Accelerated Proximal Optimization in Wasserstein and Sobolev Spaces

William Chang[1] Taro Iyadomi[2] Tianyu Wang[3]

[1]University of California, Los Angeles [2]Northwestern University [3]Columbia University

Abstract

We propose a unified accelerated proximal framework for optimization in Wasserstein and $\dot{\mathbb{H}}^1$ spaces. First, we develop an accelerated Jordan–Kinderlehrer–Otto (JKO) scheme that applies Nesterov-style momentum directly in Wasserstein space. Unlike prior work, which requires *strong* convexity of G along generalized geodesics, our analysis requires only (non-strong) geodesic convexity, yet achieves an $O(1/t^2)$ rate on the objective gap, with the iterate-to-optimum squared Wasserstein distance bounded uniformly in t. Second, we introduce an Accelerated Sobolev Gradient Ascent (ASGA) algorithm for Wasserstein barycenters. Via a proximal step, ASGA achieves $O(1/t^2)$ convergence, improving on the $O(1/\sqrt{T})$ rate of the non-accelerated Sobolev Gradient Ascent of [34] while retaining its projection-free structure, which eliminates expensive c-concavity projections. Together these results unify optimal transport geometry with accelerated first-order methods, yielding both theoretical guarantees and practical algorithms for large-scale transport and generative modeling.

1 Introduction

Recent years have seen growing interest in optimization over spaces of probability measures, driven by applications in generative modeling and optimal transport. The Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$, equipped with the W_2 metric, provides a natural Riemannian-type geometry for analyzing the evolution of distributions [8]. A landmark result in this area is the Jordan–Kinderlehrer–Otto (JKO) scheme [33], which interprets the Fokker–Planck equation as a Wasserstein gradient flow and discretizes it via implicit Euler (proximal) steps. This insight has spurred a line of work on scalable JKO approximations using neural networks, yielding flow-based generative models with tractable likelihoods [61, 19, 7, 47, 59]. However, theoretical guarantees for these models have so far required strong convexity of the objective functional along generalized geodesics [19], which restricts their applicability and leaves the non-strongly-convex regime poorly understood.

A key recent motivation is to characterize the probabilistic behavior underlying the training dynamics of generative models, encompassing GANs [27, 29], VAEs [35, 36], normalizing flows [37], and diffusion models [30, 55, 56, 3, 45], within which sharper convergence guarantees for flow-based methods remain a central theoretical objective. The strong-convexity assumption in existing JKO analyses cannot be relaxed without fundamentally changing the proof strategy. In the Euclidean setting, gradient descent on a λ-strongly convex objective contracts at the geometric rate $(1-\gamma\lambda)^t$, but for $\lambda = 0$ the rate degrades to $O(1/t)$; the standard JKO analysis exhibits the same dichotomy, with exponential contraction available only when $\lambda > 0$ and only the slower $O(1/t)$ rate available otherwise [52]. In Euclidean optimization, this rate deficit is remedied by Nesterov acceleration [48], which attains the optimal $O(1/t^2)$ rate for convex objectives via a momentum recursion and admits geometric [12], dynamical [58, 60], and coupling-based [5] interpretations, with subsequent

extensions to nonconvex [32], step-size-based [28], and continuous-time probability-measure [18] settings. No analogous accelerated discrete-time guarantee for the JKO scheme is presently available under merely geodesically convex objectives. Our first contribution is to establish such a guarantee.

A parallel question arises in the dual geometry of optimal transport. The Wasserstein barycenter [1], defined as the minimizer of

$$\mathcal{B}(\rho) \triangleq \sum_{i=1}^{m} \frac{w_i}{2} W_2^2(\mu_i, \rho), \tag{1}$$

provides a natural notion of averaging for distributions and has been widely applied to clustering [31, 65], Bayesian inference [57], and beyond. Computing barycenters is challenging: the problem is NP-hard in high dimensions [6] and even in fixed dimension requires sophisticated primal–dual formulations [66, 62, 64, 34]. The Sobolev gradient ascent (SGA) algorithm of [34] reformulates the barycenter problem as the maximization of a concave Kantorovich-type functional over the homogeneous Sobolev space $\dot{\mathbb{H}}^1$ and attains the standard non-smooth rate $O(1/\sqrt{T})$, without the restrictive c-concavity projections required by earlier dual methods. Our second contribution shows that proper acceleration for the dual functional in $\dot{\mathbb{H}}^1$ achieves an $O(1/t^2)$ rate.

Our Contributions. We study *acceleration in Wasserstein and* $\dot{\mathbb{H}}^1$ *spaces.*

- **Accelerated proximal descent in Wasserstein space (Section 4.1).** We propose and analyze an accelerated JKO scheme that applies Nesterov-style momentum in Wasserstein space, with applications to flow-based generative models. Unlike [19], which requires strong convexity ($\lambda > 0$) for exponential convergence, we require only geodesic convexity of G, yet achieve an $O(1/t^2)$ convergence rate on the objective gap $G(x_t) - G(\rho^*)$, with the squared Wasserstein distance $W_2^2(z_t, \rho^*)$ bounded uniformly in t.
- **Accelerated Sobolev Gradient Ascent for barycenters (Section 4.2).** We introduce ASGA, an accelerated variant of the SGA algorithm of [34] for the Wasserstein barycenter problem. With a proximal oracle, ASGA achieves an $O(1/t^2)$ rate, improving quadratically over the $O(1/\sqrt{T})$ rate of (non-accelerated) SGA while retaining its constraint-free structure.
- **Unified framework.** Both algorithms instantiate a common three-sequence accelerated proximal recursion—Nesterov's method generalized to non-Euclidean geometries—highlighting the broad applicability of acceleration beyond the Euclidean setting.

2 Related Works

Wasserstein Gradient Flows and Proximal Methods. The Wasserstein gradient flow framework of [8] provides the analytical foundation for the JKO scheme [33] and its variants. On the discretization side, [52] analyzes a Forward–Backward proximal scheme on $\mathcal{P}_2$ with Euclidean-style $O(1/t)$ guarantees, and [53] establishes strong L^2–H^2 convergence of the JKO sequence to the continuous Fokker–Planck solution. [13] prove Γ-convergence of entropic-regularized JKO steps to their unregularized limit. Wasserstein gradient flows have also been used to analyze nonconvex problems beyond sampling, including overparameterized two-layer networks via mean-field flow [20] and discretization-free inference [25]. Our contribution sits within this proximal line and extends it by accommodating geodesically (rather than strongly) convex objectives at an accelerated rate.

Generative Models, Flow-Based Methods, and Neural Optimal Transport. Beyond the generative-model literature cited in Section 1, score-based diffusion has seen rapid progress in convergence guarantees [11, 14, 16, 22, 40, 41, 50], whereas rigorous results for flow-based processes remain scarcer. Recent analyses address continuous-time flow matching [4, 3] and ODE-based samplers linked to SDE-trained scores [15, 17, 42]. A complementary line learns optimal transport maps directly using neural architectures: stochastic dual formulations [26, 54], input-convex networks [9], and their use in optimal transport [46, 38, 39]. Closest to our work are JKO-inspired flow architectures [19, 61] and related proposals [7, 47, 59], which establish convergence by discretizing Wasserstein gradient flows under *strong* convexity of the objective.

Barycenter Computation in Optimal Transport. Computational methods for the Wasserstein barycenter [1] split into *primal* and *dual* approaches. Primal formulations recast the problem as a linear program [66], tractable in theory but poorly scaling to high resolution. Fixed-point iterations [66] and first-order Wasserstein descent [62] have been proposed, with quantitative guarantees largely limited to Gaussian settings. Even the fixed-support variant is computationally hard: [43] show that the standard linear-programming representation is not totally unimodular for $m \geq 3$ marginals and provide a fast iterative-Bregman variant. Dual formulations avoid explicit OT maps. The multimarginal OT perspective [49, 10] reduces computation but introduces $O(m^2)$ constraints without convergence guarantees. More recently, [34] proposed a nonconvex–concave minimax formulation leading to the Wasserstein Descent–$\dot{H}^1$ Ascent (WDHA) algorithm, achieving $O(T^{-1/2})$ rates at the cost of an expensive projection. Their SGA algorithm eliminates this projection at the same rate. A separate line of work on entropic regularization [21, 10, 51, 23, 24, 43, 44] offers scalable approximations at the cost of bias.

Acceleration in Non-Euclidean Geometries. Our contributions instantiate Nesterov-style acceleration in non-Euclidean settings, a topic with a substantial parallel history on Riemannian manifolds. [63] establishes the foundational rates for first-order methods under geodesic convexity, and [2] develop a Riemannian analog of Nesterov's estimate-sequence framework. [18] study acceleration in $\mathcal{P}_2$ at the continuous-time level via a Hamiltonian-flow formulation; our work contributes a discrete-time analog for the JKO recursion in $\mathcal{P}_2$ and the dual barycenter ascent in $\dot{\mathbb{H}}^1$.

3 Preliminaries

3.1 Notation

We work with probability measures over $\mathbb{R}^d$, denoted $\mathcal{P}_2(\mathbb{R}^d)$, the set of distributions with finite second moment:

$$\mathcal{P}_2 := \left\{ P \text{ on } \mathbb{R}^d \;:\; M_2(P) := \int_{\mathbb{R}^d} \|x\|^2 \, dP(x) < \infty \right\}.$$

We denote by $\mathcal{P}_2^r \subset \mathcal{P}_2$ the subset of measures absolutely continuous with respect to Lebesgue measure (i.e., those possessing densities). We identify a density p with its corresponding distribution and write $X \sim p$.

Given a measurable vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ and a distribution P, the $L^2(P)$-norm is $\|v\|_P := \left(\int \|v(x)\|^2 \, dP(x)\right)^{1/2}$ and the inner product is $\langle u, v \rangle_P := \int u(x)^T v(x) \, dP(x)$. Given a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$, the pushforward of P under T is denoted $T_\# P$, defined by $T_\# P(A) := P(T^{-1}(A))$ for measurable sets A. For two maps S, T, we write $S \circ T$ for their composition, and Id for the identity map.

3.2 Wasserstein-2 Distance and Optimal Transport

Theorem 1 (Brenier). *Let $\mu \in \mathcal{P}_2^r$ and $\nu \in \mathcal{P}_2$. Then:*

1. *There is a unique optimal coupling π, realized by a map $T_\mu^\nu = \nabla\varphi$ for some convex φ, with $\pi = (\mathrm{Id}, T_\mu^\nu)_\# \mu$.*
2. *$W_2^2(\mu,\nu) = \inf_{T:T_\#\mu=\nu} \int \|x - T(x)\|^2 \, d\mu(x)$.*
3. *If $\nu \in \mathcal{P}_2^r$, then T_μ^ν is invertible μ-a.e. and $T_\nu^\mu \circ T_\mu^\nu = \mathrm{Id}$.*

3.3 Convexity Along Generalized Geodesics

We define convexity along a generalized geodesic, which allows proximal methods to converge to a global minimum.

