ACCELERATED-JKO: DISCRETIZING WASSERSTEIN ACCELERATED-INFORMATION GRADIENT FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

Wasserstein gradient flows (WGF), discretized by the JKO scheme, are widely adopted in optimization, sampling and generative model. JKO inherits the $O(e^{-\beta t})$ convergence rate of WGF, which degrades for densities that are not strongly log-concave. Accelerated Information Gradient Flow augments the density with a momentum for improved convergence rate, but existing realizations integrate the accelerated particle ODE explicitly which does not fully leverage the well-posed benefit of the JKO proximal step. We introduce A-JKO and its folding-free variant IA-JKO, momentum-augmented JKO schemes that discretize the AIGF momentum dynamics while keeping each step a well-posed variational problem, and prove that, as the step size vanishes, both schemes converge to the Wasserstein AIGF. Building on this scheme, we construct a neural sampler that transports a prior to the target through a flow-based velocity field, trained with a off-policy matching loss derived from the JKO optimality condition. On synthetic (DW-4, LJ-13, LJ-55) and molecular (alanine dipeptide) benchmarks, our sampler outperforms conventional JKO-based neural samplers and is competitive with several recent neural samplers.

1 INTRODUCTION

Wasserstein gradient flows (WGF) describe steepest descent of an energy functional E on the space of probability measures $\mathcal{P}_2(\mathbb{R}^d)$, and underlie a broad range of optimization (Fan et al., 2022; Nguyen et al., 2025), sampling algorithms (Lambert et al., 2022; Hertrich & Gruhlke, 2025), and generative models (Bunne et al., 2022; Xu et al., 2023; 2025). Their canonical time discretization is the Jordan–Kinderlehrer–Otto (JKO) scheme (Jordan et al., 1998), which replaces the flow by a sequence of proximal problems which can be solved both in classical computational methods and learning-based approach. Each step is a well-posed variational problem, the energy decreases for every step size, and the iterates provably converge to the gradient flow as $\tau \to 0$ (Ambrosio et al., 2008). This minimizing-movement structure has made the JKO scheme the backbone of many computational methods, including neural implementations that realize each proximal step with a continuous normalizing flow (CNF) (Xu et al., 2023; Vidal et al., 2023; Hertrich & Gruhlke, 2025).

The JKO scheme inherits the dynamic of WGF: for λ-convex energy function, the energy decays at first-order with $O(e^{-\beta t})$ rate which degenerates as $\lambda \to 0$. Nesterov momentum in Euclidean space alleviates this linear rate from β to $\sqrt{\beta}$ based on the damped second-order ordinary differential equation (Su et al., 2016). On $\mathcal{P}_2$ and L_2, Taghvaei & Mehta (2019) and Chow et al. (2019) proposes the Accelerated flow and Hamiltonian flow by augmenting the density with potential. Inspired by these potential-augmented gradient flow methods, Accelerated Information Gradient Flow (AIGF) generalizes them to different information metric with additional friction to recover the Nesterov style momentum demonstrates the improved $O(e^{-\sqrt{\beta} t})$ rate convergence (Wang & Li, 2022). Existing implementations integrate the particle ODE explicitly with interacting particles and estimate the entropic force $\nabla \log \rho_t$ from samples, e.g., by kernel density estimation.

Blessing et al. (2025) introduced underdamped diffusion as an alternative of the overdamped diffusion, augmenting the state space with the momentum variable and aligning the path in the joint space with degenerate noise in the sampling problem. Motivated by Ma et al. (2021), the Nesterov-type momentum variable provides a higher-order convergence rate compared to that of the overdamped Langevin which demonstrates both theoretical benefit and an empirical success. Concurrently, the application of JKO scheme to simulate the gradient flow demonstrated its empirical success over

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

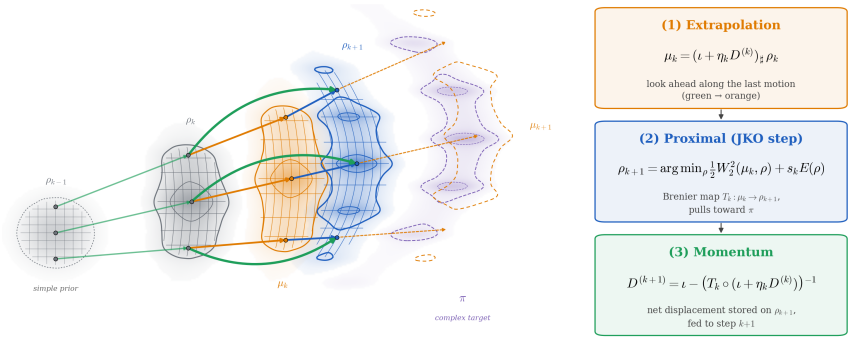

Figure 1: Visualization of our Accelerated-JKO scheme.

neural samplers (Hertrich & Gruhlke, 2025). These two lines of work, however, remain separate: the acceleration benefits of underdamped diffusion have so far been realized only through explicit-integration samplers, while the stability benefits of the JKO scheme have so far been realized only for the overdamped flow. The integration of a variational, minimizing-movement time discretization that simultaneously inherits the acceleration of the AIGF and the well-posedness of the JKO proximal step has not yet been investigated.

To address above gaps we ask the following question:

> *Does the Wasserstein AIGF admit a JKO scheme, i.e., a sequence of well-posed variational problems on $\mathcal{P}_2$ that simultaneously discretizes the momentum dynamics and provably converges to the accelerated flow?*

We answer this question with two variational schemes: A-JKO and its folding-free variant IA-JKO, that discretize the momentum dynamics while preserving the well-posedness of the JKO proximal step with theoretical guarantee on their convergence. We summarize our contributions below:

Contributions.

- We propose Accelerated-JKO scheme, a momentum-augmented counterpart of the JKO scheme, that asymptotically converges to the Wasserstein-AIGF. We further provide the theoretical guarantee that the Accelerated JKO converges to the AIGF.
- We construct a neural sampler that transports samples from prior to the target density via iterative transformation by flow-based velocity field. Particularly, we propose an off-policy matching loss by leveraging the JKO optimality condition to leverage replay buffer for better training efficiency.
- We demonstrate its empirical success over the conventional JKO-based samplers and comparable performance to the many of recent neural samplers on various synthetic target distribution (DW-4, LJ-13/55) and the task of sampling conformers of peptides (Alanine Dipeptides).

2 Related Work

Wasserstein Gradient Flows and JKO Scheme JKO is a variational framework to discretize WGF, which is the steepest descent of an energy across the probability measure space $\mathcal{P}_2(\mathbb{R}^d)$ (Jordan et al., 1998; Ambrosio et al., 2008). For KL divergence, an explicit step on the potential energy followed by the exact heat flow on the entropy recovers the unadjusted Langevin algorithm (Wibisono, 2018; Bernton, 2018), while a JKO step on the entropy instead yields an unbiased forward–backward scheme (Salim et al., 2020). Earlier work computed this transport map through numerical optimal transport techniques such as Sinkhorn iteration (Peyré & Cuturi, 2019; Carlier et al., 2017). Recent neural JKO methods parameterize the map as the gradient of an input-convex network (Mokrov et al., 2021; Alvarez-Melis et al., 2022; Bunne et al., 2022), as the flow map $v_\theta(\cdot)$ of a continuous

normalizing flow whose kinetic energy surrogates $W_2^2(\cdot)$ (Onken et al., 2021; Vidal et al., 2023; Xu et al., 2023), or through density-free variational reformulations (Fan et al., 2022). All of these discretize a first-order flow which we propose an alternative discretization of the Wasserstein AIG flow of Wang & Li (2022), whose momentum is a potential on position space.

Neural Sampler Path-guided samplers have since developed into path alignment methods (Zhang & Chen, 2022; Berner et al., 2024; Richter & Berner, 2024), with recent work further improving training efficiency by leveraging off-policy replay-buffer (Liu et al., 2025; Blessing et al., 2026; Havens et al., 2026; Berner et al., 2026). The application of our method on sampling problem is aligned with Blessing et al. (2025) which leverages the underdamped diffusion bridge to generate samples which is known to demonstrate a faster order convergence to the target by incorporating Nesterov-type momentum (Ma et al., 2021). Particularly, Hertrich & Gruhlke (2025) provides a preceding line of work of using JKO formulation and Continuous Normalizing Flow to transport the prior samples to the target with simulation-based training. Yet, this work correspond to the overdamped diffusion case without augmentation of momentum and requires expensive rejection-based importance correction which exhibits limited scalability to the high dimensions.

3 PRELIMINARIES

Accelerated Information Gradient Flow We define the probability space of finite second order momentum $\mathcal{P}_2(\mathbb{R}^d)$ and define a respective metric of Wasserstein-2 distance as

$$W_2^2(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)} \int \|x-y\|^2 d\gamma$$

where $\Pi(\mu,\nu)$ is the space of couplings between probability measures μ,ν. Define the functional $E : L_2(\mathbb{R}^d) \to \mathbb{R}$ with a well-defined functional derivative $\delta E/\delta\rho$. The most generic form of gradient flow with respect to the target functional of minimization E is given by

$$\partial_t \rho_t = -G(\rho_t)^{-1} \frac{\delta E}{\delta \rho_t}$$

for the information metric $G(\cdot)$ subject to a choice. In this work, we specifically focus on the Wasserstein metric $G^W(\rho)^{-1}\Phi = -\nabla\cdot(\rho\nabla\Phi)$ that the variational discretization has been previously studied as JKO that naturally admits a velocity field form. Particularly, (overdamped) Langevin dynamics is a particle-wise implementation of the gradient flow over KL divergence.

Remark 1 (Fisher-Rao Metric Gradient Flow)**.** *The gradient flow is not only discussed in the context of Wasserstein metric, but also in various different metrics such as Fisher-Rao, Kalman-Wasserstein, Stein. These alternative geometries have previously been exploited in optimization (Yao et al., 2024), inverse problem (Garbuno-Inigo et al., 2020) and variational inference (Liu & Wang, 2016; Duncan et al., 2023). We leave the neural network parameterization of these alternative geometries to future work, and focus in this paper on the Wasserstein metric.*

Its alternative is the Hamiltonian Flow (Chow et al., 2019) which given metric $G(\cdot)$, the Hamiltonian flow of E in probability space is the canonical system of gradient flow

$$\begin{cases} \partial_t\rho_t - G(\rho_t)^{-1}\Phi_t = 0, \\ \partial_t\Phi_t + \frac{1}{2}\frac{\delta}{\delta\rho_t}\left(\int \Phi_t G(\rho_t)^{-1}\Phi_t \,\mathrm{d}x\right) + \frac{\delta E}{\delta\rho_t} = 0 \end{cases}$$

where $\frac{\delta H(\rho_t,\Phi_t)}{\delta\rho_t} = \frac{1}{2}\frac{\delta}{\delta\rho_t}\left(\int \Phi_t G(\rho_t)^{-1}\Phi_t \,\mathrm{d}x\right) + \frac{\delta e}{\delta\rho_t}$ is the functional derivative of the Hamiltonian.

Our approach can be summarized as generating the samples from the target density by the Accelerated Information Gradient Flow (AIGF) which is a generalization of Hamiltonian Gradient Flow by incorporating an auxiliary damping term $\alpha_t\Phi_t$ resulting in the AIGF

Proposition 1 (AIGF; (Wang & Li, 2022) Proposition 1)**.** *The accelerated information gradient flow satisfies*

$$\begin{cases} \partial_t\rho_t - G(\rho_t)^{-1}\Phi_t = 0, \\ \partial_t\Phi_t + \alpha_t\Phi_t + \frac{1}{2}\frac{\delta}{\delta\rho_t}\left(\int \Phi_t G(\rho_t)^{-1}\Phi_t \,dx\right) + \frac{\delta E}{\delta\rho_t} = 0, \end{cases} \tag{1}$$

with initial values $\rho_t|_{t=0} = \rho_0$ *and* $\Phi_t|_{t=0} = 0$.

Proposition 1 is a mean-field formulation of the AIGF which demonstrates the evolution of the marginal ρ_t over the flow. Its particle-wise simulation under Wasserstein metric can be written in system of equations with tuple of random variables (X_t, V_t) as

$$\begin{cases} \frac{d}{dt}X_t = V_t, \\ \frac{d}{dt}V_t = -\alpha_t V_t - \nabla\left(\frac{\delta E}{\delta \rho_t}\right)(X_t). \end{cases} \tag{2}$$

JKO-scheme and Flow-based Modeling A common approach to discretize the gradient flow is by leveraging its variational framework (Jordan et al., 1998) or the JKO scheme. This approach to model an optimal transport map has been widely adopted via Continuous Normalizing Flow (CNF) to model the invertible map (Onken et al., 2021; Vidal et al., 2023; Xu et al., 2023). Specifically, the block-wise training of the single JKO step with family of grid time $\{s_k\}_{k=1}^n$ will be achieved by minimizing a single step JKO objective

$$\rho_{k+1} = \arg\min_{\rho \in \mathcal{P}} \frac{1}{2} W_2^2(\rho_k, \rho) + s_k E[\rho],$$

which can be re-interpreted as the proximal step. Hence, a single JKO step transports the marginal from ρ_k to ρ_{k+1} such that ρ_{k+1} minimizes the sum of Wasserstein-2 distance from the reference density ρ and the functional E which a canonical choice will be the KL divergence with the target density π, which is $E(\rho) = \int V d\rho + \int \rho \log \rho dx = \mathrm{KL}(\rho \| \pi) - \log Z$. The famous Benamou-Brenier formulation guarantees an existence of a transport map and its correspondence with a velocity field $v^\star(\cdot)$ that minimizes the W_2^2 metric over the transport while satisfying the continuity equation (Brenier, 1991; Hertrich & Gruhlke, 2025). To model the velocity field, the CNF network is trained to minimize the objective function

$$\mathcal{L}(\theta) = \mathbb{E}_{x \sim \mu_\tau^k}[-\log(\pi(z_\theta(x,\tau))) - \ell_\theta(x,\tau) + \omega_\theta(x,\tau)], \tag{3}$$

subject to the system of ODEs

$$\dot{z}_\theta(x,t) = v_\theta(z_t,t), \quad \dot{\ell}_\theta(x,t) = \mathrm{tr}(\nabla v_\theta(z_t,t)), \quad \dot{\omega}_\theta(x,t) = \|v_\theta(z_t,t)\|^2$$

This system of ODEs is often solved by numerical solvers such as Runge-Kutta which composes one of the two pillars of bottleneck in training.

4 Discretization of Wasserstein-AIGF

4.1 Accelerated-JKO

We here first present the Accelerated-JKO (JKO) which becomes the core formulation of our work. The provided A-JKO is a Nesterov-style discretization of the AIGF that iteratively updates the prior via proximal-iteration based optimization scheme to allow the marginal at the terminal timestep iteratively converges to the target density.

Definition 1 (Accelerated JKO). *Let $(\tau_k)_{k\geq 0}, (\alpha_k)_{k\geq 0} \in \mathbb{R}^+$ be the step-size and friction schedules, and set $\beta_k := (1+\alpha_k\tau_k)^{-1}$, $s_k := \beta_k\tau_k^2$ and $\eta_k := \beta_k\tau_k/\tau_{k-1}$. Given ρ_k and a momentum field $D^{(k)} \in L^2(\rho_k; \mathbb{R}^d)$ with $D^{(0)} := \mathbf{0}$ (equivalently $\mu_0 = \rho_0$), the Accelerated-JKO (A-JKO) scheme discretizes the AIGF (1) by*

$$\begin{aligned} &\text{(Extrapolation)} && \mu_k = \left(\iota + \eta_k D^{(k)}\right)_\sharp \rho_k, \\ &\text{(Proximal)} && \rho_{k+1} = \arg\min_\rho \left\{\tfrac{1}{2} W_2^2(\mu_k, \rho) + s_k E(\rho)\right\} = (T_k)_\sharp \mu_k, \\ &\text{(Momentum)} && D^{(k+1)} = \iota - \left(T_k \circ (\iota + \eta_k D^{(k)})\right)^{-1}, \end{aligned} \tag{4}$$

where T_k denotes the Brenier map realizing (4), and ι is the identify map.

To derive this, we discretize the particle-form AIGF (2) with representation as a finite time difference equation and leveraging the first order stationary condition of the proximal operator. The step size control parameter $\beta_k \in (0,1]$, naturally forces the effective step $s_k \leq \tau_k^2$ at all time step k which

naive τ_k^2 an unaccelerated scheme would use at the same grid. The full derivation is provided at appendix A.

In contrast to the original JKO, the Accelerated JKO consists of three different steps: extrapolation, proximal and momentum advance steps. The extrapolation step 1 correspond to the particle advance via momentum $\dot{X}_t = V_t$, the proximal step 4 correspond to the dissipative half-step of the momentum $\dot{V}_t = -\alpha_t V_t - \nabla(\delta E/\delta \rho_t)$. The two steps are exact analogue of the particle-wise discretization of the AIGF with the variational formulation of the functional. Lastly, the momentum advance 1 recovers the updated momentum by comparing the extrapolated and proximal-updated particle state. The momentum advance step can be considered as a composition of the backward map with the extrapolation map rather than independent updates.

