

Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem

Andre Wibisono

*Department of Electrical & Computer Engineering
University of Wisconsin – Madison*

AYWIBISONO@WISC.EDU

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study sampling as optimization in the space of measures. We focus on gradient flow-based optimization with the Langevin dynamics as a case study. We investigate the source of the bias of the unadjusted Langevin algorithm (ULA) in discrete time, and consider how to remove or reduce the bias. We point out the difficulty is that the heat flow is exactly solvable, but neither its forward nor backward method is implementable in general, except for Gaussian data. We propose the symmetrized Langevin algorithm (SLA), which should have a smaller bias than ULA, at the price of implementing a proximal gradient step in space. We show SLA is in fact consistent for Gaussian target measure, whereas ULA is not. We also illustrate various algorithms explicitly for Gaussian target measure with Gaussian data, including gradient descent, proximal gradient, and Forward-Backward, and show they are all consistent.

1. Introduction

We study sampling as optimization in the space of measures. In this paper we focus on gradient flow-based optimization with the Langevin dynamics as a case study. Our starting point is the key result of [Jordan et al. \(1998\)](#) that the Langevin dynamics in space corresponds to the gradient flow of the relative entropy functional in the space of measures with the Wasserstein metric. This is why running the Langevin dynamics is useful for sampling: It is the steepest descent flow that attracts any initial distribution to the stationary target measure. Our motivating question is:

*Is there an **implementable** discretization of the Langevin dynamics that is consistent and converges exponentially fast under the logarithmic Sobolev inequality?*

Recall in general, gradient flow converges exponentially fast under a gradient-domination condition that is weaker than strong convexity. In the space of measures with the relative entropy functional, this gradient-domination condition is the logarithmic Sobolev inequality (LSI) ([Otto and Villani, 2000](#)). Therefore, in continuous time, the Langevin dynamics converges exponentially fast under LSI. In discrete time, the situation is less clear. A basic discretization known as the unadjusted Langevin algorithm (ULA) is *biased*, which means it converges to a limit different from the target measure. This bias exists for arbitrarily small (fixed) step size, even for a Gaussian target measure. This led to proposals to correct the bias, such as the Metropolis-Hastings correction step ([Roberts and Tweedie, 1996](#); [Dwivedi et al., 2018](#)) or decreasing the step size; however, the resulting algorithms become more complicated, so here we focus on ULA in order to see the basic structure.

Even with the bias, we can use ULA to get a one-time sampling algorithm as follows: given an error threshold, first choose a small enough step size so the bias is within the threshold, then run ULA with that step size to approximate convergence. See for example Dalalyan (2017a,b); Dalalyan and Karagulyan (2017); Durmus and Moulines (2016); Cheng and Bartlett (2018).

The bias in the Langevin algorithm is actually puzzling. The bias is typically attributed to the fact that ULA is a discretization of the continuous-time Langevin dynamics, so it necessarily has a discretization error. However, this attribution is misplaced, because the discretization error is the deviation between ULA and the Langevin dynamics at small time, while we are concerned with the asymptotic bias of ULA at large time. And indeed it is possible for a discretization algorithm to be consistent (unbiased). For example, gradient descent is a discretization of gradient flow; under the gradient-domination condition, both gradient flow and gradient descent converge to the minimizer exponentially fast. Since the Langevin dynamics is a gradient flow and it converges exponentially fast under LSI, we expect the gradient descent version to also converge exponentially fast under LSI, hence our motivating question. We note that the class of measures satisfying LSI is rather large and closed under bounded perturbation, so even multimodal distributions satisfy LSI.

Why is ULA biased? It cannot be the gradient descent discretization of the Langevin dynamics. Rather, it is performing the *Forward-Flow* (FFI) discretization, which in general is biased. Here we observe that the problem of minimizing relative entropy in the space of measures is a composite optimization problem, which means the objective function is a sum of two terms. Indeed, relative entropy can be written as the sum of negative entropy and the expected function value, where the function is the negative log density of the target measure. ULA (FFI) is a two-step algorithm: First, it applies the forward method to the expected function value, which is implemented by the usual gradient descent step in space. Second, it applies the exact gradient flow of the negative entropy, the heat flow, which is implemented by adding independent Gaussian noise in space. Therefore, ULA is *implementable*, which means we can run it in space with a given sample, but it is biased. We can compute the bias explicitly for the Ornstein-Uhlenbeck (OU) process, which is the case of Gaussian target measure; see Example 2 below.

For a composite optimization problem, the algorithm of choice is the *Forward-Backward* (FB) algorithm, which is a composition of the forward method for one term and the backward method for the other. The FB algorithm is consistent, which crucially uses the fact that the backward method is the adjoint of the forward method. Furthermore, FB converges at exponential rate under the gradient-domination condition and some smoothness assumptions. For our problem of optimizing relative entropy, FB means we want to run the forward method for the expected function value (which we can do via the usual gradient descent), and run the backward method for negative entropy (which we cannot do in general). Indeed, the main interesting difficulty is that for negative entropy, its exact gradient flow (the heat flow) is implementable via Gaussian noise, but neither its forward nor backward method is implementable.¹ However, in one case, namely for Gaussian initial data, we can solve the backward method for negative entropy. Therefore, for the OU process with Gaussian initial data, we can solve the FB algorithm and see it is indeed consistent; see Example 8.

Finally, if we cannot remove the bias, we can try to reduce it. We propose the *symmetrized Forward-Flow* (SFFI) algorithm, which is a composition of FFI and its adjoint. The SFFI algorithm is symmetric and has order 2; therefore, its bias is also of order 2, smaller than FFI. Applying SFFI to the Langevin dynamics yields the *symmetrized Langevin algorithm* (SLA), which is a composition

1. This is the opposite of what typically happens in optimization, where we cannot run gradient flow but we can run the gradient descent algorithm.

of ULA and its adjoint. SLA requires being able to run the backward method for the expected function value, which is implemented by the proximal gradient step for the function in space. This may require some numerical computation in each iteration, but the prize is a sampling algorithm that should have a smaller bias. As an example, we show how to solve SLA explicitly for the OU process. We see in this case SLA is in fact consistent; see Example 3. As another example, we show how to implement SLA for a mixture of two Gaussians as target measure, by solving an explicit one-dimensional calculation in each iteration. We show using synchronous coupling under strong log-concavity that SLA converges exponentially fast to its limiting measure.

2. Sampling as optimization in the space of measures

Sampling can be formulated as optimization in the space of measures. Indeed, to sample from a target distribution, it suffices to optimize an objective function in the space of measures that is minimized at the target distribution. Thus, we can translate methods from optimization to sampling by applying them to the optimization problem in the space of measures, provided the resulting methods are implementable as (possibly stochastic) algorithms in space.

In principle, we can choose any objective function that is minimized at the target distribution. But in the space of measures, there is a special function that works, which is the *relative entropy*:

$$H_\nu(\rho) = \int \rho \log \frac{\rho}{\nu}. \quad (1)$$

This is also known as the Kullback-Leibler (KL) divergence. Here ρ is a probability measure on \mathbb{R}^n absolutely continuous with respect to ν , and the integral above is a shorthand for $\int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\nu(x)} dx$. Relative entropy is nonnegative: $H_\nu(\rho) \geq 0$; and it is minimized at the target measure: $H_\nu(\rho) = 0$ if and only if $\rho = \nu$. Furthermore, ν is the only stationary point of H_ν , even when ν is multimodal. Therefore, if we can minimize H_ν , then we can sample from ν .

Let $\nu = e^{-f}$, or equivalently $f = -\log \nu$. Relative entropy decomposes as a sum of two terms:

$$H_\nu(\rho) = \mathbb{E}_\rho[f] - H(\rho) \quad (2)$$

where $\mathbb{E}_\rho[f] = \int \rho f$ is the expected value of f and $-H(\rho) = \int \rho \log \rho$ is the negative entropy of ρ .

2.1. The Langevin dynamics as the gradient flow of relative entropy

We wish to minimize the relative entropy functional (1) in the space of measures. A general strategy to minimize a function is to run the gradient flow dynamics. This requires a metric structure (Ambrosio et al., 2008). In the space of measures over \mathbb{R}^n there is a nice choice of metric, which is the Wasserstein metric induced by the quadratic distance function (Villani, 2003, 2008).

In the space of measures with the Wasserstein metric, the gradient flow of relative entropy (1) is the following partial differential equation, known as the *Fokker-Planck equation*:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho \nabla \log \frac{\rho}{\nu} \right) = \nabla \cdot (\rho \nabla f) + \Delta \rho. \quad (3)$$

Here $\rho = \rho(x, t)$ is a smooth positive density evolving over time. This is the key result of Jordan et al. (1998), which has been extended to vast generalities (Villani, 2008). So if we can follow the flow of the Fokker-Planck equation in the space of measures, then we converge to the target measure

$\nu = e^{-f}$. Furthermore, if ν satisfies the logarithmic Sobolev inequality (LSI), then the convergence is exponentially fast, see §3.1.1. However, can we implement this in space?

It turns out the Fokker-Planck equation is the continuity equation of the *Langevin dynamics*, which is the following stochastic differential equation in space:

$$dX = -\nabla f(X) dt + \sqrt{2} dW. \quad (4)$$

Here $X = (X_t)_{t \geq 0}$ is a stochastic process and $W = (W_t)_{t \geq 0}$ is the standard Brownian motion in \mathbb{R}^n . That is, if $X_t \sim \rho_t$ evolves following the Langevin dynamics (4) in space, then $\rho(x, t) = \rho_t(x)$ evolves following the Fokker-Planck equation (3) in the space of measures; see for example (Mackey, 1992, §11).

This means the Fokker-Planck flow (3) is implementable by the Langevin dynamics (4), as long as we can follow the stochastic process (4) exactly. In one case, namely for the Gaussian target measure, we know the exact solution; see Example 1 below. In general, we need to discretize to obtain an algorithm in space; we discuss this further below.

Example 1 (Ornstein-Uhlenbeck.) Let $\nu = \mathcal{N}(\mu, \Sigma)$ be the Gaussian target measure with mean $\mu \in \mathbb{R}^n$ and covariance $\Sigma \succ 0$, so $f(x) = \frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu) + \frac{1}{2} \log \det(2\pi\Sigma)$ is quadratic. The Langevin dynamics (4) has a linear drift, and is known as the Ornstein-Uhlenbeck (OU) process:

$$dX = -\Sigma^{-1}(X - \mu) dt + \sqrt{2} dW.$$

This has an exact solution as an Itô integral. In particular, the solution at each time $t \geq 0$ satisfies

$$X_t - \mu \stackrel{d}{=} e^{-t\Sigma^{-1}}(X_0 - \mu) + \Sigma^{\frac{1}{2}} \left(I - e^{-2t\Sigma^{-1}} \right)^{\frac{1}{2}} Z$$

where $Z \sim \mathcal{N}(0, I)$ is independent of X_0 . Thus, $X_t \sim \rho_t$ converges to $\mathcal{N}(\mu, \Sigma)$ exponentially fast.

2.2. The unadjusted Langevin algorithm

A practical discretization of the Langevin dynamics (4) is the *unadjusted Langevin algorithm* (ULA):

$$x_{k+1} = x_k - \epsilon \nabla f(x_k) + \sqrt{2\epsilon} z_k \quad (5)$$

where $z_k \sim \mathcal{N}(0, I)$ is independent of x_k . Here $\epsilon > 0$ is a step size, which is equal to the time step in the discretization.

As $\epsilon \rightarrow 0$, the ULA iteration (5) converges to the Langevin dynamics (4) (Mackey, 1992, §11). However, when $\epsilon > 0$, ULA is biased, which means it does not converge to the target distribution $\nu = e^{-f}$. This bias is present even for Gaussian target measure, see Example 2.

Example 2 (ULA for OU.) Let $\nu = \mathcal{N}(\mu, \Sigma)$. ULA is $x_{k+1} - \mu = (I - \epsilon\Sigma^{-1})(x_k - \mu) + \sqrt{2\epsilon}z_k$. Unfolding and using the fact that the sum of independent Gaussians is Gaussian, we can write

$$x_k - \mu \stackrel{d}{=} A_\epsilon^k (x_0 - \mu) + \sqrt{2\epsilon} (I - A_\epsilon^2)^{-\frac{1}{2}} (I - A_\epsilon^{2k})^{\frac{1}{2}} z$$

where $A_\epsilon = I - \epsilon\Sigma^{-1}$ and $z \sim \mathcal{N}(0, I)$ is independent of x_0 . For $0 < \epsilon < 2\lambda_{\min}(\Sigma)$, $\lim_{k \rightarrow \infty} A_\epsilon^k = 0$.

Therefore, $x_k \xrightarrow{d} \mu + \sqrt{2\epsilon} (I - A_\epsilon^2)^{-\frac{1}{2}} z$. Thus, ULA for OU has the limit measure

$$\nu_\epsilon = \mathcal{N}\left(\mu, \Sigma(I - \frac{\epsilon}{2}\Sigma^{-1})^{-1}\right).$$

At $\epsilon = 0$, this is the target measure $\nu = \mathcal{N}(\mu, \Sigma)$. For $\epsilon > 0$, $\nu_\epsilon \neq \nu$, so ULA is biased. The bias is

$$W_2(\nu, \nu_\epsilon) = \|\Sigma^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}(I - \frac{\epsilon}{2}\Sigma^{-1})^{-\frac{1}{2}}\|_{\text{HS}} = \frac{\epsilon}{4}\sqrt{\text{Tr}(\Sigma^{-1})} + O(\epsilon^2).$$

where $\|B\|_{\text{HS}} = \sqrt{\text{Tr}(B^2)}$ is the Hilbert-Schmidt norm of a symmetric matrix B .

Above, we have used the formula for the Wasserstein distance between Gaussians (Takatsu, 2011).

2.2.1. CONVERGENCE TO THE BIASED LIMIT

We recall the following contraction result using synchronous coupling under strong log-concavity; see (Dalalyan, 2017b, Lemma 1) or Appendix A.1. Note the similarity with the corresponding result in strongly convex optimization, e.g., (Nesterov, 2004, Theorem 2.1.15).

Lemma 1 *Suppose ν is α -strongly log-concave and L -log-smooth ($\alpha I \preceq -\nabla^2 \log \nu \preceq LI$) for some $0 < \alpha \leq L$. Let ρ_k, ρ'_k be any two distributions evolving following the ULA algorithm (5). Then for $0 < \epsilon \leq \frac{2}{\alpha+L}$, $W_2(\rho_k, \rho'_k)^2 \leq (1 - 2\epsilon\frac{\alpha L}{\alpha+L})^k W_2(\rho_0, \rho'_0)^2$.*

The above implies that ULA has a unique stationary measure ν_ϵ . The bias should be of order 1 since ULA is a first-order discretization method, as suggested by Example 2. Using a simple synchronous coupling argument with the same smoothness assumption as in Lemma 1 only yields a bias of order $\frac{1}{2}$ (Dalalyan, 2017b, Theorem 1). But with an additional smoothness assumption, an expansion within the synchronous coupling argument yields a bias of order 1 (Dalalyan and Karagulyan, 2017, Theorem 4). It is interesting to ask whether it is possible to prove a bias of order 1 with the minimal smoothness assumption as in Lemma 1.

