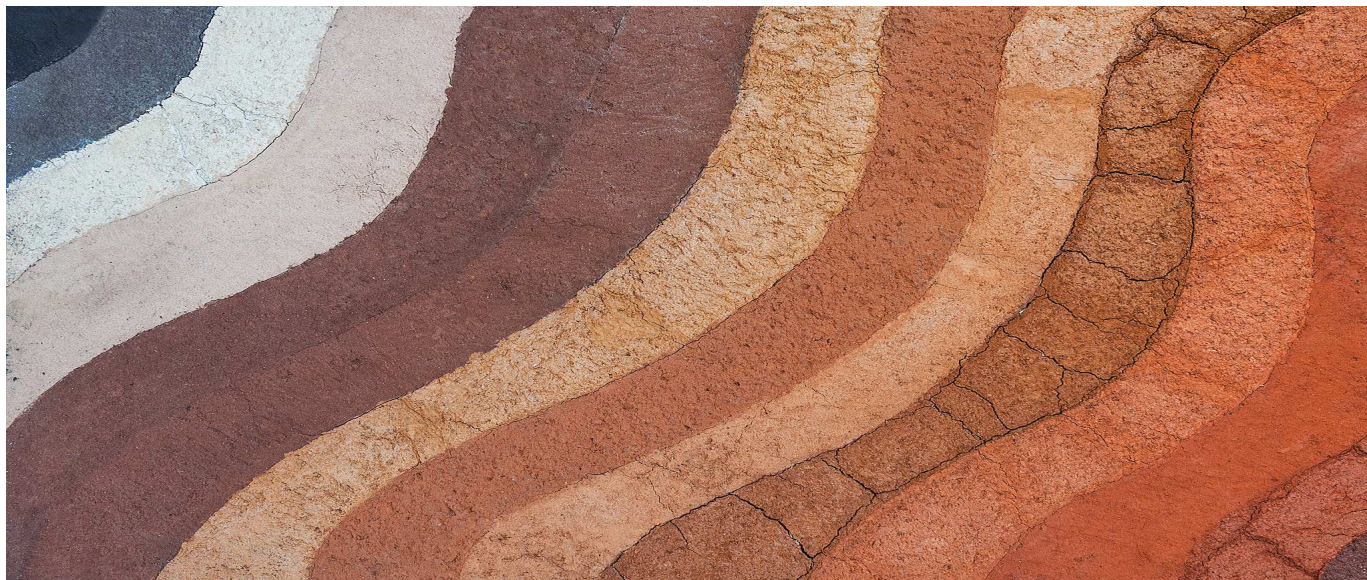# From Optimization to Sampling Through Gradient Flows

N. García Trillos, B. Hosseini, and D. Sanz-Alonso

Optimization and sampling algorithms play a central role in science and engineering as they enable finding optimal predictions, policies, and recommendations, as well as expected and equilibrium states of complex systems. The notion of "optimality" is formalized by the choice of an objective function, while the notion of an "expected" state is specified by a probabilistic model for the distribution of states. Optimizing rugged objective functions and sampling multimodal distributions is computationally challenging, especially in high-dimensional problems. For this reason, many optimization and sampling methods have

N. García Trillos is an assistant professor in the Department of Statistics at the University of Wisconsin-Madison. His email address is garciatrillo@wisc.edu.

B. Hosseini is an assistant professor in the Department of Applied Mathematics at the University of Washington. His email address is bamdadh@uw.edu.

D. Sanz-Alonso is an assistant professor in the Department of Statistics at the University of Chicago. His email address is sanzalonso@uchicago.edu.

been developed by researchers working in disparate fields such as Bayesian statistics, molecular dynamics, genetics, quantum chemistry, machine learning, weather forecasting, econometrics, and medical imaging.

State-of-the-art algorithms for optimization and sampling often rely on *ad-hoc* heuristics and empirical tuning, but some unifying principles have emerged that greatly facilitate the understanding of these methods and the communication of algorithmic innovations across scientific communities. This article is concerned with one such principle: the use of gradient flows, and discretizations thereof, to design and analyze optimization and sampling algorithms. The interplay between optimization, sampling, and gradient flows is an active research area and a thorough review of the extant literature is beyond the scope of this article.[1] Our goal is to provide an accessible and lively introduction to some core ideas, emphasizing that gradient flows uncover the conceptual unity behind several existing algorithms and give a rich mathematical framework for their rigorous analysis.

---

[1]*AMS* Notices *limits to 20 the references per article; we refer to* [GTSA20, CLGL$^+$20, GIHLS20] *for further pointers to the literature.*

We present motivating applications in section 1. Section 2 is focused on the gradient descent approach to optimization and introduces fundamental ideas such as preconditioning, convergence analysis, and time discretization of gradient flows. Sampling is discussed in section 3 in the context of Langevin dynamics viewed as a gradient flow of the Kullback–Leibler (KL) divergence with respect to (w.r.t) the Wasserstein geometry. Some modern applications of gradient flows for sampling are discussed in section 4, followed by concluding remarks in section 5.

1. Motivating Applications

We outline two applications in Bayesian statistics and molecular dynamics that illustrate some important challenges in optimization and sampling.

1.1. Bayesian statistics.
In Bayesian statistics [GCSR95], an initial belief about an unknown parameter is updated as data becomes available. Let θ denote the unknown parameter of interest belonging to the parameter space Θ and let $\pi_0(\theta)$ denote the *prior* distribution reflecting our initial belief. Furthermore, let y be the observed data also belonging to an appropriate space $\mathcal{Y}$. Then Bayes's rule identifies the distribution of θ conditioned on the data y:

$$\pi(\theta|y) \propto \pi(y|\theta)\pi_0(\theta), \qquad (1.1)$$

where $\propto$ indicates that the right-hand side should be normalized to define a probability distribution. Here $\pi(y|\theta)$ is called the *likelihood* function and $\pi(\theta|y)$ is called the *posterior* distribution. Bayesian inference on θ is based on the posterior, which blends the information in the prior and the data.

The choice of prior and likelihood is a modeling task which, perhaps surprisingly, is often not the most challenging aspect of Bayesian inference. The main challenge is to extract information from the posterior since (i) it typically does not belong to a standard family of distributions, unless in the restrictive case of conjugate models [GCSR95]; (ii) the parameter θ can be high dimensional; and (iii) the normalizing constant $\int_\Theta \pi(y|\theta)\pi_0(\theta)\,d\theta$ in (1.1) (known as the *marginal likelihood*) is rarely available and it can be expensive to compute. These practical hurdles inform the design of optimization and sampling algorithms to find posterior statistics.

Of particular importance is the posterior mode or *maximum a posteriori (MAP)* estimator

$$\theta_{\text{MAP}} := \arg\max_\theta \pi(\theta|y).$$

Many optimization algorithms for MAP estimation start from an initial guess θ_0 and produce iterates $\{\theta_n\}_{n=1}^N$ by discretizing a gradient flow with the property that $\theta_n \approx \theta_{\text{MAP}}$ for large n. Such gradient flows in parameter space will be discussed in section 2.

Computing MAP estimators is closely related to classic regularization techniques such as penalized least squares and Tikhonov regularization [SAST18]. In order to fully leverage the Bayesian framework, it is often desirable to consider other posterior statistics such as mean, variance, credible intervals, and task-specific functional moments, which can be written in the form

$$\mathbb{E}^{\pi(\cdot|y)}[\phi(\theta)] := \int \phi(\theta)\pi(\theta|y)\,d\theta,$$

where ϕ is a suitable test function. Since θ is often high dimensional, the standard approach to compute these expectations is to use Monte Carlo [Liu01]: obtain N samples $\{\theta_n\}_{n=1}^N$ from the posterior $\pi(\theta|y)$, and then approximate

$$\mathbb{E}^{\pi(\cdot|y)}[\phi(\theta)] \approx \frac{1}{N}\sum_n \phi(\theta_n).$$

While Monte Carlo integration is in principle scalable to high dimensions, the task of generating posterior samples is still highly nontrivial. To that end, one may consider sampling $\theta_n \sim \rho_n$, where the sequence $\{\rho_n\}_{n=1}^N$ arises from discretizing a gradient flow with the property that $\rho_n \approx \pi(\cdot|y)$ for large n. Such gradient flows in the space of probability distributions will be discussed in sections 3 and 4.

To relate the discussion above to subsequent developments, we note that dropping the data y from the notation, the posterior density can be written as

$$\pi(\theta) = \frac{1}{Z}\exp(-V(\theta)), \qquad (1.2)$$

where $Z = \int_\Theta \pi(y|\theta)\pi_0(\theta)\,d\theta$ is the marginal likelihood and $V : \Theta \times \mathcal{Y} \to \mathbb{R}$ is the negative logarithm of the posterior density.