This construction of Accelerated-JKO scheme is also an exact analogue of the classical Nesterov-style momentum recursion at a constant $\tau_k = \tau, \alpha_k = \alpha$

$$y_k = x_k + \beta(x_k - x_{k-1}), \qquad x_{k+1} = y_k - \tau \nabla f(y_k),$$

where the first y_k update correspond to the extrapolation, the second x_{k+1} update correspond to the proximal step. The momentum is then recorded by $v_{k+1} = x_{k+1} - y_k$ which correspond to our momentum advance step. Furthermore, we see that the infinite friction $\alpha_k \to \infty$ makes the momentum to be zero $D^{(k)} = 0$ for all $k \in 0, 1, 2, \dots$ collapsing the Accelerated JKO to a standard JKO by leaving the proximal step only.

4.2 Implicitly Accelerated-JKO

The structural limitation underlying in the A-JKO scheme is the ill-posed implementation as CNF: unlike proximal update in the JKO scheme, the extrapolation map 1 does not guarantee the well-posedness in terms of the injectivity. The extrapolation step itself is agnostic to the injectivity of the velocity field, but transports the sample under the fixed time grid, thus the velocity field may *fold* under vanishing Jacobian determinant $\det \nabla(\iota + \eta_k D^{(k)})$. This makes the CNF that models the velocity field solution of the A-JKO non-invertible diffeomorphism which is an ill-posed target that may not exist (Chen et al., 2018; Dupont et al., 2019).

To prevent folding, we propose the Implicitly Accelerated JKO (IA-JKO) that optimizes the subsequent marginal density ρ_{k+1} by a single proximal update that incorporates momentum. To construct the single-step non-invertible diffeomorphism, we introduce the following action:

Definition 2 (Inertial transport action)**.** *Let $\log_\sigma : \mathcal{P}(\Omega) \to T_\sigma\mathcal{P}(\Omega)$ denote the Wasserstein logarithmic map at $\sigma \in \mathcal{P}(\Omega)$ (the inverse of the exponential map, cf. Wang & Li (2022, Appendix D.1-D.2), satisfying $\|\log_\sigma \rho\|_\sigma = W_2(\sigma, \rho)$. For $m \in T_\sigma\mathcal{P}(\Omega)$, the momentum-corrected transport action is given by*

$$\mathcal{A}(\sigma, \rho; m) := \|\log_\sigma \rho - m\|_\sigma^2 = W_2^2(\sigma, \rho) - 2\langle m, \log_\sigma \rho\rangle_\sigma + \|m\|_\sigma^2.$$

The inertial transport action quantifies the difference between the actual optimal-transport displacement between σ and ρ and prescribed transport by X. They are compared as elements of the same Hilbert subspace of $T_\sigma\mathcal{P}(\Omega) \subset L^2(\sigma; \mathbb{R}^d)$. In the definition, the quadratic term with respect to momentum is independent of solution space ρ, only depending on deterministic current marginal σ. Subsequently, the momentum-marginal interaction only enters through the inner product term of the final expression which is represented as an inner product $-2\langle m, \log_{\rho_k} \rho\rangle_{\rho_k}$.

The corresponding Implicit Accelerated-JKO scheme is defined for the natural velocity $v_k := D_k/\tau_{k-1}$ by the following modified Proximal operator:

Definition 3 (Implicit Accelerated JKO (IA-JKO))**.** *With β_k, s_k as in definition 1 and m_k satisfying $u_k = \beta_k \tau_k v_k = \frac{\tau_k}{\tau_{k-1}} \beta_k D_k$, the Implicit Accelerated JKO scheme is*

$$\rho_{k+1} = \arg\min_\rho \left\{ E(\rho) + \frac{1}{2s_k} \mathcal{A}(\rho_k, \rho; u_k) \right\}. \tag{5}$$

By folding the extrapolation directly into an action-based penalty rather than applying it as a separate push-forward step, IA-JKO never requires $\iota + u_k$ itself to be injective. No image measure, and hence no Jacobian determinant, ever needs to exist, so the folding that incurs CNF ill-posedness identified

above is relaxed entirely since there is no μ_k whose log-density the flow would have to model. Concretely, we realize the IA-JKO step definition 3 by parametrizing a velocity field $v_\theta(x_t, t) \approx v_k$ and integrating it as a continuous normalizing flow, so that the induced transport map approximates the IA-JKO iteration at equation (5) directly.

We now show that IA-JKO is well posed without the no-folding condition.

Proposition 2 (Well-posedness of IA-JKO; abbreviated)**.** *Let Assumptions 1 and 2 hold and let $\rho_k \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ with $\mathrm{Ent}(\rho_k) < \infty$. Then the IA-JKO step (Definition 3) admits a unique minimizer $\rho_{k+1} \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ with finite entropy. Moreover, the forward map $M_k := T_{\rho_k \to \rho_{k+1}}$ is a Brenier map, hence*

$$v_{k+1} := \tau_k^{-1}(\iota - M_k^{-1}) = -\tau_k^{-1} \log_{\rho_{k+1}} \rho_k \in \mathrm{Tan}_{\rho_{k+1}}$$

is a gradient field with $\|v_{k+1}\|_{\rho_{k+1}} = W_2(\rho_k, \rho_{k+1})/\tau_k$.

The well-posedness of A-JKO in the proximal step is relatively natural based on the fact that it has the same proximal structure as the vanilla JKO proximal operation. In contrary, the proximal step for IA-JKO replaces the W_2 distance into an inertial transport action $\mathcal{A}$ which admits the well-posedness in a non-trivial way. We provide the theoretical justification to the well-posedness of A-JKO and IA-JKO Proposition 6 and 7 in Appendix B.

Remark 2 (Comparison with Wang & Li (2022)'s discretization)**.** *Wang & Li (2022) simulates the discretization of the AIGF by particle-wise iteration instead of variational formulation. However, the particle base implementation does not guarantee well-posedness and may suffer from folding (our A-JKO well-posedness assumes folding while IA-JKO does not). The solution proposed is adaptive restart, which re-initializes the algorithm starting by setting the momentum as 0. However, this may undermine the benefit of the momentum augmentation which the alternative IA-JKO allows the simulation without the adaptive restart with well-posedness guarantee.*

Lastly, we show that the the A-JKO and IA-JKO iteration converges to the AIGF by the informal theorem:

Theorem 1 (Convergence to the AIGF; informal)**.** *As the discretization step vanishes, the A-JKO and IA-JKO iterates converge to the exact solution of the continuous-time AIGF both in measure transport and particle path sense. In particular, the particles asymptotically follow the accelerated dynamics $\ddot{X}_t = -\alpha_t \dot{X}_t - \nabla(\delta E/\delta \rho_t)(X_t)$ which is the Nesterov-style momentum in Wasserstein space.*

A formal version of Theorem 1, stated as Theorem 2, along with its complete proof, is provided in Appendix B.

5 AIGF Sampler

5.1 Training of Accelerated JKO

To train a velocity field that is parameterized by a neural network, we may consider to train a velocity field that transports the samples from ρ_k to ρ_{k+1} by single A-JKO step. Following literature on the CNF-based proximal mapping modeling, it is possible to consider to train a velocity field v_θ that solves the single proximal update for both A-JKO and IA-JKO. The key difference between A-JKO and IA-JKO formulation is that the loss should be formulated as the exact same format for JKO-IC loss at (3) to model the proximal mapping while the the IA-JKO requires a modification to minimize the inertial action concurrently (Hertrich & Gruhlke, 2025). In other words, the goal is to train a velocity field v_θ that satisfy an IA-JKO step

$$\underset{\substack{v:\, \partial_t v + (v\cdot\nabla)v = 0 \\ \dot{z}(x,t) = v(z(x,t),t),\ z(x,0) = x}}{\arg\min} \underbrace{\frac{1}{2s_k} \int_{\mathbb{R}^d} \|v(x,0) - u_k(x)\|^2 \, \mathrm{d}\rho_k(x)}_{\text{inertial action}} + \underbrace{\mathcal{F}(z(\cdot,1)_{\#}\rho_k)}_{\text{free energy}}.$$

This velocity field is learned by simulating the CNF trajectory of the particle and solving for the target velocity field by using an external ODE solver. Because the numerical error in the ODE solver contributes the bias in the training target, concurrently with increased number of function evaluation in local error control, the computation of the target becomes costly at scale. The reverse KL loss at the IA-JKO training via CNF is naturally dependent on the path that the neural network simulates which

enforces the training on-policy. Recent studies argue the effectiveness of off-policy training with a replay buffer for improved training efficiency of neural sampler (Liu et al., 2025; Holderrieth et al., 2025; Berner et al., 2026). To leverage the effectiveness, we consider off-policy training objective that allows efficient training with buffer and many solution per step.

Matching Loss We first derive the optimality condition that an ideal velocity field must achieve. Motivated by the JKO optimality lemma (Santambrogio, 2015), we first derive the optimality condition of the IA-JKO step as

Proposition 3 (first-order optimality for the IA-JKO step; abbreviated)**.** *Under Sufficient regularity, let ρ_{k+1} be the minimizer of the IA-JKO optimality Equation 5. Then, for every $h \in C_c^\infty(\mathbb{R}^d)$ with $\xi_h := \nabla h \circ M_k^{-1}$, we have*

$$\int \langle \nabla V, \xi_h \rangle \, \mathrm{d}\rho_{k+1} - \int \nabla \cdot \xi_h \, \mathrm{d}\rho_{k+1} + \frac{1}{s_k} \langle M_k - \iota - u_k, \, \nabla h \rangle_{\rho_k} = 0.$$

Furthermore, there is a function Ψ_k with $\nabla \Psi_k \in \mathrm{Tan}_{\rho_{k+1}}$, determined by u_k and M_k alone and a constant C such that

$$\log \rho_{k+1} + V + \frac{\varphi^{\mathrm{IA}} - \Psi_k}{s_k} = C \quad \textit{a.e.},$$

$$v_{k+1} := \frac{\iota - M_k^{-1}}{\tau_k} = \frac{\nabla \varphi^{\mathrm{IA}}}{\tau_k} = \frac{\nabla \Psi_k}{\tau_k} - \beta_k \tau_k \nabla \big(\log \rho_{k+1} + V \big) \quad \textit{a.e.}$$

where $W_{\rho_{k+1}} = \nabla \left(\log \rho_{k+1} + V \right)$ denote the estimated field, Ψ_k is a scalar potential whose gradient $\nabla \Psi_k$ is defined via an orthogonal projection of $u_k \circ M_k^{-1}$ onto $\mathrm{Tan}_{\rho_{k+1}}$. From the optimality condition, we obtain the regression target as

$$M_k - \iota = \arg \min_{v \in \mathrm{Tan}_{\rho_k}} \mathbb{E}_{x \sim \rho_k} \left[|v_\theta(x) - \xi_k(x, M_k(x))|^2 \right]$$

with the target velocity field $\xi_k(x, M_k(x)) = u_k(x) - s_k \nabla W_{\rho_{k+1}}(y)$. The derivation of the regression target is deferred to appendix C.

The score term $\nabla \log \rho_{k+1}$ inside the next block $\nabla W_{\rho_{k+1}}$ is concurrently predicted via neural network (trained via DSM) with gradient step blocked by stop-grad operator when matching the target. We further apply polish to the small number of terminal blocks such that predicted score can be corrected by predictor-corrector scheme. The corrector uses the computation of the $\log \rho_{k+1}$ by exact log determinant of the CNF at a small additional cost (at most 30% in the effective polishing block). This loss is off-policy with respect to velocity field v_θ that is the training target. The score is predicted by fixed score estimator network which the labels are then stored in a replay buffer which allows multiple gradient steps per label. IA-JKO condition is a fixed point in ρ_{k+1}, which allows the generator to be refreshed from the current solution. Finally, we obtain the matching loss function as

$$\mathcal{L}_{\text{matching}} = \mathbb{E}_{x \sim \rho_k} \left[\left\| v_\theta(x) - \xi_k(x, M_k(s)) \right\|^2 \right]. \tag{6}$$

We summarize the full algorithm in the appendix.

Remark 3. *We categorize our sampler as a **JKO-iterative sampler** which is an alternative class to the **fixed-edge sampler**. The key differentiation is that the fixed edge sampler consider a single drift v_θ that iterates a single differential equation to transform the prior to target given by*

$$\dot{x}_t = v_\theta(x_t, t), \qquad x_0 \sim \rho_{\textit{prior}}, \quad x_1 \sim \rho_{\textit{target}}.$$

In contrast, the JKO-iterative sampler leverages a small drift blocks that simulate the single A-JKO step which is stacked up to a specific depth. In other words, our goal is to learn a sequence of velocity fields $v_{\theta_1}, v_{\theta_2}, \ldots, v_{\theta_K}$ such that each v_{θ_k} trained to realize the accelerated JKO recursion by

$$x_1 = x_0 + \int_0^{\tau_k} v_{\theta_k}(x_t, t) \, \mathrm{d}t, \qquad x_0 \sim \rho_k, \quad \rho_{k+1} := \mathrm{Law}(x_1) = (\Phi_{\theta_k})_{\#} \, \rho_k$$

while fixed-edge sampler learns a single velocity/drift field simulates the path transport where $x \sim \rho_{\textit{prior}}$ from an easy-to-sample distribution such as standard Gaussian or Dirac-delta.

Table 1: The performance of A-JKO and IA-JKO samplers compared to the JKO-IC and other fixed-edge optimization samplers on synthetic energy. JKO-IC represents the Importance Corrected JKO sampler. **Bold** : best overall; underline : best flow-iterative method.

	Method	DW-4		LJ-13		LJ-55	
		$E(\cdot)W_2$	W_2	$E(\cdot)W_2$	W_2	$E(\cdot)W_2$	W_2
MCMC (GT)†		$0.10_{\pm 0.03}$	$0.31_{\pm 0.01}$	$0.35_{\pm 0.03}$	$1.57_{\pm 0.01}$	$0.57_{\pm 0.05}$	$3.85_{\pm 0.05}$
JKO-Iterative	JKO-IC	$\underline{0.41}_{\pm 0.05}$	$1.46_{\pm 0.01}$	$60.56_{\pm 30.44}$	$3.96_{\pm 1.28}$	$4328.5_{\pm 912.47}$	$78.0_{\pm 12.38}$
	A-JKO	$0.78_{\pm 0.12}$	$0.80_{\pm 0.10}$	$8.18_{\pm 0.52}$	$1.71_{\pm 0.05}$	$113.1_{\pm 10.7}$	$4.64_{\pm 0.07}$
	IA-JKO	$0.53_{\pm 0.09}$	$\underline{0.70}_{\pm 0.06}$	$\underline{3.74}_{\pm 1.77}$	$\underline{1.66}_{\pm 0.06}$	$\mathbf{19.54}_{\pm 3.18}$	$\underline{4.58}_{\pm 0.13}$
Fixed-Edge	PIS	$0.65_{\pm 0.25}$	$0.68_{\pm 0.23}$	$18.02_{\pm 1.12}$	$1.93_{\pm 0.07}$	$228.70_{\pm 131.27}$	$4.79_{\pm 0.45}$
	DDS	$0.90_{\pm 0.37}$	$0.92_{\pm 0.11}$	$24.61_{\pm 8.99}$	$1.99_{\pm 0.13}$	$173.09_{\pm 18.01}$	$4.60_{\pm 0.09}$
	iDEM	$0.55_{\pm 0.14}$	$0.70_{\pm 0.06}$	$30.78_{\pm 24.46}$	$1.61_{\pm 0.01}$	$93.53_{\pm 16.31}$	$4.69_{\pm 1.52}$
	AS	$0.55_{\pm 0.12}$	$0.62_{\pm 0.06}$	$2.40_{\pm 1.25}$	$1.67_{\pm 0.01}$	$30.83_{\pm 8.19}$	$4.04_{\pm 0.05}$
	ASBS	$0.20_{\pm 0.11}$	$0.43_{\pm 0.05}$	$1.99_{\pm 1.01}$	$1.59_{\pm 0.03}$	$28.10_{\pm 8.15}$	$4.00_{\pm 0.03}$
	FS	$\mathbf{0.11}_{\pm 0.04}$	$\mathbf{0.36}_{\pm 0.03}$	$\mathbf{0.97}_{\pm 0.53}$	$\mathbf{1.58}_{\pm 0.01}$	$21.32_{\pm 0.63}$	$\mathbf{3.98}_{\pm 0.01}$

6 Experiments

We test the effectiveness of our A-JKO and IA-JKO on various sampling problem of synthetic energy and sampling conformers of peptides from Alanine-Dipeptide-2. For A-JKO, we use the simulation loss with folding prevented by controlling the step size instead of the IA-JKO. For IA-JKO, we use the matching loss at Equation (6) with the predictor-corrector based score estimation network. We further provide the details of the experiments at Appendix E.

6.1 Synthetic Density

We show the performance of the A-JKO and IA-JKO sampler on the synthetic energy. We categorize the JKO-IC, A-JKO and IA-JKO as flow-iterative samplers which generates sample via iterative transformation by small networks.

We compare the result of the samplers with different class of sampling methods: JKO-iterative sampler and fixed-edge sampler. The result is presented at table 1. We first compare the performance on small distribution of DW-4 distribution. Given relatively small dimensionality ($d = 8$), the JKO-IC provides a sufficiently good proposal and as a result the energy-W2 is sufficiently low to 0.41. In contrast, it can be seen that the geometry is relatively poorly matched based on the W2 (geometric-W2 distance) which demonstrates the accelerated variant outperforms the JKO-IC. In the low dimensional case, the advantage of acceleration is not clearly visible as a first-order convergence may demonstrate sufficient convergence with the rejection step and first-order proposal network.