Lemma 2 *Suppose ν is α -strongly log-concave and L -log-smooth, and $-\nabla^2 \log \nu$ is M -Lipschitz. For $0 < \epsilon < \frac{2}{\alpha+L}$, the bias of ULA is $W_2(\nu_\epsilon, \nu) \leq \frac{\epsilon}{\alpha}(\frac{1}{2}Mn + \frac{11}{5}\sqrt{L^3n})$.*

The bias from Lemma 2 and the exponential contraction from Lemma 1 imply an $O(\frac{1}{\delta} \log \frac{1}{\delta})$ iteration complexity bound for sampling from ν up to Wasserstein error $O(\delta)$, by choosing $\epsilon = O(\delta)$ and running ULA for $k = \Omega(\frac{1}{\epsilon} \log \frac{1}{\delta})$ iterations to get $W_2(\rho_k, \nu) \leq W_2(\rho_k, \nu_\epsilon) + W_2(\nu_\epsilon, \nu) = O(\epsilon)$. On the other hand, for the ideal unbiased discretization of the Langevin dynamics, exponential contraction would imply a logarithmic iteration complexity bound.

2.2.2. ULA AS THE FORWARD-FLOW DISCRETIZATION OF LANGEVIN DYNAMICS

ULA is biased because it is the Forward-Flow (FFI) discretization of the Langevin dynamics (4). Concretely, we can write ULA (5) as a composition of two operations:

$$x_{k+\frac{1}{2}} = x_k - \epsilon \nabla f(x_k) \tag{6a}$$

$$x_{k+1} = x_{k+\frac{1}{2}} + \sqrt{2\epsilon} z_k. \tag{6b}$$

The first step (6a) is a gradient descent step or the forward method for f ; the second step (6b) is the exact solution for the heat flow. In the space of measures, the iterations in (6) correspond to

$$\rho_{k+\frac{1}{2}} = (I - \epsilon \nabla f)_\# \rho_k \tag{7a}$$

$$\rho_{k+1} = \mathcal{N}(0, 2\epsilon I) * \rho_{k+\frac{1}{2}} \tag{7b}$$

where $\#$ is the pushforward operator and $*$ is the convolution. When f is smooth ($\nabla^2 f \preceq \frac{1}{\epsilon}I$), the first step (7a) is the gradient descent for the expected function value $\mathbb{E}_\rho[f]$ (see §E.1.3). The second step (7b) is the exact gradient flow for negative entropy $-H(\rho)$ (see §E.2.1). These are the two components of the decomposition (2) of relative entropy. Therefore, ULA—which in the space of measures takes the form (7)—is the Forward-Flow method applied to the composite optimization problem of minimizing relative entropy (2). The source of the bias is that the flow method is not the adjoint of the forward method, so the Forward-Flow method does not conserve the stationary point of the overall flow, which is the target measure ν . See §F for a review on composite optimization.

In a recent work, [Bernton \(2018\)](#) proposes the proximal version of ULA, obtained by replacing the forward method (gradient descent) in (7a) with the backward method (proximal gradient) for $\mathbb{E}_\rho[f]$, which is also implemented by the proximal gradient for f in space. This is the Backward-Flow discretization of the Langevin dynamics, which has similar convergence guarantees as ULA, but is also still biased.

2.3. The Forward-Backward method for Langevin dynamics

A general algorithm for a composite optimization problem is the Forward-Backward (FB) method,² which means running the forward method (gradient descent) for one component and running the backward method (proximal gradient) for the other. The FB algorithm is consistent (unbiased) because the backward method is the adjoint of the forward method, so the minimizer is conserved; see §F.1. Furthermore, FB converges exponentially fast under the gradient-domination condition and some smoothness assumptions ([Garrigos et al., 2017](#)), which is formally applicable in our case. Therefore, we wish to run the FB method for composite optimization.

However, for our problem of optimizing relative entropy, FB means we need to replace (7b) with the backward method for negative entropy in the space of measures. This is well-defined, but cannot be solved explicitly—unlike the exact flow (7b)—except for Gaussian initial data ([Carlen and Gangbo, 2003](#)). In general this exception does not help, since Gaussianity is not preserved under the forward step (7a) when ∇f is nonlinear. However, for the OU process, ∇f is linear, so a Gaussian initial data stays Gaussian under the FB algorithm, and in this case we see FB is indeed consistent; see Example 8 in §4.1.

2.4. The symmetrized Langevin algorithm

Now that we know the unbiased FB algorithm for the Langevin dynamics is not implementable in general, we can try to reduce the bias of what we can implement.

Recall that the bias of an optimization algorithm is of the same size as its discretization order (see §B for a review of discretization methods). Recall also that if an algorithm is symmetric, which means it is equal to its adjoint, then its order is even; in particular, it must be of order at least 2. In general we can symmetrize an algorithm by composing it with its adjoint. Therefore, given any algorithm with a first-order bias, we can upgrade it to an algorithm with a second-order bias by symmetrizing it.

Applying this idea to FFI as the base algorithm, we obtain the *symmetrized Forward-Flow* (SFFI) algorithm, which is the composition of FFI and its adjoint, the Flow-Backward (FIB) algorithm.

2. Or its adjoint, the Backward-Forward (BF) method. However, the forward method for the heat flow is also not implementable, except for Gaussian data. We can also run the forward method (gradient descent) or backward method (proximal gradient) for OU with Gaussian data, and see they are consistent; see Examples 4 and 5.

Applying SFFI to the Langevin dynamics yields the *symmetrized Langevin algorithm* (SLA):

$$x_{k+1} = (I + \epsilon \nabla f)^{-1}(x_k - \epsilon \nabla f(x_k) + \sqrt{4\epsilon} z_k) \quad (8)$$

where $z_k \sim \mathcal{N}(0, I)$ is independent of x_k . Here $(I + \epsilon \nabla f)^{-1}$ is the proximal gradient operator of f , i.e., $y = (I + \epsilon \nabla f)^{-1}(x)$ if and only if $y + \epsilon \nabla f(y) = x$, or $y = \arg \min_{y' \in \mathbb{R}^n} \{f(y') + \frac{1}{2\epsilon} \|y' - x\|^2\}$. This is not analytically solvable for general f , and may require numerical computation in each iteration. However, if we can do this, then we can run SLA which in principle has a smaller bias. When the target measure is Gaussian, namely for the OU process, we can write the SLA iteration explicitly. In this case we see that SLA is in fact consistent,³ and converges exponentially fast.

Example 3 (SLA for OU.) Let $\nu = \mathcal{N}(\mu, \Sigma)$. The SLA iteration is

$$x_{k+1} - \mu = (I + \epsilon \Sigma^{-1})^{-1}(I - \epsilon \Sigma^{-1})(x_k - \mu) + \sqrt{4\epsilon}(I + \epsilon \Sigma^{-1})^{-1} z_k.$$

Unfolding and using the fact that the sum of independent Gaussians is Gaussian, we can write

$$x_k - \mu \stackrel{d}{=} A_\epsilon^k(x_0 - \mu) + \sqrt{4\epsilon} B_\epsilon (I - A_\epsilon^2)^{-\frac{1}{2}} (I - A_\epsilon^{2k})^{\frac{1}{2}} z$$

where $A_\epsilon = (I + \epsilon \Sigma^{-1})^{-1}(I - \epsilon \Sigma^{-1})$, $B_\epsilon = (I + \epsilon \Sigma^{-1})^{-1}$, and $z \sim \mathcal{N}(0, I)$ is independent of x_0 . For all $\epsilon > 0$, $\lim_{k \rightarrow \infty} A_\epsilon^k = 0$. Therefore, $x_k \xrightarrow{d} \mu + \sqrt{4\epsilon} B_\epsilon (I - A_\epsilon^2)^{-\frac{1}{2}} z \sim \mathcal{N}(\mu, \Sigma)$. This shows SLA converges to the correct target measure $\nu = \mathcal{N}(\mu, \Sigma)$.

We also note that there are other discretizations of the Langevin dynamics that are unbiased for the Gaussian target measure, for example the Ozaki discretization which uses Hessian information; see for example (Dalalyan, 2017a). We can also implement SLA for a mixture of two Gaussians, which requires solving a one-dimensional numerical problem in each iteration; see §A.4.

2.4.1. CONVERGENCE TO THE BIASED LIMIT

Similar to ULA, we have the following contraction result for SLA under strong log-concavity; see Appendix A.3.

Lemma 3 Suppose ν is α -strongly log-concave and L -log-smooth ($\alpha I \preceq -\nabla^2 \log \nu \preceq LI$) for some $0 < \alpha \leq L$. Let ρ_k, ρ'_k be any two distributions evolving following the SLA algorithm (8). Then for $0 < \epsilon \leq \frac{2}{\alpha+L}$,

$$W_2(\rho_k, \rho'_k)^2 \leq \left(\frac{1 - 2\epsilon \frac{\alpha L}{\alpha+L}}{1 + 2\epsilon \frac{\alpha L}{\alpha+L}} \right)^k W_2(\rho_0, \rho'_0)^2.$$

The above implies that SLA has a unique stationary measure $\tilde{\nu}_\epsilon$. Despite Example 3, in general SLA is biased, and the bias should be of order 2 since SLA is a second-order discretization method. However, using the expansion within synchronous coupling as in ULA seems to still yield a bias of order 1 due to the stochastic terms.

3. This is surprising because SFFI is biased even for minimizing a sum of two quadratic functions in space. Here the consistency of SLA for OU relies on the property that variance adds linearly when we sum independent Gaussians.

Question 1 *Is it true that if ν is strongly log-concave, then $W_2(\nu, \tilde{\nu}_\epsilon) = O(\epsilon^2)$?*

If Question 1 is true, then combined with Lemma 3, it implies an $O(\frac{1}{\sqrt{\delta}} \log \frac{1}{\delta})$ iteration complexity bound for sampling from ν up to Wasserstein error $O(\delta)$, by choosing $\epsilon = O(\sqrt{\delta})$ and running SLA for $k = O(\frac{1}{\epsilon} \log \frac{1}{\delta})$ iterations to get $W_2(\rho_k, \nu) \leq W_2(\rho_k, \tilde{\nu}_\epsilon) + W_2(\tilde{\nu}_\epsilon, \nu) = O(\epsilon^2) = O(\delta)$.

We also note that a bias of order 2 may be the best we can hope for, since in general any higher order necessitates running the algorithm with negative step size (Hairer et al., 2006, Theorem 3.18). For the Langevin dynamics, this means running the heat flow backward in time, which is not only non-implementable, but also not well-posed mathematically (except in special cases, e.g., see Mikranker (1961)). But as entropy is a very special functional, it is possible this apparent difficulty may be circumvented and implementable algorithms with higher-order bias may be found.

3. Optimization in the space of measures

Let us discuss further how to optimize in the space of measures. We review optimization in a smooth Riemannian manifold in §C, including gradient flow and the forward and backward methods, and conditions for exponential convergence such as gradient domination and strong convexity. We also briefly recall the Wasserstein metric; see §D for a review or Villani (2008) for more detail.

Let $\mathcal{P} \equiv \mathcal{P}_2(\mathbb{R}^n)$ denote the space of probability measures on \mathbb{R}^n with finite second moments, endowed with the Wasserstein metric W_2 induced by the quadratic distance. Every element $\rho \in \mathcal{P}$ is a probability measure represented by its density function $\rho: \mathbb{R}^n \rightarrow \mathbb{R}$ with respect to the Lebesgue measure dx . The tangent space $T_\rho \mathcal{P}$ consists of functions $R: \mathbb{R}^n \rightarrow \mathbb{R}$ of the form $R = -\nabla \cdot (\rho \nabla \phi)$ for some $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$; we write $R \equiv \nabla \phi$. The norm of $R \equiv \nabla \phi$ is $\|R\|_\rho = (\mathbb{E}_\rho[\|\nabla \phi\|^2])^{\frac{1}{2}}$.

Suppose we wish to solve the optimization problem

$$\min_{\rho \in \mathcal{P}} F(\rho)$$

where $F: \mathcal{P} \rightarrow \mathbb{R}$ is a smooth functional. There are two basic classes of interesting functionals:

1. **Expected value:** Suppose $F(\rho) = \mathbb{E}_\rho[f]$ is the expected value of a smooth function $f: \mathbb{R}^n \rightarrow \mathbb{R}$. This is an example of a “potential energy” (Villani, 2003, §5.2.2). It involves only a scalar product of a density ρ and a function f , so can be implemented by samples in space. The gradient flow, gradient descent, and proximal gradient methods for $\mathbb{E}_\rho[f]$ are implemented by the corresponding gradient flow, gradient descent, and proximal gradient methods for f , with the same convergence guarantees. Therefore, we can view a deterministic problem from the space of measures without loss of information. See §E.1 for more detail.
2. **Negative entropy:** Suppose $F(\rho) = -H(\rho)$ is the negative entropy. This is an example of an “internal energy” where we apply a function to the density ρ before integrating, so it apparently cannot be implemented from samples in space. Interestingly, the gradient flow of negative entropy is the heat flow, which is implementable by the Brownian motion (Gaussian noise) in space. This also gives an optimization interpretation of the Fisher information as the squared gradient of entropy. See §E.2 for more detail.

There is a third class of functionals which is “interaction energy”, for example variance. We do not use interaction energy for sampling in this paper, but see §E.3 for the gradient flow of variance.

3.1. Minimizing relative entropy

Our objective function for sampling is the *relative entropy*, which is a combination of the potential and internal energies:

$$F(\rho) = H_\nu(\rho) = \int \rho \log \frac{\rho}{\nu} = \mathbb{E}_\rho[f] - H(\rho) \quad (9)$$

where $\nu = e^{-f}$ is the target measure. Relative entropy is nonnegative, $H_\nu(\rho) \geq 0$, and it is minimized at the target measure: $H_\nu(\rho) = 0$ if and only if $\rho = \nu$.

