Tamed Langevin sampling under weaker conditions

Iosif Lytras, Panayotis Mertikopoulos

TAMED LANGEVIN SAMPLING UNDER WEAKER CONDITIONS

IOSIF LYTRAS[c,*,◇] AND PANAYOTIS MERTIKOPOULOS[♯]

Abstract. Motivated by applications to deep learning which often fail standard Lipschitz smoothness requirements, we examine the problem of sampling from distributions that are not log-concave and are only weakly dissipative, with log-gradients allowed to grow superlinearly at infinity. In terms of structure, we only assume that the target distribution satisfies either a Log-Sobolev or a Poincare inequality and a local Lipschitz smoothness assumption with modulus growing possibly polynomially at infinity. This set of assumptions greatly exceeds the operational limits of the "vanilla" ULA, making sampling from such distributions a highly involved affair. To account for this, we introduce a taming scheme which is tailored to the growth and decay properties of the target distribution, and we provide explicit non-asymptotic guarantees for the proposed sampler in terms of the KL divergence, total variation, and Wasserstein distance to the target distribution.

1. Introduction

A broad array of modern and emerging machine learning architectures relies on being able to sample efficiently from a target distribution π on $\mathbb{R}^d$, typically expressed in Gibbs form as $\pi(x) \propto \exp(-u(x))$ for some potential function $u\colon \mathbb{R}^d \to \mathbb{R}$. Under suitable assumptions for u, this distribution arises naturally as the invariant measure of the Langevin stochastic differential equation

$$dX_t = -\nabla u(X_t)\,dt + \sqrt{2}\,dB_t \tag{LSDE}$$

where B_t is a canonical Wiener process (Brownian motion) in $\mathbb{R}^d$ with unit volatility. Based on this key property of (LSDE), one of the most – if not *the* most – widely used algorithmic schemes for sampling from π is the so-called *unadjusted Langevin algorithm* (ULA), given in recursive form as

$$\theta_{n+1}^{\mathrm{ULA}} = \theta_n^{\mathrm{ULA}} - \lambda h\big(\theta_n^{\mathrm{ULA}}\big) + \sqrt{2\lambda}\xi_{n+1} \tag{ULA}$$

where $\theta_n \in \mathbb{R}^d$, $n = 1, 2, \dots$, is the algorithm's state variable, ξ_n is an independent and identically distributed (i.i.d.) sequence of standard d-dimensional random variables with unit covariance, $\lambda > 0$ is a step-size parameter, and $h := \nabla u$ denotes the gradient of u. The idea behind (ULA) is that θ_n^{ULA} can be seen as an Euler-Maruyama discretization of (LSDE) so, for sufficiently large n and small enough λ, θ_n^{ULA} will be distributed according to some approximate version of the invariant measure of (LSDE), which is precisely the target distribution π.

This simple idea has generated a vast corpus of literature and techniques for proving the non-asymptotic convergence rate of (ULA) in different probability metrics, the most popular ones being the Wasserstein and total variation distances, as well as the Kullback–Leibler (KL) and/or Rényi divergence. Much of this literature has focused on the case where the target distribution π is log-concave and has Lipschitz continuous log-gradients, corresponding respectively to convexity and Lipschitz smoothness of the potential u; for some representative

2010 *Mathematics Subject Classification.* Primary 65C05, 60H10; secondary 68Q32.

Key words and phrases. Langevin sampling; taming; isoperimetry; Poincaré inequality; log-Sobolev inequality; weak dissipativity.

recent works, see Dalalyan [12], Durmus & Moulines [13, 14], Barkhagen et al. [5] and references therein.

Beyond these works, especially when the target distribution is multimodal, there has been significant effort to relax the (strong) convexity requirement for u by means of a combination of "convexity at infinity" and "dissipativity" assumptions – that is, convexity outside a compact set, and a drift coercivity condition of the form $\langle h(x), x\rangle = \Omega(|x|^2)$ for the drift $h = \nabla u$ of (LSDE) respectively, cf. Cheng et al. [10], Majka et al. [23], Erdogdu et al. [16], as well as a recent thread of results on the related stochastic gradient Langevin dynamics (SGLD) scheme by Raginsky et al. [30], Chau et al. [9] and Zhang et al. [34].

At the same time, building on an important insight of Vempala & Wibisono [33], a parallel thread in the literature has explored at depth the role of isoperimetric inequalities in establishing the (rapid) convergence of (ULA) when the potential of π is Lipschitz smooth, either via the use of a *logarithmic Sobolev inequality* (LSI) in the case of Mou et al. [26] and Chewi et al. [11], or a *Poincaré inequality* (PI) in the case of Balasubramanian et al. [4], and even weaker inequalities in Mousavi-Hosseini et al. [27] possibly reducing the degree of smoothness to (global) Hölder continuity of the drift of (LSDE), cf. Nguyen et al. [29] and Erdogdu & Hosseinzadeh [15],Mousavi-Hosseini et al. [27].

Our contributions in the context of related work. Our paper seeks to bridge these branches of the sampling literature – the relaxation of global Lipschitz smoothness requirements and the relaxation of convexity requirements via the use of isoperimetric inequalities – and, in so doing, to bring together the best of both worlds. Specifically, motivated by applications to the optimization and sampling of deep learning models (which are notoriously non-Lipschitz), we seek to answer the following question:

> *How to sample efficiently in the absence of log-concavity and linear gradient growth properties? Can one derive bounds in different distances with weaker assumptions?*

This is a difficult setting for sampling because, as has been noted in several works, both (ULA) and its SGLD variants may be highly unstable in such scenarios; in particular, when the drift coefficient of (ULA) exhibits superlinear growth, the Euler-Marauyama scheme – which forms the core component of (ULA) – diverges in a very strong sense. A key result in this direction was obtained by Hutzenthaler et al. [17] who showed that the difference of the exact solution of a stochastic differential equation (SDE) and its numerical approximation, diverges to infinity in the strong mean square sense, even at a finite point in time. This negative result has shown that superlinear growth of the drift coefficient directly results in a blow-up of the moments of the numerical approximation scheme used to generate samples, which thus explains the failure of these algorithms.

Providing an efficient work-around to this issue is not easy, and one needs to explore the roots of (ULA) for a possible answer – specifically, going all the way back to the initial observation that (ULA) is an Euler-Maruyama numerical apporximation scheme for the trajectories of (LSDE), and using the theory of numerical solutions of SDEs to explore a different angle of attack. In this regard, a technology tailored to solving SDEs with superlinearly growing drifts first emerged in the works of Hutzenthaler et al. [18] and Sabanis [31, 32], revolving around a technique known as "taming". The idea of these schemes is to create an adaptive Euler scheme with a new drift coefficient μ^λ which is a "tamed", rescaled version of the original drift, with the step-size of the algorithm appearing in the rescaling factor, and with the aim of ensuring the following dove-tailing properties:

(P1) μ^λ has at most linear growth, that is, $\mu^\lambda(x) = \mathcal{O}(|x|)$ for large x.

(P2) μ^λ converges pointwise to μ in the limit $\lambda \to 0$.

	Convexity	Dissipativity
Brosse et al. [7]	strongly convex (for W_2)	$\langle x \vert \nabla u(x)\rangle \geq c\vert\nabla u(x)\vert\vert x\vert$
Neufeld et al. [28]	convex at infinity	$(2+r)$-dissipative $(r>0)$
Lytras & Sabanis [22]	LSI	2-dissipative
Current work	PI + WC / LSI	1-dissipative

Table 1: Comparison of convexity and dissipativity assumptions in related works.

	KL Divergence	Total Variation	Wasserstein	Constants
Brosse et al. [7]	—	$\tilde{\mathcal{O}}(1/\varepsilon^2)$	$\tilde{\mathcal{O}}(1/\varepsilon)$ $[W_2]$	$\exp(\mathcal{O}(d))$
Neufeld et al. [28]	—	—	$\tilde{\mathcal{O}}(1/\varepsilon^2)$ $[W_2]$	$\exp(\mathcal{O}(d))$
Lytras & Sabanis [22]	$\tilde{\mathcal{O}}(1/\varepsilon)$	$\tilde{\mathcal{O}}(1/\varepsilon^2)$	$\tilde{\mathcal{O}}(1/\varepsilon^2)$ $[W_2]$	$\mathrm{poly}(d)/C_{\mathrm{LSI}}$
Current work(under LSI)	$\tilde{\mathcal{O}}(1/\varepsilon)$	$\tilde{\mathcal{O}}(1/\varepsilon^2)$	$\tilde{\mathcal{O}}(1/\varepsilon^2)$ $[W_2]$	$\mathrm{poly}(d)/C_{\mathrm{LSI}}$
Current work (under PI)	$\tilde{\mathcal{O}}(1/\varepsilon^3)$	$\tilde{\mathcal{O}}(1/\varepsilon^6)$	$\tilde{\mathcal{O}}(1/\varepsilon^8)$ $[W_1]$	$\mathrm{poly}(d)/C_{\mathrm{PI}}$

Table 2: Comparison of convergence rates under different assumptions; factors that are logarithmic in $1/\varepsilon$ have been absorbed in the tilded $\tilde{\mathcal{O}}(\cdot)$ notation. The observed drop relative to the concurrent work of Lytras & Sabanis [22] is due to the weaker assumptions made in our paper – Poincaré vs. log-Sobolev and weak dissipativity vs. 2-dissipativity (or higher), cf. Table 1. The constants C_{LSI} and C_{PI} refer to the (positive) constants that appear in the log-Sobolev and Poincaré inequalities respectively.

This technique has been applied previously in the setting of Langevin-based sampling in multiple works under strong dissipativity or convexity assumptions [7; 19], in the stochastic gradient case [21] under a "convexity at infinity" assumption [28] and, in a concurrent work, under a logarithmic Sobolev inequality coupled with a 2-dissipativity assumption [22]. However, even though it is fairly common for distributions with superlinearly growing log-gradients to satisfy a logarithmic Sobolev inequality, it is not always possible obtain a bound that remains well-behaved with respect to the dimension – and similar limitations also hold for the 2-dissipativity condition.

The main contribution of our paper is to provide a bridge between these two worlds and bring to the forefront the best properties of both: sampling efficiently from potentials with locally Lipschitz log-gradients that may grow polynomially at infinity, with a drift coefficient that is only 1-dissipative (instead of 2-dissipative) and a considerably lighter Poincaré inequality requirement – as opposed to the more rigid framework imposed by the use of LSIs. Specifically, we propose two novel algorithmic schemes, the *weakly dissipative tamed unadjusted Langevin algorithm* (wd-TULA) and the *regularized tamed unadjusted Langevin algorithm* (reg-TULA) which allow us to simultaneously treat superlinearly growing drift coefficients for target distributions satisfying a Poincaré inequality, the former under a weak convexity (WC) requirement, the latter without. For completeness, we also show that the proposed taming schemes achieve optimal convergence rates in the presence of stronger LSI conditions.

To position these contributions in the context of related work, Table 1 summarizes our paper's assumptions relative to the most closely related works in the literature, and Table 2 provides a side-by-side comparison of the achieved rates. To the best of our knowledge, the work closest to our own is the concurrent work of Lytras & Sabanis [22], who provide a tamed algorithmic scheme achieving an $\tilde{\mathcal{O}}(1/\varepsilon)$ rate of convergence to π in the KL divergence metric (respectively $\tilde{\mathcal{O}}(1/\varepsilon^2)$ in terms of the total variation and Wasserstein W_2 distance). Due to the relaxation from a logarithmic Sobolev inequality to a considerably weaker Poincaré

inequality, our rates do not match those of Lytras & Sabanis [22] in the case of the KL and total variation (TV) metrics – where wd-TULA achieves a rate of $\tilde{\mathcal{O}}(1/\varepsilon^3)$ and $\tilde{\mathcal{O}}(1/\varepsilon^6)$ respectively, cf. Table 2 (the Wassserstein metrics are otherwise incomparable; see also Theorem 3 for the rates without any WC requirements). This also applies to the "convexity at infinity" assumption of Neufeld et al. [28], which has been shown to imply an LSI, and is thus considerably more stringent than the PI setting of our paper. Importantly, our analysis still carries a polynomial dependence on the dimensionality of the problem, in contrast to the analysis of Neufeld et al. [28] where the dependence is exponential. These aspects of our results are particularly intriguing for future work on the subject, as they open a hitherto unexplored link between the isoperimetric inequalities, the role of coercivity in the target distribution in a superlinearly-growing gradient setting.

2. Setup and blanket assumptions

In this section, we provide the necessary groundwork for stating the proposed algorithmic schemes and our main results.

2.1. Notational conventions. We begin by fixing notation and terminology. Throughout our paper, $|\cdot|$ denotes the Euclidean norm of a vector, $\|\cdot\|$ the spectral norm of matrix; the Frobenius norm will be denoted by $\|\cdot\|_F$, and the total variation distance by $\|\cdot\|_{\mathrm{TV}}$. For a sufficiently smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$, we will write ∇f, $\nabla^2 f$ and Δf for its gradient, Hessian matrix, and Laplacian respectively, and $J^{(i)} f$ for its i-th order Jacobian. We also write $\mathcal{H}^k$ for the usual Sobolev space.

For any two probability measures μ, ν on a measurable space Ω with a σ-algebra understood from the context, we will write $d\mu/d\nu$ for the Radon-Nikodym derivative of μ with respect to ν when μ is absolutely continuous relative to ν ($\mu \ll \nu$). In this case, the *Kullback–Leibler* (KL) divergence of μ with respect to ν is defined as

$$H_\nu(\mu) = \int_\Omega \frac{d\mu}{d\nu} \log\left(\frac{d\mu}{d\nu}\right) d\nu. \tag{KL}$$

We say that ζ is a *transference plan* of μ and ν if it is a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ (endowed with the standard Borel algebra) and we have $\zeta\left(A \times \mathbb{R}^d\right) = \mu(A)$ and $\zeta\left(\mathbb{R}^d \times A\right) = \nu(A)$ for every Borel subset A of $\mathbb{R}^d$. We denote by $\Pi(\mu,\nu)$ the set of transference plans of μ and ν. Furthermore, we say that a couple of $\mathbb{R}^d$-valued random variables (X,Y) is a coupling of μ and ν if there exists $\zeta \in \Pi(\mu,\nu)$ such that (X,Y) is distributed according to ζ. Finally, for two probability measures μ and ν on $\mathbb{R}^d$, the Wasserstein distance of order $p \geq 1$ is defined as

$$W_p(\mu,\nu) = \left(\inf_{\zeta\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^p \, d\zeta(x,y)\right)^{1/p}. \tag{1}$$

2.2. Blanket assumptions. Throughout what follows, we will write $\pi := e^{-u}/\int e^{-u}$ for the target distribution to be sampled, and $Lf = \Delta f - \Gamma(u,f)$ for the infinitesimal generator of (LSDE), where $\Gamma(f,g) = \langle \nabla f | \nabla g \rangle$ denotes the carré du champ operator for $f, g \in \mathcal{H}^1$.

We begin by stating our blanket assumptions for (LSDE):

Assumption 1. The drift $h = \nabla u$ of (LSDE) satisfies the following conditions:

(A1) *Polynomial Lipschitz continuity:*

$$|h(x) - h(y)| \leq L'(1 + |x| + |y|)^{l'} |x-y| \tag{PLC}$$

for some $L', l' > 0$ and for all $x, y \in \mathbb{R}^d$

(A2) *Weak dissipativity:*

$$\langle h(x)|x\rangle \geq A|x|^a - b \tag{WD}$$

for some $a \geq 1$, $A, b > 0$ and for all $x \in \mathbb{R}^d$.

