ACCELERATED GRADIENT FLOW FOR PROBABILITY DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents a methodology and numerical algorithms for constructing accelerated gradient flows on the space of probability distributions. In particular, we extend the recent variational formulation of accelerated gradient methods in Wibisono et al. (2016) from vector valued variables to probability distributions. The variational problem is modeled as a mean-field optimal control problem. The maximum principle of optimal control theory is used to derive Hamilton's equations for the optimal gradient flow. The Hamilton's equation are shown to achieve the accelerated form of density transport from any initial probability distribution to a target probability distribution. A quantitative estimate on the asymptotic convergence rate is provided based on a Lyapunov function construction, when the objective functional is displacement convex. Two numerical approximations are presented to implement the Hamilton's equations as a system of N interacting particles. The continuous limit of the Nesterov's algorithm is shown to be a special case with $N = 1$. The algorithm is illustrated with numerical examples.

1 INTRODUCTION

Optimization on the space of probability distributions is important to a number of machine learning models including variational inference (Blei et al., 2017), generative models (Goodfellow et al., 2014; Arjovsky et al., 2017), and policy optimization in reinforcement learning (Sutton et al., 2000). A number of recent studies have considered solution approaches to these problems based upon a construction of gradient flow on the space of probability distributions (Zhang et al., 2018; Liu & Wang, 2016; Frogner & Poggio, 2018; Chizat & Bach, 2018; Richemond & Maginnis, 2017). Such constructions are useful for convergence analysis as well as development of numerical algorithms.

In this paper, we propose a methodology and numerical algorithms that achieve accelerated gradient flows on the space of probability distributions. The proposed numerical algorithms are related to yet distinct from the accelerated stochastic gradient descent (Jain et al., 2017) and Hamiltonian Markov chain Monte-Carlo (MCMC) algorithms (Neal et al., 2011; Cheng et al., 2017). The proposed methodology extends the variational formulation of (Wibisono et al., 2016) from vector valued variables to probability distributions. The original formulation of Wibisono et al. (2016) was used to derive and analyze the convergence properties of a large class of accelerated optimization algorithms, most significant of which is the continuous-time limit of the Nesterov's algorithm (Su et al., 2014). In this paper, the limit is referred to as the Nesterov's ordinary differential equation (ODE).

The extension proposed in our work is based upon a generalization of the formula for the Lagrangian in Wibisono et al. (2016): (i) the kinetic energy term is replaced with the expected value of kinetic

energy; and (ii) the potential energy term is replaced with a suitably defined functional on the space of probability distributions. The variational problem is to obtain a trajectory in the space of probability distributions that minimizes the action integral of the Lagrangian.

The variational problem is modeled as a mean-field optimal problem. The maximum principle of the optimal control theory is used to derive the Hamilton's equations which represent the first order optimality conditions. The Hamilton's equations provide a generalization of the Nesterov's ODE to the space of probability distributions. A candidate Lyapunov function is proposed for the convergence analysis of the solution of the Hamilton's equations. In this way, quantitative estimates on convergence rate are obtained for the case when the objective functional is displacement convex (McCann, 1997). Table 1 provides a summary of the relationship between the original variational formulation in Wibisono et al. (2016) and the extension proposed in this paper.

We also consider the important special case when the objective functional is the relative entropy functional $\mathsf{D}(\rho|\rho_\infty)$ defined with respect to a target probability distribution ρ_∞. In this case, the accelerated gradient flow is shown to be related to the continuous limit of the Hamiltonian Monte-Carlo algorithm (Cheng et al., 2017) (Remark 2). The Hamilton's equations are finite-dimensional for the special case when the initial and the target probability distributions are both Gaussian. In this case, the mean evolves according to the Nesterov's ODE. For the general case, the Lyapunov function-based convergence analysis applies when the target distribution is log-concave.

As a final contribution, the proposed methodology is used to obtain a numerical algorithm. The algorithm is an interacting particle system that empirically approximates the distribution with a finite but large number of N particles. The difficult part of this construction is the approximation of the interaction term between particles. For this purpose, two types of approximations are described: (i) Gaussian approximation which is asymptotically (as $N \to \infty$) exact in Gaussian settings; and (ii) Kernel approximation which is computationally more demanding but asymptotically exact for a general class of distributions.

The outline of the remainder of this paper is as follows: Sec. 2 provides a brief review of the variational formulation in Wibisono et al. (2016). The proposed extension to the space of probability distribution appears in Sec. 3 where the main result is also described. The numerical algorithm appears in Sec. 4. The Appendix contains the details of the proof of the main result.

Notation: The gradient and divergence operators are denoted as ∇ and $\nabla\cdot$, respectively. With multiple independent variables, ∇_z denotes the gradient with respect to the variable z. The space of absolutely continuous (with respect to the Lebesgue measure) probability measures on $\mathbb{R}^d$ with finite second moments is denoted by $\mathcal{P}_{\text{ac},2}(\mathbb{R}^d)$. The Wasserstein gradient and the Gâteaux derivative of a functional $F : \mathcal{P}_{\text{ac},2}(\mathbb{R}^d) \to \mathbb{R}$ is denoted as $\nabla_\rho F(\rho)$ and $\frac{\partial F}{\partial \rho}(\rho)$ respectively (Carmona & Delarue, 2017, Ch. 5). The probability distribution of a random variable Z is denoted as $\text{Law}(Z)$.