3.1.1. LOG-SOBOLEV INEQUALITY AS GRADIENT DOMINATION OF RELATIVE ENTROPY

The squared gradient of relative entropy $F(\rho) = H_\nu(\rho)$ is the *relative Fisher information*:

$$J_\nu(\rho) = \mathbb{E}_\rho \left[\left\| \nabla \log \frac{\rho}{\nu} \right\|^2 \right].$$

Note that $\rho = \nu$ is the only stationary point of H_ν ; because if ρ is a stationary point of H_ν , then $\| \text{grad}_\rho H_\nu \|_\rho^2 = J_\nu(\rho) = 0$, and it is clear that $J_\nu(\rho) = 0$ if and only if $\rho = \nu$.

The gradient domination condition $\| \text{grad} F \|^2 \geq 2\alpha(F - \min F)$, $\alpha > 0$, for relative entropy $F = H_\nu$ becomes the *logarithmic Sobolev inequality* (LSI) (Gross, 1975; Otto and Villani, 2000):

$$J_\nu(\rho) \geq 2\alpha H_\nu(\rho) \quad \forall \rho \in \mathcal{P}.$$

The gradient flow identity $\frac{d}{dt} F(\rho) = -\| \text{grad}_\rho F \|^2_\rho$ becomes $\frac{d}{dt} H_\nu(\rho) = -J_\nu(\rho)$, which is a generalization of the De Bruijn's identity.

The set of measures satisfying LSI includes all strongly log-concave measures, and it is closed under bounded perturbation with a constant that decays exponentially with the size of the perturbation (Holley and Stroock, 1987). Thus, even a multimodal distribution such as a mixture of Gaussian satisfies LSI, and hence the Langevin dynamics converges exponentially fast.

The Hessian of relative entropy $F(\rho) = -H_\nu(\rho)$ is, for a tangent function $R \equiv \nabla \phi \in \mathbb{T}_\rho \mathcal{P}$,

$$(\text{Hess}_\rho H_\nu)(R, R) = \mathbb{E}_\rho \left[\|\nabla^2 \phi\|_{\text{HS}}^2 + \langle \nabla \phi, (\nabla^2 f) \nabla \phi \rangle \right].$$

Therefore, if f is strongly convex (ν is strongly log-concave), then $F = H_\nu$ is also strongly convex, in which case any two co-evolving solutions are contracting exponentially fast. However, note that the Hessian of H_ν is not bounded above.

3.1.2. LANGEVIN DYNAMICS AS GRADIENT FLOW OF RELATIVE ENTROPY

The gradient of $F(\rho) = H_\nu(\rho)$ is $\text{grad}_\rho F = -\nabla \cdot (\rho \nabla \log \frac{\rho}{\nu}) \equiv \nabla \log \frac{\rho}{\nu}$. Therefore, the gradient flow equation $\dot{\rho} = -\text{grad}_\rho F$ of relative entropy is the Fokker-Planck equation (3):

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho \nabla \log \frac{\rho}{\nu} \right) = \nabla \cdot (\rho \nabla f) + \Delta \rho. \quad (10)$$

This is implementable in space as the Langevin dynamics stochastic differential equation (4):

$$dX = -\nabla f(X) dt + \sqrt{2} dW.$$

When the target measure is Gaussian, namely for the Ornstein-Uhlenbeck (OU) process, we have an exact solution, as we have seen in Example 1. However, in general we need to discretize.

If ν satisfies α -LSI, then along the gradient flow (10),

$$\frac{d}{dt}H_\nu(\rho) = -J_\nu(\rho) \leq -2\alpha H_\nu(\rho),$$

which implies exponential convergence in relative entropy:

$$H_\nu(\rho_t) \leq e^{-2\alpha t} H_\nu(\rho_0).$$

This also implies exponential convergence of the distance $W_2(\rho_t, \nu)$, since LSI implies the *Talagrand inequality*: $W_2(\rho, \nu)^2 \leq \frac{2}{\alpha} H_\nu(\rho), \forall \rho \in \mathcal{P}$ (Talagrand, 1996; Otto and Villani, 2000). This is because in general gradient domination implies the sufficient growth property, see Appendix C.1.

3.1.3. FORWARD METHOD FOR LANGEVIN DYNAMICS

The forward method $\rho_{k+1} = \exp_{\rho_k}(-\epsilon \operatorname{grad}_{\rho_k} F)$ for relative entropy $F = H_\nu$ is

$$\rho_{k+1} = \exp_{\rho_k} \left(-\epsilon \nabla \log \frac{\rho_k}{\nu} \right)$$

where recall $\operatorname{grad}_\rho F \equiv \nabla \log \frac{\rho}{\nu}$. If ρ_k is K -log-semiconcave relative to ν , which means $-\nabla^2 \log \frac{\rho_k}{\nu} \succeq KI$ for some $K \in \mathbb{R}$, then for $\epsilon \leq \frac{1}{\max\{0, -K\}}$, the exponential map above is given by:

$$\rho_{k+1} = \left(I - \epsilon \nabla \log \frac{\rho_k}{\nu} \right)_{\#} \rho_k. \quad (11)$$

If we know the analytic form of ρ_k , then we can implement one step of the algorithm by $x_{k+1} = x_k - \epsilon \nabla f(x_k) - \epsilon \nabla \log \rho_k(x_k)$. However, we cannot iterate this algorithm because at the next round we do not know what ρ_{k+1} is, we only have x_{k+1} .

For Gaussian target measure, we can solve the forward method with Gaussian initial data.

Example 4 (Forward method for OU with Gaussian data.) Let $\nu = \mathcal{N}(\mu, \Sigma)$ as in Example 1. Let $\rho_0 = \mathcal{N}(\mu, \Sigma_0)$ with $\Sigma_0^{-1} \succeq \Sigma^{-1}$. Along the forward method (11) for OU, $\rho_k = \mathcal{N}(\mu, \Sigma_k)$ stays Gaussian. Further, (11) becomes $x_{k+1} - \mu = (I + \epsilon(\Sigma_k^{-1} - \Sigma^{-1}))(x_k - \mu)$. Therefore, $\Sigma_{k+1} = \Sigma_k(I + \epsilon(\Sigma_k^{-1} - \Sigma^{-1}))^2$. The only fixed point is $\Sigma_k = \Sigma_{k+1} = \Sigma$. Thus, the forward method is consistent for OU with Gaussian data.

3.1.4. BACKWARD METHOD FOR LANGEVIN DYNAMICS

The backward method $\rho_{k+1} = \arg \min_{\rho \in \mathcal{P}} \{F(\rho) + \frac{1}{2\epsilon} W_2(\rho, \rho_k)^2\}$ for relative entropy $F = H_\nu$ is

$$\exp_{\rho_{k+1}} \left(\epsilon \nabla \log \frac{\rho_{k+1}}{\nu} \right) = \rho_k.$$

If ρ_{k+1} is L -log-smooth with respect to ν , which means $-\nabla^2 \log \frac{\rho_{k+1}}{\nu} \preceq LI$ for some $L > 0$, then for $\epsilon \leq \frac{1}{L}$, the backward method above is implemented (implicitly) by:

$$\left(I + \epsilon \nabla \log \frac{\rho_{k+1}}{\nu} \right)_{\#} \rho_{k+1} = \rho_k. \quad (12)$$

In general this is not solvable analytically.

For Gaussian target measure, we can solve the backward method with Gaussian initial data.

Example 5 (Backward method for OU with Gaussian data.) Let $\nu = \mathcal{N}(\mu, \Sigma)$ as in Example 1. Let $\rho_0 = \mathcal{N}(\mu, \Sigma_0)$ for simplicity. Along the backward method (12) for OU, $\rho_k = \mathcal{N}(\mu, \Sigma_k)$ stays Gaussian. Further, (12) becomes $(I - \epsilon(\Sigma_{k+1}^{-1} - \Sigma^{-1}))(x_{k+1} - \mu) = x_k - \mu$. Therefore, $\Sigma_{k+1}(I - \epsilon(\Sigma_{k+1}^{-1} - \Sigma^{-1}))^2 = \Sigma_k$. The only fixed point is $\Sigma_k = \Sigma_{k+1} = \Sigma$. Thus, the backward method is consistent for OU with Gaussian data.

4. Langevin dynamics as composite optimization in the space of measures

In §3.1 we have seen how to optimize $H_\nu(\rho)$ by considering it as a single function. We now study how to optimize $H_\nu(\rho)$ as a composite function when we write it as a sum of two functions:

$$H_\nu(\rho) = \mathbb{E}_\rho[f] - H(\rho).$$

4.1. Forward-Backward for Langevin dynamics

In general, the algorithm of choice for composite optimization is the Forward-Backward (FB) algorithm, which means we run the forward method for one component and the backward method for the other. (Equivalently, we can run the Backward-Forward algorithm, which is the adjoint version.) The FB algorithm is consistent because the backward method is adjoint to the forward method, so the FB algorithm preserves the stationary point; see §F for a review. FB can be shown to converge exponentially fast under gradient domination condition and some smoothness assumptions (Garri-gos et al., 2017); see also §F.2. In principle, the FB algorithm for the Langevin dynamics is the answer we are seeking.

For optimizing relative entropy, the FB algorithm means running the forward method for $\mathbb{E}_\rho[f]$ (which is implemented by the gradient descent for f), followed by the backward method for $-H(\rho)$:

$$\rho_{k+\frac{1}{2}} = (I - \epsilon \nabla f) \# \rho_k \tag{13a}$$

$$\rho_{k+1} = \arg \min_{\rho \in \mathcal{P}} \left\{ -H(\rho) + \frac{1}{2\epsilon} W_2(\rho, \rho_{k+\frac{1}{2}})^2 \right\}. \tag{13b}$$

We cannot implement the backward method for the heat flow in (13b), so the FB algorithm above is not implementable in general. For Gaussian target measure, we can solve FB for Gaussian initial data. In this case the distributions stay Gaussian, and FB is indeed consistent; see Example 8 in §G.1.

4.2. Backward-Forward for Langevin dynamics

Similarly, we can run the Backward-Forward (BF) algorithm for the Langevin dynamics.

For optimizing relative entropy, BF means running the backward method for $\mathbb{E}_\rho[f]$ (which is implemented by the proximal gradient step for f), followed by the forward method for $-H(\rho)$:

$$\rho_{k+\frac{1}{2}} = ((I + \epsilon \nabla f)^{-1}) \# \rho_k \tag{14a}$$

$$\rho_{k+1} = \exp_{\rho_{k+\frac{1}{2}}}(-\epsilon \nabla \log \rho_{k+\frac{1}{2}}). \tag{14b}$$

The BF algorithm above is not implementable in general, since we cannot implement the forward method for the heat flow beyond one step. For Gaussian target measure, we can solve BF for Gaussian initial data. In this case the distributions stay Gaussian, and BF is indeed consistent; see Example 9 in §G.2.

5. Discussion and future work

In this paper we have studied sampling as optimization in the space of measures. We started with the question of whether we can have a consistent discretization of the Langevin dynamics that converges exponentially fast under LSI. We have seen that the difficulties are twofold: First, relative entropy is a composite optimization problem in the space of measures, so we have to work with composite algorithms. Second, the heat flow is exactly solvable, but neither its forward nor backward methods are implementable. Therefore, unbiased algorithms such as the FB algorithm are not implementable for the Langevin dynamics. The basic discretization known as ULA is implementable but biased. We also proposed a symmetrized variant of ULA which should have a smaller bias, at the price of implementing the proximal gradient step in space.

We have focused on the Langevin dynamics, which is the gradient flow dynamics for minimizing relative entropy. More generally, we can apply more sophisticated optimization techniques, such as acceleration, to sampling. There is a second-order variant of the Langevin dynamics known as the underdamped Langevin dynamics, which is the stochastic version of the second-order heavy ball dynamics for optimization, and has been shown to have better convergence properties than the Langevin dynamics (Cheng et al., 2018). However, it is interesting to consider whether we can also apply the acceleration principle directly in the space of measures, for example via the variational Lagrangian approach (Wibisono et al., 2016).

References

- Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
- Espen Bernton. Langevin Monte Carlo and JKO splitting. In *Proceedings of the 2018 Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, Stockholm, Sweden, 06–09 Jul 2018. PMLR.
- Eric A. Carlen and Wilfrid Gangbo. Constrained steepest descent in the 2-Wasserstein metric. *Annals of Mathematics*, pages 807–846, 2003.
- Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR, 07–09 Apr 2018.
- Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 2018 Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, Stockholm, Sweden, 06–09 Jul 2018. PMLR.
- Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017a.
- Arnak S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the 2017 Conference on Learning Theory*,

- volume 65 of *Proceedings of Machine Learning Research*, pages 678–689, Amsterdam, Netherlands, 07–10 Jul 2017b. PMLR.
- Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095v2*, 2017.
- Glaydston de Carvalho Bento, João Xavier da Cruz Neto, and Paulo Roberto Oliveira. A new approach to the proximal point method: Convergence on general Riemannian manifolds. *Journal of Optimization Theory and Applications*, 168(3):743–755, 2016.
- Amir Dembo, Thomas M. Cover, and Joy A. Thomas. Information theoretic inequalities. *IEEE Transactions on Information Theory*, 37(6):1501–1518, 1991.
- Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559v2*, 2016.
- Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proceedings of the 2018 Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, Stockholm, Sweden, 06–09 Jul 2018. PMLR.
- O.P. Ferreira. Proximal subgradient and a characterization of Lipschitz function on Riemannian manifolds. *Journal of Mathematical Analysis and Applications*, 313(2):587–597, 2006.
- O.P. Ferreira and P.R. Oliveira. Subgradient algorithm on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 97(1):93–104, 1998.
- O.P. Ferreira and P.R. Oliveira. Proximal point algorithm on Riemannian manifolds. *Optimization*, 51(2):257–270, 2002.
- Guillaume Garrigos, Lorenzo Rosasco, and Silvia Villa. Convergence of the forward-backward algorithm: Beyond the worst case with the help of geometry. *arXiv preprint arXiv:1703.09477v3*, 2017.
- Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.
- Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, second edition, 2006.
- Richard Holley and Daniel Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5):1159–1194, 1987.
- Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.
- Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

- Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.
- Michael C. Mackey. *Time’s Arrow: The Origins of Thermodynamics Behavior*. Springer-Verlag, 1992.
- Willard L. Miranker. A well posed problem for the backward heat equation. *Proceedings of the American Mathematical Society*, 12(2):243–247, 1961.
- Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Kluwer, Boston, 2004.
- Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.
- Boris T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel’noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.
- Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.
- A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959.
- Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.
- Michel Talagrand. Transportation cost for Gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.
- Cédric Villani. A short proof of the ”concavity of entropy power”. *IEEE Transactions on Information Theory*, 46(4):1695–1696, 2000.
- Cédric Villani. *Topics in optimal transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Society, 2003.
- Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2008.
- Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.
- Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Appendix A. Details for §2

A.1. Proof of Lemma 1 (Contraction of ULA)

We use synchronous coupling to show iteratively that

$$W_2(\rho_{k+1}, \rho'_{k+1})^2 \leq \left(1 - \frac{2\epsilon\alpha L}{\alpha + L}\right) W_2(\rho_k, \rho'_k)^2$$

which will imply the desired claim.