Definition 2 (Convexity along generalized geodesics, [8, Def. 9.2.4]). *A proper, lower semicontinuous functional $G : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is said to be* λ-convex along generalized geodesics *(a.g.g.) if for every $\nu \in \mathcal{P}_2^r(\mathbb{R}^d)$ and every $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, denoting by T_ν^i the optimal transport map from ν to μ_i $(i = 1, 2)$, the generalized geodesic*

$$\mu_t^{1\to 2} \;:=\; \big((1-t)\,T_\nu^1 + t\,T_\nu^2\big)_\#\nu$$

satisfies

$$G\big(\mu_t^{1\to 2}\big) \;\le\; (1-t)\,G(\mu_1) + t\,G(\mu_2) \;-\; \tfrac{\lambda}{2}\,t(1-t)\,W_2^2(\mu_1,\mu_2), \qquad \forall\, t \in [0,1].$$

When $\lambda = 0$ we say G is convex along generalized geodesics*; when $\lambda > 0$ we say G is* strongly convex along generalized geodesics*.*

Proposition 3 (First-order inequality along generalized geodesics, [8, Thm. 10.3.6, Lem. 10.3.8]). *Let $G : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ be convex along generalized geodesics in the sense of Definition 2, and fix $\mu_1 \in \mathrm{Dom}(G)$. If $\xi \in \partial_{W_2} G(\mu_1)$, then for every $\nu \in \mathcal{P}_2^r$ and every $\mu_2 \in \mathcal{P}_2$, denoting by T_ν^i the OT map from ν to μ_i $(i = 1, 2)$,*

$$G(\mu_2) \;\ge\; G(\mu_1) \;+\; \big\langle \xi \circ T_\nu^1,\; T_\nu^2 - T_\nu^1 \big\rangle_\nu.$$

Remark 4 (Generalized geodesics with arbitrary plans). *[8, Def. 9.2.4] parametrizes generalized geodesics by arbitrary 3-plans $\boldsymbol{\mu} \in \Gamma(\nu; \mu_1, \mu_2)$, not only OT plans; Definition 2 is the OT specialization. Under the original AGS definition, Proposition 3 extends to any transport maps S_ν^i pushing ν to μ_i, and standard gg-convex functionals (relative entropy, internal and potential energies, and sums thereof; see [8, Ch. 9]) satisfy this stronger property. Our analysis uses Proposition 3 in this extended form.*

3.4 The Sobolev Space $\dot{\mathbb{H}}^1(\Omega)$

Following the notation of [34], let $\Omega \subset \mathbb{R}^d$ be a compact convex set and $c(x,y) := \frac{1}{2}\|x-y\|_2^2$ be the quadratic transport cost. For a function $f : \Omega \to \mathbb{R}$, the c-transform is

$$f^c(x) := \inf_{y\in\Omega} \{c(x,y) - f(y)\}.$$

The homogeneous Sobolev space is

$$\dot{\mathbb{H}}^1(\Omega) := \left\{ f : \Omega \to \mathbb{R} \;\middle|\; \int_\Omega \|\nabla f(x)\|_2^2 \, dx < \infty \right\},$$

endowed with inner product $\langle f, g \rangle_{\dot{\mathbb{H}}^1} := \int_\Omega \langle \nabla f(x), \nabla g(x) \rangle \, dx$ and norm $\|f\|_{\dot{\mathbb{H}}^1} := \big(\langle f, f \rangle_{\dot{\mathbb{H}}^1}\big)^{1/2} = \big(\int_\Omega \|\nabla f(x)\|_2^2 \, dx\big)^{1/2}$. Elements are defined modulo additive constants. For a functional $\mathcal{J} : \dot{\mathbb{H}}^1 \to \mathbb{R}$, its $\dot{\mathbb{H}}^1$-gradient $\nabla \mathcal{J}(f)$ is defined through the Riesz identity

$$\langle \nabla \mathcal{J}(f), h \rangle_{\dot{\mathbb{H}}^1} = \delta \mathcal{J}_f(h), \qquad \forall h \in \dot{\mathbb{H}}^1,$$

where $\delta \mathcal{J}_f(h)$ denotes the Gâteaux derivative at f in the direction h.

Unified Accelerated Framework We adopt a unified accelerated framework that extends Nesterov's method from the Euclidean setting to Wasserstein and $\dot{\mathbb{H}}^1$ geometries. The general scheme is:

$$y_t = (1 - \alpha_t) x_t + \alpha_t z_t, \tag{2}$$
$$x_{t+1} = \mathrm{Prox}(G, y_t, \gamma), \tag{3}$$
$$z_{t+1} = \mathrm{Update}(z_t, y_t, x_{t+1}, \alpha_t, \gamma), \tag{4}$$

with $\alpha_t = \frac{3}{t+3}$. The operators $\mathrm{Prox}(\cdot)$ and $\mathrm{Update}(\cdot)$ depend on the underlying geometry:

- **Case 0 (Euclidean / Nesterov).**

$$\mathrm{Prox}(G, y_t, \gamma) = \arg\min_u \Big\{ G(u) + \tfrac{1}{2\gamma} \|u - y_t\|_2^2 \Big\},$$

$$\mathrm{Update}(z_t, y_t, x_{t+1}, \alpha_t, \gamma) = z_t + \tfrac{1}{\alpha_t}(x_{t+1} - y_t).$$

- **Case I (Wasserstein-2 setting).** Iterates are realized as pushforwards of a common base z_t, with Y_t and X_{t+1} denoting maps from z_t such that $y_t = (Y_t)_\# z_t$ and $x_{t+1} = (X_{t+1})_\# z_t$.

$$\mathrm{Prox}(G, y_t, \gamma) = \arg\min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \Big\{ G(\rho) + \tfrac{1}{2\gamma} W_2^2(\rho, y_t) \Big\},$$

$$\mathrm{Update}(z_t, y_t, x_{t+1}, \alpha_t, \gamma) = \big(\mathrm{Id} + \tfrac{1}{\alpha_t}(X_{t+1} - Y_t)\big)_\# z_t.$$

- **Case II ($\dot{\mathbb{H}}^1$ setting).**

$$\mathrm{Prox}(G, y_t, \gamma) = \arg\min_u \Big\{ G(u) + \tfrac{1}{2\gamma} \|u - y_t\|_{\dot{\mathbb{H}}^1}^2 \Big\},$$

$$\mathrm{Update}(z_t, y_t, x_{t+1}, \alpha_t, \gamma) = z_t + \tfrac{1}{\alpha_t}(x_{t+1} - y_t).$$

The same three-line recursion (2)–(4) thus simultaneously captures Nesterov's acceleration in $\mathbb{R}^n$ (Case 0), the accelerated JKO scheme (Case I), and the accelerated dual barycenter update (Case II). Our main results, Theorems 8 and 11, are the $O(1/t^2)$ guarantees for Cases I and II; Case 0 is the classical Euclidean Nesterov result, included to make explicit that the other two are genuine generalizations. In all three cases, the proof skeleton (a Lyapunov sequence Δ_t/λ_t shown to be nonincreasing) is identical. The cases differ only in how inner products, distances, and momentum updates are defined.

4 Main Results

In this section we collect all of our main convergence theorems for both the accelerated JKO scheme and the accelerated Sobolev gradient ascent algorithm. All proofs are deferred to Appendix A and Appendix B, respectively.

4.1 Accelerated JKO

The JKO scheme [33] constructs an implicit time-discretization for the Fokker–Planck equation

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \beta^{-1} \Delta \rho, \tag{5}$$

via the recursion: given step size $h > 0$ and initial density $\rho^{(0)}$,

$$\rho^{(k)} = \arg\min_{\rho \in K} \left\{ \frac{1}{2h} W_2^2(\rho, \rho^{(k-1)}) + F(\rho) \right\}, \tag{6}$$

where $F(\rho) = \int V(x)\rho(x)\,dx + \beta^{-1} \int \rho(x) \log \rho(x)\,dx$ is the free energy functional and K is the set of probability densities with unit mass and finite second moment.

Assumption 5 (Geodesic convexity)**.** *$G : \mathcal{P}_2 \to (-\infty, +\infty]$ is lower semi-continuous,* $\mathrm{Dom}(G) \subset \mathcal{P}_2^r$*, and G is convex along generalized geodesics in the sense of Definition 2 (i.e., $\lambda = 0$).*

Remark 6 (Non-strong convexity: the $\lambda = 0$ case)**.** *In [19], strong convexity along generalized geodesics ($\lambda > 0$) gives an exponential contraction $(1 + \gamma\lambda/2)^{-1}$ per step; at $\lambda = 0$ their analysis breaks down. Assumption 5 requires only geodesic convexity. Just as Nesterov acceleration restores $O(1/t^2)$ in the Euclidean setting without strong convexity, Theorem 8 extends this principle to the Wasserstein geometry.*

The Accelerated JKO Scheme. Analogous to Nesterov acceleration, Algorithm 1 augments JKO with a momentum sequence. Let $G : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ be proper, lower semi-continuous, and convex along generalized geodesics, with minimizer ρ^*. Fix $\gamma > 0$ and initialize $x_0 = z_0 \in \mathcal{P}_2(\mathbb{R}^d)$. The scheme maintains the iterate x_t, momentum buffer z_t, and extrapolated point y_t. At each step, with $\alpha_t = 3/(t+3)$:

$$y_t = (Y_t)_{\#} z_t, \qquad Y_t := (1 - \alpha_t) X_t + \alpha_t \,\mathrm{Id}, \tag{7}$$

$$x_{t+1} = \arg\min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ G(\rho) + \frac{1}{2\gamma} W_2^2(\rho, y_t) \right\}, \tag{8}$$

$$X_{t+1} = T_t \circ Y_t, \tag{9}$$

$$z_{t+1} = (Z_{t+1})_{\#} z_t, \qquad Z_{t+1} := \mathrm{Id} + \frac{1}{\alpha_t}(X_{t+1} - Y_t), \tag{10}$$

where T_t is the OT map from y_t to x_{t+1}, and X_t is the OT map from z_t to x_t (recomputed at the start of each iteration). Note that $X_{t+1} := T_t \circ Y_t$ is a transport map from z_t to x_{t+1} but in general *not* the OT map between them; this is sufficient for our analysis (cf. Remark 4).