(A3) *Polynomial Jacobian growth:*

$$\max\{|h(x)|, \|J^{(i)}(h)(x)\|\} \leq L(1 + |x|^{2l}) \tag{PJG}$$

for some $L, l > 0$ and for all $x \in \mathbb{R}^d$.

Of the above, Assumption (A1) posits that h is locally Lipschitz continuous, with the modulus of Lipschitz continuity (essentially the largest eigenvalue of the Hessian of u) growing possibly at a polynomial rate at infinity. As such, Assumption (A1) allows us to capture a very broad spectrum of applications with superlinear Hessian growth (especially in the context of deep learning landscapes that exhibit polynomial growth with a degree equal to the depth of the underlying network).

Assumption (A2) is at the core of our analysis, as it enables us to provide moment bounds that are uniform in time: it is essentially a coercivity assumption, but with a relaxed exponent relative to the 2-dissipativity framework of other works, which can be fairly restrictive if the tails of the target distribution are thicker than sub-Gaussians. For generality, we treat not only the case of 1-dissipative gradients, but all dissipativity exponents $a \geq 1$. In practice, although it is quite possible that 2- dissipativity may hold for a given distribution, a-dissipativity for $a < 2$ may be much easier to verify and leverage to produce more favourable constants A,b.

Finally, Assumption (A3) is a strictly technnical requirement intended to streamline our presentation when rigorously differentiating under the integral sign – that is, exchanging the order of integration and time derivatives – in the use of a divergence theorem when establishing a differential inequality later in our paper.

Our next blanket assumption concerns the target distribution π:

Assumption 2. The target distribution π satisfies a *Poincaré inequality* (PI) of the form

$$\mathrm{Var}_\pi(f) := \int_{\mathbb{R}^d} \left(f - \int_{\mathbb{R}^d} f\, d\pi\right)^2 d\pi \leq \frac{1}{C_{\mathrm{PI}}} \int |\nabla f|^2\, d\pi \tag{PI}$$

for some positive constant $C_{\mathrm{PI}} > 0$ and all test functions $f \in \mathcal{H}^1(\mathbb{R}^d)$.

This assumption is weaker than the widely used (but more stringent) logarithmic Sobolev inequality.

$$H_\pi(\nu) := \int_{\mathbb{R}^d} f \log f\, d\pi \leq \frac{1}{2C_{\mathrm{LSI}}} \int_{\mathbb{R}^d} \frac{\Gamma(f,f)}{f}\, d\pi =: \frac{1}{2C_{\mathrm{LSI}}} I_\pi(\nu), \tag{LSI}$$

for some positive constant $C_{\mathrm{LSI}} > 0$ and for every probability measure $\nu \ll \pi$ with $f := d\nu/d\pi$. One should note that PI can be given as a consequence of the weak dissipativity condition (see [2]). In the case of diffusion processes, the importance of these inequalities lies in the fact that (PI) implies exponential ergodicity with respect to the χ^2 divergence, while (LSI) implies exponential ergodicity in relative entropy. In particular, by the Bakry–Emery theorem [1], (LSI) was established for strongly convex potentials and is stable under bounded pertrubations, Lipschitz mappings and convolutions. It also implies Talagrand's transporation-cost inequality: if μ satisfies (LSI), then

$$W_2(\mu, \nu) \leq \sqrt{\frac{2}{C_{\mathrm{LSI}}} H_\mu(\nu)}. \tag{2}$$

By comparison, (PI) is significantly less stringent than (LSI): to begin, (LSI) implies (PI) with the same constant but, moreover, (PI) has been shown to hold for dissipative potentials

where (LSI) fails, and is also stable under pertubations, Lipschitz mappings and convolutions. It also implies exponential moments of some order i.e., $\mathbb{E}_\mu e^{q|x|} < \infty$ whenever μ satisfies (PI).

Our last assumption concerns the convexity characteristics of the potential $u(x)$ and will be used only in the case where (PI) is the strongest condition in place.

Assumption 3. u is weakly convex, i.e.,

$$\nabla^2 u(x) \succcurlyeq -KI \quad \text{for some } K > 0 \text{ and for all } x \in \mathbb{R}^d. \tag{WC}$$

This assumption simply means that the eigenvalues of $\nabla^2 u(x)$ do not become arbitrarily negative, and is widely used in the numerical approximation of SDEs. For our purposes, we will mostly use it in the context of the famous HWI inequality

$$H_\pi(\nu) \leq \sqrt{I_\pi(\nu)} W_2(\pi,\nu) + \frac{K}{2} W_2^2(\pi,\nu) \quad \text{for all } \pi,\nu \in \mathcal{P}_2(\mathbb{R}^d). \tag{HWI}$$

To provide a glimpse of the analysis to come, we will develop and examine two novel taming schemes, one non-regularized and one regularized, that are tailored to the dissipativity profile of the initial potential, and which cover the following cases:

(1) When the potential satisfies (LSI).
(2) For the non-regularized case: when (LSI) fails and the potential satisfies (PI) along with (WC).
(3) For the regularized case: when (LSI) fails and the potential only satisfies (PI).

2.3. The failure of the unadjusted Langevin algorithm. Before moving forward with the development of the taming schemes mentioned above, we conclude this section with a simple – but not simplistic – 1-dimensional example that satisfies our range of assumptions, but where the "vanilla" unadjusted Langevin algorithm fails.

Setting $u(x) = x^3/3$ and applying (ULA) with step-size λ and initial condition $X_0 = \mathcal{N}(0, \frac{4}{\lambda})$, we get

$$X_{n+1} = X_n - \lambda X_n^2 + \sqrt{2\lambda}\xi_{n+1}$$

Then, since X_n is independent of ξ_{n+1}

$$\mathbb{E}[X_{n+1}^2] = \mathbb{E}\left[X_n^2(1-\lambda X_n)^2\right] + 2\lambda = \mathbb{E}X_n^2(1 - 2\lambda\mathbb{E}X_n + \lambda^2\mathbb{E}X_n^2) + 2\lambda$$

using the inequality $1 - 2x + x^2 \geq -1 + \frac{1}{2}x^2$ one obtains

$$\mathbb{E}[X_{n+1}^2] \geq -\mathbb{E}X_n^2 + \frac{1}{2}\lambda^2\mathbb{E}X_n^6 \geq -\mathbb{E}X_n^2 + \frac{\lambda^2}{2}\left(\mathbb{E}X_n^2\right)^3 + 2\lambda$$

where the last step was obtain by Jensen's inequality. Applying for $n = 0$ it is easy to see that

$$\mathbb{E}X_1^2 \geq (\mathbb{E}X_0^2)\left(\frac{\lambda^2}{2}\mathbb{E}(X_0^2)^2 - 1\right) + 2\lambda \geq \mathbb{E}X_0^2 + 2\lambda.$$

Iterating over n yields

$$\mathbb{E}X_{n+1}^2 \geq \mathbb{E}X_n^2 + 2\lambda n. \tag{3}$$

We thus see that the second moment of the algorithm's iterates diverges as $n \to \infty$, indicating in this way that (ULA) cannot be used to sample from the target distribution.

3. Tamed schemes and main results

We now proceed to state our tamed algorithmic schemes and main results.

3.1. Taming without regularization. To streamline our presentation and ease the introduction of the various components of the analysis, we begin with the case where the eigenvalues of $\nabla^2 u$ do not become arbitrarily negative, i.e., u satisfies the weak convexity assumption (WC). In this case, we will consider the *weakly dissipative tamed unadjusted Langevin algorithm* that iterates as

$$\bar{\theta}^\lambda_{n+1} = \bar{\theta}^\lambda_n - \lambda h_\lambda(\bar{\theta}^\lambda_n) + \sqrt{2\lambda}\xi_{n+1} \quad \text{for all } n = 0, 1, \dots \tag{wd-TULA}$$

where θ_0 is initialized randomly according to a Gaussian distribution π_0,[1] ξ_n is an i.i.d. sequence of Gaussian d-dimensional vectors with unit covariance, and the *tamed drift* h_λ is given by

$$h_\lambda(x) = \frac{Ax}{(1+|x|^2)^{1-a/2}} + f_\lambda(x) \quad \text{with} \quad f_\lambda(x) = \frac{f(x)}{1+\sqrt{\lambda}|x|^{2\ell}} \tag{4}$$

where $f(x) = h(x) - \frac{Ax}{(1+|x|^2)^{1-a/2}}$ and the various constants defined as in Assumption 1.

To connect (wd-TULA) with the existing literature on tamed schemes, we note here that the majority of taming factors are either of the form $h(x)/[1+(\lambda)^c|h(x)|]$ for $c = 1$ or $c = 1/2$ [7], or of the form $h(x)/[1+\lambda^c|x|^{2l-1}]$ [21]. In this regard, the taming scheme (4) is more intricate: we first split the original gradient drift into a part which has at most linear growth, and we then proceed to tame the superlinearly growing part. The drift coefficient of this scheme has the property that it grows at most as $|x|^{\frac{a}{2}}$ (so it grows at most linearly) while inheriting the dissipativity condition of the initial gradient. For more details, see Lemma A.1. In this regard, when the potential satisfies a stronger 2-dissipativity condition, we recover the taming scheme of Lytras & Sabanis [22].

Our main result for (wd-TULA) may then be stated as follows:

Theorem 1. *Suppose that Assumptions 1–3 hold and let ρ_n denote the distribution of the n-th iterate of (wd-TULA) run with $\lambda < \lambda_{\max} = \min\{\frac{1}{4(2AC^*+2L+1)^2}, \frac{1}{\dot{c}_0 H_\pi(\rho_0)}, \frac{2}{\mu^2}\}$ where the constants are given in the proof of Proposition A.2 and Lemmas A.2, Proposition A.3 .*

Then ρ_n enjoys the convergence guarantee

$$H_\pi(\rho_n) \le \left(1 - \frac{\dot{c}_0}{2}\lambda^{3/2}\right)^n H_\pi(\rho_0) + \left(1 + \frac{4c_1}{c_0}\right)\sqrt{\lambda} \tag{5}$$

where c_1 depends polynomially on d and $\dot{c}_0$ is an explicit function of the Poincaré constant C_{PI} of (PI). In particular, given a tolerance level $\varepsilon > 0$, (wd-TULA) achieves $H_\pi(\rho_n) \le \varepsilon$ within $n \ge \dot{c}_0^{-1}(1 + c_1/\dot{c}_0)^3 \log(2/\varepsilon)/\varepsilon^3 = \tilde{\Theta}(1/\varepsilon^3)$ if run with step-size $\lambda \le \varepsilon^2/[4(1 + 4C_1/\dot{c}_0)^2]$.

This theorem ensures that the algorithm converges at a polynomial rate, even in the absence of (LSI). This is very important in practice as, even if (LSI) holds, it is usually difficult to derive explicit bounds with nice dependence on the problem's defining parameters. More to the point, if (LSI) holds and C_{LSI} is known, the proposed algorithm exhibits optimal convergence rates, achieving in this way the best of both worlds:

Theorem 2. *Suppose that Assumption 1 and (LSI) hold. Let ρ_n be the distribution of $n-th$ iterate of the algorithm (wd-TULA). Then, for $\lambda \le \lambda_{\max}$, we have*

$$H_\pi(\rho_n) \le e^{-\frac{3}{2}C_{LSI}\lambda(n-1)} H_\pi(\rho_0) + \frac{\hat{C}}{\frac{3}{2}C_{LSI}}\lambda$$

[1]The Gaussian requirement could be relaxed by positing that $|\nabla \log \pi_0|$ and $\|\nabla^2 \log \pi_0\|$ grow at most polynomially, but we will not need this level of generality.

where $\hat{C}$ depends polynomially on the dimension with leading term at most $\mathcal{O}\left(d^{\max\{2,l'+1\}(2l+1)}\right)$. In particular, given a tolerance level $\varepsilon > 0$, if (wd-TULA) is run with step-size $\lambda \leq \frac{3\epsilon C_{LSI}}{2\hat{C}}$, for $n \geq \frac{2\hat{C}}{\epsilon} C_{LSI}^{-1} \log(\frac{2}{\epsilon} H_\pi(\rho_0)) = \tilde{\Theta}(1/\varepsilon)$ iterations, we have $H_\pi(\rho_n) \leq \epsilon$.

Theorems 1 and 2 are our main results for (wd-TULA). The proof of both theorems is fairly arduous and involves a series of intricate steps, so, to streamline our presentation and facilitate our comparison with the case where (WC) is dropped altogether, we proceed directly to the regularized version of (wd-TULA) and defer the discussion of the proof of the theorem to the next section.

3.2. Regularized taming. We consider now the case where the eigenvalues of $\nabla^2 u(x)$ become arbitrarily negative (i.e., Assumption 3 fails altogether). To account for this negative growth, we are going to regularize the tamed potential by anchoring it close to the original target. This will require care to ensure that the new sampling potential satisfies Assumption 3 with a controllable constant, as well as the corresponding regularity requirements of Assumption 1.

Without further ado, these considerations lead to the *regularized potential*

$$u_{r,\lambda}(x) = u(x) + \lambda |x|^{2r+2} \tag{6}$$

where, with a fair degree of hindsight, the exponent r is chosen so that $r > l/2$ and $r(2+l')/[(r+1)(2r-l')] < 1$ (a moment's reflection shows that this is not the empty set).

In view of the above, the *regularized taming* scheme that we will consider involves rescaling by the factor $(1+\sqrt{\lambda}|x|^{2r+1})$, leading to the regularized drift:

$$h_{r,\lambda}(x) = \frac{Ax}{(1+|x|^2)^{1-\frac{a}{2}}} + \frac{\nabla u_{r,\lambda}(x) - Ax(1+|x|^2)^{a/2-1}}{1+\sqrt{\lambda}|x|^{2r+1}}. \tag{7}$$

In this way, we obtain the *regularized tamed unadjusted Langevin algorithm*

$$\bar{x}^\lambda_{n+1} = \bar{x}^\lambda_n - \lambda h_{r,\lambda}(\bar{x}^\lambda_n) + \sqrt{2\lambda}\Xi_{n+1} \tag{reg-TULA}$$

where Ξ_n is an i.i.d. sequence of Gaussian d-dimensional vectors. Our main result for this regularized sampling scheme may then be stated as follows:

Theorem 3. *Suppose that Assumptions 1 and 2 hold and let ρ_n^{reg} denote the distribution of the n-th iterate of (reg-TULA) run with $\lambda < \lambda_{\max,2} := \min\{\lambda_{\max}, \frac{ln2}{R_2^{2r+2}}\}$ where R_2 is given in Lemma A.8. Then ρ_n^{reg} enjoys the convergence guarantee*

$$H_\pi(\rho_n^{\mathrm{reg}}) \leq \left(1 - c\lambda^{1+\frac{1}{r+1}+\frac{l}{2r-l}}\right)^{n-1} H_{\pi_{\mathrm{reg}}}(\rho_0) + (\hat{C}/c)\lambda^{1-\frac{1}{r+1}-\frac{l}{2r-l}} + C_3\lambda \tag{8}$$

where $c, C_3, \hat{C}$ depend polynomially on d. In particular, if $c_{l,r} := \frac{r(2+l)}{(r+1)(2r-l)}$, then, for $\lambda < \mathcal{O}\left(\epsilon^{\frac{1}{1-c_{l,r}}}\right)$, we have

$$H_{\pi_{\mathrm{reg}}}(\rho_n^{\mathrm{reg}}) \leq \epsilon \quad \textit{after} \quad n = \Theta\left(\log(1/\epsilon)\cdot \epsilon^{-\frac{1+c_{l,r}}{1-c_{l,r}}}\right) \quad \textit{iterations.}$$

4. Proof outline and technical innovations

To give an idea of the main ideas and technical innovations required for the proof of Theorems 1–3, we provide below a brief roadmap of our proof strategy.