1.2. Molecular dynamics.
Another important source of challenging optimization and sampling problems is statistical mechanics, and in particular the simulation of molecular dynamics (see chapter 9 in [Liu01]). According to Boltzmann and Gibbs, the positions q and momenta p of the atoms in a molecular system of constant size, occupying a constant volume, and in contact with a heat bath (at constant temperature), are distributed according to

$$\pi(q,p) = \frac{1}{Z}\exp\left(-\beta(U(q) + K(p))\right), \qquad (1.3)$$

where Z is a normalizing constant known as the partition function, β represents the inverse temperature, U is a potential energy describing the interaction of the particles in the system, and K represents the kinetic energy of the system. Letting $\theta := (q,p)$, we can write the Boltzmann–Gibbs distribution (1.3) in the form (1.2), with

$$V(\theta) = -\beta(U(q) + K(p)). \qquad (1.4)$$

As in Bayesian statistics, it is important to determine the most likely configuration of particles (i.e., the mode of π), along with expectations of certain test functions w.r.t.

the Boltzmann distribution. These two tasks motivate the need for optimization and sampling algorithms that acknowledge that the potential U is often a rough function with many local minima, that the dimension of q and p is large, and that finding the normalizing constant Z is challenging.

2. Optimization

In this section we discuss gradient flows for solution of the unconstrained minimization problem

$$\text{minimize } V(\theta) \quad \text{s.t. } \theta \in \Theta, \tag{2.1}$$

where $V(\theta)$ is a given objective function. Henceforth we take $\Theta := \mathbb{R}^d$ unless otherwise noted. As guiding examples, consider computing the mode of a posterior or Boltzmann distribution by minimizing V given by (1.2) or (1.4). The methods described in this section are applicable beyond the specific problem of finding the modes, however, this interpretation will be of particular interest in relating the material in this section to our discussion of sampling in section 3.

2.1. Gradient systems.
One of the most standard approaches to solve (2.1) is *gradient descent*, an optimization scheme that is based on the discretization of the gradient system

$$\dot{\theta}_t = -\nabla V(\theta_t), \quad t > 0, \tag{2.2}$$

with user-defined initial value θ_0; throughout this article $\nabla V(\theta_t)$ will denote the gradient of the function V at the point θ_t, which will be tacitly assumed to exist wherever needed.

While equation (2.2) is perhaps the most popular formulation of the continuous-time gradient descent dynamics, the equivalent integral form below reveals more transparently some of its properties:

$$V(\theta_t) = V(\theta_s) - \frac{1}{2}\int_s^t |\nabla V(\theta_r)|^2 dr - \frac{1}{2}\int_s^t |\dot{\theta}_r|^2 dr, \tag{2.3}$$

for all $t \geq s > 0$. Indeed, notice that from (2.3) it is apparent that $V(\theta_t) \leq V(\theta_s)$ for $s \leq t$, i.e., the value of the function V decreases in time, and in all but a few trivial situations the decrease is strict. Another advantage of the reformulation (2.3) (or its inequality form (2.5) below) is that it can be adapted to more general settings with less mathematical structure than the one needed to make sense of (2.2). In particular, (2.3) can be used to motivate a notion of gradient flow in arbitrary metric spaces; see [AGS08] for an in-depth discussion of this topic.

Proposition 2.1. *Suppose V is a C^1 function. Then* (2.2) *and* (2.3) *are equivalent and they both imply*

$$|\nabla V(\theta_t)|^2 = |\dot{\theta}_t|^2. \tag{2.4}$$

Proof. By Cauchy–Schwarz and Young's inequalities, for any $t > 0$ it holds that

$$-\langle \nabla V(\theta_t), \dot{\theta}_t \rangle \leq |\nabla V(\theta_t)||\dot{\theta}_t|$$
$$\leq \frac{1}{2}|\nabla V(\theta_t)|^2 + \frac{1}{2}|\dot{\theta}_t|^2,$$

and both inequalities are equalities iff $-\nabla V(\theta_t) = \dot{\theta}_t$. Therefore,

$$V(\theta_t) = V(\theta_s) + \int_s^t \langle \nabla V(\theta_r), \dot{\theta}_r \rangle dr$$
$$\geq V(\theta_s) - \frac{1}{2}\int_s^t |\nabla V(\theta_r)|^2 dr - \frac{1}{2}\int_s^t |\dot{\theta}_r|^2 dr,$$

and equality holds iff $-\nabla V(\theta_t) = \dot{\theta}_t$ for all $t > 0$. The identity $|\nabla V(\theta_t)|^2 = |\dot{\theta}_t|^2$ follows directly from (2.2). □

Notice that the relationship $\dot{\theta}_t = -\nabla V(\theta_t)$ is only required in proving the *energy dissipation inequality*

$$V(\theta_t) \leq V(\theta_s) - \frac{1}{2}\int_s^t |\nabla V(\theta_r)|^2 dr - \frac{1}{2}\int_s^t |\dot{\theta}_r|^2 dr, \tag{2.5}$$

since the reverse inequality is a consequence of Cauchy–Schwarz. Notice further that (2.2) implies (2.4), but in general the converse statement is not true. For example, the flow $\dot{\theta}_t = \nabla V(\theta_t)$ satisfies (2.4), but in general does not satisfy (2.2). Likewise, (2.2) implies $\frac{d}{dt}V(\theta_t) = -|\nabla V(\theta_t)|^2$ (which follows directly from the chain rule), but not conversely. Indeed, in $\mathbb{R}^2$ we may take A to be any orthogonal matrix and consider $\dot{\theta}_t = -A^2\nabla V(\theta_t)$ so that $\frac{d}{dt}V(\theta_t) = -|\nabla V(\theta_t)|^2$ but (2.2) is not, in general, satisfied. This digression illustrates that equation (2.3) captures in one single identity of *scalar* quantities the *vectorial* identity (2.2), even if it is not as intuitive as other scalar relations.

2.2. A note on convergence.
Despite the fact that gradient descent satisfies the energy dissipation property, it is in general not true that as time goes to infinity the dynamics (2.2) converge to a global minimizer of (2.1). This could happen for different reasons. First, the problem (2.1) may not have a minimizer (e.g., take $V(\theta) = e^{-\theta}$ for $\theta \in \mathbb{R}$). Second, ∇V may have critical points associated with saddle points or local optima of V as we illustrate in the next example.

Example 2.2. Consider the double well potential

$$V(\theta) = \frac{3}{8}\theta^4 - \frac{3}{4}\theta^2, \quad \theta \in \mathbb{R}, \tag{2.6}$$

so that $\nabla V(\theta) = \frac{3}{2}\theta(\theta^2 - 1)$. Notice that $\theta = 0$ is an unstable equilibrium of (2.2), which corresponds to a local maximum of V. For each local minima $\theta = \pm 1$ of V there is an associated subregion in the space of parameters (known as
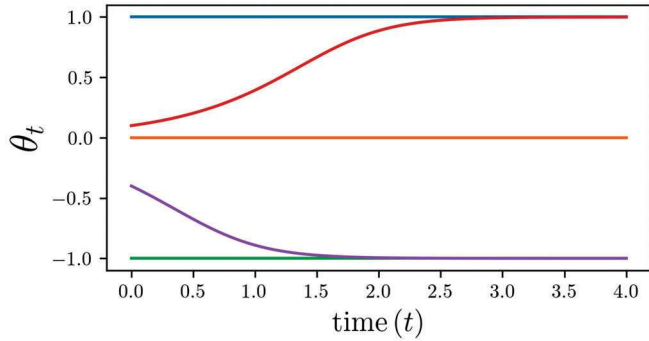
Figure 1. Five trajectories of the one-dimensional gradient system (2.2) with double well potential (2.6). The objective V has critical points at $\theta = 0$ and $\theta = \pm 1$. The intervals $(0, \infty)$ and $(-\infty, 0)$ are basins of attraction for $\theta = 1$ and $\theta = -1$, respectively.

a basin of attraction) such that any initial condition θ_0 chosen in this subregion leads the gradient dynamics toward its corresponding local minimizer, see Figure 1.

Suitable assumptions on V prevent the existence of local minimizers that are not global and also imply rates of convergence. One such assumption is the Polyak–Lojasiewic (PL) condition [KNS16]:

$$\alpha(V(\theta) - V^*) \leq \frac{1}{2}|\nabla V(\theta)|^2, \quad \forall \theta \in \Theta, \qquad (2.7)$$

where $V^* = \inf_{\theta \in \Theta} V(\theta)$ and $\alpha > 0$ is a constant. Note that the PL condition readily implies that any stationary point θ^* of V is a global minimizer, since

$$\alpha(V(\theta^*) - V^*) \leq \frac{1}{2}|\nabla V(\theta^*)|^2 = 0.$$

Under the PL condition we can easily obtain a convergence rate for continuous-time gradient descent.