However, the high-dimensional densities such as Lennard-Jones, the dimension is higher ($d = 39$, $d = 165$) which makes the tracking more complex. Conventional JKO-IC method fails to track the target distribution and results in the high W2 distance of 3.96 and energy W2 distance of 60.56 on LJ-13. In contrast, the accelerated JKO resolves this failure and demonstrates a better convergence such that the W2 improves to 1.71 and 1.66 for A- and IA-JKO respectively. This geometric W2 distance outperforms not only the JKO-IC in the JKO-iterative sampler, but also the PIS, DDS and AS ($1.93, 1.99$ and 1.67). Particularly, we find the energy-W2 distance has decreased to 13.6% and 6.2% (8.18 and 3.74) of the JKO-IC's value. Yet, a gap remains on energy-W2, where FS (0.97) and ASBS (1.99) still outperform both flow-iterative variants.

We further extend the experiment to higher dimensional sampling problem of LJ-55 density ($d = 165$). At this sclae, the IA-JKO sampler has achieved the energy-W2 of 19.54 which is better than any other sampler class. The performance in W2 falls behind the state of the art AS, ASBS and FS ($4.04, 4.00, 3.98$ respectively), while outperforming other fixed-edge samplers such as PIS, DDS and iDEM ($4.79, 4.60, 4.69$). Therefore, we broaden the study of JKO-iterative sampler by demonstrating its potential competitiveness against the fixed-edge samplers.

Method	NFE (Train)	D_{KL} 1-D (↓)		(ϕ, ψ) joint	Energy
		ϕ	ψ	JSD (↓)	$\mathcal{W}_2$ (↓)
IA-JKO	N/A	$\underline{0.188}_{\pm 0.051}$	$\underline{0.318}_{\pm 0.037}$	$0.086_{\pm 0.004}$	$4.462_{\pm 0.611}$
IA-JKO (w/ rej.)	N/A	$0.405_{\pm 0.103}$	$0.321_{\pm 0.075}$	$\underline{0.054}_{\pm 0.004}$	$\mathbf{0.492}_{\pm 0.129}$
ASBS	1024	$0.504_{\pm 0.006}$	$0.726_{\pm 0.300}$	$0.242_{\pm 0.042}$	$8.650_{\pm 0.371}$
	128	$0.663_{\pm 0.460}$	$0.550_{\pm 0.267}$	$0.285_{\pm 0.158}$	$1.0\mathrm{e}7_{\pm 1.0\mathrm{e}7}$
Flow Sampling	1024	$\mathbf{0.031}_{\pm 0.004}$	$\mathbf{0.008}_{\pm 0.002}$	$\mathbf{0.018}_{\pm 0.001}$	$\underline{0.637}_{\pm 0.267}$
	128	$0.267_{\pm 0.023}$	$0.344_{\pm 0.057}$	$0.152_{\pm 0.021}$	$3.579_{\pm 0.311}$

Table 2: Performance on Alanine Dipeptide dihedral angle distributions and energy metrics. Errors denote standard deviations across 3 runs. **Bold** : best overall; underline : second best.

Remark 4 (Benefit of momentum at high-dimensional setting). *Table 1 shows that the accelerated samplers become more competitive as the dimension grows. JKO-IC attains the best energy-W2 on DW-4 ($d = 8$), both variants surpass it on LJ-13 ($d = 39$), and IA-JKO attains the best energy-W2 of all samplers on LJ-55 ($d = 165$). We attribute this to momentum, which carries displacement across blocks so that no single velocity field must realize a large transport alone. On geometric W2, both variants outperform JKO-IC but are comparable to or behind the strongest fixed-edge samplers, likely because the finite-depth chain stops at a finite horizon and accumulates per-block errors.*

6.2 Alanine Dipeptide

We subsequently conduct the same experiment on the task of sampling the equilibrium Boltzmann distribution of alanine dipeptide (Ala2). Following the standard evaluation protocol, we assess each method along three axes: the 1-D KL divergence between model and ground-truth marginals for ϕ and ψ individually, the Jensen–Shannon divergence between the joint (ϕ, ψ) Ramachandran distributions, and the 2-Wasserstein distance between model and reference energy histograms.

Table 2 shows that Flow Sampling degrades as NFE is reduced, while ASBS becomes numerically unstable below NFE $N = 128$ with energy-W2 exploding significantly to the scale of 10^7. IA-JKO sits between these regimes on the 1-D marginals but is substantially more competitive on the joint and energy metrics. The IA-JKO outperforms ASBS at all metrics upto $N = 1024$ NFE both rejection and without rejection as a post-hoc where the KL divergence in ϕ and ψ are reduced by more than half on both cases. Particularly, the JSD is reduced to approximately 35% of the ASBS case with out rejection and energy-W2 is reduced approximately by half. With rejection, it further improves on joint metric and achieves the best energy-W2 (0.492) of all methods considered, including the highest-budget Flow Sampling run (NFE $N = 1024$). The gap between IA-JKO and IA-JKO (w/ rej.) further indicates that a modest correction step recovers most of the residual bias left by the finite-depth JKO approximation, at negligible additional cost relative to Flow Sampling's high-NFE regime.

7 Conclusion

In this work, we propose the Accelerated-JKO (A-JKO)—a momentum-augmented JKO scheme that simulates the Wasserstein-AIGF at measure optimization perspective. We further propose its implicit variant IA-JKO which absorbs the extrapolation term into an inertial trasnport action, hence prevents the folding of the velocity field. Both schemes consist of well-posed variational problems on the probability space $\mathcal{P}_2$ with energy-dissipation inequality from the Hamiltonian perspective of the AIGF. We further establish the convergence proof of the A-JKO and IA-JKO to the AIGF as the step size vanishes.

Motivated by the learning-based implementation of the JKO, we construct a neural sampler that simulates the A-JKO and IA-JKO by modeling a flow neural-network. We design an off-line training objective that enables efficient training from the single-step optimality condition of the IA-JKO which empirically outperforms other JKO-based neural samplers. Furthermore, the resulting sampler is competitive with recent neural samplers on DW-4, LJ-13/55, and alanine dipeptide.

AI USE STATEMENT

In this work, AI are used for the following purposes:

1. Generation of proof sketch for Theorems, Propositions, Lemmas in Section 4, 5 and corresponding appendices.
2. Code generation for the experiments on A-JKO and IA-JKO.
3. Anonymizing the code for submission.
4. Visualization generation of the A-JKO scheme at Figure 1.
5. Table formatting of Table 1 and 2.
6. First draft for the section: Appendix D, E, F.

ETHICS STATEMENT

This work is theoretical in nature: it establishes convergence guarantees for a damped Jacobi–Kantorovich–Otto (JKO) discretization scheme toward a Wasserstein accelerated information gradient flow, together with the associated algorithmic and numerical analysis. It does not introduce new datasets, does not involve human subjects, and the numerical experiments use only standard synthetic and benchmark sampling targets. As a general-purpose accelerated sampling method, the techniques developed here could in principle be applied downstream to generative models trained on sensitive data, but the paper itself makes no such application and we are not aware of any direct ethical concerns, societal risks, or foreseeable negative applications specific to this work. We have read the ICLR Code of Ethics and believe this work conforms to it.

REPRODUCIBILITY STATEMENT

We provide the anonymized github repository for reproducibility and leave the experimental details at Appendix E. https://anonymous.4open.science/r/Accelerated-JKO

REFERENCES

Tara Akhound-Sadegh, Jarrid Rector-Brooks, Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Alexander Tong. Iterated denoising energy matching for sampling from boltzmann densities. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=gVjMwLDFoQ.

David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. Optimizing functionals on the space of probabilities with input convex neural networks. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=dpOYN7o8Jm.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2 edition, 2008. ISBN 978-3-7643-8722-8. doi: 10.1007/978-3-7643-8722-8.

Julius Berner, Lorenz Richter, and Karen Ullrich. An optimal control perspective on diffusion-based generative modeling. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=oYIjw37pTP.

Julius Berner, Lorenz Richter, Marcin Sendera, Jarrid Rector-Brooks, and Nikolay Malkin. From discrete-time policies to continuous-time diffusion samplers: Asymptotic equivalences and faster training. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=xLE3xJUuDO. Expert Certification.

Espen Bernton. Langevin Monte Carlo and JKO splitting. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1777–1798. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/bernton18a.html.

Denis Blessing, Julius Berner, Lorenz Richter, and Gerhard Neumann. Underdamped diffusion bridges with applications to sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Q1QTxFm0Is.

Denis Blessing, Lorenz Richter, Julius Berner, Egor Malitskiy, and Gerhard Neumann. Bridge matching sampler: Scalable sampling via generalized fixed-point diffusion matching. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=gUbNUNuJbO.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44:375–417, 1991. doi: 10.1002/cpa.3160440402.

Charlotte Bunne, Laetitia Meng-Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal Optimal Transport Modeling of Population Dynamics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 25, 2022.

Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2): 1385–1418, 2017. doi: 10.1137/15M1050264.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Shui-Nee Chow, Wuchen Li, and Haomin Zhou. Wasserstein hamiltonian flows, 2019. URL https://arxiv.org/abs/1903.01088.

Andrew Duncan, Nikolas NÃ¼sken, and Lukasz Szpruch. On the geometry of stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023. URL http://jmlr.org/papers/v24/20-602.html.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/21be9a4bd4f81549a9d1d241981cec3c-Paper.pdf.

Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659, 2017. doi: 10.1371/journal.pcbi.1005659.

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational wasserstein gradient flow, 2022. URL https://arxiv.org/abs/2112.02424.

RÃ©mi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, AurÃ©lie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, LÃ©o Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, and Andrew M. Stuart. Interacting langevin diffusions: Gradient structure and ensemble kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020. doi: 10.1137/19M1251655.

Aaron J Havens, Brian Karrer, and Neta Shaul. Flow sampling : Learning to sample from unnormalized densities via denoising conditional processes. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=YlcyOMTPNl.

Johannes Hertrich and Robert Gruhlke. Importance corrected neural JKO sampling. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=yQBZZeWPdQ.

Peter Holderrieth, Michael Samuel Albergo, and Tommi Jaakkola. LEAPS: A discrete neural sampler via locally equivariant networks. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Hq2RniQAET.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359.

Jonas Köhler, Leon Klein, and Frank Noe. Equivariant flows: Exact likelihood generative learning for symmetric densities. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5361–5370. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/kohler20a.html.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=K2PTuvVTF1L.

Guan-Horng Liu, Jaemoo Choi, Yongxin Chen, Benjamin Kurt Miller, and Ricky T. Q. Chen. Adjoint schrödinger bridge sampler. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=rMhQBlhh4c.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2378–2386. Curran Associates, Inc., 2016.

Yi-An Ma, Niladri S Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Is there an analog of nesterov acceleration for gradient-based mcmc? *Bernoulli*, 27(3), 2021.

Laurence Illing Midgley, Vincent Stimper, Gregor N. C. Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XCTVFJwS9LJ.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=nlLjIuHsMHp.

Juno Nam, Bálint Máté, Artur P. Toshev, Manasa Kaniselvan, Rafael Gomez-Bombarelli, Ricky T. Q. Chen, Brandon M. Wood, Guan-Horng Liu, and Benjamin Kurt Miller. Enhancing diffusion-based sampling with molecular collective variables. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1bJN1EQByS.

Dai Hai Nguyen, Tetsuya Sakurai, and Hiroshi Mamitsuka. Wasserstein gradient flow over variational parameter space for variational inference. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1756–1764. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/nguyen25d.html.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9223–9232, 2021.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11 (5-6):355–602, 2019. URL https://arxiv.org/abs/1803.00567.

Lorenz Richter and Julius Berner. Improved sampling via learned diffusions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=h4pNROsO06.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill Science, Engineering & Mathematics, 1974.

Adil Salim, Anna Korba, and Giulia Luise. The wasserstein proximal gradient algorithm. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12356–12366. Curran Associates, Inc., 2020.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Cham, 1 edition, 2015. ISBN 978-3-319-20827-5. doi: 10.1007/978-3-319-20828-2.

Weijie Su, Stephen Boyd, and Emmanuel 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. URL http://jmlr.org/papers/v17/15-084.html.

Amirhossein Taghvaei and Prashant Mehta. Accelerated flow for probability distributions. In *International conference on machine learning*, pp. 6076–6085. PMLR, 2019.

Alexander Vidal, Samy Wu Fung, Luis Tenorio, Stanley Osher, and Levon Nurbekyan. Taming hyperparameter tuning in continuous normalizing flows using the jko scheme. *Scientific reports*, 13(1):4501, 2023.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Yifei Wang and Wuchen Li. Accelerated information gradient flow. *J. Sci. Comput.*, 90(1), January 2022. ISSN 0885-7474. doi: 10.1007/s10915-021-01709-3. URL https://doi.org/10.1007/s10915-021-01709-3.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2093–3027. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/wibisono18a.html.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by JKO scheme. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ZQMlfNijY5.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Computing high-dimensional optimal transport by flow neural networks. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2872–2880. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/xu25f.html.

Rentian Yao, Linjun Huang, and Yun Yang. Minimizing convex functionals over space of probability measures via KL divergence gradient flow. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2530–2538. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/yao24a.html.

Qinsheng Zhang and Yongxin Chen. Path integral sampler: A stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_uCb2ynRu7Y.

A Derivation of Accelerated JKO Scheme

In this section, we present how the Accelerated-JKO scheme has been derived based on the AIGF and defer the convergence result for later.

A.1 Preliminaries on Accelerated Information Gradient Flow

The Accelerated Information Gradient Flow (AIGF) in Wang & Li (2022) is constructed by adding a damping term $\alpha\Phi$ to the Hamiltonian flow such that it solves

$$\begin{cases} \partial_t \rho_t - G(\rho_t)^{-1}\Phi_t = 0, \\ \partial_t \Phi_t + \alpha_t \Phi_t + \frac{1}{2}\frac{\delta}{\delta \rho_t}\left(\int \Phi_t G(\rho_t)^{-1}\Phi_t \, dx\right) + \frac{\delta E}{\delta \rho_t} = 0, \end{cases} \tag{7}$$

with initial values $\rho_t|_{t=0} = \rho_0$ and $\Phi_t|_{t=0} = 0$. $G(\cdot)$ denotes the respective information metric such as Wasserstein metric

$$G^W(\rho)^{-1}\Phi = -\nabla \cdot (\rho \nabla \Phi), \qquad \Phi \in T^*_\rho \mathcal{P}(\Omega)$$

or Fisher Information metric

$$G^F(\rho)^{-1}\Phi = \rho\left(\Phi - \mathbb{E}_\rho[\Phi]\right), \qquad \Phi \in T^*_\rho \mathcal{P}(\Omega).$$

We specifically consider Wasserstein-AIGF, which its particle-form expression is given as

$$\begin{cases} \frac{d}{dt} X_t = V_t, \\ \frac{d}{dt} V_t = -\alpha_t V_t - \nabla\left(\frac{\delta E}{\delta \rho_t}\right)(X_t). \end{cases} \tag{8}$$

And the measure-level Euler-Lagrange form is Appendix A.3 of Wang & Li (2022):

$$\frac{D^2}{Dt^2}\rho_t + \alpha_t \partial_t \rho_t + G^W(\rho_t)^{-1}\frac{\delta E}{\delta \rho_t} = 0.$$

By substituting the first equation to the second equation at Equation (8), we obtain

$$\ddot{X}_t = \dot{V}_t = -\alpha_t V_t - \nabla \frac{\delta E}{\delta \rho_t}(X_t) = -\alpha_t \dot{X}_t - \nabla \frac{\delta E}{\delta \rho_t}(X_t),$$

which is the Nesterov ODE in mean-field perspective. Lastly, its respective velocity form with ρ_0 given and zero initial velocity field is given by

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \qquad \partial_t v_t + (v_t \cdot \nabla) v_t = -\alpha_t v_t - \nabla W_{\rho_t}. \tag{9}$$

A.2 Accelerated JKO

We first discretize Equation (8) on a time grid $(t_k, t_{k+1}]$ and let $\tau = t_{k+1} - t_k$. We shall define the natural velocity as $v_{k+1} = (X_{k+1} - X_k)/\tau$ and discrete acceleration $(v_{k+1} - v_k)/\tau$. We then can consider the evaluation of the damping on the new velocity, force at the new point

$$\frac{v_{k+1} - v_k}{\tau} + \alpha_k v_{k+1} + \nabla \frac{\delta E}{\delta \rho_{k+1}}(X_{k+1}) = 0.$$

Multiply by τ and collect v_{k+1}:

$$(1 + \alpha_k \tau) v_{k+1} = v_k - \tau \nabla \frac{\delta E}{\delta \rho_{k+1}}(X_{k+1}).$$

Define the *extrapolation weight* $\beta_k := \frac{1}{1+\alpha_k \tau}$. Then

$$v_{k+1} = \beta_k v_k - \beta_k \tau \nabla \frac{\delta E}{\delta \rho_{k+1}}(X_{k+1}).$$

Substitute into $X_{k+1} = X_k + \tau v_{k+1}$:

$$X_{k+1} = X_k + \underbrace{\tau\beta_k v_k}_{=:\,\tau m_k} - \underbrace{\tau^2\beta_k}_{=:\,s_k} \nabla \frac{\delta E}{\delta \rho_{k+1}}(X_{k+1}). \tag{10}$$

Equation (10) is first-order stationarity condition of the proximal operator

$$X_{k+1} = \arg\min_x \left\{ \frac{\delta E}{\delta \rho_{k+1}}(x) + \frac{1}{2s_k} |x - (X_k + \tau m_k)|^2 \right\}. \tag{11}$$

We already know that $|x - (X_k + \tau m_k)|^2 = |(x - X_k) - \tau m_k|^2$ indicating the proximal center is displaced from X_k by the inertial tangent vector τm_k. Equation (11) is the pointwise (Lagrangian) form of the JKO optimality condition at a single transported particle $X_{k+1} = X_k + \tau m_k$. Aggregation of this over all particles by integration with respect to μ_k gives the JKO minimizer's transport is the Brenier map, where the quadratic term integrates to $\frac{1}{2s_k} W_2^2(\mu_k, \rho_{k+1})$ and the first-variation term to the respective functional contribution. Therefore, Equation (11) is consistent with Equation (12)'s optimality condition.