	Vector	Probability distribution
State-space	$\mathbb{R}^d$	$\mathcal{P}_2(\mathbb{R}^d)$
Gradient flow	$\dot{X}_t = -\nabla f(X_t)$	$\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$
Lagrangian	$e^{\alpha_t+\gamma_t}\left(\frac{1}{2}\lvert e^{-\alpha_t}u\rvert^2 - e^{\beta_t}f(x)\right)$	$e^{\alpha_t+\gamma_t}\mathsf{E}\left[\frac{1}{2}\lvert e^{-\alpha_t}U\rvert^2 - e^{\beta_t}\tilde{F}(\rho, X)\right]$
Lyapunov funct.	$\frac{1}{2}\left\lvert x + e^{-\gamma_t}y - \bar{x}\right\rvert^2$ $+e^{\beta_t}(F(\rho) - F(\rho_\infty))$	$\frac{1}{2}\mathsf{E}[\lvert X_t + e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t)\rvert^2]$ $+e^{\beta_t}(F(\rho) - F(\rho_\infty))$

Table 1: Summary of the variational formulations for vectors and probability distributions.

2 REVIEW OF THE VARIATIONAL FORMULATION OF WIBISONO ET AL. (2016)

The basic problem is to minimize a C^1 smooth convex function f of d real variables: $\min_{x\in\mathbb{R}^d} f(x)$. The standard form of the gradient descent algorithm for this problem is an ODE:

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = -\nabla f(X_t), \quad t \geq 0 \tag{1}$$

Accelerated forms of this algorithm are obtained based on a variational formulation due to Wibisono et al. (2016). The formulation is briefly reviewed here using an optimal control formalism. The Lagrangian $L : \mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$L(t,x,u) := e^{\alpha_t+\gamma_t}\Bigg(\underbrace{\frac{1}{2}|e^{-\alpha_t}u|^2}_{\text{kinetic energy}} - \underbrace{e^{\beta_t}f(x)}_{\text{potential energy}}\Bigg) \tag{2}$$

where $t \geq 0$ is the time, $x \in \mathbb{R}^d$ is the state, $u \in \mathbb{R}^d$ is the velocity or control input, and the time-varying parameters $\alpha_t, \beta_t, \gamma_t$ satisfy the following scaling conditions: $\alpha_t = \log p - \log t$, $\beta_t = p\log t + \log C$, and $\gamma_t = p \log t$ where $p \geq 2$ and $C > 0$ are constants.

The variational problem is

$$\begin{aligned} \underset{u}{\text{Minimize}} \quad & J(u) = \int_0^\infty L(t, X_t, u_t)\,\mathrm{d}t \\ \text{Subject to} \quad & \frac{\mathrm{d}X_t}{\mathrm{d}t} = u_t, \quad X_0 = x_0 \end{aligned} \tag{3}$$

The Hamiltonian function is

$$H(t,x,y,u) = y \cdot u - L(t,x,u) \tag{4}$$

where $y \in \mathbb{R}^d$ is dual variable and $y \cdot u$ denotes the dot product between vectors y and u.

According to the Pontryagin's Maximum Principle, the optimal control $u_t^* = \arg\max_v H(t, X_t, Y_t, v) = e^{\alpha_t-\gamma_t}Y_t$. The resulting Hamilton's equations are

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = +\nabla_y H(t, X_t, Y_t, u_t) = e^{\alpha_t-\gamma_t}Y_t, \quad X_0 = x_0 \tag{5a}$$

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -\nabla_x H(t, X_t, Y_t, u_t) = -e^{\alpha_t+\beta_t+\gamma_t}\nabla f(X_t), \quad Y_0 = y_0 \tag{5b}$$

The system (5) is an example of accelerated gradient descent algorithm. Specifically, if the parameters $\alpha_t, \beta_t, \gamma_t$ are defined using $p = 2$, one obtains the continuous-time limit of the Nesterov's accelerated algorithm. It is referred to as the Nesterov's ODE in this paper.

For this system, a Lyapunov function is as follows:

$$V(t,x,y) = \frac{1}{2}\left|x + e^{-\gamma_t}y - \bar{x}\right|^2 + e^{\beta_t}(f(x) - f(\bar{x})) \tag{6}$$

where $\bar{x} \in \arg\min_x f(x)$. It is shown in Wibisono et al. (2016) that upon differentiating along the solution trajectory, $\frac{\mathrm{d}}{\mathrm{d}t}V(t, X_t, Y_t) \leq 0$. This yields the following convergence rate:

$$f(X_t) - f(\bar{x}) \leq O(e^{-\beta_t}), \quad \forall t \geq 0 \tag{7}$$

3 VARIATIONAL FORMULATION FOR PROBABILITY DISTRIBUTIONS

3.1 MOTIVATION AND BACKGROUND

Let $\mathsf{F} : \mathcal{P}_{\text{ac},2}(\mathbb{R}^d) \to \mathbb{R}$ be a functional on the space of probability distributions. Consider the problem of minimizing $\mathsf{F}(\rho)$. The (Wasserstein) gradient flow with respect to $\mathsf{F}(\rho)$ is

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla_\rho \mathsf{F}(\rho_t)) \tag{8}$$

where $\nabla_\rho \mathsf{F}(\rho)$ is the Wasserstein gradient of F.

An important example is the relative entropy functional where $\mathsf{F}(\rho) = \mathsf{D}(\rho|\rho_\infty) := \int_{\mathbb{R}^d} \log(\frac{\rho(x)}{\rho_\infty(x)})\rho(x)\,\mathrm{d}x$ where $\rho_\infty \in \mathcal{P}_{\text{ac},2}(\mathbb{R}^d)$ is referred to as the target distribution. The gradient of relative entropy is given by $\nabla_\rho \mathsf{F}(\rho) = \nabla \log(\frac{\rho}{\rho_\infty})$. The gradient flow

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t \nabla \log(\rho_\infty)) + \Delta \rho_t \tag{9}$$

is the Fokker-Plank equation (Jordan et al., 1998). The gradient flow achieves the density transport from an initial probability distribution ρ_0 to the target (here, also equilibrium) probability distribution ρ_∞; and underlies the construction and the analysis of Markov chain Monte-Carlo (MCMC) algorithms. The simplest MCMC algorithm is the Langevin stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad X_0 \sim \rho_0$$

where B_t is the standard Brownian motion in $\mathbb{R}^d$.