Algorithm 1 Accelerated JKO — Nesterov-accelerated proximal descent in W_2

Input: step size $\gamma > 0$, coefficients $\{\alpha_t\}_{t=0}^{T-1}$, horizon T
Initialize: $x_0 = z_0 \in \mathcal{P}_2(\mathbb{R}^d)$
for $t = 0, 1, \dots, T-1$ **do**
  Compute OT map $X_t = T_{z_t}^{x_t}$ from z_t to x_t
  Form the extrapolated measure

$$y_t \leftarrow \big((1-\alpha_t)X_t + \alpha_t \,\mathrm{Id}\big)_{\#} z_t$$

  Solve the proximal step

$$x_{t+1} \leftarrow \arg\min_{\rho \in \mathcal{P}_2} \left\{ G(\rho) + \frac{1}{2\gamma} W_2^2(\rho, y_t) \right\}$$

  Let $T_t = T_{y_t}^{x_{t+1}}$ and set $X_{t+1} \leftarrow T_t \circ Y_t$
  Update the momentum sequence

$$z_{t+1} \leftarrow \big(\mathrm{Id} + \tfrac{1}{\alpha_t}(X_{t+1} - Y_t)\big)_{\#} z_t$$

end for

Convergence Results

Lemma 7 (One-step accelerated descent in Wasserstein space)**.** *Assume G is convex along generalized geodesics and let $\rho^* \in \arg\min G$. Suppose that at step t there exist optimal transport maps*

$$X_t = T_{z_t}^{x_t}, \qquad P_t = T_{z_t}^{\rho^*},$$

and define the generalized-geodesic interpolation as in (7)–(10). Then:

$$\frac{\alpha_t^2}{2\gamma} W_2^2(z_{t+1}, \rho^*) + G(x_{t+1}) - G(\rho^*) \le \frac{\alpha_t^2}{2\gamma} W_2^2(z_t, \rho^*) + (1-\alpha_t)(G(x_t) - G(\rho^*)).$$

The proof is given in Appendix A.2. The key technical ingredients are the proximal subdifferential identity (Lemma 13) and the monotonicity inequality for functionals convex along generalized geodesics.

Theorem 8 (Accelerated JKO convergence)**.** *Under the same assumptions as Lemma 7, run Algorithm 1 with $\alpha_t = 3/(t+3)$ and $x_0 = z_0$. Then for all $t \ge 1$,*

$$G(x_t) - G(\rho^*) \le \frac{9}{2\gamma(t+2)^2} W_2^2(z_0, \rho^*) = O(t^{-2}).$$

More generally, defining $\lambda_{t+1} := \alpha_t^2/\gamma$ and the Lyapunov functional $\Delta_t := \frac{\lambda_t}{2} W_2^2(z_t, \rho^*) + G(x_t) - G(\rho^*)$, one has $\Delta_{t+1}/\lambda_{t+1} \le \Delta_1/\lambda_1$ for all $t \ge 0$, provided $\lambda_t(1-\alpha_t) \le \lambda_{t+1}$. Since $\alpha_0 = 1$, the first step gives $\Delta_1 \le \frac{\lambda_1}{2} W_2^2(z_0, \rho^*)$, so the initial objective gap $G(x_0) - G(\rho^*)$ does not enter the bound.

The proof is given in Appendix A.3. The argument defines the Lyapunov sequence Δ_t/λ_t, shows that it is nonincreasing via Lemma 7, and verifies the coefficient condition for $\alpha_t = 3/(t+3)$.

In contrast, under λ-strong convexity along generalized geodesics ($\lambda > 0$) and per-step error $\|\xi_{n+1}\|_{p_{n+1}} \le \varepsilon$, [19]'s Theorem 4.3 gives

$$W_2^2(p_n, q) \le \big(1 + \tfrac{\gamma\lambda}{2}\big)^{-n} W_2^2(p_0, q) + \tfrac{4\varepsilon^2}{\lambda^2},$$

which provides exponential contraction only when $\lambda > 0$ and an error floor $O(\varepsilon^2/\lambda^2)$ that diverges as $\lambda \to 0$. Our bound is uniform in λ at the $O(1/t^2)$ rate—optimal for first-order methods on convex non-strongly-convex problems [48]—and applies under mere geodesic convexity.

Remark 9 (Approximate JKO steps)**.** *When each proximal step is solved only approximately, the one-step lemma acquires an additional error term that, like the per-step error in standard inexact Nesterov methods, may accumulate under acceleration. A sharp analysis of error propagation through the Lyapunov sequence Δ_t/λ_t is left to future work; here we restrict attention to the exact-prox case analyzed in Theorem 8.*

4.2 Accelerated Sobolev Gradient Ascent

We consider the Kantorovich dual functional

$$\mathcal{I}(f) := \int_\Omega f \, d\mu + \int_\Omega f^c \, d\nu.$$

Its maximizers are Kantorovich potentials for the transport problem between μ and ν. A key point, already implicit in the Sobolev-gradient approach to optimal transport, is that the first variation of $\mathcal{I}$ is meaningful even away from the class of c-concave functions. In particular, if $g = \nabla\mathcal{I}(f)$, then g is characterized by the Neumann problem

$$\begin{cases} -\Delta g = \mu - (\mathrm{Id} - \nabla f^c)_\# \nu, & \text{in } \Omega, \\ \frac{\partial g}{\partial n} = 0, & \text{on } \partial\Omega, \end{cases}$$

or, equivalently,

$$\nabla\mathcal{I}(f) = (-\Delta)^{-1}\big(\mu - (\mathrm{Id} - \nabla f^c)_\# \nu\big),$$

with zero Neumann boundary conditions and the understanding that additive constants are irrelevant. This observation allows one to optimize $\mathcal{I}$ directly in the Sobolev geometry, without projecting after every step onto the cone of c-concave functions.

The accelerated version is obtained by applying the usual Nesterov estimate-sequence idea in [34], but now in the Hilbert space $\dot{\mathbb{H}}^1$. Since $\dot{\mathbb{H}}^1$ is a Hilbert space, all Euclidean algebraic manipulations carry over verbatim once inner products and norms are interpreted in $\dot{\mathbb{H}}^1$. See the full pseudo-code in Algorithm 2.

We now state the one step iteration lemma for Algorithm 2.

Lemma 10 (One-step accelerated ascent in $\dot{\mathbb{H}}^1$)**.** *Let $\tilde{f} \in \dot{\mathbb{H}}^1$ be a maximizer of $\mathcal{I}$. Then the iterates of Algorithm 2 satisfy*

$$\frac{\alpha_t^2}{2\eta}\|z^{(t+1)} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + \mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(t+1)}) \le \frac{\alpha_t^2}{2\eta}\|z^{(t)} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + (1-\alpha_t)\big(\mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(t)})\big).$$

The proof is given in Appendix B.1. Using this lemma we can prove our main convergence result.

Theorem 11 (Accelerated convergence of ASGA)**.** *Run Algorithm 2 with $\alpha_t = 3/(t+3)$, step size $\eta > 0$, and $z^{(0)} = f^{(0)}$. Then for every $t \ge 1$,*

$$\mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(t)}) \le \frac{9}{2\eta(t+2)^2}\,\|z^{(0)} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 = O(t^{-2}),$$

where $\tilde{f}$ is any maximizer of $\mathcal{I}$. Since $\alpha_0 = 1$, the initial dual gap $\mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(0)})$ does not enter the bound.

Algorithm 2 ASGA — Accelerated Sobolev Proximal Gradient Ascent

1: **Input:** step size $\eta > 0$, coefficients $\{\alpha_t\}_{t=0}^{T-1}$, horizon T
2: **Initialize:** $f^{(0)} \in \dot{\mathbb{H}}^1(\Omega)$, $z^{(0)} \leftarrow f^{(0)}$
3: **for** $t = 0, 1, \ldots, T-1$ **do**
4: Form the extrapolated point

$$y^{(t)} \leftarrow (1-\alpha_t) f^{(t)} + \alpha_t z^{(t)}$$

5: Perform the proximal accelerated ascent update

$$f^{(t+1)} = \arg\max_f \left\{ \mathcal{I}(f) - \frac{1}{2\eta} \|f - y^{(t)}\|_{\dot{\mathbb{H}}^1}^2 \right\}.$$

▷ Notice this proximal step implicitly gives $f^{(t+1)} = y^{(t)} + \eta \nabla \mathcal{I}(f^{(t+1)})$

6: Update $z^{(t+1)}$ via

$$z^{(t+1)} \leftarrow z^{(t)} + \frac{\eta}{\alpha_t} \nabla \mathcal{I}(f^{(t+1)}).$$

7: **end for**

The proof is given in Appendix B.2. The argument parallels the JKO case: a Lyapunov sequence Φ_t/λ_t with $\lambda_{t+1} = \alpha_t^2/\eta$ is shown nonincreasing via Lemma 10, with the coefficient condition verified for $\alpha_t = 3/(t+3)$.

Remark 12 (Smooth case via explicit gradient)**.** *Theorem 11 assumes an exact proximal oracle for $\mathcal{I}$, which has no closed form for the unregularized Kantorovich dual. When $\mathcal{I}$ is additionally L-smooth in $\dot{\mathbb{H}}^1$ (i.e., $\nabla\mathcal{I}$ is L-Lipschitz)—for example, after entropic regularization, where the smoothed dual $\mathcal{I}^\varepsilon$ has a Lipschitz gradient—the same Lyapunov argument yields the same $O(1/t^2)$ rate for the explicit-gradient variant of Algorithm 2 that replaces the proximal step on Line 5 with the forward step $f^{(t+1)} = y^{(t)} + \eta\, \nabla\mathcal{I}(y^{(t)})$ at any step size $\eta \le 1/L$. This explicit variant requires only a single Sobolev-gradient evaluation (one Poisson solve) per iteration, and is the variant implemented in our experiments (Appendix C.5).*

Table 1 summarizes the comparison. The $O(T^{-1/2})$ rate of WDHA and SGA [34] arises from running gradient descent–ascent on the nonconvex–concave minimax barycenter formulation, and bounds the squared Wasserstein gradient norm of the outer problem. WDHA attains this rate via projection onto $\mathcal{F}_{\alpha,\beta}$, while SGA replaces the projection with the second convex conjugate at the cost of dropping the strong-convexity and local smoothness regularity. ASGA improves the rate quadratically to $O(t^{-2})$ via Nesterov acceleration (available directly because $\dot{\mathbb{H}}^1$ is a Hilbert space) while retaining the projection-free structure of SGA. Per-iteration cost depends on which version of Algorithm 2 is used: the exact-prox formulation (Theorem 11) requires solving an inner non-quadratic optimization for the proximal step, while the explicit-gradient smooth variant (Remark 12) requires only a single Poisson solve per iteration, matching the per-iteration cost of SGA.