Let $x_k \sim \rho_k$, $x'_k \sim \rho'_k$ be coupled with the optimal coupling, so $W_2(\rho_k, \rho'_k) = \mathbb{E}[\|x_k - x'_k\|^2]$. We evolve x_k, x'_k via ULA (5) with the same Gaussian noise z_k (this is the synchronous coupling):

$$\begin{aligned} x_{k+1} &= x_k - \epsilon \nabla f(x_k) + \sqrt{2\epsilon} z_k \\ x'_{k+1} &= x'_k - \epsilon \nabla f(x'_k) + \sqrt{2\epsilon} z_k. \end{aligned}$$

Subtracting and taking the squared norm, we obtain

$$\begin{aligned} \|x_{k+1} - x'_{k+1}\|^2 &= \|x_k - x'_k - \epsilon(\nabla f(x_k) - \nabla f(x'_k))\|^2 \\ &= \|x_k - x'_k\|^2 - 2\epsilon \langle \nabla f(x_k) - \nabla f(x'_k), x_k - x'_k \rangle + \epsilon^2 \|\nabla f(x_k) - \nabla f(x'_k)\|^2. \end{aligned}$$

Since $f = -\log \nu$ is α -strongly convex and L -smooth, we have by (Nesterov, 2004, Theorem 2.1.12):

$$\langle \nabla f(x_k) - \nabla f(x'_k), x_k - x'_k \rangle \geq \frac{\alpha L}{\alpha + L} \|x_k - x'_k\|^2 + \frac{1}{\alpha + L} \|\nabla f(x_k) - \nabla f(x'_k)\|^2. \quad (15)$$

Therefore, we have the bound

$$\|x_{k+1} - x'_{k+1}\|^2 \leq \left(1 - \frac{2\epsilon\alpha L}{\alpha + L}\right) \|x_k - x'_k\|^2 + \epsilon \left(\epsilon - \frac{2}{\alpha + L}\right) \|\nabla f(x_k) - \nabla f(x'_k)\|^2.$$

If $0 < \epsilon \leq \frac{2}{\alpha + L}$, then the second term above is nonpositive, so we may drop it:

$$\|x_{k+1} - x'_{k+1}\|^2 \leq \left(1 - \frac{2\epsilon\alpha L}{\alpha + L}\right) \|x_k - x'_k\|^2.$$

Now we take expectation and use the fact that x_k, x'_k have the optimal coupling:

$$\mathbb{E}[\|x_{k+1} - x'_{k+1}\|^2] \leq \left(1 - \frac{2\epsilon\alpha L}{\alpha + L}\right) \mathbb{E}[\|x_k - x'_k\|^2] = \left(1 - \frac{2\epsilon\alpha L}{\alpha + L}\right) W_2(\rho_k, \rho'_k)^2.$$

Finally, by the definition of Wasserstein distance as the infimum over all coupling, we conclude

$$W_2(\rho_{k+1}, \rho'_{k+1})^2 \leq \left(1 - \frac{2\epsilon\alpha L}{\alpha + L}\right) W_2(\rho_k, \rho'_k)^2$$

as desired. \square

A.2. Proof of Lemma 2 (Bias of ULA)

This follows from the noiseless case ($\delta = \sigma = 0$) of (Dalalyan and Karagulyan, 2017, Theorem 4).

A.3. Proof of Lemma 3 (Contraction of SLA)

We follow the same outline as §A.1. We use synchronous coupling to show iteratively that

$$W_2(\rho_{k+1}, \rho'_{k+1})^2 \leq \left(\frac{1 - 2\epsilon \frac{\alpha L}{\alpha + L}}{1 + 2\epsilon \frac{\alpha L}{\alpha + L}} \right) W_2(\rho_k, \rho'_k)^2$$

which will imply the desired claim.

Let $x_k \sim \rho_k$, $x'_k \sim \rho'_k$ be coupled with the optimal coupling, so $W_2(\rho_k, \rho'_k) = \mathbb{E}[\|x_k - x'_k\|^2]$. We evolve x_k, x'_k via SLA (8) with the same Gaussian noise z_k (this is the synchronous coupling):

$$\begin{aligned} x_{k+1} + \epsilon \nabla f(x_{k+1}) &= x_k - \epsilon \nabla f(x_k) + \sqrt{2\epsilon} z_k \\ x'_{k+1} + \epsilon \nabla f(x'_{k+1}) &= x'_k - \epsilon \nabla f(x'_k) + \sqrt{2\epsilon} z_k. \end{aligned}$$

Subtracting and taking the squared norm, we obtain

$$\|x_{k+1} - x'_{k+1} + \epsilon(\nabla f(x_{k+1}) - \nabla f(x'_{k+1}))\|^2 = \|x_k - x'_k - \epsilon(\nabla f(x_k) - \nabla f(x'_k))\|^2.$$

We expand both sides and use the inequality (15). As in §A.1, for $0 < \epsilon \leq \frac{2}{\alpha + L}$, the right hand side is upper bounded by $(1 - \frac{2\epsilon\alpha L}{\alpha + L})\|x_k - x'_k\|^2$. Similarly, the left hand side is lower bounded by $(1 + \frac{2\epsilon\alpha L}{\alpha + L})\|x_{k+1} - x'_{k+1}\|^2$. Combining and taking expectation, we obtain

$$\mathbb{E}[\|x_{k+1} - x'_{k+1}\|^2] \leq \left(\frac{1 - 2\epsilon \frac{\alpha L}{\alpha + L}}{1 + 2\epsilon \frac{\alpha L}{\alpha + L}} \right) \mathbb{E}[\|x_k - x'_k\|^2] = \left(\frac{1 - 2\epsilon \frac{\alpha L}{\alpha + L}}{1 + 2\epsilon \frac{\alpha L}{\alpha + L}} \right) W_2(\rho_k, \rho'_k)^2.$$

Finally, by the definition of Wasserstein distance as the infimum over all coupling, we conclude

$$W_2(\rho_{k+1}, \rho'_{k+1})^2 \leq \left(\frac{1 - 2\epsilon \frac{\alpha L}{\alpha + L}}{1 + 2\epsilon \frac{\alpha L}{\alpha + L}} \right) W_2(\rho_k, \rho'_k)^2$$

as desired. \square

A.4. SLA and ULA for mixture of Gaussians

Let the target measure be an equal mixture of two Gaussians:

$$\nu = \frac{1}{2}\mathcal{N}(-a, I) + \frac{1}{2}\mathcal{N}(a, I)$$

for some $a \in \mathbb{R}^n$. Then we can write $f = -\log \nu$ as

$$f(x) = \frac{1}{2}\|x\|^2 - \log \cosh(\langle x, a \rangle) + \frac{1}{2}\|a\|^2 + \frac{n}{2} \log(2\pi).$$

The gradient of f is $\nabla f(x) = x - \tanh(\langle x, a \rangle)a$.

ULA. The ULA iteration for mixture of Gaussians is:

$$x_{k+1} = (1 - \epsilon)x_k + \epsilon \tanh(\langle x_k, a \rangle)a + \sqrt{2\epsilon} z_k$$

where $z_k \sim \mathcal{N}(0, I)$ is independent of x_k . We can run this directly.

SLA. The ULA iteration for mixture of Gaussians is:

$$(1 + \epsilon)x_{k+1} - \epsilon \tanh(\langle x_{k+1}, a \rangle) a = (1 - \epsilon)x_k + \epsilon \tanh(\langle x_k, a \rangle) a + \sqrt{4\epsilon} z_k \quad (16)$$

where $z_k \sim \mathcal{N}(0, I)$ is independent of x_k .

Let $v_k = \langle x_k, a \rangle \in \mathbb{R}$. Taking the inner product of both sides of (16) with a gives us

$$(1 + \epsilon)v_{k+1} - \epsilon \|a\|^2 \tanh(v_{k+1}) = (1 - \epsilon)v_k + \epsilon \|a\|^2 \tanh(v_k) + \sqrt{4\epsilon} \langle z_k, a \rangle.$$

Given v_k and z_k , we can invert the equation above to solve for v_{k+1} , which is well-defined for small ϵ . Once we have $v_{k+1} = \langle x_{k+1}, a \rangle$, we can substitute it to (16) to solve for x_{k+1} :

$$x_{k+1} = \frac{1 - \epsilon}{1 + \epsilon} x_k + \frac{\epsilon}{1 + \epsilon} (\tanh(v_k) + \tanh(v_{k+1})) a + \frac{\sqrt{4\epsilon}}{1 + \epsilon} z_k.$$

Appendix B. A review of discretization methods for a flow in space

We provide brief review of discretization methods, and refer to (Hairer et al., 2006) for more detail.

B.1. Integrator, order, and adjoint

Let $\varphi = (\varphi_t)_{t \in \mathbb{R}}$ be the flow of the differential equation $\dot{x} = v(x)$ for a smooth vector field v on \mathbb{R}^n . An *integrator* for φ is a family $A = (A_\epsilon)_{\epsilon \in \mathbb{R}}$ of algorithms $A_\epsilon: \mathbb{R}^n \rightarrow \mathbb{R}^n$, indexed by a *step size* $\epsilon \in \mathbb{R}$ (or ϵ in a neighborhood of 0), such that $A_0 = I$ is the identity map, $(\epsilon, x) \mapsto A_\epsilon(x)$ is smooth, and

$$\lim_{\epsilon \rightarrow 0} \frac{A_\epsilon(x) - x}{\epsilon} = v(x) \quad \forall x \in \mathbb{R}^n.$$

We say that the integrator A has *order* p , for some $p \in \mathbb{N} \cup \{+\infty\}$, if

$$\|\varphi_\epsilon(x) - A_\epsilon(x)\| \leq O(\epsilon^{p+1}) \quad \text{as } \epsilon \rightarrow 0, \forall x \in \mathbb{R}^n$$

where the bound on the right hand side above may depend on x .

The *adjoint* of an integrator $A = (A_\epsilon)_{\epsilon \in \mathbb{R}}$ is another integrator $A^* = (A_\epsilon^*)_{\epsilon \in \mathbb{R}}$ defined by

$$A_\epsilon^* = (A_{-\epsilon})^{-1}.$$

(If A_ϵ is defined for ϵ in a symmetric interval around 0, then A_ϵ^* is defined for ϵ in the same interval.) This definition of adjoint satisfies $(A^*)^* = A$ and $(A \circ B)^* = B^* \circ A^*$ for any integrators A and B , where $A \circ B = (A_\epsilon \circ B_\epsilon)_{\epsilon \in \mathbb{R}}$ is the composition of two integrators.

We say that an integrator A is *symmetric* if it is self-adjoint: $A^* = A$. We can symmetrize any integrator A by composing it with its adjoint in either order; this gives two symmetric integrators $A \circ A^*$ and $A^* \circ A$, which in general are not equal. We recall the following results.

Lemma 4 ((Hairer et al., 2006, Theorem 3.2))

1. An integrator A and its adjoint A^* have the same order.
2. If A is symmetric, then its order is even.

For example, the exact flow $\varphi = (\varphi_t)_{t \in \mathbb{R}}$ is a symmetric integrator of order $+\infty$. This is the ideal integrator, but typically not computable in practice. There are two first-order integrators that come from basic discretizations: the forward (explicit) method, and the backward (implicit) method. We can also symmetrize them to obtain second-order integrators. We describe them further below.

B.2. The forward method

We wish to approximate the differential equation $\dot{x} = v(x)$. The forward method $F = (F_\epsilon)_{\epsilon \in \mathbb{R}}$ uses the approximation $\frac{1}{\epsilon}(F_\epsilon(x) - x) = v(x)$, or equivalently,

$$F_\epsilon(x) = x + \epsilon v(x). \tag{17}$$

We also write

$$F_\epsilon = I + \epsilon v \tag{18}$$

where I is the identity map. This is known as the forward or explicit Euler method, because given where we are now, we can determine where to go next with little computation. If v is smooth and Lipschitz, then the forward method F_ϵ is a diffeomorphism for ϵ in a small enough neighborhood around 0, therefore F is an integrator. Furthermore, F has order 1 because F_ϵ is performing a first-order approximation.

B.3. The backward method

We wish to approximate the differential equation $\dot{x} = v(x)$. The backward method $B = (B_\epsilon)_{\epsilon \in \mathbb{R}}$ uses the approximation $\frac{1}{\epsilon}(B_\epsilon(x) - x) = v(B_\epsilon(x))$, or equivalently,

$$B_\epsilon(x) - \epsilon v(B_\epsilon(x)) = x. \tag{19}$$

Therefore,

$$B_\epsilon = (I - \epsilon v)^{-1}. \tag{20}$$

This is known as the backward or implicit Euler method, because to determine where to go next we need to solve an implicit equation (19), or equivalently compute the inverse of an operator (20). As in the forward method, if v is smooth and Lipschitz, then for small $|\epsilon|$, the map $I - \epsilon v$ is a diffeomorphism. So $B_\epsilon = (I - \epsilon v)^{-1}$ is also a diffeomorphism, and therefore B is an integrator.

Furthermore, observe that we can write

$$B_\epsilon = (F_{-\epsilon})^{-1} = F_\epsilon^*.$$

Therefore, the backward method is the adjoint of the forward method, and thus by Lemma 4 they have the same order 1.

B.4. Symmetrized methods

We can consider the symmetrized versions of the basic methods. There are two versions: the symmetrized forward method—which applies the forward method followed by the backward method—is known as the trapezoid rule, while the symmetrized backward method—which applies the backward method followed by the forward method—is known as the implicit midpoint rule. Both have order 2, being symmetric methods. We can also consider compositions of basic methods with varying step sizes, properly chosen to increase the order of the resulting algorithm; see (Hairer et al., 2006, §II.II.4).

B.5. On order of error and order of bias

We show that the order of the discretization error of an algorithm is the same as the order of the bias, at least under strong convexity. Thus, if we can find a discretization algorithm of high order, then we are guaranteed the bias will be small.

Suppose the vector field v satisfies the following monotonicity property for some $\alpha > 0$:

$$\langle v(x) - v(y), x - y \rangle \leq -\alpha \|x - y\|^2 \quad \forall x, y \in \mathbb{R}^n. \quad (21)$$

This is satisfied, for example, for gradient flow of a strongly convex function, i.e., $v(x) = -\nabla f(x)$ where $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is α -strongly convex. The condition above implies any two solutions of the flow of v contract exponentially fast, and in particular there is a unique fixed point x^* . We now show that any integrator for the flow of v will also converge exponentially fast to a biased limit of the same order as the discretization error.