The main problem of this paper is to construct an accelerated form of the gradient flow (8). The proposed solution is based upon a variational formulation. As tabulated in Table 1, the solution represents a generalization of Wibisono et al. (2016) from its original deterministic finite-dimensional to now probabilistic infinite-dimensional settings.

The variational problem can be expressed in two equivalent forms: (i) The probabilistic form is described next in the main body of the paper; and (ii) The partial differential equation (PDE) form appears in the Appendix. The probabilistic form is stressed here because it represents a direct generalization of the Nesterov's ODE and because it is closer to the numerical algorithm.

3.2 Probabilistic form of the variational problem

Consider the stochastic process $\{X_t\}_{t\geq 0}$ that takes values in $\mathbb{R}^d$ and evolves according to:

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = U_t, \quad X_0 \sim \rho_0$$

where the control input $\{U_t\}_{t\geq 0}$ also takes values in $\mathbb{R}^d$, and $\rho_0 \in \mathcal{P}_{\text{ac},2}(\mathbb{R}^d)$ is the probability distribution of the initial condition X_0. It is noted that the randomness here comes only from the random initial condition.

Suppose the objective functional is of the form $\mathsf{F}(\rho) = \int \tilde{F}(\rho, x)\rho(x)\,\mathrm{d}x$. The Lagrangian $\mathsf{L} : \mathbb{R}^+ \times \mathbb{R}^d \times \mathcal{P}_{\text{ac},2}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathsf{L}(t, x, \rho, u) := e^{\alpha_t + \gamma_t}\Bigg(\underbrace{\frac{1}{2}|e^{-\alpha_t}u|^2}_{\text{kinetic energy}} - \underbrace{e^{\beta_t}\tilde{F}(\rho, x)}_{\text{potential energy}}\Bigg) \tag{10}$$

where the parameters $\alpha_t, \beta_t, \gamma_t$ are defined exactly the same as in the finite-dimensional case. The stochastic optimal control problem is:

$$\begin{aligned} &\text{Minimize} \quad \mathsf{J}(u) = \mathsf{E}\left[\int_0^\infty \mathsf{L}(t, X_t, \rho_t, U_t)\,\mathrm{d}t\right] \\ &\text{Subject to} \quad \frac{\mathrm{d}X_t}{\mathrm{d}t} = U_t, \quad X_0 \sim \rho_0 \end{aligned} \tag{11}$$

where $\rho_t = \text{Law}(X_t) \in \mathcal{P}_{\text{ac},2}(\mathbb{R}^d)$ is the probability density function of the random variable X_t.

The Hamiltonian function $\mathsf{H} : \mathbb{R}^+ \times \mathbb{R}^d \times \mathcal{P}_{\text{ac},2}(\mathbb{R}^d) \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for this problem is given by (Carmona & Delarue, 2017, Sec. 6.2.3):

$$\mathsf{H}(t, x, \rho, y, u) := u \cdot y - L(t, x, \rho, u) \tag{12}$$

where $y \in \mathbb{R}^d$ is the dual variable.

Remark 1. *The variational problem* (11) *is an example of a mean-field (McKean-Vlasov) optimal control problem. This is because the Lagrangian depends also upon the law of the stochastic process; cf., (Carmona & Delarue, 2017, Ch. 6).*

3.3 MAIN RESULT

Theorem 1. *Consider the variational problem* (11).

(i) *The optimal control* $U_t^* = e^{\alpha_t - \gamma_t} Y_t$ *where the optimal trajectory* $\{(X_t, Y_t)\}_{t \geq 0}$ *evolves according to the Hamilton's equations:*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = U_t^* = e^{\alpha_t - \gamma_t} Y_t, \quad X_0 \sim \rho_0 \tag{13a}$$

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -e^{\alpha_t + \beta_t + \gamma_t} \nabla_\rho \mathsf{F}(\rho_t)(X_t), \quad Y_0 = \nabla \phi_0(X_0) \tag{13b}$$

where ϕ_0 *is any convex function and* $\rho_t := \mathit{Law}(X_t)$.

(ii) *Suppose also that the functional* F *is displacement convex and* ρ_∞ *is its minimizer. Define the energy along the optimal trajectory*

$$V(t) = \frac{1}{2} \mathsf{E}[|X_t + e^{-\gamma_t} Y_t - T_{\rho_t}^{\rho_\infty}(X_t)|^2] + e^{\beta_t}(\mathsf{F}(\rho) - \mathsf{F}(\rho_\infty)) \tag{14}$$

where the map $T_{\rho_t}^{\rho_\infty} : \mathbb{R}^d \to \mathbb{R}^d$ *is the optimal transport map from* ρ_t *to* ρ_∞. *Suppose also that the following technical assumption holds:* $\mathsf{E}[(X_t + e^{-\gamma_t} Y_t - T_{\rho_t}^{\rho_\infty}(X_t)) \cdot \frac{\mathrm{d}}{\mathrm{d}t} T_{\rho_t}^{\rho_\infty}(X_t)] = 0$. *Then* $\frac{\mathrm{d}V}{\mathrm{d}t}(t) \leq 0$. *Consequently, the following rate of convergence is obtained along the optimal trajectory:*

$$\mathsf{F}(\rho_t) - \mathsf{F}(\rho_\infty) \leq O(e^{-\beta_t}), \quad \forall t \geq 0 \tag{15}$$

Proof sketch. The Hamilton's equations are derived using the standard mean-field optimal control theory Carmona & Delarue (2017). The Lyapunov function argument is based upon the variational inequality characterization of a displacement convex function (Ambrosio et al., 2008, Eq. 10.1.7). The detailed proof appears in the Appendix. We expect that the technical assumption is not necessary. This is the subject of the continuing work. □

