Langevin Monte Carlo and JKO splitting

Espen Bernton EBERNTON@G.HARVARD.EDU
Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA.

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Algorithms based on discretizing Langevin diffusion are popular tools for sampling from high-dimensional distributions. We develop novel connections between such Monte Carlo algorithms, the theory of Wasserstein gradient flow, and the operator splitting approach to solving PDEs. In particular, we show that a proximal version of the Unadjusted Langevin Algorithm corresponds to a scheme that alternates between solving the gradient flows of two specific functionals on the space of probability measures. Using this perspective, we derive some new non-asymptotic results on the convergence properties of this algorithm.

Keywords: Langevin Monte Carlo, Fokker–Planck, Wasserstein gradient flow, operator splitting, proximal operators

1. Introduction

In this paper, we shed new light on Langevin-based Monte Carlo algorithms by drawing connections to the Wasserstein gradient flow literature and the operator splitting approach to solving PDEs. In a seminal paper, Jordan et al. (1998) expressed the solution of the Fokker–Planck equation as the gradient flow of the relative entropy functional (otherwise known as the KL-divergence) with respect to the 2-Wasserstein distance. Their constructive proof used a time discretization approach that has since become known as the JKO scheme. We show that applying the JKO scheme in conjunction with a splitting approach to solving the Fokker–Planck equation reduces to a proximal version of the Unadjusted Langevin Algorithm. Our proofs rely heavily on the theory developed by Ambrosio et al. (2005), and have the benefit of holding for potentials that are not necessarily differentiable. In turn, this allows us to provide some new results regarding the convergence of the algorithm. Our work is related to Durmus et al. (2016), and we will make comparisons to their theoretical results.

To motivate the use of Langevin-based Monte Carlo algorithms, consider a log-concave target distribution π, given in terms of the Lebesgue density $\pi(x) = Z^{-1}e^{-V(x)}$, where $V : \mathbb{R}^d \to \mathbb{R}$ is a convex function, $d \in \mathbb{N}$ is an integer, and Z is the normalizing constant. In the case where V is differentiable, we can associate with it the Langevin diffusion, given in terms of the Itô stochastic differential equation

$$dX(t) = -\nabla V(X(t))dt + \sqrt{2}dW(t), \quad X(0) = X_0 \sim \rho_0. \tag{1}$$

It represents the position $X(t) \in \mathbb{R}^d$ of a particle at time $t > 0$, initialized at the random location $X_0 \sim \rho_0$, with drift according to the gradient of the potential V and subject to random perturbations $dW(t)$. The process $W(t)$ is the standard Wiener process. The density of $X(t)$ at time t, written $\rho(t)$, satisfies the linear Fokker–Planck equation:

$$\frac{d\rho}{dt} = \text{div}(\rho\nabla V) + \Delta\rho, \quad \rho(0) = \rho_0. \tag{2}$$

A classical result says that under quite weak convexity and smoothness conditions on V, the unique stationary solution of (2) is equal to π, and that convergence to π is exponentially fast (see for example Pavliotis, 2014, Chapter 4). These attractive properties have spawned a range of sampling algorithms targeting π based on time discretizations of the process in (1). Notably, the Unadjusted Langevin Algorithm (ULA) and its Metropolis adjusted counterpart MALA have received much attention.

The Unadjusted Langevin Algorithm is simply an explicit Euler discretization of (1): for a time-step $h > 0$ and for $k \geq 0$,

$$X_h^{k+1} = X_h^k - h\nabla V(X_h^k) + \sqrt{2h}\eta^{k+1}, \quad X_h^0 = X_0, \tag{3}$$

where $(\eta^k)_{k\geq 1}$ is a sequence of independent $\mathcal{N}(0, \mathcal{I}_d)$ random variables and $\mathcal{I}_d$ is the d-dimensional identity matrix. In MALA, X_h^{k+1} is either accepted or rejected in a Metropolis step with the purpose of removing the asymptotic bias of ULA stemming from discretization error.

Originating with Roberts and Tweedie (1996), there has been a lot of interest in quantifying the performance of these algorithms, with early work primarily focusing on MALA (see e.g. Jarner and Hansen, 2000; Roberts and Stramer, 2002; Pillai et al., 2012; Xifara et al., 2014). It was not until Dalalyan (2014), who gave precise bounds for the total variation distance between the law of X_h^k and π in terms of d, k, and h, that ULA garnered similar attention. His results were further improved and extended to other metrics and discrepancies by Durmus and Moulines (2016b, 2017); Cheng and Bartlett (2017); Dalalyan (2017). For instance, Dalalyan and Karagulyan (2017) show that if V is strongly convex and has Lipschitz continuous gradient, then $\Omega(d/\varepsilon^2)$ iterations are sufficient for ULA to achieve an error of ε in the 2-Wasserstein distance. Similar results also hold in situations where only a (sufficiently regular) approximation of the gradient is available.

In what follows, we will view Langevin-based Monte Carlo through the lens of Wasserstein gradient flow, and show that this perspective can lead to interesting results on the computational complexity of such algorithms. Wasserstein gradient flow was also used by Cheng and Bartlett (2017) as a theoretical tool to study ULA, but our approach makes closer connections to the operator splitting literature, and as such leads to different results. We hope that further connections can have methodological implications in these fields, by considering the wide variety of JKO schemes, splitting schemes, and Langevin Monte Carlo algorithms that exist.

The rest of this paper is structured as follows. Section 1.1 defines the notation and states some important definitions, Section 2 reviews some concepts from the Wasserstein gradient flow literature, Section 3 briefly discusses the operator splitting approach to solving PDEs, Section 4 establishes connections between Wasserstein gradient flow, operator splitting and Langevin Monte Carlo and includes some convergence results on the proximal version of the ULA algorithm, and Section 5 concludes. Proofs are given in the Appendix.

1.1. Notation and definitions

Let $\|\cdot\|_p$ be the ℓ_p-norm on $\mathbb{R}^d$, unless $p = 2$, in which case it reduces to the Euclidean distance and is denoted by $\|\cdot\|$. Define $\mathcal{P}_2(\mathbb{R}^d)$ to be the set of probability measures on $\mathbb{R}^d$ with finite second moments with respect to the Euclidean distance. The 2-Wasserstein distance is a metric on $\mathcal{P}_2(\mathbb{R}^d)$, and is for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ defined by

$$\mathcal{W}_2(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\gamma(x, y) \right)^{\frac{1}{2}}, \tag{4}$$

where $\Gamma(\mu, \nu)$ is the set of all joint distributions with marginals μ and ν. A desirable feature of the 2-Wasserstein distance is that $\mathcal{W}_2(\mu_n, \mu) \to 0$ as $n \to \infty$ if and only if μ_n converges weakly to μ and the corresponding sequence of second moments also converges (Villani, 2008, Theorem 6.9).

The entropy and potential energy functionals, $\rho \mapsto \mathcal{H}(\rho)$ and $\rho \mapsto \mathcal{V}(\rho)$ respectively, are given by

$$\mathcal{H}(\rho) = \begin{cases} \int \log \rho d\rho & \text{for } \rho \ll \mu_{\text{Leb}}, \\ +\infty & \text{otherwise,} \end{cases} \tag{5}$$

where μ_{Leb} denotes the Lebesgue measure on $\mathbb{R}^d$, and

$$\mathcal{V}(\rho) = \int V d\rho. \tag{6}$$

The relative energy functional $\rho \mapsto \mathcal{H}(\rho|\pi)$, also called the KL-divergence, is given by

$$\mathcal{H}(\rho|\pi) = \mathcal{H}(\rho) + \mathcal{V}(\rho) + \log Z. \tag{7}$$

An important concept in optimal transport, which will play a significant role later, is the notion of displacement convexity. A functional $\rho \mapsto \mathcal{F}(\rho)$ is said to be λ-displacement convex for some $\lambda \in \mathbb{R}$ if, for all $t \in [0, 1]$,

$$\mathcal{F}(\mu_t) \leq (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1) - \frac{\lambda}{2} t(1-t)\mathcal{W}_2^2(\mu_0, \mu_1) \tag{8}$$

for any constant speed geodesic $\mu : [0, 1] \to \mathcal{P}_2(\mathbb{R}^d)$. A curve $\mu : [0, 1] \to \mathcal{P}_2(\mathbb{R}^d)$ is a constant speed geodesic if, for any $0 \leq s \leq t \leq 1$, we have that $\mathcal{W}_2(\mu_s, \mu_t) = (t-s)\mathcal{W}_2(\mu_0, \mu_1)$.

We use the following notation for the density of a Gaussian distribution with zero mean and covariance matrix $2t\mathcal{I}_d$:

$$\phi_t(x) = \frac{1}{(4\pi t)^{d/2}} \exp\left(-\frac{\|x\|^2}{4t}\right). \tag{9}$$

By a Markov operator, we mean a linear functional R that maps the set of non-negative Lebesgue integrable functions into itself. A family of Markov operators $(R_t)_{t \geq 0}$ is called a Markov semigroup if R_0 is the identity map, $R_{t+s} = R_t R_s$ for any $s, t \geq 0$, and the map $t \mapsto R_t f$ is continuous for any non-negative and Lebesgue integrable f.

2. Wasserstein gradient flow

The theory of gradient flows in the space of probability measures was pioneered by Ambrosio, Gigli and Savaré in their book Ambrosio et al. (2005), generalizing the variational structure Jordan et al. (1998) had used to describe the diffusion and Fokker–Planck equations. With Langevin Monte Carlo in mind, we provide only a brief introduction to this theory, and refer to the aforementioned references and the accessible review of Santambrogio (2016) for further details.

