ACCELERATED WASSERSTEIN-2 GRADIENT FLOW WITH STEERING

WILLIAM WANG

Abstract. We extend upon previous works of accelerated gradient flows (with momentum) in the Wasserstein-2 metric. The new flow imitates a gradient-based descent method used for structural relaxation and adds an addition "steering" term to the momentum updates. We provide simple numerical examples which show competitive convergence rates with this new method.

1. Introduction

In optimization we typically seek the solution to

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

where f is a convex function from $\Omega \to \mathbb{R}$. The most popular method, gradient descent, can be modelled by the following trajectory (discrete and in the continuous limit):

$$\begin{aligned} x_{n+1} &= \operatorname*{argmin}_x \{f(x) + \frac{1}{2\eta}\text{dist}(x, x_n)\} \\ \frac{d}{dt}x &= -\nabla f(x). \end{aligned} \tag{1}$$

However, convergence with gradient descent is typically slow near minima due to a vanishing gradient. Several augmentations to gradient descent have been proposed which may improve the convergence rates, the most notable being momentum-based approaches (e.g., Nesterov's algorithm). These methods introduce a velocity/momentum term which is updated by the gradient, and gradually damped so that a stationary point is eventually reached. We will show in Section 2 that accelerated gradient flows can be formulated as the superposition of Hamiltonian flow and a damped flow.

In this work we consider optimization of functionals over the space of probability distributions (with finite second order moments):

$$\min_{\mu\in\mathcal{P}_2(\Omega)} F(\mu),$$

where $F : \mathcal{P}_2(\Omega) \to \mathbb{R}$ is a functional over the space of probability distributions. We extend several gradient-based optimization methods with the gradient flows of metric spaces, considering the convergence rates of these approaches.

FIRE Minimization. [1] proposes a novel minimization algorithm for local atomic structure optimization, known as FIRE. The algorithm utilizes only gradient information, but has been shown to be competitive with conjugate gradient and quasi-Newton methods. The authors of FIRE recommend the following equation of motion, analogous to a "blind skier" search:

$$
\begin{aligned}
\frac{d}{dt}\vec{q} &= \vec{p}(t) \\
\frac{d}{dt}\vec{p} &= \frac{\vec{F(t)}}{m} - \gamma(t)|\vec{p}(t)|(\hat{p}(t) - \hat{F}(t))
\end{aligned}
\tag{2}
$$

where $\gamma(t)$ is a damping factor, $\vec{F} = -\nabla f(x)$, and mass m (we will consider only $m = 1$). In comparison to standard accelerated gradient-descent approaches, FIRE not only includes damping but also "steering," which introduces new dynamics as compared to the flow of these accelerated gradient methods (see Section 2).

2. Hamiltonian Dynamics of Optimization Methods

As shown in [2], the class of accelerated gradient descent methods (e.g., Nesterov's method [3]) can be modeled by the combination of Hamiltonian dynamics with a damping field.

Review of Hamiltonian Dynamics. Hamilton's principle, or the principle of least action, describes the law governing the dynamics of mechanical systems. The Lagrangian is commonly used, coupled with the Euler-Lagrange equations, to describe the equations of motion.

Similarly, we can describe the same dynamics through Hamiltonian mechanics. Let $\vec{q} = \{q_i\}_{i=1}^N$ denote coordinate/position variables, and $\vec{p} = \{p_i\}_{i=1}^N$ represent the corresponding momentum. The dynamics are governed by the following system:

$$
\frac{d\vec{q}}{dt} = \frac{\partial \mathcal{H}}{\partial \vec{p}}, \quad \frac{d\vec{p}}{dt} = -\frac{\partial \mathcal{H}}{\partial \vec{p}}. \tag{3}
$$

Assuming that the kinetic energy T has a homogeneous quadratic dependence and the potential energy V is independent of velocity, the Hamiltonian is equivalent to the total energy of the system; i.e., $\mathcal{H}(\vec{q}, \vec{p}, t) = \mathcal{H}(\vec{q}, \vec{p}) = T(\vec{p}) + V(\vec{q})$. In the case of optimization, we set V to be the function we with to minimize over (i.e., $V(x) = f(x)$).

By conservation of energy, starting with any given $(\vec{q_0}, \vec{p_0})$, the dynamics must evolve along a fixed level set, for which $\mathcal{H}(\vec{q}, \vec{p}) = \mathcal{H}(\vec{q_0}, \vec{p_0})$. That is, the trajectory is along a fixed subset in (p, q) phase space.