Extrapolation We first exploit the displaced center as an actual measure $\rho_k^{\text{ext}} := \exp_{\rho_k}(\tau m_k) = (\iota + \tau m_k)_\sharp \rho_k$ followed by penalizing $\frac{1}{2s_k} W_2^2(\rho, \rho_k^{\text{ext}})$. Since the extended density ρ_k^{ext} is an extrapolation of the current marginal via momentum and the push-forward map $S = (\iota + \tau m_k)_\sharp$ requires injectivity, equivalently $\det(\iota + \tau \nabla m_k) > 0$. For the displacement $D_k = X_{k+1} - X_k$, we have the relationship $v_k = (X_k - X_{k-1})/\tau_{k-1} = D_k/\tau_{k-1}$ and $\tau_k m_k = \tau_k \beta_k v_k = \beta_k \frac{\tau_k}{\tau_{k-1}} D_k$. Defining a parameter $\eta_k = \beta_k \frac{\tau_k}{\tau_{k-1}}$, it reformulates the determinant to become $\det(\iota + \eta_k \nabla D_k)$.

Therefore, we shall derive the simple Accelerated-JKO scheme by two-step method composed of 1) extrapolation which evolves the marginal based on the momentum and 2) proximal step which penalizes the marginal to retrieve the AIGF. Therefore, the formal Accelerated-JKO scheme can be given by:

Proposition 4 (Accelerated JKO). *Let $(\tau_k)_{k\geq 0} \in \mathbb{R}^+$ be a step-size schedule (geometric in practice) and $(\alpha_k)_{k\geq 0} \in \mathbb{R}^+$ the friction. Define the schedule $\beta_k := \frac{1}{1+\alpha_k\tau_k}$ and $s_k := \frac{\tau_k^2}{1+\alpha_k\tau_k} = \beta_k \tau_k^2$.*

For $k \geq 0$, given ρ_k and the momentum field $D_{(\cdot)} \in L^2(\rho_k; \mathbb{R}^d)$, let $D_0 := \mathbf{0}$ (equivalently $\mu_0 = \rho_0$) and $\eta_k = \beta_k \frac{\tau_k}{\tau_{k-1}}$. The Accelerated-JKO (A-JKO) scheme is the following discretization of the AIGF Equation at (7):

Step 1 (Extrapolation).

$$\mu_k = \big(\iota + \eta_k D_k\big)_\sharp \rho_k.$$

Step 2 (Proximal / JKO update).

$$\rho_{k+1} = \arg\min_\rho \left\{ \frac{1}{2} W_2^2\big(\mu_k, \rho\big) + s_k\, E(\rho) \right\}, \tag{12}$$

realized by its associated Brenier map T_k, i.e. $\rho_{k+1} = (T_k)_{\#}\, \mu_k$.

Step 3 (Momentum advance). *Tracking the coupled particle through update*

$$D_{k+1} = \iota - \big(T_k \circ (\iota + \eta_k D_k)\big)^{-1}.$$

Implicit Lift One problem of the extrapolation-based vanilla Accelerated-JKO is that there is possibility of *folding* which represents when $\det(\nabla S)$ crosses zero and makes ρ_k^{ext} naturally ill-defined.

Alternatively, we use the exact tangent-space expansion of the penalty via the log map. On W_2-geodesic, the flat vector $x - X_k$ becomes the log map (inverse of an exponential map) $\log_{\rho_k} \rho := \exp_{\rho_k}^{-1} \rho \in T_{\rho_k}\mathcal{P}$ and τm_k is a tangent vector at ρ_k. We then can keep the penalty as a norm of tangent vectors via the log map:

$$\begin{aligned} \mathcal{A}_{m_k}(\rho_k, \rho) :&= \big\|\log_{\rho_k} \rho - \tau m_k\big\|_{\rho_k}^2 \\ &= \underbrace{W_2^2(\rho_k, \rho)}_{\|\log_{\rho_k} \rho\|^2} - 2\tau \langle m_k, \log_{\rho_k} \rho \rangle_{\rho_k} + \tau^2 \|m_k\|_{\rho_k}^2, \end{aligned}$$

Here, the momentum enters only through the second term, $-2\tau\langle m_k, \log_{\rho_k}\rho\rangle_{\rho_k}$, a functional of m_k as a fixed tangent vector and of the free variable ρ through its log map. The quadratic term is constant independent of ρ which allows the interpret the influence on the choice of measure ρ only using the bilinear momentum term. The singular pushforward $\exp_{\rho_k}(\tau m_k)$ never appears, so there is no determinant to vanish. The resultant Implicit Accelerated-JKO scheme is given by

$$\rho_{k+1} = \arg\min_{\rho}\left\{E(\rho) + \frac{1}{2s_k}\left\|\log_{\rho_k}\rho - \tau m_k\right\|^2_{\rho_k}\right\}, \quad s_k = \frac{\tau^2}{1+\alpha_k\tau}, \quad \tau m_k = \beta_k D.$$

The marginal transport $\rho_k \to \rho_{k+1}$ is achieved by a flow-map $\dot{X}_t = v_t(X_t)$ which prevents folding. The resultant Implicit Accelerated-JKO is given by:

Proposition 5 (Implicit Accelerated JKO (IA-JKO)). *With β_k, s_k as in proposition 4 and m_k satisfying $\tau_k m_k = \beta_k D_k$, the Implicit Accelerated JKO scheme is*

$$\rho_{k+1} = \arg\min_{\rho}\left\{E(\rho) + \frac{1}{2s_k}\mathcal{A}_{m_k}(\rho_k, \rho)\right\} \tag{13}$$

which is a single-step transport that is *folding*-free.

B Proofs

In this section, we demonstrate the theoretical analysis of the Accelerated JKO (A-JKO) and Implicitly-Accelerated JKO (IA-JKO) and its convergence to the targetted Accelerated Information Gradient Flow. The notations for forward optimal map is given by $M_k = T_{\rho_k \to \rho_{k+1}}$ while the backward optimal map is $S_k = T_{\rho_{k+1}\to\mu_k}$.

B.1 Assumptions

Assume the following for the proof.

Assumption 1 (Target Regularity). *The target density $\exp(-V)$ satisfies the following regularity: $V \in C^1(\mathbb{R}^d)$; ∇V is L_V-Lipschitz; V is λ-convex for some constant $\lambda \in \mathbb{R}$; $\int \exp(-V) < \infty$. We denote $\lambda^- = \min(\lambda, 0)$.*

Assumption 2 (Schedules). *Let $\alpha_k = \alpha(t_k)$ for a continuous function $\alpha : [0,T] \to [0, \alpha_{max}]$. The steps taken are small enough such that $1 + \lambda^- s_k \geq \frac{1}{2}$ and $|\lambda^-|\tau_k^2 \leq \frac{1}{2}$.*

Assumption 3 (Initialization). *We have well-defined probability density at initialization (i.e. $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ absolutely continuous with finite second moment and entropy. Additionally, $\rho_0 > 0$ a.e. with locally integrable $\log\rho_0$ (i.e. $\int_{B_R(0)} |\log\rho_0(x)|\,\mathrm{d}x < \infty$) satisfied by any Gaussian initialization (positivity).*

Assumption 4 (No Folding). *v_k has a Lipschitz version with $\beta_k\tau_k \operatorname{Lip}(v_k) \leq \frac{1}{2}$. Then $\iota + u_k$ is injective with Jacobian determinant in $[2^{-d}, (3/2)^d]$ a.e., so $\mu_k \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ with $\operatorname{Ent}(\mu_k) \leq \operatorname{Ent}(\rho_k) + d\log 2$.*

Assumption 5 (Velocity Regularity). *There are constant $L, C \in \mathbb{R}$ such that*

$$\sup_k \operatorname{Lip}(v_k) \leq L, \qquad |\nabla W_{\rho_{k+1}}(y)| \leq C(1+|y|) \quad \rho_{k+1} - \text{a.e., all } k,$$

and $\bar{\tau}L \leq \frac{1}{2}$.

Assumption 6 (AIGF Solution). *There exists a solution to the AIGF, equation 7, $(\rho_t^*, v_t^*)_{t\leq T}$ with $\rho_t^* \in \mathcal{P}_2(\mathbb{R}^d)$, $\bar{v} \in C^1$ where the flow maps $\dot{\bar{X}}_t = v_t(\bar{X}_t), \bar{X}_0 = \iota$, with finite constants*

$$\bar{K} = \sup_t \|\bar{v}_t\|_{L^2(\bar{\rho}_t)}, \quad \bar{A} = \sup_t \|(\partial_t + \bar{v}_t\cdot\nabla)\bar{v}_t\|_{L^2(\bar{\rho}_t)}, \quad \bar{B} = \sup_t \left\|\frac{\mathrm{d}}{\mathrm{d}t}\left[\nabla\log\bar{\rho}_t \circ \bar{X}_t\right]\right\|_{L^2(\bar{\rho}_t)}.$$

Assumption 7 (Bounded Residual). *The residuals of the single-step form*

$$M_k = \iota + u_k - s_k\nabla W_{\rho_{k+1}} \circ M_k + r_k$$

satisfy

$$\|r_k\|_{L^2(\rho_k)} \le \varepsilon(\bar\tau)\tau_k^2 \quad \textit{for all } k, \qquad \sum_k \zeta_k \le \delta_0(\bar\tau),$$

with $\varepsilon(\bar\tau) \to 0$, $\delta_0(\bar\tau) \to 0$ *as* $\bar\tau \to 0$. $\mathcal{O}(\bar\tau)$ *rate required* $\varepsilon(\bar\tau) \le c\bar\tau$. *Both are automatically zero when* $d = 1$.

In this manuscript, we assume the residual upper bound condition 7. However, it is possible to show assumption is valid if the stronger assumption holds on Brenier potential and the score function: 1) The Brenier potential ϕ_k is uniformly strongly convex and uniformly smooth 2) $\rho_{k+1} \in W^{1,1}_{\text{loc}}$ which is the local Sobolev function, also $\nabla \log \rho_{k+1} \in L^2(\rho_{k+1})$ i.e. ρ_{k+1} has finite Fisher information.

We leave the proof with assumption 7 first, and show that it is derivable from stronger but natural assumption on Brenier potential and the score at appendix B.7.

B.2 Definitions of W-AIGF Solution

We show that the key reformulation sits on the fact that the first-order optimality condition of the proximal step provides the semi-implicit Euler discretization of AIGF. In the proof, we specifically focus on the Wasserstein Accelerated Information Gradient Flow (W-AIG Flow) defined with respect to the Wasserstein metric $G^W(\rho)^{-1}\Phi = -\nabla\cdot(\rho\nabla\Phi)$ as

$$\partial_t\rho_t + \nabla\cdot(\rho_t\nabla\Phi_t) = 0, \qquad \partial_t\Phi_t + \alpha_t\Phi_t + \frac{1}{2}|\nabla\Phi_t|^2 + \frac{\delta E}{\delta\rho_t} = 0. \tag{14}$$

We define the weak solution of the Wasserstein-AIGF which is the main target of study in our work.

Definition 4 (Weak solution of the W-AIG flow–velocity form)**.** *A pair* (ρ, Φ) *is a weak solution of Equation* (9) *with data* $(\rho_0, 0)$ *if:*

- ***(a) Regularity.*** $t \mapsto \rho_t$ *is continuous into* $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ *with uniformly finite (with* t*) second moment;* $\Phi(\cdot, t) \in C^{1,1}(\mathbb{R}^d)$ *with* $\nabla\Phi_t$ *L-Lipschitz in* x *uniformly in* t *and of linear growth;* Φ *is normalized as a family of addition-invariant equivalent class in* $E(\Omega)/\mathbb{R}$ *by* $\Phi_t(0) = 0$*,* $\Phi_0 \equiv 0$*th. Set* $v_t := \nabla\Phi_t$*; (a) forces* $\sup_t \|v_t\|_{L^2(\rho_t)} < \infty$.

- ***(b) Continuity equation.*** *For all* $\varphi \in C_c^\infty(\mathbb{R}^d \times [0, T))$*,*

$$\int_0^T \int [\partial_t\varphi + \langle\nabla\varphi, v_t\rangle]\,\mathrm{d}\rho_t\,\mathrm{d}t + \int \varphi(\cdot, 0)\,\mathrm{d}\rho_0 = 0.$$

- ***(c) Momentum equation.*** *For all* $\xi \in C_c^\infty(\mathbb{R}^d \times [0, T); \mathbb{R}^d)$*,*

$$\int_0^T \int [\langle\partial_t\xi, v_t\rangle + \langle(D_x\xi)v_t, v_t\rangle - \alpha_t\langle\xi, v_t\rangle - \langle\xi, \nabla V\rangle + \nabla\cdot\xi]\,\mathrm{d}\rho_t\,\mathrm{d}t = 0.$$

which correspond to the first and second equation of 14.

Definition 5 (Weak solution of the W-AIG flow)**.** *A pair* (ρ, Φ) *is a weak solution of 7 with data* $(\rho_0, \Phi_0 = 0)$ *if* $\Phi(\cdot, t) \in C^{1,1}_{loc}(\mathbb{R}^d)$ *with* $\nabla\Phi$ *Lipschitz in* x *uniformly in* t *of linear growth, with normalization* $\Phi(0, t) = 0, \Phi(\cdot, 0) \equiv 0$*. Furthermore,* $(\rho, \nabla\Phi)$ *is a weak solution in the sense of Definition 4.*

We say the second equation holds pointwise up to gauge if there is an absolutely continuous $g : [0, T] \to \mathbb{R}$*,* $g(0) = 0$ *such that* $[\Phi] := \Phi - g$ *satisfies*

$$\partial_t[\Phi] + \alpha_t[\Phi] + \frac{1}{2}|\nabla[\Phi]|^2 + \frac{\delta E}{\delta\rho_t} = 0 \qquad \textit{a.e. on } \mathbb{R}^d \times (0, T),$$

with $\frac{\delta E}{\delta\rho_t} = V + \log\rho_t + textconst.$

B.3 One-step Score Regularization

Here, we provide a regularity condition of the score:

Lemma 1. *Let $\rho \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ with $\rho > 0$ a.e., $\log \rho \in W^{1,1}_{\mathrm{loc}}(\mathbb{R}^d)$ and score is at most linearly growing for a.e. x. Let $v \in C^{1,1}$ with L-Lipschitz and L_2-Hölder, and $\tau L \le \frac{1}{2}$. Put $S := \iota - \tau v$, $M := S^{-1}$ and $\rho_{k+1} := M_{\#}\mu_k$. Then:*

1. *S and M are bi-Lipschitz by $1+\tau L$ and $(1-\tau L)^{-1}$ respectively. For $\det DS \in [(1-\tau L)^d, (1+\tau L)^d]$*
$$\rho_{k+1}(y) = \mu_k(S(y)) \det(DS(y)) \qquad \text{a.e.}$$
2. *$\rho_{k+1} > 0$ a.e., $\log \rho_{k+1}, \rho_{k+1} \in W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^d)$, and for a.e. y*
$$\nabla \log \rho_{k+1}(y) = DS(y)^\top \nabla \log \mu(S(y)) + \nabla[\log\det DS](y), \quad |\nabla[\log\det DS](y)| \le 2d\tau L_2;$$
3. *$|\nabla \log \rho_{k+1} + (y)| \le a^+ + b^+|y|$ with*
$$a^+ := (1+\tau L)(a + \tau b|v(0)|) + 2d\tau L_2$$
$$b^+ := (1+\tau L)^2 b$$
and in particular $\nabla \log \rho_{k+1} \in L^2(\rho_{k+1})$ whenever $\rho_{k+1} \in \mathcal{P}_2$.

Proof. (1) Since
$$(1-\tau L)|y-y'| \le |S(y) - S(y')| \le (1+\tau L)|y-y'|,$$
S is injective, Lipschitz. The image is all of $\mathbb{R}^d$ because $y \mapsto z + \tau v(y)$ is a contraction and has a fixed point $y = M(z)$ for all z. The change of variables formula for bi-Lipschitz maps gives the density of $M_\sharp \mu$, and $DS = I - \tau Dv$ with $\|\tau Dv\| \le \tau L \le \frac{1}{2}$ gives the determinant bounds.