Multi-marginal extension. For the barycenter with m marginals (Eq. (1)), the dual formulation of [34] introduces m Kantorovich potentials $f_1, \ldots, f_m \in \dot{\mathbb{H}}^1(\Omega)$ with sum-to-zero constraint $\sum_i w_i f_i = 0$ and functional $\mathcal{I}_m(f_1, \ldots, f_m) := \sum_i w_i \int f_i \, d\mu_i + \int \left(\sum_i w_i f_i^c\right) d\rho_*$, where ρ_* is the barycenter. The i-th Sobolev gradient is $\nabla_{f_i} \mathcal{I}_m = (-\Delta)^{-1}(\mu_i - (\mathrm{Id} - \nabla f_i^c)_{\#}\rho_*)$. Algorithm 2 extends component-wise (one momentum sequence $(z_i^{(t)})$ per i); Theorem 11 and Remark 12 carry over to the product Hilbert

Table 1: Comparison of WDHA, SGA, and ASGA for the Kantorovich dual. ASGA is the only method achieving the $O(t^{-2})$ rate, and does so without requiring the c-concavity projection of WDHA. ASGA admits two variants: the exact-prox version (Theorem 11) and the smooth-explicit version (Remark 12).

Method	Assumption	Rate	Projection-free?
WDHA [34]	Strong concavity + smoothness	$O(T^{-1/2})$	No ($\mathcal{F}_{\alpha,\beta}$ proj.)
SGA [34]	Concavity (via φ^{**})	$O(T^{-1/2})$	Yes
ASGA-prox (ours)	Concavity + proximal oracle	$O(t^{-2})$	**Yes**
ASGA-smooth (ours)	Concavity + L-smoothness	$O(t^{-2})$	**Yes**

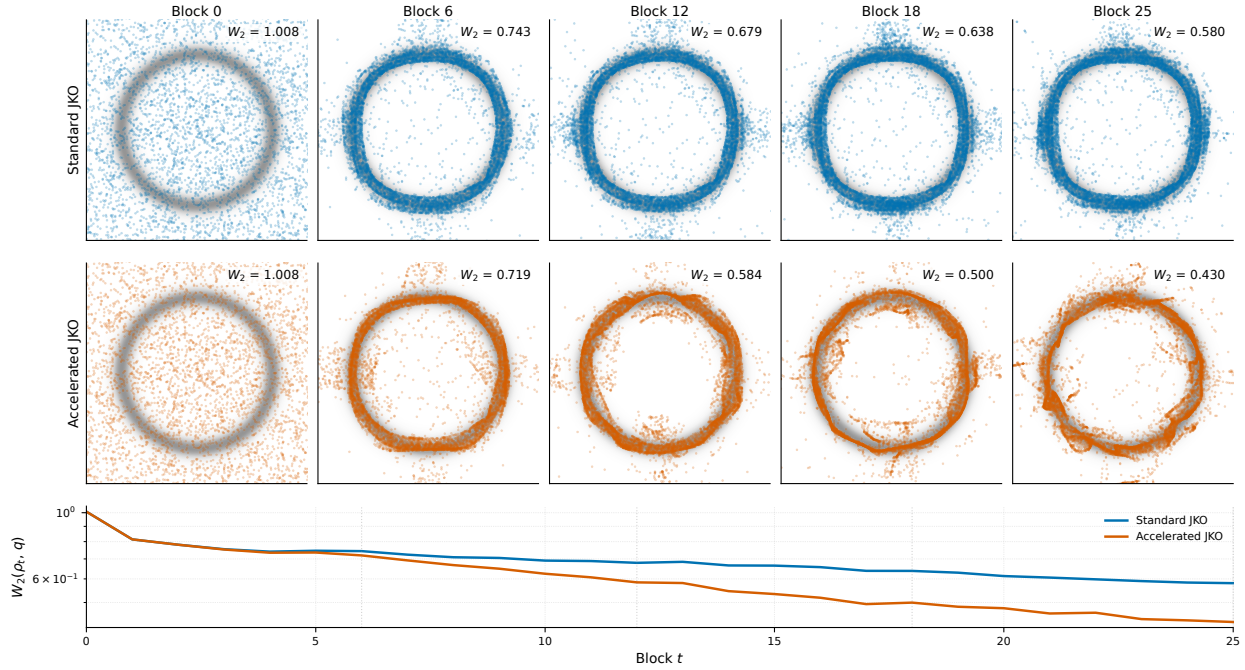


Figure 1: Particle transport on the outer-ring target ($\gamma = 0.04$, 25 blocks, 12,000 particles). Accelerated JKO ($W_2 = 0.429$) versus standard JKO ($W_2 = 0.611$) at block 25, on a non-log-concave target outside the strongly-convex regime where prior guarantees do not apply.

space $(\dot{\mathbb{H}}^1)^m$ with weighted norm $\sum_i w_i \|\cdot\|^2_{\dot{\mathbb{H}}^1}$. Discretization details, including how ρ_* enters via entropic transport plans, are in Appendix C.5.

5 Experiments

We validate both algorithms empirically. Full details appear in Appendix C.

Figure 1 shows particle transport on the outer-ring target $q \propto \exp\big(-(r-0.75)^2/(2\cdot 0.07^2)\big)$, a non-log-concave distribution outside the strongly-convex regime $\lambda > 0$ (cf. Remark 6). From $p_0 = \mathcal{N}(0, 1.5^2 I)$ with $\gamma = 0.04$, 25 blocks, and 12,000 particles, accelerated JKO reaches $W_2 = 0.429$ at block 25 versus $W_2 = 0.611$ for the standard scheme—a 30% reduction, widening with depth, consistent with Theorem 8. Implementation details (network architecture, training schedule, W_2 evaluation) are in Appendix C.1.

On 1-D Gaussian targets with closed-form proximal steps (Appendix C.2, Figure 2), the acceler-

ated gap oscillates around and stays uniformly below the $O(t^{-2})$ reference while the standard gap tracks the $O(t^{-1})$ reference. In the strongly convex regime, $\lambda = 1$, both converge exponentially at comparable rates, with $O(e^{-\gamma\lambda t/2})$ dominating any asymptotic gain from acceleration. A depth-scaling experiment (Appendix C.3, Figure 3) shows the standard final gap tracking the $O(1/N)$ reference and the accelerated final gap decaying at least as fast as $O(1/N^2)$, illustrating the λ-uniform bound of Theorem 8. On the non-log-concave double-well target (Appendix C.4, Figure 4), standard JKO makes negligible progress for the first 25 blocks, while the accelerated scheme begins converging around block 13 and descends below the $O(t^{-2})$ bound. Finally, for the Wasserstein barycenter problem (Appendix C.5, Figure 5), the ASGA dual gap tracks the $O(t^{-2})$ reference while SGA decays much more slowly over 400 iterations, with the gap ratio reaching $56\times$ at $t = 400$, consistent with Theorem 11.

6 Conclusion

We developed a unified accelerated proximal framework for optimization in Wasserstein and $\dot{\mathbb{H}}^1$ spaces, instantiated as an accelerated JKO scheme and an Accelerated Sobolev Gradient Ascent (ASGA) algorithm. Both follow the same three-sequence Nesterov-style recursion, differing only in how the proximal operator and momentum update are realized in their respective geometries. Accelerated JKO achieves $O(1/t^2)$ on the objective gap (with the squared Wasserstein distance to the optimum bounded uniformly in t) under mere geodesic convexity of G, removing the strong-convexity requirement of [19] and giving a bound uniform in λ where the nonaccelerated $O(\varepsilon^2/\lambda^2)$ floor diverges as $\lambda \to 0$. ASGA achieves $O(1/t^2)$ for the Kantorovich dual problem, a significant improvement over the $O(1/\sqrt{T})$ rate of [34] while retaining its projection-free structure. Experiments illustrate these rates and reveal a pronounced gap in the weakly- and non-strongly-convex regimes.

Limitations and future work. Our guarantees assume exact proximal steps and exact Sobolev gradients. Sharp error propagation under momentum is left open, and is the natural next step toward analyzing practical neural-JKO implementations where each proximal step incurs a learning error. The ASGA rate is established under either an exact proximal oracle for the unregularized Kantorovich dual (Theorem 11) or L-smoothness of the dual with explicit gradient (Remark 12). The latter is what our experiments instantiate via entropic regularization, but an unregularized proximal-oracle implementation is not addressed numerically. Our experiments are low-dimensional and use Gaussian marginals; high-dimensional and more structured targets remain to be studied.

References

[1] M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] K. Ahn and S. Sra. From nesterov's estimate sequence to riemannian acceleration. In *Conference on Learning Theory*, pages 84–118. PMLR, 2020.

[3] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[4] M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[5] Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

[6] J. Altschuler and E. Boix-Adsera. Wasserstein barycenters are np-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

[7] D. Alvarez-Melis, Y. Schiff, and Y. Mroueh. Optimizing functionals on the space of probabilities with input convex neural networks. *arXiv preprint arXiv:2106.00774*, 2021.

[8] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[9] B. Amos, L. Xu, and J. Z. Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

[10] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[11] J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

[12] S. Bubeck, Y. T. Lee, and M. Singh. A geometric alternative to nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[13] G. Carlier and M. Laborde. Convergence of entropic schemes for gradient flows in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 52(6):6291–6326, 2020.

[14] H. Chen, H. Lee, and J. Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

[15] S. Chen, S. Chewi, H. Lee, Y. Li, J. Lu, and A. Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36:68552–68575, 2023.

[16] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

[17] S. Chen, G. Daras, and A. Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023.

[18] S. Chen, Q. Li, O. Tse, and S. J. Wright. Accelerating optimization over the space of probability measures. *arXiv preprint arXiv:2310.04006*, 2023.

[19] X. Cheng, J. Lu, Y. Tan, and Y. Xie. Convergence of flow-based generative models via proximal gradient descent in wasserstein space. *IEEE Transactions on Information Theory*, 70(11):8087–8106, 2024.

[20] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[21] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300, 2013.

[22] V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

[23] P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. *arXiv preprint arXiv:1802.04367*, 2018.

[24] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

[25] C. Frogner and T. Poggio. Approximate inference with wasserstein gradient flows. In *International Conference on Artificial Intelligence and Statistics*, pages 2581–2590. PMLR, 2020.

[26] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *arXiv preprint arXiv:1605.08527*, 2016.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[28] B. Grimmer, K. Shu, and A. L. Wang. Accelerated gradient descent via long steps. *arXiv preprint arXiv:2309.09961*, 2023.

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[30] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[31] N. Ho, X. Nguyen, M. Yurochkin, H. Bui, V. Huynh, and D. Phung. Multilevel clustering via wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1501–1509, 2017.

[32] C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference on Learning Theory*, pages 1042–1085. PMLR, 2018.

[33] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[34] K. Kim, R. Yao, C. Zhu, and X. Chen. Optimal transport barycenter via nonconvex concave minimax optimization. In *International Conference on Machine Learning*, 2025.

[35] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[36] D. P. Kingma, M. Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[37] I. Kobyzev, S. J. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[38] A. Korotin, V. Egiazarian, A. Asadulaev, A. Safin, and E. Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021.