The cornerstone of our approach is the derivation of a differential inequality in the spirit of Vempala & Wibisono [33]. The tricky part here is that the drift coefficient is not the original gradient but a tamed one, which yields additional complexity when one tries to

prove the exchange of integrals and derivatives starting from the Fokker-Planck equation. In so doing, we ultimately obtain a "template inequality" of the form

$$\frac{d}{dt}H_\pi(\hat{\pi}_t) \leq -\frac{3}{4}I_\pi(\hat{\pi}_t) + \mathbb{E}|h(\theta_t) - h_\lambda(\theta_{k\lambda})|^2 \tag{9}$$

where $\hat{\pi}_t$ is the continuous-time interpolation of the algorithm.

The first term of the template inequality (9) is connected to the relative entropy via an isoperimtric inequality, either directly – if (LSI) holds – or by means of another inequality – under (PI) and (WC). In this last case, the connection is achieved by establishing a "modified" version of (LSI) as a consequence of (PI) and (HWI). As for the second term of (9), its contribution can be controlled by the one-step error of the algorithm (by local Lipschitzness) and the L_2 approximation of the tamed scheme to the original gradient at grid points. A key difficulty here is that these bounds must be uniform in the number of iterations moment bounds for our algorithm which is obtained by the careful construction of the tamed coefficient, which satisfies a dissipativity condition.

This is an important novetlyt of our work, as we are able to achieve a uniform-in-time exponential moment bound for the algorithm, even in the 1− dissipative case (contrary to other works such as [15] where the moments bounds are not uniform in time or [34] which leverages a "convexity at infinity" assumption). We achieve this by means of a Herbst argument for Gaussians (since each iterate is a Gaussian when conditioned to the previous step), to pass from the conditional expectation of the exponential to the exponential of the conditional expectation. We then use the contraction structure (which is provided by the inherited dissipativity of our scheme and the growth of the tamed drift coefficient as given in Lemma A.1) to create an induction.

In the absence of *both* (LSI) *and* (WC), we employ a similar method to sample from the regularized potential – for which we prove the equivalent of (LSI) – and we then proceed to compute the KL divergence of the algorithm's iterates relative to the target distribution. The main challenge here is to show that the regularized potential also satisfies (PI) with a constant that is explicitly connected to the Poincaré constant of the *original* target and is *independent* of λ. In general, these are mutually antagonistic properties, which are ultimately achieved in our case by leveraging the dissipativity properties of the regularized potential to find a Lyapunov function W such that $LW \leq -\theta$ for some $\theta > 0$ outside of a ball. By using the inequality for the infinitesimal generator of the Langevin SDE with drift coeffient the regularized potential, and the fact that, inside said ball, the regularized measure inherits the Poincaré constant of the original (as a bounded perturbation thereof), and by employing a variation of a shrewd argument of Cattiaux et al. [8, Proof of Theorem 2.3], we are finally able to establish (PI) on the whole space. The conditions of 2-dissipativity and weak convexity are then easier to prove, eventually leading to the upgrade of (PI) to a suitably modified form of (LSI).

A major obstacle in the above strategy is determining the explicit constants while simultaneously trying to minimize their dependence on the step-size as much as possible. For example, one can establish a version of (LSI) for the regularized potential in a more direct manner, by simply using convexity at infinity, or by a technique similar to Lytras & Sabanis [22, Corollary 5.4]. However, the version of (LSI) obtained in this would involve a catastrophic exponential dependence on $1/\lambda$, which would thus render it unusable for deriving finite-time convergence rates. Albeit (significantly) more involved, our method completely circumvents the exponential dependence, which in turn enables the polynomial-time convergence rates of the proposed schemes (and recoups the technical investment described above).

For convenience, we summarize the main steps below, in decreasing order of the assumptions made.

Case 1: Analysis under (LSI)**.** This case concerns Theorem 2, and the analysis unfolds as follows:

(1) We prove that the tamed coefficients exhibit linear growth, and inherit the dissipativity property of the original gradient, cf. Lemma A.1.
(2) We use the properties of the tamed scheme to derive exponential (and subsequently polynomial) moments for our algorithm, cf. Lemmas A.2 and A.3.
(3) We establish a differential template inequality for the relative entropy between the continuous-time interpolation of the algorithm and target measure, cf. Corollary 2. This involves a rigorous treatment of the exchange of derivatives and integrals.
(4) We bound the remaining terms using the local Lipschitz property of ∇u and the approximation properties of the tamed scheme.
(5) We employ (LSI) to conntect the Fisher information term to the relative entropy, and we backsolve to produce convergence rates in terms of the KL divergence.
(6) Using Pinsker's inequality, we obtain a result for total variation distance and by Talagrand's inequality for the W_2 distance.

Case 2: (PI) **and** (WC)**.** This case concerns Theorem 1, and the analysis unfolds as follows:

(1) We employ the same scheme to tame ∇u and repeat the first steps as in the proof of Theorem 2.
(2) Lacking (LSI), our analysis branches out as follows: we use Assumption 3 and (PI) to produce a different template inequality between the KL divergence and the Fisher information distance between the continuous-time interpolation of the algorithm and target measure (Proposition A.2).
(3) We backsolve the derived differential inequality to produce non-exponential convergence rates relative to the KL divergence, cf. Proposition A.3 and Theorem 6.
(4) Finally, by using Pinsker's inequality and the exponential moments of our alogirthm and the bound in relative entropy, we are able to derive bounds TV and W_1 distance under (PI), cf. Corollary 3.

Case 3: (PI) **only.** This case concerns Theorem 3, and the analysis unfolds as follows:

(1) We introduce a regularized potential to sample from, and we show that it has a range of desirable properties as described in Lemma A.1.
(2) We use the same scheme to tame the gradient of the regularized potential and, through similar arguments, we derive exponential (and polynomial) moments for (reg-TULA).
(3) We show that the regularized measure satisfies a version of (PI), cf. Lemma A.8, and this inequality can be upgraded to a version of (LSI) with manageable constants (Proposition A.4).
(4) We branch back to the analysis using (LSI) to sample from the regularized potential, and we use the relation between the regularized and the original one to derive the algorithm's convergence rate, as outlined in Theorem 7.
(5) Using Pinsker's and Talagrand's inequalities, we convert these bounds to TV and W_2 (Corollary 4).

The details of all the above are provided in full in the paper's appendix.

5. Numerical Experiments

We proceed by providing some numerical experiments that validate our results. The focus of our attention will be the invariant measure π generated by the double-well potential

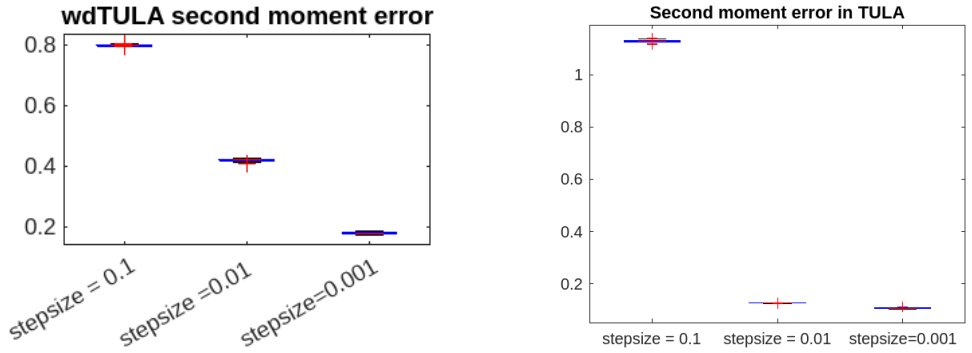

Figure 1: Performance of (wd-TULA) compared to TULA (left and right respectively); lower values are better.

$u = (|x|^2 - 1)^2$. Since $\nabla u = 4x(|x|^2 - 1)$ one easily observes that the potential satisfies our smoothness and dissipativity assumptions. Since the 1-dissipativity holds one can also deduce a Poincare inequality for the potential. Finally, it is immediate to see that the potential satisfies a convexity at infinity assumption, which of course is way stronger that our weak convexity assumption. For explicit calculations the interested reader can point to [28]. We perform the experiment for our algorithm for 10^6 iterations and 10^5 samples for 100 independent iterations. The dimension is $d = 100$ and we start the algorithm from a constant where every coordinate is zero, but the first coordinate is 200. We should note that starting from a constant doesn't contradict our analysis as we can still perform the analysis of the algorithm starting from the result of the 1st iteration which is a Gaussian. When one runs 'vanilla' ULA for stepsize=$\{0.1, 0.01\}$ all experiments show that the algorithm explodes (the moments exceed the infinity value of the computer) so the need to use an alternative becomes apparent. Here we present a boxplot which describes the the second moment of the first coordinate for different values. The second moment of each coordinate is given by

$$\mathbb{E}\left[X_i^2\right] = d^{-1} \int_{\mathbb{R}_+} r^2 \nu(r) \mathrm{d}r \Big/ \int_{\mathbb{R}_+} \nu(r) \mathrm{d}r, \quad \nu(r) = r^{d-1} \exp\left\{\left(r^2/2\right) - \left(r^4/4\right)\right\}$$

and is estimated by a random walk of 10^7 samples as $\mathbb{E}[X_i^2] = 0.104$. The figure below shows the behaviour of the algorithm for different stepsizes. One can see that the for stepsize=$\{0.1, 0.01\}$ the error is of order 10^{-1} while for stepsize $=0.001$ the error is of order 10^{-2}.

We also present a similar figure for the well-known TULA algorithm devoleped in [7]. We can see that the TULA algorithm is not as efficient as wd-TULA for large stepsize as it gives an error of approximately 1.1 but performs better for stepsize 0.1 (gives error of order 10^{-2}) and similarly to wdTULA for $\lambda = 0.001$.

A. Preliminary steps and lemmas

A.1. Moment bounds for (wd-TULA). In order to prove the moment bounds for our algorithm, the following properties of our the tamed coeffient will play a pivotal role.

Lemma A.1. *For all $x \in \mathbb{R}^d$, we have*

$$\langle h_\lambda(x), x\rangle \geq A_1|x|^a - B_1 \tag{A.1}$$

where $A_1 = A/2$ and $B_1 = \max\{A_1, B\}$. In addition, we have

$$|h_\lambda(x)|^2 \leq 4A^2|x|^a + 2L^2/\lambda + 4A^2 \quad \text{for all } x \in \mathbb{R}^d. \tag{A.2}$$

Proof. Postponed to the proof section. ∎

Lemma A.2. *Let* $M := \big(2(2d + 4A^2 + 2L^2 + A)\big)^{\frac{1}{a}}$ *and* $\mu = \frac{AaM^a}{16(1+M^2)^{1-\frac{a}{2}}}$. *Let* $V_\mu(x) = e^{\mu(1+|x|^2)^{\frac{a}{2}}}$. *There holds, for* $\lambda < \min\{1, \frac{A}{4}, \frac{2}{\mu^2}\}$

$$\sup_n \mathbb{E}V_\mu(\bar{\theta}_n^\lambda) \leq C_\mu$$

where $C_\mu \leq \mathcal{O}(e^{\mu d})$.

Proof. Postponed to the Appendix. The main tools used in the proof is the fact that conditionally on the previous step the algorithm is a Gaussian distribution and therefore satisfies a Log-Sobolev. We proceed our proof by using the fact that the function $f(x) = |x|/(1 + |x|^2)$ is 1-Lipschitz and we proceed using Herbst argument. The dissipativity and growth condition of our scheme enables to control the key quantity $|x - h_\lambda(x)|^2$. ∎

Lemma A.3. *Let* $p \in \mathbb{N}$. *There holds*

$$\sup_n |\bar{\theta}_n^\lambda|^{2p} \leq C_p$$

where $C_p \leq \mathcal{O}(d^p) + 2p$.

A.2. Establishing a key differential inequality regarding KL- divergence. The goal of this Section is to establish a differential inequality that will be the basis for our analysis. We define the continuous-time interpolation of our algorithm given as

$$\theta_t = \theta_{k\lambda} - (t - k\lambda)h_\lambda(\theta_{k\lambda}) + \sqrt{2}(B_t - B_{k\lambda}), \quad \forall t \in [k\lambda, (k+1)\lambda] \tag{A.3}$$

and $\theta_0 = \bar{\theta_0}$.

That way

$$\mathcal{L}(\theta_{k\lambda}) = \mathcal{L}(\bar{\theta}_k^\lambda) \quad \forall k \in \mathbb{N}.$$

We define the marginal distribution of θ_t as $\hat{\pi}_t$. One notices that, since conditioned on $\theta_{k\lambda}$, θ_t is a Gaussian its conditional distribution is given by

$$\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y) = Ce^{-\frac{\sqrt{t-k\lambda}}{2}|x-\mu(t,y)|^2}$$

where $\mu(t, y) = y - (t - k\lambda)h_\lambda(y)$ and C some normalizing constant. One further notes that, as $\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)$ can be viewed as a distribution of a process satisfying a Langevin SDE with constant drift $-h_\lambda(y)$ and initial condition y, i.e

$$\begin{aligned} d\hat{\mu}_t &= -h_\lambda(y)dt + \sqrt{2}dB_t, \quad \forall t \in (k\lambda, (k+1)\lambda] \\ \hat{\mu}_{k\lambda} &= y \end{aligned}$$

it satisfies the following Fokker-Planck PDE:

$$\frac{\partial \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)}{\partial t} = div\left(\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)h_\lambda(y)\right) + \Delta_x \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y). \tag{A.4}$$

Based on rigorous work done in the Appendix we are able to prove the analogous differential in time relative entropy inequality (that originally appeared in [33] for the vanilla ULA) for our tamed scheme.

Proposition A.1. *Let* $k \in \mathbb{N}$. *Then, for every* $t \in [k\lambda, (k+1)\lambda]$,

$$\frac{d}{dt}H_\pi(\hat{\pi}_t) = -\int_{\mathbb{R}^d} \langle \hat{\pi}_t(x)E\left(h_\lambda(\theta_{k\lambda})\big|\theta_t = x\right) + \nabla\hat{\pi}_t(x), \nabla \log \hat{\pi}_t(x) - \nabla \log \pi\rangle dx.$$

Proof. See Appendix. The idea of the proof is the same as the one in Lytras & Sabanis [22]. We see that our scheme has nice properties such as $||\nabla h_\lambda|| \leq \mathcal{O}(\frac{1}{\sqrt{\lambda}})$ and has polynomially growing higher derivatives. Working as in Lytras & Sabanis [22] we are able to produce the following inequality and show that in a small neighbourghood of t

$$\hat{\pi}_t(x) \leq Ce^{-r|x|^2}$$

and $|\nabla \log \hat{\pi}_t|$ and $||\nabla^2 \log \hat{\pi}_t||$ grow polynomially in x. This enables some change of the integrals and the derivatives along with application of the divergence theorem. By using Bayes theorem on the conditional Fokker-Planck equation with the divergence theorem we produce the following inequality. ■

Theorem 4. *Then, for $\lambda < \lambda_{\max}$ and for every $t \in [k\lambda, (k+1)\lambda]$,$k \in \mathbb{N}$, there holds*

$$\frac{d}{dt}H_\pi(\hat{\pi}_t) \leq -\frac{3}{4}I_\pi(\hat{\pi}_t) + \mathbb{E}|h(\theta_t) - h_\lambda(\theta_{k\lambda})|^2.$$

Lemma A.4. *There holds*

$$\mathbb{E}|\theta_t - \theta_{k\lambda}|^{2p} \leq C_{1,p}\lambda^p$$

The proof follows by using the growth of property of h_λ and the uniform in time moment bounds of the algorithm.