We first consider continuous time flows, which will lead to a useful perspective on generalizations of the continuous time processes in (1) and (2). Secondly, we consider the time discretizations through which the existence and uniqueness of gradient flows are typically established. Although they were originally introduced as theoretical tools in the literature, it will later become clear that Langevin Monte Carlo in fact numerically approximates such a time discretization.

2.1. Continuous time flows

In Euclidean space, a curve $x : [0, \infty) \to \mathbb{R}^d$ is the gradient flow, or steepest descent, of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ if

$$\frac{dx}{dt} = -\nabla f(x), \quad x(0) = x_0. \tag{10}$$

By analogy, one can interpret the gradient flow of a functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ to be a curve $\rho : [0, \infty) \to \mathcal{P}_2(\mathbb{R}^d)$ that satisfies

$$\frac{d\rho}{dt} = -\nabla_{\mathcal{W}_2} \mathcal{F}(\rho), \quad \rho(0) = \rho_0, \tag{11}$$

for some generalized notion of gradient $\nabla_{\mathcal{W}_2}$, in terms of the $\mathcal{W}_2$ metric. For sufficiently regular ρ and $\mathcal{F}$, $\nabla_{\mathcal{W}_2}\mathcal{F}(\rho)$ corresponds to $-\mathrm{div}(\rho \nabla \frac{\delta \mathcal{F}}{\delta \rho})$, where $\delta\mathcal{F}/\delta\rho$ is the first variation of $\mathcal{F}$. Applied to the functional of interest, namely $\mathcal{F}(\rho) = \mathcal{H}(\rho|\pi)$, one has that $\delta\mathcal{F}/\delta\rho = V + \log\rho + 1$. Thus, if V is differentiable one recovers (2) (see e.g. Ambrosio et al., 2005, Lemma 10.4.1).

Due to the technically challenging nature of defining Wasserstein gradients this way when V is not differentiable, we instead adopt the definition given in Ambrosio et al. (2009), inspired by the characterization of gradient flows in terms of evolution variational inequalities (EVIs) shown in Ambrosio et al. (2005, Theorem 11.1.4). In particular, we say that a continuous curve $\rho : (0, +\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ is a gradient flow of a λ-displacement convex functional $\mathcal{F}$ if

$$\frac{d}{dt}\frac{1}{2}\mathcal{W}_2^2(\rho(t), \nu) + \frac{\lambda}{2}\mathcal{W}_2^2(\rho(t), \nu) + \mathcal{F}(\rho(t)) \leq \mathcal{F}(\nu), \tag{12}$$

holds in the sense of distributions, for all $\nu \in \mathcal{D}(\mathcal{F}) = \{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mathcal{F}(\mu) < +\infty\}$. The flow is said to start from ρ_0 if $\mathcal{W}_2(\rho(t), \rho_0) \to 0$ as $t \to 0$. Here, "in the sense of distributions" means that for all infinitely differentiable and compactly supported test functions, denoted $f \in C_c^\infty((0, \infty); \mathbb{R})$, such that $f \geq 0$, we have

$$-\frac{1}{2}\int_0^\infty \mathcal{W}_2^2(\rho(t), \nu) f'(t) dt \leq \int_0^\infty \left[\mathcal{F}(\nu) - \mathcal{F}(\rho(t)) - \frac{\lambda}{2}\mathcal{W}_2^2(\rho(t), \nu)\right] f(t) dt. \tag{13}$$

The connection between (12) and (13) can be seen by imagining the left hand side of (13) being integrated by parts.

One of the most attractive features of gradient flows are their convergence properties. For any λ-displacement convex functional $\mathcal{F}$ with $\lambda > 0$, the map $\rho \mapsto \mathcal{F}(\rho)$ has a unique minimum $\bar{\rho}$, and Theorem 11.2.1 of Ambrosio et al. (2005) states that there exists a unique gradient flow $t \mapsto \rho(t)$, which satisfies

$$\mathcal{W}_2(\rho(t), \bar{\rho}) \leq \mathcal{W}_2(\rho_0, \bar{\rho}) e^{-\lambda t} \quad \text{and} \quad \mathcal{F}(\rho(t)) - \mathcal{F}(\bar{\rho}) \leq [\mathcal{F}(\rho_0) - \mathcal{F}(\bar{\rho})]\, e^{-2\lambda t}, \tag{14}$$

or any $t \geq 0$. Convergence results also exist in the case where $\lambda = 0$, but do not yield the exponential convergence observed above.

This result can be applied to the relative entropy by making the following observations: when V is λ-strongly convex with $\lambda > 0$, it follows that $\rho \mapsto \mathcal{V}(\rho)$ is λ-displacement convex (Ambrosio et al., 2005, Proposition 9.3.2). In turn, this implies that $\rho \mapsto \mathcal{H}(\rho|\pi)$ is λ-displacement convex.

Recall that $\mathcal{H}(\rho|\pi) \geq 0$ for any ρ, and that $\rho \mapsto \mathcal{H}(\rho|\pi)$ is uniquely minimized at π due to the strict convexity of the function $x \mapsto x \log x$ for $x > 0$ appearing in $\mathcal{H}(\rho)$, and Jensen's inequality. The result in (14) can then be formulated as

$$\mathcal{W}_2(\rho(t), \pi) \leq \mathcal{W}_2(\rho_0, \pi)e^{-\lambda t} \quad \text{and} \quad \mathcal{H}(\rho(t)|\pi) \leq \mathcal{H}(\rho_0|\pi)e^{-2\lambda t}. \tag{15}$$

This is a more general statement of the exponential convergence to π of the solution to the Fokker–Planck equation mentioned in the introduction, and is as such one of the main motivations for studying Langevin Monte Carlo algorithms.

2.2. Time discretized flows

An important theoretical tool in establishing the existence of gradient flows is the minimizing movement scheme, often also called the JKO scheme. For a time-step $h > 0$, $k \geq 0$, and $\rho_h^0 = \rho_0$, consider the iterated minimization problems

$$\rho_h^{k+1} = \operatorname*{argmin}_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\rho) + \frac{1}{2h}\mathcal{W}_2^2(\rho, \rho_h^k). \tag{16}$$

Such minimizers exist and are unique under weak assumptions, such as lower semi-continuity and (strong) displacement convexity of $\mathcal{F}$ (see e.g. Ambrosio et al., 2009, Proposition 4.2). Both of these conditions hold for the relative entropy functional $\rho \mapsto \mathcal{H}(\rho|\pi)$ when V is convex: the first property holds in more generality and is well-known, whereas the second was proved in McCann (1997).

In the Euclidean setting, the sequence $(x_h^k)_{k\geq 0}$ is an implicit Euler discretization with step-size h of the gradient flow of $f : \mathbb{R}^d \to \mathbb{R}$ given in (10) with initial condition $x_h^0 = x_0$ if

$$x_h^{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^d} f(y) + \frac{1}{2h}\|x_h^k - y\|^2. \tag{17}$$

The map defined by the right hand side of (17) is often written $\operatorname{prox}_f^h(x_h^k)$ in the optimization literature, and is referred to as the proximal operator (see e.g. Parikh and Boyd, 2014).

By analogy, the JKO scheme (16) can be seen as an implicit Euler discretization of the flow in (11). It was this time discretization scheme applied to the functional $\rho \mapsto \mathcal{H}(\rho|\pi)$ that Jordan et al. (1998) employed, showing that the interpolation

$$\rho^h(t) = \rho_h^{k+1} \quad \text{for } t \in (kh, (k+1)h] \tag{18}$$

converges (in some formal sense) to the solution of the Fokker–Planck equation as $h \to 0$, in the case where V is smooth and satisfies certain growth conditions.

Building on results by Cépa (1998), Ambrosio et al. (2009) used a minimizing movement scheme to show existence and uniqueness of the gradient flow of the relative entropy functional given any convex V. In particular, they show that there exists a semigroup $(P_t)_{t\geq 0}$ and a unique Markov family $\{\mathbb{P}_x : x \in \mathbb{R}^d\}$ of probability measures on $(\mathbb{R}^d)^{[0,+\infty)}$ such that $\mathbb{E}_x f(X_t) = P_t f(x)$ for all bounded Borel functions f and all $x \in \mathbb{R}^d$. Moreover, it is shown that $\{\mathbb{P}_x : x \in \mathbb{R}^d\}$ is reversible with respect to π, and that π is uniquely invariant for $(P_t)_{t\geq 0}$. Restricting $(P_t)_{t\geq 0}$ to indicator functions of Borel sets $B \in \mathcal{B}(\mathbb{R}^d)$, we define $(R_t)_{t\geq 0}$ by $R_t\rho_0(B) = \int P_t 1_B d\rho_0$.

The process $\rho(t) = R_t\rho_0$ then uniquely satisfies (12) and the associated properties outlined in the previous section.

After originally being introduced as a theoretical tool, there has recently been interest in developing numerical implementations of the JKO scheme for solving PDEs. Several Eulerian grid-based approaches exist, see e.g. Burger et al. (2012); Carrillo et al. (2015a); Peyré (2015). By virtue of being grid-based, these have limited application in the high-dimensional sampling setting.

It will later be seen that Langevin-based Monte Carlo can be considered a Lagrangian scheme using a particle approximation to the gradient flow. Other Lagrangian approaches have been considered by e.g. Carrillo et al. (2015b); Benamou et al. (2016); Carrillo et al. (2017). These methods are typically adapted to accurately solving PDEs in two or three dimensions, and do not scale well with d. For instance, Carrillo et al. (2017) used the modified relative entropy functional

$$\mathcal{F}_\gamma(\rho) = \int \log(\phi_\gamma * \rho) d\rho + \int V d\rho + \log Z, \tag{19}$$

where $\varphi_\gamma = \gamma^{-d}\varphi(x/\gamma)$ denotes a mollifier, typically a Gaussian kernel with standard deviation $\gamma > 0$. This modification makes the functional well-behaved when evaluated at an empirical measure, with the first term providing a kernel-based estimate of the entropy of the underlying distribution. For small time steps h, their algorithm reduces to solving a system of ODEs to evolve the particles in the empirical measure. The application of this approach to the high-dimensional setting is limited by the kernel-based estimate of entropy.