Proposition 2.3. *Suppose V satisfies* (2.7). *Then,*

$$V(\theta_t) - V^* \leq (V(\theta_0) - V^*)\exp(-2\alpha t), \quad \forall t \geq 0. \qquad (2.8)$$

Proof. Using condition (2.7) in equation (2.5) and recalling (2.4) we conclude that

$$V(\theta_t) - V^* \leq (V(\theta_0) - V^*) - 2\alpha \int_0^t (V(\theta_r) - V^*)dr.$$

The result follows by Gronwall's inequality. $\qquad\square$

One can verify the PL condition under various assumptions on the function V [KNS16]. Here we present, as an important example, the case of α-strong convexity. For $\alpha > 0$, one says that V is α-*strongly convex* if for any $\theta, \theta' \in \Theta$ it holds that

$$V(t\theta + (1-t)\theta') \leq$$
$$tV(\theta) + (1-t)V(\theta') - \frac{\alpha}{2}t(1-t)|\theta - \theta'|^2,$$
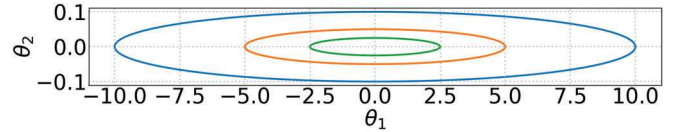

Figure 2. Level curves of a two-dimensional potential of the form $V(\theta) = \alpha_1\theta_1^2 + \alpha_2\theta_2^2$ with $0 < \alpha_1 \ll \alpha_2$. The anisotropy of this potential causes the gradient system (2.2) to converge slowly.

for all $t \in [0, 1]$. This condition can be shown to be equivalent to

$$V(\theta') \geq V(\theta) + \langle \nabla V(\theta), \theta' - \theta \rangle + \frac{\alpha}{2}|\theta' - \theta|^2, \quad \forall \theta, \theta' \in \Theta,$$

from which we can see, after minimizing both sides w.r.t. θ', that

$$V^* \geq V(\theta) - \frac{1}{2\alpha}|\nabla V(\theta)|^2, \qquad (2.9)$$

which is equivalent to (2.7). From this we conclude that α-strong convexity implies the PL condition with the same constant α.

Note that strong convexity is a stronger condition than the PL condition. For example, the function $V(\theta) = \frac{1}{2}\theta_1^2$ (where $\theta = (\theta_1, \theta_2)$) satisfies the PL condition with $\alpha = 1$, but it is not strongly convex.

2.3. Choice of the metric. Let us consider a function of the form $V(\theta) = \alpha_1\theta_1^2 + \alpha_2\theta_2^2$, where $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$. Suppose that $0 < \alpha_1 \ll \alpha_2$ and that α_1 is very close to zero, as in Figure 2. We can now apply Proposition 2.3 with $\alpha = \alpha_1$, but since we assumed α_1 is small we see that the right-hand side of (2.8) decreases very slowly. This suggests that gradient descent may take a long time to reach V's global minimum when initialized arbitrarily.

The poor behavior of gradient descent described above arises whenever there are regions of points away from the minimizer at which the gradient of V is very small. One approach to remedy this issue is to introduce a more general version of gradient descent that accelerates the dynamics in those regions where the gradient of V is small. This is the goal of *preconditioning*. Let $H : \Theta \to \mathcal{S}_{++}^d$ be a continuous field of positive definite matrices, i.e., a function that assigns to every point $\theta \in \Theta$ a $d \times d$ positive definite matrix $H(\theta)$. The *preconditioned gradient descent* dynamics induced by H is defined as:

$$\dot{\theta}_t = -H(\theta_t)^{-1}\nabla V(\theta_t), \quad t > 0. \qquad (2.10)$$

Observe that (2.10) coincides with the original gradient descent dynamics (2.2) when H is constant and equal to the $d \times d$ identity matrix. On the other hand, when V is convex and twice differentiable, choosing $H(\theta) = \nabla^2 V(\theta)$, the Hessian of V, results in a continuous-time analog of *Newton's algorithm*. In the example in Figure 2, we can directly compute $\nabla^2 V = \begin{pmatrix} 2\alpha_1 & 0 \\ 0 & 2\alpha_2 \end{pmatrix}$, i.e., the Hessian is a

fixed matrix since the potential V is quadratic. Substituting this choice of H in (2.10) for that example gives the dynamics $\dot{\theta}_t = -\theta_t$, a scheme that achieves a much faster convergence rate.

The reader may wonder if we could have in fact chosen $H = \frac{1}{r}D^2V$ for large constant r in order to induce a system that converges to equilibrium at a faster rate. However, as implied by our discussion in section 2.4 and, specifically, Remark 2.6 there is no benefit in doing so, as the cost of discretizing becomes correspondingly higher; rescaling the Hessian may be simply interpreted as a change of units. In general, there is a natural tension between accelerating continuous-time dynamics by changing the metric of the space, and producing accurate time discretizations for the resulting flows; see [GTSA20] for a related discussion in the context of sampling algorithms. In a similar vein, we notice that the superior convergence rate and affine invariance of Newton's algorithm comes at the price of utilizing the Hessian matrix ∇^2V, which in many applications can be prohibitively costly to compute or store. To this end, constructing matrix fields $H(\theta)$ that are good proxies for the Hessian and that can be computed efficiently is the goal of preconditioning. Perhaps the most well-known family of such algorithms is the family of *Quasi-Newton* algorithms [NW99] and in particular the *Broyden–Fletcher–Goldfarb–Shanno* (BFGS) algorithm, which approximates the Hessian using gradients calculated at previous iterates.

2.3.1. *Geometric interpretation.*
As a step toward introducing the material in section 3, here we give a geometric interpretation of equation (2.10). Specifically, we will show that (2.10) can still be understood as a gradient descent equation but w.r.t. a different metric on the parameter space Θ. For this purpose it is convenient to recall that a *Riemannian manifold* $(\mathcal{M}, g)$ is a manifold $\mathcal{M}$ endowed with a family of inner products $g = \{g_\theta\}_{\theta\in\mathcal{M}}$ (often referred to as the *metric*), one for each point on the manifold, and which can be used to measure angles between vectors at every tangent plane of $\mathcal{M}$. We will use $\mathcal{T}_\theta\mathcal{M}$ to denote the tangent plane at a given $\theta \in \mathcal{M}$. The standard example of a Riemannian manifold is $\mathcal{M} = \mathbb{R}^d$ with g_θ the Euclidean inner product at every point. More general examples of Riemannian manifolds with $\mathcal{M} = \mathbb{R}^d$ can be generated from a field of positive definite matrices H. Consider the family of inner products:

$$g_\theta(u, v) := \langle H(\theta)u, v \rangle, \quad u, v \in \mathbb{R}^d,$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product. Notice that g_θ is indeed an inner product since $H(\theta)$ is positive definite. In what follows we often suppress the dependence of g on θ for brevity.

We now proceed to define the notion of the gradient of a function F defined over an arbitrary Riemannian manifold. Let $(\mathcal{M}, g)$ be a Riemannian manifold and let $F : \mathcal{M} \to \mathbb{R}$ be a smooth enough function. The gradient of F at the point $\theta \in \mathcal{M}$ relative to the metric g, denoted $\nabla_gF(\theta)$, is defined as the vector in $\mathcal{T}_\theta\mathcal{M}$ for which the following identity holds:

$$\frac{d}{dt}F(\gamma(t))\big|_{t=0} = g(\nabla_gF(\theta), \dot{\gamma}(0)) \qquad (2.11)$$

for any differentiable curve $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}$ with $\gamma(0) = \theta$; by $\dot{\gamma}(0)$ we mean the velocity of the curve γ at time 0, which is an element in $\mathcal{T}_\theta\mathcal{M}$. In the example of $\mathcal{M} = \mathbb{R}^d$ with inner products induced by a field H (which we denote with g_H), we have that

$$\begin{aligned}\frac{d}{dt}V(\gamma(t))\big|_{t=0} &= \langle \nabla V(\theta), \dot{\gamma}(0) \rangle \\ &= \langle H(\theta)H(\theta)^{-1}\nabla V(\theta), \dot{\gamma}(0) \rangle \\ &= g(H(\theta)^{-1}\nabla V(\theta), \dot{\gamma}(0)),\end{aligned}$$

for any curve $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}$ with $\gamma(0) = \theta$, from where it follows that $\nabla_{g_H}V(\theta) = H(\theta)^{-1}\nabla V(\theta)$, where we recall ∇ denotes the usual gradient in $\mathbb{R}^d$.

In this light, (2.10) can be interpreted as a gradient descent algorithm, only that the gradient is taken w.r.t. a metric that is different from the standard Euclidean one. In section 3, where we discuss sampling, we will return to some of the insights that we have developed in this section. In particular, in order to define gradient descent dynamics of a functional over a manifold we need to specify two ingredients: 1) an energy V to optimize, and 2) a metric g_θ under which we define the gradient. For the last item, it will be convenient to have a clear understanding of how to represent smooth curves in the manifold of interest and characterize their velocities appropriately. Indeed, equation (2.11) explicitly relates the metric g_θ of the manifold, the target energy V, the rate of change of the energy along arbitrary smooth curves γ, and the gradient ∇_gV of the energy relative to the chosen metric.

2.3.2. *Geodesic convexity.*
There are analogous conditions to the PL and strong convexity assumptions discussed in section 2.1 that guarantee the convergence of the flow (2.10) toward global minima of V. First, write (2.10) as an energy dissipation equality of the form

$$V(\theta_t) = V(\theta_s) - \frac{1}{2}\int_s^t |\nabla_gV(\theta_r)|^2_{\theta_r}dr - \frac{1}{2}\int_s^t |\dot{\theta}_r|^2_{\theta_r}dr,$$
$$(2.12)$$

where we have used $|\cdot|^2_\theta$ to denote $g_\theta(\cdot, \cdot)$. The equivalence between (2.10) and (2.12) follows from an identical argument as in Proposition 2.1 applied to an arbitrary inner product. The analogous PL condition in the preconditioned setting takes the form:

$$\alpha(V(\theta) - V^*) \le \frac{1}{2}|\nabla_gV(\theta)|^2_\theta, \qquad (2.13)$$

which generalizes the PL condition in the Euclidean setting to general inner products and gradients.