Damped Hamiltonian Flow. In optimization, we require that the solution reaches a stationary/fixed point, which is generally incompatible with systems lacking dissipative forces. For example, a simple harmonic oscillator without damping has a maximum displacement that is invariant with time.

A system with damping/dissipative forces cannot be described by a Hamiltonian dynamics with a Hamiltonian which equals the system's total energy; for example, the following equation of motion:

$$
\frac{d^2 q}{dt^2} + \gamma(t)\frac{dq}{dt} + \frac{dV(q)}{dq} = 0. \tag{4}
$$

We manually introducing a dissipation field (as in [2]); we extend upon this idea by also adding velocity modifications.

Dissipation Field and Steering. Following [2] and [4], we introduce a dissipation field which affects the equation of motion for momentum:

$$\frac{d}{dt}\begin{bmatrix}\vec{q}\\ \vec{p}\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}\nabla_{\vec{p}}\mathcal{H}\\ \nabla_{\vec{q}}\mathcal{H}\end{bmatrix} - \begin{bmatrix}0\\ \gamma(t)\vec{p}\end{bmatrix}. \tag{5}$$

This combines the Hamiltonian flow with a dissipative force, ensuring that $\|\vec{p}\| \to 0$ as $t \to \infty$. Notice that without damping and with a kinetic term that is quadratic in momentum, Equation 5 reduces to the velocity updating under traditional gradient descent.

Figure 1 represents the flow fields for the right-hand side of this equation for a particle with a one-dimensional quadratic kinetic and potential energy: $T(p,q) = \frac{p^2}{2m}$ and $V(q) = \frac{1}{2}q^2$.