3. Operator splitting

In the previous section, we alluded to the idea that Langevin Monte Carlo numerically approximates the time discretizations used to theoretically study Wasserstein gradient flows. Before making this connection clear, we first need to introduce the concept of operator splitting.

Consider the generic Cauchy problem

$$\frac{df}{dt} = \mathcal{A}(f), \quad f(0) = f_0, \tag{20}$$

with solution given by $f(t) = S_t f_0$ in semigroup notation. In many situations, the operator $\mathcal{A}$ can be split into the sum of two simpler operators: $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$. Let $f_j(t) = S_t^j f_0$ for $j = 1, 2$ denote the solutions to the problems

$$\frac{df_j}{dt} = \mathcal{A}_j(f_j), \quad f_j(0) = f_0. \tag{21}$$

One can hope to estimate the solution f of (20) via $f(t) \approx (S^2_{t/n} S^1_{t/n})^n f_0$ for some large positive integer n, which can be justified if a Lie–Trotter–Kato product formula of the form

$$f(t) = \lim_{n \to +\infty} (S^2_{t/n} S^1_{t/n})^n f_0 \tag{22}$$

holds. The book of Holden et al. (2010) contains a thorough overview of such results.

Returning to the Fokker–Planck equation (2), there is a natural split between the transport part of the equation:

$$\frac{d\rho}{dt} = \operatorname{div}(\rho \nabla V), \quad \rho(0) = \rho_0, \tag{23}$$

and the diffusion part:

$$\frac{d\rho}{dt} = \Delta\rho, \quad \rho(0) = \rho_0. \tag{24}$$

In his Ph.D. thesis, Stojković (2011) considers such a split for the Fokker–Planck equation with smooth drift satisfying a monotonicity property, but which is not necessarily a gradient. Bowles and Agueh (2015) also consider this split for the fractional Fokker–Planck equation, where the Laplacian in the diffusion equation (24) is substituted for a fractional Laplacian. In both of these works, operator splitting is introduced as a theoretical tool to establish the existence of solutions to generalized Fokker–Planck equations, but they do not consider numerical aspects nor the general case of convex V.

The splitting interpretation carries over to the Wasserstein gradient flow formulation, where the transport equation (23) can be interpreted as the gradient flow of the potential energy functional $\rho \mapsto \mathcal{V}(\rho)$, and the diffusion equation (24) can be interpreted as the gradient flow of the entropy functional $\rho \mapsto \mathcal{H}(\rho)$. We now take a brief closer look at these two gradient flows.

3.1. The transport equation

In addition to the formulation in (12), the gradient flow of $\rho \mapsto \mathcal{V}(\rho)$ can be characterized by the semigroup $(T_t)_{t\geq 0}$, induced by the differential inclusion

$$\frac{d}{dt}T_t(x) \in -\partial V(T_t(x)), \quad T_0(x) = x \quad \text{for all } x \text{ s.t. } V(x) < +\infty. \tag{25}$$

According to Theorem 11.2.3 of Ambrosio et al. (2005), there exists a unique gradient flow of $\rho \mapsto \mathcal{V}(\rho)$ and solution to (25). This gradient flow satisfies $\rho(t) = (T_t)_{\#}\rho_0$, where $(T_t)_{\#}$ denotes the push-forward map associated with T_t.

The corresponding JKO scheme performs minimizations of the form

$$\rho_h^{k+1} = \operatorname*{argmin}_{\rho\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{V}(\rho) + \frac{1}{2h}\mathcal{W}_2^2(\rho, \rho_h^k). \tag{26}$$

By the proof of Proposition 10.4.2 in Ambrosio et al. (2005), it is clear that these steps are well-defined. Moreover, the map $\mathcal{T}_h(x) = \operatorname{prox}_V^h(x)$ is such that $\rho_h^{k+1} = (\mathcal{T}_h)_{\#}\rho_h^k$. Since the proximal operator satisfies $y = \operatorname{prox}_V^h(x) \iff (x-y)/h \in \partial V(x)$ (see e.g. Parikh and Boyd, 2014), this can be seen as an implicit Euler step for the evolution of T_t given in (25).

3.2. The diffusion equation

The classical diffusion equation (24), also known as the heat equation, was first described as the gradient flow of the entropy functional $\rho \mapsto \mathcal{H}(\rho)$ on the set of densities in $\mathcal{P}_2(\mathbb{R}^d)$ by Jordan et al. (1998). Note that $\mathcal{H}(\rho)$ is the negative Gibbs–Boltzmann entropy of ρ. As pointed out in the aforementioned paper, the interpretation of the diffusion equation as the gradient flow of $\mathcal{H}$ therefore provides a natural interpretation of diffusion as the tendency of a system to maximize entropy.

Unlike the other gradient flows we have discussed, the flow of $\rho \mapsto \mathcal{H}(\rho)$ is known in closed form: it is well-known that the solution of the diffusion equation (24) is given by the density $\rho(t) = \phi_t * \rho_0$, where ϕ_t is the Gaussian kernel defined in (9).

4. Proximal Langevin Monte Carlo

We are now ready to describe connections between JKO discretized gradient flows, operator splitting, and Langevin-based Monte Carlo algorithms. For a time-step $h > 0$ and for $k \geq 0$, consider the iterative scheme

$$\rho_h^{k+1/2} = (\mathcal{T}_h)_{\#}\rho_h^k, \qquad \rho_h^{k+1} = \phi_h * \rho_h^{k+1/2}, \tag{27}$$

which can be seen as alternating between performing a JKO step for the gradient flow of $\rho \mapsto \mathcal{V}(\rho)$ and solving the exact gradient flow of $\rho \mapsto \mathcal{H}(\rho)$. Taking instead the particle perspective, let $X_h^0 \sim \rho_0$ and perform

$$X_h^{k+1/2} = \mathcal{T}_h(X_h^k) = \mathrm{prox}_V^h(X_h^k), \qquad X_h^{k+1} = X_h^{k+1/2} + \sqrt{2h}\eta^{k+1}, \tag{28}$$

where $(\eta^k)_{k\geq 1}$ is a sequence of independent $\mathcal{N}(0, \mathcal{I}_d)$ random variables. For each k, the laws of $X_h^{k+1/2}$ and X_h^{k+1} are equal to $\rho_h^{k+1/2}$ and ρ_h^{k+1} respectively. A generalization of this algorithm was proposed by Pereyra (2016) and studied further in Durmus et al. (2016).

Note that $\mathrm{prox}_V^h(x) = x - h\nabla M_V^h(x)$, where

$$M_V^h(x) = \inf_{y\in\mathbb{R}^d}\left\{V(y) + \frac{1}{2h}\|x-y\|^2\right\} \tag{29}$$

is the Moreau–Yosida regularization of V. Moreover, in the case where V is twice differentiable with positive definite Hessian $D^2V(x)$ for every $x \in \mathbb{R}^d$, it is known that $\mathrm{prox}_V^h(x) = x - h\nabla V(x) + o(h)$ as $h \to 0$ (see e.g. Parikh and Boyd, 2014, Section 3.3). Hence, for small h, the steps in (28) can be thought of as approximating the Unadjusted Langevin Algorithm.

4.1. Convergence analysis

We follow the approach of Clément and Maas (2011), which itself is an adaptation of the methods in Ambrosio et al. (2005, Chapter 4), to establish that the scheme in (27) satisfies a Lie–Trotter–Kato formula. We will also derive an upper bound on the 2-Wasserstein distance between the interpolation $\rho^h(t) = \rho_h^{k+1}$ for $t \in (kh, (k+1)h]$ and the gradient flow $\rho(t)$ of $\rho \mapsto \mathcal{H}(\rho|\pi)$. In turn, this allows us to bound the quantity of interest, $\mathcal{W}_2(\rho^h(t), \pi)$. Before stating the main results, we introduce some notation.

For any $n \geq 1$ and any $0 \leq k \leq n-1$, define the quantities

$$\delta_h^{k+1} = \mathcal{V}(\rho_h^{k+1}) - \mathcal{V}(\rho_h^{k+1/2}), \qquad \Delta_h^{k+1} = \sum_{j=1}^{k+1}\delta_h^j. \tag{30}$$

Note that δ_h^{k+1} can also be expressed

$$\delta_h^{k+1} = \mathbb{E}V(X+\eta) - \mathbb{E}V(X), \tag{31}$$

where $X \sim \rho_h^{k+1/2}$ and $\eta \sim \mathcal{N}(0, 2h\mathcal{I}_d)$ independently. By convexity of V and Jensen's inequality, it is clear that $\delta_h^{k+1} \geq \mathbb{E}V(\mathbb{E}(X+\eta|X)) - \mathbb{E}V(X) \geq 0$. The next results show that controlling these quantities is sufficient to establish convergence. We also remark that if one has access to independent runs of the algorithm given in (28), one can estimate δ_h^{k+1} by averaging $V(X_h^{k+1}) - V(X_h^k)$ across those runs.

Theorem 1 *Let $(\rho^{h_m}(t))_{m\geq 1}$ be a sequence of discrete solutions generated from ρ_0, such that $h_m\Delta^m_{h_m} \to 0$ and $h_m m \to T$ for some $T > 0$, as $m \to \infty$. Then, $\rho^{h_m}(t)$ converges uniformly on $[0, T]$ to $\rho(t)$, the gradient flow of $\rho \mapsto \mathcal{H}(\rho|\pi)$ started from ρ_0. Moreover, if $h > 0$ and $n \geq 1$ are such that $hn \leq T$, then for any $t \in [0, hn]$,*

$$\mathcal{W}_2(\rho^h(t), \rho(t)) \leq \sqrt{6h\left(\mathcal{H}(\rho_0|\pi) + \Delta^n_h\right)}. \tag{32}$$

The corollary below follows from combining (15) and (32) via the triangle inequality.

