int_{\mathbb{R}^d} q_2^2(\vartheta) n_0(\vartheta) p_{0,t}(\vartheta,\theta)\mathrm{d}\vartheta\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} n_0(\vartheta) p_{0,t}(\vartheta,\theta)\mathrm{d}\vartheta\right)^{\frac{1}{2}} \\ &\leq \sqrt{A_T n_t(\theta)},\end{aligned}$$

where we used that the first integral is bounded by a constant A_T since $p_{0,t}(\vartheta,\theta)$ is bounded by a polynomial of $\|\vartheta\|_2$ and n_0 is the density function of a Gaussian distribution.

We have proven that for any $t \in [T^{-1}, T]$:

$$\begin{aligned}\left|\partial_t\left\{\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right) n_t(\theta)\right\}\right| &\leq (1+\log(\|h_0\|_\infty))\sqrt{A_T n_t(\theta)} \\ &\leq (1+\log(\|h_0\|_\infty))\sqrt{A_T\|h_0\|_\infty}\sqrt{\mu_n(\theta)} = g(\theta).\end{aligned}$$

The integrability of the function $\theta \mapsto \sqrt{\mu_n(\theta)}$ is a consequence of Propositions 1.4 and 5.2, from which we deduce that U_{ν_n} satisfies a $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL+\ell_0)$ condition, then for all $\theta \in \mathbb{R}^d$:

$$\sqrt{\mu_n(\theta)} \leq \mathcal{Z}_n^{-1/2} \exp\left\{-a_r\|\theta - \arg\min U_{\nu_n}\|^{2/(1+r)}\right\},$$

where a_r depends of r and n. The inequality above guarantees that $\int_{\mathbb{R}^d} g(\theta)\mathrm{d}\theta < \infty$.

After proving the exchange of derivative and integral sign when $t \in [T^{-1}, T]$, for any $T > 1$, we can conclude it for any $t > 0$. The second part of the statement is immediate from Proposition 5.5. □

5.5 Evolution of the weighted $\mathbb{L}^2$ distance I_t

The quantity I_t defined in (2.9) measures how close to ν_n the conditional distribution of $X_t|\theta_t$ is. Then, to study I_t, we first remark that it may be rewritten in a simpler way.

$$\begin{aligned}I_t &= \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i=1}^n\left(\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)} - 1\right)^2 \mathrm{d}n_t(\theta) \\ &= \int_{\mathbb{R}^d}\sum_{i=1}^n\left(\frac{m_t^2(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)} - 2m_t(\mathbf{X}_i|\theta) + \nu_n(\mathbf{X}_i)\right)\mathrm{d}n_t(\theta) \\ &= \int_{\mathbb{R}^d}\left(\sum_{i=1}^n\frac{m_t^2(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)} - 1\right)\mathrm{d}n_t(\theta) \\ &= \int_{\mathbb{R}^d}\sum_{i=1}^n\frac{m_t^2(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}\mathrm{d}n_t(\theta) - 1.\end{aligned}$$

Using that $m_t(\mathbf{X}_i|\theta)n_t(\theta) = m_t(\theta,\mathbf{X}_i)$ and $\nu_n(\mathbf{X}_i) = \frac{1}{n}$ for $i = 1,2,\ldots,n$, we obtain that:

$$I_t = n\int_{\mathbb{R}^d}\sum_{i=1}^n\frac{m_t^2(\theta,\mathbf{X}_i)}{n_t(\theta)}\mathrm{d}\theta - 1. \tag{5.4}$$

The next proposition then assesses how fast I_t decreases to 0 as $t \longrightarrow +\infty$.

Proposition 5.7. *Assume $\mathcal{H}_{\pi_0}(\ell_0)$ and for each $\mathbf{X}_i$, $\theta \to -\log p_\theta(\mathbf{X}_i)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$. Then, for any initial distribution n_0 and $t \geq 0$:*

$$I_t \leq I_0 e^{-2\alpha_n t} \leq (n-1)e^{-2\alpha_n t}.$$

Proof. Our starting point is Equation (5.4). The existence of $\partial_t\{I_t\}$ and the exchange of derivative and integral sign could be justified as in Section 5.4. We then compute its derivative with respect to t as follows:

$$\begin{aligned}\partial_t\{I_t\} &= 2n\int_{\mathbb{R}^d}\sum_{i=1}^n \frac{m_t(\theta,\mathbf{X}_i)}{n_t(\theta)}\partial_t\{m_t(\theta,\mathbf{X}_i)\}\mathrm{d}\theta - n\int_{\mathbb{R}^d}\sum_{i=1}^n \frac{m_t^2(\theta,\mathbf{X}_i)}{n_t^2(\theta)}\partial_t\{n_t(\theta)\}\mathrm{d}\theta \\ &= 2n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t(\mathbf{X}_i|\theta)\partial_t\{m_t(\theta,\mathbf{X}_i)\}\mathrm{d}\theta - n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t\{n_t(\theta)\}\mathrm{d}\theta.\end{aligned}$$

Using the Kolmogorov forward equation in the first line and $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ in the second one where $\mathcal{L}_1$ and $\mathcal{L}_2$ are defined in Equation (2.1), we have that:

$$\begin{aligned}\partial_t\{I_t\} &= 2n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}m_t(\mathbf{X}_i|\theta)\; m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta - n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t\{n_t(\theta)\}\mathrm{d}\theta \\ &= \underbrace{2n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}_1 m_t(\mathbf{X}_i|\theta)\; m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta}_{I_{3,t}} + \underbrace{2n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}_2 m_t(\mathbf{X}_i|\theta)\; m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta}_{I_{1,t}} \\ &\quad \underbrace{-n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t\{n_t(\theta)\}\mathrm{d}\theta}_{I_{2,t}}. \qquad (5.5)\end{aligned}$$

Then, $\partial_t\{I_t\}$ may be splitted into three terms that are studied separately.