Lemma 5 *Let $A = (A_\epsilon)_{\epsilon \in \mathbb{R}}$ be an integrator of order $p \in \mathbb{N}$ for the flow of v satisfying (21). Let $x_0 \in \mathbb{R}^n$ and define iteratively $x_{k+1} = A_\epsilon(x_k)$. Then*

$$\|x_k - x^*\| \leq O(\epsilon^p) + e^{-\alpha \epsilon k} \|x_0 - x^*\|.$$

In particular, the bias of A_ϵ is $O(\epsilon^p)$.

Proof Let $\tilde{x}(t) = \varphi_t(x_0)$ be the flow of v starting from $\tilde{x}(0)$. We compare x_k with $\tilde{x}(\epsilon k)$. By triangle inequality,

$$\|x_k - x^*\| \leq \|x_k - \tilde{x}(\epsilon k)\| + \|\tilde{x}(\epsilon k) - x^*\|.$$

The first term is upper bounded by $e^{-\alpha \epsilon k} \|\tilde{x}(0) - x^*\|$ by property (21). We need to show the second term $\delta_k := \|x_k - \tilde{x}(\epsilon k)\|$ is $O(\epsilon^p)$.

Consider the flow $\varphi_\epsilon(x_{k-1})$ of v starting at x_{k-1} . We also write $x_k = A_\epsilon(x_{k-1})$ and $\tilde{x}(\epsilon k) = \varphi_\epsilon(\tilde{x}(\epsilon(k-1)))$. By triangle inequality,

$$\delta_k = \|x_k - \tilde{x}(\epsilon k)\| \leq \|A_\epsilon(x_{k-1}) - \varphi_\epsilon(x_{k-1})\| + \|\varphi_\epsilon(x_{k-1}) - \varphi_\epsilon(\tilde{x}(\epsilon(k-1)))\|.$$

The first term above is $O(\epsilon^{p+1})$ since A_ϵ is an order- p integrator. The second term above is upper bounded by $e^{-\alpha \epsilon} \|x_{k-1} - \tilde{x}(\epsilon(k-1))\| = e^{-\alpha \epsilon} \delta_{k-1}$ by the contraction property of φ_t , by (21). Then

$$\delta_k \leq O(\epsilon^{p+1}) + e^{-\alpha \epsilon} \delta_{k-1}.$$

Unfolding the recursion with $\delta_0 = 0$ (since $\tilde{x}(0) = x_0$), we get

$$\delta_k \leq \frac{O(\epsilon^{p+1})(1 - e^{-\alpha \epsilon k})}{1 - e^{-\alpha \epsilon}} = O(\epsilon^p)$$

as desired. ■

Appendix C. A review of optimization in Riemannian manifold

We review optimization in a smooth Riemannian manifold. We focus on gradient flow and simple discretization methods including the forward and backward methods. We discuss sufficient conditions ensuring exponential convergence rate including gradient domination and strong convexity.

Let M be a complete smooth Riemannian manifold of dimension $n \geq 1$. Let

$$f: M \rightarrow \mathbb{R}$$

be a smooth objective function. We assume f is bounded below and achieves its minimum in M (not necessarily at a unique point), so the set of minimizers

$$x^*(f) = \arg \min_{x \in M} f(x)$$

is not empty. Let $\min f \equiv \min_{x \in M} f(x)$ be the minimum value of f .

We want to solve the optimization problem

$$\min_{x \in M} f(x)$$

which is equivalent to the problem of reaching the set of minimizers $x^*(f)$. The basic dynamics that achieves this task is gradient flow. We first recall some conditions that ensure exponential convergence rate.

C.1. Conditions ensuring exponential convergence

We say that f is α -strongly convex for some $\alpha > 0$ if

$$\text{Hess}_x f \succeq \alpha I \quad \forall x \in M.$$

Here recall the Hessian $\text{Hess}_x f$ is the bilinear form on the tangent space $\mathbb{T}_x M$ that measures the second rate of change of f along geodesics, and the condition above means $(\text{Hess}_x f)(v) \equiv (\text{Hess}_x f)(v, v) \geq \alpha \|v\|_x^2$ for all $v \in \mathbb{T}_x M$.

We say that f is α -gradient dominated for some $\alpha > 0$ if

$$\|\text{grad}_x f\|_x^2 \geq 2\alpha(f(x) - \min f) \quad \forall x \in M.$$

For $M = \mathbb{R}^n$, this is known as the Polyak-Łojaciewicz condition or the Kurdyka-Łojaciewicz condition (Polyak, 1963; Łojasiewicz, 1963). In the setting of $M = \mathcal{P}_2(\mathbb{R}^n)$ with the relative entropy functional, this is known as the logarithmic Sobolev inequality (Otto and Villani, 2000). Observe that if f is gradient dominated, then any stationary point of f (where the gradient vanishes) must be a global minimum.

We say that f has α -sufficient growth for some $\alpha > 0$ if

$$f(x) - \min f \geq \frac{\alpha}{2} d(x, x^*(f))^2 \quad \forall x \in M.$$

Here $d(x, x^*(f)) = \inf_{x^* \in x^*(f)} d(x, x^*)$ is the distance from x to the minimizer set $x^*(f)$. In the setting of $M = \mathcal{P}_2(\mathbb{R}^n)$ with the relative entropy functional, this is known as the Talagrand inequality (Otto and Villani, 2000).

We recall the following implications from (Otto and Villani, 2000, Propositions 1' & 2'), which follow by interpolating the inequalities along the gradient flow.

Lemma 6 *Let $\alpha > 0$.*

1. *If f is α -strongly convex, then f is α -gradient dominated.*
2. *If f is α -gradient dominated, then f has α -sufficient growth.*

C.2. Gradient flow

The basic dynamics for minimizing a function f is the *gradient flow*:

$$\dot{x} = -\text{grad}_x f. \quad (22)$$

Here $\text{grad}_x f \in \mathbb{T}_x M$ is the metric gradient of f at x , which in local coordinate is given by multiplying the vector of partial derivatives of f by the inverse metric at x . Gradient flow is a descent flow:

$$\frac{d}{dt}f(x) = \langle \text{grad}_x f, \dot{x} \rangle_x = -\|\text{grad}_x f\|_x^2 \leq 0.$$

We are interested in quantifying how fast the convergence occurs.

C.2.1. EXPONENTIAL CONTRACTION OF SOLUTIONS UNDER STRONG CONVEXITY

Strong convexity is the weakest condition needed for exponential contraction between solutions. Indeed, suppose f is α -strongly convex for some $\alpha > 0$, and let $x(t), y(t)$ be two solutions of the gradient flow (22). Assume $x(t)$ and $y(t)$ are close enough so that the geodesic connecting them is minimizing. Then the gradient of the squared distance $x \mapsto d(x, y(t))^2$ is $\text{grad}_{x(t)} d(\cdot, y(t))^2 = -2 \log_{x(t)}(y(t))$ (see for example (Ferreira, 2006)). Here $\log_x y \in \mathbb{T}_x M$ is the logarithm map, which is the inverse of the exponential map, i.e., $v = \log_x y$ if and only if $y = \exp_x(v)$. Similarly, the gradient of $y \mapsto d(y, x(t))^2$ is $\text{grad}_{y(t)} d(\cdot, x(t))^2 = -2 \log_{y(t)}(x(t))$. Then we have (hiding dependence on t for simplicity):

$$\begin{aligned} \frac{d}{dt}d(x(t), y(t))^2 &= -2\langle \log_x(y), \dot{x} \rangle_x - 2\langle \log_y(x), \dot{y} \rangle_y \\ &= 2\langle \log_x(y), \text{grad}_x f \rangle_x + 2\langle \log_y(x), \text{grad}_y f \rangle_y. \end{aligned} \quad (23)$$

On the other hand, by the strong convexity of f , we have

$$\begin{aligned} f(y) &\geq f(x) + \langle \text{grad}_x f, \log_x y \rangle_x + \frac{\alpha}{2}d(x, y)^2 \\ f(x) &\geq f(y) + \langle \text{grad}_y f, \log_y x \rangle_y + \frac{\alpha}{2}d(x, y)^2. \end{aligned}$$

Summing yields $\langle \text{grad}_x f, \log_x y \rangle_x + \langle \text{grad}_y f, \log_y x \rangle_y \leq -\alpha d(x, y)^2$. Substituting to (23) yields

$$\frac{d}{dt}d(x(t), y(t))^2 \leq -2\alpha d(x(t), y(t))^2.$$

Therefore, $d(x(t), y(t))^2 \leq e^{-2\alpha t} d(x(0), y(0))^2$, as desired.

C.2.2. EXPONENTIAL CONVERGENCE OF FUNCTION VALUE UNDER GRADIENT DOMINATION

Gradient domination is the weakest condition needed for exponential convergence of function value. Indeed, suppose f is α -gradient dominated for some $\alpha > 0$. Then along the gradient flow,

$$\frac{d}{dt}(f(x) - \min f) = -\|\text{grad}_x f\|_x^2 \leq -2\alpha(f(x) - \min f).$$

Therefore, $f(x(t)) - \min f \leq e^{-2\alpha t}(f(x(0)) - \min f)$. This also implies exponential convergence to the minimizer, since gradient domination implies sufficient growth.

C.3. Gradient descent

Gradient descent algorithm for minimizing f with step size $\epsilon > 0$ is the iteration

$$x_{k+1} = \exp_{x_k}(-\epsilon \text{grad}_{x_k} f).$$

Here we assume ϵ is small enough so the geodesic from x_k to x_{k+1} is minimizing.

As in gradient flow, in general we can get exponential contraction between solutions of gradient descent under strong convexity, and get exponential convergence in function value under gradient domination. However, we now also need a smoothness assumption on f . Specifically, suppose f is α -gradient dominated for some $\alpha > 0$. Assume further f is L -smooth for some $L > 0$, which means $\text{Hess } f \preceq LI$. This implies, for x, y sufficiently close,

$$f(y) \leq f(x) + \langle \text{grad}_x f, \log_x y \rangle_x + \frac{L}{2}d(x, y)^2.$$

Plugging in $x = x_k, y = x_{k+1}$ along gradient descent with $d(x, y)^2 = \epsilon^2 \|\text{grad}_{x_k} f\|_{x_k}^2$, we obtain

$$f(x_{k+1}) \leq f(x_k) - \epsilon \left(1 - \frac{\epsilon L}{2}\right) \|\text{grad}_{x_k} f\|_{x_k}^2. \quad (24)$$

If $0 < \epsilon \leq \frac{2}{L}$, then we can chain the last term above with the gradient domination inequality to conclude that

$$f(x_{k+1}) - \min f \leq \left(1 - 2\alpha\epsilon \left(1 - \frac{\epsilon L}{2}\right)\right) (f(x_k) - \min f).$$

Unrolling the recursion gives the exponential convergence in function value:

$$f(x_k) - \min f \leq \left(1 - 2\alpha\epsilon \left(1 - \frac{\epsilon L}{2}\right)\right)^k (f(x_0) - \min f).$$

For example, if $\epsilon = \frac{1}{L}$, then the rate is $(1 - \frac{\alpha}{L})^k$. This also implies exponential convergence in the distance to minimizer by the sufficient growth property. See also, for example, (Ferreira and Oliveira, 1998; Zhang and Sra, 2016; Karimi et al., 2016).

C.4. Proximal gradient

The proximal gradient method for minimizing f with step size $\epsilon > 0$ is

$$x_{k+1} = \arg \min_{x \in M} \left\{ f(x) + \frac{1}{2\epsilon} d(x, x_k)^2 \right\}. \quad (25)$$

Assume $\epsilon > 0$ is small enough so the minimizer above is unique and within the injectivity radius of x_k . Then the minimizer x_{k+1} is characterized by $\text{grad}_{x_{k+1}} (f + \frac{1}{2\epsilon} d(\cdot, x_k)^2) = 0$, or equivalently,

$$\log_{x_{k+1}} x_k = \epsilon \text{grad}_{x_{k+1}} f.$$

This is equivalent to the implicit update $\exp_{x_{k+1}} (\epsilon \text{grad}_{x_{k+1}} f) = x_k$, which generalizes the usual update $x_{k+1} + \epsilon \nabla f(x_{k+1}) = x_k$ in the Euclidean case.

As before, in general we can get exponential contraction between solutions under strong convexity, and get exponential convergence in function value under gradient domination. Unlike in gradient descent, here we do not need a smoothness assumption on f , but we need to solve the implicit update above.

Suppose f is α -gradient dominated for some $\alpha > 0$. Since x_{k+1} is the minimizer of (25),

$$f(x_{k+1}) + \frac{1}{2\epsilon} d(x_{k+1}, x_k)^2 \leq f(x_k).$$

Equivalently, since $d(x_{k+1}, x_k)^2 = \epsilon^2 \|\text{grad}_{x_{k+1}} f\|_{x_{k+1}}^2$,

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2\epsilon} d(x_{k+1}, x_k)^2 = -\frac{\epsilon}{2} \|\text{grad}_{x_{k+1}} f\|_{x_{k+1}}^2.$$

Now using the gradient domination inequality and collecting the terms give us

$$f(x_{k+1}) - \min f \leq \frac{1}{1 + \alpha\epsilon} (f(x_k) - \min f). \quad (26)$$

Unfolding the recursion, we conclude that

$$f(x_k) - \min f \leq \frac{1}{(1 + \alpha\epsilon)^k} (f(x_0) - \min f).$$

See also, for example, (Ferreira and Oliveira, 2002; de Carvalho Bento et al., 2016)

C.5. Symmetrized forward method

The symmetrized forward method for minimizing f is the composition of the gradient descent and the proximal gradient methods:

$$x_{k+\frac{1}{2}} = \exp_{x_k} (-\epsilon \text{grad}_{x_k} f) \quad (27a)$$

$$x_{k+1} = \arg \min_{x \in M} \left\{ f(x) + \frac{1}{2\epsilon} d(x, x_{k+\frac{1}{2}})^2 \right\}. \quad (27b)$$

Note that this is the Forward-Backward algorithm for composite optimization, applied to the self-decomposition $2f = f + f$.