[39] A. Korotin, L. Li, J. Solomon, and E. Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*, 2021.

[40] H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

[41] H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

[42] G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards non-asymptotic convergence for diffusion-based generative models. The Twelfth International Conference on Learning Representations, 2024.

[43] T. Lin, N. Ho, X. Chen, M. Cuturi, and M. Jordan. Fixed-support wasserstein barycenters: Computational hardness and fast algorithm. *Advances in neural information processing systems*, 33:5368–5380, 2020.

[44] T. Lin, N. Ho, and M. Cuturi. Projection robust wasserstein barycenters. In *International Conference on Machine Learning*, pages 6719–6730. PMLR, 2021.

[45] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[46] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

[47] P. Mokrov, A. Korotin, L. Li, A. Genevay, J. M. Solomon, and E. Burnaev. Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34:15243–15256, 2021.

[48] Y. Nesterov. *A method for solving the convex programming problem with convergence rate O(1/k2̂)*. Soviet Mathematics Doklady, 1983.

[49] B. Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1771–1790, 2015.

[50] F. Pedrotti, J. Maas, and M. Mondelli. Improved convergence of score-based diffusion models via prediction-correction. *arXiv preprint arXiv:2305.14164*, 2023.

[51] G. Peyré and M. Cuturi. *Computational optimal transport*. Foundations and Trends® in Machine Learning, 2019.

[52] A. Salim, A. Korba, and G. Luise. The wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366, 2020.

[53] F. Santambrogio and G. Toshpulatov. Strong l 2 h 2 convergence of the jko scheme for the fokker–planck equation. *Archive for Rational Mechanics and Analysis*, 248(6):99, 2024.

[54] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

[55] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[56] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[57] S. Srivastava, C. Li, and D. B. Dunson. Scalable bayes via barycenter in wasserstein space. *Journal of Machine Learning Research*, 19(1):312–346, 2018.

[58] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[59] A. Vidal, S. Wu Fung, L. Tenorio, S. Osher, and L. Nurbekyan. Taming hyperparameter tuning in continuous normalizing flows using the jko scheme. *Scientific reports*, 13(1):4501, 2023.

[60] A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. In *Proceedings of the National Academy of Sciences*, volume 113, pages E7351–E7358, 2016.

[61] C. Xu, X. Cheng, and Y. Xie. Normalizing flow neural networks by jko scheme. *Advances in Neural Information Processing Systems*, 36:47379–47405, 2023.

[62] Y. Zemel and V. M. Panaretos. Fréchet means and procrustes analysis in wasserstein space. *Bernoulli*, 25(2):932–976, 2019.

[63] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on learning theory*, pages 1617–1638. PMLR, 2016.

[64] B. Zhou and M. Parno. Efficient and exact multimarginal optimal transport with pairwise costs. *Journal of Scientific Computing*, 100(1):25, 2024.

[65] Y. Zhuang, X. Chen, and Y. Yang. Wasserstein k-means for clustering probability distributions. In *Advances in Neural Information Processing Systems*, 2022.

[66] P. C. Álvarez Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

A Proofs for the Accelerated JKO Scheme

A.1 Proximal Subdifferential Identity

Lemma 13 (Exact proximal optimality for the accelerated JKO step)**.** *Let $G : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ be proper and lower semicontinuous, with $\mathrm{Dom}(G) \subset \mathcal{P}_2^r(\mathbb{R}^d)$. Fix $\gamma > 0$ and let*

$$x_{t+1} \in \operatorname*{argmin}_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ G(\rho) + \frac{1}{2\gamma} W_2^2(\rho, y_t) \right\}.$$

Assume $y_t, x_{t+1} \in \mathcal{P}_2^r(\mathbb{R}^d)$, and let $T_t := T_{y_t}^{x_{t+1}}$ be the optimal transport map from y_t to x_{t+1}. Then

$$\eta_{t+1} := \frac{Y_t - X_{t+1}}{\gamma} \circ X_{t+1}^{-1} = \frac{T_{x_{t+1}}^{y_t} - \mathrm{Id}}{\gamma} \in \partial_{W_2} G(x_{t+1}),$$

where $X_{t+1} := T_t \circ Y_t$ is any realization of x_{t+1} over a common base measure z_t and Y_t is any realization of y_t over the same base. Equivalently,

$$\frac{T_{x_{t+1}}^{y_t} - \mathrm{Id}}{\gamma} \in \partial_{W_2} G(x_{t+1}).$$

Proof. Define

$$F_t(\rho) := G(\rho) + \frac{1}{2\gamma} W_2^2(\rho, y_t).$$

Since x_{t+1} is an exact minimizer of F_t, we have

$$0 \in \partial_{W_2} F_t(x_{t+1}).$$

Now apply the proximal subdifferential identity (Lemma A.1 of [19], following Lemma 10.1.2 of [8]) with $p = y_t$ and $\rho = x_{t+1}$: if $\xi \in \partial_{W_2} F_t(x_{t+1})$, then

$$\xi + \frac{T_{x_{t+1}}^{y_t} - \mathrm{Id}}{\gamma} \in \partial_{W_2} G(x_{t+1}).$$

Taking $\xi = 0$ yields

$$\frac{T_{x_{t+1}}^{y_t} - \mathrm{Id}}{\gamma} \in \partial_{W_2} G(x_{t+1}).$$

Finally, if $y_t = (Y_t)_{\#} z_t$ and $x_{t+1} = (X_{t+1})_{\#} z_t$ with $X_{t+1} = T_t \circ Y_t$, then

$$T_{x_{t+1}}^{y_t} \circ X_{t+1} = Y_t \qquad z_t\text{-a.e.},$$

hence

$$\frac{T_{x_{t+1}}^{y_t} - \mathrm{Id}}{\gamma} \circ X_{t+1} = \frac{Y_t - X_{t+1}}{\gamma},$$

which is the stated formula for η_{t+1}. □

A.2 Proof of Lemma 7: One-Step Accelerated Descent

Proof. By definition of z_{t+1} and since $(Z_{t+1}, P_t)_{\#} z_t$ is a coupling of z_{t+1} and ρ^*,

$$W_2^2(z_{t+1}, \rho^*) \le \|Z_{t+1} - P_t\|_{z_t}^2.$$

Therefore

$$\begin{aligned}\frac{\alpha_t^2}{2\gamma} W_2^2(z_{t+1}, \rho^*) &\le \frac{\alpha_t^2}{2\gamma}\|Z_{t+1} - P_t\|_{z_t}^2 \\ &= \frac{\alpha_t^2}{2\gamma}\|\mathrm{Id} - P_t\|_{z_t}^2 + \frac{1}{2\gamma}\|X_{t+1} - Y_t\|_{z_t}^2 + \frac{\alpha_t}{\gamma}\langle \mathrm{Id} - P_t,\, X_{t+1} - Y_t\rangle_{z_t}. \end{aligned} \tag{11}$$

Since

$$\alpha_t(\mathrm{Id} - P_t) = Y_t - \big((1-\alpha_t)X_t + \alpha_t P_t\big),$$

the cross term can be rewritten as

$$\frac{\alpha_t}{\gamma}\langle \mathrm{Id} - P_t,\, X_{t+1} - Y_t\rangle_{z_t} = -\frac{1}{\gamma}\Big\langle (1-\alpha_t)X_t + \alpha_t P_t - Y_t,\, X_{t+1} - Y_t\Big\rangle_{z_t}.$$

Hence (11) becomes

$$\begin{aligned}\frac{\alpha_t^2}{2\gamma} W_2^2(z_{t+1}, \rho^*) \le \frac{\alpha_t^2}{2\gamma} W_2^2(z_t, \rho^*) &+ \frac{1}{2\gamma}\|X_{t+1} - Y_t\|_{z_t}^2 \\ &- \frac{1}{\gamma}\Big\langle (1-\alpha_t)X_t + \alpha_t P_t - Y_t,\, X_{t+1} - Y_t\Big\rangle_{z_t}. \end{aligned} \tag{12}$$

Now use Lemma 13. It gives

$$\eta_{t+1} := \frac{Y_t - X_{t+1}}{\gamma} \circ X_{t+1}^{-1} \in \partial_{W_2} G(x_{t+1}).$$

By Lemma 13 (the proximal subdifferential identity), $\eta_{t+1} \in \partial_{W_2} G(x_{t+1})$. We apply Proposition 3, in its extended form (Remark 4), twice, in each case with base $\nu = z_t$ and current point $\mu_1 = x_{t+1}$ realized via the transport map X_{t+1} from z_t to x_{t+1} (which is not in general the OT map):

- Taking $\mu_2 = \rho^*$ with OT map P_t from z_t,

$$G(\rho^*) - G(x_{t+1}) \;\ge\; \big\langle \eta_{t+1} \circ X_{t+1},\; P_t - X_{t+1}\big\rangle_{z_t}.$$

- Taking $\mu_2 = x_t$ with OT map X_t from z_t,

$$G(x_t) - G(x_{t+1}) \;\ge\; \big\langle \eta_{t+1} \circ X_{t+1},\; X_t - X_{t+1}\big\rangle_{z_t}.$$

Note that, in this extended form, Proposition 3 requires only that X_{t+1}, P_t, and X_t be transport maps from z_t to their respective targets (P_t is in fact the OT map from z_t to ρ^*, and X_t the OT map from z_t to x_t, but $X_{t+1} = T_t \circ Y_t$ need not be optimal); the induced maps $P_t \circ X_{t+1}^{-1}$ and $X_t \circ X_{t+1}^{-1}$ likewise need not be OT. Multiplying the first inequality by α_t, the second by $(1-\alpha_t)$, and summing gives

$$(1-\alpha_t)G(x_t) + \alpha_t G(\rho^*) - G(x_{t+1}) \ge \langle \eta_{t+1} \circ X_{t+1},\, (1-\alpha_t)X_t + \alpha_t P_t - X_{t+1}\rangle_{z_t}.$$

Using

$$\eta_{t+1} \circ X_{t+1} = \frac{Y_t - X_{t+1}}{\gamma},$$

this becomes

$$\begin{aligned}(1-\alpha_t)G(x_t) + \alpha_t G(\rho^*) - G(x_{t+1}) &\geq -\frac{1}{\gamma}\left\langle (1-\alpha_t)X_t + \alpha_t P_t - X_{t+1},\, X_{t+1} - Y_t\right\rangle_{z_t} \\ &= -\frac{1}{\gamma}\left\langle (1-\alpha_t)X_t + \alpha_t P_t - Y_t,\, X_{t+1} - Y_t\right\rangle_{z_t} \\ &\quad + \frac{1}{\gamma}\|X_{t+1} - Y_t\|_{z_t}^2. \end{aligned} \tag{13}$$