- Study of $I_{1,t}$. We observe that for any i and $\theta \in \mathbb{R}^d$:

$$\mathcal{L}_2 m_t(\mathbf{X}_i|\theta) = \frac{\alpha_n}{n}\sum_{j=1}^n [m_t(\mathbf{X}_j|\theta) - m_t(\mathbf{X}_i|\theta)] = \frac{\alpha_n}{n} - \alpha_n m_t(\mathbf{X}_i|\theta). \qquad (5.6)$$

  We then use this last equation in the definition of $I_{1,t}$ and obtain that:

$$\begin{aligned}I_{1,t} &= 2\alpha_n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta - 2\alpha_n n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t(\mathbf{X}_i|\theta)m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta \\ &= 2\alpha_n - 2\alpha_n n\int_{\mathbb{R}^d}\sum_{i=1}^n \frac{m_t^2(\theta,\mathbf{X}_i)}{n_t(\theta)}\mathrm{d}\theta \\ &= -2\alpha_n I_t. \qquad (5.7)\end{aligned}$$

- Study of $I_{2,t}$. Using the definition of n_t, we get that:

$$\begin{aligned}I_{2,t} &= -n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t\{n_t(\theta)\}\mathrm{d}\theta \\ &= -n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t\left\{\sum_{j=1}^n m_t(\theta,\mathbf{X}_j)\right\}\mathrm{d}\theta \\ &= -n\int_{\mathbb{R}^d}\sum_{j=1}^n\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t\{m_t(\theta,\mathbf{X}_j)\}\mathrm{d}\theta \\ &= -n\int_{\mathbb{R}^d}\sum_{j=1}^n\left(\sum_{i=1}^n \mathcal{L}m_t^2(\mathbf{X}_i|\theta)\right) m_t(\theta,\mathbf{X}_j)\mathrm{d}\theta \\ &= -n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}m_t^2(\mathbf{X}_i|\theta)\mathrm{d}n_t(\theta).\end{aligned}$$

where we used the Kolmogorov forward equation in the fourth line and again the definition of n_t in the last line. Again, the decomposition $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ yields:

$$I_{2,t} = -n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta) \mathrm{d}n_t(\theta) - n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \mathcal{L}_2 m_t^2(\mathbf{X}_i|\theta) \mathrm{d}n_t(\theta).$$

We repeat some similar computations as those developed in Equation (5.6) to study the action of the jump component induced by $_2$ on $m_t^2(\mathbf{X}_i|\theta)$. Then, we deduce that:

$$\mathcal{L}_2 m_t^2(\mathbf{X}_i|\theta) = \frac{\alpha_n}{n} \sum_{k=1}^{n} m_t^2(\mathbf{X}_k|\theta) - \alpha_n m_t^2(\mathbf{X}_i|\theta).$$

We use this last equation to show that:

$$\begin{aligned}
I_{2,t} =& - n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta) \mathrm{d}n_t(\theta) - \alpha_n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \sum_{k=1}^{n} m_t^2(\mathbf{X}_k|\theta) \mathrm{d}n_t(\theta) \\
&+ \alpha_n n \int_{\mathbb{R}^d} \sum_{i=1}^{n} m_t^2(\mathbf{X}_i|\theta) \mathrm{d}n_t(\theta) \\
=& - n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta) \mathrm{d}n_t(\theta) - \alpha_n n \int_{\mathbb{R}^d} \sum_{k=1}^{n} m_t^2(\mathbf{X}_k|\theta) \mathrm{d}n_t(\theta) \\
&+ \alpha_n n \int_{\mathbb{R}^d} \sum_{i=1}^{n} m_t^2(\mathbf{X}_i|\theta) \mathrm{d}n_t(\theta) \\
=& - n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta) \mathrm{d}n_t(\theta).
\end{aligned}$$

- Study of $I_{2,t} + I_{3,t}$. We observe that this sum involves only $_1$, which was defined in Equation (2.1) . We first compute:

$$\mathcal{L}_1 m_t(\mathbf{X}_i|\theta) = -\langle \nabla_\theta U_{\mathbf{X}_i}(\theta), \nabla_\theta m_t(\mathbf{X}_i|\theta) \rangle + \Delta_\theta m_t(\mathbf{X}_i|\theta),$$

Then:

$$\begin{aligned}
I_{3,t} &= 2n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \mathcal{L}_1 m_t(\mathbf{X}_i|\theta)\ m_t(\theta, \mathbf{X}_i) \mathrm{d}\theta \\
&= 2n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \left[-\langle \nabla_\theta U_{\mathbf{X}_i}(\theta), \nabla_\theta m_t(\mathbf{X}_i|\theta) \rangle + \Delta_\theta m_t(\mathbf{X}_i|\theta) \right] m_t(\theta, \mathbf{X}_i) \mathrm{d}\theta.
\end{aligned}$$

Similarly, we deduce that:

$$\begin{aligned}
\mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta) =& - \langle \nabla_\theta U_{\mathbf{X}_i}(\theta), \nabla_\theta m_t^2(\mathbf{X}_i|\theta), \rangle + \Delta_\theta m_t^2(\mathbf{X}_i|\theta) \\
=& 2 m_t(\mathbf{X}_i|\theta) \left(-\langle \nabla_\theta U_{\mathbf{X}_i}(\theta), \nabla_\theta m_t(\mathbf{X}_i|\theta) \rangle + \Delta_\theta m_t(\mathbf{X}_i|\theta) \right) \\
&+ 2 \|\nabla_\theta m_t(\mathbf{X}_i|\theta)\|_2^2.
\end{aligned}$$

Using that $m_t(\mathbf{X}_i|\theta) n_t(\theta) = m_t(\theta, \mathbf{X}_i)$, we get that:

$$\begin{aligned}
I_{2,t} =& - n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta) \mathrm{d}n_t(\theta) \\
=& - 2n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \left[-\langle \nabla_\theta U_{\mathbf{X}_i}(\theta), \nabla_\theta m_t(\mathbf{X}_i|\theta) \rangle + \Delta_\theta m_t(\mathbf{X}_i|\theta) \right] m_t(\theta, \mathbf{X}_i) \mathrm{d}\theta \\
&- 2n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \|\nabla_\theta m_t(\mathbf{X}_i|\theta)\|_2^2 \mathrm{d}n_t(\theta).
\end{aligned}$$

It is immediate that:

$$I_{2,t} + I_{3,t} = -2n \int_{\mathbb{R}^d} \sum_{i=1}^{n} \|\nabla_\theta m_t(\mathbf{X}_i|\theta)\|_2^2 \mathrm{d}n_t(\theta) \leq 0.$$