Suppose f is α -gradient dominated and L -smooth for some $0 < \alpha \leq L$, and let $0 < \epsilon \leq \frac{2}{L}$. Then as in (24), the first update (27a) above implies

$$f(x_{k+\frac{1}{2}}) - f(x_{k+1}) \leq -\epsilon \left(1 - \frac{\epsilon L}{2}\right) \|\text{grad}_{x_k} f\|_{x_k}^2.$$

Similarly, as in (26), the second update (27b) above implies

$$f(x_{k+1}) - f(x_{k+\frac{1}{2}}) \leq -\frac{\epsilon}{2} \|\text{grad}_{x_{k+1}} f\|_{x_{k+1}}^2.$$

Summing the two inequalities above and using the gradient domination inequality give us

$$f(x_{k+1}) - f(x_k) \leq -2\alpha\epsilon \left(1 - \frac{\epsilon L}{2}\right) (f(x_k) - \min f) - \alpha\epsilon (f(x_{k+1}) - \min f).$$

Collecting terms and unrolling the recursion, we conclude the exponential convergence rate

$$f(x_k) - \min f \leq \left(\frac{1 - 2\alpha\epsilon \left(1 - \frac{\epsilon L}{2}\right)}{1 + \alpha\epsilon}\right)^k (f(x_0) - \min f).$$

For example, if $\epsilon = \frac{1}{L}$, then the rate is $\left(\frac{1 - \frac{\alpha}{L}}{1 + \frac{\alpha}{L}}\right)^k$, which is slightly better than for gradient descent.

Appendix D. A brief review of the Wasserstein metric in the space of measures

Let $\mathcal{P} \equiv \mathcal{P}_2(\mathbb{R}^n)$ be the space of probability measures ρ on \mathbb{R}^n with finite second moments. We provide a brief review of the Wasserstein metric, and refer the reader for more detail to Villani (2003, 2008); Carlen and Gangbo (2003); Otto and Villani (2000).

The Wasserstein metric W_2 on \mathcal{P} is defined as

$$W_2(\rho, \nu)^2 = \inf_{X \sim \rho, Y \sim \nu} \mathbb{E}[\|X - Y\|^2] \quad (28)$$

where the infimum is over all coupling of random variables (X, Y) with $X \sim \rho$ and $Y \sim \nu$. This formally endows \mathcal{P} with an infinite-dimensional smooth Riemannian metric.

A tangent vector/function $R \in T_\rho \mathcal{P}$ is of the form

$$R = -\nabla \cdot (\rho \nabla \phi)$$

for some function $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$. We also write $R \equiv \nabla \phi$. The squared norm of $R \equiv \nabla \phi$ is

$$\|R\|_\rho^2 = \|\nabla \phi\|_\rho^2 = \int \rho \|\nabla \phi\|^2 = \mathbb{E}_\rho[\|\nabla \phi\|^2].$$

The gradient of a functional $F: \mathcal{P} \rightarrow \mathbb{R}$ is

$$\text{grad}_\rho F = -\nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho} \right)$$

where $\frac{\delta F}{\delta \rho}(x)$ is the naive (L^2) derivative of F with respect to $\rho(x)$. There are also Hessian formulae for specific functional classes (Villani, 2008, §15).

When ρ and ν are smooth and absolutely continuous, the optimal coupling in the Wasserstein distance (28) is unique and induced by an optimal transport map which is the gradient of a convex function: $(X, Y) = (X, \nabla\phi(X))$, where $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and satisfies $(\nabla\phi)_\# \rho = \nu$. Therefore, if there is a convex function whose gradient pushes forward one measure to another, then that gradient must be the optimal transport map. Furthermore, the geodesic in the Wasserstein metric connecting one measure to another is obtained by linearly interpolating the optimal transport map in space and taking the pushforward. We summarize this fact in the following lemma, which we use in our discussion in this paper.

Lemma 7 *Let $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ satisfy $\nabla^2\phi \succeq KI$ for some $K \in \mathbb{R}$. For $0 \leq \epsilon < \frac{1}{\max\{0, -K\}}$, the exponential map $\exp_\rho(\epsilon\nabla\phi)$ in the Wasserstein metric is given by the pushforward of the linear map $(I + \epsilon\nabla\phi)_\#$.*

Proof The map $I + \epsilon\nabla\phi = \nabla\left(\frac{\|\cdot\|^2}{2} + \epsilon\phi\right)$ is the gradient of a convex function that pushes forward ρ to $\exp_\rho(\epsilon\nabla\phi)$, so it must be the optimal transport map. \blacksquare

Appendix E. Details for §3: Optimization in the space of measures

E.1. Expected value

Let

$$F(\rho) = \mathbb{E}_\rho[f] = \int_{\mathbb{R}^n} \rho(x)f(x) dx$$

for some smooth function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ in space. If f is minimized at $x^*(f) = \arg \min_{x \in \mathbb{R}^n} f(x)$ with minimum value $\min f$, then F is minimized at any measure ρ supported on $x^*(f)$ with the same minimum value, $\min F = \min f$. Essentially, the behavior of F is the same as that of f .

E.1.1. GRADIENT DOMINATION AND STRONG CONVEXITY

The gradient of $F(\rho) = \mathbb{E}_\rho[f]$ is $\text{grad}_\rho F = -\nabla \cdot (\rho\nabla f) \equiv \nabla f$, so $\|\text{grad}_\rho F\|^2 = \mathbb{E}_\rho[\|\nabla f\|^2]$.

The gradient-domination condition $\|\text{grad} F\|^2 \geq 2\alpha(F - \min F)$, $\alpha > 0$, becomes

$$\mathbb{E}_\rho[\|\nabla f\|^2] \geq 2\alpha \mathbb{E}_\rho[f - \min f] \quad \forall \rho \in \mathcal{P}.$$

Therefore, F satisfies gradient domination in the space of measures if and only if f satisfies the gradient-domination condition $\|\nabla f\|^2 \geq 2\alpha(f - \min f)$ in space.

Similarly, the Hessian of F is given by $(\text{Hess}_\rho F)(R, R) = \mathbb{E}_\rho[\langle \nabla\phi, (\nabla^2 f)\nabla\phi \rangle]$, for a tangent function $R \equiv \nabla\phi \in \mathbb{T}_\rho\mathcal{P}$. Therefore, F is strongly convex in the space of measures if and only if f is strongly convex in space. In particular, if f (and hence F) is strongly convex, then f has a unique minimizer $x^* \in \mathbb{R}^n$, and F has a unique minimizer $\delta_{x^*} \in \mathcal{P}$ which is the point mass at x^* .

E.1.2. GRADIENT FLOW

The gradient flow equation $\dot{\rho} = -\text{grad}_\rho F$ is

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho\nabla f).$$

This is the continuity equation of the gradient flow equation of f in space:

$$\dot{x} = -\nabla f(x).$$

Therefore, the gradient flow of $F(\rho) = \mathbb{E}_\rho[f]$ is implementable by the gradient flow of f in space. Furthermore, the rate of convergence for F is the same as the rate of convergence for f . If f (and hence F) is gradient dominated, then the convergence is exponential in the function value, and also in the distance to the set of minimizers. If f (and hence F) is strongly convex, then any two coevolving solutions are contracting exponentially fast. If f (and hence F) is nonconvex, then the limiting measure is supported on the set of local minima, with weights equal to the measure of the corresponding basins of attraction under the initial measure.

E.1.3. GRADIENT DESCENT

The forward method or gradient descent $\rho_{k+1} = \exp_{\rho_k}(-\epsilon \operatorname{grad}_{\rho_k} F)$ becomes

$$\rho_{k+1} = \exp_{\rho_k}(-\epsilon \nabla f) \tag{29}$$

where recall $\operatorname{grad}_\rho F \equiv \nabla f$. Assume f is L -smooth, which means $\nabla^2 f \preceq LI$. Then for $\epsilon < \frac{1}{L}$, the gradient descent for F (29) is implemented by gradient descent for f in space:

$$\rho_{k+1} = (I - \epsilon \nabla f) \# \rho_k.$$

Furthermore, if f (and hence F) is α -gradient-dominated for some $\alpha > 0$, then gradient descent converges exponentially fast in the function value:

$$F(\rho_k) - \min F \leq (1 - \alpha\epsilon(2 - \epsilon L))^k (F(\rho_0) - \min F).$$

For example, for $\epsilon = \frac{1}{2L}$, the rate is $(1 - \frac{3}{4}\frac{\alpha}{L})^k$. This also implies an exponential convergence rate in the distance to minimizer. Similarly, if f (and hence F) is strongly convex, then any two solutions are contracting exponentially fast.

E.1.4. PROXIMAL GRADIENT

The backward method or proximal gradient algorithm for $F(\rho) = \mathbb{E}_\rho[f]$ is

$$\rho_{k+1} = \arg \min_{\rho \in \mathcal{P}} \left\{ \mathbb{E}_\rho[f] + \frac{1}{2\epsilon} W_2(\rho, \rho_k)^2 \right\}.$$

Equivalently, ρ_{k+1} solves the adjoint version of the equation (29):

$$\exp_{\rho_{k+1}}(\epsilon \nabla f) = \rho_k.$$

This is implemented by the backward method or proximal gradient algorithm for f in space:

$$\rho_{k+1} = ((I + \epsilon \nabla f)^{-1}) \# \rho_k.$$

See below for detail. Here $(I + \epsilon \nabla f)^{-1}$ is the proximal gradient operator for f ; i.e., $x_{k+1} = (I + \epsilon \nabla f)^{-1}(x_k)$ if and only if

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\epsilon} \|x - x_k\|^2 \right\}.$$

This is well-defined if f satisfies $-LI \preceq \nabla^2 f \preceq LI$ for some $L > 0$, and $0 < \epsilon < \frac{1}{L}$. Furthermore, if f (and hence F) is α -gradient-dominated for some $\alpha > 0$, then proximal gradient converges exponentially fast in the function value:

$$F(\rho_k) - \min F \leq \frac{1}{(1 + \alpha\epsilon)^k} (F(\rho_0) - \min F).$$

This also implies an exponential convergence rate in the distance to minimizer. Similarly, if f (and hence F) is strongly convex, then any two solutions are contracting exponentially fast.

E.1.5. DETAIL FOR §E.1.4: PROXIMAL GRADIENT FOR EXPECTED VALUE

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a smooth function.

The *proximal step* for f with step size $\epsilon > 0$ is the map $P_{\epsilon, f}: \mathbb{R}^n \rightarrow \mathbb{R}^n$ given by

$$P_{\epsilon, f}(x) = \arg \min_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2\epsilon} \|y - x\|^2 \right\}.$$

Taking derivative and setting it to zero, the solution $y = P_{\epsilon, f}(x)$ satisfies $\nabla f(y) + \frac{1}{\epsilon}(y - x) = 0$, or equivalently, $y + \epsilon \nabla f(y) = x$. Thus,

$$P_{\epsilon, f}(x) = (I + \epsilon \nabla f)^{-1}(x).$$

This is well-defined if f is L -smooth, i.e., $-LI \preceq \nabla^2 f(x) \preceq LI$, and $0 < \epsilon < \frac{1}{L}$.

In the space of measure, this corresponds to the pushforward map

$$\tilde{P}_{\epsilon, f}(\nu) = ((I + \epsilon \nabla f)^{-1})_{\#} \nu.$$

We will show this is also the proximal step in the space of measure of the corresponding expected value functional:

$$\tilde{P}_{\epsilon, f}(\nu) = \arg \min_{\rho \in \mathcal{P}} \left\{ \mathbb{E}_{\rho}[f] + \frac{1}{2\epsilon} W(\rho, \nu)^2 \right\}.$$

We calculate the derivative and set it to zero at stationary point. At the minimizer $\rho = \tilde{P}_{\epsilon, f}(\nu)$, the gradient of $\tilde{F}(\rho) = \mathbb{E}_{\rho}[f] + \frac{1}{2\epsilon} W(\rho, \nu)^2$ vanishes. This means for any motion of ρ given by $\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \xi) = 0$, we have

$$0 = \frac{d}{dt} \Big|_{t=0} \tilde{F}(\rho) = \int \rho(y) \left\langle \nabla f(y) + \frac{1}{\epsilon}(y - \nabla \varphi^*(y)), \xi(y) \right\rangle dy.$$

Here $\nabla \varphi$ is the optimal transport map that sends ν to ρ , so $I - \nabla \varphi^*$ is the optimal map that sends ρ to ν . Since the integral above is zero for any ξ , we must have (ρ -almost surely)

$$y + \epsilon \nabla f(y) = \nabla \varphi^*(y).$$

Thus,

$$\nabla \varphi^* = I + \epsilon \nabla f$$

so the optimal map that sends ν to ρ is

$$\nabla \varphi = (I + \epsilon \nabla f)^{-1}$$

which is the proximal step map. Therefore, the solution $\rho = \tilde{P}_{\epsilon, f}(\nu)$ is given by

$$\rho = (\nabla \varphi)_{\#} \nu = ((I + \epsilon \nabla f)^{-1})_{\#} \nu$$

which is the same as the pushforward of the proximal map above, as desired.

E.2. Negative entropy

Suppose

$$F(\rho) = -H(\rho) = \int_{\mathbb{R}^n} \rho(x) \log \rho(x) dx$$

is the negative entropy. It takes values in $\mathbb{R} \cup \{+\infty\}$. The value $+\infty$ is achieved by a point mass, or any distribution supported on a set of Lebesgue measure zero. Negative entropy is also unbounded below, with the limiting value $-\infty$ achieved by the Lebesgue measure (which is not a probability measure). For example, for a Gaussian distribution $\rho = \mathcal{N}(\mu, \Sigma)$, the negative entropy is $F(\rho) = -\frac{1}{2} \log \det(2\pi e\Sigma)$.

E.2.1. THE HEAT FLOW AS GRADIENT FLOW OF NEGATIVE ENTROPY

The gradient of $F(\rho) = -H(\rho)$ is $\text{grad}_\rho F = -\nabla \cdot (\rho \nabla \log \rho) = -\Delta \rho$, or $\text{grad}_\rho F \equiv \nabla \log \rho$. Therefore, the gradient flow equation $\dot{\rho} = -\text{grad}_\rho F$ of negative entropy is the *heat equation*:

$$\frac{\partial \rho}{\partial t} = \Delta \rho.$$

This has an exact solution which is the *heat flow*:

$$\rho_t = \rho_0 * \mathcal{N}(0, 2tI).$$

In space, this is implemented via addition of independent Gaussian noise:

$$X_t = X_0 + \sqrt{2t}Z \tag{30}$$

where $Z \sim \mathcal{N}(0, I)$ is independent of X_0 .⁴ Note that ρ_t tends to the Lebesgue measure as $t \rightarrow \infty$. Note also that variance grows linearly along the heat flow.

For example, if $\rho_0 = \mathcal{N}(\mu, \Sigma)$, then $\rho_t = \mathcal{N}(\mu, \Sigma + 2tI)$. In this case negative entropy decreases logarithmically: $F(\rho_t) = -\frac{n}{2} \log(2\pi e) - \frac{1}{2} \sum_{i=1}^n \log(\lambda_i + 2t)$, where $\lambda_1, \dots, \lambda_n > 0$ are the eigenvalues of $\Sigma \succ 0$. This logarithmic rate is also the fastest rate at which negative entropy decreases along the heat flow. This follows from the fact that Gaussian minimizes negative entropy for a given covariance, so $F(\rho_t) \geq F(\mathcal{N}(0, \Sigma + 2tI))$ where Σ is the covariance of ρ_0 .