Corollary 2 *Suppose V is λ-strongly convex. Then, under the assumptions of Theorem 1, we have*

$$\mathcal{W}_2(\rho^h(t), \pi) \leq \sqrt{6h\left(\mathcal{H}(\rho_0|\pi) + \Delta^n_h\right)} + \mathcal{W}_2(\rho_0, \pi)e^{-\lambda t}, \tag{33}$$

for any $t \in [0, hn]$, where $h > 0$ and $n \geq 1$.

4.2. Explicit rates

It is clear that the rate at which $h\Delta^n_h \to 0$ as $h \to 0$ is crucial in determining the quality of the approximation $\rho^h(t)$. Under some assumptions on ρ_0 and V, we can obtain explicit bounds on Δ^n_h in terms of h, n, and d, as will be seen below.

Suppose $V = f + g$, where f is λ-strongly convex and has Lipschitz continuous gradient, and g is convex and Lipschitz. That is, assume that there exist $M(d)$ and $L(d)$ such that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leq M(d)\|x - y\| \tag{34}$$
$$|g(x) - g(y)| \leq L(d)\|x - y\|, \tag{35}$$

where the notation $M(d)$ and $L(d)$ reflects potential dependence of the Lipschitz constants on dimension. Under this assumption, we can bound δ^{k+1}_h as follows:

$$\mathbb{E}V(X + \eta) - \mathbb{E}V(X) = \mathbb{E}[f(X + \eta) - f(X)] + \mathbb{E}[g(X + \eta) - g(X)] \tag{36}$$
$$\leq \mathbb{E}\left[\nabla f(X)^\top \eta + \frac{M(d)}{2}\|\eta\|^2\right] + L(d)\mathbb{E}\|\eta\| \tag{37}$$
$$\leq M(d)hd + L(d)\sqrt{2hd}, \tag{38}$$

where (37) follows from the basic property that

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{M(d)}{2}\|x - y\|^2, \tag{39}$$

for all $x, y \in \mathbb{R}^d$, see for example Nesterov (2013). Then, $h\Delta^n_h \leq M(d)hd \cdot hn + L(d)\sqrt{2hd} \cdot hn$. Hence, for any $T > 0$ we could take $h_m = T/m$ and satisfy the conditions of Corollary 2.

Next, we can use these bounds to derive explicit rates for n and h that yield a desired approximation error. When selecting the initial distribution, it is not unreasonable to assume that one can choose ρ_0 such that $\mathcal{W}_2(\rho_0, \pi) = \mathcal{O}(\sqrt{d})$ and $\mathcal{H}(\rho_0|\pi) = \mathcal{O}(d)$. See Appendix B for justifications and an explicit example where these assumptions hold.

Now, if we want $\mathcal{W}_2(\rho^h(hn), \pi) = \mathcal{O}(\varepsilon)$ for a threshold $\varepsilon > 0$, we could require that both $h\mathcal{H}(\rho_0|\pi) + h\Delta^n_h = \mathcal{O}(\varepsilon^2)$ and $\mathcal{W}_2(\rho_0, \pi)e^{-\lambda hn} = \mathcal{O}(\varepsilon)$. Under the assumptions above, to ensure

$\mathcal{W}_2(\rho_0, \pi)e^{-\lambda hn} = \mathcal{O}(\varepsilon)$, it is sufficient to take $hn = \Omega(\log(d/\varepsilon^2))$. To get $h\mathcal{H}(\rho_0|\pi) = \mathcal{O}(\varepsilon^2)$, one can require that $h = \mathcal{O}(\varepsilon^2/d)$. Lastly, to get $h\Delta_h^n = \mathcal{O}(\varepsilon^2)$, one can in turn require that both $M(d)hd\log(\sqrt{d}/\varepsilon) = \mathcal{O}(\varepsilon^2)$ and $L(d)\sqrt{2hd}\log(\sqrt{d}/\varepsilon) = \mathcal{O}(\varepsilon^2)$. The former can be achieved if

$$n = \Omega\left(\frac{dM(d)\log(\sqrt{d}/\varepsilon)^2}{\varepsilon^2}\right) \quad \text{and} \quad h = \mathcal{O}\left(\frac{\varepsilon^2}{dM(d)\log(\sqrt{d}/\varepsilon)}\right), \tag{40}$$

while maintaining $hn = \Omega(\log(d/\varepsilon^2))$. Similarly, the latter can be achieved if

$$n = \Omega\left(\frac{dL(d)^2\log(\sqrt{d}/\varepsilon)^3}{\varepsilon^4}\right) \quad \text{and} \quad h = \mathcal{O}\left(\frac{\varepsilon^4}{dL(d)^2\log(\sqrt{d}/\varepsilon)^2}\right), \tag{41}$$

still keeping $hn = \Omega(\log(d/\varepsilon^2))$.

In the case where $g = 0$ (or equivalently $L(d) = 0$) and $M(d) = \mathcal{O}(1)$, we recover the assumptions on V that were made in e.g. Dalalyan (2017); Dalalyan and Karagulyan (2017). Using (40), we see that $n = \Omega(d\varepsilon^{-2}\log(d\varepsilon^{-2})^2)$ iterations with a step-size of $h = \log(d/\varepsilon^2)/n$ are sufficient to achieve a 2-Wasserstein error of $\mathcal{O}(\varepsilon)$. Up to log-terms, this is the same rate as those derived for ULA in the aforementioned papers.

In the case where $g(x) \propto \|x\|_1$ so that $L(d) = \mathcal{O}(\sqrt{d})$, we get that $n = \Omega(d^2/\varepsilon^4)$ iterations are sufficient (ignoring the log-terms). This improves upon the recent results of Grappin (2018), who showed that if additionally f is quadratic, then $n = \Omega(d^3/\varepsilon^4)$ iterations are sufficient to yield a 2-Wasserstein error of $\mathcal{O}(\varepsilon)$. Comparing to the remark accompanying Theorem 3 of Durmus et al. (2016), our results appears less sharp than the TV bounds they derive, in which n depends linearly on d (up to log-terms) whenever V is strongly convex. As can be seen in Appendix A, this likely stems from not optimally accounting for λ-displacement convexity in Lemma 6.

5. Conclusion

In this paper, we have developed novel connections between the fields of Wasserstein gradient flow, operator splitting, and Langevin Monte Carlo. We have demonstrated that the gradient flow perspective allows us to derive new convergence results about a proximal version of the Unadjusted Langevin Algorithm. Under certain assumptions on the potential V, we derive results that are on par with the contemporary literature on ULA. However, we point out that there is room for improvement in our current proofs. In particular, they could be improved by better accounting for the condition that V is λ-strongly convex, allowing us to obtain sharper bounds when that assumption is present. On the other hand, the proof of Theorem 1 generalizes to any convex V. Hence, to obtain control over the proximal ULA algorithm in such a case, one would only need to formulate conditions under which one can still derive a rate of convergence of the exact gradient flow to π, though one should no longer expect this convergence to be exponentially fast. Some recent progress in this direction based on Lojasiewicz inequalities was made by Blanchet and Bolte (2016).

We also hope that these connections can have implications on methodology. The many other splitting schemes discussed by Holden et al. (2010) and in the optimization literature can potentially lead to new sampling algorithms. The same holds for other numerical schemes, such as the alternative JKO algorithm developed by Legendre and Turinici (2017). For the Fokker–Planck equation, they show that their new scheme is second-order convergent, improving the original JKO scheme's

first-order convergence. Recently, Plazotta (2018) developed a variational formulation of the BDF2 scheme applicable to the estimation of gradient flows. It is also likely that the growing literature on Langevin Monte Carlo and its variations can lead to new time discretization schemes that are of both practical and theoretical interest to the gradient flow community.

Acknowledgments

I'm greatly indebted to Nicolas Chopin and Marco Cuturi for hosting my visit to ENSAE ParisTech and CREST, where the material in this paper was developed. I'd also like to thank Lénaïc Chizat, Arnak Dalalyan, Jeremy Heng, Pierre E. Jacob, Boris Muzellec and Gabriel Peyré for interesting conversations about optimal transport, gradient flows, and Monte Carlo sampling. This material is based upon research supported by the Chateaubriand Fellowship of the Office for Science & Technology of the Embassy of France in the United States.

References

Luigi Ambrosio, Nicola Gigli, and Guiseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser Verlag AG, Basel, second edition, 2005.

Luigi Ambrosio, Giuseppe Savaré, and Lorenzo Zambotti. Existence and stability for Fokker–Planck equations with log-concave reference measure. *Probability theory and related fields*, 145 (3):517–564, 2009.

Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the Monge–Ampère operator. *Numerische mathematik*, 134(3):611–636, 2016.

Adrien Blanchet and Jérôme Bolte. A family of functional inequalities: Lojasiewicz inequalities and displacement convex functions. *arXiv preprint arXiv:1612.02619*, 2016.

Malcolm Bowles and Martial Agueh. Weak solutions to a fractional Fokker–Planck equation via splitting and Wasserstein gradient flow. *Applied Mathematics Letters*, 42:30–35, 2015.

Martin Burger, Marzena Franek, and Carola-Bibiane Schönlieb. Regularized regression and density estimation based on optimal transport. *Applied Mathematics Research eXpress*, 2012(2):209–253, 2012.

José A Carrillo, Alina Chertock, and Yanghong Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17(1):233–258, 2015a.