Gathering this last inequality with (5.7) into Equation (5.5) yields:

$$\partial_t\{I_t\} \leq -2\alpha_n I_t.$$

We conclude with a direct application of the Gronwall lemma while observing that $I_0 \leq n-1$. □

5.6 From Poincaré inequality to WLSI

We will prove that if a measure m verifies a Poincaré inequality then it verifies a WLSI. Although this result is part of Proposition 3.1 in [9], we rewrite its proof to explicitly find a function φ. We recall the Remark 1.3 in [9]: by Rothaus lemma, for any function f:

$$Ent_m(f^2) \leq Ent_m(\tilde{f}^2) + 2Var_m(f),$$

where $\tilde{f} = f - m[f]$. Moreover, Popoviciu's inequality establishes that $Var_m(f) \leq \frac{1}{4}Osc^2(f)$. We also notice that $Osc(f) = 1$ implies $\sup \tilde{f}^2 \leq 1$ and then $Ent_m(\tilde{f}^2) \leq \frac{1}{e}$. Hence, by homogeneity:

$$Ent_m(f^2) \leq \left(\frac{1}{e} + \frac{1}{2}\right) Osc^2(f).$$

For $s \geq \frac{1}{e} + \frac{1}{2}$, we could take $\varphi(s)$ as a constant, as was mentioned in [9].

In order to recall a measure - capacity inequality and some results obtained in [4], we first define the capacity of a measurable set.

Definition 5.8 (Capacity)**.** *Let $\mathcal{A}$ and Ω be two measurable sets of $\mathbb{R}^d$ such that $\mathcal{A} \subset \Omega$, the capacity $Cap_m(\mathcal{A}, \Omega)$ is defined as*

$$Cap_m(\mathcal{A}, \Omega) = \inf\left\{\mathcal{E}_m(f), \mathbb{1}_{\mathcal{A}} \leq f \leq \mathbb{1}_\Omega\right\},$$

where f is a Lipschitz function on $\mathbb{R}^d$. If $m(\mathcal{A}) \leq \frac{1}{2}$, then we denote

$$Cap_m(\mathcal{A}) = \inf\left\{Cap_m(\mathcal{A}, \Omega), \mathcal{A} \subset \Omega, m(\Omega) \leq \frac{1}{2}\right\}.$$

The measure - capacity inequalities are a class of inequalities that, as their name indicates, involve the capacity of measurable sets. They are commonly used to describe some functional inequalities as Poincaré and LSI, for a more in-depth study we refer to [4]. So, in order to prove Proposition 3.6 we state the following lemma which was taken from [9] and it shows a sufficient condition to verify a WLSI.

Lemma 5.9. *Let $\phi : (0, +\infty) \to \mathbb{R}^+$ be a non-increasing function such that for every measurable subset $\mathcal{A}$ with $0 < m(\mathcal{A}) \leq 1/2$, one has*

$$\frac{m(\mathcal{A}) \log\left(1 + \frac{e^2}{m(\mathcal{A})}\right) - s}{\phi(s)} \leq Cap_m(\mathcal{A}), \quad \forall s > 0. \tag{5.8}$$

Then the measure m satisfies a WLSI with the function $\varphi(s) = 16\phi(3s/14)$ for $s > 0$.

From Lemma 5.9, the proof of Proposition 3.6 is reduced to finding a function $\phi(s)$ that satisfies inequality 5.8.

Proof of Proposition 3.6. Let $\mathcal{A}$ be a measurable subset with $0 < m(\mathcal{A}) \leq 1/2$. From Proposition 8.3.1 of [4], if a probability measure satisfies a Poincaré inequality with constant $C_{\mathrm{PI}}(m)$, then:

$$\frac{m(\mathcal{A})C_{\mathrm{PI}}(m)}{2} \leq Cap_m(\mathcal{A}).$$

Using the inequality above, a positive function ϕ verifies (5.8) if, for any $s > 0$,

$$2C_{\mathrm{PI}}(m)\left[\log\left(1+\frac{e^2}{m(A)}\right) - \frac{s}{m(A)}\right] \leq \phi(s).$$

Let us fix $s > 0$ small and define the function $g_s(x) = \log\left(1+\frac{e^2}{x}\right) - \frac{s}{x}$, for $0 < x \leq \frac{1}{2}$. Using that g_s reaches its maximum at

$$x_{\max} = \begin{cases} \frac{se^2}{e^2-s}, & 0 < s \leq \frac{e^2}{2e^2+1} \\ \frac{1}{2}, & s > \frac{e^2}{2e^2+1} \end{cases},$$

then $g_s(x) \leq g_s(x_{\max})$ where

$$g_s(x_{\max}) = \begin{cases} \log\left(\frac{1}{s}\right) + \frac{s}{e^2} + 1, & 0 < s \leq \frac{e^2}{2e^2+1} \\ \log(1+2e^2) - 2s, & s > \frac{e^2}{2e^2+1} \end{cases}.$$

We define $\phi(s) = 2C_{\mathrm{PI}}(m)g_s(x_{\max})$ if $s > e^{-1}$, while we choose

$$\phi(s) = \mathfrak{a}C_{\mathrm{PI}}(m)\log(1/s) \geq 2C_{\mathrm{PI}}(m)g_s(x_{\max}),$$

for any $0 < s \leq e^{-1}$, where $\mathfrak{a} = 2 + e^{-3}$. □

5.7 Moments upper bounds

Proposition 5.10. *Assume $\mathcal{H}_{\mathbf{n_0}}(L, \ell_0)$, $\mathcal{H}_{\pi_\mathbf{0}}(\ell_0)$, $\mathcal{H}_{\min}$ and that for each $\mathbf{X}_i$, $\theta \mapsto -\log p_\theta(\mathbf{X}_i)$ satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$. Then:*

i) Three positive constants C_1, C_2 and C_3, independent from n and d, exist such that for any $t > 0$:

$$\mathbb{E}_{n_t}\left[e^{\frac{(1+r)n\mathfrak{c}}{16}\frac{1}{1+r}(\|\theta_t\|_2^2+1)^{\frac{1}{1+r}}}\right] \leq C_1\left(d\log^{2\beta}(n)\right)^r e^{C_2 nd\log^{2\beta}(n)} + C_3^d e^{\frac{(1+r)n\mathfrak{c}}{16}\frac{1}{1+r}}.$$

ii) For any $t > 0$ and for any $\alpha \geq 1$:

$$\mathbb{E}_{n_t}[U^\alpha_{\nu_n}(\theta_t)] \lesssim_{uc} n^\alpha\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)}.$$