To introduce an appropriate notion of convexity that allows us to generalize the results of section 2.1 we need to introduce a few more ideas from Riemannian geometry. Given a Riemannian manifold $(\mathcal{M}, g)$, we define the geodesic distance d_g induced by the metric g as:

$$d_g^2(\theta, \theta') = \inf_{t \in [0,1] \mapsto (\gamma_t, \dot{\gamma}_t)} \int_0^1 |\dot{\gamma}_t|^2_{\gamma_t} dt. \quad (2.14)$$

We will say that $\gamma : [0,1] \to \mathcal{M}$ is a constant speed geodesic between θ and θ' if γ is a minimizer of the right-hand side of the above expression. Equivalently, a constant speed geodesic between θ and θ' is any curve with $\gamma(0) = \theta$ and $\gamma(1) = \theta'$ such that $d_g(\gamma(s), \gamma(t)) = |t - s| d_g(\theta, \theta')$ for all $s, t \in [0,1]$. The advantage of the latter definition is that it is completely described in terms of the distance function d_g and in particular does not require explicit mention of the Riemannian structure of the space.

We can now define the notion of α-*geodesic convexity*. We say $V : \mathcal{M} \to \mathbb{R}$ is α-geodesically convex if for all $\theta, \theta' \in \mathcal{M}$ there exists a constant speed geodesic $\gamma : [0,1] \to \mathcal{M}$ between them, such that

$$V(\gamma(t)) \le tV(\theta) + (1-t)V(\theta') - \frac{\alpha}{2}t(1-t)d_g(\theta, \theta')^2, \quad (2.15)$$

for all $t \in [0,1]$.

Notice that α-geodesic convexity reduces to α-strong convexity when $(\mathcal{M}, g)$ is an Euclidean space. Also, it can be shown that α-geodesic convexity for $\alpha > 0$ implies the generalized PL condition (2.13) (see Lemma 11.28 in [Bou23]), which in turn implies, following the proof of Proposition 2.3, exponential decay rates for the energy V along its gradient flow, in direct analogy with Proposition 2.3.

Remark 2.4. Equation (2.14) relates the distance function d_g with the family of inner products g. This formula is very useful as it allows us to recover the metric g from its distance function d_g. This observation will be relevant when discussing the formal Riemannian structures on spaces of probability measures in the context of sampling in section 3.3.1.

2.4. Time discretizations.

2.4.1. *Standard gradient descent.*
In this section we discuss how to obtain practical optimization algorithms by discretizing the gradient system (2.2) in time. First, the *explicit Euler* scheme gives the standard gradient descent iteration

$$\theta_{n+1} = \theta_n - \tau \nabla V(\theta_n). \quad (2.16)$$

In numerical analysis of differential equations, $\tau > 0$ is interpreted as a small time-step; then, if (2.16) and (2.2) are initialized at the same point θ_0, it holds that $\theta_n \approx \theta_t$ for $t = n\tau$. In the optimization context of interest, τ is referred to as a learning rate and it is insightful to note that (2.16)

can be defined variationally as

$$\theta_{n+1} = \mathrm{argmin}_\theta \left(\langle \nabla V(\theta_n), \theta - \theta_n \rangle + \frac{1}{2\tau}|\theta - \theta_n|^2 \right).$$

Thus, θ_{n+1} is found by minimizing $V(\theta_n) + \langle \nabla V(\theta_n), \theta - \theta_n \rangle + \frac{1}{2\tau}|\theta - \theta_n|^2$, noticing that the first two terms form the first order approximation of the objective V around the most recent iterate θ_n. In practice, the learning rate may be chosen adaptively using a line search [NW99].

Compared to the continuous-time setting, energy dissipation and convergence of the explicit Euler scheme require further assumptions on the function V. The following proposition is analogous to Proposition 2.8 but relies on a *smoothness* assumption on the gradient of V additional to the PL condition.

Proposition 2.5. *Suppose that V has L-Lipschitz gradient, has minimum V^*, and satisfies the PL condition. Then the gradient descent algorithm defined by (2.16) with step-size $\tau := \frac{1}{L}$ has a linear convergence rate. More precisely, it holds that*

$$V(\theta_{n+1}) - V^* \le \left(1 - \frac{\alpha}{L}\right)^n \left(V(\theta_0) - V^*\right). \quad (2.17)$$

Proof. A classical result in convex analysis ensures that the assumption that ∇V is L-Lipschitz implies

$$V(\theta_{n+1}) \le V(\theta_n) + \langle \nabla V(\theta_n), \theta_{n+1} - \theta_n \rangle + \frac{L}{2}|\theta_{n+1} - \theta_n|^2.$$

Using (2.16), we then deduce that

$$V(\theta_{n+1}) \le V(\theta_n) - \frac{1}{2L}|\nabla V(\theta_n)|^2,$$

which combined with the PL condition gives

$$V(\theta_{n+1}) - V^* \le V(\theta_n) - V^* - \frac{\alpha}{L}\left(V(\theta_n) - V^*\right)$$

$$= \left(1 - \frac{\alpha}{L}\right)\left(V(\theta_n) - V^*\right).$$

The result follows by induction. $\qquad \square$

From the proof of Proposition 2.5 we see that, under the L-smoothness condition and assuming that the step size τ is sufficiently small, the explicit Euler scheme dissipates the energy V. Moreover, this condition helps us quantify the amount of dissipation in one iteration of the scheme in terms of the norm of the gradient of V at the current iterate.

As an alternative discretization, one can consider the *implicit Euler* scheme:

$$\theta_{n+1} = \theta_n - \tau \nabla V(\theta_{n+1}), \quad (2.18)$$

which coincides with the first order optimality conditions for

$$\theta_{n+1} = \mathrm{argmin}_\theta \left(V(\theta) + \frac{1}{2\tau}|\theta - \theta_n|^2 \right). \quad (2.19)$$

It follows directly from the definition of the implicit Euler scheme that it satisfies a dissipation inequality analogous

to (2.5) without imposing any additional smoothness conditions on V. However, an important caveat is that determining θ_{n+1} from θ_n requires solving a new optimization problem (2.19) or finding a root for the (in general) nonlinear equation (2.18). On the other hand, from a theoretical perspective the implicit Euler scheme, or *minimizing movement scheme* as it is called in [AGS08], is an important tool for proving existence of gradient flow equations in general metric spaces; see Chapters 1–2 in [AGS08].

2.4.2. *Discretizations and preconditioning.* One possible time discretization for (2.10) is given by

$$\theta_{n+1} = \theta_n - \tau H(\theta_n)^{-1}\nabla V(\theta_n),$$

which is a direct adaptation of (2.16) to the preconditioned setting. Proposition 2.5 can be readily adapted using the PL condition and L-smoothness of V relative to the geometry induced by the field H.

Remark 2.6. In line with the discussion at the end of section 2.3, we notice that the effect of scaling the field H by a constant $1/r$ is to scale the constants in both the PL condition and the L-smoothness condition by a factor of r. The net gain in (2.17) from rescaling the metric is thus null.

Another possible time discretization for (2.10) when the function H is the Hessian of a strictly convex function $h : \Theta \to \mathbb{R}$ (not necessarily equal to the objective function V) is the *mirror descent* scheme:

$$\begin{cases} z_{n+1} = z_n - \tau\nabla V(\theta_n), \\ \theta_{n+1} = (\nabla h)^{-1}(z_{n+1}). \end{cases} \quad (2.20)$$

The idea in mirror descent is to update an associated mirror variable (a transformation of θ by a mirror map, in this case ∇h) using a gradient step, as opposed to directly updating the variable θ as in the standard explicit Euler scheme. Using a Taylor approximation of $(\nabla h)^{-1}$ around z_t we see that

$$\theta_{n+1} = (\nabla h)^{-1}(z_n - \tau\nabla V(\theta_n))$$
$$\approx \theta_n - \tau H(\theta_n)^{-1}\nabla V(\theta_n),$$

revealing why mirror descent can be regarded as an approximation of (2.10) when H is the Hessian of h.

It is worth remarking that the update rule (2.20) has the following variational characterization:

$$\theta_{n+1} := \mathrm{argmin}_{\theta\in\Theta}\left(\langle\nabla V(\theta_n),\theta\rangle + \frac{1}{\tau}D_h(\theta\|\theta_n)\right), \quad (2.21)$$

where the function $D_h(\theta\|\theta')$ has the form

$$D_h(\theta\|\theta') := h(\theta) - h(\theta') - \langle\nabla h(\theta'),\theta - \theta'\rangle,$$

and is often referred to as *Bregman divergence*. This variational characterization was discovered and used in [BT03] to deduce convergence properties of mirror descent. Notice that the strict convexity of h guarantees that the function $D_h(\cdot\|\cdot)$ is non-negative and zero only when both of
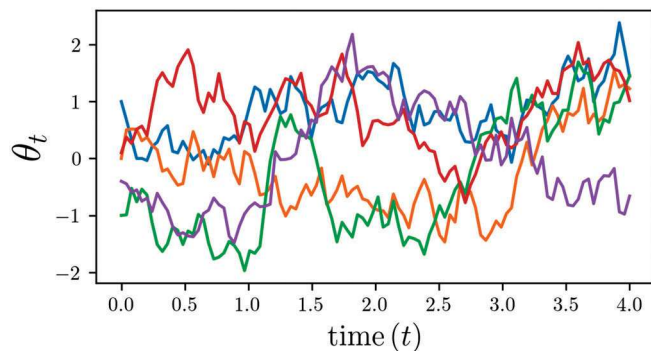

Figure 3. Five trajectories of Langevin dynamics with double well potential V given by (2.6).

its arguments coincide. Bregman divergences thus play a similar role to the one played by the quadratic function $\frac{1}{2}|\cdot - \cdot|^2$ in the variational form of the standard explicit Euler scheme.