Jose Antonio Carrillo, Yanghong Huang, Francesco Saverio Patacchini, and Gershon Wolansky. Numerical study of a particle method for gradient flows. *arXiv preprint arXiv:1512.03029*, 2015b.

José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *arXiv preprint arXiv:1709.09195*, 2017.

Emmanuel Cépa. Problème de Skorohod multivoque. *The Annals of Probability*, 26(2):500–532, 1998.

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

Philippe Clément and Jan Maas. A Trotter product formula for gradient flows in metric spaces. *Journal of Evolution Equations*, 11(2):405–427, 2011.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *arXiv preprint arXiv:1412.7392*, 2014.

Arnak S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. *arXiv preprint arXiv:1704.04752*, 2017.

Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *arXiv preprint arXiv:1605.01559*, 2016a.

Alain Durmus and Eric Moulines. Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. *arXiv preprint arXiv:1605.01559*, 2016b.

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

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *arXiv preprint arXiv:1612.07471*, 2016.

Edwin Grappin. *Model Averaging in Large Scale Learning*. PhD thesis, Université Paris-Saclay, 2018.

Helge Holden, Kenneth H Karlsen, Knut-Andreas Lie, and Nils Henrik Risebro. *Splitting Methods for Partial Differential Equations with Rough Solutions*. European Mathematical Society, 2010.

Søren F. Jarner and Ernst Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361, 2000.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Guillaume Legendre and Gabriel Turinici. Second-order in time schemes for gradient flows in Wasserstein and geodesic metric spaces. *Comptes Rendus Mathematique*, 355(3):345–353, 2017.

Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1): 153–179, 1997.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1 (3):127–239, 2014.

Grigorios A. Pavliotis. *Stochastic processes and applications*. Springer, 2014.

Marcelo Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26 (4):745–760, 2016.

Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

Natesh S. Pillai, Andrew M. Stuart, and Alexandre H. Thiéry. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *The Annals of Applied Probability*, 22(6):2320–2356, 2012.

Simon Plazotta. A BDF2-approach for the non-linear Fokker-Planck equation. *arXiv preprint arXiv:1801.09603*, 2018.

Gareth O. Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Filippo Santambrogio. {Euclidean, Metric, and Wasserstein} Gradient Flows: an overview. *arXiv preprint arXiv:1609.03890*, 2016.

Igor Stojković. *Geometric approach to evolution problems in metric spaces*. PhD thesis, Mathematical Institute, Faculty of Science, Leiden University, 2011.

Cédric Villani. *Optimal transport, old and new*. Springer-Verlag New York, 2008.

Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91: 14–19, 2014.

Appendix A. Proofs

Closely following Clément and Maas (2011) and Ambrosio et al. (2005), we start by proving a discrete version of the evolution variational inequality used to characterize gradient flows. Using interpolations of the discrete solutions, we use the discrete EVI to build a continuous approximation to the desired EVI. With this approximation, we derive a bound that quantifies the closeness of two discrete solutions. This bound is used to show that under appropriate assumptions on a sequence of discrete solutions, this sequence is Cauchy and therefore has a limit. Lastly, this limit is shown to be the desired gradient flow.

Lemma 3 (Discrete Evolution Variation Inequality) *For any $n \geq 1$, $h > 0$, $\nu \ll \mu_{Leb}$ and $k = 0, \dots, n-1$ we have*

$$\begin{aligned} &\frac{1}{2h}\left[\mathcal{W}_2^2(\rho_h^{k+1}, \nu) - \mathcal{W}_2^2(\rho_h^k, \nu)\right] + \frac{\lambda}{2}\mathcal{W}_2^2(\rho_h^{k+1/2}, \nu) \\ &\quad \leq \mathcal{H}(\nu|\pi) - \mathcal{H}(\rho_h^{k+1}|\pi) - \frac{1}{2h}\mathcal{W}_2^2(\rho_h^{k+1/2}, \rho_h^k) + \delta_h^{k+1}. \end{aligned} \tag{42}$$

Proof By Corollary 4.1.3 of Ambrosio et al. (2005) (see also their Lemma 9.2.7), for any $\rho_h^k \ll \mu_{\text{Leb}}$, we have

$$\begin{aligned}\frac{1}{2h}\left[\mathcal{W}_2^2(\rho_h^{k+1/2},\nu)-\mathcal{W}_2^2(\rho_h^k,\nu)\right]+\frac{\lambda}{2}\mathcal{W}_2^2(\rho_h^{k+1/2},\nu)\\ \leq \mathcal{V}(\nu)-\mathcal{V}(\rho_h^{k+1/2})-\frac{1}{2h}\mathcal{W}_2^2(\rho_h^{k+1/2},\rho_h^k).\end{aligned} \tag{43}$$

Recall that $t \mapsto \phi_t * \rho_h^{k+1/2}$ is the gradient flow of the 0-displacement convex entropy functional $\rho \mapsto \mathcal{H}(\rho)$. Therefore,

$$\frac{d}{dt}\frac{1}{2}\mathcal{W}_2^2(\phi_t * \rho_h^{k+1/2},\nu)+\mathcal{H}(\phi_t * \rho_h^{k+1/2}) \leq \mathcal{H}(\nu), \tag{44}$$

in the sense of distributions. By Remark 1.2 of Clément and Maas (2011), an equivalent condition is: for all $0 < a < b < \infty$,

$$\begin{aligned}\frac{1}{2}\left[\mathcal{W}_2^2(\phi_b * \rho_h^{k+1/2},\nu)-\mathcal{W}_2^2(\phi_a * \rho_h^{k+1/2},\nu)\right]\\ \leq (b-a)\mathcal{H}(\nu)-\int_a^b \mathcal{H}(\phi_t * \rho_h^{k+1/2})dt.\end{aligned} \tag{45}$$

Noting that $t \mapsto \mathcal{H}(\phi_t * \rho_h^{k+1/2})$ is non-increasing by Theorem 11.2.1 of Ambrosio et al. (2005) (see equation 11.2.4), we have that for all $0 < a < b < \infty$,

$$\begin{aligned}\frac{1}{2}\left[\mathcal{W}_2^2(\phi_b * \rho_h^{k+1/2},\nu)-\mathcal{W}_2^2(\phi_a * \rho_h^{k+1/2},\nu)\right]\\ \leq (b-a)\mathcal{H}(\nu)-(b-a)\mathcal{H}(\phi_b * \rho_h^{k+1/2}).\end{aligned} \tag{46}$$

Letting $a \to 0$, $b = h$, we have

$$\frac{1}{2h}\left[\mathcal{W}_2^2(\rho_h^{k+1},\nu)-\mathcal{W}_2^2(\rho_h^{k+1/2},\nu)\right] \leq \mathcal{H}(\nu)-\mathcal{H}(\rho_h^{k+1}). \tag{47}$$

Adding inequalities (43) and (47), as well as adding and subtracting $\mathcal{V}(\rho_h^{k+1})$ to the right hand side to make δ_h^{k+1} appear, yields the result. ■

It can be deduced from Lemma 3 that

$$\frac{1}{2h}\mathcal{W}_2^2(\rho_h^{k+1},\rho_h^k) \leq \mathcal{H}(\rho_h^k|\pi)-\mathcal{H}(\rho_h^{k+1}|\pi)-\frac{1+\lambda h}{2h}\mathcal{W}_2^2(\rho_h^{k+1/2},\rho_h^k)+\delta_h^{k+1}, \tag{48}$$

by taking $\nu = \rho_h^k$, so that

$$\sum_{k=0}^{n-1}\mathcal{W}_2^2(\rho_h^{k+1},\rho_h^k) \leq 2h\left[\mathcal{H}(\rho_h^0|\pi)-\mathcal{H}(\rho_h^n|\pi)+\Delta_h^n\right], \tag{49}$$

$$\leq 2h\left[\mathcal{H}(\rho_h^0|\pi)+\Delta_h^n\right]. \tag{50}$$

Similarly,

$$\mathcal{W}_2^2(\rho_h^{k+1/2},\rho_h^k) \leq \frac{2h}{1+\lambda h}\left[\mathcal{H}(\rho_h^k|\pi)-\mathcal{H}(\rho_h^{k+1}|\pi)+\delta_h^{k+1}\right], \tag{51}$$

so that

$$\sum_{k=0}^{n-1} \mathcal{W}_2^2(\rho_h^{k+1/2}, \rho_h^k) \leq \frac{2h}{1+\lambda h} \left[\mathcal{H}(\rho_h^0|\pi) + \Delta_h^n\right]. \tag{52}$$

Before proceeding, we introduce some more notation. Introduce the delayed interpolation $\rho_h(t) = \rho_h^k$ if $t \in [hk, (k+1)h)$, and note that $\rho^h(t)$ and $\rho_h(t)$ are left and right continuous respectively. Introduce also an interpolation of the half-steps, denoted by $\rho_{1/2}^h(t) = \rho_h^{k+1/2}$ if $t \in [hk, (k+1)h)$.

Define the piecewise affine function

$$\ell_h(t) = \frac{t - hk}{h} \qquad \text{if } t \in [hk, (k+1)h), \tag{53}$$

and in turn let

$$\mathcal{W}_h^2(t, \nu) = (1 - \ell_h(t))\mathcal{W}_2^2(\rho_h(t), \nu) + \ell_h(t)\mathcal{W}_2^2(\rho^h(t), \nu), \tag{54}$$

$$\mathcal{H}_h(t) = (1 - \ell_h(t))\mathcal{H}(\rho_h(t)|\pi) + \ell_h(t)\mathcal{H}(\rho^h(t)|\pi). \tag{55}$$

Let also

$$R_h(t) = 2(1 - \ell_h(t)) \left(\mathcal{H}(\rho_h^k|\pi) - \mathcal{H}(\rho_h^{k+1}|\pi) + \delta_h^{k+1}\right) + 2\ell_h(t)\delta_h^{k+1} \tag{56}$$

for $t \in [hk, (k+1)h)$. By (48) and $\delta_h^{k+1} \geq 0$, it is clear that $R_h(t) \geq 0$. The following result is an analog of Theorem 4.1.4 of Ambrosio et al. (2005).