Proof of i). The proof is based on a Lyapunov argument. Consider the twice differentiable function:

$$f(\theta) = \exp\left(\frac{a}{2}(\|\theta\|_2^2+1)^\rho\right), \quad \theta \in \mathbb{R}^d,$$

where $0 < \rho < 1$ and $a > 0$ are two constant to fix later on. For any $\theta \in \mathbb{R}^d$, the gradient of f is:

$$\nabla f(\theta) = a\rho(\|\theta\|_2^2+1)^{\rho-1}f(\theta)\theta,$$

and the Laplacian of f satisfies the following inequality:

$$\begin{aligned}\Delta f(\theta) &= a\rho(\|\theta\|_2^2+1)^{\rho-2}f(\theta)\left[a\rho(\|\theta\|_2^2+1)^\rho\|\theta\|_2^2 + d(\|\theta\|_2^2+1) + 2(\rho-1)\|\theta\|_2^2\right] \\ &\leq a\rho(\|\theta\|_2^2+1)^{\rho-1}f(\theta)\left[a\rho(\|\theta\|_2^2+1)^\rho + d\right],\end{aligned}$$

where we used that $0 < \rho < 1$ and $\|\theta\|_2^2 \leq \|\theta\|_2^2 + 1$.

We then deduce that for any $t > 0$ and $\theta \in \mathbb{R}^d$:

$$\begin{aligned}\frac{\mathcal{G}_t f(\theta)}{f(\theta)} &= \frac{1}{f(\theta)}\left[-\sum_{i=1}^n \langle \nabla U_{\mathbf{X}_i}, \nabla f(\theta)\rangle m_t(\mathbf{X}_i|\theta) + \Delta f(\theta)\right] \\ &\leq a\rho(\|\theta\|_2^2+1)^{\rho-1}\Big[-\sum_{i=1}^n \langle \theta, \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle m_t(\mathbf{X}_i|\theta) + a\rho(\|\theta\|_2^2+1)^\rho + d\Big] \\ &\leq a\rho(\|\theta\|_2^2+1)^{\rho-1}\left[-\sum_{i=1}^n \left(U_{\mathbf{X}_i}(\theta) - U_{\mathbf{X}_i}(0)\right) m_t(\mathbf{X}_i|\theta) + a\rho(\|\theta\|_2^2+1)^\rho + d\right], \quad (5.9)\end{aligned}$$

where we used the convexity of U_x for any position x.

First considerations: For any i, we denote by $\theta_i = \arg\min U_{\mathbf{X}_i}$ and from hypothesis $\mathcal{H}_{\min}$, there exist two positive constants $\mathcal{K}_1$ and $\mathcal{K}_2$ independent on n and d such that:

$$\max_i \|\theta_i\|_2^2 \leq \mathcal{K}_1 d\log^{2\beta}(n) \quad \text{and} \quad \max_i U_{\mathbf{X}_i}(\theta_i) \leq \mathcal{K}_2 d\log^{2\beta}(n).$$

So, in order to lower bound the term $\sum_{i=1}^n \left(U_{\mathbf{X}_i}(\theta) - U_{\mathbf{X}_i}(0)\right) m_t(\mathbf{X}_i|\theta)$, let us establish the bounds of $U_{\mathbf{X}_i}(\theta)$ and $U_{\mathbf{X}_i}(0)$ separately.

• In Proposition 1.4 we proved that each non-negative function $U_{\mathbf{X}_i}$ satisfies a $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL+\ell_0)$-condition, then we are able to apply Proposition 5.2 and obtain that for any $U_{\mathbf{X}_i}$:

$$U_{\mathbf{X}_i}(\theta) \geq n\left[\frac{(1+r)\mathfrak{c}}{2}\right]^{\frac{1}{1+r}} \|\theta - \theta_i\|_2^{\frac{2}{1+r}}.$$

Since $\frac{2}{1+r} > 1$, the Jensen inequality yields $(u+v)^{\frac{2}{1+r}} \leq 2^{\frac{1-r}{1+r}}\left[u^{\frac{2}{1+r}} + v^{\frac{2}{1+r}}\right]$, for all $(u,v) \in \mathbb{R}_+^2$ and we deduce that for any $\theta \in \mathbb{R}^d$:

$$\|\theta - \theta_i\|_2^{\frac{2}{1+r}} \geq 2^{\frac{r-1}{1+r}}\|\theta\|_2^{\frac{2}{1+r}} - \|\theta_i\|_2^{\frac{2}{1+r}} \geq 2^{\frac{r-1}{1+r}}\|\theta\|_2^{\frac{2}{1+r}} - \left(\mathcal{K}_1 d\log^{2\beta}(n)\right)^{\frac{1}{1+r}}.$$

Then we use this inequality to obtain a lower bound of $U_{\mathbf{X}_i}$:

$$\begin{aligned}U_{\mathbf{X}_i}(\theta) &\geq 2n\left[\frac{(1+r)\mathfrak{c}}{8}\right]^{\frac{1}{1+r}}\|\theta\|_2^{\frac{2}{1+r}} - n\left[\frac{(1+r)\mathfrak{c}}{2}\right]^{\frac{1}{1+r}}\left(\mathcal{K}_1 d\log^{2\beta}(n)\right)^{\frac{1}{1+r}} \\ &\geq \frac{n\mathfrak{c}^{\frac{1}{1+r}}}{4}\|\theta\|_2^{\frac{2}{1+r}} - n\mathfrak{c}^{\frac{1}{1+r}}\left(\mathcal{K}_1 d\log^{2\beta}(n)\right)^{\frac{1}{1+r}},\end{aligned}$$

where we used some uniform upper bounds when $r \in [0,1)$.