Lemma A.5. *Then,*

$$\mathbb{E}|h_\lambda(\theta_{k\lambda}) - h(\theta_t)|^2 \leq C_{err}\lambda \quad \forall k \in \mathbb{N},$$

where C_{err} is given explicitly in the proof and depends at most polynomially in the dimension.

For this proof we have split the split the difference as follows

$$|h(\theta_t) - h_\lambda(\theta_{k\lambda})|^2 \leq 2|h(\theta_t) - h(\theta_{k\lambda})|^2 + 2|h(\theta_{k\lambda}) - h_\lambda(\theta_{k\lambda})|^2.$$

The first term is bounded using Assumption (A1) Lemma A.4 and the uniform in time moment bounds of the algorithm (since $\mathcal{L}(\theta_{k\lambda}) = \mathcal{L}(\bar{\theta}_n^\lambda)$ and the squared taming error which is of order λ.
Now we are going to provide an inequality between the relative entropy and the Fisher information. When LSI is assumed the connection is immediate.

A.3. Convergence analysis.

A.3.1. Convergence under LSI.

Theorem 5. *Let ρ_n be the distribution of $n-th$ iterate of the algorithm* (wd-TULA). *Then, for $\lambda \leq \lambda_{max}$,*

$$H_\pi(\rho_n) \leq e^{-\frac{3}{2}C_{LSI}\lambda(n-1)}H_\pi(\rho_0) + \frac{\hat{C}}{\frac{3}{2}C_{LSI}}\lambda$$

where $\hat{C}$ is given explicitly in the proof and depends polynomially on the dimension.

Using Talagrand's inequality one deduces the following result regarding the convergence in Wasserstein distance.

Corollary 1. *There holds,*

$$W_2(\mathcal{L}(\bar{\theta}_n^\lambda), \pi) \leq \frac{\sqrt{2}}{\sqrt{C_{LSI}}}\left(e^{-\frac{3}{4}(C_{LSI}\lambda(n-1)}H_\pi(\rho_0) + \sqrt{\frac{\beta\hat{C}}{\frac{3}{2}(C_{LSI})}\lambda}\right).$$

A.3.2. **Convergence under PI and weak convexity.** When one does not assume any Log-Sobolev inequality other tools are needed to describe their connection.

Proposition A.2. *There exists $\dot{c}_0 > 0$ such that*

$$H_\pi(\hat{\pi}_t) \leq \frac{1}{\dot{c}_0}\sqrt{I_\pi(\hat{\pi}_t)}.$$

The relies on a combination of an inequality between the W_2 distance and the Fischer information which stems from the Poincare inequality, combined with the HWI inequality.

Corollary 2. *There holds,*

$$\frac{d}{dt}H_\pi(\hat{\pi}_t) \leq -\dot{c}_0 H_\pi^2(\hat{\pi}_t) + C_{err}\lambda \quad \forall t \in [k\lambda, (k+1)\lambda]$$

Proof. Combining Theorem 4, Proposition A.2 and Lemma A.5 yields the result. ∎

Proposition A.3. *Let ρ_n be the n-th iteration of our algorithm. Then, there holds*

$$H_\pi(\rho_{k+1}) \leq (H_\pi(\rho_k)^{-1} + \dot{c}_0\lambda)^{-1} + 2C_1\lambda^2$$

The proof proceeds by using a comparison theorem for ODEs to solve the differential inequality. By using elementary inequalites we reach a simpler recurrent condition which we iterate over n to reach the following theorem.

Theorem 6. *Suppose that λ satisfies the stepsize restrictions given in the moment bounds. In addition, we assume that $\lambda \leq \frac{1}{4\dot{c}_0 C_1}$. Then, There holds*

$$H_\pi(\rho_n) \leq (1 - \frac{\dot{c}_0}{2}\lambda^{\frac{3}{2}})^n H_\pi(\rho_0) + (1 + \frac{4C_1}{\dot{c}_0})\sqrt{\lambda}$$

Suppose that $\lambda < \epsilon^2/4(1 + 4\frac{C_1}{\dot{c}_0})^2$ Then, $H_\pi(\rho_n) \leq \epsilon$ after $n \geq \mathcal{O}\left(\log(\frac{1}{\epsilon})\frac{1}{\epsilon^3}\right)$ iterations.

Corollary 3. *Let $\bar{\theta}_n^\lambda$ be the $n-th$ iterate of the algorithm. Then, there holds*

$$\|\mathcal{L}(\bar{\theta}_n^\lambda), \pi) - \pi\|_{TV} \leq \frac{\sqrt{2}}{2}\left(\sqrt{H_\pi(\rho_0)}(1 - \frac{\dot{c}_0}{2}\lambda^{\frac{3}{2}})^{\frac{n}{2}} + \sqrt{(1 + 4\frac{C_1}{\dot{c}_0})}\lambda^{\frac{1}{4}}\right)$$

and

$$W_1(\mathcal{L}(\bar{\theta}_n^\lambda), \pi) \leq C_W\left(H_\pi(\rho_n) + H_\pi(\rho_n)^{\frac{1}{2}}\right)$$

A.4. Proving convergence without Assumption 3 using a regularized potential. In the case where **B**3 is missing, we are going to sample from a regularized potential which is close to the original target. The new regularized potential will inherit the important Local Lipschitzness, growth and dissipativity properties of the original and it will also satisfy **B**3 with constant depending on λ.

We first state some important properties of the new potential which are related to the properties of the original target.

A.4.1. **Properties of the regularized potential.**

Lemma A.6. *For $\lambda \leq 1$, the regularized potential $u_{\mathrm{reg},\lambda}$ satisfies Assumption (A1) with constants independent of λ. As a result, the tamed scheme inherits the a-dissipitivity condition, and has at most linear growth.*

Lemma A.7. *Let $p > 1$. There exist, $C_{reg,2p} \leq \mathcal{O}(d^p)$ such that*

$$\sup_n \mathbb{E}|x_n^\lambda|^{2p} \leq C_{reg,2p}$$

Since the regularized potential satisfies potential satisfies Assumption 1 with different constants independent of λ and the tamed scheme inherits the dissipativity condition, the proof follows in the same way as in the unregularized case.

Lemma A.8. *Let $\lambda \leq \frac{ln2}{2R_2^{2r+2}}$. and $R_2 \leq \mathcal{O}(d)$. The measure π_{reg} satisfies a Poincare inequality with constant $C_{P,r}^{-1}$ independent of λ.*

Proposition A.4. *The regularized measure satisfies LSI with constant $C_{LSI}^{-1} \leq \mathcal{O}\left((\frac{1}{\lambda})^{\frac{1}{r+1}+\frac{l'}{2r-l'}}\right)$*

To prove this theorem we use the fact that the regularized measure satisfies a Poincare inequality. By proving that it also satisfies a 2-dissipativity condition and has a lower bounded Hessian, by using some classic theorems depending on Lyapunov functions we show the Poincare inequality can be upgraded to a Log-Sobolev.

Theorem 7. *Let ρ_n^{reg} be the distribution of the n-th iterate of the tamed algorithn with the regularized gradient* (reg-TULA)*. There holds*

$$H_\pi(\rho_n^{\text{reg}}) \leq H_{\pi_{\text{reg}}}(\rho_n^{\text{reg}}) + \mathcal{O}(\lambda) \leq e^{-\dot{c}\lambda(n-1)} H_{\pi_{\text{reg}}}(\rho_0) + \frac{\hat{C}}{\dot{c}}\lambda$$

where $\dot{c} = C_{LS}$ given in Lemma A.4 and $\hat{C}$ depends polynomially on the dimension.

By using the same arguments as in the unregularized case one reaches a differential inequality. This time we make use of the Log-Sobolev inequality to get a differential inequality for $H_{\pi_{\text{reg}}}(\rho_n^{reg})$. Then, the proof of the connection between $H_\pi(\rho_n^{\text{reg}})$ and $H_{\pi_{\text{reg}}}(\rho_n^{\text{reg}})$ is a simple application of the definition of the relative entropy and the moments of the algorithm and the invariant measure.

Corollary 4. *Let $c_{l,r} := \frac{r(2+l)}{(r+1)(2r-l)}$ For $\lambda < \mathcal{O}\left(\epsilon^{\frac{1}{1-c_{l,r}}}\right)$, there holds*

$$H_{\pi_{\text{reg}}}(\rho_n) \leq \epsilon \quad \textit{after} \quad n \geq \mathcal{O}\left(\log(\frac{1}{\epsilon})(\frac{1}{\epsilon})^{\frac{1+c_{l,r}}{1-c_{l,r}}}\right) \quad \textit{iterations}$$

and

$$||\mathcal{L}(\bar{x}_n^\lambda) - \pi||_{TV} \leq \epsilon \quad n \geq \mathcal{O}\left(\log(\frac{1}{\epsilon}^2)(\frac{1}{\epsilon^2})^{\frac{1+c_{l,r}}{1-c_{l,r}}}\right) \quad \textit{iterations}$$

In addition, for $\lambda \leq \mathcal{O}\left(\epsilon^{\frac{2+c_{r,l}}{1-c_{l,r}}}\right)$ there holds

$$W_2(\mathcal{L}(\bar{x}_n^\lambda), \pi) \leq \epsilon \quad \textit{after} \quad n \geq \mathcal{O}\left(\log(\frac{1}{\epsilon})(\frac{1}{\epsilon})^{(2+c_{r,l})\frac{1+c_{l,r}}{1-c_{l,r}}}\right) \quad \textit{iterations.}$$

B. Proof section

B.1. Proof of preliminary statements.

Proof of Lemma A.1. It easy to see that if $\langle f(x), x\rangle < 0$, then $\langle f_\lambda(x), x\rangle \geq \langle f(x), x\rangle$ which implies that

$$\langle h_\lambda(x), x\rangle \geq A|x|^a - B.$$

Suppose that $\langle f(x), x\rangle \geq 0$. Then,

$$\langle h_\lambda(x), x\rangle \geq A\frac{|x|^2}{(1+|x|^2)^{1-\frac{a}{2}}} \geq A|x|^a\frac{|x|^2}{1+|x|^2}$$

If $|x| > 1$ then

$$\langle h_\lambda(x), x\rangle \geq \frac{A}{2}|x|^a$$

and if $|x| < 1$

$$\langle h_\lambda(x), x\rangle \geq \frac{A}{2}|x|^a - \frac{A}{2}.$$

To prove the second part one notices that

$$\begin{aligned}
|h_\lambda(x)|^2 &\leq 2(1 - \frac{1}{1+\sqrt{\lambda}|x|^{2l}})^2 A^2 \frac{|x|^2}{(1+|x|^2)^{2-a}} + 2(\frac{h(x)}{1+\sqrt{\lambda}|x|^{2l}})^2 \\
&\leq 2A^2(1+|x|^2)^{a-1} + 2\frac{L^2}{\lambda} \\
&\leq 2A^2(1+|x|^2)^{\frac{a}{2}} + 2\frac{L^2}{\lambda} \\
&\leq 4A^2 + 4A^2|x|^a + 2\frac{L^2}{\lambda}
\end{aligned}$$

■

B.2. Moment bounds.

Proof of Lemma A.2. The proof starts by noticing that conditioned on $\bar\theta_n^\lambda$, θ_{n+1}^λ is a Gaussian, with covariance matrix $\frac{\lambda}{2}I_d$. Thus conditioned on the previous step since the function $(1+|x|^2)^{\frac{1}{2}})$ is 1-Lipschitz by t Proposition 5.5.1 in Bakry et al. [3] there holds so for $\mu^2 \leq \frac{2}{\lambda}$,

$$\begin{aligned}
\mathbb{E}[V_\mu(\theta_{n+1}^\lambda)|\bar\theta_n^\lambda] &\leq e^{\mu^2\lambda} e^{\mu\mathbb{E}[(1+|\theta_{n+1}^\lambda|^2)^{\frac{1}{2}}|\bar\theta_n^\lambda]} \\
&\leq e^{\mu^2\lambda} e^{\mu\left(1+\mathbb{E}[|\theta_{n+1}^\lambda|^2|\bar\theta_n^\lambda]\right)^{\frac{1}{2}}} \\
&= e^{\mu^2\lambda} e^{\mu\left(1+|\bar\theta_n^\lambda - \lambda h_\lambda(\bar\theta_n^\lambda)|^2 + 2\lambda d\right)^{\frac{1}{2}}}
\end{aligned}$$

where the penultimate step was obtained by Jensen's inequality. Since

$$\begin{aligned}
|x - \lambda h_\lambda(x)|^2 &\leq |x|^2 - 2\lambda\langle x, h_\lambda(x)\rangle + \lambda^2|h_\lambda(x)|^2 \\
&\leq |x|^2 + \lambda(4\lambda A^2 - A|x|^a) + \lambda(A + 2L^2) + \lambda^2 4A^2
\end{aligned} \tag{B.1}$$

Since for $\lambda \leq \max\{1, \frac{1}{2A}\}$, and $|x| \geq M := \left(2(2d + 4A^2 + 2L^2 + A)\right)^{\frac{1}{a}}$ by (B.1) one deduces

$$\begin{aligned}
(1+|x-h_\lambda(x)|^2 + 2\lambda d)^{\frac{1}{2}} &\leq (1+|x|^2 - \lambda\frac{A}{4}|x|^a)^{\frac{1}{2}} \\
&= (1+|x|^2)^{\frac{1}{2}}\left(1 - \lambda\frac{A}{4}\frac{|x|^a}{(1+|x|^2)}\right)^{\frac{1}{2}} \\
&\leq (1+|x|^2)^{\frac{1}{2}}\left(1 - \frac{Aa}{8}\lambda\frac{|x|^a}{(1+|x|^2)}\right) \quad \text{using} \quad (1-t)^{\frac{1}{2}} \leq 1 - \frac{1}{2}t \\
&= (1+|x|^2)^{\frac{1}{2}} - \lambda\frac{A}{8}\frac{|x|^a}{(1+|x|^2)^{1-\frac{1}{2}}} \\
&= (1+|x|^2)^{\frac{1}{2}} - \lambda\frac{A}{8}\frac{|x|^a}{(1+|x|^2)^{\frac{1}{2}}} \\
&\leq (1+|x|^2)^{\frac{1}{2}} - \lambda\frac{AM^a}{8(1+M^2)^{\frac{1}{2}}}
\end{aligned} \tag{B.2}$$

and the last step was deduced using that the function $g(x) = \frac{x^a}{(1+|x|^2)^{\frac{1}{2}}}$ is increasing for $a \geq 1$ and $x \geq 0$ Using (B.2) one deduces that if $|\bar{\theta}_n^\lambda| \geq M$,

$$e^{\mu\left(1+|\bar{\theta}_n^\lambda - \lambda h_\lambda(\bar{\theta}_n^\lambda)|^2 + 2\lambda d\right)^{\frac{1}{2}}} \leq V_\mu(\bar{\theta}_n^\lambda) e^{-\mu^2\lambda}$$