Lemma 4 (Gradient flow approximation) *For any $n \geq 1$, $h > 0$, $\nu \ll \mu_{Leb}$ and $t \in [0, hn] \setminus \{kh : k = 0, \ldots, n\}$, we have*

$$\frac{d}{dt}\frac{1}{2}\mathcal{W}_h^2(t, \nu) + \frac{\lambda}{2}\mathcal{W}_2^2(\rho_{1/2}^h(t), \nu) + \mathcal{H}_h(t) - \mathcal{H}(\nu|\pi) \leq \frac{1}{2}R_h(t), \tag{57}$$

where d/dt denotes the pointwise derivative.

Proof If $t \in (hk, (k+1)h)$, then

$$\frac{d}{dt}\frac{1}{2}\mathcal{W}_h^2(t, \nu) = \frac{1}{2h}\left[\mathcal{W}_2^2(\rho_h^{k+1}, \nu) - \mathcal{W}_2^2(\rho_h^k, \nu)\right]. \tag{58}$$

By Lemma 3, this means

$$\begin{aligned}
&\frac{d}{dt}\frac{1}{2}\mathcal{W}_h^2(t, \nu) + \frac{\lambda}{2}\mathcal{W}_2^2(\rho_{1/2}^h(t), \nu) + \mathcal{H}_h(t) - \mathcal{H}(\nu|\pi) && (59)\\
&\quad = \frac{1}{2h}\left[\mathcal{W}_2^2(\rho_h^{k+1}, \nu) - \mathcal{W}_2^2(\rho_h^k, \nu)\right] + \frac{\lambda}{2}\mathcal{W}_2^2(\rho_{1/2}^h(t), \nu) + \mathcal{H}_h(t) - \mathcal{H}(\nu|\pi) && (60)\\
&\quad \leq \mathcal{H}_h(t) - \mathcal{H}(\rho_h^{k+1}|\pi) + \delta_h^{k+1} && (61)\\
&\quad = (1 - \ell_h(t))\mathcal{H}(\rho_h^k|\pi) + \ell_h(t)\mathcal{H}(\rho_h^{k+1}|\pi) - \mathcal{H}(\rho_h^{k+1}|\pi) + \delta_h^{k+1} && (62)\\
&\quad = (1 - \ell_h(t))\left(\mathcal{H}(\rho_h^k|\pi) - \mathcal{H}(\rho_h^{k+1}|\pi)\right) + \delta_h^{k+1} && (63)\\
&\quad = \frac{1}{2}R_h(t). && (64)
\end{aligned}$$

■

Lemma 5 *For any $n \geq 1$, $h > 0$ and $k = 0, \ldots, n-1$, we have the estimate*

$$0 \leq \int_0^{(k+1)h} R_h(t)dt \leq h\left(\mathcal{H}(\rho_h^0|\pi) + 2\Delta_h^n\right). \tag{65}$$

Proof The lower bound follows from $R_h(t) \geq 0$ for all $t \in [0, hn]$. Observe that

$$\int_{kh}^{(k+1)h} \ell_h(t)dt = \int_{kh}^{(k+1)h} (1 - \ell_h(t))dt = \frac{1}{2}h, \tag{66}$$

which in turn implies that

$$\int_0^{(k+1)h} R_h(t)dt = \sum_{j=0}^{k-1} \int_{jh}^{(j+1)h} R_h(t)dt \tag{67}$$

$$= \sum_{j=0}^{k-1} h\left(\mathcal{H}(\rho_h^j|\pi) - \mathcal{H}(\rho_h^{j+1}|\pi) + \delta_h^{j+1}\right) + \sum_{j=0}^{k-1} h\delta_h^{j+1} \tag{68}$$

$$\leq h\left(\mathcal{H}(\rho_h^0|\pi) - \mathcal{H}(\rho_h^{k+1}|\pi) + \Delta_h^{k+1}\right) + h\Delta_h^{k+1} \tag{69}$$

$$\leq h\left(\mathcal{H}(\rho_h^0|\pi) + 2\Delta_h^n\right). \tag{70}$$

■

Let $(\gamma_r^j)_{j=0}^m$ denote a trajectory corresponding to another time-step r, and define the quantities $\gamma_r(s), \gamma^r(s), \ell_r(s), \mathcal{H}_r(s)$ and $R_r(s)$ analogously to those defined in terms of h. Define

$$\mathcal{W}_{h,r}^2(t,s) = (1 - \ell_r(s))\mathcal{W}_h^2(t, \gamma_r(s)) + \ell_r(s)\mathcal{W}_h^2(t, \gamma^r(s)), \tag{71}$$

and observe that this function is continuous in t and s.

Lemma 6 *For any $n, m \geq 1$, $h, r > 0$ and $t \in [0, \min\{hn, rm\}]$,*

$$\mathcal{W}_{h,r}^2(t,t) \leq \mathcal{W}_2^2(\rho_h^0, \gamma_r^0) + \int_0^t R_h(t) + R_r(t)dt. \tag{72}$$

Proof Let $s \in [0, rm]$ and $t \in [0, hn] \setminus \{kh : k = 0, \ldots, n\}$. By Lemma 4,

$$\frac{\partial}{\partial t}\frac{1}{2}\mathcal{W}_{h,r}^2(t,s) + \mathcal{H}_h(t) - \mathcal{H}_r(s) \leq \frac{1}{2}R_h(t). \tag{73}$$

Similarly, for $s \in [0, rm] \setminus \{jr : j = 0, \ldots, m\}$ and $t \in [0, hn]$,

$$\frac{\partial}{\partial s}\frac{1}{2}\mathcal{W}_{r,h}^2(s,t) + \mathcal{H}_r(s) - \mathcal{H}_h(t) \leq \frac{1}{2}R_r(s). \tag{74}$$

Note the symmetry

$$\mathcal{W}_{h,r}^2(t,s) = \mathcal{W}_{r,h}^2(s,t), \tag{75}$$

so that for $s \in [0, rm] \setminus \{jr : j = 0, \ldots, m\}$ and $t \in [0, hn] \setminus \{kh : k = 0, \ldots, n\}$,

$$\frac{\partial}{\partial t}\mathcal{W}_{h,r}^2(t,s) + \frac{\partial}{\partial s}\mathcal{W}_{h,r}^2(t,s) \leq R_h(t) + R_r(s), \tag{76}$$

by adding the inequalities above. Setting $s = t$ and letting $t \in [0, \min\{hn, rm\}] \setminus (\{kh : k = 0, \dots, n\} \cup \{jr : j = 0, \dots, m\})$,

$$\frac{d}{dt}\mathcal{W}_{h,r}^2(t,t) \le R_h(t) + R_r(t). \tag{77}$$

Since $t \mapsto \mathcal{W}_{h,r}^2(t,t)$ is continuous and piecewise differentiable, the Fundamental Theorem of Calculus implies that

$$\mathcal{W}_{h,r}^2(t,t) \le \mathcal{W}_{h,r}^2(0,0) + \int_0^t R_h(t) + R_r(t)dt \tag{78}$$

$$= \mathcal{W}_2^2(\rho_h^0, \gamma_r^0) + \int_0^t R_h(t) + R_r(t)dt. \tag{79}$$

■

Lemma 7 *For any $n, m \ge 1$, $h, r > 0$ and $t \in [0, \min\{hn, rm\}]$,*

$$\begin{aligned}&\mathcal{W}_2^2(\rho^h(t), \gamma^r(t)) \\ &\quad \le 6\left[\mathcal{W}_2^2(\rho_h^0, \gamma_r^0) + h\left(\mathcal{H}(\rho_h^0|\pi) + \Delta_h^n\right) + r\left(\mathcal{H}(\gamma_r^0|\pi) + \Delta_r^m\right)\right].\end{aligned} \tag{80}$$

Proof Suppose j and k are such that $t \in [kh, (k+1)h) \cap [jr, (j+1)r)$. Then,

$$\begin{aligned}
&\mathcal{W}_2^2(\rho^h(t), \gamma^r(t)) = \mathcal{W}_2^2(\rho_h^{k+1}, \gamma_r^{j+1}) \\
&= (1-\ell_h(t))(1-\ell_r(t))\mathcal{W}_2^2(\rho_h^{k+1}, \gamma_r^{j+1}) \\
&\quad + (1-\ell_h(t))\ell_r(t)\mathcal{W}_2^2(\rho_h^{k+1}, \gamma_r^{j+1}) \\
&\quad + \ell_h(t)(1-\ell_r(t))\mathcal{W}_2^2(\rho_h^{k+1}, \gamma_r^{j+1}) \\
&\quad + \ell_h(t)\ell_r(t)\mathcal{W}_2^2(\rho_h^{k+1}, \gamma_r^{j+1}) \\
&\le 3(1-\ell_h(t))(1-\ell_r(t))\left[\mathcal{W}_2^2(\rho_h^{k+1}, \rho_h^k) + \mathcal{W}_2^2(\rho_h^k, \gamma_r^j) + \mathcal{W}_2^2(\gamma_r^{j+1}, \gamma_r^j)\right] \\
&\quad + 2(1-\ell_h(t))\ell_r(t)\left[\mathcal{W}_2^2(\rho_h^{k+1}, \rho_h^k) + \mathcal{W}_2^2(\rho_h^k, \gamma_r^{j+1})\right] \\
&\quad + 2\ell_h(t)(1-\ell_r(t))\left[\mathcal{W}_2^2(\gamma_r^{j+1}, \gamma_r^j) + \mathcal{W}_2^2(\rho_h^{k+1}, \gamma_r^j)\right] \\
&\quad + \ell_h(t)\ell_r(t)\mathcal{W}_2^2(\rho_h^{k+1}, \gamma_r^{j+1}) \\
&\le 3(1-\ell_h(t))(1-\ell_r(t))\left[\mathcal{W}_2^2(\rho_h^{k+1}, \rho_h^k) + \mathcal{W}_2^2(\rho_h^k, \gamma_r^j) + \mathcal{W}_2^2(\gamma_r^{j+1}, \gamma_r^j)\right] \\
&\quad + 3(1-\ell_h(t))\ell_r(t)\left[\mathcal{W}_2^2(\rho_h^{k+1}, \rho_h^k) + \mathcal{W}_2^2(\rho_h^k, \gamma_r^{j+1})\right] \\
&\quad + 3\ell_h(t)(1-\ell_r(t))\left[\mathcal{W}_2^2(\gamma_r^{j+1}, \gamma_r^j) + \mathcal{W}_2^2(\rho_h^{k+1}, \gamma_r^j)\right] \\
&\quad + 3\ell_h(t)\ell_r(t)\mathcal{W}_2^2(\rho_h^{k+1}, \gamma_r^{j+1}) \\
&= 3\mathcal{W}_{h,r}^2(t,t) + 3(1-\ell_h(t))\mathcal{W}_2^2(\rho_h^{k+1}, \rho_h^k) + 3(1-\ell_r(t))\mathcal{W}_2^2(\gamma_r^{j+1}, \gamma_r^j).
\end{aligned}$$