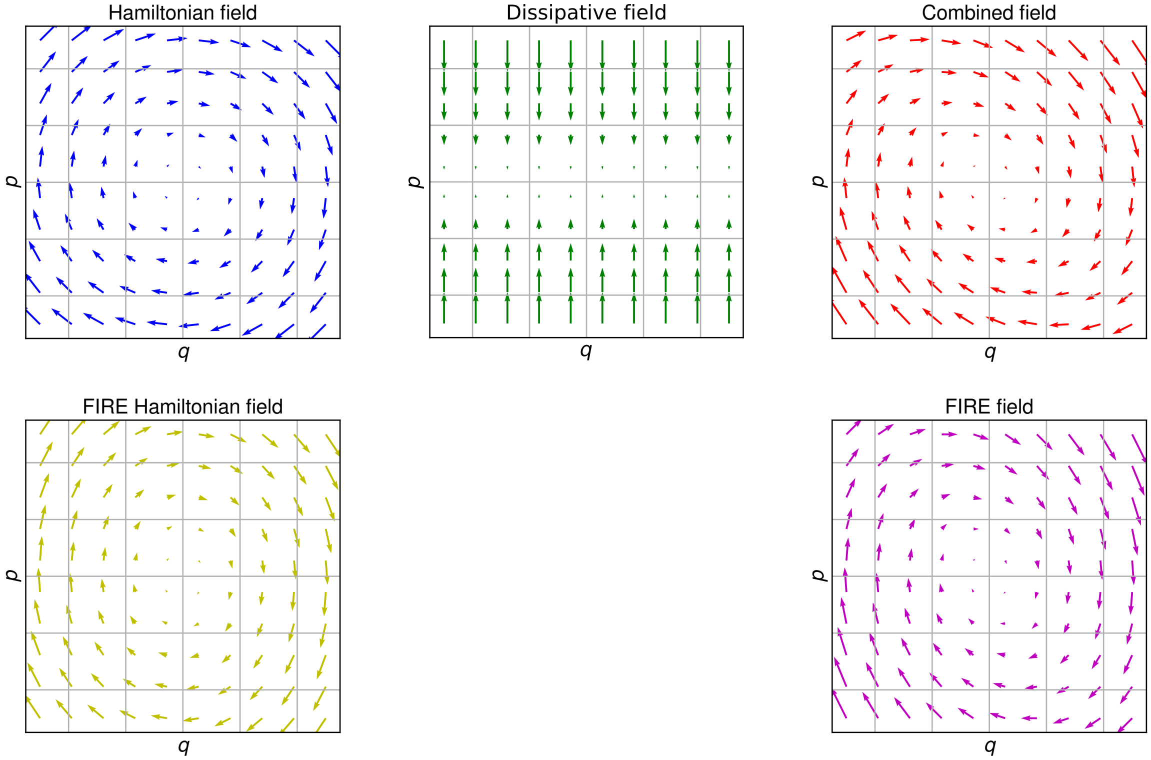

Figure 1. **Left to right: Hamiltonian, dissipative flow, accelerated Hamiltonian flow, FIRE-Hamiltonian (no dissipation), and FIRE vector fields for a particle with quadratic potential.** Position q on the horizontal axis and momentum p on the vertical axis, damping is $\gamma(t) = \frac{1}{2}$.

We consider the equation of motion prescribed by FIRE (Equation 2), and augment the flow in Equation 5 to mimic the motion by introducing an additional velocity modification (i.e., steering):

$$\frac{d}{dt}\begin{bmatrix}\vec{q}\\ \vec{p}\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & -(1+\gamma(t)\frac{\|\vec{p}\|}{\|\nabla_{\vec{q}}\mathcal{H}\|})\end{bmatrix}\begin{bmatrix}\nabla_{\vec{p}}\mathcal{H}\\ \nabla_{\vec{q}}\mathcal{H}\end{bmatrix} - \begin{bmatrix}0\\ \gamma(t)\vec{p}\end{bmatrix}. \tag{6}$$

Equations 5 and 6 define our the setup for accelerated gradient flow and accelerated gradient flow with steering, which we generalize to descent methods over the space of probability distributions.

3. Accelerated Gradient Flow

Hamiltonian with Wasserstein Gradient Flow. Rather than considering the equation of motion for a particle, we generalize to probability distributions, replacing $\vec{q}$ with ρ (position) and p with ν (velocity), and update the functional we wish to minimize over $V : \mathcal{P}_2(\Omega) \to \mathbb{R}$. That is, Equation 5 now is updated to:

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho \\ \nu\end{bmatrix} = \begin{bmatrix}1 & 0 \\ 0 & -1\end{bmatrix}\begin{bmatrix}\frac{\delta}{\delta \nu}\mathcal{H} \\ \frac{\delta}{\delta \rho}\mathcal{H}\end{bmatrix} - \begin{bmatrix}0 \\ \gamma(t)\nu\end{bmatrix}. \tag{7}$$

We consider Equation 7 only in the Wasserstein metric, though this can be extended to various other metrics. The Wasserstein Gradient Flow, with respect to some functional $F : \mathcal{P}_2(\Omega) \to \mathbb{R}$, satisfies:

$$\frac{\partial}{\partial t}\rho = \nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho}\right), \tag{8}$$

where $\nabla \frac{\delta F}{\delta \rho}$ is the Wasserstein gradient and $\frac{\delta F}{\delta \rho}$ can be found through a small variation, considering $F(\rho + \epsilon\xi)$ for some $\xi \in \mathcal{P}_2(\Omega)$ [5].

We adapt the Hamiltonian flow by finding the kinetic term of the Hamiltonian in terms of ν. We posit $\mathcal{H}(\rho, \nu) = T(\rho, \nu) + V(\rho)$ and that one should recover updates according to the Wasserstein Gradient Flow; i.e.,

$$\frac{\partial}{\partial t}\rho = \frac{\delta \mathcal{H}}{\delta \nu} = \frac{\delta T}{\delta \nu} := -\nabla \cdot (\rho \nabla \nu),$$

implying that the kinetic term has the form:

$$T(\rho, \nu) := -\frac{1}{2}\int (\nabla \cdot (\rho \nabla \nu))\, \mathrm{d}\nu. \tag{9}$$

Using this new kinetic term results in the following Hamiltonian dynamics for ν:

$$\frac{\delta \mathcal{H}}{\delta \rho} = \frac{\delta T}{\delta \rho} + \frac{\delta V}{\delta \rho} = \frac{1}{2}(\nabla \nu)^2 + \frac{\delta V}{\delta \rho} \tag{10}$$

see Appendix A for details of the proof.