3. Sampling

While direct sampling from certain distributions, e.g., Gaussians, may be rather straightforward, sampling from a general target distribution can be challenging, especially in high-dimensional settings. In this section we consider the problem of sampling a target density $\pi(\theta) \propto \exp(-V(\theta))$. As guiding examples, one may consider sampling a posterior or Boltzmann distribution, see (1.2)–(1.3). We start by introducing Langevin dynamics in section 3.1, a stochastic differential equation that resembles the gradient system (2.2), but which incorporates a Brownian motion that makes the solution trajectories $\{\theta_t\}_{t\geq 0}$ random. In section 3.2 we present some results that state that, under suitable assumptions, the density ρ_t of θ_t converges to the desired target density π as $t \to \infty$. In section 3.3 we discuss how the Langevin dynamics define a gradient flow in the space of probability distributions. Finally, section 3.4 discusses how to use discretizations of Langevin dynamics to obtain practical sampling algorithms. Our presentation here parallels that of section 2.

3.1. **Langevin dynamics.** Consider the *overdamped Langevin* diffusion [Pav14]

$$d\theta_t = -\nabla V(\theta_t)\,dt + \sqrt{2}\,dB_t, \quad (3.1)$$

where $\{B_t\}_{t\geq 0}$ is a Brownian motion on $\Theta = \mathbb{R}^d$. Langevin dynamics can be interpreted as a stochastic version of the gradient descent dynamics (2.2). This is illustrated in the following example, which also provides intuition on the connection between Langevin dynamics and sampling.

Example 3.1. Consider Langevin dynamics with the double well potential V introduced in (2.6). Figure 3 shows five trajectories, initialized as in Figure 1. For each $t > 0$, θ_t is now a random variable, whose Lebesgue density will be denoted by ρ_t in what follows. Figure 4 shows an
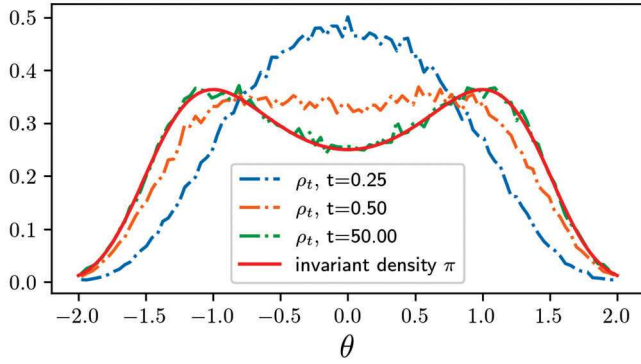
Figure 4. Histograms of ρ_t at $t = 0.25, 0.5, 50$. For large t, ρ_t is close to the target density $\pi \propto e^{-V}$.

approximation of ρ_t for $t \in \{0.25, 0.5, 50\}$ obtained by simulating $N = 10^5$ solution trajectories. Notice that at $t = 50$, ρ_t is exceedingly close to the target density $\pi(\theta) \propto \exp(-V(\theta))$, so that θ_t can be viewed as a sample from π. Thus, while θ_t is random due to the Brownian motion, the density $\rho_t(\theta)$ is larger at points θ where $V(\theta)$ is small. □

3.2. A note on convergence. It is natural to ask whether the law $\{\rho_t\}_{t \geq 0}$ of a given stochastic process converges to an *invariant* distribution π. For the Langevin diffusion, the positive answer illustrated in Example 3.1 holds under suitable assumptions on V that are analogous to the PL and strong convexity conditions in section 2. A natural way to study the long-time behavior of ρ_t is to derive a differential equation for its evolution. To that end, one may characterize the time derivative of the action of ρ_t on suitable test functions $\phi : \Theta \to \mathbb{R}$. More precisely, we compute $\frac{d}{dt} \int \phi(\theta) d\rho_t(\theta)$, which is a standard derivative of a function from the real line to itself. For the Langevin diffusion (3.1) it can be proved that:

$$\frac{d}{dt} \int_\Theta \phi(\theta) d\rho_t(\theta) = - \int_\Theta \nabla\phi \cdot \nabla(V + \log(\rho_t)) d\rho_t(\theta), \quad (3.2)$$

$\forall t > 0, \forall \phi \in C_c^\infty(\Theta)$. The above condition is the *weak formulation* of the Fokker–Planck equation:

$$\partial_t \rho_t = \mathrm{div}(\rho_t \nabla(V + \log(\rho_t))) =: \mathcal{L}\rho_t. \quad (3.3)$$

From now on we interpret (3.3) in its weak form (3.2).

We observe that $\pi \propto e^{-V}$ is a stationary point of the dynamics (3.3). That is, if we initialize the dynamics at $\rho_0 = \pi$, then $\rho_t := \rho_0$ for all $t > 0$ is a solution to the equation. The next result describes the long-time behavior of a solution to the Fokker–Planck equation when initialized at more general ρ_0.

Theorem 3.2. *Let ρ_t be the solution to the Fokker–Planck equation with $\rho_0 \in L^2(\pi^{-1})$, where $L^2(\pi^{-1})$ is the L^2 space with the weight function π^{-1}. Suppose that there is $\alpha > 0$ such that the following Poincaré inequality holds: for every*

$f \in C^1 \cap L^2(\pi)$ *that has zero mean under π, it holds that $\alpha\|f\|_{L^2(\pi)}^2 \leq \|\nabla f\|_{L^2(\pi)}^2$. Then it holds that*

$$\|\rho_t - \pi\|_{L^2(\pi^{-1})} \leq e^{-\alpha t}\|\rho_0 - \pi\|_{L^2(\pi^{-1})}.$$

Proof. Define u_t by $\rho_t = u_t \pi$. We can verify that

$$\partial_t u_t = -\nabla V \cdot \nabla u_t + \mathrm{div}(\nabla u_t), \quad u_0 = \rho_0 \pi^{-1}.$$

Therefore, the zero-mean function $u_t - 1$ satisfies

$$\frac{\partial(u_t - 1)}{\partial t} = \mathcal{L}(u_t - 1). \quad (3.4)$$

Multiplying by $(u_t - 1)\pi$, integrating, and using that by assumption $\alpha\|u_t - 1\|_{L^2(\pi)}^2 \leq \|\nabla u_t\|_{L^2(\pi)}^2$, we deduce that

$$\frac{1}{2}\frac{d}{dt}\|u_t - 1\|_{L^2(\pi)}^2 \leq -\alpha\|u_t - 1\|_{L^2(\pi)}^2. \quad (3.5)$$

Gronwall's inequality gives the desired result. □

The above result implies that, as time t goes to infinity, the distribution ρ_t converges toward the target density $\pi \propto e^{-V}$ exponentially fast. The notion of convergence implied by Theorem 3.2, however, is not as strong as other notions such as the convergence in KL divergence that will be discussed in the next section. In particular, Theorem 3.2 should be contrasted with the discussion in section 3.3.2.

3.3. Choice of objective and metric. Here we discuss a concrete variational interpretation for the Langevin system (3.3). In essence, this entails viewing sampling as an optimization algorithm (in particular, as a gradient flow) that aims at recovering the target density π. As discussed toward the end of section 2.3, to realize this interpretation it is important to identify precisely the geometric objects involved in the definition of a gradient flow (energy, metric, etc.). In all subsequent sections, $\mathcal{M}$ will be the space of probability measures over Θ, which we will denote with $\mathcal{P}(\Theta)$, and F will be the *KL divergence* relative to π, which we recall is defined as:

$$D_{\mathrm{KL}}(\nu\|\pi) := \int_\Theta \log\left(\frac{\nu(\theta)}{\pi(\theta)}\right)\nu(\theta)\, d\theta, \quad (3.6)$$

whenever ν is absolutely continuous w.r.t. π and $D_{\mathrm{KL}}(\nu\|\pi) = \infty$ otherwise. Different choices of metric over $\mathcal{P}(\Theta)$ will induce different evolution equations (recall our discussion of preconditioning in the context of optimization over the parameter space Θ), and thus different optimization schemes. In section 3.3.1 we discuss a specific geometric structure for $\mathcal{P}(\Theta)$ that realizes the Fokker–Planck equation of the Langevin diffusion as a gradient flow of F, and in section 4 we discuss other gradient flow structures that motivate other sampling algorithms.