On the other hand, if $|\bar{\theta}_n^\lambda| \leq M$, using the inequality $(1+z+y)^{\frac{1}{2}} \leq 1 + z^{\frac{1}{2}} + \frac{y}{2}$ one deduces

$$\begin{aligned}
\mu\left(1+|\bar{\theta}_n^\lambda - \lambda h_\lambda(\bar{\theta}_n^\lambda)|^2 + 2\lambda d\right)^{\frac{1}{2}} &= \mu\left(1+|\bar{\theta}_n^\lambda|^2 + (2|\bar{\theta}_n^\lambda||h_\lambda(\bar{\theta}_n^\lambda)| + \lambda^2|h_\lambda(\bar{\theta}_n^\lambda)|^2 + 2\lambda d)\right)^{\frac{1}{2}} \\
&\leq \mu(1+|\bar{\theta}_n^\lambda|^2)^{\frac{1}{2}} + (\frac{\mu}{2}(2|\bar{\theta}_n^\lambda|\lambda|h_\lambda(\bar{\theta}_n^\lambda)| + \lambda^2|h_\lambda(\bar{\theta}_n^\lambda)|^2 + 2\lambda d)^{\frac{1}{2}} \\
&\leq \mu(1+|\bar{\theta}_n^\lambda|^2)^{\frac{1}{2}} + \lambda C_M
\end{aligned} \tag{B.3}$$

where $C_M \leq C_0 + M^{2l+1} + M^{4l+2} + 2d$ where C_0 is an absolute constant independent of the dimension. As a result, if $|\bar{\theta}_n^\lambda| \leq M$

$$e^{\mu\left(1+|\bar{\theta}_n^\lambda - \lambda h_\lambda(\bar{\theta}_n^\lambda)|^2 + 2\lambda d\right)^{\frac{1}{2}}} \leq V_\mu(\bar{\theta}_n^\lambda) e^{C_M\lambda}$$

which leads to

$$\begin{aligned}
\mathbb{E}[V_\mu(\theta_{n+1}^\lambda)|\bar{\theta}_n^\lambda] &\leq V_\mu(\bar{\theta}_n^\lambda) e^{(C_M+\mu^2)\lambda} \\
&= e^{-\mu^2\lambda} V_\mu(\bar{\theta}_n^\lambda) + \left(e^{(C_M+\mu^2)\lambda} - e^{-\mu^2\lambda}\right) V_\mu(\bar{\theta}_n^\lambda) \\
&\leq e^{-\mu^2\lambda} V_\mu(\bar{\theta}_n^\lambda) + e^{(C_M+\mu^2)\lambda} V_\mu(\bar{\theta}_n^\lambda)\left(1 - e^{-(2\mu^2 - C_M)\lambda}\right) \\
&\leq e^{-\mu^2\lambda} V_\mu(\bar{\theta}_n^\lambda) + \lambda C
\end{aligned}$$

where the last step was derived from the inequality $1 - e^{-t} \leq t$. Putting all together one deduces,

$$\mathbb{E}V_\mu(\bar{\theta}_{n+1}^\lambda) \leq e^{-\lambda\mu^2 n}\mathbb{E}V_\mu(\theta_0) + \bar{C}.$$

■

Proof of Lemma A.3. The proof starts by noticing that the function $g(x) = (ln(x))^{2p}$ is concave for $x \geq e^{2p}$. As a result, for $n \in \mathbb{N}$

$$\begin{aligned}
\mathbb{E}(\mu|\bar{\theta}_n^\lambda| + 2p)^{2p} &\leq \mathbb{E}g(e^{\mu(1+|\bar{\theta}_n^\lambda|^2)^{\frac{1}{2}} + 2p}) \\
&\leq g(\mathbb{E}e^{\mu(1+|\bar{\theta}_n^\lambda|^2)^{\frac{1}{2}} + 2p}) \quad \text{(Jensen)} \\
&\leq 2^{2p-1}(ln\mathbb{E}e^{\mu(1+|\bar{\theta}_n^\lambda|^2)^{\frac{1}{2}}} + 2p^{2p}) \\
&\leq 2^{2p-1}(lnC_\mu^{2p} + 2p^{2p})
\end{aligned}$$

■

B.3. Rigorous proofs of integral and derivative exchange.

Lemma B.1. *Let $\lambda < \frac{1}{4(2AC^*+2L+1)^2}$. Then, the following hold: Let $k \in \mathbb{N}$ and $t \in [k\lambda, (k+1)\lambda]$. Then, there exist C, r, q such that*

- $$\hat{\pi}_t \leq Ce^{-r|x|^2} \quad \forall x \in \mathbb{R}^d$$
- $$|\nabla \log \hat{\pi}_t(x)| \leq C(1+|x|^q) \quad \forall x \in \mathbb{R}^d$$
- $$||\nabla^2 \log \hat{\pi}_t(x)| \leq C'(1+|x|^{q'}) \quad \forall x \in \mathbb{R}^d.$$

Proof. Let

$$\phi(x) = x - (t-k\lambda)h_\lambda(x) = x - (t-k\lambda)AR(x) - (t-\kappa\lambda)f_\lambda(x)$$

where $R(x) = \frac{x}{(1+|x|^2)^{1-\frac{\varrho}{2}}}$, $f_\lambda = \frac{h(x)-AR(x)}{g_\lambda}$ and $g_\lambda = \frac{1}{1+\sqrt{\lambda}|x|^{2l}}$. Since for the derivatives of R there holds

$$||J_R|| \leq C^*$$

and for $H := J_h$, $||H|| \leq g_\lambda L\frac{1}{\sqrt{\lambda}}$,

$$\begin{aligned}(t-k\lambda)||AJ_R + J_{f_\lambda}|| &\leq \lambda(A||J_R(x)|| + ||(H - AJ_R)g_\lambda + \nabla g_\lambda \otimes (h(x) - AR(x))|| \\ &\leq \lambda(AC^* + (||H|| + AC^*)g_\lambda + |\nabla g_\lambda(x)||h(x) - AR(x)|) \\ &\leq (2AC^* + 2L + 1)\sqrt{\lambda} \leq \frac{1}{2}.\end{aligned} \tag{B.4}$$

Thus,

$$\frac{1}{2}I_d \leq J_\phi \leq \frac{3}{2}I_d.$$

In addition, using the fact that the high derivatives of h and R have at most polynomial growth one can easily see that $||J_\phi^{(2)}||$ and $||J_\phi^{(3)}||$ have at most polynomial growth. From then, on we proceed with same arguments as in Lemmas A.5-A.7 in Lytras & Sabanis [22]. ■

Lemma B.1.

$$\mathbb{E}\left(\frac{\partial\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|\theta_{k\lambda})}{\partial t}\right) = \frac{\partial\hat{\pi}_t}{\partial t}(x).$$

Proof. Analysing the left hand side of the equation one deduces the following:

In a neighbourhood of t, for fixed x, $\frac{\partial\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)}{\partial t}$ decays exponentially with y and since $\hat{\pi}_{k\lambda}(y) \leq Ce^{-r|y|^2}$ due to Lemma B.1 one can exchange the derivative with the integral in the following expression

$$\frac{\partial}{\partial t}\int_{\mathbb{R}^d}\hat{\pi}_{k\lambda}(y)\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)dy = \int_{\mathbb{R}^d}\hat{\pi}_{k\lambda}(y)\frac{\partial\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)}{\partial t}dy.$$

Noticing that

$$\frac{\partial\hat{\pi}_t}{\partial t}(x) = \frac{\partial}{\partial t}\int_{\mathbb{R}^d}\hat{\pi}_{k\lambda}(y)\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)dy$$

and

$$\int_{\mathbb{R}^d}\hat{\pi}_{k\lambda}(y)\frac{\partial\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)}{\partial t}dy = \mathbb{E}\left(\frac{\partial\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|\theta_{k\lambda})}{\partial t}\right)$$

yields the result. ■

Lemma B.2.

$$\mathbb{E}\left(div_x\left(\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|\theta_{k\lambda})h_\lambda(\theta_{k\lambda})\right)\right) = div_x\left(\hat{\pi}_t(x)\mathbb{E}\left(h_\lambda(\theta_{k\lambda})\middle|\theta_t = x\right)\right).$$

Proof. Since $\hat{\pi}_t$ decays exponentially with y and for fixed t, in a neighbourhood of x $\nabla\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)$ is at most linear in y and h_λ has at most linear growth, this enables the interchange of integral and derivative with respect to x in the following expression

$$\int_{\mathbb{R}^d}\hat{\pi}_{k\lambda}(y)div_x\left(\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)h_\lambda(y)\right)dy = div_x\int_{\mathbb{R}^d}\hat{\pi}_{k\lambda}(y)\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)h_\lambda(y)dy$$

Since

$$\mathbb{E}\left(div\left(\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|\theta_{k\lambda})h_\lambda(\theta_{k\lambda})\right)\right) = \int_{\mathbb{R}^d}\hat{\pi}_{k\lambda}(y)div_x\left(\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)h_\lambda(y)\right)dy$$

and due to Bayes theorem

$$\begin{aligned} div_x \int_{\mathbb{R}^d} \hat{\pi}_{k\lambda}(y)\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)h_\lambda(y)dy &= div_x \int_{\mathbb{R}^d} \hat{\pi}_t(x)\hat{\pi}_{\theta_{k\lambda}|\theta_t}(y|x)h_\lambda(y)dy \\ &= div_x \left(\hat{\pi}_t(x)\mathbb{E}\left(h_\lambda(\theta_{k\lambda})\middle|\theta_t = x\right)\right) \end{aligned}$$

and the result immediately follows. ■

Lemma B.3.

$$\mathbb{E}\left(\Delta_x \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|\theta_{k\lambda})\right) = \Delta\hat{\pi}_t(x).$$

Proof. Noting that by definition

$$\mathbb{E}\left(\Delta_x \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|\theta_{k\lambda})\right) = \int_{\mathbb{R}^d} \Delta_x(\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y))\hat{\pi}_{k\lambda}(y)dy$$

and

$$\Delta_x\hat{\pi}_t(x) = \Delta_x \int_{\mathbb{R}^d} \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)\hat{\pi}_{k\lambda}(y)dy$$

it suffices to prove that

$$\int_{\mathbb{R}^d} \Delta_x(\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y))\hat{\pi}_{k\lambda}(y)dy = \Delta_x \int_{\mathbb{R}^d} \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)\hat{\pi}_{k\lambda}(y)dy.$$

By simple computations for the Gaussian distribution one deduces that $|\nabla_x \log \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)|$, $\Delta_x \log \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)$ have at most linear growth with respect to y in a neighbourhood of x . Writing

$$\Delta_x\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y) = \left(\Delta_x \log \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y) + |\nabla_x \log \hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)|^2\right)\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)$$

one deduces that in a neighbourhood of x, the integrand in the first term is dominated by a function of the form $C(1+|y|^2)e^{-c|y|^2}$. Applying the dominated convergence theorem enables the exchange of the integral and the Laplacian which completes the proof. ■

Corollary 5.

$$\frac{\partial\hat{\pi}_t}{\partial t}(x) = div_x \left(\hat{\pi}_t(x)\mathbb{E}\left(h_\lambda(\theta_{k\lambda})\middle|\theta_t = x\right)\right) + \Delta\hat{\pi}_t(x) \quad \forall t \in [k\lambda, (k+1)\lambda]$$

Proof. Taking expectations in (A.4) and combining Lemmas B.1,B.2 and B.3 yields the result. ■

Lemma B.4. *There exist $C, k, r' > 0$ indepent of x, uniform in a small neighbourghood of t such that*

$$div_x \left(\hat{\pi}_t(x)\mathbb{E}\left(h_\lambda(\theta_{k\lambda})\middle|\theta_t = x\right)\right) + \Delta\hat{\pi}_t \leq C(1+|x|^k)e^{-r'|x|^2}$$

Proof. Writing, due to Bayes' theorem,

$$\begin{aligned} &div_x \left(\hat{\pi}_t(x)\mathbb{E}\left(h_\lambda(\theta_{k\lambda})\middle|\theta_t = x\right)\right) = \int_{\mathbb{R}^d} \hat{\pi}_{k\lambda}(y) div_x \left(\hat{\pi}_{t|\mathcal{F}_{k\lambda}}(x|y)h_\lambda(y)\right) dy \\ &\leq Ce^{-c|x|^2+|x|} \int_{\mathbb{R}^d} e^{-r|y|^2}|y|dy \end{aligned}$$

for some $C, c, r > 0$ where the last step is a result of the Gaussian expression of the conditional density, the linear growth of h_λ and the exponential decay of $\hat{\pi}_{k\lambda}$ given in Lemma B.1.

For the second term, writing

$$\Delta\hat{\pi}_t = \hat{\pi}_t\left(|\nabla \log \hat{\pi}_t|^2 + \Delta \log \hat{\pi}_t\right)$$

the result follows due to Lemma B.1. ■

Corollary 6.

$$\frac{d}{dt}H_\pi(\hat{\pi}_t) = \int_{\mathbb{R}^d} \frac{\partial \hat{\pi}_t(x)}{\partial t}(1 + \log \hat{\pi}_t(x) - \log \pi)dx$$

Proof. Noting that $\log \hat{\pi}_t, \log \pi$ have polynomial growth, due to Lemma B.4, $\frac{\partial \hat{\pi}_t(x)}{\partial t}(1 + \log \hat{\pi}_t(x) - \log \pi)$ can be dominated by an L^1 integrable function over small neighbourhood of t, thus using the dominated convergence theorem one deduces the exchange of derivative and integration i.e

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{\partial \hat{\pi}_t(x)}{\partial t}(1 + \log \hat{\pi}_t(x) - \log \pi(x))dx &= \int_{\mathbb{R}^d} \frac{\partial}{\partial t}\left(\hat{\pi}_t(x) \log \frac{\hat{\pi}_t(x)}{\pi(x)}\right) dx \\ &= \frac{d}{dt}\int_{\mathbb{R}^d} \hat{\pi}_t(x) \log \frac{\hat{\pi}_t(x)}{\pi(x)}dx \\ &= \frac{d}{dt}H_\pi(\hat{\pi}_t).\end{aligned}$$

■

Proof of Corollary A.1. Recall that from Lemma 6, there holds

$$\frac{d}{dt}H_\pi(\hat{\pi}_t) = \int_{\mathbb{R}^d} \frac{\partial \hat{\pi}_t(x)}{\partial t}(1 + \log \hat{\pi}_t(x) - \log \pi)dx. \tag{B.5}$$

Let

$$F_t(x) = \hat{\pi}_t(x)E\left(h_\lambda(\theta_{k\lambda})\middle|\theta_t = x\right) + \nabla \hat{\pi}_t(x)$$

and

$$g_t(x) = 1 + \log \hat{\pi}_t - \log \pi.$$

Recall from Corollary 5 that

$$\frac{\partial \hat{\pi}_t}{dt}(x) = div_x(F_t)(x)$$

Since $\nabla \hat{\pi}_t = \hat{\pi}_t \nabla \log \hat{\pi}_t$ using the Lemma B.1,B.4 one deduces that there exists constants C, q ,r>0 independent of x, uniform in a small neighbourhood of t, such that

$$\max\{|F_t(x)g_t(x)|, |div(F_t)(x)g_t(x)|, |\langle F_t(x)\nabla g_t(x)\rangle|\} \leq C(1 + |x|^q)e^{-r|x|^2}. \tag{B.6}$$

We drop the dependence of the constants on t since we want to integrate with respect to x. Let $R > 0$ and $v(x)$ the normal unit vector on $\partial B(0, R)$. Due to (B.6)

$$\int_{\partial B(0,R)} \langle g_t(x)F_t(x), v(x)\rangle dx \leq R^d C(1 + |R|^q)e^{-r|R|^2}. \tag{B.7}$$

Since $div(F_t)g_t$, $\langle F_t, \nabla_x g_t\rangle$ are integrable (in view of (B.6)) applying the divergence theorem on $B(0, R)$ there holds

$$\int_{B(0,R)} div_x(F_t)(x)g_t(x)dx = \int_{\partial B(0,R)} \langle g_t(x)F_t(x), v(x)\rangle dx - \int_{B(0,R)} \langle F_t(x)\nabla_x g_t(x)\rangle dx. \tag{B.8}$$