Incorporating Steering. We are not aware of a suitable norm for probability spaces (since a metric does necessarily induce a norm), so to adapt Equation 6 for gradient flows, we use a "norm" which only considers the square-root of the second-moment of a distribution. Note that this does not meet all the requirements of a norm, but the goal is to approximate the behavior of Equation 6. One interpretation of the dynamics is:

$$\begin{aligned} \frac{\partial}{\partial t}\rho &= -\nabla \cdot (\rho \nabla \nu) \\ \frac{\partial}{\partial t}\nu &= -\gamma(t)\nu - \left(1 + \gamma(t)\sqrt{\frac{\mathbb{E}_{x\sim\nu}[z^2]}{\mathbb{E}_{x\sim\tau}[x^2]}}\right)\tau \end{aligned} \tag{11}$$

where τ is $\frac{1}{2}(\nabla \nu)^2 + \frac{\delta V}{\delta \rho}$.

We note that for a convex function, the strict minimizer is clearly a stationary point for standard gradient flow and accelerated gradient flow (damping sets the velocity to zero). For our flow with steering, the same may not hold true if the τ does not vanish faster than its second moment.

Discretization of Accelerated Gradient Flow (No Steering). We seek to find a discretization of the Accelerated Wasserstein gradient flow, in terms of particles $\vec{x} = \{x_i\}_{i=1}^N$ with $x_i \sim \rho$ and velocities $\vec{v} = \{v_i\}_{i=1}^N$.

It is unclear how to discretize Equation 11. We model similar to [4] and will attempt the following discretization:

$$\begin{aligned} \frac{d}{dt}x_i &= v_i \\ \frac{d}{dt}v_i &= -\gamma(t)v_i - \left(1 + \gamma(t)\sqrt{\frac{\mathbb{E}_{x\sim\nu}[z^2]}{\mathbb{E}_{x\sim F_i}[x^2]}}\right) F_i \\ F_i &= \nabla\left(\frac{\delta V}{\delta\rho}\right)(x_i) \end{aligned} \tag{12}$$

and estimate the moments after the discretization.

Wasserstein Gradient and Convexity of Example Functions. We now consider example potential energies $V(\rho)$ and consider their corresponding Wasserstein gradient and convexity under the Wasserstein-2 geometry.

Example 1: $V(\rho) = \int \mathcal{V}(x)\rho(x)\ \mathrm{d}x$ for $\mathcal{V} : \Omega \to \mathbb{R}$.

The Wasserstein gradient is $\nabla\frac{\delta V}{\delta\rho} = \nabla\mathcal{V}$. In addition, V is always convex under the L^2 metric, but is only convex under W^2 when $\mathcal{V}$ is also convex.

Example 2: $V(\rho) = \frac{1}{2}\int W(x,y)\rho(x)\rho(y)\ \mathrm{d}x\ \mathrm{d}y$ for $\mathcal{W} : \Omega \times \Omega \to \mathbb{R}$.

The Wasserstein gradient is $\nabla\frac{\delta V}{\delta\rho} = \int \nabla W(x,y)\rho(y)\ \mathrm{d}y$. Furthermore, in W^2 metric, V is convex only when W is convex.

Example 3: $V(\rho) = -\int \rho(x)f(\rho(x))\ \mathrm{d}x$ for $f : \mathbb{R}^+ \to \mathbb{R}$.

The Wasserstein gradient is $\nabla\frac{\delta V}{\delta\rho} = -\nabla(f(\rho) + \rho f'(\rho))$ and is convex when f is convex.

4. Numerical Examples

Example: Quadratic Function. We consider $\Omega = \mathbb{R}^3$ and the following potential:

$$V(\rho) = \int_\Omega \|x\|^2\rho(x)\ \mathrm{d}x,$$

which clearly has a strict minimizing measure $\rho^\star$ (delta about zero) with minimum value 0.

In our numerical experiments we take a Lagrangian discretization of the distribution, taking N points, rewriting the potential as:

$$V(\rho) \approx \frac{1}{N}\sum_i^N \|x_i\|^2$$

Figure 2 shows the convergence rate comparing the three approaches: standard gradient flow, accelerated flow, and accelerated flow with steering.