By the nonnegativity of mutual information, we also have a logarithmic upper bound on negative entropy. Indeed, since $I(X_0; X_t) = H(X_t) - H(X_t | X_0) \geq 0$ and $H(X_t | X_0) = H(\mathcal{N}(0, 2tI))$, we have $F(\rho_t) \leq F(\mathcal{N}(0, 2tI))$. However, the upper bound is $+\infty$ at $t = 0$. Using the entropy power inequality, we can improve this to $F(\rho_t) \leq -\frac{n}{2} \log(e^{-\frac{2}{n}F(\rho_0)} + e^{-\frac{2}{n}F(\mathcal{N}(0, 2tI))})$, which has the same logarithmic behavior for large t , and takes the correct value $F(\rho_0)$ at $t = 0$.

E.2.2. FISHER INFORMATION AND CONVEXITY OF NEGATIVE ENTROPY

The squared norm of the gradient of negative entropy $F = -H$ is the *Fisher information*:

$$J(\rho) = \mathbb{E}_\rho[\|\nabla \log \rho\|^2].$$

4. The true solution of the heat flow is the Brownian motion in space. However, at each time, the solution has the same distribution as (30) above.

Therefore, the gradient flow identity $\frac{d}{dt}F(\rho) = -\|\text{grad}_\rho F\|^2$ becomes $\frac{d}{dt}H(\rho) = J(\rho)$, which is known as De Bruijn's identity in information theory (Stam, 1959).

The Hessian of $F(\rho) = -H(\rho)$ is, on a tangent function $R = -\nabla \cdot (\rho \nabla \phi) \in \mathbb{T}_\rho \mathcal{P}$,

$$(\text{Hess}_\rho F)(R, R) = \mathbb{E}_\rho[\|\nabla^2 \phi\|_{\text{HS}}^2].$$

The expression above is nonnegative; therefore, negative entropy is a convex functional in the space of measures. It is not strictly convex in general, but it is strongly convex along geodesics that preserve the mean, with strong convexity parameter the Poincaré constant of the distribution; see (Carlen and Gangbo, 2003, Corollary 2.5). Note that the Hessian of negative entropy is not bounded above.

We also note that the gradient flow identity $\frac{d^2}{dt^2}F(\rho) = 2(\text{Hess } F)(\text{grad } F, \text{grad } F)$ becomes $\frac{d^2}{dt^2}H(\rho) = -2K(\rho)$, where $K(\rho)$ is the *second-order Fisher information*:

$$K(\rho) = \mathbb{E}_\rho[\|\nabla^2 \log \rho\|_{\text{HS}}^2].$$

This is used in the proof of the concavity of entropy power along the heat flow (Dembo et al., 1991; Villani, 2000). Moreover, since $\frac{d}{dt}H(\rho) = J(\rho) \geq 0$ and $\frac{d^2}{dt^2}H(\rho) = -2K(\rho) \leq 0$, entropy is increasing and concave along the heat flow; this is also a consequence of the general fact that a convex function is decreasing in a convex manner along its own gradient flow.

E.2.3. FORWARD METHOD FOR HEAT FLOW

The forward method $\rho_{k+1} = \exp_{\rho_k}(-\epsilon \text{grad}_{\rho_k} F)$ for negative entropy $F = -H$ is

$$\rho_{k+1} = \exp_{\rho_k}(-\epsilon \nabla \log \rho_k)$$

where recall $\text{grad}_\rho F \equiv \nabla \log \rho$. If ρ_k is K -log-semiconcave, which means $-\nabla^2 \log \rho_k \succeq KI$ for some $K \in \mathbb{R}$, then for $\epsilon \leq \frac{1}{\max\{0, -K\}}$, the forward method above is implemented by:

$$\rho_{k+1} = (I - \epsilon \nabla \log \rho_k) \# \rho_k. \quad (31)$$

If we know the analytic form of $\nabla \log \rho_k$, then we can run one step of the algorithm above by $x_{k+1} = x_k - \epsilon \nabla \log \rho_k(x_k)$. However, in the next step we do not know the analytic form of $\nabla \log \rho_{k+1}$, so we cannot continue running it.

For Gaussian initial data, we can solve the forward method. In this case the distributions stay Gaussian, but the variance grows faster than in the heat flow.

Example 6 (Forward method for heat flow with Gaussian data.) Let $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$, $\epsilon > 0$. Along the forward method (31) for the heat flow, $\rho_k = \mathcal{N}(\mu_k, \Sigma_k)$ stays Gaussian. Furthermore, the algorithm (31) becomes the iteration $x_{k+1} = (I + \epsilon \Sigma_k^{-1})x_k - \epsilon \Sigma_k^{-1} \mu_k$. Since $x_k \sim \mathcal{N}(\mu_k, \Sigma_k)$, $x_{k+1} \sim \mathcal{N}(\mu_k, \Sigma_k(I + \epsilon \Sigma_k^{-1})^2)$. Therefore, $\mu_{k+1} = \mu_k = \mu_0$ and $\Sigma_{k+1} = \Sigma_k(I + \epsilon \Sigma_k^{-1})^2$. Note that $\Sigma_k \succ \Sigma_{k-1} + 2\epsilon I \succ \Sigma_0 + 2\epsilon kI$, so variance grows faster than in the heat flow (with $t = \epsilon k$).

E.2.4. BACKWARD METHOD FOR HEAT FLOW

The backward method $\rho_{k+1} = \arg \min_{\rho \in \mathcal{P}} \{F(\rho) + \frac{1}{2\epsilon} W_2(\rho, \rho_k)^2\}$ for negative entropy $F = -H$ is

$$\exp_{\rho_{k+1}}(\epsilon \nabla \log \rho_{k+1}) = \rho_k.$$

If ρ_{k+1} is L -log-smooth, which means $-\nabla^2 \log \rho_{k+1} \preceq LI$ for some $L > 0$, then for $\epsilon \leq \frac{1}{L}$, the backward method above is implemented (implicitly) by:

$$(I + \epsilon \nabla \log \rho_{k+1})_{\#} \rho_{k+1} = \rho_k. \quad (32)$$

In general this is not solvable analytically. For Gaussian initial data, we can solve the backward method. In this case the distributions stay Gaussian, but the variance grows slower than in the heat flow (and in contrast to the forward method), see also (Carlen and Gangbo, 2003, Remark, §3).

Example 7 (Backward method for heat flow with Gaussian data.) Let $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and $0 < \epsilon \leq \lambda_{\min}(\Sigma_0)$. Along the backward method (32) for the heat flow, $\rho_k = \mathcal{N}(\mu_k, \Sigma_k)$ stays Gaussian. Furthermore, the algorithm (32) becomes the iteration $(I - \epsilon \Sigma_{k+1}^{-1})x_{k+1} + \epsilon \Sigma_{k+1}^{-1} \mu_{k+1} = x_k$. Since $x_{k+1} \sim \mathcal{N}(\mu_{k+1}, \Sigma_{k+1})$, $x_k \sim \mathcal{N}(\mu_{k+1}, \Sigma_{k+1}(I - \epsilon \Sigma_{k+1}^{-1})^2)$. Therefore, $\mu_{k+1} = \mu_k = \mu_0$ and $\Sigma_{k+1}(I - \epsilon \Sigma_{k+1}^{-1})^2 = \Sigma_k$. Equivalently, $\Sigma_{k+1} = \frac{1}{2}(\Sigma_k + 2\epsilon I + \{\Sigma_k(\Sigma_k + 4\epsilon I)\}^{\frac{1}{2}}) = \Sigma_k + 2\epsilon I - \epsilon^2 \Sigma_k^{-1} + O(\epsilon^3)$. In particular, $\Sigma_k \prec \Sigma_{k-1} + 2\epsilon I \prec \Sigma_0 + 2\epsilon k I$, so variance grows slower than in the heat flow.

E.3. Gradient flow of variance

The variance of a probability density ρ on \mathbb{R}^n can be written as

$$\text{Var}(\rho) = \frac{1}{2} \int_{\mathbb{R}^n \times \mathbb{R}^n} \rho(x)\rho(y)\|x - y\|^2 dx dy.$$

This is an example of an ‘‘interaction energy’’ of (Villani, 2003, §5.2.2), which in general is of the form $F(\rho) = \frac{1}{2} \int_{\mathbb{R}^n \times \mathbb{R}^n} \rho(x)\rho(y)W(x - y) dx dy$ where $W: \mathbb{R}^n \rightarrow \mathbb{R}$ is a symmetric convex function. Variance is the case when $W(x) = \|x\|^2$.

In general, the gradient of the interaction energy is $\text{grad}_{\rho} F = -\nabla \cdot (\rho \nabla(\rho * W)) \equiv \nabla(\rho * W)$, where $*$ is the convolution. For variance with $W(x) = \|x\|^2$, we have $(\rho * W)(x) = \int_{\mathbb{R}^n} \rho(y)\|x - y\|^2 dy$. Then $\nabla(\rho * W)(x) = 2 \int_{\mathbb{R}^n} \rho(y)(x - y) dy = 2(x - \mu(\rho))$ where $\mu(\rho) = \mathbb{E}_{\rho}[X]$ is the mean of ρ . Therefore, the gradient of variance is $\text{grad}_{\rho} \text{Var} = -\nabla \cdot (\rho \nabla(\rho * W)) = -2\nabla \cdot (\rho(x - \mu(\rho)))$. Notice that

$$\|\text{grad}_{\rho} \text{Var}\|^2 = \int \rho \|\nabla(\rho * W)\|^2 = 4 \int \rho \|x - \mu(\rho)\|^2 = 4 \text{Var}(\rho).$$

Therefore, along the gradient flow of variance, $\frac{d}{dt} \text{Var}(\rho) = -\|\text{grad}_{\rho} \text{Var}\|^2 = -4 \text{Var}(\rho)$, which implies $\text{Var}(\rho_t) = e^{-4t} \text{Var}(\rho_0)$.

The gradient flow of variance $\dot{\rho} = -\text{grad}_{\rho} \text{Var}$ is

$$\frac{\partial \rho}{\partial t} = 2\nabla \cdot (\rho(x - \mu(\rho))).$$

Observe that the mean is preserved:

$$\frac{d}{dt}\mu(\rho) = \int_{\mathbb{R}^n} \frac{\partial \rho}{\partial t} x = 2 \int_{\mathbb{R}^n} \nabla \cdot (\rho(x - \mu(\rho))) x = -2 \int_{\mathbb{R}^n} \rho(x - \mu(\rho)) = -2(\mu(\rho) - \mu(\rho)) = 0$$

where in the calculation above we have used integration by parts. Therefore, $\mu(\rho) = \mu(\rho_0) = \mu_0$ is fixed along the gradient flow of variance. Thus, the gradient flow of variance becomes

$$\frac{\partial \rho}{\partial t} = 2\nabla \cdot (\rho(x - \mu_0)).$$

This has an exact solution

$$\rho(t, x) = e^{2nt} \rho_0(e^{2t}x + (I - e^{2t})\mu_0).$$

Furthermore, this is implemented in space by shrinking around the mean:

$$x(t) = \mu_0 + e^{-2t}(x - \mu_0).$$

Appendix F. A review of composite optimization

Let M be an n -dimensional smooth Riemannian manifold. We consider the composite optimization problem

$$\min_{x \in M} f(x) + g(x)$$

where $f, g: M \rightarrow \mathbb{R}$ are smooth functions. One way to solve this optimization problem is to apply basic algorithms—gradient flow, gradient descent, or proximal gradient—to the objective function $f + g$. However, sometimes we can compute these algorithms on f and g individually, but not on their sum; an example is the Langevin dynamics in discrete time.

One algorithm for solving this optimization problem is the *Forward-Backward* (FB) algorithm, which alternates between a forward step (gradient descent) for f and a backward step (proximal gradient) for g . The iteration can be written as a composition of two steps, with a step size $\epsilon > 0$:

$$x_{k+\frac{1}{2}} = \exp_{x_k}(-\epsilon \operatorname{grad}_{x_k} f) \tag{33a}$$

$$x_{k+1} = \arg \min_{x \in M} \left\{ g(x) + \frac{1}{2\epsilon} d(x, x_{k+\frac{1}{2}})^2 \right\}. \tag{33b}$$

Recall that the optimality condition for the second equation (33b) is $\exp_{x_{k+1}}(\epsilon \operatorname{grad}_{x_{k+1}} g) = x_{k+\frac{1}{2}}$. Combining this with (33a), we can write the FB algorithm as defined implicitly by the identity

$$\exp_{x_{k+1}}(\epsilon \operatorname{grad}_{x_{k+1}} g) = \exp_{x_k}(-\epsilon \operatorname{grad}_{x_k} f). \tag{34}$$

F.1. Why does Forward-Backward work?

The minimizer (or any stationary point) of $f + g$ is a fixed point of the forward-backward algorithm. This is because the backward step (proximal gradient) is the inverse of the negative forward step (gradient descent).

Concretely, let $x^* \in M$ be a stationary point of $f + g$, so $\operatorname{grad}_{x^*} f + \operatorname{grad}_{x^*} g = 0$. Suppose we start at $x_k = x^*$. Then in the first half-step we move to $x_{k+\frac{1}{2}} = \exp_{x^*}(-\epsilon \operatorname{grad}_{x^*} f)$, which

in general is different from x^* (because x^* may not be a stationary point of f , so $\text{grad}_{x^*} f \neq 0$). However, the next half-step brings us back to $x_{k+1} = x^*$; this is because x_{k+1} by definition satisfies the consistency equation

$$\exp_{x_{k+1}}(\epsilon \text{grad}_{x_{k+1}} g) = x_{k+\frac{1}{2}} = \exp_{x^*}(-\epsilon \text{grad}_{x^*} f)$$

and we see that $x_{k+1} = x^*$ is a solution since $\text{grad}_{x^*} g = -\text{grad}_{x^*} f$.

This pairing between methods which are inverses of each other is important to make the algorithm converge to the true minimizer of the composite function. By symmetry, we can also use the backward and forward method, which will also converge to the correct minimizer. However, if we choose any other pairing, then the algorithm will have a bias, i.e., it will converge to a point that is different from the true minimizer.