• An upper bound of $\max_i U_{\mathbf{X}_i}(0)$ comes from Proposition 1.4 and 5.2 as follows:

$$U_{\mathbf{X}_i}(0) \leq U_{\mathbf{X}_i}(\theta_i) + \frac{nL+\ell_0}{2}\|\theta_i\|_2^2 \leq \left(\mathcal{K}_2 + \frac{\mathcal{K}_1(nL+\ell_0)}{2}\right) d\log^{2\beta}(n) \leq \mathcal{K}nd\log^{2\beta}(n),$$

where $\mathcal{K}$ is a constant independent of n and d and could proportionally change from line to line.

Using the previous bounds and the fact that $\sum_{i=1}^n m_t(\mathbf{X}_i|\theta) = 1$, it yields:

$$\sum_{i=1}^n \left(U_{\mathbf{X}_i}(\theta) - U_{\mathbf{X}_i}(0)\right) m_t(\mathbf{X}_i|\theta) \geq \frac{n\mathfrak{c}^{\frac{1}{1+r}}}{4}\|\theta\|_2^{\frac{2}{1+r}} - \mathcal{K}nd\log^{2\beta}(n).$$

We omit the term $n\mathfrak{c}^{\frac{1}{1+r}}\left(\mathcal{K}_1 d\log^{2\beta}(n)\right)^{\frac{1}{1+r}}$ since it could be bounded by $\mathcal{K}nd\log^{2\beta}(n)$ if we proportionally change the value of $\mathcal{K}$. Returning to inequality (5.9), we now get:

$$\frac{\mathcal{G}_t f(\theta)}{f(\theta)} \leq a\rho(\|\theta\|_2^2+1)^{\rho-1}\left[-\frac{n\mathfrak{c}^{\frac{1}{1+r}}}{4}\|\theta\|_2^{\frac{2}{1+r}} + \mathcal{K}nd\log^{2\beta}(n) + a\rho(\|\theta\|_2^2+1)^\rho\right].$$

Values of a and ρ: We choose $\rho = \frac{1}{1+r}$ and deduce that:

$$\frac{\mathcal{G}_t f(\theta)}{f(\theta)} \leq \frac{a}{1+r}(\|\theta\|_2^2+1)^{-\frac{r}{1+r}}\left[-\frac{n\mathfrak{c}^{\frac{1}{1+r}}}{4}\|\theta\|_2^{\frac{2}{1+r}} + \mathcal{K}nd\log^{2\beta}(n) + \frac{a}{1+r}(\|\theta\|_2^2+1)^{\frac{1}{1+r}}\right],$$

using the inequality $(\|\theta\|_2^2+1)^{\frac{1}{1+r}} \leq \|\theta\|_2^{\frac{2}{1+r}} + 1$, then:

$$\frac{\mathcal{G}_t f(\theta)}{f(\theta)} \leq \frac{a}{1+r}(\|\theta\|_2^2+1)^{-\frac{r}{1+r}}\left[-\left(\frac{n\mathfrak{c}^{\frac{1}{1+r}}}{4} - \frac{a}{1+r}\right)\|\theta\|_2^{\frac{2}{1+r}} + \mathcal{K}nd\log^{2\beta}(n)\right].$$

Let us fix $a = \frac{n(1+r)\mathfrak{c}^{\frac{1}{1+r}}}{8}$, then for any $t>0$ and $\theta \in \mathbb{R}^d$,

$$\frac{\mathcal{G}_t f(\theta)}{f(\theta)} \leq -\frac{n^2\mathfrak{c}^{\frac{2}{1+r}}}{64}(\|\theta\|_2^2+1)^{-\frac{r}{1+r}}\left[\|\theta\|_2^{\frac{2}{1+r}} - \mathcal{K}d\log^{2\beta}(n)\right]. \tag{5.10}$$

Lyapunov contraction: We study two complementary situations and below, we denote by $K_{n,d}$ the radius of the key compact set involved by the previous Lyapunov contraction:

$$K_{n,d}^{\frac{2}{1+r}} = Cd\log^{2\beta}(n).$$

• When $\|\theta\|_2$ is large enough ($\|\theta\|_2 \geq K_{n,d}$), we use the fact that for any two fixed constants $k \geq 0$ and $0 \leq r < 1$, the function $x \mapsto -(x+1)^{-\frac{r}{1+r}}(x^{\frac{1}{1+r}} - k)$, for $x \geq k^{1+r}$, is decreasing and we observe that a large enough $C>0$ independent of n and d exists such that:

$$\begin{aligned}\|\theta\|_2^{\frac{2}{1+r}} \geq Cd\log^{2\beta}(n) &\Longrightarrow \frac{\mathcal{G}_t f(\theta)}{f(\theta)} \leq -\frac{n^2\mathfrak{c}^{\frac{2}{1+r}}}{64}\frac{(C-\mathcal{K})}{2C^r}(d\log^{2\beta}(n))^{1-r}\\ &\Longrightarrow \frac{\mathcal{G}_t f(\theta)}{f(\theta)} \leq -\frac{n^2\mathfrak{c}^{\frac{2}{1+r}}}{128}(d\log^{2\beta}(n))^{1-r} = -a_{n,d}.\end{aligned} \tag{5.11}$$

• When $\|\theta\|_2$ is upper bounded ($\|\theta\|_2 \leq K_{n,d}$), we use the upper bound stated in Equation (5.10) and obtain that a universal C_1 (whose value may change from line to line) exists such that :

$$\begin{aligned}\|\theta\|_2^{\frac{2}{1+r}} \leq Cd\log^{2\beta}(n) &\Longrightarrow \mathcal{G}_t f(\theta) \leq C_1 n^2 d\log^{2\beta}(n) f(\theta)\\ &\Longrightarrow \mathcal{G}_t f(\theta) \leq C_1 n^2 d\log^{2\beta}(n)\exp\left(C\mathfrak{c}^{\frac{1}{1+r}}nd\log^{2\beta}(n)\right)\\ &\Longrightarrow \mathcal{G}_t f(\theta) \leq b_{n,d}e^{\delta_{n,d}},\end{aligned} \tag{5.12}$$

where we denoted $b_{n,d} = C_1 n^2 d\log^{2\beta}(n)$ and $\delta_{n,d} = C\mathfrak{c}^{\frac{1}{1+r}}nd\log^{2\beta}(n)$.