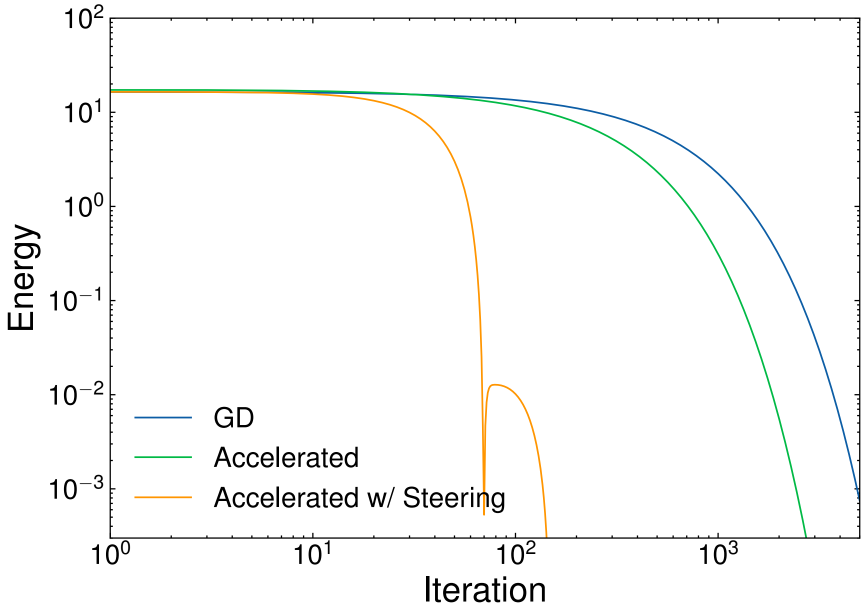

FIGURE 2. **Simple quadratic function gradient flow convergence comparison.** Energy versus iteration number.

Example: Fokker–Planck. We consider $\Omega = \mathbb{R}^2$ and the following potential:

$$V(\rho) = \int_\Omega \|x\|^2 \rho(x)\ \mathrm{d}x - \int_\Omega \rho(x)(\log(x) - 1)\ \mathrm{d}x,$$

which is the sum of a linear term and an entropy (with respect to Lebesque measure on $\mathbb{R}^2$). As we have shown in, this is the sum of two convex functions.

Under our Lagrangian discretization, we rewrite the potential as:

$$V(\rho) \approx \frac{1}{N}\left(\sum_i^N \|x_i\|^2 - \log(\min_{j\neq i} \mathrm{dist}(x_i, x_j))\right).$$

Notice that the right expression is an estimator of the entropy, which sums the logarithm of the nearest-neighbor distance [5]. In our implementation, we build a K-D Tree for fast nearest-neighbor lookup. The gradients are also computed analytically.

We randomly same $N = 10^3$ random points, with half sampled from the first quadrant and the other half the third quadrant. Figure 3 shows the initial configuration, after 500, 1000, 2000, and 5000 iterations of the discretization gradient flow for standard descent, accelerated flow without steering, and accelerated flow with steering. The points converge along a Gaussian distribution.