F.1.1. EXAMPLE: QUADRATIC FUNCTION IN \mathbb{R}

Let $M = \mathbb{R}$ and $f(x) = \frac{1}{2}(x - a)^2$, for some $a \in \mathbb{R}$. The basic algorithms with step size $\epsilon > 0$ are:

Algorithm	Iteration
Gradient descent (GD)	$x_{k+1} = (1 - \epsilon)x_k + \epsilon a$
Gradient flow (GF)	$x_{k+1} = e^{-\epsilon}x_k + (1 - e^{-\epsilon})a$
Proximal gradient (PG)	$x_{k+1} = \frac{1}{1+\epsilon}x_k + \frac{\epsilon}{1+\epsilon}a$

Table 1: Basic algorithms applied to $f(x) = \frac{1}{2}(x - a)^2$ in \mathbb{R} .

Observe that $1 - \epsilon \leq e^{-\epsilon} \leq \frac{1}{1+\epsilon}$, so in the quadratic case we see that PG is faster than GF which in turn is faster than GD.

Now consider the composite optimization problem $\min_{x \in \mathbb{R}} f(x) + g(x)$ where

$$f(x) = \frac{1}{2}(x - 1)^2 \quad \text{and} \quad g(x) = \frac{1}{2}(x + 1)^2.$$

Then $f(x) + g(x) = x^2 + 1$, which is minimized at $x^* = 0$; but note that $x^* = 0$ is not a stationary point of f or g .

We consider composite algorithms to solve the composite optimization problem in which we alternately apply a basic algorithm to each of f and g :

$$\begin{aligned} x_{k+\frac{1}{2}} &= A_{\epsilon, f}(x_k) \\ x_{k+1} &= A'_{\epsilon, g}(x_{k+\frac{1}{2}}) \end{aligned}$$

for some $A, A' \in \{\text{GD}, \text{GF}, \text{PG}\}$. For each combination, we use Table 1 to compute the exact iteration from x_k to x_{k+1} and determine the limit point. If the limit point is not $x^* = 0$, then the composite algorithm is biased; else, it is unbiased. The results are in Table 2. Note that all are biased except for two: the Forward-Backward algorithm, and the Backward-Forward algorithm.

Alg. for f	Alg. for g	Iteration	Limit
GD	GD	$x_{k+1} = (1 - \epsilon)^2 x_k - \epsilon^2$	$-\frac{\epsilon}{2-\epsilon}$
	GF	$x_{k+1} = e^{-\epsilon}(1 - \epsilon)x_k + e^{-\epsilon}(1 + \epsilon) - 1$	$\frac{e^{-\epsilon}(1+\epsilon)-1}{1-e^{-\epsilon}(1-\epsilon)}$
	PG	$x_{k+1} = \frac{1-\epsilon}{1+\epsilon}x_k$	0
GF	GD	$x_{k+1} = e^{-\epsilon}(1 - \epsilon)x_k + (1 - e^{-\epsilon})(1 - \epsilon) - \epsilon$	$\frac{(1-e^{-\epsilon})(1-\epsilon)-\epsilon}{1-e^{-\epsilon}(1-\epsilon)}$
	GF	$x_{k+1} = e^{-2\epsilon}x_k - (1 - e^{-\epsilon})^2$	$-\frac{(1-e^{-\epsilon})^2}{1-e^{-2\epsilon}}$
	PG	$x_{k+1} = \frac{e^{-\epsilon}}{1+\epsilon}x_k + \frac{1-e^{-\epsilon}-\epsilon}{1+\epsilon}$	$\frac{1-e^{-\epsilon}-\epsilon}{1-e^{-\epsilon}+\epsilon}$
PG	GD	$x_{k+1} = \frac{1-\epsilon}{1+\epsilon}x_k$	0
	GF	$x_{k+1} = \frac{e^{-\epsilon}}{1+\epsilon}x_k + \frac{e^{-\epsilon}\epsilon}{1+\epsilon} - 1 + e^{-\epsilon}$	$\frac{e^{-\epsilon}\epsilon - (1-e^{-\epsilon})(1+\epsilon)}{1+\epsilon - e^{-\epsilon}}$
	PG	$x_{k+1} = \frac{1}{(1+\epsilon)^2}x_k + \frac{\epsilon^2}{(1+\epsilon)^2}$	$\frac{\epsilon}{2-\epsilon}$

 Table 2: Composite algorithms applied to $f(x) = \frac{1}{2}(x-1)^2$ and $g(x) = \frac{1}{2}(x+1)^2$ in \mathbb{R} .

F.2. Convergence rate of FB under gradient domination in Euclidean space

We review the exponential convergence rate of the Forward-Backward (FB) algorithm under gradient domination condition and partial smoothness assumptions. In this section we are working in the Euclidean case $M = \mathbb{R}^n$. The following is adapted from (Garrigos et al., 2017, Theorem 4.2.b.ii).

Lemma 8 *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be K -semiconvex and L -smooth ($KI \preceq \nabla^2 f \preceq LI$) for some $K \in \mathbb{R}$, $L > \max\{0, -K\}$, and let $g: \mathbb{R}^n \rightarrow \mathbb{R}$ be convex ($\nabla^2 g \succeq 0$). Assume $\tilde{f} = f + g$ is α -gradient dominated for some $\alpha > 0$. Consider the FB iteration*

$$x_{k+\frac{1}{2}} = x_k - \epsilon \nabla f(x_k)$$

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left\{ g(x) + \frac{1}{2\epsilon} \|x - x_{k+\frac{1}{2}}\|^2 \right\}.$$

Then for $0 < \epsilon \leq \min\{\frac{2}{L}, \frac{2}{K+L}\}$,

$$\tilde{f}(x_k) - \min \tilde{f} \leq \left(1 + \frac{\alpha\epsilon(2-\epsilon L)}{1-2\epsilon\frac{KL}{K+L}} \right)^{-k} (\tilde{f}(x_0) - \min \tilde{f}).$$

Proof By the L -smoothness of f and the convexity of g ,

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

$$g(x_{k+1}) \leq g(x_k) - \langle \nabla g(x_{k+1}), x_k - x_{k+1} \rangle.$$

Adding the two inequalities above yields

$$\tilde{f}(x_{k+1}) \leq \tilde{f}(x_k) + \langle \nabla f(x_k) + \nabla g(x_{k+1}), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2.$$

Since $x_k - \epsilon \nabla f(x_k) = x_{k+\frac{1}{2}} = x_{k+1} + \epsilon \nabla g(x_{k+1})$, we have $\nabla f(x_k) + \nabla g(x_{k+1}) = \frac{1}{\epsilon}(x_k - x_{k+1})$. Note that here the linearity \mathbb{R}^n is essential. Substituting this equality to the inequality above yields

$$\tilde{f}(x_{k+1}) \leq \tilde{f}(x_k) - \left(\frac{1}{\epsilon} - \frac{L}{2}\right) \|x_{k+1} - x_k\|^2.$$

Since f is K -semiconvex and L -smooth, we have $\|x_{k+1} - x_k\|^2 \geq (1 - 2\epsilon \frac{KL}{K+L})^{-1} \|x_{k+\frac{3}{2}} - x_{k+\frac{1}{2}}\|^2$ by Lemma 9 below, so

$$\tilde{f}(x_{k+1}) \leq \tilde{f}(x_k) - \frac{1}{\left(1 - 2\epsilon \frac{KL}{K+L}\right)} \left(\frac{1}{\epsilon} - \frac{L}{2}\right) \|x_{k+\frac{3}{2}} - x_{k+\frac{1}{2}}\|^2.$$

Furthermore, $x_{k+\frac{3}{2}} - x_{k+\frac{1}{2}} = (x_{k+1} - \epsilon \nabla f(x_{k+1})) - (x_{k+1} + \epsilon \nabla g(x_{k+1})) = -\epsilon(\nabla f(x_{k+1}) + \nabla g(x_{k+1})) = -\epsilon \nabla \tilde{f}(x_{k+1})$. Note that here the linearity of \mathbb{R}^n is also essential. Therefore,

$$\tilde{f}(x_{k+1}) \leq \tilde{f}(x_k) - \frac{\epsilon^2}{\left(1 - 2\epsilon \frac{KL}{K+L}\right)} \left(\frac{1}{\epsilon} - \frac{L}{2}\right) \|\nabla \tilde{f}(x_{k+1})\|^2.$$

Now using the α -gradient domination assumption on \tilde{f} , we get

$$\tilde{f}(x_{k+1}) \leq \tilde{f}(x_k) - \frac{2\alpha\epsilon^2}{\left(1 - 2\epsilon \frac{KL}{K+L}\right)} \left(\frac{1}{\epsilon} - \frac{L}{2}\right) (\tilde{f}(x_{k+1}) - \min \tilde{f}).$$

Collecting terms and unrolling the recursion give

$$\tilde{f}(x_k) - \min \tilde{f} \leq \left(1 + \frac{\alpha\epsilon(2 - \epsilon L)}{1 - 2\epsilon \frac{KL}{K+L}}\right)^{-k} (\tilde{f}(x_0) - \min \tilde{f})$$

as desired. ■

The proof of Lemma 8 uses the following lemma.

Lemma 9 *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be K -semiconvex and L -smooth ($KI \preceq \nabla^2 f \preceq LI$) for some $K \in \mathbb{R}$, $L > \max\{0, -K\}$. Let $x, y \in \mathbb{R}^n$, and for $\epsilon \geq 0$, let $x(\epsilon) = x - \epsilon \nabla f(x)$ and $y(\epsilon) = y - \epsilon \nabla f(y)$. Then for $0 \leq \epsilon \leq \frac{2}{K+L}$,*

$$\|x(\epsilon) - y(\epsilon)\|^2 \leq \left(1 - 2\epsilon \frac{KL}{K+L}\right) \|x - y\|^2.$$

Proof Since f is K -semiconvex and L -smooth, the function $\phi(x) = f(x) - \frac{K}{2}\|x\|^2$ is convex and $(L - K)$ -smooth. If $K = L$, then we are done. Else, by a standard property of smooth convex function (Nesterov, 2004, Theorem 2.1.5) we have $\langle \nabla \phi(x) - \nabla \phi(y), x - y \rangle \geq \frac{1}{L-K} \|\nabla \phi(x) - \nabla \phi(y)\|^2$. Since $\nabla \phi(x) = \nabla f(x) - Kx$, this is equivalent to

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{KL}{K+L} \|x - y\|^2 + \frac{1}{K+L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Note this is the same as the bound for a smooth and strongly convex function (Nesterov, 2004, Theorem 2.1.12), but here K may be negative (but not too negative, as $K + L > 0$). Then:

$$\begin{aligned} \|x(\epsilon) - y(\epsilon)\|^2 &= \|x - y - \epsilon(\nabla f(x) - \nabla f(y))\|^2 \\ &= \|x - y\|^2 - 2\epsilon\langle \nabla f(x) - \nabla f(y), x - y \rangle + \epsilon^2\|\nabla f(x) - \nabla f(y)\|^2 \\ &\leq \left(1 - 2\epsilon\frac{KL}{K+L}\right)\|x - y\|^2 - \epsilon\left(\frac{2}{K+L} - \epsilon\right)\|\nabla f(x) - \nabla f(y)\|^2. \end{aligned}$$

If $0 \leq \epsilon \leq \frac{2}{K+L}$, then the last term on the right hand side above is nonpositive, so we may drop it to get the desired result $\|x(\epsilon) - y(\epsilon)\|^2 \leq \left(1 - 2\epsilon\frac{KL}{K+L}\right)\|x - y\|^2$. \blacksquare

Appendix G. Details for §4

G.1. Forward-Backward for Langevin dynamics

Example 8 (FB for OU with Gaussian data.) Let $\nu = \mathcal{N}(\mu, \Sigma)$ as in Example 1, and $\epsilon \leq \lambda_{\min}(\Sigma)$. Let $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and let $\Sigma_0 = I$ for simplicity, so Σ_0 commutes with Σ . Along FB (13) for OU, $\rho_k = \mathcal{N}(\mu_k, \Sigma_k)$ stays Gaussian, and Σ_k commutes with Σ . Furthermore, (13) becomes:

$$\begin{aligned} x_{k+\frac{1}{2}} &= \mu + (I - \epsilon\Sigma^{-1})(x_k - \mu) \\ x_{k+1} &= \mu_{k+1} + (I - \epsilon\Sigma_{k+1}^{-1})^{-1}(x_{k+\frac{1}{2}} - \mu_{k+1}). \end{aligned}$$

Since $x_k \sim \mathcal{N}(\mu_k, \Sigma_k)$, this yields a system of updates (the second one is implicit):

$$\begin{aligned} \mu_{k+1} &= \mu + (I - \epsilon\Sigma^{-1})(\mu_k - \mu) \\ \Sigma_{k+1}(I - \epsilon\Sigma_{k+1}^{-1})^2 &= \Sigma_k(I - \epsilon\Sigma^{-1})^2. \end{aligned}$$

The mean converges exponentially fast to the correct mean: $\mu_k = \mu + (I - \epsilon\Sigma^{-1})^k(\mu_0 - \mu) \rightarrow \mu$. For the covariance, note that $\Sigma_k = \Sigma_{k+1} = \Sigma$ is the only fixed point of the update. Therefore, the FB algorithm is consistent for OU with Gaussian data.

G.2. Backward-Forward for Langevin dynamics

Example 9 (BF for OU with Gaussian data.) Let $\nu = \mathcal{N}(\mu, \Sigma)$ as in Example 1, and $\epsilon \leq \lambda_{\min}(\Sigma)$. Let $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and let $\Sigma_0 = I$ for simplicity, so Σ_0 commutes with Σ . Along BF (14) for OU, $\rho_k = \mathcal{N}(\mu_k, \Sigma_k)$ stays Gaussian, and Σ_k commutes with Σ . Furthermore, (14) becomes:

$$\begin{aligned} x_{k+\frac{1}{2}} &= \mu + (I + \epsilon\Sigma^{-1})^{-1}(x_k - \mu) \\ x_{k+1} &= \mu_k + (I + \epsilon\Sigma_k^{-1})(\mu - \mu_k) + (I + \epsilon\Sigma_k^{-1})(I - \epsilon\Sigma^{-1})^{-1}(x_k - \mu). \end{aligned}$$

Since $x_k \sim \mathcal{N}(\mu_k, \Sigma_k)$, this yields a system of updates

$$\begin{aligned} \mu_{k+1} &= \mu + (I + \epsilon\Sigma^{-1})^{-1}(\mu_k - \mu) \\ \Sigma_{k+1} &= \Sigma_k(I + \epsilon\Sigma_k^{-1})^2(I - \epsilon\Sigma^{-1})^{-2}. \end{aligned}$$

The mean converges exponentially fast to the correct mean: $\mu_k = \mu + (I + \epsilon\Sigma^{-1})^{-k}(\mu_0 - \mu) \rightarrow \mu$. For the covariance, note that $\Sigma_k = \Sigma_{k+1} = \Sigma$ is the only fixed point of the update. Therefore, the BF algorithm is consistent for OU with Gaussian data.