Rearranging (13),

$$-\frac{1}{\gamma}\left\langle (1-\alpha_t)X_t + \alpha_t P_t - Y_t,\, X_{t+1} - Y_t\right\rangle_{z_t} \leq (1-\alpha_t)G(x_t) + \alpha_t G(\rho^*) - G(x_{t+1}) - \frac{1}{\gamma}\|X_{t+1} - Y_t\|_{z_t}^2.$$

Insert this into (12):

$$\begin{aligned}\frac{\alpha_t^2}{2\gamma}W_2^2(z_{t+1},\rho^*) &\leq \frac{\alpha_t^2}{2\gamma}W_2^2(z_t,\rho^*) + \frac{1}{2\gamma}\|X_{t+1} - Y_t\|_{z_t}^2 \\ &\quad + (1-\alpha_t)G(x_t) + \alpha_t G(\rho^*) - G(x_{t+1}) - \frac{1}{\gamma}\|X_{t+1} - Y_t\|_{z_t}^2 \\ &= \frac{\alpha_t^2}{2\gamma}W_2^2(z_t,\rho^*) + (1-\alpha_t)G(x_t) + \alpha_t G(\rho^*) - G(x_{t+1}) - \frac{1}{2\gamma}\|X_{t+1} - Y_t\|_{z_t}^2 \\ &\leq \frac{\alpha_t^2}{2\gamma}W_2^2(z_t,\rho^*) + (1-\alpha_t)G(x_t) + \alpha_t G(\rho^*) - G(x_{t+1}).\end{aligned}$$

Move $G(\rho^*)$ to the left-hand side to conclude

$$\frac{\alpha_t^2}{2\gamma}W_2^2(z_{t+1},\rho^*) + G(x_{t+1}) - G(\rho^*) \leq \frac{\alpha_t^2}{2\gamma}W_2^2(z_t,\rho^*) + (1-\alpha_t)\big(G(x_t) - G(\rho^*)\big).$$

□

A.3 Proof of Theorem 8: Accelerated JKO Convergence

Proof. By Lemma 7,

$$\Delta_{t+1} = \frac{\lambda_{t+1}}{2}W_2^2(z_{t+1},\rho^*) + G(x_{t+1}) - G(\rho^*) \leq \frac{\lambda_{t+1}}{2}W_2^2(z_t,\rho^*) + (1-\alpha_t)\big(G(x_t) - G(\rho^*)\big).$$

Since $\lambda_t(1-\alpha_t) \leq \lambda_{t+1}$,

$$\Delta_{t+1} \leq \frac{\lambda_{t+1}}{\lambda_t}\left(\frac{\lambda_t}{2}W_2^2(z_t,\rho^*) + G(x_t) - G(\rho^*)\right) = \frac{\lambda_{t+1}}{\lambda_t}\Delta_t.$$

Dividing by λ_{t+1} yields

$$\frac{\Delta_{t+1}}{\lambda_{t+1}} \leq \frac{\Delta_t}{\lambda_t}.$$

Iterating from $t = 1$,

$$\frac{\Delta_t}{\lambda_t} \leq \frac{\Delta_1}{\lambda_1}, \qquad \text{hence} \qquad \Delta_t \leq \lambda_t\frac{\Delta_1}{\lambda_1}, \qquad t \geq 1.$$

At $t = 0$, the choice $\alpha_0 = 3/(0+3) = 1$ makes the coefficient $(1-\alpha_0)$ in Lemma 7 vanish, so

$$\Delta_1 \le \frac{\lambda_1}{2} W_2^2(z_0, \rho^*), \qquad \text{i.e.,} \qquad \frac{\Delta_1}{\lambda_1} \le \frac{1}{2} W_2^2(z_0, \rho^*).$$

Combining,

$$\Delta_t \le \frac{\lambda_t}{2}\, W_2^2(z_0, \rho^*).$$

Since $G(x_t) - G(\rho^*) \le \Delta_t$, the same bound holds for the objective gap. For the choice $\alpha_t = 3/(t+3)$, $\lambda_t = 9/(\gamma(t+2)^2)$, so

$$G(x_t) - G(\rho^*) \le \frac{9}{2\gamma(t+2)^2}\, W_2^2(z_0, \rho^*).$$

This proves the claim. □

B Proofs for the Accelerated Sobolev Gradient Ascent

B.1 Proof of Lemma 10: One-Step Accelerated Ascent

Proof. For brevity write

$$y_t := y^{(t)}, \quad f_t := f^{(t)}, \quad f_{t+1} := f^{(t+1)}, \quad z_t := z^{(t)}, \quad z_{t+1} := z^{(t+1)}, \quad g_t := \nabla \mathcal{I}(f_{t+1}),$$

and define the Lyapunov increment

$$\Delta_{t+1} := \mathcal{I}(\tilde{f}) - \mathcal{I}(f_{t+1}) + \frac{\alpha_t^2}{2\eta} \|z_{t+1} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2.$$

Our goal is to show $\Delta_{t+1} \le \frac{\alpha_t^2}{2\eta}\|z_t - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + (1-\alpha_t)\big(\mathcal{I}(\tilde{f}) - \mathcal{I}(f_t)\big)$.

Step 1 (expand the squared norm). Since $z_{t+1} = z_t + \frac{\eta}{\alpha_t} g_t$,

$$\frac{\alpha_t^2}{2\eta}\|z_{t+1} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 = \frac{\alpha_t^2}{2\eta}\|z_t - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + \alpha_t \langle z_t - \tilde{f}, g_t\rangle_{\dot{\mathbb{H}}^1} + \frac{\eta}{2}\|g_t\|_{\dot{\mathbb{H}}^1}^2. \tag{14}$$

The relation $y_t = (1-\alpha_t) f_t + \alpha_t z_t$ gives $\alpha_t(z_t - \tilde{f}) = y_t - (1-\alpha_t) f_t - \alpha_t \tilde{f}$, so the cross term in (14) becomes

$$\alpha_t \langle z_t - \tilde{f}, g_t\rangle = \langle y_t - (1-\alpha_t) f_t - \alpha_t \tilde{f}, g_t\rangle = \langle f_{t+1} - \eta g_t - (1-\alpha_t) f_t - \alpha_t \tilde{f}, g_t\rangle,$$

where the last equation uses the update rule on line 6 in Algorithm 2. Plugging this into the definition of Δ_{t+1},

$$\begin{aligned}
\Delta_{t+1} &= \mathcal{I}(\tilde{f}) - \mathcal{I}(f_{t+1}) + \frac{\alpha_t^2}{2\eta}\|z_t - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + \langle f_{t+1} - \eta g_t - (1-\alpha_t) f_t - \alpha_t \tilde{f}, g_t\rangle_{\dot{\mathbb{H}}^1} + \frac{\eta}{2}\|g_t\|_{\dot{\mathbb{H}}^1}^2 \\
&= \mathcal{I}(\tilde{f}) - \mathcal{I}(f_{t+1}) + \frac{\alpha_t^2}{2\eta}\|z_t - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + \langle f_{t+1} - (1-\alpha_t) f_t - \alpha_t \tilde{f}, g_t\rangle_{\dot{\mathbb{H}}^1} + \left(\frac{\eta}{2} - \eta\right)\|g_t\|_{\dot{\mathbb{H}}^1}^2 \\
&\le \mathcal{I}(\tilde{f}) - \mathcal{I}(f_{t+1}) + \frac{\alpha_t^2}{2\eta}\|z_t - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + \langle f_{t+1} - (1-\alpha_t) f_t - \alpha_t \tilde{f}, g_t\rangle_{\dot{\mathbb{H}}^1}
\end{aligned} \tag{15}$$

Step 2 (bound the cross term by concavity). Concavity of $\mathcal{I}$ at f_{t+1} gives, for any $w \in \dot{\mathbb{H}}^1$, $\mathcal{I}(w) \le \mathcal{I}(f_{t+1}) + \langle g_t, w - f_{t+1}\rangle$. Applying this with $w = f_t$ and $w = \tilde{f}$ and combining with weights $(1-\alpha_t)$ and α_t,

$$(1-\alpha_t)\mathcal{I}(f_t) + \alpha_t \mathcal{I}(\tilde{f}) \le \mathcal{I}(f_{t+1}) + \langle g_t,\, (1-\alpha_t)f_t + \alpha_t \tilde{f} - f_{t+1}\rangle.$$

Rearranging,

$$\langle f_{t+1} - (1-\alpha_t)f_t - \alpha_t \tilde{f}, g_t\rangle \le \mathcal{I}(f_{t+1}) - (1-\alpha_t)\mathcal{I}(f_t) - \alpha_t \mathcal{I}(\tilde{f}). \tag{16}$$

Step 3 (combine). Substituting this bound into (15),

$$\Delta_{t+1} \le \frac{\alpha_t^2}{2\eta}\|z_t - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + \mathcal{I}(\tilde{f}) - \mathcal{I}(f_{t+1}) + \mathcal{I}(f_{t+1}) - (1-\alpha_t)\mathcal{I}(f_t) - \alpha_t \mathcal{I}(\tilde{f}).$$

The $\mathcal{I}(f_{t+1})$ terms cancel and $\mathcal{I}(\tilde{f}) - \alpha_t \mathcal{I}(\tilde{f}) = (1-\alpha_t)\mathcal{I}(\tilde{f})$, yielding

$$\Delta_{t+1} \le \frac{\alpha_t^2}{2\eta}\|z_t - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + (1-\alpha_t)\big(\mathcal{I}(\tilde{f}) - \mathcal{I}(f_t)\big),$$

which is the claimed estimate. □

B.2 Proof of Theorem 11: Accelerated ASGA Convergence

Proof. Define

$$\lambda_{t+1} := \frac{\alpha_t^2}{\eta}$$

and the Lyapunov quantity

$$\Phi_t := \frac{\lambda_t}{2}\|z^{(t)} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + \mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(t)}).$$

Lemma 10 states precisely that

$$\Phi_{t+1} \le \frac{\lambda_{t+1}}{2}\|z^{(t)} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + (1-\alpha_t)\big(\mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(t)})\big).$$

If

$$\lambda_t(1-\alpha_t) \le \lambda_{t+1}, \tag{17}$$

then

$$\Phi_{t+1} \le \frac{\lambda_{t+1}}{\lambda_t}\left(\frac{\lambda_t}{2}\|z^{(t)} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2 + \mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(t)})\right) = \frac{\lambda_{t+1}}{\lambda_t}\Phi_t.$$

Dividing by λ_{t+1} yields

$$\frac{\Phi_{t+1}}{\lambda_{t+1}} \le \frac{\Phi_t}{\lambda_t},$$

so by iteration from $t = 1$,

$$\Phi_t \le \lambda_t \frac{\Phi_1}{\lambda_1}, \qquad t \ge 1.$$

At $t = 0$, the choice $\alpha_0 = 3/(0+3) = 1$ makes the coefficient $(1-\alpha_0)$ in Lemma 10 vanish, so

$$\Phi_1 \le \frac{\lambda_1}{2}\|z^{(0)} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2, \qquad \text{i.e.,} \qquad \frac{\Phi_1}{\lambda_1} \le \frac{1}{2}\|z^{(0)} - \tilde{f}\|_{\dot{\mathbb{H}}^1}^2.$$