As a result,

$$\begin{aligned}\int_{\mathbb{R}^d} div(F_t)(x)g_t(x)dx &= \lim_{R\to\infty} \int_{B(0,R)} div_x(F_t)(x)g_t(x)dx \\ &= \lim_{R\to\infty} \left(\int_{\partial B(0,R)} \langle g_t(x)F_t(x), v(x)\rangle dx - \int_{B(0,R)} \langle F_t(x)\nabla_x g_t(x)\rangle\right) dx \\ &= 0 - \lim_{R\to\infty} \int_{B(0,R)} \langle F_t(x)\nabla_x g_t(x)\rangle dx = -\int_{\mathbb{R}^d} \langle F_t(x)\nabla_x g_t(x)\rangle dx.\end{aligned}$$

■

Proof of theorem 4. Using Proposition A.1, for all $t \in [k\lambda, (k+1)\lambda]$,

$$\begin{aligned}
\frac{d}{dt}H_\pi(\hat{\pi}_t) &= -\int_{\mathbb{R}^d} \langle \hat{\pi}_t(x) E\left(h_\lambda(\theta_{k\lambda})\middle|\theta_t = x\right) + \nabla\hat{\pi}_t(x), \nabla\log\hat{\pi}_t(x) - \nabla\log\pi(x)\rangle dx \\
&= -\int_{\mathbb{R}^d} \hat{\pi}_t(x)\langle E\left(h_\lambda(\theta_{k\lambda})\middle|\theta_t = x\right) + \nabla\log\hat{\pi}_t(x), \nabla\log\hat{\pi}_t(x) - \nabla\log\pi(x)\rangle dx \\
&= -\int_{\mathbb{R}^d} \hat{\pi}_t(x)\langle E\left(h_\lambda(\theta_{k\lambda})\middle|\theta_t = x\right) + \nabla\log\pi, \nabla\log\hat{\pi}_t(x) - \nabla\log\pi(x)\rangle dx \\
&- \int_{\mathbb{R}^d} \hat{\pi}_t|\nabla\log\hat{\pi}_t(x) - \nabla\log\pi(x)|^2 dx \\
&= -I_\pi(\hat{\pi}_t) - \int_{\mathbb{R}^d} \hat{\pi}_t(x)\langle E\left(h_\lambda(\theta_{k\lambda}) - h(x)\middle|\theta_t = x\right), \nabla\log\hat{\pi}_t(x) - \nabla\log\pi(x)\rangle dx \\
&= -I_\pi(\hat{\pi}_t) - \int_{\mathbb{R}^d} \hat{\pi}_t(x)\langle E\left(h_\lambda(\theta_{k\lambda}) - h(\theta_t))\middle|\theta_t = x\right), \nabla\log\hat{\pi}_t(x) - \nabla\log\pi(x)\rangle dx \\
&\leq -I_\pi(\hat{\pi}_t) + \int_{\mathbb{R}^d} \hat{\pi}_t(x)\left|E\left(h_\lambda(\theta_{k\lambda}) - h(\theta_t))\middle|\theta_t = x\right)\right|^2 dx + \frac{1}{4}I_\pi(\hat{\pi}_t) \\
&= -\frac{3}{4}I_\pi(\hat{\pi}_t) + \int_{\mathbb{R}^d} \hat{\pi}_t(x)\left|\int_{\mathbb{R}^d} \hat{\pi}_{\theta_{k\lambda}|\theta_t}(y|x)(h_\lambda(y) - h(x))dy\right|^2 dx \\
&\leq -\frac{3}{4}I_\pi(\hat{\pi}_t) + \int_{\mathbb{R}^d} \hat{\pi}_t(x)\int_{\mathbb{R}^d} \hat{\pi}_{\theta_{k\lambda}|\theta_t}(y|x)\left|h_\lambda(y) - h(x)\right|^2 dydx \\
&= -\frac{3}{4}I_\pi(\hat{\pi}_t) + \mathbb{E}|h_\lambda(\theta_{k\lambda}) - h(\theta_t)|^2
\end{aligned}$$

where the first inequality was obtained using Young inequality and the second using Jensen's. ■

Proof of Lemma A.4. Let $t \in [k\lambda, (k+1)\lambda]$. First of all, one needs to bound the one step error $\mathbb{E}|\theta_t - \theta_{k\lambda}|^{2p}$ for different values of $p \in \mathbb{N}$.

$$\begin{aligned}
\mathbb{E}|\theta_t - \theta_{k\lambda}|^{2p} &\leq 2^{2p}\lambda^{2p}\mathbb{E}|h_\lambda(\theta_{k\lambda})|^{2p} + 2^p\lambda^p\mathbb{E}|Z|^{2p} \\
&\leq 2^p\lambda^p\mathbb{E}\left(4A^2|\theta_{k\lambda}|^a + 4(L^2 + A^2)\right)^p + 2^p\lambda^p\mathbb{E}|Z|^{2p} \\
&\leq \lambda^p C_{1,p}
\end{aligned}$$

where $C_{1,p} = \mathcal{O}\left(d^{p(2l+1)}\right)$, which is derived by the moment bounds of the Gaussian, the fact that $\mathcal{L}(\theta_{k\lambda}) = \mathcal{L}(\bar{\theta}_k^\lambda)$ and the moment bounds of the algorithm. ■

Proof of Lemma A.5. For every $x \in \mathbb{R}^d$,

$$|h_\lambda(x) - h(x)| = \left|(h(x) - A\frac{x}{(1+|x|^2)^{1-\frac{a}{2}}})(1 - \frac{1}{1+\sqrt{\lambda}|x|^{2l}})\right|^2 \leq \lambda\left|(|h(x)| + |x|)|x|^{2l}\right|^2$$

so

$$\mathbb{E}\left|h(\theta_{k\lambda}) - h_\lambda(\theta_{k\lambda})\right|^2 \leq \lambda\mathbb{E}\left|\left(|h(\bar{\theta}_k)| + |\bar{\theta}_k|\right)|\bar{\theta}_k|\right|^2 \leq 16\left(L^2(\bar{C}_{4l} + 1) + \bar{C}_{2l+1}\right)\lambda. \tag{B.9}$$

where the constants are given in Lemma A.3. In addition, using Assumption (A1) one deduces that

$$\begin{aligned}\mathbb{E}|h(\theta_{k\lambda}) - h(\theta_t)|^2 &\leq \mathbb{E}(1 + |\theta_{k\lambda} + |\theta_t|)^{2l'}|\theta_t - \theta_{k\lambda}|^2 \\ &\leq \sqrt{3^{4l'}(1 + \mathbb{E}|\theta_{k\lambda}^{4l'} + \mathbb{E}|\theta_{k\lambda} - \theta_t|^{4l'})}\sqrt{\mathbb{E}|\theta_{k\lambda} - \theta_t|^4} \\ &\leq \sqrt{3^{4l'}}\sqrt{1 + \lambda^{2l'}C_{1,2l'} + \sup_n \mathbb{E}|\bar{\theta}_n^{\lambda}|^{4l'}}\lambda\sqrt{C_{2,p}} \quad \text{(Lemma A.4} \\ &\leq \sqrt{1 + \lambda^{2l'}C_{1,2l'} + C_{2l'}}\lambda\sqrt{C_{2,p}}\end{aligned} \tag{B.10}$$

where the last step was derived by Lemma A.3. Combining (B.9) and (B.10), yields the result. ∎

Proof of Theorem 2.

$$\begin{aligned}\frac{d}{dt}H_\pi(\hat{\pi}_t) &\leq -\frac{3}{4}I_\pi(\hat{\pi}_t) + \beta\mathbb{E}|h_\lambda(\theta_{k\lambda} - h(\theta_t)|^2 \\ &\leq -\dot{c}H_\pi(\hat{\pi}_t) + 2\beta\mathbb{E}|h_\lambda(\theta_{k\lambda}) - h(\theta_{k\lambda})|^2 + 2\beta\mathbb{E}|h(\theta_{k\lambda}) - h(\theta_t)|^2 \\ &\leq -\dot{c}H_\pi(\hat{\pi}_t) + \beta\hat{C}\lambda\end{aligned}$$

where $\hat{C} = 2C_{onestep} + 2C_{tam}$ where the first term has been bounded using the Log-Sobolev inequality and the rest of the terms using the one-step error in Lemma A.4 and the taming error in Lemma A.5. Splitting the terms one obtains

$$\left(\frac{d}{dt}H_\pi(\hat{\pi}_t) + \dot{c}H_\pi(\hat{\pi}_t)\right)e^{\dot{c}t} \leq e^{\dot{c}t}\beta\hat{C}\lambda$$

Integrating over $[k\lambda, t]$ yields

$$e^{\dot{c}t}H_\pi(\hat{\pi}_t) - e^{\dot{c}k\lambda}H_\pi(\hat{\pi}_{k\lambda}) \leq \frac{\beta\hat{C}}{\dot{c}}\lambda(e^{\dot{c}t} - e^{\dot{c}k\lambda})$$

which implies

$$H_\pi(\hat{\pi}_t) \leq e^{\dot{c}(k\lambda - t)}H_\pi(\hat{\pi}_{k\lambda}) + \frac{\beta\hat{C}}{\dot{c}}\lambda(1 - e^{\dot{c}(k\lambda - t)}). \tag{B.11}$$

Setting $t = n\lambda$ and $k = (n-1)$ leads to

$$H_\pi(\hat{\pi}_{n\lambda}) \leq e^{-\dot{c}\lambda}H_\pi(\hat{\pi}_{(n-1)\lambda}) + \frac{\beta\hat{C}}{\dot{c}}\lambda(1 - e^{-\dot{c}\lambda})$$

so by iterating over n,

$$H_\pi(\hat{\pi}_{n\lambda}) \leq e^{-\dot{c}\lambda(n-1)}H_\pi(\pi_0) + \frac{\beta\hat{C}}{\dot{c}}\lambda$$

which completes the proof. ∎

Proof of Lemma A.2. Let $f = \pi/\hat{\pi}_t$ Then,

$$\begin{aligned}
\frac{1}{2}\int(\sqrt{\hat{\pi}_t}-\sqrt{\pi})^2dx &\leq \left(1-\mathbb{E}_\nu(\sqrt{f})\right)\left(1+\mathbb{E}_\nu(\sqrt{f})\right)\\
&\leq 1-\left(\mathbb{E}_\nu(\sqrt{f})\right)^2\\
&= \mathbb{E}_\nu((\sqrt{f})^2)-\left(\mathbb{E}_\pi(\sqrt{f})\right)^2\\
&= Var_\pi(\sqrt{f})\\
&\leq \frac{1}{C_P}\mathbb{E}_\pi|\nabla\sqrt{f}|^2\\
&\leq \frac{1}{4C_P}\mathbb{E}_\pi|\nabla f|^2/f\\
&= \frac{1}{4C_P}\int\left(\pi f|\frac{\nabla f}{f}|^2\right)dx\\
&= \frac{1}{4C_p}\int\left(\hat{\pi}_t|-\nabla\log\frac{\hat{\pi}_t}{\pi}|^2\right)dx\\
&= \frac{1}{C_p}I_\pi(\hat{\pi}_t).
\end{aligned} \tag{B.12}$$

In addition, since both $\hat{\pi}_t$ and π have finite polynomial moments, there holds

$$\begin{aligned}
W_2^2(\mathcal{L}(\theta_t),\pi) &= 2\int|x|^2|\pi(x)-\hat{\pi}_t(x)|dx\\
&\leq 2\left(\int|x|^4(\sqrt{\pi}+\sqrt{\hat{\pi}_t})^2dx\right)^{\frac{1}{2}}\left(\int(\sqrt{\hat{\pi}_t}-\sqrt{\pi})^2dx\right)^{\frac{1}{2}}\\
&\leq 4(\sqrt{\mathbb{E}_{\hat{\pi}_t}|x|^4}+\sqrt{\mathbb{E}_\pi|x|^4})\sqrt{I_\pi(\hat{\pi}_t)} \quad \text{derived from (B.12)}\\
&32(\sqrt{\sup\mathbb{E}|\bar{\theta}_n^\lambda|^4}+\sqrt{\mathbb{E}|\theta_t-\theta_{k\lambda}|^4}+\sqrt{\mathbb{E}_\pi|x|^4})\sqrt{I_\pi(\hat{\pi}_t)}\\
&\leq C\sqrt{I_\pi(\hat{\pi}_t)}
\end{aligned} \tag{B.13}$$

where the last step was derived from Lemmas A.4 and A.3 We are going to use our assumption to connect the relative entropy to W_2 distance. Since Assumption 3 holds and π has finite second moments, the HWI can be applied, so

$$\begin{aligned}
H_\pi(\hat{\pi}_t) &\leq \sqrt{I_\pi(\hat{\pi}_t)}W_2(\mathcal{L}(\theta_t),\pi)+\frac{\kappa}{2}W_2^2(\mathcal{L}(\bar{\theta}_n^\lambda),\pi)\\
&\leq \sqrt{2}(\sqrt{\mathbb{E}_\pi}|x|^2+\sqrt{\mathbb{E}|\theta_{k\lambda}-\theta_t|^2}+\sqrt{\mathbb{E}|\bar{\theta}_n^\lambda|^2})\sqrt{I_\pi(\hat{\pi}_t)}+\frac{\kappa}{2}W_2^2(\mathcal{L}(\bar{\theta}_n^\lambda),\pi).
\end{aligned} \tag{B.14}$$

Combining (B.13) with (B.14) yields the result. ∎

Proof of Proposition A.3. Let $\phi(t,x) = -\dot{c}_0x^2+k_2$ where $k_2 := 2C_1\lambda$. Then, from Corollary 2 there holds

$$\frac{dH_\pi(\hat{\pi}_t)}{dt} < \phi(H_\pi(\hat{\pi}_t),t).$$

Let $\delta < H_\pi^{-1}(\rho_k)/2$ Setting $g_\delta(t) = \left(H_\pi(\rho_k)^{-1}-\delta+\dot{c}_0(t-k\lambda)\right)^{-1}+k_2(t-k\lambda)$ one deduces

$$g_\delta'(t) = -\dot{c}_0\left(H_\pi(\rho_k)^{-1}-\delta+\dot{c}_0(t-k\lambda)\right)^{-2}+k_2 \tag{B.15}$$

Since $(H_\pi^{-1}(\rho_k)-\delta)^2 \leq \left(H_\pi(\rho_k)^{-1}-\delta+\dot{c}_0(t-k\lambda)\right)^2$ one obtains

$$g'_\delta(t)-\phi(g_\delta(t),t)\geq 0 > \frac{dH_\pi(\hat{\pi}_t)}{dt}-\phi(H_\pi(\hat{\pi}_t),t) \quad \forall t\in[k\lambda,(k+1)\lambda] \tag{B.16}$$

Using (B.16) and the fact that $g_\delta(k\lambda)=\left(H_\pi(\rho_k)^{-1}-\delta\right)^{-1} > H_\pi(\rho_k)=H_\pi(\hat{\pi}_{k\lambda})$ By comparison theorem for differential inequalities, see McNabb [24], there holds

$$(H_\pi(\rho_k)^{-1}+\dot{c}_0\lambda)^{-1}+2C_1\lambda^2=\lim_{\delta\to 0^+} g_\delta((k+1)\lambda)\geq H_\pi(\rho_{k+1}) \tag{B.17}$$