3.4 RELATIVE ENTROPY AS THE FUNCTIONAL

In the remainder of this paper, we assume that the functional $\mathsf{F}(\rho) = \mathsf{D}(\rho|\rho_\infty)$ is the relative entropy where $\rho_\infty \in \mathcal{P}_{\mathrm{ac},2}(\mathbb{R}^d)$ is a given target probability distribution. In this case the Hamilton's equations are given by

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = e^{\alpha_t - \gamma_t} Y_t, \quad X_0 \sim \rho_0 \tag{16a}$$

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -e^{\alpha_t + \beta_t + \gamma_t} (\nabla f(X_t) + \nabla \log(\rho_t(X_t)), \quad Y_0 = \nabla \phi_0(X_0) \tag{16b}$$

where $\rho_t = \mathrm{Law}(X_t)$ and $f = -\log(\rho_\infty)$. Moreover, if f is convex (or equivalently ρ_∞ is log-concave), then F is displacement convex with the unique minimizer at ρ_∞ and the convergence estimate is given by $\mathsf{D}(\rho_t|\rho_\infty) \leq O(e^{-\beta_t})$.

Remark 2. *The Hamilton's equations* (16) *with the relative entropy functional is related to the under-damped Langevin equation (Cheng et al., 2017). The difference is that the deterministic term* $\nabla \log(\rho)$ *in* (16) *is replaced with a random Brownian motion term in the under-damped Langevin equation.*

3.5 QUADRATIC GAUSSIAN CASE

Suppose the initial distribution ρ_0 and the target distribution ρ_∞ are both Gaussian, denoted as $\mathcal{N}(m_0, \Sigma_0)$ and $\mathcal{N}(\bar{x}, Q)$, respectively. This is equivalent to the objective function $f(x)$ being quadratic of the form $f(x) = \frac{1}{2}(x - \bar{x})^\top Q^{-1}(x - \bar{x})$. Therefore, this problem is referred to as the *quadratic Gaussian case*. The following Proposition shows that the mean of the stochastic process (X_t, Y_t) evolves according to the Nesterov ODE (5):

Proposition 1. *(Quadratic Gaussian case) Consider the variational problem* (11) *for the quadratic Gaussian case. Then*

(i) *The stochastic process (X_t, Y_t) is Gaussian. The Hamilton's equations are given by:*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = e^{\alpha_t - \gamma_t} Y_t,$$
$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -e^{\alpha_t + \beta_t + \gamma_t}(Q^{-1}(X_t - \bar{x}) - \Sigma_t^{-1}(X_t - m_t))$$

where m_t and Σ_t are the mean and the covariance of X_t.

(ii) *Upon taking the expectation of both sides, and denoting $n_t := \mathsf{E}[Y_t]$*

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = e^{\alpha_t - \gamma_t} n_t,$$
$$\frac{\mathrm{d}n_t}{\mathrm{d}t} = -e^{\alpha_t + \beta_t + \gamma_t} \underbrace{Q^{-1}(m_t - \bar{x})}_{\nabla f(m_t)}$$

which is identical to Nesterov ODE (5)*.*

Proof sketch. Fix ρ_t. Consider the resulting pair (X_t, Y_t) from (23) and let $\tilde{\rho}_t = \mathrm{Law}(X_t)$. The proof follows by observing that ρ_t Gaussian is a fixed-point of the map $\rho_t \mapsto \tilde{\rho}_t$. □

4 NUMERICAL ALGORITHM

The proposed numerical algorithm is based upon an interacting particle implementation of the Hamilton's equation (16). Consider a system of N particles $\{(X_t^i, Y_t^i)\}_{i=1}^N$ that evolve according to:

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = e^{\alpha_t - \gamma_t} Y_t^i, \quad X_0^i \sim \rho_0$$
$$\frac{\mathrm{d}Y_t^i}{\mathrm{d}t} = -e^{\alpha_t + \beta_t + \gamma_t}(\nabla f(X_t^i) + \underbrace{I_t^{(N)}(X_t^i)}_{\text{interaction term}}), \quad Y_0^i = \nabla \phi_0(X_0^i)$$

The interaction term $I_t^{(N)}$ is an empirical approximation of the $\nabla \log(\rho_t)$ term in (16). We propose two types of empirical approximations as follows:

1. Gaussian approximation: Suppose the density is approximated as a Gaussian $\mathcal{N}(m_t, \Sigma_t)$. In this case, $\nabla \log(\rho_t(x)) = -{\Sigma_t}^{-1}(x - m_t)$. This motivates the following empirical approximation of the interaction term:

$$I_t^{(N)}(x) = -{\Sigma_t^{(N)}}^{-1}(x - m_t^{(N)}) \tag{18}$$

where $m_t^{(N)} := N^{-1}\sum_{i=1}^N X_t^i$ is the empirical mean and $\Sigma_t^{(N)} := \frac{1}{N-1}\sum_{i=1}^N (X_t^i - m_t^{(N)})(X_t^i - m_t^{(N)})^\top$ is the empirical covariance.

Algorithm 1 Interacting particle implementation of the accelerated gradient flow

Input: ρ_0, ϕ_0, N, t_0, Δt, p, C, K
Output: $\{X_k^i\}_{i=1,k=0}^{N,K}$
  Initialize $\{X_0^i\}_{i=1}^N \overset{\text{i.i.d}}{\sim} \rho_0$, $Y_0^i = \nabla\phi_0(X_0^i)$
  Compute $I_0^{(N)}(X_0^i)$ with (18) or (19)
  for $k = 0$ to $K-1$ **do**
    $t_{k+\frac{1}{2}} = t_k + \frac{1}{2}\Delta t$
    $Y_{k+\frac{1}{2}}^i = Y_k^i - \frac{1}{2}Cpt_{k+\frac{1}{2}}^{2p-1}(\nabla f(X_k^i) + I_k^{(N)}(X_k^i))\Delta t$
    $X_{k+1}^i = X_k^i + \frac{p}{t_{k+\frac{1}{2}}^{p+1}} Y_k^i \Delta t$
    Compute $I_{k+1}^{(N)}(X_{k+1}^i)$ with (18)or (19)
    $Y_{k+1}^i = Y_{k+\frac{1}{2}}^i - \frac{1}{2}Cpt_{k+\frac{1}{2}}^{2p-1}(\nabla f(X_{k+1}^i) + I_{k+1}^{(N)}(X_{k+1}^i))\Delta t$
    $t_{k+1} = t_{k+\frac{1}{2}} + \frac{1}{2}\Delta t$
  end for