Combining,

$$\mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(t)}) \le \Phi_t \le \frac{\lambda_t}{2}\|z^{(0)} - \tilde{f}\|^2_{\dot{\mathbb{H}}^1}.$$

For $\alpha_t = 3/(t+3)$ we have $\lambda_{t+1} = 9/(\eta(t+3)^2)$, so $\lambda_t = 9/(\eta(t+2)^2)$. The coefficient condition (17) reduces to

$$\lambda_t(1-\alpha_t) = \frac{9}{\eta(t+2)^2}\cdot\frac{t}{t+3} \le \frac{9}{\eta(t+3)^2} = \lambda_{t+1} \quad\Longleftrightarrow\quad t(t+3) \le (t+2)^2 \quad\Longleftrightarrow\quad 0 \le t+4,$$

which holds for all $t \ge 0$. We conclude that

$$\mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(t)}) \le \frac{9}{2\eta(t+2)^2}\,\|z^{(0)} - \tilde{f}\|^2_{\dot{\mathbb{H}}^1},$$

as claimed. □

C Experiments

We present five experiments validating the guarantees of Sections 4.1–4.2: the 2D outer-ring particle transport from Figure 1 (Appendix C.1), three further experiments on the accelerated JKO scheme (Theorem 8), and one for ASGA (Theorem 11). All experiments were run on CPU and the code is included in the supplementary material.

C.1 Accelerated JKO: 2D Outer-Ring Particle Transport (Figure 1)

Setup. We compare standard and accelerated JKO on a 2D non-log-concave target

$$q(x) \propto \exp\Big(-(\|x\| - 0.75)^2/(2\cdot 0.07^2)\Big), \qquad x \in [-1,1]^2,$$

with $G(\rho) = \mathrm{KL}(\rho\|q)$, step size $\gamma = 0.04$, and $N = 25$ blocks.

Iterate representation. The iterate ρ_t is a fixed cloud of $n = 12{,}000$ particles whose locations are updated each block (no reweighting; uniform empirical mass throughout the run). The accelerated variant additionally maintains an auxiliary cloud z_t of the same size, with particle indices shared across the two clouds (so the implicit coupling between z_t and x_t is the index-identity coupling, and $T_t \circ Y_t$ is realized as the index-aligned application of the trained map). Initialization: $x_0 \sim \mathcal{N}(0, 1.5^2 I_2)$ with 12,000 i.i.d. samples; $z_0 = x_0$.

Proximal step. Each block solves

$$x_{t+1} = \arg\min_{\rho}\Big\{\mathrm{KL}(\rho\|q) + \tfrac{1}{2\gamma}W_2^2(\rho, y_t)\Big\}$$

by training a fresh residual MLP $T_\theta(x) = x + \mathrm{MLP}_\theta(x)$ with three hidden layers of width 256 and SiLU activations (final layer zero-initialized so T_θ starts as the identity). The block-level objective is

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim y_t}\Big[-\log q(T_\theta(x)) - \log\big|\det \nabla T_\theta(x)\big|\Big] + \frac{1}{2\gamma}\,\mathbb{E}_{x\sim y_t}\Big[\|T_\theta(x) - x\|^2\Big],$$

the first term being $\mathrm{KL}((T_\theta)_{\#}y_t \,\|\, q)$ up to an x-independent constant. The Jacobian determinant is computed via `vmap(jacrev(...))` and `torch.linalg.slogdet`. Optimization: 800 epochs of mini-batch Adam, batch size 1024, learning rate cosine-annealed from 2×10^{-3} to 10^{-4}, gradient-norm clipping at 5. After training, particles are clamped to $[-2.5, 2.5]^2$ (x_t) and $[-3.5, 3.5]^2$ (z_t) for numerical stability.

Pushforward and momentum. The extrapolated measure y_t is realized as a particle-wise convex combination $y_t = (1-\alpha_t)x_t + \alpha_t z_t$ over index-matched particles, with $\alpha_t = 3/(t+3)$. After training T_θ to approximate the proximal map at y_t, the updates are $x_{t+1} = T_\theta(y_t)$ and $z_{t+1} = z_t + (x_{t+1} - y_t)/\alpha_t$, applied per particle. The standard baseline uses the same MLP architecture, training schedule, and clamping bounds, but omits the z_t cloud and the momentum interpolation.

Target representation and $\log q$. We discretize q as a 256×256 image on $[-1, 1]^2$, normalize to sum to 1, and take logs (with the maximum shifted to 0; this affects only an additive constant and not $\nabla \log q$). $\log q(x)$ is then evaluated for arbitrary x via bilinear `grid_sample` with border padding, yielding a smooth and differentiable surrogate. Samples from q are drawn via inverse-CDF on the flattened grid plus uniform within-cell jitter.

Reporting. The reported W_2 at block 25 is computed as the Sinkhorn divergence (`geomloss.SamplesLoss` with $p = 2$, blur $= 0.01$, scaling $= 0.95$) between a uniform subsample of 4,096 iterate-particles and a fixed reference of 4,096 i.i.d. samples from q drawn once at the start of the run. Reported numbers are therefore Sinkhorn-divergence approximations, not exact W_2.

Reproducibility. Random seed 0 is applied at the start of each run via `torch.manual_seed` and `numpy.random.seed`; the standard and accelerated runs re-seed independently so they begin from the same initial particle cloud. The reference sample for W_2 uses seed 777. Code is included in the supplementary material.

Results. The accelerated scheme reaches $W_2 = 0.429$ at block 25 versus $W_2 = 0.611$ for the standard scheme—a 30% reduction that widens with depth (Figure 1). The target is non-log-concave, falling outside the strongly-convex regime where the analysis of [19] provides guarantees, while Theorem 8 applies under the geodesic-convexity assumption alone.

C.2 Accelerated JKO: Convergence Rates on Gaussian Targets

Setup. We work in $d = 1$ with target $q = \mathcal{N}(0, \lambda^{-1})$ and objective $G(\rho) = \mathrm{KL}(\rho \| q)$. The JKO proximal step for this Gaussian family admits a closed-form solution [19], so every iterate is exact ($\varepsilon = 0$). We initialize at $\rho_0 = \mathcal{N}(5, 2.5^2)$, set $\gamma = 0.5$, and compare standard JKO ($\rho_{n+1} = \mathrm{Prox}_{\gamma G}(\rho_n)$) against Algorithm 1 with $\alpha_t = 3/(t+3)$.

Results. Figure 2 shows results for two values of λ.

At $\lambda = 0.04$ (panel (a), log-log scale), λ is small enough that, over the iteration horizon shown, standard JKO is in a transient regime where the gap tracks the $O(t^{-1})$ reference, before the geometric rate $(1 + \gamma\lambda/2)^{-t}$ would eventually dominate. The accelerated gap exhibits oscillatory behavior but stays uniformly below the $O(t^{-2})$ reference. The ratio of gaps reaches over 90,000 by $t = 300$. At $\lambda = 1$ (panel (b), log-linear scale), both methods converge exponentially, consistent with [19]. Standard JKO converges faster in this strongly convex regime, as expected: the exponential rate $O(e^{-\gamma\lambda t/2})$ dominates any polynomial gain from acceleration when λ is large.

C.3 Accelerated JKO: Depth Scaling and Behavior as $\lambda \to 0$

Setup. Using the same Gaussian setting, we run two further comparisons. For the depth sweep we fix $\lambda = 0.04$, $\gamma = 0.5$, and vary $N \in \{4, 8, 16, 32, 64, 128\}$, recording $G(\rho_N) - G(\rho^*)$. For the λ sweep we fix $N = 200$, $\gamma = 0.5$, and vary $\lambda \in \{1.0, 0.5, \ldots, 0.001\}$.

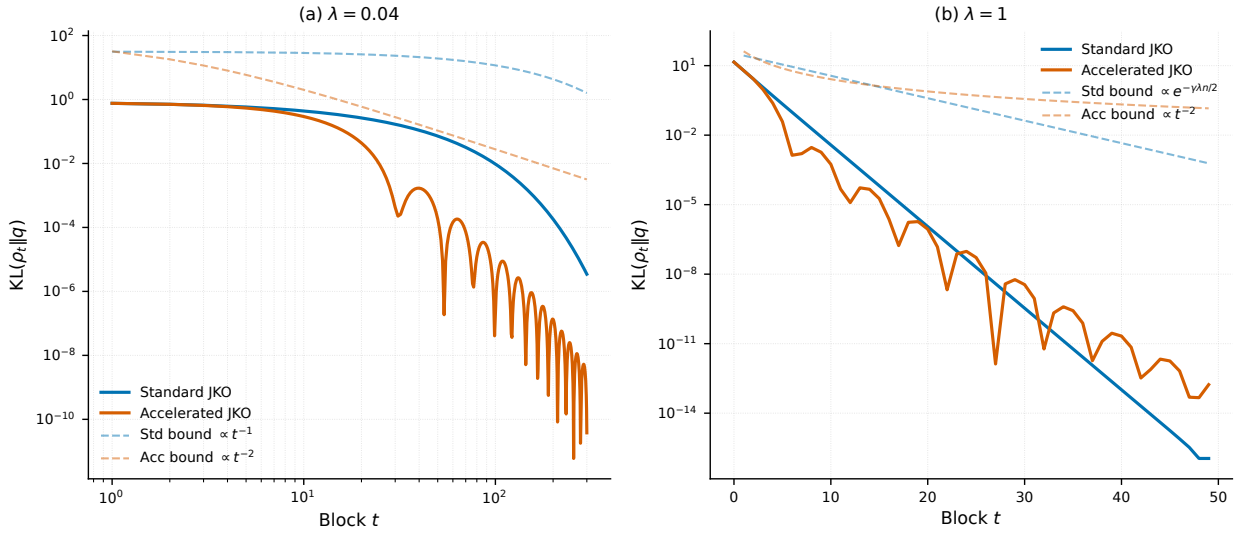


Figure 2: Standard vs. accelerated JKO on 1-D Gaussian targets ($\gamma = 0.5$, initialized at $\mathcal{N}(5, 2.5^2)$). (a) $\lambda = 0.04$, log-log scale. (b) $\lambda = 1$, log-linear scale. Dashed lines show $O(t^{-1})$ and $O(t^{-2})$ references. The accelerated gap stays uniformly below the $O(t^{-2})$ reference in the weakly convex regime, consistent with Theorem 8, while the standard scheme tracks the $O(t^{-1})$ reference. Both methods converge exponentially at comparable rates in the strongly convex regime.