■

Proof of theorem 6. We begin the proof by noticing that

$$H_\pi(\rho_n)\leq \frac{1}{\dot{c}_0\lambda} \quad \forall n. \tag{B.18}$$

This will be done by induction. Since for $\lambda<\lambda_{max}$,

$$H_\pi(\rho_0)\leq \frac{1}{\dot{c}_0\lambda}$$

it holds for $n=0$. Suppose that

$$H_\pi(\rho_k)\leq \frac{1}{\dot{c}_0\lambda} \tag{B.19}$$

Then

$$H_\pi(\rho_{k+1})\leq \frac{1}{\dot{c}_0\lambda}\left(\dot{c}_0\lambda H_\pi(\rho_k)(1+\dot{c}_0\lambda H_\pi(\rho_k))^{-1}\right)+2C_1\lambda^2$$

Since the function $\phi(x)=\frac{x}{1+x}$ is increasing, then $\phi(\dot{c}_0\lambda H_\pi(\rho_k))<\phi(1)$ so

$$H_\pi(\rho_{k+1})\leq \frac{1}{2\dot{c}_0\lambda}+2C_1\lambda^2\leq \frac{1}{\dot{c}_0\lambda}.$$

which proves (B.18) by induction.
We proceed with two cases:
Case 1: $H_\pi(\rho_{k_0})\geq \frac{4c_1}{\dot{c}_0}\sqrt{\lambda} \quad \forall k_0\leq k$:
Making use of (B.18) and the inequality $\frac{1}{x+1}\leq(1-\frac{x}{2})$ for $x\leq 1$, one obtains

$$H_\pi(\rho_{k+1})\leq H_\pi(\rho_k)(1-\frac{\dot{c}_0}{2}\lambda H_\pi(\rho_k))+2C_1\lambda^2\leq H_\pi(\rho_k)(1-2C_1\lambda^{\frac{3}{2}})+2C_1\lambda^2$$

Summing over k one deduces

$$H_\pi(\rho_k)\leq H_\pi(\rho_0)(1-2C_1)\lambda^{\frac{3}{2}})^k+\sqrt{\lambda}. \tag{B.20}$$

Case 2: There exist $k_0\leq k$ such that $H_\pi(k_0)\leq\sqrt{\frac{4c_1}{\dot{c}_0}}\sqrt{\lambda}$. Suppose that $\frac{4c_1}{\dot{c}_0}\sqrt{\lambda}\geq H_\pi(\rho_{k_0})\geq \frac{1}{2}\frac{4c_1}{\dot{c}_0}\sqrt{\lambda}$. Then,

$$H_\pi(\rho_{k_0+1})\leq H_\pi(\rho_{k_0})-\frac{\dot{c}_0}{2}\lambda H_\pi^2(\rho_{k_0})+C_1\lambda^2\leq H_\pi(\rho_{k_0})$$

On the other hand, if $H_\pi(k_0)\leq\frac{1}{2}\frac{4c_1}{\dot{c}_0}\sqrt{\lambda}$. it is easy to see that $H_\pi(\rho_{k+1})\leq\frac{4c_1}{\dot{c}_0}\sqrt{\lambda}$.
This implies that

$$\exists k_0<k: H_\pi(\rho_{k_0})\leq\frac{4c_1}{\dot{c}_0}\sqrt{\lambda}\implies H_\pi(\rho_k)\leq\frac{4c_1}{\dot{c}_0}\sqrt{\lambda}$$

Combining case 1 and case 2 together yields the result. ■

Proof of Corollary 3. For the bound in total variation, using Theorem 6 and Pinsker's inequality gives the result.
For the bound in W_1 distance, using Lemma A.2 and Corollary 2.3 in [6], one deduces that

$$C_W := \frac{2}{\mu}(\frac{3}{2} + \log \mathbb{E}e^{\mu|\bar{\theta}_n^\lambda|}) < \infty$$

and

$$W_1(\mathcal{L}(\bar{\theta}_n^\lambda), \pi) \leq C_W \left(H_\pi(\rho_n) + H_\pi(\rho_n)^{\frac{1}{2}} \right).$$

Applying the bound on C_W in Lemma A.2 and Theorem 6 yields the result. ■

C. Proofs for the convergence of regularized scheme

C.1. Properties for the regularized potential.

Proof of Lemma A.6. It is easy to see that the function $G(x) = (r+1)|x|^{2r}$ is Locally Lipschitz since

$$J_G = (r+1)|x|^{2r} I_d + (r+1) r x^t x |x|^{2r-2}$$

then,

$$||J_G(x)|| \leq (r+1)^2 |x|^{2r}.$$

Using the mean value theorem

$$|\lambda G(x) - \lambda G(y)| \leq \lambda \int_0^1 ||J_G(tx + (1-t)y|||x-y|dt \leq \lambda (r+1)^2 (1 + |x| + |y|)^{2r} |x-y|.$$

As a result,

$$|\nabla u_{\mathrm{reg},\lambda}(x) - \nabla u_{\mathrm{reg},\lambda}(y)| \leq ((r+1)^2 + L)(1 + |x| + |y|)^{2r} |x-y| \quad \forall x, y \in \mathbb{R}^d.$$

It is also easy to see that the higher derivatives of $u_{\mathrm{reg},\lambda}$ have polynomial growth less than $2r+1$. With respect to the dissipativity it is easy to see that

$$\langle \nabla u_{\mathrm{reg},\lambda}(x), x \rangle \geq \langle \nabla u(x), x \rangle,$$

so Assumption (A2) is satisified with the same A and b. For the tamed scheme, by it is definition it easy to see that

$$|\nabla u_{r,\lambda}| \leq A + \sqrt{\lambda} + A|x|^{\frac{a}{2}} + \frac{(L+1)}{\sqrt{\lambda}}.$$

■

Proof of Lemma A.8. The proof starts by noticing that there exists R_1 depending on A, B of Assumption (A2) such that

$$\langle \nabla u_{\mathrm{reg},\lambda}(x), x \rangle \geq \frac{A}{2}|x| \quad \forall |x| \geq R_1.$$

In addition, picking a smooth Lyapunov function $W \geq 1$ such

$$W = e^{\frac{A}{4}|x|} \quad \forall |x| \geq R_1,$$

one deduces that for generator of the Langevin SDE with drift coeffient the regularized gradient,

$$\begin{aligned} LW(x) = \Delta W(x) - \langle \nabla W(x), \nabla u_{\mathrm{reg},\lambda}(x) &\leq \frac{A}{4} W \left(\frac{d-1}{|x|} + \frac{A}{4} - \langle \nabla u_{\mathrm{reg},\lambda}(x), x \rangle \} \right) \\ &\leq \frac{A}{4} W (\frac{d-1}{|x|} + \frac{A}{4} - \frac{A}{4}|x|) \end{aligned}$$

So there exists $R_0 \leq \mathcal{O}(d)$ such that

$$LW \leq -\theta W \quad \forall |x| \geq R_2 := \max\{R_0, R_1\}.$$

Setting $B = B(0, R_2)$ and $B_2 = B(0, R_2 + 2)$. Let a smooth function $\chi = \psi(|x|)$ (see Lemma B.13 of Li & Erdogdu [20] for the construction) such that $\chi = 0$ on B nad $\chi = 1$ on B_2^c and $|\nabla \chi| \leq 1$.

$$\begin{aligned}
\int \frac{-LW}{W} f^2 d\pi_{\text{reg}} &= \int \Gamma\left(\frac{f^2}{W}, W\right) d\pi_{\text{reg}} \\
&= 2\int \frac{f}{W}\Gamma(f, W) d\pi_{\text{reg}} - \int \frac{f^2}{W^2}\Gamma(W, W) d\pi_{\text{reg}} \\
&= -\int \left|\frac{f}{W}\nabla W - \nabla f\right|^2 d\pi_{\text{reg}} + \int \Gamma(f, f) d\pi_{\text{reg}} \\
&\leq \int \Gamma(f, f) d\pi_{\text{reg}}
\end{aligned} \tag{C.1}$$

Writing for a smooth f,

$$\begin{aligned}
\int f^2 d\pi_{\text{reg}} &= \int (f(1-\chi) + f\chi)^2 d\pi_{\text{reg}} \\
&\leq 2\int f^2(1-\chi)^2 d\pi_{\text{reg}} + 2\int f^2\chi^2 d\pi_{\text{reg}} \\
&\leq \frac{2}{\theta}\int \frac{-LW}{W} f^2(1-\chi)^2 d\pi_{\text{reg}} + 2\int_{B_2} f^2 d\pi_{\text{reg}} \\
&\leq \frac{2}{\theta}\int \Gamma(f(1-\chi), f(1-\chi)) d\pi_{\text{reg}} + 2\int_{B_2} f^2 d\pi_{\text{reg}}
\end{aligned}$$

. Since $\Gamma(fg, fg) \leq 2\left(f^2\Gamma(g, g) + g^2\Gamma(f, f)\right)$, we get:

$$\begin{aligned}
\int f^2 d\pi_{\text{reg}} &\leq \frac{4}{\theta}\int \Gamma(f, f) d\pi_{\text{reg}} + \frac{4}{\theta}\int f^2\Gamma(\chi, \chi) d\pi_{\text{reg}} + 2\int_{B_2} f^2 d\pi_{\text{reg}} \\
&\leq \frac{4}{\theta}\int \Gamma(f, f) d\pi_{\text{reg}} + \left(\frac{4}{\theta} + 2\right)\int_{B_2} f^2 d\pi_{\text{reg}}
\end{aligned} \tag{C.2}$$

Applying the previous inequality for $\tilde{f} = f - \int_{B_2} f d\pi_{\text{reg}}$ and using the fact that

$$Var_{\pi_{\text{reg}}}(f) \leq \int \tilde{f}^2 d\pi_{\text{reg}}$$

yields

$$\begin{aligned}
Var_{\pi_{\text{reg}}}(f) \leq \int \tilde{f}^2 d\pi_{\text{reg}} &\leq \frac{4}{\theta}\int \Gamma(\tilde{f}, \tilde{f}) d\pi_{\text{reg}} + \left(\frac{4}{\theta} + 2\right)\int_{B_2} \tilde{f}^2 d\pi_{\text{reg}} \\
&= \frac{4}{\theta}\int \Gamma(f, f) d\pi_{\text{reg}} + \left(\frac{4}{\theta} + 2\right)\int_{B_2} \tilde{f}^2 d\pi_{\text{reg}}
\end{aligned} \tag{C.3}$$

When restricted to the ball B_2 $u_{\text{reg},\lambda}$ is a bounded petrubation u on the same ball since

$$|u(x) - u_{\text{reg},\lambda}(x)| \leq \lambda(R_2 + 2)^{2r+2} \quad \forall x \in B_2.$$

Using Hooley-Strook pertubation theorem one deduces that π_{reg} satisfies Poincare inequality when resticted to B_2 with constant $k_{B_2}^{-1} \leq e^{2\lambda(R_2+2)^{2r+2}} C_P^{-1} \leq 3C_P - 1$. Thus,

$$\int_{B_2} \tilde{f}^2 d\pi_{\text{reg}} \leq k_{B_2}\int \Gamma(f, f) d\pi_{\text{reg}}.$$

Applying this to (C.3) completes the proof. ■

Lemma C.1. *The function u given by $u_{r,\lambda}(x) := u(x) + \lambda|x|^{2r+2}$ satisfies*

$$\langle \nabla u_{r,\lambda}(x) - \nabla u_{r,\lambda}(y), x-y\rangle \geq \Big(c_1(|x|^{2r}+|y|^{2r}) - c_2(|x|^{l'}+|y|^{l'}) - c_3\Big)\,|x-y|^2 \quad \forall x,y\in\mathbb{R}^d.$$

where $c_1 := \lambda(r+1)$, $c_2 = c_3 = L$.

The proof follows by using Assumption (A1) and the fact that the regularized term $2r$ dominates l for large values. This will yield a lower bound for the minimum eigenvalue of $\nabla^2 u_{\mathrm{reg},\lambda}$.

Proof. Let $f(x) = |x|^{2r+2}$. Then $\nabla f(x) = 2(r+1)|x|^{2r}x$. Writing

$$\begin{aligned}
\langle \nabla f(x) - \nabla f(y), x-y\rangle &= \langle \nabla f(x), x\rangle + \langle \nabla f(y), y\rangle - \langle \nabla f(x), y\rangle - \langle \nabla f(y), x\rangle \\
&= (2r+2)(|x|^{2r+2}+|y|^{2r+2}) - (r+1)(|x|^{2r}+|y|^{2r})2\langle x,y\rangle \\
&= (2r+2)(|x|^{2r+2}+|y|^{2r+2}) \\
&\quad + (r+1)(|x|^{2r}+|y|^{2r})\left(|x-y|^2 - |x|^2 - |y|^2\right) \\
&= (r+1)\left(|x|^{2r+2}+|y|^{2r+2} - |x|^{2r}|y|^2 - |y|^{2r}|x|^2\right) \\
&\quad + (r+1)(|x|^{2r}+|y|^{2r})|x-y|^2.
\end{aligned}$$

Since

$$\begin{aligned}
|x|^{2r+2}+|y|^{2r+2} - |x|^{2r}|y|^{2r+2} - |y|^{2r}|x|^{2r+2} &= |x|^2(|x|^{2r}-|y|^{2r}) - |y|^2(|x|^{2r}-|y|^{2r}) \\
&= (|x|^2-|y|^2)(|x|^{2r}-|y|^{2r}) \\
&\geq 0,
\end{aligned}$$

one deduces

$$\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq (r+1)(|x|^{2r}+|y|^{2r})|x-y|^2 \quad \forall x,y\in\mathbb{R}^d. \tag{C.4}$$

Noting that by the gradient local Lipschitz assumption on g, there holds

$$\langle \nabla u(x) - \nabla u(y), x-y\rangle \geq -L(1+|x|^l+|y|^l)|x-y|^2 \quad \forall x,y\in\mathbb{R}^d,$$

the result immediately follows. ■

Proof of Proposition A.4. It is easy to see that the regularized measure satisfies a $2-$ dissipativity condition with constant $A_{reg} = \lambda^{\frac{1}{r+1}}$ i.e

$$\langle \nabla u_{r,\lambda}(x), x\rangle \geq A_{reg}|x|^2 - (b+1). \tag{C.5}$$

In addition, using Lemma C.1, it is easy to see that when $|x|, |y| \geq \frac{L}{r+1}(\frac{1}{\lambda})$

$$\langle \nabla u_{r,\lambda}(x) - \nabla u_{r,\lambda}(y), x-y\rangle \geq 0$$

so one concludes that

$$\langle \nabla u_{r,\lambda}(x) - \nabla u_{r,\lambda}(y), x-y\rangle \geq -K_\lambda|x-y|^2 \quad \forall x,y\in\mathbb{R}^d$$

which leads to

$$\nabla^2 u_{r,\lambda}(x) \geq -K_\lambda I_d \quad \forall x\in\mathbb{R}^d. \tag{C.6}$$

Let $W := e^{\frac{A_{reg}|x|^2}{4}}$. Then, since $\nabla W = W\frac{A_{reg}}{2}x$ and $\Delta W \leq (\frac{A_{reg}d}{2} + \frac{A_{reg}^2 d}{4})W$ one observes that

$$LW = \Delta W - \langle \nabla u_{r,\lambda}(x), \nabla W\rangle = (\frac{A_{reg}d}{2} + \frac{A_{reg}^2 d}{4} - \frac{A_{reg}}{2}|x|^2)W \tag{C.7}$$

Since π_{reg} also satisfies a Poincare inequality with consant $C_{P,r}$ using Theorem 3.15 in [25], it can be upgraded to a Log-Sobolev inequality with constant

$$\begin{aligned}\frac{1}{C_{LS}} \leq 2\sqrt{\frac{1}{A_{reg}}\left(\frac{1}{2}+\frac{+\frac{A_{reg}^2 d}{4}-\frac{A_{reg}}{2}+\frac{A_{reg}}{2}\pi_{\mathrm{reg}}\left(|x|^2\right)}{C_{P,reg}}\right)} \\ +\frac{K_\lambda}{A_{reg}\lambda}+\frac{K_\lambda\left(\frac{A_{reg}d}{2}+\frac{A_{reg}^2 d}{4}+\frac{A_{reg}}{2}\pi_{\mathrm{reg}}\left(|x|^2\right)\right)+2\frac{A_{reg}}{2}}{C_{P,reg}}\end{aligned} \tag{C.8}$$

where $\pi_{\mathrm{reg}}(|x|^2)$ is given by Lemma C.6. ■

Proof of A.7. The proof is the same, as in the proof of the moment bounds for the regularized potential, since all we need is the $a-$ dissipativity and the growth condition which are almost the same. It is done through providing first exponential moments and then produce polynomial. ■

Lemma C.2. *There holds*

$$\mathbb{E}|x_t - x_{k\lambda}|^{2p} \leq \mathcal{O}(\lambda^p)$$

Proof. The proof follows in the same way is the respective one for the unregulazized algorithm. ■

Lemma C.3. *There holds*

$$\mathbb{E}|\nabla u_{r,\lambda}(x_{k\lambda}) - h_{r,\lambda}(x_{k\lambda})|^2 \leq C_{tam}^{reg}\lambda$$

where $C_{tam}^{reg} \leq \mathcal{O}(d^{4r+2})$.