Even though the approximation is asymptotically (as $N \to \infty$) exact only under the Gaussian assumption, it may be used in a more general settings, particularly when the density ρ_t is unimodal. The situation is analogous to the (Bayesian) filtering problem, where an ensemble Kalman filter is used as an approximate solution for non-Gaussian distributions (Evensen, 2003).

2. Kernel approximation: This is based upon the so-called kernel-approximation of the weighted Laplacian operator (Coifman & Lafon, 2006; Hein et al., 2007). For a C^2 function f, the weighted Laplacian is defined as $\Delta_\rho f := \frac{1}{\rho}\nabla\cdot(\rho\nabla f)$. Denote $e(x) = x$ as the coordinate function on $\mathbb{R}^d$. It is a straightforward calculation to show that $\nabla\log(\rho) = \Delta_\rho e$. This allows one to use the kernel-approximation of the weighted Laplacian to approximate the interaction term as follows:

$$I_t^{(N)}(X_t^i) = \frac{1}{\epsilon}\frac{\sum_{j=1}^N k_\epsilon(X_t^i, X_t^j)(X^j - X^i)}{\sum_{j=1}^N k_\epsilon(X_t^i, X_t^j)} \tag{19}$$

where the kernel $k_\epsilon(x,y) = \frac{g_\epsilon(x,y)}{\sqrt{\sum_{i=1}^N g_\epsilon(y,X^i)}}$ is constructed empirically in terms of the Gaussian kernel $g_\epsilon(x,y) = \exp(-|x-y|^2/(4\epsilon))$. The positive parameter ϵ is referred to as the kernel bandwidth. It is known that the approximation is asymptotically exact as $\epsilon \downarrow 0$ and $N \to \infty$ (Hein et al., 2007).

The resulting interacting particle algorithm with a Gaussian or kernel approximation of the interaction term is tabulated in Table 1. The symplectic method proposed in (Betancourt et al., 2018) is used to carry out the numerical integration. The algorithm is applied to two examples as described in the following sections.

Remark 3. *For the case where there is only one particle ($N = 1$), the interaction term is zero and the system* (17) *reduces to the Nesterov ODE* (5)*.*

4.1 Gaussian Example

Consider the Gaussian example as described in Sec. 3.5. The simulation results for the scalar ($d = 1$) case with initial distribution $\rho_0 = \mathcal{N}(2,4)$ and target distribution $\mathcal{N}(\bar{x}, Q)$ where $\bar{x} = -5.0$ and $Q = 0.25$ is depicted in Figure 1-(a)-(b). For this simulation, the numerical parameters are as follows: $N = 100$, $\phi_0(x) = 0.5(x-2)$, $t_0 = 1$, $\Delta t = 0.1$, $p = 2$,$C = 0.625$, and $K = 400$. The numerical result illustrates the convergence result for the case of Gaussian target distribution.

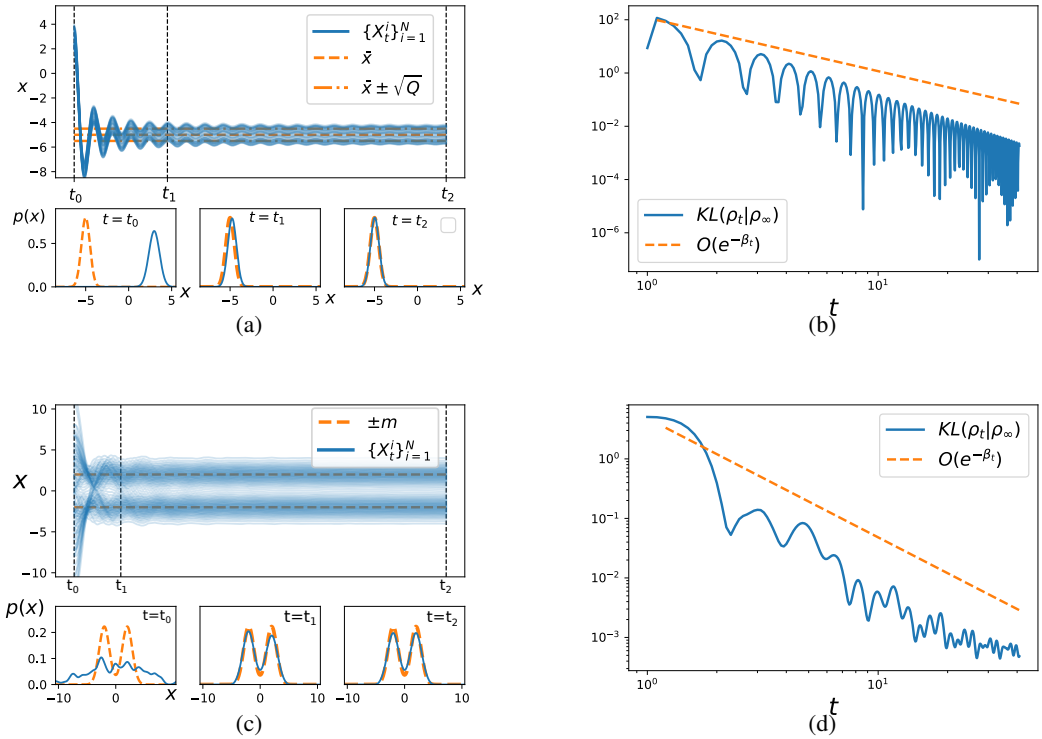

Figure 1: Simulation result for the Gaussian case (Example 4.1): (a) The time traces of the particles; (b) The KL-divergence as a function of time. Simulation result for the non-Gaussian case (Example 4.2): (c) The time traces of the particles; (d) The KL-divergence as a function of time.