(2) Since the map is bi-Lipschitz, we have $\{\rho_{k+1} = 0\} = S^{-1}\{\mu = 0\}$ is Lebesgue null,

We naturally have the regularity of the second summand:
$$y \mapsto DS(y) = I - \tau Dv(y)$$
is τL_2-Lipschitz with values in the compact convex set $\mathcal{K} := \{X : \|X - I\|_{\mathrm{op}} \le \frac{1}{2}\}$, on which the map to log determinant is smooth, with
$$d(\log\det)_X[Y] = \mathrm{tr}(X^{-1}Y), \qquad |\mathrm{tr}(X^{-1}Y)| \le d\|X^{-1}\|_{\mathrm{op}}\|Y\|_{\mathrm{op}} \le 2d\|Y\|_{\mathrm{op}}.$$
Hence, $\log\det DS$ is Lipschitz and, at every point of Dv that is differentiable, we have
$$\partial_i \log\det DS(y) = -\tau\,\mathrm{tr}(DS(y)^{-1}\partial_i Dv(y)),$$
which gives
$$|\partial_i \log\det DS(y)| \le 2d\tau\|\partial_i Dv(y)\|_{\mathrm{op}} \le 2d\tau L_2.$$
The first summand is regular because $\log\mu \in W^{1,1}_{loc}$ and S bi-Lipschitz imply that $\log\mu \circ S \in W^{1,1}_{loc}$ with the chain rule
$$\nabla(\log\mu \circ S) = DS^T(\nabla \log\mu \circ S) \quad \text{a.e.}$$
Adding these two, we have the formula.

(iii) We have
$$\|DS(y)^\top\|_{\mathrm{op}} \le 1 + \tau L, \qquad |S(y)| \le |y| + \tau|v(y)| \le (1+\tau L)|y| + \tau|v(0)|.$$
Inserting this into the formula obtained at (ii), we have
$$\int (a^+ + b^+|y|)^2 \mathrm{d}\rho_{k+1} \le 2(a^+)^2 + 2(b^+)^2 M_2(\rho_{k+1}) < \infty.$$
Then, we conclude with $\nabla \log\rho_{k+1} \in L^\infty_{loc}$, hence $\log\rho_{k+1} \in W^{1,\infty}_{loc}$ is locally Lipschitz, hence so is $\rho_{k+1} = \exp(\log\rho_{k+1})$ with $\nabla\rho_{k+1} = \rho_{k+1}\nabla\log\rho_{k+1}$ directly. □

Corollary 1. *Assuming 1-5 and letting $(\rho_k)_{k\le K}$ be the A-JKO or IA-JKO chain with*

$$M_2(\rho_k)^{1/2} \le M_2(\rho_0)^{1/2} + C_0 T, \qquad \|v_k\|_{\rho_k} \le C_0.$$

Then, every ρ_k satisfies $\rho_k > 0$ a.e. $\rho_k \in W^{1,\infty}_{loc}$ and

$$|\nabla \log \rho_k(x)| \le a_k + b_k|x|, \qquad b_k \le e^{2Lt_k} b_0, \qquad a_k \le e^{Lt_k}\Big(a_0 + t_k\big(e^{2LT} b_0 A + 2dL_2\big)\Big),$$

where $A := C_0 + L(M_2(\rho_0)^{1/2} + C_0 T)$ bounds $|v_k(0)|$ for all k. Consequently there is a uniform C_W with

$$|\nabla W_{\rho_k}(x)| \le C_W(1+|x|) \quad a.e., \text{ all } k, \qquad \|\nabla W_{\rho_k}\|_{L^2(\rho_k)} \le C_W(1 + M_2(\rho_k)^{1/2}) \le C'_W.$$

Proof. $|v_k(0)| \le |v_k(x)| + L|x|$ for every x. Integrating this against ρ_k and using Cauchy-Schwarz, we have

$$|v_k(0)| \le \|v_k\|_{\rho_k} + L M_2(\rho_k)^{1/2} \le A.$$

Applying lemma 1, with $(\rho, v, \tau) = (\rho_k, v_{k+1}, \tau_k)$, the recursion becomes

$$a_{k+1} = (1+\tau_k L)(a_k + \tau_k b_k A) + 2d\tau_k L_2, \qquad b_{k+1} = (1+\tau_k L)^2 b_k.$$

This is solved by the Grönwall inequality. The bound on ∇W_{ρ_k} further adds $|\nabla V(x)| \le |\nabla V(0)| + L_v|x|$ and L^2 bound follows from

$$\int (1+|x|)^2 \mathrm{d}\rho_k \le 2(1 + M_2(\rho_k)).$$

This completes the proof. □

B.4 Well-posedness of Single A-JKO and IA-JKO Step

We first establish the single A-JKO and IA-JKO step optimality to show the well-posedness and useful results.

Proposition 6 (Well-posedness of Single A-JKO Step)**.** *Assume 1-5 and $\rho_k \in \mathcal{P}_2(\mathbb{R}^d)$ absolutely continuous with $\mathrm{Ent}(\rho_k) < \infty$, with bounded second order moment $M_2(\mu)$. Then, the following statements hold:*

1. *The proximal problem at Equation (12) accepts a unique minimizer ρ_{k+1} which is absolutely continuous with respect to the finite entropy*
2. *$\rho_{k+1} \in W^{1,1}_{loc}(\mathbb{R}^d)$ (weakly differentiable with locally L^1 gradient), and $\nabla W_{\rho_{k+1}} \in L^2(\rho_{k+1}; \mathbb{R}^d)$, and the backward optimal map $S_k = T_{\rho_{k+1}\to\mu_k}$ satisfies*
$$S_k = \iota + s_k \nabla W_{\rho_{k+1}} \qquad \rho_{k+1}\text{-a.e.} \tag{15}$$
3. *Let $M_k := T_k \circ (\iota + u_k)$ where $T_k = T_{\mu_k \to \rho_{k+1}}$. The identity for the forward map*
$$M_k = \iota + u_k - s_k \nabla W_{\rho_{k+1}} \circ M_k + r_k, \tag{16}$$
holds with $r_k = 0$. Moreover, v_{k+1} defined by $v_{k+1} \circ M_k = (M_k - \iota)/\tau_k$ is well-defined ρ_{k+1}-a.e. with $\|v_{k+1}\|_{\rho_{k+1}} = \|M_k - \iota\|_{\rho_k}/\tau_k$.

Proof. **Existence.** Let $J_A(\rho) := E(\rho) + \frac{1}{2s_k} W_2^2(\mu_k, \rho)$ on $\mathcal{P}_2(\mathbb{R}^d)$ with $E(\rho) := \int V d\rho + \mathrm{Ent}(\rho)$ if ρ is absolutely continuous and $E(\rho) := \infty$ otherwise. By assumption 1, ∇V is L_V-Lipschitz which gives

$$|V(x)| \le |V(0)| + |\nabla V(0)||x| + \frac{L_V}{2}|x|^2,$$

so $\int V d\rho$ is finite on $\mathcal{P}_2$ and $E(\rho) \ge -\log Z$. Consider the minimizing sequence (ρ^n). We have bounded W_2 as

$$W_2^2(\mu_k, \rho^n) \le 2s_k(J_A(\rho^n) + \log Z),$$

which gives

$$M_2(\rho^n) \le 2M_2(\mu_k) + 2W_2^2(\mu_k, \rho^n)$$

bounded. The sequence is tight and subsequence narrowly converges to some ρ^* with

$$M_2(\rho^*) \le \liminf M_2(\rho^n) < \infty \qquad \text{(Fatou's lemma, (Rudin, 1974)).}$$

Narrow l.s.c. of $\rho \mapsto W_2^2(\mu_k, \rho)$ and $\rho \mapsto \mathrm{KL}(\rho\|\pi)$ (Ambrosio et al., 2008, Proposition 7.1.3, 9.4.3) gives

$$J_A(\rho^*) \le \liminf J_A(\rho^n) = \inf J_A.$$

In particular $\mathrm{KL}(\rho^*\|\pi) < \infty$, so ρ^* is absolutely continuous and

$$\mathrm{Ent}(\rho^*) = \mathrm{KL}(\rho^* \parallel \pi) - \log Z - \int V \mathrm{d}\rho^* < \infty.$$

The same argument with J_A replaced by J_{IA} and $W_2^2(\mu_k, \cdot)$ by the coercive bound gives the first existence proof of IA-JKO convergence existence.

Uniqueness. Let ρ^0 and ρ^1 be the minimizer of the optimal maps T^0, T^1 from μ_k. Consider the convex combination of maps $\rho_s = ((1-s)T^0 + sT^1)_\sharp \mu_k$. Along the curve provided, the metric term can be computed as an exact quadratic function

$$\frac{1}{2}W_2^2(\mu_k, \rho_s) = \frac{1}{2}\left\|(1-s)T^0 + sT^1 - \iota\right\|_{\mu_k}^2,$$

from the fact that ρ_s is again an optimal map and E is λ^--convex along the convex combination (Ambrosio et al., 2008, Section 9.2). Assumption 2 gives the strong convexity along the curvature which for objective $G(\cdot)$, we obtain the bound

$$G(\rho_s) \le (1-s)G(\rho^0) + sG(\rho^1) - \frac{1+\lambda^- s_k}{2} s(1-s)\|T^1 - T^0\|_{\mu_k}^2.$$

Since ρ^0 and ρ^1 are minimizer of optimal maps and ρ_s is a value in between, by letting $G(\rho^0) = G(\rho^1) = m$, we have

$$m \le G(\rho_s) \le m - \frac{1+\lambda^- s_k}{2} s(1-s)\|T^1 - T^0\|_{\mu_k}^2,$$

resulting in the $T^0 = T^1$ μ_k-a.e. as $s(1-s) > 0$ and $\frac{1+\lambda^- s_k}{2} > 0$ at $s \in \mathrm{Int}([0,1])$.

Optimality Condition. Fix $\xi \in C_c^\infty$, let $\rho_\varepsilon = (\iota + \varepsilon\xi)_\sharp \rho_{k+1}$ for small ε. We expand the both terms of the objective function Equation (12) at ρ_ε. For the functional we consider the standard energy functional $E = \int V d\rho + \mathrm{Ent}(\rho)$. Since ∇V is L_v-Lipschitz, we obtain the double-sided bound for the perturbation

$$\int V d\rho_\varepsilon = \int V(y + \varepsilon\xi(y)) d\rho_{k+1}(y),$$

given by

$$|V(y+\varepsilon\xi(y)) - V(y) - \varepsilon\langle\nabla V(y), \xi(y)\rangle| \le \frac{L_V}{2}\varepsilon^2|\xi(y)|^2.$$

Therefore, the difference in the first term is

$$\int V \, d\rho_\varepsilon - \int V \, d\rho_{k+1} = \varepsilon \int \langle \nabla V, \xi \rangle \, d\rho_{k+1} + O(\varepsilon^2),$$

Subsequently, we obtain the representation of the entropy with respect to perturbed variable $\phi(y) = y + \varepsilon\xi(y)$ from the diffeomorphism relation $\rho_\varepsilon(\phi_\varepsilon(y)) \cdot \det(D\phi_\varepsilon(y)) = \rho_{k+1}(y)$ as

$$\begin{aligned}\mathrm{Ent}(\rho_\varepsilon) &= \int \rho_\varepsilon(\Phi(y)) \log \rho_\varepsilon(\Phi(y)) \cdot \det(I + \varepsilon D\xi(y)) \, dy \\ &= \int \rho_{k+1}(y) \Big[\log \rho_{k+1}(y) - \log\det(I + \varepsilon D\xi(y)) \Big] dy.\end{aligned}$$

Leveraging the fact that $\rho_\varepsilon(y + \varepsilon\xi(y)) \det(I + \varepsilon D\xi(y)) = \rho_{k+1}(y)$, we obtain the similar difference for the entropy as

$$\mathrm{Ent}(\rho_\varepsilon) - \mathrm{Ent}(\rho_{k+1}) = -\int \log\det(I + \varepsilon D\xi) \, d\rho_{k+1}.$$

Jacobi's formula $\det(I + \varepsilon A) = 1 + \varepsilon \operatorname{tr}(A) + O(\varepsilon^2)$, gives the equality $\log\det(I + \varepsilon D\xi(y)) = \varepsilon\nabla \cdot \xi(y) + O(\varepsilon^2)$, which results in

$$\mathrm{Ent}(\rho_\varepsilon) - \mathrm{Ent}(\rho_{k+1}) = -\varepsilon \int \nabla \cdot \xi d\rho_{k+1} + O(\varepsilon^2).$$

Therefore, for a small variation $y \mapsto y + \varepsilon\xi(y)$, the energy functional term gives

$$E(\rho_\varepsilon) - E(\rho_{k+1}) = \varepsilon \int \left[\langle \nabla V, \xi\rangle - \nabla \cdot \xi\right] d\rho_{k+1} + \mathcal{O}(\varepsilon^2).$$

For the map from $S_k = T$ as $\rho_{k+1} \to \mu_k$, we have the inequality

$$\begin{aligned} W_2^2(\rho_\varepsilon, \mu_k) &\le \int \left|(y - S_k(y)) + \varepsilon\xi(y)\right|^2 d\rho_{k+1}(y) \\ &= \int |y - S_k(y)|^2 \, d\rho_{k+1} + 2\varepsilon \int \langle \xi, \iota - S_k\rangle \, d\rho_{k+1} + \varepsilon^2 \int |\xi|^2 \, d\rho_{k+1} \\ &= W_2^2(\rho_{k+1}, \mu_k) + 2\varepsilon \int \langle \xi, \iota - S_k\rangle \, d\rho_{k+1} + \varepsilon^2 \|\xi\|^2_{\rho_{k+1}}. \end{aligned}$$

The metric term is then bounded above by the competitor coupling that move each point y to $y + \varepsilon\xi(y)$ while keeping its partner $S_k(y)$:

$$\frac{1}{2} W_2^2(\rho_\varepsilon, \mu_k) \le \frac{1}{2} W_2^2(\rho_{k+1}, \mu_k) + \varepsilon \int \langle \xi, \iota - S_k\rangle \, \mathrm{d}\rho_{k+1} + \frac{\varepsilon^2}{2} \|\xi\|^2_{\rho_{k+1}}.$$

Combining the equations together with consideration of the weight $1/s_k$, we obtain for all small ε, we have on both sides of $\pm\xi$ resulting in

$$0 \le \varepsilon \int \left[\langle \nabla V, \xi\rangle - \nabla \cdot \xi + \frac{1}{s_k}\langle \xi, \iota - S_k\rangle\right] \mathrm{d}\rho_{k+1} + C_\xi \varepsilon^2,$$

where the bracket integrates to zero at all ξ. This implies that the distributional gradient of ρ_{k+1} is the locally L_1 function; hence $\rho_{k+1} \in W^{1,1}$ locally and

$$\nabla \log \rho_{k+1} = -\nabla V - \frac{1}{s_k}(\iota - S_k),$$

ρ_{k+1}-a.e., which is the second argument of the proposition.

Zero-Residual Identity Since $\mu_k, \rho_{k+1} \in \mathcal{P}_2(\mathbb{R}^d)$ absolutely continuous, the forward and backward optimal maps satisfy $S_k \circ T_k = \iota$ μ_k-almost everywhere (Santambrogio, 2015, Theorem 1.22). Applying the map S_k on the forward map $T_k(\hat{x})$ and applying Equation (15) to $y = T_k(\hat{x})$, we have

$$\hat{x} = T_k(\hat{x}) + s_k \nabla W_{\rho_{k+1}}(T_k(\hat{x})).$$

When we substitute $\hat{x} = x + u_k(x)$ and rearranging the formulation gives the desired result with $r_k = 0$. Finally, there exists the map $M_k^{-1} = (\iota + u_k)^{-1} \circ S_k$ ρ_{k+1}-a.e., so the velocity field v_{k+1} for the momentum advance

$$v_{k+1} \circ M_k = \frac{M_k - \iota}{\tau_k} \qquad \rho_k\text{-a.e.}$$

is well defined with $\|v_{k+1}\|_{\rho_{k+1}} = \|M_k - \iota\|_{\rho_k}/\tau_k$, completing the proof. □

The provided proposition is a statement regarding the optimality of a single-step Accelerated-JKO iteration that incorporates the well-posedness, optimality and useful identity that are derived from the proximal problem that are used for JKO step. A similar single-step well-posedness guarantee can be obtained for the Implicitly-Accelerated JKO

Proposition 7 (Well-posedness of Single IA-JKO Step)**.** *Assuming 1-3,5 and $\rho_k \in \mathcal{P}_2(\mathbb{R}^d)$ absolutely continuous with $\mathrm{Ent}(\rho_k) < \infty$, with bounded second order moment $M_2(\mu)$. Then, the following statements hold:*

1. *The IA-JKO step at Equation (13) admits a unique minimizer ρ_{k+1} on $\mathcal{P}_2(\mathbb{R}^d)$ that is absolutely continuous with finite entropy. Furthermore, we have $\rho_{k+1} \in W^{1,1}_{loc}(\mathbb{R}^d)$ and $\nabla W_{\rho_{k+1}} = \nabla V + \nabla \log \rho_{k+1} \in L^2(\rho_{k+1})$.*

2. *The inertial transport action admits the following priori bounds*

$$\mathcal{A}(\rho_k, \rho_{k+1}; u_k) \le \|u_k\|^2_{\rho_k} + 2s_k\big(E(\rho_k) - E(\rho_{k+1})\big) \le \|u_k\|^2_{\rho_k} + 2s_k(E(\rho_k) - m),$$

and consequently $W_2(\rho_k, \rho_{k+1}) \le \|u_k\|_{\rho_k} + \big(\|u_k\|^2_{\rho_k} + 2s_k(E(\rho_k) - m)\big)^{1/2}$.