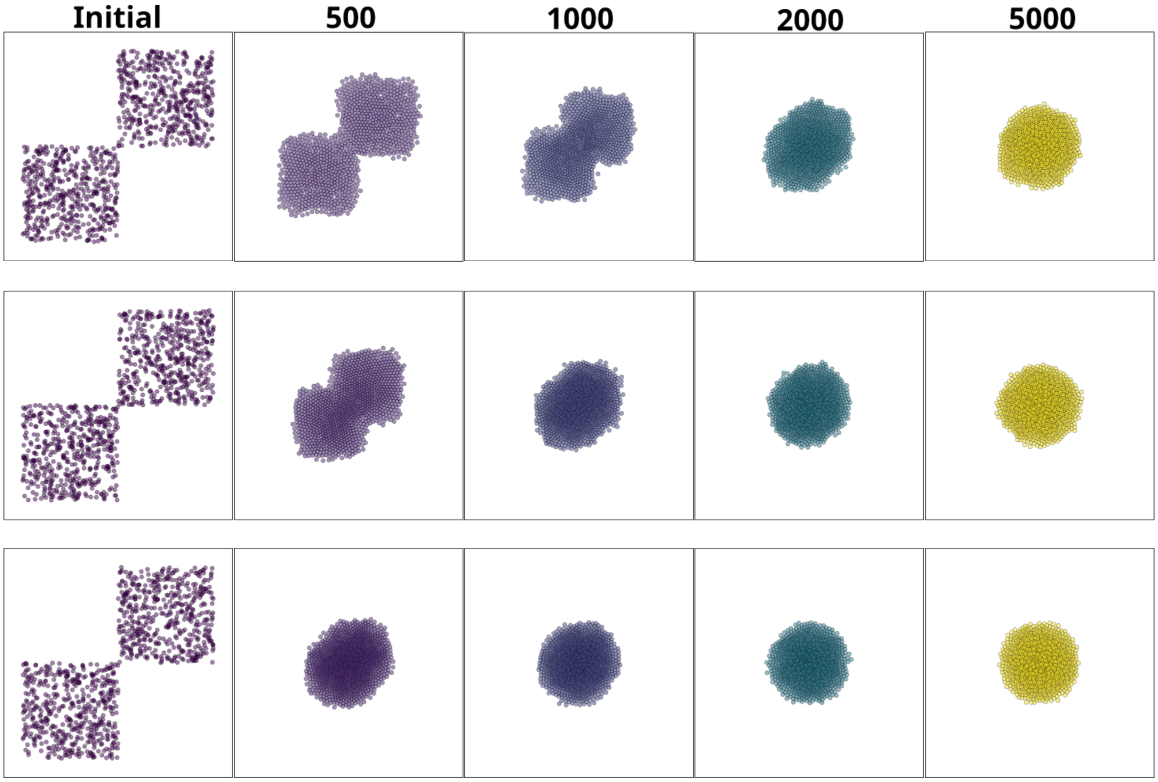

Figure 3. **Fokker-Planck gradient flow for standard gradient descent (top), accelerated flow (middle), and accelerated flow with steering (bottom).** Left to right: Initial configuration (uniformly sampled from first and third quadrant), after 500, 1000, 2000, and 5000 iterations of gradient flow with fixed step size 0.5 and damping $\gamma(t) = 0.5$. Color based on iteration number.

Animated visualizations of comparing these flows can be found here.

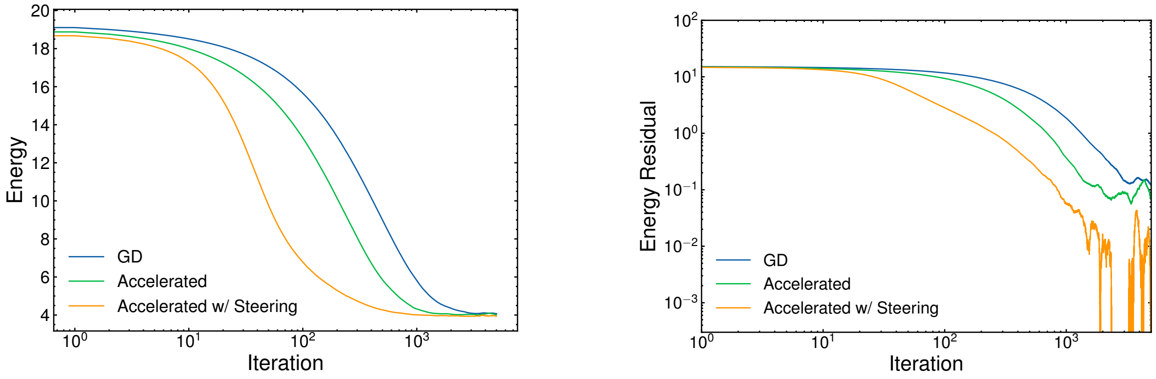

Figure 4. **Fokker-Planck gradient flow convergence comparison.** Energy versus iteration number (left) and residual (right).

Figure 4 shows the total potential energy as a function of iteration number. The accelerated flow with steering appears to converge significantly faster than the other two methods.

Though we have not proven convergence rates (it is not straightforward to find this), [4] has shown that if V is β-strongly convex, then convergence of accelerated Wasserstein gradient flow is $\mathcal{O}(e^{-\sqrt{\beta}t})$. We predict it is not significantly different for the case of accelerated flow with steering.

However, experimental results seem to dictate that there are scenarios where this accelerated gradient flow converges faster than previous methods.

Example: KL Divergence. We follow [6] to approximate the KL Divergence from samples. Our numerical experiment tests against a Cornell University logo, from which we sample $10^3 - 10^4$ points. The initial configuration is uniformly sampled from a 2-D square and we perform several iterations of our accelerated descent with steering, shown in Figure 5.