We then use Equations (5.11) and (5.12) as follows. We define the function $\psi_{n,d}$ as $\psi_{n,d}(t) = \mathbb{E}_{n_t}[f(\theta_t)]$. The existence of $\partial_t\{\psi_{n,d}(t)\}$ and the following equality are justified as in Section 5.4:

$$\begin{aligned}\partial_t\{\psi_{n,d}(t)\} &= \mathbb{E}_{n_t}[\mathcal{G}_t f(\theta_t)] = \mathbb{E}_{n_t}\left[\mathcal{G}_t f(\theta_t)\left(\mathbb{1}_{\|\theta_t\|_2 \geq K_{n,d}} + \mathbb{1}_{\|\theta_t\|_2 \leq K_{n,d}}\right)\right]\\ &\leq \mathbb{E}_{n_t}\left[-a_{n,d}f(\theta_t)\mathbb{1}_{\|\theta_t\|_2 \geq K_{n,d}} + b_{n,d}e^{\delta_{n,d}}\mathbb{1}_{\|\theta_t\|_2 \leq K_{n,d}}\right]\\ &\leq -a_{n,d}\psi_{n,d}(t) + a_{n,d}\sup_{\|\theta\|_2 \leq K_{n,d}} f(\theta) + b_{n,d}e^{\delta_{n,d}}\\ &\leq -a_{n,d}\psi_{n,d}(t) + (a_{n,d}+b_{n,d})e^{\delta_{n,d}}.\end{aligned}$$

We apply the Gronwall Lemma and obtain that:

$$\forall t>0, \qquad \psi_{n,d}(t) \leq \left(1+\frac{b_{n,d}}{a_{n,d}}\right) e^{\delta_{n,d}} + \psi_{n,d}(0) e^{-a_{n,d}t}. \tag{5.13}$$

From the $\mathcal{H}_{\mathbf{n_0}}(L,\ell_0)$ hypothesis, n_0 is a Gaussian distribution, the we find an upper bound of $\psi_{n,d}(0) = \mathbb{E}_{n_0}[f(\theta_0)] = \int_{\mathbb{R}^d} f(\theta)\mathrm{d}n_0(\theta)$ as follows:

$$\psi_{n,d}(0) = \left(2\pi\sigma^2\right)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{\frac{a}{2}\left(\|\theta\|_2^2+1\right)^{\frac{1}{1+r}} - \frac{\|\theta\|_2^2}{2\sigma^2}} \mathrm{d}\theta \leq \left(2\pi\sigma^2\right)^{-\frac{d}{2}} e^{\frac{a}{2}} \int_{\mathbb{R}^d} e^{-\frac{\|\theta\|_2^2}{2}\left(\frac{1}{\sigma^2}-a\right)} \mathrm{d}\theta,$$

if $\sigma^2 < \frac{1}{a} = \frac{8}{n(1+r)\mathfrak{c}^{\frac{1}{1+r}}}$ then the integral above is finite. From Remark 1.2, we verify that $c_2 < 1 \leq \frac{8L}{(1+r)\mathfrak{c}^{\frac{1}{1+r}}}$, which guarantees that $\sigma^2 < \frac{1}{a}$ and:

$$\psi_{n,d}(0) \leq \left(1-a\sigma^2\right)^{-\frac{d}{2}} e^{\frac{a}{2}} \leq C_3^d e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}},$$

where C_3 is a constant independent from n and d.

Finally, using the value of $a_{n,d}$ and $b_{n,d}$ in (5.13), we deduce that for any $t>0$,

$$\mathbb{E}_{n_t}\left[e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}\left(\|\theta_t\|_2^2+1\right)^{\frac{1}{1+r}}}\right] \leq C_1\left(d\log^{2\beta}(n)\right)^r e^{C_2 nd\log^{2\beta}(n)} + C_3^d e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}}.$$

where C_2 is another universal constant, which concludes the proof. □

Proof of ii). We consider $\alpha>1$ and below, $C>0$ refers to a constant independent from n and d, whose value may change from line to line. Our starting point is the upper bound of the exponential moments obtained in i). Proposition 1.4 shows that U_{ν_n} satisfies $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}n^{1+r}, nL+\ell_0)$, then thanks to Proposition 5.2:

$$\begin{aligned}\mathbb{E}_{n_t}[U^\alpha_{\nu_n}(\theta_t)] &\leq \mathbb{E}_{n_t}\left[\left(U_{\nu_n}(\theta_n^*) + Cn\|\theta_t-\theta_n^*\|_2^2\right)^\alpha\right] \\ &\leq \mathbb{E}_{n_t}\left[\left(U_{\nu_n}(\theta_n^*) + Cn\|\theta_n^*\|_2^2 + Cn\|\theta_t\|_2^2\right)^\alpha\right],\end{aligned}$$

where $\theta_n^* = \arg\min U_{\nu_n}$.

By using Proposition 5.3 and the inequality derived from the Jensen inequality $(a+b)^\alpha \leq c_\alpha(a^\alpha+b^\alpha)$ for $(a,b)\in\mathbb{R}_+^2$ and $\alpha\geq 1$, we obtain that:

$$\begin{aligned}\mathbb{E}_{n_t}[U^\alpha_{\nu_n}(\theta_t)] &\leq C\mathbb{E}_{n_t}\left[\left(nd\log^{2\beta}(n) + n\left(d\log^{2\beta}(n)\right)^{1+r} + n\|\theta_t\|_2^2\right)^\alpha\right] \\ &\leq Cn^\alpha\left[\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)} + \mathbb{E}_{n_t}\left(\|\theta_t\|_2^{2\alpha}\right)\right].\end{aligned} \tag{5.14}$$

Let us focus on the moment of order 2α. It could be rewritten as:

$$\begin{aligned}\mathbb{E}_{n_t}\left(\|\theta_t\|_2^{2\alpha}\right) &= k^{-\alpha(1+r)}\mathbb{E}_{n_t}\left(\log^{\alpha(1+r)}\left(e^{k\|\theta_t\|_2^{\frac{2}{1+r}}}\right)\right) \\ &\leq k^{-\alpha(1+r)}\mathbb{E}_{n_t}\left(\log^{\alpha(1+r)}\left(e^{\alpha(1+r)-1+k\|\theta_t\|_2^{\frac{2}{1+r}}}\right)\right).\end{aligned}$$