3. *The Brenier map $M_k := T_{\rho_k \to \rho_{k+1}} = \nabla \phi_k$ where ϕ_k is a convex potential that has the reverse direction Brenier map $M_k^{-1} = T_{\rho_{k+1} \to \rho_k} = \nabla \phi_k^*$, with $M_k^{-1} \circ M_k = \iota$ ρ_k-a.e. and $M_k \circ M_k^{-1} = \iota$ ρ_{k+1}-a.e. Hence $D_{k+1} := \iota - M_k^{-1}$ and $v_{k+1} := D_{k+1}/\tau_k$ are well defined ρ_{k+1}-a.e.,*

$$v_{k+1} \circ M_k = \frac{M_k - \iota}{\tau_k} \quad \rho_k\text{-a.e.}, \qquad \|v_{k+1}\|_{\rho_{k+1}} = \frac{\|M_k - \iota\|_{\rho_k}}{\tau_k} = \frac{W_2(\rho_k, \rho_{k+1})}{\tau_k},$$

and $v_{k+1} = -\tau_k^{-1} \log_{\rho_{k+1}} \rho_k = \tau_k^{-1} \nabla\big(\frac{1}{2}|\cdot|^2 - \phi_k^*\big) \in \mathrm{Tan}_{\rho_{k+1}}$.

4. *For every $\rho \in \mathcal{P}_2(\mathbb{R}^d)$, we have*

$$W_2^2(\mu_k, \rho) \le \mathcal{A}(\rho_k, \rho; u_k),$$

which gives

$$E(\rho) + \frac{1}{2s_k} W_2^2(\rho_k, \rho) \le E(\rho) + \frac{1}{2s_k} \mathcal{A}(\rho_k, \rho; u_k)$$

the upper bound of A-JKO by IA-JKO. If Assumption 4 holds such that μ_k is absolutely continuous, the equality holds in for above inequality if and only if $T_{\rho_k \to \rho} = T_{\mu_k \to \rho} \circ (\iota + u_k)$ ρ_k-a.e.

Proof. **Existence.** Define the single-step minimizer argument for A-JKO $J_A = E(\cdot) + \frac{1}{2s_k} W_2^2(\mu_k, \rho)$ and IA-JKO $J_{IA} = E(\cdot) + \frac{1}{2s_k} \mathcal{A}(\rho_k, \rho; u_k)$. By definition of $\mathcal{A}$ and the fact that $\log_{\rho_k} \rho_k = 0$, we have $J_{IA}(\rho_k) = E(\rho_k) + \frac{1}{2s_k}\|u_k\|^2_{\rho_k}$ is finite and lower bounded by m. Furthermore, it is coercive which the triangle inequality in $L^2(\rho_k, \mathbb{R}^d)$ space gives

$$\begin{aligned} W_2(\rho_k, \rho) &= \|\log_{\rho_k} \rho\|_{\rho_k} \\ &\le \|u_k\|_{\rho_k} + \mathcal{A}(\rho_k, \rho; u_k)^{1/2} \\ &\le \|u_k\|_{\rho_k} + \big(2s_k(J_{\mathrm{IA}}(\rho) - m)\big)^{1/2}. \end{aligned} \tag{17}$$

Consider a minimizing sequence of marginals $(\rho^n)_n$, Equation (17) gives a bounded Wasserstein $W_2(\rho_k, \rho^n)$ which gives $M_2(\rho^n) \le 2M_2(\rho_k) + 2W_2^2(\rho_k, \rho^n)$. By Markov's inequality, the sequence is tight and converges narrowly to some ρ^* by Prokhorov's theorem. Concurrently, we have $M_2(\rho^*) \le \liminf_n M_2(\rho^n) < \infty$, i.e. $\rho^* \in \mathcal{P}_2(\mathbb{R}^d)$. The l.s.c of KL on $\mathcal{P}_2(\mathbb{R}^d)$ gives $E(\rho^*) \le \liminf_n E(\rho^n)$ and the inequality:

$$E(\rho^n) \le J_{IA}(\rho^n) \le \inf J_{IA} + 1$$

gives the absolutely continuity of ρ^* with finite entropy. Denote $T_n := T_{\rho_k \to \rho^n}$ and $T := T_{\rho_k \to \rho^*}$. The transport plan $\gamma_n := (\iota, T_n)_\sharp \rho_k$ are optimal for $\rho_k, \rho^n)$ with $\int |x - y|^2 d\gamma_n = W_2^2(\rho_k, \rho^n)$ bounded, and the convergence $\rho^n \to \rho^*$ occurs narrowly.

The sequence of plans (γ_n) are right and the stability of optimal plans (Ambrosio et al., 2008, Proposition 7.1.3) gives the convergence of further subsequence to the optimal plan (ρ_k, ρ^*). This is equivalent to $(\iota, T)_\sharp \rho_k$ by uniqueness of the optimal plan. The fixed first marginal and map induced by the limit plan forces the convergence of $T_n \to T$ in ρ_k-measure. Moreover, we have $\sup_n \|T_n - T\|_{\rho_k} \le \sup_n W_2(\rho_k, \rho^n) + W_2(\rho_k, \rho^*) < \infty$. We have a sequence that is bounded in $L^2(\rho_k; \mathbb{R}^d)$ and weakly converges in $L^2(\rho_k; \mathbb{R}^d)$ such that $\psi \in L^2(\rho_k; \mathbb{R}^d)$ and $R > 0$ gives

$$\left| \int \langle T_n - T, \psi \rangle \, d\rho_k \right| \le \left| \int \langle T_n - T, \psi \, \mathbf{1}_{\{|\psi| \le R\}} \rangle \, d\rho_k \right| + \sup_j \|T_j - T\|_{\rho_k} \|\psi \, \mathbf{1}_{\{|\psi| > R\}}\|_{\rho_k}.$$

Here, the first term converges to 0 as $n \to \infty$ and the second term is negligibly small for large R with uniformity in n. This gives $\log_{\rho_k} \rho^n - u_k \to \log_{\rho_k} \rho^* - u_k$ weakly in $L^2(\rho_k; \mathbb{R}^d)$ and weak l.s.c. of the norm gives the upper bound

$$\mathcal{A}(\rho_k, \rho^*; u_k) \le \liminf_n \mathcal{A}(\rho_k, \rho^n; u_k)$$

Combining the two bounds give ρ^* as the minimizer, that is absolutely continuous with finite entropy. Since $M_k^{-1} = \iota - \tau_k v_{k+1}$, combining Lemma 1 and corresponding Corollary 1 gives the last statement of Proposition 7.1.

Uniqueness. Let $\rho^0, \rho^1 \in \mathcal{P}_2(\mathbb{R}^d)$ with $J_{IA} < \infty$ and the Brenier map gives $T_n = \nabla \phi_n$ with convex ϕ_n. For $s \in [0, 1]$, define the convex combination of the Brenier map

$$T^s := (1-s)T^0 + sT^1 = \nabla((1-s)\phi_0 + s\phi_1)$$

and $\rho^s := T^s_\sharp \rho_k \in \mathcal{P}_2(\mathbb{R}^d)$. The map T^s is the Brenier map from ρ_k to ρ^1 (Ambrosio et al., 2008, Theorem 6.2.4) which gives

$$\log_{\rho_k} \rho^s = (1-s) \log_{\rho_k} \rho^0 + s \log_{\rho_k} \rho^1.$$

Furthermore on $L^2(\rho_k; \mathbb{R}^d)$, we obtain

$$\begin{aligned}\mathcal{A}(\rho_k, \rho^s; u_k) &= \|(1-s)a_0 + sa_1\|^2_{\rho_k} \\ &= (1-s)\mathcal{A}(\rho_k, \rho^0; u_k) + s\mathcal{A}(\rho_k, \rho^1; u_k) - s(1-s)\|T^1 - T^0\|^2_{\rho_k}.\end{aligned}$$

The curve $(\rho^s)_{s\in[0,1]}$ is a generalized geodesic with base ρ_k which the potential energy is λ-convex along the geodesic curve. Integrating the pointwise inequality gives

$$V((1-s)x + sy) \le (1-s)V(x) + sV(y) - \frac{\lambda}{2}s(1-s)|x-y|^2$$

at $(x, y) = (T^0(z), T^1(z))$ against $\rho_k(dz)$ gives the bound

$$\int V \, d\rho^s \le (1-s)\int V \, d\rho^0 + s\int V \, d\rho^1 - \frac{\lambda}{2}s(1-s)\|T^1 - T^0\|^2_{\rho_k}.$$

The entropy is convex along the geodesic with absolutely continuous base which gives

$$J_{\mathrm{IA}}(\rho^s) \le (1-s)J_{\mathrm{IA}}(\rho^0) + sJ_{\mathrm{IA}}(\rho^1) - \frac{1+\lambda^- s_k}{2s_k}s(1-s)\|T^1 - T^0\|^2_{\rho_k}. \tag{18}$$

If $J_{IA}(\rho^0) = J_{IA}(\rho^1) = \min J_{IA}$, then Equation (18) at $s = 1/2$ together with $J_{IA}(\rho^{1/2}) \ge \min J_{IA}$ and $1 + \lambda^- s_k \ge \frac{1}{2} > 0$ forces $\|T^1 - T^0\|_{\rho_k} = 0$ which gives uniqueness of T^0 and T^1 at ρ_k-a.e and $\rho^0 = T^0_\sharp \rho_k = T^1_\sharp \rho_k = \rho^1$, completing the part 1 of the proof.

Priori Bound. Now, the remainders are straightforward. Considering the minimality against $\rho = \rho_k$ gives

$$E(\rho_{k+1}) + \frac{1}{2s_k}\mathcal{A}(\rho_k, \rho_{k+1}; u_k) \le E(\rho_k) + \frac{1}{2s_k}\|u_k\|^2_{\rho_k},$$

which gives the first inequality. Using the fact that $E(\rho_{k+1}) \ge m$, we obtain the second bound with the W_2 bound with Equation (17) at $\rho = \rho_{k+1}$.

Brenier Map. By (1), ρ_k and ρ_{k+1} are absolutely continuous with finite second moments, such that the Brenier map $M_k = \nabla \phi_k$ is injective. Its inverse is the Brenier map in the reverse direction such that $M_k^{-1} = \nabla \phi_k^* = T_{\rho_{k+1} \to \rho_k}$ and

$$M_k^{-1} \circ M_k = \iota \quad \rho_k\text{-a.e.}, \qquad M_k \circ M_k^{-1} = \iota \quad \rho_{k+1}\text{-a.e.}$$

(Santambrogio, 2015, Theorem 1.22). Hence, $v_{k+1} - (\iota - M_k^{-1})/\tau_k$ is defined ρ_{k+1}-a.e. such that

$$v_{k+1} \circ M_k = (M_k - M_k^{-1} \circ M_k)/\tau_k = (M_k - \iota)/\tau_k$$

ρ_k-a.e. Then by change of variables with $(M_k)_\sharp \rho_k = \rho_{k+1}$,

$$\|v_{k+1}\|^2_{\rho_{k+1}} = \int |v_{k+1} \circ M_k|^2 \, d\rho_k = \tau_k^{-2}\|M_k - \iota\|^2_{\rho_k} = \tau_k^{-2} W_2^2(\rho_k, \rho_{k+1}).$$

Finally, we have the relationship $v_{k+1} = -\tau_k^{-1}(T_{\rho_{k+1}\to\rho_k} - \iota) = -\tau_k^{-1}\log_{\rho_{k+1}}\rho_k$, which lies in $\mathrm{Tan}_{\rho_{k+1}}$ because $\log_\rho \nu \in \mathrm{Tan}_\rho$ for every absolutely continuous ρ, and $\nu \in \mathcal{P}_2$. The gradient representation is then naturally given by $\iota - \nabla\phi_k^* = \nabla(\frac{1}{2}|\cdot|^2 - \phi_k^*)$. This completes the proof for the third part.

A-JKO and IA-JKO bound. Now, to show the comparison with A-JKO, we set $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ and put $\gamma := (\iota + u_k, T_{\rho_k\to\rho})_\sharp\rho_k$. Its marginals then are $(\iota + u_k)_\sharp\rho_k = \mu_k$ and $(T_{\rho_k\to\rho})_\sharp\rho_k = \rho$. So, for the coupling $\gamma \in \Pi(\mu_k, \rho)$, and

$$W_2^2(\mu_k, \rho) \le \int |y - z|^2 \, d\gamma(y, z) = \int |T_{\rho_k\to\rho}(x) - x - u_k(x)|^2 \, d\rho_k(x) = \mathcal{A}(\rho_k, \rho; u_k). \quad (19)$$

Adding $E(\rho)$ and dividing both sides by $2s_k$ gives the inequality $J_A \le J_{IA}$. This gives the inequality between the infimum. Assuming 4 such that μ_k is absolutely continuous and the optimal plan between μ_k and ρ is unique and equal to

$$(\iota, T_{\mu_k\to\rho})_\sharp\mu_k = (\iota + u_k, T_{\mu_k\to\rho} \circ (\iota + u_k))_{\rho_k}.$$

Equality in Equation (19) implies γ is optimal. If $T_{\rho_k\to\rho} = T_{\mu_k\to\rho} \circ (\iota + u_k)$ ρ_k-a.e., the two push-forward maps coincide. Conversely, if γ is concentrated on $\{(y, z) : z = T_{\mu_k\to\rho}(y)\}$ then, since γ is the law of $x \mapsto (x + u_k(x), T_{\rho_k\to\rho}(x)$ under ρ_k, we obtain the map $T_{\rho_k\to\rho}(x) = T_{\mu_k\to\rho}(x + u_k(x))$ for ρ_k-a.e. This completes the last section of the proof. □

Based on the single-step optimality, it is possible to then extend this to the crucial inequality required for the proof of the convergence. One useful observation is that is it possible to define some quantities based on the physics of the system such as total energy.

B.5 Total Energy Quantities of (Implicitly) Accelerated-JKO Minimizer

Proposition 8. *Define the total energy and the one-step quantities*

$$E_k = E(\rho_k) + \tfrac{1}{2}\|v_k\|_{\rho_k}^2, \qquad w_k = \frac{M_k - \iota}{\tau_k} = v_{k+1} \circ M_k, \qquad e_k = \|w_k - v_k\|_{\rho_k}$$

so that $\|w_k\|_{\rho_k} = \|v_{k+1}\|_{\rho_{k+1}}$. *The energy* E_k *is the discrete AIGF Hamiltonian* $E(\rho) + \frac{1}{2}\int |\nabla\Phi|^2 \, \mathrm{d}\rho$ *of Wang & Li (2022). Let* ρ_{k+1} *be the A-JKO or IA-JKO minimizer under well-posedness of the respective problems. Then,*

$$E_{k+1} + \frac{\alpha_k\tau_k}{2}\|v_{k+1}\|_{\rho_{k+1}}^2 + \frac{1}{2}e_k^2 \le E_k + \frac{|\lambda^-|\tau_k^2}{2}\|v_{k+1}\|_{\rho_{k+1}}^2 + \zeta_k \quad (20)$$

where $\zeta_k = 0$ *for IA-JKO and A-JKO has*

$$\zeta_k = \frac{\|M_k - \iota\|_{\rho_k}^2 - W_2^2(\rho_k, \rho_{k+1})}{2\tau_k^2},$$

vanishing when M_k *is an optimal map from* ρ_k.

This is an inequality that relates the energy between the two different steps of A-JKO and IA-JKO iteration.

Proof. We give the IA-JKO computation, while A-JKO case differs in the geodesic term.

Implicitly-Accelerated Case We first write $a = M_k - \iota$ and recall $u_k = \beta_k\tau_k v_k, s_k = \beta_k\tau_k^2$. This gives the formulation

$$\frac{\tau_k^2}{2s_k} = \frac{1}{2\beta_k}, \qquad \frac{\|u_k\|_{\rho_k}^2}{2s_k} = \frac{\beta_k}{2}\|v_k\|_{\rho_k}^2. \quad (21)$$

For $s \in [0, 1]$, let $T_s = (1-s)T_{\rho_k\to\rho_{k+1}} + s\iota$ as an convex interpolant between T and ι, $\rho_s = (T_s)_\sharp\rho_k$ being the interpolation of optimal maps from the minimizer back to ρ_k. The action along this curve is the quadratic

$$\mathcal{A}_k(\rho_s) := \mathcal{A}(\rho_k, \rho_s; u_k) = \|\log_{\rho_k}\rho_s - u_k\|_{\rho_k}^2 = \|(1-s)a - u_k\|_{\rho_k}^2,$$

and E is λ^--convex. When we write $h(s) := E(\rho_s) + \frac{1}{2s_k}\mathcal{A}_k(\rho_s)$, minimality condition $h(0) \le h(s)$ combined with the semi-convexity upper bound for $h(s)$ gives

$$\underbrace{E(\rho_k) + \frac{\|u_k\|^2_{\rho_k}}{2s_k}}_{=h(1)} \ge \underbrace{E(\rho_{k+1}) + \frac{\|a-u_k\|^2_{\rho_k}}{2s_k}}_{=h(0)} + \underbrace{\frac{1+\lambda^- s_k}{2s_k}\|a\|^2_{\rho_k}}_{\text{strong-convexity gain}}. \tag{22}$$

Substitute $a = \tau_k w_k, u_k = \beta_k\tau_k v_k$, and Equation (21) into Equation (22) which gives

$$E(\rho_{k+1}) + \frac{1}{2\beta_k}\|w_k - \beta_k v_k\|^2_{\rho_k} + \frac{1+\lambda^- s_k}{2\beta_k}\|w_k\|^2_{\rho_k} \le E(\rho_k) + \frac{\beta_k}{2}\|v_k\|^2_{\rho_k}.$$

Expanding the squares and cancel $\frac{\beta_k}{2}\|v_k\|^2_{\rho_k}$ on both sides, we obtain

$$E(\rho_{k+1}) + \frac{1}{2\beta_k}\|w_k\|^2_{\rho_k} - \langle w_k, v_k\rangle_{\rho_k} + \frac{1+\lambda^- s_k}{2\beta_k}\|w_k\|^2_{\rho_k} \le E(\rho_k).$$

Now, use polarization identity of an inner product space

$$-\langle w_k, v_k\rangle_{\rho_k} = -\frac{1}{2}\|w_k\|^2_{\rho_k} - \frac{1}{2}\|v_k\|^2_{\rho_k} + \frac{1}{2}e_k^2$$

and collecting the terms with $\|w_k\|^2_{\rho_k}$, we have

$$\frac{1}{2\beta_k} + \frac{1+\lambda^- s_k}{2\beta_k} - \frac{1}{2} = \frac{(2+\lambda^- s_k)(1+\alpha_k\tau_k)}{2} - \frac{1}{2} = \frac{1}{2} + \alpha_k\tau_k + \frac{\lambda^-\tau_k^2}{2},$$

using $\beta_k^{-1} = 1 + \alpha_k\tau_k$ and $s_k/\beta_k = \tau_k^2$. Substituting this into equation B.5 and reorganizing, we obtain the inequality between the energy between the single implicitly-damped JKO step as

$$E(\rho_{k+1}) + \left(\frac{1}{2} + \alpha_k\tau_k + \frac{\lambda^-\tau_k^2}{2}\right)\|w_k\|^2_{\rho_k} + \frac{1}{2}e_k^2 \le E(\rho_k) + \frac{1}{2}\|v_k\|^2_{\rho_k},$$

which $\|w_k\|_{\rho_k} = \|v_{k+1}\|_{\rho_{k+1}}$ is the first equation of the proposition.