Proof. Writing

$$\begin{aligned}|\nabla u_{r,\lambda}(x_{k\lambda}) - h_{r,\lambda}(x_{k\lambda})|^2 &\leq 2|h(x_{k\lambda}) - h_\lambda(x_{k\lambda})|^2 + 2\lambda^2\left|(r+1)x_{k\lambda}|x_{k\lambda}|^{2r}(1-\frac{1}{1+\sqrt{\lambda}|x_{k\lambda}|^{2r+1}}\right|^2 \\ &\leq 2|h(x_{k\lambda}) - h_\lambda(x_{k\lambda})|^2 + 2\lambda^2(r+1)^2|x_{k\lambda}|^{4r+2}.\end{aligned}$$

Taking expectations, the first term can be treated as in the proof of Lemma A.5 with the moment bounds in Lemma A.7. ■

Lemma C.4. *There holds*

$$\mathbb{E}|\nabla u_{r,\lambda}(x_t) - \nabla u_{r,\lambda}(x_{k\lambda}|^2 \leq C_{onestep}^{reg}\lambda$$

Proof. The proof follows in the same way as the respective one for the unregularized algorithm with the new moment bounds, and setting the Local Lipschitz constants as in Lemma A.6. ■

Lemma C.5. *Let x_t the continuous time interpolation of the algorithm. Then, for $\lambda < \lambda_{\max}$ and for every $t \in [k\lambda, (k+1)\lambda]$,$k \in \mathbb{N}$, there holds*

$$\frac{d}{dt}H_\pi(\hat{\pi}_t^{reg}) \leq -\frac{3}{4}I_\pi(\hat{\pi}_t^{reg}) + \mathbb{E}|\nabla u_{r,\lambda}(x_t) - h_{r,\lambda}(x_{k\lambda})|^2$$

.

Proof. Since the regularized potential has the same key properties as the unregularized, the interpolation inequality holds with exactly the same arguments. ■

Lemma C.6. *Let $p \in \mathbb{N}$. There holds*

$$\mathbb{E}_{\pi}|x|^{2p} \leq C_{\pi}$$

and

$$\mathbb{E}_{\pi_{\text{reg}}}|x|^{2p} \leq C_{\pi_{\text{reg}}}$$

where $C_{\pi_{\text{reg}}}$ and C_{π} are $\mathcal{O}(d^{2p})$.

Proof. Using the fact that both measures satisfy the dissipativity condition with constant a we will proceed with the same arguments. We will show it only for π. Since π and π_{reg} satisfy a Poincare inequality they have finite polynomial moments of all orders.

$$|x|^{2p-1}\langle h(x), x\rangle \geq |x|^{2p-1}(A|x|^{a} - b) \geq A|x|^{2p} - A - b|x|^{2p-1}. \tag{C.9}$$

Setting $V(x) = |x|^{2p}$ one notices that $\nabla V(x) = 2p|x|^{2p-1}x$ and $\Delta V = (2pd+4(p-1)p)|x|^{2p-2}$. Since π is the invariant measure of the Langevin SDE with generator

$$LV = \Delta V - \langle V, h\rangle,$$

there holds

$$(2pd+4(p-1)p)\mathbb{E}_{\pi}|x|^{2p-2} = \mathbb{E}_{\pi}\Delta V(x) = \mathbb{E}_{\pi}\langle V(x), h(x)\rangle \geq 2p\left(A\mathbb{E}_{\pi}|x|^{2p} - A - b\mathbb{E}_{\pi}|x|^{2p-1}\right).$$

Iterating over $2p$ yields the result. ∎

Proof of Theorem 7. Setting $\dot{c} = \frac{3}{2}C_{LSI}$ one obtains

$$\begin{aligned}
\frac{d}{dt}H_{\pi_{\text{reg}}}(\hat{\pi}_t^{reg}) &\leq -\frac{3}{4}I_{\pi_{\text{reg}}}(\hat{\pi}_t^{reg}) + \mathbb{E}|h_{r,\lambda}(x_{k\lambda} - \nabla u_{r,\lambda}(x_t)|^2 \\
&\leq -\dot{c}H_{\pi_{\text{reg}}}(\hat{\pi}_t^{reg}) + 2\mathbb{E}|h_{r,\lambda}(x_{k\lambda}) - \nabla u_{r,\lambda}(x_{k\lambda})|^2 + 2\mathbb{E}|\nabla u_{r,\lambda}(x_{k\lambda}) - \nabla u_{r,\lambda}(x_t)|^2 \\
&\leq -\dot{c}H_{\pi_{\text{reg}}}(\hat{\pi}_t^{reg}) + \hat{C}\lambda
\end{aligned}$$

where $\hat{C}$ depends polynomially on the dimension, where the first term has been bounded using the Log-Sobolev inequality and the rest of the terms using the one-step error and taming, as in the unregularized case . Splitting the terms one obtains

$$\left(\frac{d}{dt}H_{\pi_{\text{reg}}}(\hat{\pi}_t^{reg}) + \dot{c}H_{\pi_{\text{reg}}}(\hat{\pi}_t^{reg})\right)e^{\dot{c}t} \leq e^{\dot{c}t}\hat{C}\lambda$$

Integrating over $[k\lambda, t]$ yields

$$e^{\dot{c}t}H_{\pi_{\text{reg}}}(\hat{\pi}_t^{reg}) - e^{\dot{c}k\lambda}H_{\pi_{\text{reg}}}(\hat{\pi}_{k\lambda}^{reg}) \leq \frac{\hat{C}}{\dot{c}}\lambda(e^{\dot{c}t} - e^{\dot{c}k\lambda})$$

which implies

$$H_{\pi_{\text{reg}}}(\hat{\pi}_t^{reg}) \leq e^{\dot{c}(k\lambda - t)}H_{\pi_{\text{reg}}}(\hat{\pi}_{k\lambda}^{reg}) + \frac{\hat{C}}{\dot{c}}\lambda(1 - e^{\dot{c}(k\lambda - t)}). \tag{C.10}$$

Setting $t = n\lambda$ and $k = (n-1)$ leads to

$$H_{\pi_{\text{reg}}}(\hat{\pi}_{n\lambda}^{reg}) \leq e^{-\dot{c}\lambda}H_{\pi_{\text{reg}}}(\hat{\pi}_{(n-1)\lambda}^{reg}) + \frac{\hat{C}}{\dot{c}}\lambda(1 - e^{-\dot{c}\lambda})$$

so by iterating over n,

$$H_{\pi_{\text{reg}}}(\hat{\pi}_{n\lambda}^{reg}) \leq e^{-\dot{c}\lambda(n-1)}H_{\pi_{\text{reg}}}(\rho_0) + \frac{\hat{C}}{\dot{c}}\lambda$$

Noticing that

$$\begin{aligned}\int \log \frac{\rho_n^{\mathrm{reg}}}{\pi} d\rho_n^{\mathrm{reg}} &= \int \log \frac{\rho_n^{\mathrm{reg}}}{\pi_{\mathrm{reg}}} d\rho_n^{\mathrm{reg}} + \int \log \frac{\pi_{\mathrm{reg}}}{\pi} d\rho_n^{\mathrm{reg}} \\ &= H_{\pi_{\mathrm{reg}}}(\rho_n^{\mathrm{reg}}) + \int \log \frac{\pi_{\mathrm{reg}}}{\pi} d(\rho_n^{\mathrm{reg}} - \pi) - H_{\pi_{\mathrm{reg}}}(\pi) \\ &\leq H_{\pi_{\mathrm{reg}}}(\rho_n^{\mathrm{reg}}) + \lambda \mathbb{E}_{\rho_n^{\mathrm{reg}}}[|x|^{2r+2}] + \lambda \mathbb{E}_{\pi}[|x|^{2r+2}].\end{aligned}$$

The result follows by Lemma C.6. ∎

Proof of Corollary 4. Using Pinsker's inequality and the bound in Theorem 7 one obtains the bound in total variation. Recall that for the W_2 distance, since π_{reg} satisfies a Log-Sobolev inequality then, it satisfies a Talagrand inequality with same constant.

$$\begin{aligned}W_2(\mathcal{L}(\bar{x}_n^\lambda, \pi) &\leq W_2(\mathcal{L}(\bar{x}_n^\lambda, \pi_{\mathrm{reg}}) + W_2(\pi, \pi_{\mathrm{reg}}) \\ &\leq \sqrt{2C_{\mathrm{LSI}}^{-1}(H_{\pi_{\mathrm{reg}}}(\rho_n^{\mathrm{reg}})} + 2C_{\mathrm{LSI}}^{-1}\sqrt{I_{\pi_{\mathrm{reg}}}(\pi)})\end{aligned}$$

The first term can be bounded by Theorem 7 while the second term is

$$\sqrt{\int |\nabla \log \pi_{\mathrm{reg}}(x) - \nabla \log \pi(x)|^2 d\pi} \leq \lambda \sqrt{\mathbb{E}_\pi[|x|^{4r+2}]}.$$

Using the bound on the Log Sobolev constant and Lemma C.6 leads to the result. ∎

Acknowledgments

This research was supported in part by the French National Research Agency (ANR) in the framework of the PEPR IA FOUNDRY project (ANR-23-PEIA-0003), the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003). PM is also a member of the Archimedes Research Unit, Athena RC, Department of Mathematics, University of Athens. This research was also supported in part by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

[1] Bakry, D. and Émery, M. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pp. 177–206. Springer, 2006.

[2] Bakry, D., Barthe, F., Cattiaux, P., and Guillin, A. A simple proof of the poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.

[3] Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

[4] Balasubramanian, K., Chewi, S., Erdogdu, M. A., Salim, A., and Zhang, S. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. In *Conference on Learning Theory*, pp. 2896–2923. PMLR, 2022.

[5] Barkhagen, M., Chau, N. H., Moulines, É., Rásonyi, M., Sabanis, S., and Zhang, Y. On stochastic gradient langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.

[6] Bolley, F. and Villani, C. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pp. 331–352, 2005.

[7] Brosse, N., Durmus, A., Moulines, É., and Sabanis, S. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

[8] Cattiaux, P., Guillin, A., and Zitt, P. A. Poincaré inequalities and hitting times. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pp. 95–118, 2013.

[9] Chau, N. H., Moulines, É., Rásonyi, M., Sabanis, S., and Zhang, Y. On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986, 2021.

[10] Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

[11] Chewi, S., Erdogdu, M. A., Li, M. B., Shen, R., and Zhang, M. Analysis of langevin monte carlo from poincar\'e to log-sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

[12] Dalalyan, A. S. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[13] Durmus, A. and Moulines, E. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

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

[15] Erdogdu, M. A. and Hosseinzadeh, R. On the convergence of langevin monte carlo: The interplay between tail growth and smoothness. In *Conference on Learning Theory*, pp. 1776–1822. PMLR, 2021.

[16] Erdogdu, M. A., Hosseinzadeh, R., and Zhang, S. Convergence of langevin monte carlo in chi-squared and rényi divergence. In *International Conference on Artificial Intelligence and Statistics*, pp. 8151–8175. PMLR, 2022.

[17] Hutzenthaler, M., Jentzen, A., and Kloeden, P. E. Strong and weak divergence in finite time of euler's method for stochastic differential equations with non-globally lipschitz continuous coefficients. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 467 (2130):1563–1576, 2011. ISSN 1364-5021.

[18] Hutzenthaler, M., Jentzen, A., and Kloeden, P. E. Strong convergence of an explicit numerical method for sdes with nonglobally lipschitz continuous coefficients. *Ann. Appl. Probab.*, 22(4):1611–1641, 08 2012.

[19] Johnston, T., Lytras, I., and Sabanis, S. Kinetic langevin mcmc sampling without gradient lipschitz continuity–the strongly convex case. *arXiv preprint arXiv:2301.08039*, 2023.

[20] Li, M. B. and Erdogdu, M. A. Riemannian langevin algorithm for solving semidefinite programs. *arXiv preprint arXiv:2010.11176*, 2020.

[21] Lovas, A., Lytras, I., Rásonyi, M., and Sabanis, S. Taming neural networks with tusla: Nonconvex learning via adaptive stochastic gradient langevin algorithms. *SIAM Journal on Mathematics of Data Science*, 5(2):323–345, 2023.

[22] Lytras, I. and Sabanis, S. Taming under isoperimetry. *arXiv preprint arXiv:2311.09003*, 2023.

[23] Majka, M. B., Mijatović, A., and Szpruch, Ł. Nonasymptotic bounds for sampling algorithms without log-concavity. *The Annals of Applied Probability*, 30(4):1534–1581, 2020.

[24] McNabb, A. Comparison theorems for differential equations. *Journal of mathematical analysis and applications*, 119(1-2):417–428, 1986.

[25] Menz, G. and Schlichting, A. Poincaré and logarithmic sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, 42(5):1809–1884, 2014.

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

[27] Mousavi-Hosseini, A., Farghly, T. K., He, Y., Balasubramanian, K., and Erdogdu, M. A. Towards a complete analysis of langevin monte carlo: Beyond poincaré inequality. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1–35. PMLR, 2023.

[28] Neufeld, A., En, M. N. C., and Zhang, Y. Non-asymptotic convergence bounds for modified tamed unadjusted langevin algorithm in non-convex setting. *arXiv preprint arXiv:2207.02600*, 2022.

[29] Nguyen, D., Dang, X., and Chen, Y. Unadjusted langevin algorithm for non-convex weakly smooth potentials. *arXiv preprint arXiv:2101.06369*, 2021.

[30] Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703, 2017.

[31] Sabanis, S. A note on tamed euler approximations. *Electron. Commun. Probab.*, 18(47):1–10, 2013.

[32] Sabanis, S. Euler approximations with varying coefficients: the case of superlinearly growing diffusion coefficients. *Ann. Appl. Probab.*, 26(4):2083–2105, 2016.

[33] Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

[34] Zhang, Y., Akyildiz, Ö. D., Damoulas, T., and Sabanis, S. Nonasymptotic estimates for stochastic gradient langevin dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87(2):25, 2023.

[c] Corresponding author.

[*] Archimedes/Athena RC, Greece.

[◇] The University of Edinburgh, Edinburgh, UK.
Email address: i.lytras@sms.ed.ac.uk

[♯] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France.
Email address: panayotis.mertikopoulos@imag.fr