Now, by Lemmas 6 and 5,

$$\mathcal{W}_{h,r}^2(t,t) \leq \mathcal{W}_2^2(\rho_h^0, \gamma_r^0) + \int_0^t R_h(t) + R_r(t) dt \tag{81}$$

$$\leq \mathcal{W}_2^2(\rho_h^0, \gamma_r^0) + h\left(\mathcal{H}(\rho_h^0|\pi) + 2\Delta_h^n\right) + r\left(\mathcal{H}(\gamma_r^0|\pi) + 2\Delta_r^m\right). \tag{82}$$

Lastly, we know by Lemma 3 that

$$\mathcal{W}_2^2(\rho_h^{k+1}, \rho_h^k) \leq 2h\left(\mathcal{H}(\rho_h^0|\pi) + \Delta_h^n\right), \tag{83}$$

$$\mathcal{W}_2^2(\gamma_r^{j+1}, \gamma_r^j) \leq 2r\left(\mathcal{H}(\gamma_r^0|\pi) + \Delta_r^m\right). \tag{84}$$

In conclusion, and without optimizing the constant, we get

$$\begin{aligned} &\mathcal{W}_2^2(\rho^h(t), \gamma^r(t)) \\ &\quad \leq 6\left[\mathcal{W}_2^2(\rho_h^0, \gamma_r^0) + h\left(\mathcal{H}(\rho_h^0|\pi) + \Delta_h^n\right) + r\left(\mathcal{H}(\gamma_r^0|\pi) + \Delta_r^m\right)\right]. \end{aligned} \tag{85}$$

■

Before giving its proof, we restate the main theorem of the paper:

Theorem 1 *Let $(\rho^{h_m}(t))_{m\geq 1}$ be a sequence of discrete solutions generated from ρ_0, such that $h_m \Delta_{h_m}^m \to 0$ and $h_m m \to T$ for some $T > 0$, as $m \to \infty$. Then, $\rho^{h_m}(t)$ converges uniformly on $[0,T]$ to $\rho(t)$, the gradient flow of $\rho \mapsto \mathcal{H}(\rho|\pi)$ started from ρ_0, as $m \to \infty$. Moreover, if $h > 0$ and $n \geq 1$ are such that $hn \leq T$, then for any $t \in [0, hn]$,*

$$\mathcal{W}_2(\rho^h(t), \rho(t)) \leq \sqrt{6h\left(\mathcal{H}(\rho_0|\pi) + \Delta_h^n\right)}. \tag{86}$$

Proof Let the discrete solutions $\rho^{h_n}(t)$ and $\rho^{h_m}(t)$ be members of the sequence. From Lemma 7, we know that $\mathcal{W}_2^2(\rho^{h_m}(t), \rho^{h_n}(t)) \to 0$ as $m, n \to \infty$, for any $t \in [0,T]$. This implies that $(\rho^{h_m}(t))_{m\geq 1}$ is a Cauchy sequence. Since $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is complete, this means that the sequence converges to a function $\rho(t)$. Since the bound in Lemma 7 does not depend on t, this convergence is uniform on $[0,T]$.

Since the convergence is uniform and $\rho^{h_n}(t)$ is left continuous, then so is the limit $\rho(t)$. Moreover, since if $t \in [kh, (k+1)h)$ for some $k = 0, \ldots, n-1$,

$$\mathcal{W}_2^2(\rho^{h_n}(t), \rho_{h_n}(t)) \leq \mathcal{W}_2^2(\rho_{h_n}^{k+1}, \rho_{h_n}^k) \leq 2h_n\left(\mathcal{H}(\rho_0|\pi) + \Delta_{h_n}^n\right) \to 0 \quad \text{as } n \to \infty. \tag{87}$$

Hence, $\rho_{h_n}(t)$ converges to $\rho(t)$ in the same manner as $\rho^{h_n}(t)$, meaning that the limit $\rho(t)$ is right continuous also. Combining these facts, it is clear that $\rho(t)$ is continuous.

Similarly,

$$\mathcal{W}_2^2(\rho_{h_n}(t), \rho_{1/2}^{h_n}(t)) \leq \mathcal{W}_2^2(\rho_{h_n}^{k+1/2}, \rho_{h_n}^k) \leq 2h_n\left(\mathcal{H}(\rho_0|\pi) + \Delta_{h_n}^n\right) \to 0 \quad \text{as } n \to \infty, \tag{88}$$

by the bound in (52). This implies that $\rho_{1/2}^{h_n}(t)$ converges to $\rho(t)$ in the same manner as $\rho_{h_n}(t)$ and $\rho^{h_n}(t)$.

It remains to show that $\rho(t)$ is the gradient flow of $\rho \mapsto \mathcal{H}(\rho|\pi)$. Indeed, let $f \in C_c^\infty((0,\infty);\mathbb{R})$ be non-negative and $\nu \ll \mu_{\text{Leb}}$. Note that $\lim_{n\to\infty} \mathcal{W}_{h_n}^2(t,\nu) = \mathcal{W}_2^2(\rho(t),\nu)$ uniformly on $[0,T]$.

Since $t \mapsto \mathcal{W}_{h_n}^2(t,\nu)$ is continuous, so is the limit $\mathcal{W}_2^2(\rho(t),\nu)$. Thus, $t \mapsto f'(t)\mathcal{W}_2^2(\rho(t),\nu)$ is continuous, i.e. integrable, on [0,T]. The continuity of f' implies that there exists an $M > 0$ such that $|f'(t)| \leq M$. In combination with the aforementioned uniform convergence, we know that

$$\lim_{n\to\infty} \int_0^T f'(t)\mathcal{W}_{h_n}^2(t,\nu)dt = \int_0^T f'(t)\mathcal{W}_2^2(\rho(t),\nu)dt. \tag{89}$$

By the same reasoning, and the fact that $\lim_{n\to\infty} \mathcal{W}_2^2(\rho_{1/2}^{h_n}(t),\nu) = \mathcal{W}_2^2(\rho(t),\nu)$ uniformly on $[0,T]$, we have

$$\lim_{n\to\infty} \int_0^T f(t)\mathcal{W}_2^2(\rho_{1/2}^{h_n}(t),\nu)dt = \int_0^T f(t)\mathcal{W}_2^2(\rho(t),\nu)dt. \tag{90}$$

Now, since f and $\mathcal{H}(\cdot|\pi)$ are non-negative, so is the function $t \mapsto f(t)\mathcal{H}_h(t)$. Thus, by Fatou's lemma,

$$\liminf_{n\to\infty} \int_0^T f(t)\mathcal{H}_{h_n}(t)dt \geq \int_0^T \liminf_{n\to\infty} f(t)\mathcal{H}_{h_n}(t)dt. \tag{91}$$

By Lemma 2.8 of Clément and Maas (2011),

$$\int_0^T \liminf_{n\to\infty} f(t)\mathcal{H}_{h_n}(t)dt \geq \int_0^T f(t)\mathcal{H}(\rho(t)|\pi)dt. \tag{92}$$

So,

$$\begin{aligned}
&\int_0^T \left[-f'(t)\frac{1}{2}\mathcal{W}_2^2(\rho(t),\nu) + f(t)\frac{\lambda}{2}\mathcal{W}_2^2(\rho(t),\nu) + f(t)\mathcal{H}(\rho(t)|\pi)\right] dt && (93)\\
&\leq \liminf_{n\to\infty} \int_0^T \left[-f'(t)\frac{1}{2}\mathcal{W}_{h_n}^2(t,\nu) + f(t)\frac{\lambda}{2}\mathcal{W}_2^2(\rho_{1/2}^{h_n}(t),\nu) + f(t)\mathcal{H}_{h_n}(t)\right] dt && (94)\\
&= \liminf_{n\to\infty} \int_0^T \left[f(t)\frac{d}{dt}\frac{1}{2}\mathcal{W}_{h_n}^2(t,\nu) + f(t)\frac{\lambda}{2}\mathcal{W}_2^2(\rho_{1/2}^{h_n}(t),\nu) + f(t)\mathcal{H}_{h_n}(t)\right] dt && (95)\\
&\leq \liminf_{n\to\infty} \int_0^T f(t)\left[\frac{1}{2}R_{h_n}(t) + H(\nu|\pi)\right] dt && (96)\\
&= \int_0^T f(t)H(\nu|\pi)dt + \liminf_{n\to\infty} \int_0^T f(t)\frac{1}{2}R_{h_n}(t)dt && (97)\\
&\leq \int_0^T f(t)H(\nu|\pi)dt + \sup_{t\in[0,T]} f(t) \liminf_{n\to\infty} \int_0^T \frac{1}{2}R_{h_n}(t)dt && (98)\\
&\leq \int_0^T f(t)H(\nu|\pi)dt + \sup_{t\in[0,T]} f(t) \liminf_{n\to\infty} \left[\frac{1}{2}h_n(\mathcal{H}(\rho_0|\pi) + 2\Delta_{h_n}^n)\right] && (99)\\
&= \int_0^T f(t)H(\nu|\pi)dt, && (100)
\end{aligned}$$

where (94) follows from (91) and (92), (95) follows by integration by parts, (96) follows by Lemma 4, (98) follows by f being non-negative and continuous, and $R_{h_n}(t) \geq 0$, (99) follows by Lemma 5, and (100) follows by the assumption. This concludes the proof that $\rho(t)$ is indeed the gradient flow.