Accelerated JKO Case For A-JKO, we have the endpoint of the map T_k and $S_k = T_{\mu_k\to\rho_k}$. The value at $s=1$ is again bounded by $E(\rho_k) + \|u_k\|^2_{\rho_k}/(2s_k)$ while the strong-convexity gain term is now modified to

$$\frac{1+\lambda^- s_k}{2s_k}W_2^2(\rho_k,\rho_{k+1}) = \frac{1+\lambda^- s_k}{2s_k}\left(\|a\|^2_{\rho_k} - \Delta_k\right).$$

The geodesic sees the Wasserstein displacement, not the displacement of the composed map $M_k = T_k \circ (\iota + u_k)$ and their difference is exactly $\Delta_k \ge 0$. Similar computation shows same formulation with $\zeta_k = \Delta_k/(2\tau_k^2)$. □

We establish a simple bound based on the total energy bounds. This proposition provides the energy bound of the A- and IA-JKO over the iteration.

Corollary 2. *For $m = \inf_{\mathcal{P}_2}\mathcal{F} = -\log Z$, set*

$$\frac{1}{2}C_0^2 = e^{2|\lambda^-|T}\left(\mathcal{F}(\rho_0) - m + \delta_0(\bar\tau)\right), \qquad C_1^2 = C_0^2(1 + |\lambda^-|T) + 2\delta_0(\bar\tau).$$

Then for all $k \le K$, the following bounds hold:

$$\max_k\left[E_k - m\right] \le \frac{1}{2}C_0^2, \qquad \max_k\|v_k\|_{\rho_k} \le C_0, \qquad \sum_{k<K}\left[\alpha_k\tau_k\|v_{k+1}\|^2_{\rho_{k+1}} + e_k^2\right] \le C_1^2.$$

Moreover, the force field ∇W_ρ satisfy the bound

$$\tau_k\|\nabla W_{\rho_{k+1}}\|_{L^2(\rho_{k+1})} \le 2e_k + C\tau_k.$$

Proof. From Equation (20), we know $\alpha_k, e_k \geq 0$ which provides the looser inequality

$$E_{k+1} \leq E_k + \frac{|\lambda^-|\tau_k^2}{2}\|v_{k+1}\|^2_{\rho_{k+1}} + \zeta_k.$$

Since $\mathcal{F}(\rho_{k+1}) + \frac{1}{2}\|v_{k+1}\|^2_{\rho_{k+1}}$, we have

$$\|v_{k+1}\|^2_{\rho_{k+1}} = 2(E_{k+1} - \mathcal{F}(\rho_{k+1})) \leq 2(E_{k+1} - m),$$

rearranging gives

$$(1 - |\lambda^-|\tau_k^2)(E_{k+1} - m) \leq (E_k - m) + \zeta_k.$$

Since $|\lambda^-|\tau_k^2 \leq \frac{1}{2}$, $1 - |\lambda^-|\tau_k^2 \geq \frac{1}{2} > 0$, the elementary bound of $(1-x)^{-1} \leq e^{2x}$ for $x \in [0, \frac{1}{2}]$, we have a Grönwall-type bound of

$$E_{k+1} - m \leq e^{2|\lambda^-|\tau_k^2}\big[(E_k - m) + \zeta_k\big].$$

Applying this from $k = 0$ iteratively, it gives

$$E_K - m \leq e^{2|\lambda^-|\sum_{j<K}\tau_j^2}\Big[(E_0 - m) + \sum_{j<K}\zeta_j\Big],$$

where the same argument holds for any $k \leq K$.

Using the cumulative transport time $\sum_j \tau_j^2 \leq \bar{\tau}\sum_j \tau_j \leq \bar{\tau}T \leq T$, the residual budget $\sum_k \zeta_k \leq \delta_0(\bar{\tau})$, and $E_0 = \mathcal{F}(\rho_0)$ which gives the first bound

$$\max_k(E_k - m) \leq e^{2|\lambda^-|T}\big[\mathcal{F}(\rho_0) - m + \delta_0(\bar{\tau})\big] =: \frac{1}{2}C_0^2.$$

Finally, $\|v_k\|^2_{\rho_k} = 2\big(E_k - \mathcal{F}(\rho_k)\big) \leq 2(E_k - m) \leq C_0^2$, i.e.

$$\|v_k\|_{\rho_k} \leq C_0 \qquad \text{for every } k \leq K.$$

The last kinetic bound naturally follows from the summation after rearranging Equation (20)

$$\sum_{k<K}\left[\frac{\alpha_k\tau_k}{2}\|v_{k+1}\|^2_{\rho_{k+1}} + \frac{1}{2}e_k^2\right] \leq \underbrace{\sum_{k<K}(E_k - E_{k+1})}_{=E_0 - E_K} + \underbrace{\frac{|\lambda^-|}{2}\sum_{k<K}\tau_k^2\|v_{k+1}\|^2_{\rho_{k+1}}}_{\leq \frac{|\lambda^-|}{2}C_0^2T} + \underbrace{\sum_{k<K}\zeta_k}_{\delta_0(\bar{\tau})}$$

and combining to have

$$\sum_{k<K}\Big[\alpha_k\tau_k\|v_{k+1}\|^2_{\rho_{k+1}} + e_k^2\Big] \leq 2(E_0 - m) + |\lambda^-|C_0^2T + 2\delta_0(\bar{\tau}) =: C_1^2.$$

The force field bound is subsequently obtained by first considering the one step form Equation (16)

$$s_k\nabla W_{\rho_{k+1}} \circ M_k = \iota + u_k - M_k + r_k = -(M_k - \iota) + u_k + r_k.$$

Since $M_k - \iota = \tau_k w_k$, $u_k = \beta_k\tau_k v_k$ and $s_k = \beta_k\tau_k^2$,

$$\beta_k\tau_k\,\nabla W_{\rho_{k+1}} \circ M_k = -(w_k - \beta_k v_k) + \frac{r_k}{\tau_k} = -(1-\beta_k)v_k - (w_k - v_k) + \frac{r_k}{\tau_k},$$

from $-(w_k - \beta_k v_k) = -(1-\beta_k)v_k - (w_k - v_k)$. Take $L^2(\rho_k)$-norms and applying the triangle inequality, we obtain:

$$\beta_k\tau_k\|\nabla W_{\rho_{k+1}} \circ M_k\|_{\rho_k} \leq (1-\beta_k)\|v_k\|_{\rho_k} + \|w_k - v_k\|_{\rho_k} + \frac{\|r_k\|_{\rho_k}}{\tau_k}.$$

By pushforward M_k, we have $\|\nabla W_{\rho_{k+1}} \circ M_k\|_{L^2(\rho_k)} = \|\nabla W_{\rho_{k+1}}\|_{L^2(\rho_{k+1})}$. We have $\|w_k - v_k\|_{\rho_k}$ and $1 - \beta_k = \alpha_k\tau_k\beta_k$ which gives

$$\beta_k\tau_k\|\nabla W_{\rho_{k+1}}\|_{L^2(\rho_{k+1})} \leq \alpha_k\tau_k\beta_k\|v_k\|_{\rho_k} + e_k + \varepsilon(\bar{\tau})\tau_k \leq \bar{\alpha}C_0\tau_k + e_k + \varepsilon(\bar{\tau})\tau_k,$$

combined with the with the residual bound Assumption 7 and $\|v_k\|_{\rho_k} \leq C_0$, $\beta_k \leq 1$ and $\alpha_k \leq \bar{\alpha}$. Dividing by $\beta_k \geq \frac{1}{2}$ to provide the bound on $1/\beta_k \leq 2$ gives

$$\tau_k\|\nabla W_{\rho_{k+1}}\|_{L^2(\rho_{k+1})} \leq 2\big(\bar{\alpha}C_0\tau_k + e_k + \varepsilon(\bar{\tau})\tau_k\big) = 2e_k + C\tau_k, \quad C := 2\bar{\alpha}C_0 + 2\sup_{\bar{\tau}}\varepsilon(\bar{\tau}).$$

□

B.6 Convergence of (Implicitly) Accelerated-JKO to the AIGF

Lemma 2 (Uniform modulus of continuity in time)**.** *For every $R > 0$ there is a function $C_R : [0,\infty) \to [0,\infty)$ with $C_R(\delta) \to 0$ as $\delta \to 0^+$, depending only on ρ_0, L, C, C_0, T (the constants already fixed by Assumption 3, 5 and Corollary 2), and not on the mesh, such that for all $j \le k$ and $|x| \le R$,*

$$|v_k(x) - v_j(x)| \le C_R(t_k - t_j).$$

Consequently, for the interpolant $\bar{v}_{\bar\tau}$,

$$|\bar{v}_{\bar\tau}(x,t) - \bar{v}_{\bar\tau}(x,s)| \le 3\, C_R(|t-s| + \bar\tau) \qquad (x \in \bar{B}_R),$$

and the family $\{\bar{v}_{\bar\tau}\}_{\bar\tau}$ is uniformly bounded and equicontinuous on $\bar{B}_R \times [0,T]$.

Proof. Fix $|x| \le R$ and put $z := \Psi_j^{-1}(x)$. By the growth bound already established for Ψ_k,

$$|\Psi_i(z)| \le R' := e^{2LT}(R + CT) \qquad \text{for all } i.$$

Split the increment along the particle labeled z:

$$v_k(x) - v_j(x) = [v_k(x) - v_k(\Psi_k(z))] + [V_k - V_j](z), \qquad V_i := v_i \circ \Psi_i.$$

By the same spatial-transport estimate used for $|M_k x - x|$,

$$|x - \Psi_k(z)| = |\Psi_j(z) - \Psi_k(z)| \le \sum_{i=j}^{k-1} \tau_i |v_{i+1}(\Psi_{i+1}(z))| \le (C + LR')(t_k - t_j),$$

so the first bracket is at most $L(C + LR')(t_k - t_j)$.

The map $g := V_k - V_j$ is $2Le^{2LT}$-Lipschitz, using the Lipschitz bound already established for Ψ_k, and satisfies

$$\|g\|_{L^2(\rho_0)} \le C(t_k - t_j),$$

which strengthens the summable bound $\sum_k \tau_k e_k \to 0$ used elsewhere in the proof to theper-step rate $e_k \le C\tau_k$. This uses the uniform $C^{1,1}$ velocity bound now carried by Assumption 5. Since $\rho_0 > 0$ a.e. by Assumption 3, for every $r > 0$ and $y \in B_r(z)$,

$$|g(z)| \le |g(y)| + 2Le^{2LT} r,$$

by the Lipschitz bound on g and the triangle inequality. Averaging over $B_r(z)$ against ρ_0 and applying Cauchy–Schwarz,

$$\begin{aligned} |g(z)| &\le \Big(\inf_{|z| \le R'} \rho_0(B_r(z)) \Big)^{-1/2} \|g\|_{L^2(\rho_0)} + 2Le^{2LT} r \\ &\le \Big(\inf_{|z| \le R'} \rho_0(B_r(z)) \Big)^{-1/2} C(t_k - t_j) + 2Le^{2LT} r. \end{aligned}$$

To combine them together, set

$$C_R(\delta) := L(C + LR')\,\delta + \inf_{r>0} \left[\Big(\inf_{|z| \le R'} \rho_0(B_r(z)) \Big)^{-1/2} C\,\delta + 2Le^{2LT} r \right].$$

The infimum over $\rho_0(B_r(z))$ is strictly positive for each fixed $r > 0$: the map $z \mapsto \rho_0(B_r(z))$ is continuous (dominated convergence, since $\rho_0 > 0$ a.e. makes the boundary $\partial B_r(z)$ null) and strictly positive, hence bounded below on the compact set $\bar{B}_{R'}$. Taking $r \to 0$ slowly as $\delta \to 0$ (e.g. $r = \sqrt{\delta}$) gives $C_R(\delta) \to 0$.

For the interpolant, $\bar{v}_{\bar\tau}(x,t)$ agrees with $v_k(x)$ for $t \in [t_k, t_{k+1})$, and any two grid indices within $|t-s| + \bar\tau$ of each other in time satisfy the bound above, giving the stated factor of 3. Uniform boundedness of $\bar{v}_{\bar\tau}$ follows from the growth bound on v_k; spatial equicontinuity is immediate from $\mathrm{Lip}(v_k(\cdot,t)) \le L$. □

From the preliminary results above, we finally have the theorem of convergence for the (Implicitly) Accelerated JKO scheme to the W-AIG Flow

Theorem 2. *Assume 1-7. Let the time $\bar{\tau} \to 0$. Then, there exist a subsequence and a limit (ρ, ν) such that:*

1. *$\sup_{t\in[0,T]} W_2(\bar{\rho}_{\bar{\tau}}(t), \rho_t) \to 0$ and $\bar{\nu}_{\bar{\tau}} \to \nu$ uniformly on compact subsets of $\mathbb{R}^d \times [0, T]$ with the L-Lipschitz limit velocity in x and locally Lipschitz in t. We have $t \mapsto \rho_t$ Lipschitz into $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ as well as the convergence of the kinetic energy in Proposition 8 for every t.*
2. *(ρ, ν) is a weak solution of the velocity form W-AIG Flow in Definition 4.*
3. *Assuming the velocity field admits a gradient form $v_t = \nabla\Phi_t$ for the potential $\Phi(x, t) = \int_0^1 \langle v(sx, t), x\rangle ds$, and (ρ, Φ) is a weak solution of the W-AIG Flow in the sense of Definition 5. Under the positivity clause of Assumption 3, the second equation of (14) moreover holds pointwise up to gauge: a time-dependent, spatially constant shift $g(t)$, $g(0) = 0$.*
4. *Under the same hypotheses as Theorem 3, the limit is a particle W-AIG flow: $\rho_t = (X_t)_\sharp \rho_0$ for the Lipschitz flow $\dot{X}_t = \nabla\Phi_t(X_t), X_0 = \iota$ and for ρ_0-a.e. x the pair $(X_t(x), V_t(x) := \nabla\Phi_t(X_t(x)))$ is Lipschitz with respect to t and satisfy the Particle W-AIG Flow:*

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = V_t, \qquad \frac{\mathrm{d}}{\mathrm{d}t} V_t = -\alpha_t V_t - \nabla\left(\frac{\delta E}{\delta \rho_t}\right)(X_t),$$

for a.e. $t \in (0, T)$, with $V_0 = 0$.

Recognize that the Lipschitz condition is required both at the limit velocity field as well as the path $t \mapsto \rho_t$. More importantly, the convergence guarantee requires the curl-free differentiation map D_{k+1} which holds trivially for IA-JKO while does not hold in general for A-JKO due to the folding from extrapolation. The performance gap between the IA-JKO and A-JKO emprically validates this justification. The detailed analysis of the folding is left for further discussion at Section D.