4.2 Non-Gaussian example

This example involves a non-Gaussian target distribution $\rho_\infty = \frac{1}{2}\mathcal{N}(-m, \sigma^2) + \frac{1}{2}\mathcal{N}(m, \sigma^2)$ which is a mixture of two one-dimensional Gaussians where $m = 2.0$ and $\sigma^2 = 0.8$. The simulation results are depicted in Figure 1-(c)-(d). The numerical parameters are same as in the Example 4.1. The interaction term is approximated using a kernel approximation with $\epsilon = 0.01$. The numerical result illustrates the performance of the kernel-approximation algorithm for non-Gaussian distributions and shows the convergence result for the case of non-Gaussian target distribution.

5 Conclusion

The main contribution of this paper is to extend the variational formulation of Wibisono et al. (2016) to obtain theoretical results and numerical algorithms for accelerated gradient flow in the space of probability distributions. In continuous-time settings, bounds on convergence rate are derived based on a Lyapunov function argument. Two numerical algorithms based upon an interacting particle representation are presented and illustrated with examples. Analysis of these algorithms in finite-N and finite Δt is a subject of continuing investigation. As has been the case in finite-dimensional settings, the theoretical framework is expected to be useful in this regard.

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.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.

Michael Betancourt, Michael I Jordan, and Ashia C Wilson. On symplectic optimization. arXiv preprint arXiv:1802.03653, 2018.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518):859–877, 2017.

Rene Carmona and François Delarue. Probabilistic Theory of Mean Field Games with Applications I-II. Springer, 2017.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. arXiv preprint arXiv:1707.03663, 2017.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. arXiv preprint arXiv:1805.09545, 2018.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. Applied and computational harmonic analysis, 21(1):5–30, 2006.

Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. Ocean dynamics, 53(4):343–367, 2003.

Charlie Frogner and Tomaso Poggio. Approximate inference with wasserstein gradient flows. arXiv preprint arXiv:1806.04542, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672–2680, 2014.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. Journal of Machine Learning Research, 8(Jun):1325–1368, 2007.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent. arXiv preprint arXiv:1704.08227, 2017.

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

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In Advances In Neural Information Processing Systems, pp. 2378–2386, 2016.

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

Radford M Neal et al. Mcmc using hamiltonian dynamics. Handbook of Markov Chain Monte Carlo, 2(11):2, 2011.

Pierre H Richemond and Brendan Maginnis. On wasserstein reinforcement learning and the fokker-planck equation. arXiv preprint arXiv:1712.07185, 2017.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In Advances in Neural Information Processing Systems, pp. 2510–2518, 2014.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems, pp. 1057–1063, 2000.

Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences, pp. 201614734, 2016.

Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy optimization as wasserstein gradient flows. arXiv preprint arXiv:1808.03030, 2018.

A PDE FORMULATION OF THE VARIATIONAL PROBLEM

An equivalent pde formulation is obtained by considering the stochastic optimal control problem (11) as a deterministic optimal control problem on the space of the probability distributions. Specifically, the process $\{\rho_t\}_{t\geq 0}$ is a deterministic process that takes values in $\mathcal{P}_{\text{ac},2}(\mathbb{R}^d)$ and evolves according to the continuity equation

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t u_t)$$

where $u_t : \mathbb{R}^d \to \mathbb{R}^d$ is now a time-varying vector field. The Lagrangian $\mathcal{L} : \mathbb{R}^+ \times \mathcal{P}_{\text{ac},2}(\mathbb{R}^d) \times L^2(\mathbb{R}^d; \mathbb{R}^d) \to \mathbb{R}$ is defined as:

$$\mathcal{L}(t, \rho, u) := e^{\alpha_t + \gamma_t} \left[\int_{\mathbb{R}^d} \frac{1}{2} |e^{-\alpha_t} u(x)|^2 \rho(x) \, \mathrm{d}x - e^{\beta_t} F(\rho) \right] \tag{20}$$

The optimal control problem is:

$$\begin{aligned} \text{Minimize} \quad & \int_0^\infty \mathcal{L}(t, \rho_t, u_t) \, \mathrm{d}t \\ \text{Subject to} \quad & \frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t u_t) = 0 \end{aligned} \tag{21}$$

The Hamiltonian function $\mathcal{H} : \mathbb{R}^+ \times \mathcal{P}_{\text{ac},2}(\mathbb{R}^d) \times \mathcal{C}(\mathbb{R}^d; \mathbb{R}) \times L^2(\mathbb{R}^d; \mathbb{R}^d) \to \mathbb{R}$ is

$$\mathcal{H}(t, \rho, \phi, u) := \langle \nabla \phi, u \rangle_{L^2(\rho)} - \mathcal{L}(t, \rho, u) \tag{22}$$

where $\phi \in \mathcal{C}(\mathbb{R}^d; \mathbb{R})$ is the dual variable.

B RESTATEMENT OF THE MAIN RESULT AND ITS PROOF

We restate Theorem 1 below which now includes the pde formulation as well.

Theorem 2. *Consider the variational problem* (11)-(21).