The Jensen inequality and the concavity of $x\mapsto\log^p(x)$ on $[e^{p-1},+\infty[$ when $p\geq 1$ yield:

$$\begin{aligned}\mathbb{E}_{n_t}\left(\|\theta_t\|_2^{2\alpha}\right) &\leq k^{-\alpha(1+r)}\log^{\alpha(1+r)}\left(\mathbb{E}_{n_t}\left(e^{\alpha(1+r)-1+k\|\theta_t\|_2^{\frac{2}{1+r}}}\right)\right) \\ &\leq k^{-\alpha(1+r)}\left[\alpha(1+r)-1+\log\left(\mathbb{E}_{n_t}\left(e^{k\|\theta_t\|_2^{\frac{2}{1+r}}}\right)\right)\right]^{\alpha(1+r)} \\ &\leq k^{-\alpha(1+r)}\left[\alpha(1+r)-1+\log\left(\mathbb{E}_{n_t}\left(e^{k(\|\theta_t\|_2^2+1)^{\frac{1}{1+r}}}\right)\right)\right]^{\alpha(1+r)},\end{aligned}$$

where we used in the last inequality that $\|\theta\|_2^2 \leq \|\theta\|_2^2 + 1$.

We then apply $i)$ in Proposition 5.10, we choose $k = \frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}$ and obtain that:

$$\begin{aligned}
\mathbb{E}_{n_t}\left(\|\theta_t\|_2^{2\alpha}\right) &\leq \frac{C}{n^{\alpha(1+r)}}\left[1+\log\left(\mathbb{E}_{n_t}\left(e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}(\|\theta_t\|_2^2+1)^{\frac{1}{1+r}}}\right)\right)\right]^{\alpha(1+r)} \\
&\leq \frac{C}{n^{\alpha(1+r)}}\left[1+\log\left[C_1\left(d\log^{2\beta}(n)\right)^r e^{C_2 nd\log^{2\beta}(n)} + C_3^d e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}}\right]\right]^{\alpha(1+r)} \\
&\leq C\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)}, \qquad (5.15)
\end{aligned}$$

where we used in the previous lines simple algebra and $\log(a+b) \leq \log(2)+\log(a)+\log(b)$ when $a \geq 1$ and $b \geq 1$. Replacing (5.15) in (5.14), we conclude that:

$$\mathbb{E}_{n_t}[U_{\nu_n}^{\alpha}(\theta_t)] \leq Cn^{\alpha}\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)}.$$

□

References

[1] Altschuler, J. M. and Talwar, K. : Resolving the mixing time of the Langevin algorithm to its stationary distribution for log-concave sampling. *Conference on Learning Theory*. Proceedings of Machine Learning, (2023), 195:1-2.

[2] Bakry, D. and Cattiaux, P. and Guillin, A. : Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *Journal of Functional Analysis* **254**, 3, (2008), 727–759.

[3] Bakry, D. and Emery, M. : Diffusions hypercontractives. *Séminaire de probabilités* **1123**, XIX, (1985), 177–206.

[4] Bakry, D. and Gentil, I. and Ledoux, M. : Analysis and geometry of Markov diffusion operators. *Springer*. **103**, (2014).

[5] Balasubramanian, K. and Chewi, S. and Erdogdu, M. and Salim, A. and Zhang, S.: Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. *Conference on Learning Theory* **2890–2923**, (2022).

[6] Bobkov, S. G. : Isoperimetric and analytic Inequalities for log-concave probability measures. *Annals of Probability* **27**, (1999), 1903–1921.

[7] Bolte, J. and Daniilidis, A. and Ley, O. and Mazet, L. : Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.* **362**, (2010), 3319–3363.

[8] Cattiaux, P. and Fathi, M. and Guillin, A. : Self-improvement of the Bakry-Emery criterion for Poincaré inequalities and Wasserstein contraction using variable curvature bounds. *Journal de Mathématiques Pures et Appliquées*, (2022).

[9] Cattiaux, P. and Gentil, I. and Guillin, A. : Weak logarithmic Sobolev inequalities and entropic convergence. *Probability Theory and Related Fields* **139**, 3, (2007), 563–603.

[10] Cattiaux, P. and Guillin, A. : Hitting times, functional inequalities, Lyapunov conditions and uniform ergodicity. *Journal of Functional Analysis* **272**, 6, (2017), 2361–2391.

[11] Cattiaux, P. and Guillin, A. and Wang, F. and Wu, L. : Lyapunov conditions for Super Poincaré inequalities. *Journal of Functional Analysis.* **256**, 6, (2009), 1821–1841.

[12] Chewi, S. and Erdogdu, M. A. and Li, M. B. and Shen, R. and Zhang, M. : Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. *Conference on Learning Theory*. Proceedings of Machine Learning, (2022), **178**:1-2.

[13] Chewi, S. and Lu, C. and Ahn, K. and Cheng, X. and Le Gouic, T. and Rigollet, P. : Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. *Conference on Learning Theory*. Proceedings of Machine Learning Research (2021), **134** 1–41.

[14] Chiang, T. and Hwang, C. and Sheu, S. J. : Diffusion for Global Optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, **25**, 3, (1987), 737–753.

[15] Dalalyan, A. : Theoretical guarantees for approximate sampling from a smooth and log-concave density. *Journal of the Royal Statistical Society B*,**79**, (2017), 651–676.

[16] Dalalyan, A. and Karagulyan, A. : User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, **129**, 12, (2019), 5278–5311.

[17] Dalalyan, A. and Karagulyan, A. and Riou-Durand, L. : Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *Journal of Machine Learning Research*, **23**, 235, (2022), 1–38.

[18] Dalalyan, A. and Riou-Durand, L. : On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, **26**, 3, (2020), 1956–1988.

[19] Dalalyan, A. and Tsybakov, A. : Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.* , **78**, 5, (2012), 1423–1443.

[20] Das, A. and Nagaraj, D. M. and Raj, A. : Utilising the CLT structure in stochastic gradient based sampling: Improved analysis and faster algorithms. *Conference on Learning Theory*. Proceedings of Machine Learning Research, (2023), **195**, 1–58.