Remark 3.3. It holds that $D_{\mathrm{KL}}(\nu\|\pi) \geq 0$ for all ν, and equality holds if and only if $\nu = \pi$. In particular, the unique minimizer of $\nu \in \mathcal{P}(\Theta) \mapsto D_{\mathrm{KL}}(\nu\|\pi)$ is the target π. Although trivial, this observation motivates *variational inference*—see [LCB$^+$22] and references therein—a method for

producing tractable proxies for π that relies on the minimization of $D_{KL}(\cdot\|\nu)$ over a user-chosen family of tractable distributions. $\qquad\square$

Remark 3.4. If π has a density w.r.t. the Lebesgue measure that is proportional to e^{-V}, then $D_{KL}(\nu\|\pi) < \infty$ implies that ν is also absolutely continuous w.r.t. the Lebesgue measure. In that case we will abuse notation slightly and use ν to also denote ν's corresponding density. In particular, when we write $\log(\nu)$ it is understood that ν is interpreted as the density function w.r.t. Lebesgue measure of the measure ν.

3.3.1. *KL and Wasserstein.* Following a series of seminal works that started with a paper by Jordan, Kinderlehrer, and Otto in the late 90s (see sections 8.1. and 8.2. in [Vil03]) we will interpret equation (3.3) as the gradient flow of the energy F w.r.t the *Wasserstein* metric. Given $\rho, \rho' \in \mathcal{P}(\Theta)$ with finite second moments, their Wasserstein distance $W_2(\rho, \rho')$ is given by

$$W_2^2(\rho, \rho') := \min_{\Upsilon \in \Gamma(\rho,\rho')} \int_{\Theta\times\Theta} |\theta - \theta'|^2 d\Upsilon(\theta, \theta'), \qquad (3.7)$$

where $\Gamma(\rho, \rho')$ is the set of couplings between ρ and ρ', i.e. the set of Borel probability measures on the product space $\Theta \times \Theta$ with first and second marginals equal to ρ and ρ', respectively.

Formula (3.7), although simple, does not reveal the infinitesimal geometric structure of the Wasserstein space to define gradients of functionals over $\mathcal{P}(\Theta)$. What is missing is a representation of the distance W in a form similar to (2.14). The next result by Benamou and Brenier (see section 8.1. in [Vil03]) provides the missing elements.

Proposition 3.5. *Let $\rho, \rho' \in \mathcal{P}(\Theta)$. Then*

$$W_2^2(\rho, \rho') = \inf_{t\in[0,1]\mapsto(\gamma_t, \nabla\varphi_t)} \int_0^1 \int_\Theta |\nabla\varphi_t(\theta)|^2 d\gamma_t(\theta)dt$$
$$s.t. \partial_t\gamma_t + \mathrm{div}(\gamma_t\nabla\varphi_t) = 0, \qquad (3.8)$$
$$\gamma(0) = \rho, \gamma(1) = \rho'.$$

The infimum in (3.8) is taken over all maps $t \in [0, 1] \to (\gamma_t, \nabla\varphi_t)$, where each pair $(\gamma_t, \nabla\varphi_t)$ consists of a measure $\gamma_t \in \mathcal{P}(\Theta)$ and a vector field of the form $\nabla\varphi_t$ for a smooth $\varphi_t : \Theta \to \mathbb{R}$, that satisfy the *continuity equation*:

$$\partial_t\gamma_t + \mathrm{div}(\gamma_t\nabla\varphi_t) = 0, \qquad (3.9)$$

interpreted in weak form. Notice that the Fokker–Planck equation (3.3) is a particular case of the continuity equation with $\varphi_t = V + \log(\rho_t)$. Inspired by equation (2.14) (see also Remark 2.4) we can provide a geometric interpretation of identity (3.8): the continuity equation (3.9) provides a representation of curves in the formal manifold $\mathcal{M} = \mathcal{P}(\Theta)$. In this representation, the velocity of a

curve (tangent vector) at each point in the curve can be identified with a vector field (over Θ) of the form $\nabla\varphi$. Furthermore, (3.8) motivates introducing an inner product at each $\nu \in \mathcal{P}(\Theta)$ of the form:

$$g_\nu(\nabla\varphi, \nabla\varphi') := \int_\Theta \nabla\varphi(\theta) \cdot \nabla\varphi'(\theta) d\nu(\theta).$$

With the above geometric interpretation in place, we may follow equation (2.11) and identify the gradient of $F(\cdot) = D_{KL}(\cdot\|\pi)$ at an arbitrary point ν. For this purpose take an arbitrary solution $(\gamma_t, \nabla\varphi_t)$ to the continuity equation (3.9) (i.e., take an arbitrary curve in $\mathcal{M}$) for which $F(\gamma_t) < \infty$ and compute:

$$\frac{d}{dt}F(\gamma_t) = \frac{d}{dt}\int_\Theta \log\left(\frac{\gamma_t}{e^{-V}}\right)d\gamma_t(\theta)$$
$$= \int_\Theta \nabla\varphi_t \cdot \nabla(V + \log(\gamma_t))d\gamma_t(\theta)$$
$$= g_{\gamma_t}(\nabla\varphi_t, \nabla(V + \log(\gamma_t))).$$

In the above we have gone from the first line to the second one using the weak form of the continuity equation; to go from the second to third line we have used the definition of g_{γ_t}.

From the above computation we conclude that the gradient of F (w.r.t. to the Wasserstein metric) at a point ν for which $F(\nu) < \infty$ takes the form $\nabla(V+\log(\nu))$. In particular, the curve in $\mathcal{M} = \mathcal{P}(\Theta)$ whose velocity vector agrees with the negative gradient of the functional F takes the form of the Fokker–Planck equation (3.3). In other words, (3.3) can be interpreted as the gradient flow of $F(\cdot) = D_{KL}(\cdot\|\pi)$ w.r.t. the Wasserstein metric.

Remark 3.6. To some extent, the computations in this section have been formal and some of the above derivations have been left unjustified. These computations rely on a formal adaptation of formula (2.2) to the setting of the Riemannian manifold $\mathcal{P}(\Theta)$ endowed with the Wasserstein distance. For a rigorous treatment of the topics discussed in this section the reader is referred to the second part of the book [AGS08]. There, the notion of gradient flow in $\mathcal{P}(\Theta)$ is motivated by the dissipation identity (2.3) and adapted to the metric space $(\mathcal{P}(\Theta), W_2)$.

Remark 3.7. At a high level, the ideas discussed in this section can be used to propose flows aimed at solving variational inference problems like the ones briefly mentioned in Remark 3.3. Indeed, following the geometric intuition from projected gradient descent methods, where one uses the projection of the negative gradient of the objective onto the tangent plane of the constrained set to define the projected gradient descent flow, one may consider the projection of the negative gradient of the energy F (w.r.t Wasserstein) onto the tangent planes of the submanifold $\mathcal{G}$;

naturally, in this setting the notion of (orthogonal) projection is taken w.r.t. the Riemannain metric underlying the Wasserstein space. This idea has been recently explored in [LCB+22] for certain families $\mathcal{G}$ of tractable distributions.

3.3.2. *Geodesic convexity of the relative entropy in the Wasserstein space.*

The definition of geodesic convexity introduced in (2.15) can be readily adapted to the setting of an energy defined over an arbitrary metric space, and in particular to the setting of $\mathcal{P}(\theta)$ endowed with the Wasserstein distance. Indeed, notice that equation (2.15) is completely determined by the energy of interest, the distance function, and the notion of constant speed geodesic, which in turn can be defined in terms of the distance function.

We have the following theorem by McCann relating the convexity of the function V with the α-geodesic convexity of $D_{KL}(\cdot\|\pi)$ when $\pi \propto e^{-V}$; see Theorem 5.15. in [Vil03].

Theorem 3.8. *Suppose that V is α-strongly convex and let $\pi \propto e^{-V}$. Then $D_{KL}(\cdot\|\pi)$ is α-geodesically convex w.r.t. the Wasserstein distance.*

With Theorem 3.8 in hand, the α-strong convexity of V implies that

$$D_{KL}(\rho_t\|\pi) \le e^{-2\alpha t}D_{KL}(\rho_0\|\pi), \quad \forall t \ge 0,$$

along a solution $\{\rho_t\}_{t\ge 0}$ of the Fokker–Planck equation (3.3). This notion of exponential contraction to equilibrium is not implied by the Poincaré condition from Theorem 3.2.

3.4. Time discretizations.

This section describes how to obtain sampling algorithms from time discretization of the Langevin dynamics (3.1). In analogy to the explicit Euler scheme (2.16) for (2.2), the Euler–Maruyama discretization for (3.1) is given by $\theta_0 \sim \rho_0$, and

$$\theta_{n+1} = \theta_n - \tau\nabla V(\theta_n) + \sqrt{2\tau}\xi_n, \quad \xi_n \overset{\text{i.i.d.}}{\sim} N(0,1). \quad (3.10)$$

For $t = n\tau$, the law ρ_n of θ_n approximates the law ρ_t of θ_t given by (3.1). However, the error introduced by time discretization causes ρ_n to not converge, in general, to π as $n \to \infty$. In other words, the probability kernel $q(\theta_n, \cdot) = \text{law}(\theta_{n+1}|\theta_n)$, defined by the Markov chain (3.10), does not leave π invariant.