(i) *For the probabilistic form* (11) *of the variational problem, the optimal control* $U_t^* = e^{\alpha_t - \gamma_t} Y_t$*, where the optimal trajectory* $\{(X_t, Y_t)\}_{t\geq 0}$ *evolves according to the Hamilton's odes:*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = U_t^* = e^{\alpha_t - \gamma_t} Y_t, \quad X_0 \sim \rho_0 \tag{23a}$$

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -e^{\alpha_t + \beta_t + \gamma_t} \nabla_\rho F(\rho_t)(X_t), \quad Y_0 = \nabla \phi_0(X_0) \tag{23b}$$

where ϕ_0 *is a convex function, and* $\rho_t = \mathit{Law}(X_t)$*.*

(ii) *For the pde form* (21) *of the variational problem, the optimal control is* $u_t^* = e^{\alpha_t - \gamma_t} \nabla \phi_t(x)$*, where the optimal trajectory* $\{(\rho_t, \phi_t)\}_{t\geq 0}$ *evolves according to the Hamilton's pdes:*

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t \underbrace{e^{\alpha_t - \gamma_t} \nabla \phi_t}_{u_t^*}), \quad \textit{initial condn.} \;\; \rho_0 \tag{24a}$$

$$\frac{\partial \phi_t}{\partial t} = -e^{\alpha_t - \gamma_t} \frac{|\nabla \phi_t|^2}{2} - e^{\alpha_t + \gamma_t + \beta_t} \nabla_\rho F(\rho) \tag{24b}$$

(iii) *The solutions of the two forms are equivalent in the following sense:*

$$\mathit{Law}(X_t) = \rho_t, \quad U_t = u_t(X_t), \quad Y_t = \nabla \phi_t(X_t)$$

(iv) *Suppose additionally that the functional* F *is displacement convex and* ρ_∞ *is its minimizer. Define*

$$V(t) = \frac{1}{2} \mathsf{E}(|X_t + e^{-\gamma_t} Y_t - T_{\rho_t}^{\rho_\infty}(X_t)|^2) + e^{\beta_t}(F(\rho) - F(\rho_\infty)) \tag{25}$$

where the map $T_{\rho_t}^{\rho_\infty} : \mathbb{R}^d \to \mathbb{R}^d$ *is the optimal transport map from* ρ_t *to* ρ_∞*. Suppose also that the following technical assumption holds:* $\mathsf{E}[(X_t + e^{-\gamma_t} Y_t - T_{\rho_t}^{\rho_\infty}(X_t)) \cdot \frac{\mathrm{d}}{\mathrm{d}t} T_{\rho_t}^{\rho_\infty}(X_t)] = 0$*. Then* $\frac{\mathrm{d}V}{\mathrm{d}t}(t) \leq 0$*. Consequently, the following rate of convergence is obtained along the optimal trajectory*

$$F(\rho_t) - F(\rho_\infty) \leq O(e^{-\beta_t}), \quad \forall t \geq 0$$

Proof. (i) The Hamiltonian function defined in (12) is equal to

$$H(t, x, \rho, y, u) = y \cdot u - e^{\gamma_t - \alpha_t} \frac{1}{2} |u|^2 + e^{\alpha_t + \gamma_t \beta_t} \tilde{F}(\rho, x)$$

after inserting the formula for the Lagrangian. According to the maximum principle in probabilistic form for (mean-field) optimal control problems (see (Carmona & Delarue, 2017, Sec. 6.2.3)), the optimal control law $U_t^* = \arg\min_v H(t, X_t, \rho_t, Y_t, v) = e^{\alpha_t - \gamma_t} Y_t$ and the Hamilton's equations are

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = +\nabla_y H(t, X_t, \rho_t, Y_t, U_t^*) = U_t^* = e^{\alpha_t - \gamma_t} Y_t$$

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -\nabla_x H(t, X_t, \rho_t, Y_t, U_t^*) - \tilde{\mathsf{E}}[\nabla_\rho H(t, \tilde{X}_t, \rho_t, \tilde{Y}_t, \tilde{U}_t^*)(X_t)]$$

where $\tilde{X}_t, \tilde{Y}_t, \tilde{U}_t^*$ are independent copies of X_t, Y_t, U_t^*. The derivatives

$$\nabla_x H(t, x, \rho, y, u) = e^{\alpha_t + \beta_t + \gamma_t} \nabla_x \tilde{F}(\rho, x)$$

$$\nabla_\rho H(t, x, \rho, y, u) = e^{\alpha_t + \beta_t + \gamma_t} \nabla_\rho \tilde{F}(\rho, x)$$

It follows that

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -e^{\alpha_t + \beta_t + \gamma_t} \left(\nabla_x \tilde{F}(\rho_t, X_t) + \tilde{\mathsf{E}}[\nabla_\rho \tilde{F}(\rho_t, \tilde{X}_t)(X_t)] \right) = -e^{\alpha_t + \beta_t + \gamma_t} \nabla_\rho F(\rho)(X_t)$$

where we used the definition $F(\rho) = \int \tilde{F}(x, \rho) \rho(x) \, \mathrm{d}x$ and the identity (Carmona & Delarue, 2017, Sec. 5.2.2 Example 3)

$$\nabla_\rho F(\rho)(x) = \nabla_x \tilde{F}(\rho, x) + \int \nabla_\rho \tilde{F}(\rho, \tilde{x})(x) \rho(\tilde{x}) \, \mathrm{d}\tilde{x}$$

(ii) The Hamiltonian function defined in (22) is equal to

$$H(t,\rho,\phi,u) = \int \left[\nabla\phi(x)\cdot u(x) - \frac{1}{2}e^{\gamma_t-\alpha_t}|u(x)|^2\right]\rho(x)\,\mathrm{d}x + e^{\alpha_t+\gamma_t+\beta_t}F(\rho)$$

after inserting the formula for the Lagrangian. According to the maximum principle for pde formulation of mean-field optimal control problems (see (Carmona & Delarue, 2017, Sec. 6.2.4)) the optimal control vector field is $u_t^* = \arg\min_v \mathsf{H}(t,\rho_t,\phi_t,v) = e^{\alpha_t-\gamma_t}\nabla\phi_t$ and the Hamilton's equations are:

$$\begin{aligned}\frac{\partial\rho_t}{\partial t} &= +\frac{\partial\mathcal{H}}{\partial\phi}(t,\rho_t,\phi_t,u_t) = -\nabla\cdot(\rho_t\nabla u_t^*)\\ \frac{\partial\phi_t}{\partial t} &= -\frac{\partial\mathcal{H}}{\partial\rho}(t,\rho_t,\phi_t,u_t) = -(\nabla\phi\cdot u^* - e^{\gamma_t-\alpha_t}\frac{1}{2}|u_t^*|^2 + e^{\alpha_t+\gamma_t+\beta_t}\frac{\partial F}{\partial\rho}(\rho_t))\end{aligned}$$

inserting the formula $u_t^* = e^{\alpha_t-\gamma_t}\nabla\phi_t$ concludes the result.