[21] Ding, Z. and Li, Q. and Lu, J., and Wright, S. J. : Random coordinate Langevin Monte Carlo. *Conference on Learning Theory*. Proceedings of Machine Learning Research, (2021), **134**, 1–28.

[22] Durmus, A. and Majewski, S. and Miasojedow, B. : Analysis of Langevin Monte Carlo via convex optimization, *Journal of Machine Learning Research*, **20**, 73, (2019), 1–46.

[23] Durmus, A. and Moulines, E. : High-dimensional Bayesian inference via the unadjusted Langevin algorithm, *Bernoulli*, **25**, 4A, (2019), 2854–2882.

[24] Erdogdu, M. A., and Hosseinzadeh, R. : On the convergence of Langevin Monte Carlo: The interplay between tail growth and smoothness. *Conference on Learning Theory*. Proceedings of Machine Learning Research (2021), **134**, 1–47.

[25] Ethier, S. N. and Kurtz, T. G. : Markov processes – characterization and convergence, *John Wiley & Sons Inc.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, New York, (1986).

[26] Freidlin, M. and Wentzell, A. : Random perturbations of dynamical systems, *Springer Verlag*, (1984).

[27] Gadat, S. and Gavra, I. and Risser, L. : How to calculate the barycenter of a weighted graph. *Mathematics of Operation Research*, **43**, 4, (2018).

[28] Gadat, S. and Panloup, F. : Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. *Stochastic Processes and their Applications*, **156**, (2022), 312–348.

[29] Gadat, S. and Panloup, F. and Pellegrini, C. : On the cost of Bayesian posterior mean strategy for log-concave models. *Preprint*, (2022).

[30] Gadat, S. and Panloup, F. and Pellegrini, C. : Large deviation principle for invariant distributions of memory gradient diffusions. *Electronic Journal of Probability*, **81**, (2013), 1–34.

[31] Gramacy, R. B. and Polson, N. G. : Simulation-based Regularized Logistic Regression. *Bayesian Analysis*, **7**, 3, (2012), 567–590.

[32] Gross, L. : Logarithmic Sobolev inequalities. *American Journal of Mathematics*, **4**, 97, (1975), 1061–1083.

[33] Hajeck, B.: Cooling schedules for optimal annealing. *Mathematics of Operation Research*, **12**, 2, (1988), 311–329.

[34] Held, L. and Holmes, C. C. : Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, **1**, 1, (2006), 145–168.

[35] Holley, R. and Stroock, D. : Simulated annealing via Sobolev inequalities. *Communications in Mathematical Physics* **115**, 4, (1988), 553–569.

[36] Hörmander, L. : Hypoelliptic second order differential equations. *Acta Mathematica* **119**, (1967), 147–171.

[37] Khasminskii , R. : Stochastic stability of differential equations. *Stochastic Modelling and Applied Probability, Springer*, (2012).

[38] Kurdyka, K. : On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)* **48**, 3, (1998), 769–783.

[39] Kusuoka, S. and Stroock, D. : Applications of the Malliavin calculus, Part I. *Stochastic Analysis*. Elsevier **32**, North-Holland Mathematical Library, (1984), 271–306.

[40] Łojasiewicz, S. : Une propriété topologique des sous-ensembles analytiques réels. *Editions du centre National de la Recherche Scientifique, Paris, Les Équations aux Dérivées Partielles*. (1963), 87–89.

[41] Ma, Y. and Chen, Y. and Jin, C. and Flammarion, N. and Jordan, M. I : Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences* **116**, 42, (2019), 20881–20885.

[42] Meyn, S. and Tweedie, R. : Markov chains and stochastic stability. *Springer Science & Business Media*. (2012).

[43] Miclo, L. : Recuit simulé sur $\mathbb{R}^n$. Étude de l'évolution de l'énergie libre. *Annales de l'IHP Probabilités et statistiques* **28**, 2, (1992), 235–266.

[44] Mou, W. and Flammarion, N. and Wainwright, M. J. and Bartlett, P. L. : Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity. *Bernoulli* **28**, 3, (2022), 1577–1601.

[45] Park, M. Y. and Hastie, T. : L 1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69**, 4, (2007), 659–677.

[46] Raginsky, M. and Rakhlin, A. and Telgarsky, M. : Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *Proceedings of Machine Learning Research*, **65**, (2017), 1–30.

[47] Robbins, H. and Monro, S. : A stochastic approximation method. *The Annals of Mathematical Statistics* **22**, 3, (1951): 400–407.

[48] Roberts, G. O. and Tweedie, R. L. : Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 4, (1996): 341–363.

[49] Stroock, D. W. and Varadhan, SR S. : Multidimensional diffusion processes. *Springer Science & Business Media*, **233**, (1997).

[50] Vempala, S. S. and Wibisono, A. : Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices. *Neural Information Processing Systems*, (2019).

[51] Wang, F. : Functional inequalities for empty essential spectrum. *Journal of Functional Analysis* **170**, 1, (2000), 219–245.

[52] Wang, B. and Zou, D. and Gu, Q. and Osher, S. J. : Laplacian smoothing stochastic gradient Markov chain Monte Carlo. *SIAM Journal on Scientific Computing* **43**, 1, (2021), A26–A53.

[53] Welling, M. and Teh, Y. W. : Bayesian learning via stochastic gradient Langevin dynamics. *International Conference on Machine Learning* **28**, 3, (2011), 681–688.

[54] Wibisono, A. and Yang, K.Y.: Convergence in KL Divergence of the inexact Langevin algorithm with application to score-based generative models. *NeurIPS, Workshop on Score-Based Methods* (2022).

[55] Xu, P. and Chen, J. and Zou, D. and Gu, Q. : Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *Conference on Neural Information Processing Systems.*. Curran Associates Inc. (2018), 3126—3137.

[56] Zhang, K. S., Peyré, G., Fadili, J., and Pereyra, M. : Wasserstein control of mirror Langevin Monte Carlo. *Conference on Learning Theory*. Proceedings of Machine Learning Research, **125** (2020), 1–28.

[57] Zou, D. and Xu, P. and Gu, Q. : Faster convergence of stochastic gradient Langevin dynamics for non-log-concave sampling. *Conference on Uncertainty in Artificial Intelligence.* Proceedings of Machine Learning Research, **161** (2021), 1152–1162.