To remedy this issue, one may consider using (3.10) as a *proposal kernel* within a Metropolis–Hastings algorithm [Liu01], leading to the *Metropolis Adjusted Langeving Algorithm*, often referred to as MALA. The basic idea is to use an accept/reject mechanism to turn the proposal kernel q into a new Markov kernel that leaves π invariant. Given the current state θ_n, one proposes a move $\theta_n \mapsto \theta_{n+1}^*$ by sampling $q(\theta_n, \cdot)$; the move is accepted with a probability

$$a = \min\left(1, \frac{\pi(\theta_{n+1}^*)}{\pi(\theta_n)}\frac{q(\theta_{n+1}^*, \theta_n)}{q(\theta_n, \theta_{n+1}^*)}\right). \quad (3.11)$$

If the move is accepted, we set $\theta_{n+1} := \theta_{n+1}^*$. Otherwise, we set $\theta_{n+1} := \theta_n$. The Metropolis–Hastings acceptance probability (3.11) is chosen in such a way that π is the invariant distribution of the new chain $\{\theta_n\}_{n=1}^\infty$. Notice that the two steps of the algorithm, namely, sampling from the proposal kernel defined by (3.10) and evaluating (3.11), can be implemented without knowledge of the normalizing constant of π. We refer to [RT96] for further details on the convergence of Langevin diffusions and their discretizations and we refer to [CLGL+20] for a sampling analog of the mirror descent optimization algorithm in (2.20).

4. Modern Twists on Langevin

In this section we outline recent extensions of the gradient flows of section 3.3.1 aimed at sampling. The idea is to employ gradient flows of the energy $F = D_{KL}(\cdot\|\pi)$ w.r.t. metrics beyond the Wasserstein distance. One hopes that the new dynamics lead to faster convergence to minimizers and, in turn, to more efficient sampling algorithms.

4.1. Ensemble preconditioning.

In analogy with section 2.3, we consider preconditioned variants of the Langevin diffusion (3.1),

$$d\theta_t = -H(\theta_t)^{-1}\nabla V(\theta_t)\,dt + \sqrt{2H(\theta_t)^{-1}}\,dB_t, \quad (4.1)$$

where $H(\theta)$ is once again the preconditioning matrix field. The intuition from subsection 2.3 carries over in this setting: by choosing an appropriate preconditioner we can speed up the convergence of the Langevin diffusion to the target density. However, in contrast to (3.1), choosing H as a function of θ_t can in general lead to a nonlinear evolution for the law of the process. Moreover, without additional structure, the target $\pi \propto e^{-V}$ may not be an invariant distribution for (4.1). This motivates discussing suitable choices for H. One approach to constructing the matrix H is to consider an ensemble of particles evolving according to (4.1) and use the location of the particles to construct an appropriate preconditioner. Following [GIHLS20], consider an ensemble of $J \ge 1$ interacting Langevin diffusions $\theta_t := \{\theta_t^{(j)}\}_{j=1}^J$ that are evolved according to the coupled system

$$d\theta_t^{(j)} = -H(\theta_t)^{-1}\nabla V(\theta_t^{(j)}) + \sqrt{2H(\theta_t)^{-1}}\,dB_t^{(j)},$$

where $\{B_t^{(j)}\}_{j=1}^J$ are i.i.d. Brownian motions, and

$$H(\theta_t) = \frac{1}{J}\sum_{j=1}^J \left(\theta_t^{(j)} - \bar{\theta}_t\right)\left(\theta_t^{(j)} - \bar{\theta}_t\right)^\top,$$

for $\bar{\theta}_t$ the ensemble mean of the $\theta_t^{(j)}$. In other words, the preconditioner is chosen to be the empirical covariance matrix of the ensemble at each point in time, a quantity that is convenient to compute in practice.

As in previous sections, one can show that the ensemble preconditioned Langevin dynamics has a gradient flow structure. Taking the mean-field limit, i.e., letting $J \to \infty$, we formally obtain the following diffusion for the evolution of the ensemble

$$d\theta_t = -H(\rho_t)^{-1} \nabla V(\theta_t) \, dt + \sqrt{2H(\rho_t)^{-1}} \, dB_t,$$

where ρ_t denotes the distribution of θ_t as before and $H(\rho_t)$ is the covariance matrix of ρ_t. The Fokker–Planck equation for ρ_t is given by

$$\partial_t \rho_t = \mathrm{div}\big(\rho_t H(\rho_t)^{-1}(\nabla V(\theta_t) + \nabla \log \rho_t)\big), \qquad (4.2)$$

which has the target $\pi \propto e^{-V}$ as a stationary point. The divergence form in (4.2) suggests a gradient flow structure as shown in [GIHLS20]. Indeed, the above equation is a gradient flow of F w.r.t. the *Kalman–Wasserstein* distance W_K defined as

$$W_K^2(\rho, \rho') := \inf_{\gamma_t, \varphi_t} \int_0^1 \int \langle \nabla \varphi_t, H(\rho_t) \nabla \varphi_t \rangle \rho_t \, d\theta \, dt,$$
$$\text{s.t. } \partial_t \gamma_t + \mathrm{div}(\gamma_t H(\gamma_t) \nabla \varphi_t) = 0,$$
$$\gamma(0) = \rho, \quad \gamma(1) = \rho'.$$

4.2. Langevin with birth-death. One of the shortcomings of the Langevin dynamics is that its convergence suffers when sampling from multimodal target densities: the process may get stuck around one of the modes and it may take a long time to cross the energy barrier between various modes or to overcome entropic bottlenecks. To ameliorate this metastable behavior [LLN19] proposes to consider the *birth-death accelerated Langevin (BDL)* dynamics

$$\partial_t \rho_t = \mathcal{L}\rho_t + \rho_t(\log \pi - \log \rho_t)$$
$$- \rho_t \int_\theta (\log \pi - \log \rho_t) d\rho_t(\theta), \qquad (4.3)$$

where $\mathcal{L}$ is as in (3.1). Notice that compared with the Fokker–Planck equation (3.3), equation (4.3) contains two additional terms. The second term on the right-hand side favors increasing (resp. decreasing) $\rho_t(\theta)$ whenever $\rho_t(\theta) < \pi(\theta)$ (resp. $\rho_t(\theta) > \pi(\theta)$), hence the name birth-death. The third term is only included to ensure that ρ_t remains a probability distribution through the birth-death process. The birth-death terms make the equation nonlocal (due to averaging) and allow the dynamics to explore the support of a multimodal π more efficiently: if the dynamics get stuck in one mode, the birth-death process can still transfer some mass to another mode that has not yet been thoroughly explored.

It turns out that, akin to the Langevin dynamics, BDL is also a gradient flow of the KL divergence but w.r.t. a modification of the Wasserstein distance referred to as the *Wasserstein–Fisher–Rao (or the Spherical Hellinger–Kantorovich)* distance:

$$W_{\mathrm{FR}}^2(\rho, \rho')$$
$$= \inf_{\gamma_t, \varphi_t} \int_0^1 \left(\int |\nabla \varphi_t|^2 + |\varphi_t|^2 d\gamma_t - \left(\int \varphi_t d\gamma_t \right)^2 \right) dt$$

subject to the constraints

$$\partial_t \gamma_t + \mathrm{div}(\gamma_t \nabla \varphi_t) = -\gamma_t \Big(\varphi_t - \int \varphi_t d\gamma_t\Big), \qquad (4.4a)$$
$$\gamma(0) = \rho, \quad \gamma(1) = \rho'. \qquad (4.4b)$$

As before, the BDL continuity equation (4.4a) is understood in the weak sense and plays the same role as the continuity equation in the Benamou–Brenier formulation (3.8). Equation (4.4a) thus provides an alternative representation for admissible curves in the space $\mathcal{P}(\Theta)$. Motivated by this, we will thus think of a pair $(\varphi, \nabla\varphi)$ as tangent to a given point $\nu \in \mathcal{P}(\Theta)$ and introduce the inner product at the point ν

$$g_\nu\big((\varphi, \nabla\varphi), (\varphi', \nabla\varphi')\big) := \int \nabla\varphi \cdot \nabla\varphi' d\nu$$
$$+ \int \varphi\varphi' d\nu - \int \varphi d\nu \int \varphi' d\nu.$$

With this geometric structure for the space $\mathcal{P}(\Theta)$ endowed with the WFR metric, we can proceed to carry out an analogous computation to the one at the end of section 3.3.1 and show that the BDL dynamics is the gradient flow of the energy functional $D_{\mathrm{KL}}(\cdot\|\pi)$ w.r.t. the W_{FR} geometry; see [LLN19, Thms. 3.2 and 3.3]. Furthermore, the rate of convergence of BDL is at least as good as that of Langevin dynamics and is asymptotically independent of the negative-log-density V, making it suitable for exploration of multimodal landscapes. However, an important caveat is that turning BDL into a sampling algorithm requires utilizing an ensemble of interacting Langevin trajectories to empirically approximate the mean-field dynamics (4.3) using a kernel density estimation, which may be forbiddingly expensive in high-dimensional settings.