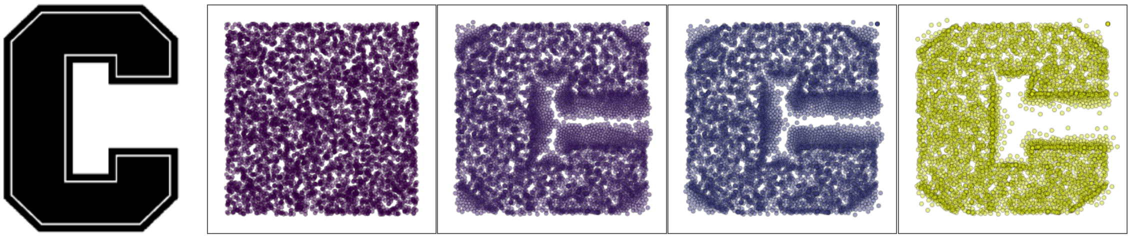

Figure 5. **KL Divergence example. From left to right: reference Cornell logo, initial configuration, resulting configuration after** 500 **iterations,** 1000 **iterations, and** 5000. Distribution of reference image taken by sampling $10^3 - 10^4$ points.

We do not cite convergence rates since the KL divergence estimate sometimes resulted in overflows. A fully animated visualization can be found here.

5. Discussion

We have developed and implemented several methods for Wasserstein gradient flow, considering flow without damping or acceleration, with damping (accelerated flow), and finally with damping and velocity "steering." We are not aware of any previous work that incorporates velocity steering in the context of gradient flow. These results may be beneficial for high-performance gradient flow (e.g., in training Bayesian neural networks).

In terms of potential future directions, the discretization of our algorithms could be improved, such as with adaptive step sizes and non-constant damping. In addition, we note that the Lagrangian discretization is not suitable for distributions in high-dimensions and does not offer a good approximation to these measures without an exceedingly high computational cost. Some work has shed light toward the use of neural networks to approximate the JKO update scheme (by estimating the proximal update) [7]. Though we were limited by computational resources, we should certainly aim for a wider variety of numerical experiments for benchmarks, perhaps considering optimization over non-convex functions.

Regarding our choice of Hamiltonian dynamics, we manually added damping to our system. This separates the Hamiltonian dynamics from the damping forces, which allows the Hamiltonian to be equal to the total energy and conserves its quantity (time-independent).

However, this is not a strict requirement; one may consider a Hamiltonian with exponential time factors, relaxing our requirement that $\mathcal{H}$ equals the total energy.

All code used to create the figures and numerical examples can be found here.

Appendix A. Gradient Flow Derivations

Kinetic Term of Hamiltonian. Following [5], we show that $T(\rho,\nu)$ in Equation 9 has the desired derivative with respect to ν. Take some $\xi \in \mathcal{P}_2(\Omega)$ and $\epsilon \in \mathbb{R}$ and consider the variation

$$\begin{aligned} T(\rho,\nu+\epsilon\xi) &= -\frac{1}{2}\int (\nu+\epsilon\xi)(\nabla\cdot(\rho\nabla(\nu+\epsilon\xi)))(x)\ \mathrm{d}x \\ &= \underbrace{-\frac{1}{2}\int (\nu(\nabla\cdot(\rho\nabla\nu))(x)\ \mathrm{d}x}_{T(\rho,\nu)} - \epsilon\int(\xi(\nabla\cdot(\rho\nabla\nu))(x)\ \mathrm{d}x + o(\epsilon^2) \\ &= T(\rho,\nu) - \epsilon\int \nabla\cdot(\rho\nabla\nu))\ \mathrm{d}\xi + o(\epsilon^2), \end{aligned}$$

so we have the desired $\frac{\delta T}{\delta \nu} = -\nabla\cdot(\rho\nabla\nu)$.

Taking the derivative of T now with respect to ρ gives the Hamiltonian dynamics for ν; we again repeat a similar variation

$$\begin{aligned} T(\rho+\epsilon\xi,\nu) &= -\frac{1}{2}\int \nu(\nabla\cdot((\rho+\epsilon\xi)\nabla\nu))\ \mathrm{d}x \\ &= T(\rho,\nu) - \frac{1}{2}\epsilon\int \nu(\nabla\cdot(\xi\nabla\nu))\ \mathrm{d}x \\ &= T(\rho,\nu) + \frac{1}{2}\epsilon\int \nabla\nu\cdot\nabla\nu\ \mathrm{d}\xi, \end{aligned}$$

giving us $\frac{\delta T}{\delta \nu} = \frac{1}{2}(\nabla\nu)^2$.