Now, fix $h > 0$ and $n \geq 1$ such that $hn \leq T$. Then, for any $m \geq 1$,

$$\mathcal{W}_2^2(\rho^h(t), \rho^{h_m}(t)) \leq 6\left[h\left(\mathcal{H}(\rho_0|\pi) + \Delta_h^n\right) + h_m\left(\mathcal{H}(\rho_0|\pi) + \Delta_{h_m}^m\right)\right], \tag{101}$$

for any $t \in [0, \min\{hn, h_m m\}]$ by Lemma 7. Taking $m \to \infty$ yields the conclusion.

■

Appendix B. Rates for $\mathcal{W}_2(\rho_0, \pi)$ and $\mathcal{H}(\rho_0|\pi)$

In this section, we provide some heuristic support for the claim that one can often assume that $\mathcal{H}(\rho_0|\pi) = \mathcal{O}(d)$ and $\mathcal{W}_2(\rho_0, \pi) = \mathcal{O}(\sqrt{d})$. These assumptions can also be shown to be hold for more general settings than those we consider below.

Let $\rho_0(x) = Z_0^{-1} e^{-V_0(x)}$, and note that

$$\begin{aligned}
\mathcal{W}_2^2(\rho_0, \pi) &= \inf_{\gamma \in \Gamma(\rho_0, \pi)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\gamma(x, y) \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|x - y\|^2 d\pi(x) d\rho_0(y) \\
&= \int_{\mathbb{R}^d} \|x - \bar{x}\|^2 d\pi(x) + \int_{\mathbb{R}^d} \|y - \bar{y}\|^2 d\rho_0(y) + \|\bar{x} - \bar{y}\|^2,
\end{aligned}$$

and where $\bar{x}$ and $\bar{y}$ are the means of π and ρ_0 respectively. The third term on the last line safely be assumed to be $\mathcal{O}(d)$. By Theorem 1 of Durmus and Moulines (2016a), the first term can be bounded by d/λ under the λ-strong convexity assumption. Under similar assumptions on ρ_0, or e.g. assuming that $V_0(x) = \sum_{i=1}^d V_0^i(x_i)$, one can also defend imposing a bound of $\mathcal{O}(d)$ for second term.

Secondly, one can easily support the assumption $\mathcal{H}(\rho_0|\pi) = \mathcal{O}(d)$ if both $V_0(x) = \sum_{i=1}^d V_0^i(x_i)$ and $V(x) = \sum_{i=1}^d V^i(x_i)$. A less restrictive condition is to assume that $0 \leq V(x) - V_0(x) \leq a\|x\|^2 + b$ for some $a \geq 0$ and $b \in \mathbb{R}$ not dependent on d. The first inequality is analogous to saying that ρ_0 has heavier tails than π, whereas the second inequality constrains exactly how much heavier these tails can be. Under this assumption, and using the proof of Lemma 3 of Dalalyan (2014), we can write

$$\begin{aligned}
\mathcal{H}(\rho_0|\pi) &= \int_{\mathbb{R}^d} \log\left(\frac{\rho_0}{\pi}\right) d\rho_0 \\
&= \int_{\mathbb{R}^d} \left[V(x) - V_0(x)\right] d\rho_0 + \log\left(\int_{\mathbb{R}^d} e^{V_0(x) - V(x)} d\rho_0\right) \\
&\leq \int_{\mathbb{R}^d} \left(a\|x\|^2 + b\right) d\rho_0,
\end{aligned}$$

by noting that $e^{V_0(x) - V(x)} \leq 1$ by the assumption. One can then proceed as in the last paragraph.

B.1. Gaussian initial distribution

Let $x^\star$ denote the minimum of V, and let $V_0(x) = \frac{\alpha}{2}\|x - \mu\|^2 + V(x^\star)$ with $\alpha < M(d)$, so that ρ_0 is a Gaussian distribution. We focus on bounding $\mathcal{H}(\rho_0|\pi)$, as bounding the Wasserstein distance

can be done as in the previous section. Then, using strong convexity, (34) and (35),

$$V(x) \leq V(x^\star) + L(d)\|x - x^\star\| + \nabla f(x^\star)^\top (x - x^\star) + \frac{M(d)}{2}\|x - x^\star\|^2,$$
$$V(x) \geq V(x^\star) - L(d)\|x - x^\star\| + \nabla f(x^\star)^\top (x - x^\star) + \frac{\lambda}{2}\|x - x^\star\|^2,$$

so that

$$\begin{aligned}
&\int_{\mathbb{R}^d} [V(x) - V_0(x)]\, d\rho_0 \\
&= \int_{\mathbb{R}^d} \left[\frac{M(d)}{2}\|x - x^\star\|^2 - \frac{\alpha}{2}\|x - \mu\|^2 + L(d)\|x - x^\star\| + \nabla f(x^\star)^\top (x - x^\star) \right] d\rho_0 \\
&\leq \frac{M(d)}{2}\|\mu - x^\star\|^2 + \frac{M(d)d}{2\alpha} - \frac{\alpha d}{2\alpha} + \nabla f(x^\star)^\top (\mu - x^\star) + L(d) \left(\int_{\mathbb{R}^d} \|x - x^\star\|^2 d\rho_0 \right)^{1/2} \\
&\leq \frac{M(d)}{2}\|\mu - x^\star\|^2 + \frac{(M(d) - \alpha)d}{2\alpha} + \nabla f(x^\star)^\top (\mu - x^\star) + L(d) \left(\|\mu - x^\star\|^2 + \frac{d}{\alpha} \right)^{1/2},
\end{aligned}$$

and

$$\begin{aligned}
\log \int_{\mathbb{R}^d} e^{V_0(x) - V(x)} d\rho_0 &\leq \log \left(\frac{1}{Z_{1/\alpha}} \int_{\mathbb{R}^d} e^{-\frac{\lambda}{2}\|x - x^\star\|^2 + L(d)\|x - x^\star\| - \nabla f(x^\star)^\top (x - x^\star)} dx \right) \\
&\leq \log \left(\frac{1}{Z_{1/\alpha}} \int_{\mathbb{R}^d} e^{-\frac{\lambda}{2}\|x - x^\star\|^2 + (L(d) + \|\nabla f(x^\star)\|)\|x - x^\star\|} dx \right) \\
&= \log \left(\frac{1}{Z_{1/\alpha}} \int_{\mathbb{R}^d} e^{-\frac{\lambda}{2}\|x - x^\star\|^2 + c\|x - x^\star\|} dx \right),
\end{aligned}$$

where $c = L(d) + \|\nabla f(x^\star)\|$ and $Z_{1/\alpha} = \int_{\mathbb{R}^d} e^{-\frac{\alpha}{2}\|x - \mu\|^2} dx$. Furthermore,

$$\begin{aligned}
\log \left(\frac{1}{Z_{1/\alpha}} \int_{\mathbb{R}^d} e^{-\frac{\lambda}{2}\|x\|^2 + c\|x\|} dx \right) &\leq \log \left(\frac{1}{Z_{1/\alpha}} \int_{\mathbb{R}^d} e^{-\frac{\lambda}{4}\|x\|^2 + \frac{c^2}{\lambda}} dx \right) \\
&= \log \left(\frac{Z_{2/\alpha}}{Z_{1/\alpha}} e^{\frac{c^2}{\lambda}} \right) \\
&= \frac{d}{2}\log(2) + \frac{(L(d) + \|\nabla f(x^\star)\|)^2}{\lambda} \\
&\leq \frac{d}{2}\log(2) + \frac{(L(d) + M(d)\|x^\star - x^f\|)^2}{\lambda},
\end{aligned}$$

where x^f is the minimum of f. Hence,

$$\begin{aligned}
\mathcal{H}(\rho_0|\pi) \leq \frac{M(d)}{2}\|\mu - x^\star\|^2 + \frac{(M(d) - \alpha)d}{2\alpha} + \|x^\star - x^f\|\|\mu - x^\star\| + L(d) \left(\|\mu - x^\star\|^2 + \frac{d}{\alpha} \right)^{1/2} \\
+ \frac{d}{2}\log(2) + \frac{(L(d) + M(d)\|x^\star - x^f\|)^2}{\lambda}.
\end{aligned}$$

Take $\alpha = \lambda$ and μ such that $\|\mu - x^\star\|^2 = \mathcal{O}(d)$, and make the safe assumption that $\|x^\star - x^f\|^2 = \mathcal{O}(d)$. If $M(d) = \mathcal{O}(1)$ and $L(d) = \sqrt{d}$ like in Section 4.2, we get $\mathcal{H}(\rho_0|\pi) = \mathcal{O}(d)$.