5. Conclusion and Discussion

This article provided a gentle introduction to gradient flows as a unifying framework for the design and analysis of optimization and sampling algorithms. Three key elements are involved in specifying a gradient flow: the space of interest (parameter or distribution); an energy function defined on that space to be minimized; and a geometric notion of a gradient. After discussing different versions of gradient flows for optimization, we mostly focused on Langevin dynamics as a sampling analog to gradient descent and discussed its generalizations through ensemble preconditioning and addition of nonlocal terms.

The flexibility of the gradient flow framework gives significant freedom in the choice of the energy function and geometry to design new algorithms beyond the examples discussed in this article. For instance, in [AKSG19] the gradient flow of the maximum mean discrepancy (as opposed to KL) w.r.t. the Wasserstein distance is used to analyze an ensemble sampling algorithm. As another example, the *Stein Variational Gradient Descent* or SVGD algorithm [Liu17] can be viewed as a gradient flow of KL w.r.t. to a modified Wasserstein distance defined from an appropriate reproducing kernel Hilbert space.

To close, we point out that many alternative approaches that do not rely on direct reference to gradient flows can be used to enhance the convergence of Langevin dynamics and ameliorate their metastable behavior. For example, *Hamiltonian Monte Carlo* or HMC (see [Liu01] Chapter 10) utilizes a proposal kernel obtained by discretization of (deterministic) Hamilton equations as opposed to (stochastic) Langevin dynamics; theoretical and empirical evidence suggests that HMC scales favorably to high-dimensional settings. As another example, momentum methods may be used to accelerate the convergence of optimization and sampling algorithms; these methods may be interpreted as arising from time discretization of higher-order systems that approximate a gradient flow structure in certain limiting regimes [KS21]. Nonreversible variants of the Langevin diffusion can potentially achieve faster convergence (see section 4.8 in [Pav14]), while umbrella, tempering, and annealing sampling strategies (see chapter 10 in [Liu01]) enable the efficient traversing of multimodal targets. Finally, recent machine learning techniques such as normalizing flows [KPB20] offer an alternative approach to sampling and density estimation by direct parameterization of transport maps using neural networks. Employing the so called neural ODE models leads to formulations that closely resemble gradient flows and the continuity equation (3.9) with the vector field $\nabla \varphi_t$ replaced by a neural network.

References

[AGS08] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*, 2nd ed., Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2008. MR2401600

[AKSG19] M. Arbel, A. Korba, A. Salim, and A. Gretton, *Maximum mean discrepancy gradient flow*, Advances in Neural Information Processing Systems **32** (2019).

[Bou23] N. Boumal, *An introduction to optimization on smooth manifolds*, Cambridge University Press, Cambridge, 2023. MR4533407

[BT03] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters **31** (2003), no. 3, 167–175.

[CLGL+20] S. Chewi, T. Le Gouic, C. Lu, T. Maunu, P. Rigollet, and A. Stromme, *Exponential ergodicity of mirror-Langevin diffusions*, Advances in Neural Information Processing Systems **33** (2020), 19573–19585.

[GCSR95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, Texts in Statistical Science Series, Chapman & Hall, London, 1995. MR1385925

[GIHLS20] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart, *Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler*, SIAM J. Appl. Dyn. Syst. **19** (2020), no. 1, 412–441, DOI 10.1137/19M1251655. MR4059375

[GTSA20] N. Garcia Trillos and D. Sanz-Alonso, *The Bayesian update: variational formulations and gradient flows*, Bayesian Anal. **15** (2020), no. 1, 29–56, DOI 10.1214/18-BA1137. MR4050876

[KNS16] H. Karimi, J. Nutini, and M. Schmidt, *Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition*, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2016, pp. 795–811.

[KPB20] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, *Normalizing flows: An introduction and review of current methods*, IEEE Transactions on Pattern Analysis and Machine Intelligence **43** (2020), no. 11, 3964–3979.

[KS21] N. B. Kovachki and A. M. Stuart, *Continuous time analysis of momentum methods*, J. Mach. Learn. Res. **22** (2021), Paper No. 17, 40, DOI 10.4995/agt.2021.13084. MR4253710

[LCB+22] M. Lambert, S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet, *Variational inference via Wasserstein gradient flows*, arXiv preprint arXiv:2205.15902 (2022).

[Liu17] Q. Liu, *Stein variational gradient descent as gradient flow*, Advances in Neural Information Processing Systems **30** (2017).

[Liu01] J. S. Liu, *Monte Carlo strategies in scientific computing*, Springer Series in Statistics, Springer-Verlag, New York, 2001. MR1842342

[LLN19] Y. Lu, J. Lu, and J. Nolen, *Accelerating Langevin sampling with birth-death*, arXiv preprint arXiv:1905.09863 (2019).

[NW99] J. Nocedal and S. Wright, *Numerical Optimization*, Vol. 35, Springer Science, 1999.

[Pav14] G. A. Pavliotis, *Stochastic processes and applications*: *Diffusion processes, the Fokker-Planck and Langevin equations*, Texts in Applied Mathematics, vol. 60, Springer, New York, 2014, DOI 10.1007/978-1-4939-1323-7. MR3288096

[RT96] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli **2** (1996), no. 4, 341–363, DOI 10.2307/3318418. MR1440273

[SAST18] D. Sanz-Alonso, A. M. Stuart, and A. Taeb, *Inverse Problems and Data Assimilation*, arXiv preprint arXiv:1810.06191 (2018).

[Vil03] C. Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, 2003, DOI 10.1090/gsm/058. MR1964483
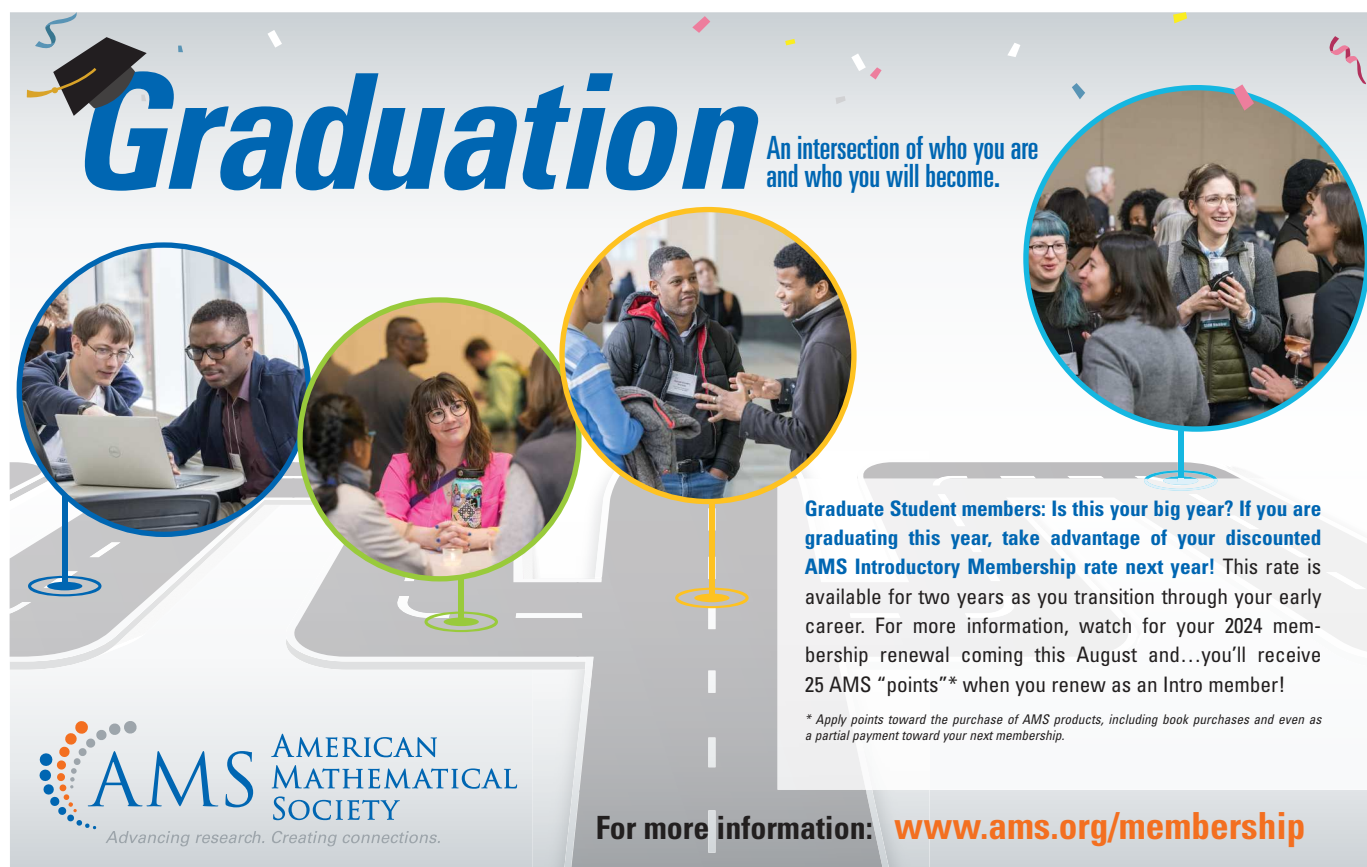
N. García Trillos

B. Hosseini

D. Sanz-Alonso