These results are consistent with previous findings [4, 8]. An equivalent formulation to the kinetic term of $T(\rho,\nu) = \frac{1}{2}\int(\nabla\nu)^2\ \mathrm{d}\rho$.

Proof of Gradients and Convexity for Example Functions.

Example 1. The Wasserstein gradient is straight-forward to compute (again via a variation):

$$V(\rho+\epsilon\xi) = \int \mathcal{V}(x)(\rho+\epsilon\xi)(x)\ \mathrm{d}x = V(\rho) + \epsilon\int \mathcal{V}(x)\ \mathrm{d}\xi,$$

resulting in $\frac{\delta V}{\delta \rho} = \mathcal{V}$.

To prove convexity under W^2, take two $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R})$ and suppose there exists an optimal transport plan γ such that there exists a Wasserstein geodesic $t \mapsto \mu_t$ for $t \in [0,1]$:

$$\mu_t = ((1-t)x + ty)\#\gamma$$

where # represents a push-forward. Then, looking at $V(\mu_t)$ shows that V is convex only when $\mathcal{V}$ is:

$$\begin{aligned}
V(\mu_t) &= \int \mathcal{V}(x)\ \mathrm{d}\mu_t(x) \\
&= \int\int \mathcal{V}((1-t)x+ty)\ \mathrm{d}\gamma(x,y) && \text{Law of Unconscious Statistician} \\
&\leq \int\int (1-t)\mathcal{V}(x)+t\mathcal{V}(y))\ \mathrm{d}\gamma(x,y) && \mathcal{V}\text{ convex} \\
&= (1-t)V(\mu_0)+tV(\mu_1)
\end{aligned}$$

Example 2. We find $\frac{\delta V(\rho)}{\delta\rho}$ by considering the following variation (for some measure ξ over Ω):

$$\begin{aligned}
V(\rho+\epsilon\xi) &= \frac{1}{2}\int(\rho+\epsilon\xi)(x)(\rho+\epsilon\xi)(y)W(x,y)\ \ \mathrm{d}x\ \mathrm{d}y \\
&= \frac{1}{2}\int\left(\rho(x)\rho(y)+\epsilon\xi(x)\rho(y)+\epsilon\xi(y)\rho(x)+o(\epsilon^2)\right)W(x,y)\ \mathrm{d}x\ \mathrm{d}y \\
&= E(\rho)+\epsilon\underbrace{\int W(x,y)\rho(y)dy}_{\frac{\delta E}{\delta\rho}}\ \mathrm{d}\xi+o(\epsilon^2)
\end{aligned}$$

Again, consider the same geodesic from the previous example. We see that convexity under W^2 depends on convexity of W:

$$\begin{aligned}
V(\mu_t) &= \int\int W(x,y)\ \mathrm{d}\mu_t(x)\ \mathrm{d}\mu_t(y) \\
&= \int\int W((1-t)x+tx'),(1-t)y+ty')\ \mathrm{d}\gamma(x,x')\ \mathrm{d}\gamma(y,y') \\
&\leq \int\int((1-t)W(x,y)+tW(x',y'))\ \mathrm{d}\gamma(x,x')\ \mathrm{d}\gamma(y,y') && W(x,y)\text{ convex} \\
&= (1-t)V(\mu_0)+tV(\mu_1)
\end{aligned}$$

so E_2 is convex under the Wasserstein geodesic when W is convex.

References

[1] Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch, *Structural relaxation made simple*, Phys. Rev. Lett. **97** (2006), 170201.

[2] Chris J. Maddison, Daniel Paulin, Yee Whye Teh, Brendan O'Donoghue, and Arnaud Doucet, *Hamiltonian descent methods* (2018), available at `1809.05042`.

[3] Yurii Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, Proceedings of the USSR Academy of Sciences **269** (1983), 543–547.

[4] Yifei Wang and Wuchen Li, *Accelerated information gradient flow*, Journal of Scientific Computing **90** (2021), 11.

[5] Gabriel Peyré and Marco Cuturi, *Computational optimal transport* (2020), available at `1803.00567`.

[6] Fernando Perez-Cruz, *Kullback-leibler divergence estimation of continuous distributions* (2008), 1666–1670.

[7] Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen, *Variational wasserstein gradient flow* (2022), available at `2112.02424`.

[8] Shui-Nee Chow, Wuchen Li, and Haomin Zhou, *Wasserstein hamiltonian flows* (2019), available at `1903.01088`.