(iii) Consider the (ρ_t,ϕ_t) defined from (24). The distribution ρ_t is identified with a stochastic process $\tilde{X}_t$ such that $\frac{\mathrm{d}\tilde{X}_t}{\mathrm{d}t} = e^{\alpha_t-\gamma_t}\nabla\phi_t(\tilde{X}_t)$ and $\mathrm{Law}(\tilde{X}_t) = \rho_t$. Then define $\tilde{Y}_t = \nabla\phi_t(\tilde{X}_t)$. Taking the time derivative shows that

$$\begin{aligned}\frac{\mathrm{d}\tilde{Y}_t}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi_t(\tilde{X}_t) = \nabla^2\phi_t(\tilde{X}_t)\frac{\mathrm{d}\tilde{X}_t}{\mathrm{d}t} + \nabla\frac{\partial\phi_t}{\partial t}(X_t)\\ &= e^{\alpha_t-\gamma_t}\nabla^2\phi_t(\tilde{X}_t)\nabla\phi_t(\tilde{X}_t) - e^{\alpha_t-\gamma_t}\nabla^2\phi_t(\tilde{X}_t)\nabla\phi_t(X_t) - e^{\alpha_t+\beta_t+\gamma_t}\nabla\frac{\partial F}{\partial\rho}(\rho_t)(\tilde{X}_t)\\ &= -e^{\alpha_t+\beta_t+\gamma_t}\nabla\frac{\partial F}{\partial\rho}(\rho_t)(\tilde{X}_t)\\ &= -e^{\alpha_t+\beta_t+\gamma_t}\nabla_\rho F(\rho_t)(\tilde{X}_t)\end{aligned}$$

with the initial condition $\tilde{Y}_0 = \nabla\phi_0(\tilde{X}_0)$, where we used the identity $\nabla\frac{\partial F}{\partial\rho}(\rho) = \nabla_\rho F(\rho)$ (Carmona & Delarue, 2017, Prop. 5.48). Therefore the equations for $\tilde{X}_t$ and $\tilde{Y}_t$ are identical. Hence one can identify (X_t,Y_t) with $(\tilde{X}_t,\tilde{Y}_t)$.

(iv) The energy functional

$$V(t) = \underbrace{\frac{1}{2}\mathsf{E}\left[|X_t + e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t)|^2\right]}_{\text{first term}} + \underbrace{e^{\beta_t}(F(\rho)-F(\rho_\infty))}_{\text{second term}}$$

Then the derivative of the first term is

$$\mathsf{E}\left[(X_t + e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t))\cdot(e^{\alpha_t-\gamma_t}Y_t - \dot{\gamma}_t e^{-\gamma_t}Y_t - e^{\alpha_t+\beta_t}\nabla_\rho F(\rho_t)(X_t) + \xi(T_{\rho_t}^{\rho_\infty}(X_t)))\right]$$

where $\xi(T_{\rho_t}^{\rho_\infty}(X_t)) := \frac{\mathrm{d}}{\mathrm{d}t}T_{\rho_t}^{\rho_\infty}(X_t)$. Using the scaling condition $\dot{\gamma}_t = e^{\alpha_t}$ the derivative of the first term simplifies to

$$\mathsf{E}\left[(X_t + e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t))\cdot(-e^{\alpha_t+\beta_t}\nabla_\rho F(\rho_t)(X_t) + \xi(T_{\rho_t}^{\rho_\infty}(X_t)))\right]$$

Upon using the technical assumption, $\mathsf{E}[(X_t + e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t))\cdot\xi(T_{\rho_t}^{\rho_\infty}(X_t))] = 0$ the derivative of the first term simplifies to

$$\mathsf{E}\left[(X_t + e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t))\cdot(-e^{\alpha_t+\beta_t}\nabla_\rho F(\rho_t)(X_t))\right]$$

The derivative of the second term is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\text{second term}) &= \dot{\beta}_t e^{\beta_t}(F(\rho_t)-F(\rho_\infty)) + e^{\beta_t}\frac{\mathrm{d}}{\mathrm{d}t}F(\rho_t)\\ &= e^{\alpha_t+\beta_t}(F(\rho_t)-F(\rho_\infty)) + e^{\beta_t}\mathsf{E}[\nabla_\rho F(\rho_t)(X_t)e^{\alpha_t-\gamma_t}Y_t]\end{aligned}$$

where we used the scaling condition $\dot{\beta}_t = e^{\alpha_t}$ and the chain-rule for the Wasserstein gradient (Ambrosio et al., 2008, Ch. 10, E. Chain rule). Adding the derivative of the first and second term yields:

$$\frac{\mathrm{d}V}{\mathrm{d}t}(t) = e^{\alpha_t+\beta_t}\left(F(\rho_t) - F(\rho_\infty) - \mathsf{E}\left[(X_t - T_{\rho_t}^{\rho_\infty}(X_t)) \cdot \nabla_\rho F(\rho_t)(X_t)\right]\right)$$

which is negative by variational inequality characterization of the displacement convex function $F(\rho)$ (Ambrosio et al., 2008, Eq. 10.1.7).

We expect that the technical assumption can be removed. This is the subject of the continuing work.

□