Results. Figure 3 shows the results.

In the depth sweep (panel (a)), the accelerated final gap shrinks much faster with N than the standard final gap, consistent with the predicted $O(1/N^2)$ vs. $O(1/N)$ rates. The ratio between them grows with N and reaches 1547× at $N = 128$. The λ sweep (panel (b)) illustrates how the finite-time contraction rate of standard JKO degrades as $\lambda \to 0$. With closed-form Gaussian proximal steps every iterate is exact ($\varepsilon = 0$), so the asymptotic error floor is zero for both methods; the widening gap reflects that for fixed $N = 200$, the geometric contraction $(1 + \gamma\lambda/2)^{-N}$ in [19] approaches 1 as $\lambda \to 0$, while the accelerated bound is uniform in λ. The two curves separate by two to four orders of magnitude across $\lambda \in [10^{-3}, 10^{-1}]$.

C.4 Accelerated JKO Beyond the Strongly Convex Regime: Double-Well Target

Setup. We demonstrate the qualitative gap of Remark 6 on a non-Gaussian, non-log-concave target outside the strongly-convex regime:

$$q = \tfrac{1}{2}\mathcal{N}(2.0,\, 0.7^2) + \tfrac{1}{2}\mathcal{N}(-2.0,\, 0.7^2).$$

We use the variational family $\rho(m, s) = \frac{1}{2}\mathcal{N}(m, s^2) + \frac{1}{2}\mathcal{N}(-m, s^2)$, which contains q exactly. Each proximal step is solved by L-BFGS-B over $(m, \log s)$ with $\mathrm{KL}(\rho\|q)$ computed by numerical quadrature; $\gamma = 0.5$, $N = 40$ blocks.

Results. Figure 4 shows $\mathrm{KL}(\rho_t\|q)$ on a log-linear scale. Standard JKO makes essentially no progress for the first 25 blocks, reflecting the absence of a strong-convexity-driven contraction factor on this target. The accelerated scheme begins converging around block 13 and descends below the $O(t^{-2})$ reference line, indicating that the theoretical bound is not tight in this setting.

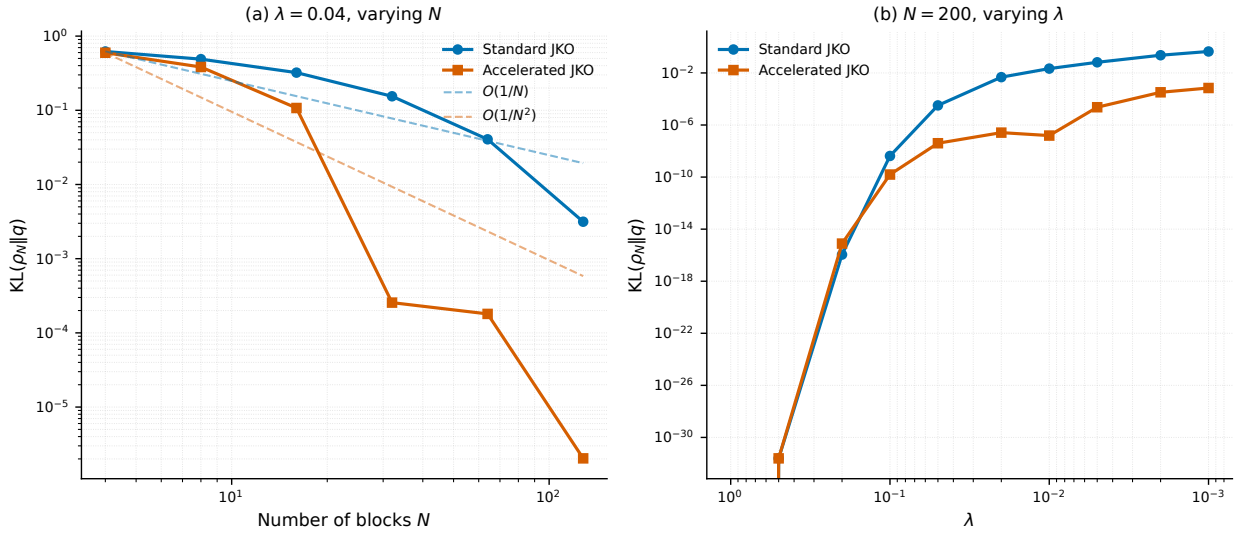


Figure 3: Further Gaussian comparisons ($\gamma = 0.5$). (a) Final gap vs. depth N at $\lambda = 0.04$. (b) Final gap after $N = 200$ blocks vs. λ. Panel (a) shows the standard final gap tracking the $O(1/N)$ reference and the accelerated final gap decaying at least as fast as $O(1/N^2)$, with a ratio of $1547\times$ at $N = 128$. Panel (b) illustrates the λ-uniform bound of Theorem 8: the accelerated gap grows only mildly as $\lambda \to 0$, while the standard gap grows by orders of magnitude.

C.5 ASGA: Convergence on the Wasserstein Barycenter Problem

Setup. We validate the $O(1/t^2)$ ASGA rate (Theorem 11 and Remark 12) on the barycenter problem with $m = 4$ equally weighted Gaussian marginals $\mu_i = \mathcal{N}(c_i, 0.07^2)$ at centers $c_i \in \{0.15, 0.35, 0.65, 0.85\}$ on a uniform grid of $n = 60$ points in $[0, 1]$, with entropic regularization $\varepsilon = 0.008$. The regularization renders the dual functional $\mathcal{I}^\varepsilon$ L-smooth, so we instantiate the smooth-explicit variant of Algorithm 2 (Remark 12): the proximal step on Line 5 is replaced by the explicit forward step $f^{(t+1)} \leftarrow y^{(t)} + \eta \nabla \mathcal{I}^\varepsilon(y^{(t)})$. Both SGA (fixed step η, last iterate) and the explicit ASGA ($\alpha_t = 3/(t+3)$, last iterate) are initialized at $f_i^{(0)} = 0$. The Sobolev gradient $\nabla \mathcal{I}^\varepsilon(f_i) = (-\Delta)^{-1}(\mu_i - \pi_i^{\text{row}})$, where π_i^{row} denotes the first marginal of the entropic optimal transport plan between μ_i and the current dual variable, is computed via a tridiagonal Poisson solve (Neumann boundary conditions). The Lipschitz constant is estimated empirically by finite differences, giving $L \approx 0.015$ and $\eta \approx 67$. The reference $\mathcal{I}^*$ is approximated by 3000 ASGA iterations.

Results. Figure 5 shows the dual gap $\mathcal{I}(\tilde{f}) - \mathcal{I}(f^{(t)})$ over 400 iterations on a log-log scale. The ASGA gap stays close to the $O(t^{-2})$ reference while the SGA gap decays much more slowly; the ratio between them grows to $56\times$ at $t = 400$.

Remark 14. *The baseline here is smooth gradient ascent at rate $O(1/T)$, not the subgradient rate $O(1/\sqrt{T})$ reported for SGA in [34].*

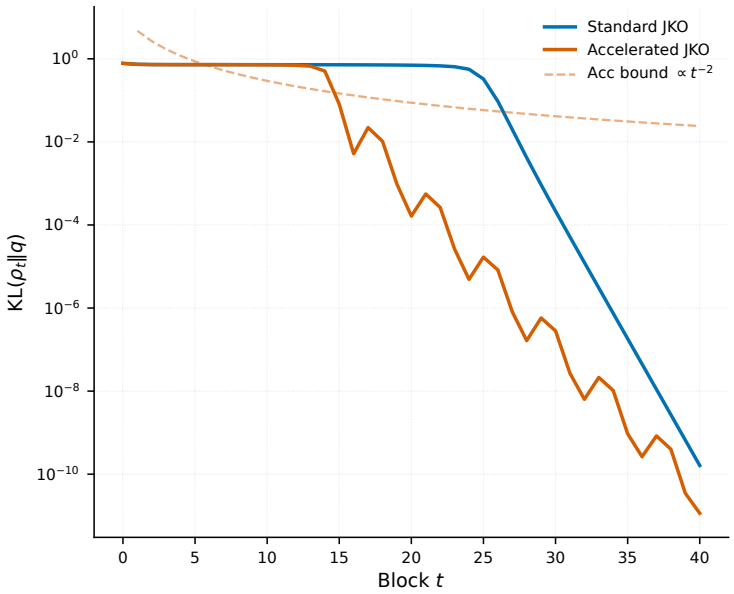


Figure 4: Standard vs. accelerated JKO on the double-well target $q = \frac{1}{2}\mathcal{N}(2, 0.7^2) + \frac{1}{2}\mathcal{N}(-2, 0.7^2)$ ($\gamma = 0.5$, 40 blocks). Log-linear scale; dashed line is the $O(t^{-2})$ theoretical bound. Standard JKO stalls for 25 blocks since no strong-convexity contraction is available on this non-log-concave target, whereas the accelerated scheme begins converging around block 13 and descends below the theoretical bound, indicating that the bound is not tight in this setting.

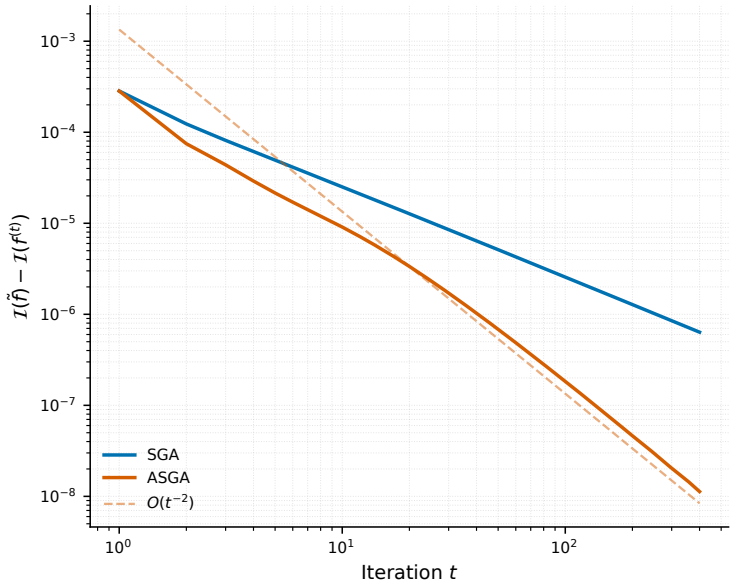


Figure 5: SGA vs. ASGA on the Wasserstein barycenter dual ($m = 4$ Gaussian marginals, $n = 60$, $\varepsilon = 0.008$, $\eta \approx 67$). Log-log scale; dashed line is the $O(t^{-2})$ reference. The ASGA gap stays close to the $O(t^{-2})$ reference, with a gap ratio of 56× over SGA at $t = 400$.