Now, we provide the proof for the main theorem of convergence:

Proof. For any k, the paired map $(\iota, M_k)_\sharp \rho_k$ is a coupling between the measures (ρ_k, ρ_{k+1}). Since W_2 is of the minimum of all coupling, the Wasserstein metric gives an upper bound

$$W_2^2(\rho_k, \rho_{k+1}) \le \int |M_k(s) - x|^2 d\rho_k(x) = \|M_k - \iota\|_{\rho_k}^2.$$

Leveraging the identity from the definition of w_k and the change of variables isometry of measure-preserving map, we obtain $M_k - \iota = \tau_k w_k$ and $\|w_k\|_{\rho_k} = \|v_{k+1}\|_{\rho_{k+1}}$. This gives the upper bound on the Wasserstein metric as

$$W_2(\rho_k, \rho_{k+1}) \le \tau_k \|v_{k+1}\|_{\rho_{k+1}} \le C_0 \tau_k$$

using the natural supremum bound of $\sup_{k\le K} \|v_k\|_{\rho_k} \le C_0$. By the telescoping triangle inequality, W_2 is a metric on $\mathcal{P}_2(\mathbb{R}^d)$. For $j < k$, it is possible to chain the one-step bounds as

$$W_2(\rho_j, \rho_k) \le \sum_{i=j}^{k-1} W_2(\rho_i, \rho_{i+1}) \le \sum_{i=j}^{k-1} C_0 \tau_i = C_0(t_k - t_j) \le C_0(t_j - t_k)$$

using the definition of grid as $\sum_{i=j}^{k-1} \tau_i = t_k - t_j$. Furthermore, we have the representation of the Wasserstein distance between ρ and dirac delta function δ_0 as

$$W_2^2(\rho, \delta_0) = \int |x - 0|^2 d\rho(x) = \mathcal{M}_2(\rho), \quad \text{i.e. } W_2(\rho, \delta_0) = \mathcal{M}_2(\rho)^{1/2}$$

where $\mathcal{M}_2(\mu) = \int |x|^2 d\mu(x)$ denote the second order moment. This identifies $\mathcal{M}_2(\cdot)^{1/2}$ as the W_2 distance to a fixed point of the metric space which the upper bound requires a triangle-inequality:

$$M_2(\rho_k)^{1/2} = W_2(\rho_k, \delta_0) \le W_2(\rho_k, \rho_0) + W_2(\rho_0, \delta_0) \le C_0\, t_k + M_2(\rho_0)^{1/2} \le M_2(\rho_0)^{1/2} + C_0 T.$$

Therefore, the curves $t \mapsto \bar{\rho}_{\bar{\tau}}(t)$ are C_0-Lipschitz into $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. Let $\Psi_{k+1} = M_k \circ \Psi_k : \mathbb{R}^d \to \mathbb{R}^d$, $\Psi_0 = \iota$ denote the discrete flow map of the particle. Under Assumption 5, the particle positions obey the pointwise bound $|\Psi_k(x)| \leq \exp(2Lt_k)(|x| + Ct_k)$, since $M_k^{-1} = \iota - \tau_k v_{k+1}$ which gives

$$|\Psi_{k+1}(z)| \leq |\Psi_k(z)| + \tau_k(A + L|\Psi_{k+1}(z)|),$$

which gives

$$(1 - \tau_k L)|\Psi_{k+1}(z)| \leq |\Psi_k(z)| + \tau_k A.$$

Furthermore, $(1-x)^{-1} \leq 1 + 2x$ on the interval $[0, 1/2]$ gives

$$|\Psi_{k+1}(z)| \leq (1 + 2\tau_k L)(|\Psi_k(z)| + \tau_k A).$$

We are also equipped with the Lipschitz bound

$$\mathrm{Lip}(M_k) \leq (1 - \tau_k L)^{-1} \leq e^{2\tau_k L} \quad \text{and} \quad \mathrm{Lip}(M_k^{-1}) \leq 1 + \tau_k L \leq e^{\tau_k L}.$$

This gives the rate

$$|\hat{X}_{\bar{\tau}}(z,t)| \leq e^{2LT}(|z| + CT).$$

Therefore, the family of measures $\{\hat{\rho}_{\bar{\tau}}(t) : \bar{\tau}, t \in [0,T]\}$ has uniformly integrable second moments, i.e.

$$\int_{\{|x|>R\}} |x|^2 \,\mathrm{d}\hat{\rho}_{\bar{\tau}}(t) = \int_{\{|\hat{X}_{\bar{\tau}}(z,t)|>R\}} |\hat{X}_{\bar{\tau}}(z,t)|^2 \,\mathrm{d}\rho_0(z) \leq e^{4LT} \int_{\{|z|>R'\}} (|z| + AT)^2 \,\mathrm{d}\rho_0(z).$$

The right hand side converges to 0 with $R \to \infty$ uniformly in $(\bar{\tau}, t)$ since $\rho_0 \in \mathcal{P}_2$. Hence, the set is compact in $(\mathcal{P}_2, W_2)$ (Ambrosio et al., 2008, Proposition 7.1.5). Therefore, the convergence is with the contribution by $\hat{v}_{\bar{\tau}}$ small uniformly in n.

Therefore, the second moments are uniformly integrable with t-pointwise compactness gives W_2 compactness of the respective space of measures (Villani et al., 2009, Theorem 6.9). The refined Arzela-Ascoli theorem (Ambrosio et al., 2008, Proposition 3.3.1) yields the convergent subsequent satisfying

$$\sup_{t\in[0,T]} W_2\left(\bar{\rho}_{\bar{\tau}}(t), \rho_t\right) \longrightarrow 0, \qquad t \mapsto \rho_t \;\; C_0\text{-Lipschitz}, \qquad \rho|_{t=0} = \rho_0.$$

Let $V_k := v_k \circ \Psi_k \in L^2(\rho_0; \mathbb{R}^d)$. Since $V_{k+1} - V_k = (v_{k+1} \circ M_k - v_k) \circ \Psi_k$ and $(\Psi_k)_\sharp \rho_0 = \rho_k$. Therefore,

$$\|V_{k+1} - V_k\|_{\rho_0} = \|v_{k+1} \circ M_k - v_k\|_{\rho_k} = e_k \leq C_a \tau_k$$

from Corollary 2. This gives

$$\|V_{k+1} - v_k\|_{L^2(\rho_0)} = C_a(t_k - t_j) \qquad \text{for } j \leq k.$$

Also, for $y \in B_r(z)$, the Lipschitz bound can be established such that

$$|g(z)| \leq |g(y)| + L'r.$$

Averaging this over the ball $B_r(z)$ against the initial density ρ_0 and applying the Cauchy-Schwarz, we have

$$|g(z)| \leq \rho_0(B_r(z))^{-1} \int_{B_r(z)} |g| \,\mathrm{d}\rho_0 + L'r \leq \rho_0(B_r(z))^{-1/2} \|g\|_{L^2(\rho_0)} + L'r.$$

Defining $m(R,r) := \inf_{z \in \bar{B}_R(0)} \rho_0(B_r(z))$, from positive m, we have continuous $z \mapsto \rho_0(B_r(z))$ by the dominated convergence theorem. Hence, it is strictly positive which gives lower bound on the compact set $\bar{B}_r$ such that

$$|g(z)| \leq m(R,r)^{-1/2} \|g\|_{L^2(\rho_0)} + L'r$$

for L'-Lipschitz $g : \mathbb{R}^d \to \mathbb{R}^d$ on $z \in \bar{B}_R(0)$. By Lemma 2 and Arzela-Ascoli, there exists $\bar{\tau}_n \to 0$ with $\hat{v}_{\bar{\tau}_n} \to v$ locally uniformly on $\mathbb{R}^d \times [0,T]$. The limit naturally inherits the L-Lipschitz-ness and linear growth by $C + L|x|$ and zero initial velocity and modulus ω_R in t. Since $|\bar{v}_{\bar{\tau}} - \hat{v}_{\bar{\tau}}| \leq \omega_R(\bar{\tau})$ on the product space $\bar{B}_R \times [0,T]$, we have the map $\bar{v}_{\bar{\tau}_n}$ locally uniformly convergent to v.

For the kinetic energy, split $\int |\hat{v}_{\bar{\tau}}|^2 d\hat{\rho}_{\bar{\tau}}(t)$ into B_R and its complement for each t. On B_R, uniform convergence of $\hat{v}_{\bar{\tau}_n}$ and narrow convergence of $\hat{\rho}_{\bar{\tau}_n}(t)$ gives convergence. Outside B_R, we have the upper bound by

$$|\hat{v}_{\bar{\tau}_n}|^2 \leq 2A^2 + 2L^2|x|^2$$

The contribution is small uniformly in n by W_2-compact measures.

Fix the continuous function $\varphi \in C_c^\infty(\mathbb{R}^d \times [0,T))$ and let it telescope using the forward map $(M_k)_\sharp \rho_k = \rho_{k+1}$:

$$0 = \int \varphi(\cdot,0) d\rho_0 + \sum_k \int \left[\varphi(M_k x, t_{k+1}) - \varphi(x,t_k)\right] d\rho_k(x).$$

which the Tayler expansion with the increment $M_k x - x = \tau_k w_k(x)$ with the modulus $\leq C_R \tau_k$ on the support region by previous statement. This turns the sum over k into a Riemann sum of $\int [\partial_t \varphi + \langle \nabla \varphi, v\rangle] d\rho_t$ with $O(\bar{\tau})$ total error which gives the B_R locally-uniform convergence to the weak solution in the sense of Definition 4.

Fix $\xi \in C_c^\infty(\mathbb{R}^d \times [0,T); \mathbb{R}^d)$ and telescope the momentum. Define $B_k := \int \langle \xi(x,t_k), v_k(x)\rangle d\rho_k(x)$ for $k = 0, \ldots, K$. Since $v_0 = 0$ and $\xi(\cdot, T) = 0$, the endpoints vanish, $B_0 = 0 = B_K$, so telescoping gives

$$0 = B_K - B_0 = \sum_k \big(B_{k+1} - B_k\big).$$

Using this, we can bound the increments such that it splits into the Equation (16) as

$$\begin{aligned}
0 = &\sum_k \int \langle \xi(M_k x, t_{k+1}) - \xi(x,t_k), w_k(x)\rangle \,\mathrm{d}\rho_k && \text{(transport)} \\
&- \sum_k \alpha_k \tau_k \beta_k \int \langle \xi(\cdot,t_k), v_k\rangle \,\mathrm{d}\rho_k && \text{(friction)} \\
&- \sum_k \beta_k \tau_k \int \langle \xi(\cdot,t_k), \nabla W_{\rho_{k+1}}\rangle \,\mathrm{d}\rho_{k+1} && \text{(force)} \\
&- \sum_k \beta_k \tau_k \int \langle \xi \circ M_k^{-1} - \xi, \nabla W_{\rho_{k+1}}\rangle \,\mathrm{d}\rho_{k+1} + \sum_k \frac{1}{\tau_k} \int \langle \xi, r_k\rangle \,\mathrm{d}\rho_k. && \text{(transfer error, residual)}
\end{aligned}$$

The derivations are available by each increment compares a ρ_{k+1}-integral to a ρ_k-integral, and $(M_k)_\# \rho_k = \rho_{k+1}$ together with Proposition 8, $w_k = v_{k+1} \circ M_k$ lets us rewrite B_{k+1} entirely as an integral against ρ_k by the change of variables $y = M_k x$:

$$B_{k+1} = \int \langle \xi(\cdot, t_{k+1}), v_{k+1}\rangle \,\mathrm{d}\rho_{k+1} = \int \langle \xi(M_k x, t_{k+1}), w_k(x)\rangle \,\mathrm{d}\rho_k(x).$$

Subsequently, adding and subtracting $\int \langle \xi(x,t_k), w_k(x)\rangle \,\mathrm{d}\rho_k$ isolates a pure transport term that actually carries the dynamics:

$$B_{k+1} - B_k = \int \langle \xi(M_k x, t_{k+1}) - \xi(x,t_k),\, w_k\rangle \,\mathrm{d}\rho_k \;+\; \int \langle \xi(x,t_k),\, w_k - v_k\rangle \,\mathrm{d}\rho_k(x).$$

Summed over k, the first piece is already the first line (transport) of the target display. The equation (16) enters through the second term by substituting $w_k(x) = \beta_k\big(v_k(x) - \tau_k \nabla W_{\rho_{k+1}}(M_k x)\big) + r_k(x)/\tau_k$ and simplifying $\beta_k - 1 = -\alpha_k \tau_k \beta_k$ gives

$$w_k(x) - v_k(x) = -\alpha_k \tau_k \beta_k\, v_k(x) \;-\; \beta_k \tau_k\, \nabla W_{\rho_{k+1}}(M_k x) \;+\; \frac{r_k(x)}{\tau_k}.$$

Pairing this with $\xi(x,t_k)$ and integrating against ρ_k hands us the friction term (second line) and the residual term (last summand of the fourth line) immediately, plus an intermediate force term $-\beta_k \tau_k \int \langle \xi(\cdot,t_k), \nabla W_{\rho_{k+1}} \circ M_k\rangle \,\mathrm{d}\rho_k$, which still mismatches its two evaluation points (ξ at x, the force at $M_k x$). One more pushforward, $y = M_k x$, converts it to an integral against ρ_{k+1} carrying $\xi \circ M_k^{-1}$:

$$\int \langle \xi(\cdot, t_k), \nabla W_{\rho_{k+1}} \circ M_k \rangle \, \mathrm{d}\rho_k = \int \langle (\xi \circ M_k^{-1})(\cdot, t_k), \nabla W_{\rho_{k+1}} \rangle \, \mathrm{d}\rho_{k+1}.$$

Writing $\xi \circ M_k^{-1} = \xi + (\xi \circ M_k^{-1} - \xi)$ splits this into the clean force term (third line) plus the transfer-error term (first summand of the fourth line). Collecting every piece and summing over k reproduces the stated identity term by term.

By Taylor expansion of ξ, we obtain the expression on the

$$\begin{aligned}\xi(M_k x, t_{k+1}) &= \xi\big(x + \tau_k w_k(x),\ t_k + \tau_k\big) \\ &= \xi(x, t_k) + \tau_k \left(D_x \xi(x, t_k)\right) w_k(x) + \tau_k \, \partial_t \xi(x, t_k) + O(\tau_k^2).\end{aligned}$$

Rearranging gives the difference between the test vector field ξ in transport term can be expanded by Taylor approximation as

$$\xi(M_k x, t_{k+1}) - \xi(x, t_k) = \tau_k \, \partial_t \xi(x, t_k) + \tau_k \left(D_x \xi(x, t_k)\right) w_k(x) + O(\tau_k^2).$$

Then, the quadratic factor $\langle (D_x \xi) w_k, w_k \rangle$ is bounded pointwise on the support region and converges to $\langle (D_x \xi) v_k, v_k \rangle$ by the B_R locally-uniform convergence to the weak solution above. The friction term converges to $-\int_0^T \alpha_t \int \langle \xi, v \rangle d\rho_t dt$. The integrand in the force term can be formulated as

$$\int \langle \xi, \nabla W_{\rho_{k+1}} \rangle d\rho_{k+1} = \int \langle \xi, \nabla V \rangle d\rho_{k+1} - \int \nabla \cdot \xi d\rho_{k+1},$$

a fixed integrand over a compact support with convergent Riemann sum. The last transfer error term is bounded per step from the Lipschitz criteria from Corollary 2,

$$\mathrm{Lip}(\xi)\tau_k \|v_{k+1}\|_{\rho_{k+1}} \cdot \beta_k \tau_k \|\nabla W_{\rho_{k+1}}\|_{L^2(\rho_{k+1})} \leq \mathrm{Lip}(\xi) C_0 \tau_k (2e_k + C\tau_k)$$

which gives the time-grid bound

$$\sum_k \tau_k e_k \leq \Big(\sum_k \tau_k^2\Big)^{1/2} \Big(\sum_k e_k^2\Big)^{1/2} \leq C_1 (\bar{\tau} T)^{1/2} \longrightarrow 0,$$

while the residual sum is upper bounded by $\|\xi\|_\infty \varepsilon(\bar{\tau}) T \to 0$ by the Assumption 7. Collecting the limits gives the weak solution in 4 sense for the momentum. Furthermore, the boundary term vanishes with $v_0 = 0$ and $\xi(\cdot, T) = 0$. It completes the proof for the part Theorem 2.1 and 2.2 of the theorem by establishing the limit. Now, it remains to identify the limits as the W-AIG Flow.

Assume the momentum can be represented in a gradient form (of a convex potential ϕ with its convex conjugate is ϕ^*) in a form of $D_{k+1} = \iota - M_k^{-1}$ satisfy

$$D_{k+1} = \nabla(q - \phi_k^*) \quad \text{with} \quad q(y) = \frac{1}{2}|y|^2 \qquad \text{therefore } v_{k+1} = \nabla \frac{q - \phi_k^*}{\tau_k}.$$

This holds true for IA-JKO while A-JKO holds true for $d = 1$, and is not a gradient map for $d \geq 2$. However, it is possible to quantify the rate of estimation error is of $O(\bar{\tau})$ which we defer the proof to later. Each $\bar{v}_{\bar{\tau}}(\cdot, t)$ is then a continuous gradient field and locally uniform limit of gradient fields is again a gradient field. Therefore, for every t, $v(\cdot, t) = \nabla \Phi(\cdot, t)$ for the line-integral potential,

$$\Phi(x, t) := \int_0^1 \langle v(sx, t), x \rangle ds, \quad \Phi(0, t) = 0, \quad \Phi(\cdot, 0) \equiv 0.$$

We have $\Phi(\cdot, t) \in C^{1,1}_{\mathrm{loc}}$ with $\nabla \Phi$ as Lipschitz x-uniformly with linear growth, Φ is locally Lipschitz in t such that $\partial_t \Phi$ exists a.e. and $\nabla \partial_t \Phi = \partial_t \nabla \Phi$ is interchangeable. Combined with the above proof, this composes the first half of the statement at Theorem 3.

Since v is Lipschitz in x and continuous, the flow can be represented as $\dot{X}_t = v_t(X_t), X_0 = \iota$, well-posed and Lipschitz with double-sided Grönwall-type bound

$$e^{-Lt}|x - y| \leq |X_t(x) - X_t(y)| \leq e^{Lt}|x - y|, \qquad e^{-dLt} \leq \det \nabla X_t \leq e^{dLt} \quad \text{a.e.}$$

The continuity equation with a Lipschitz velocity admits a unique solution $\rho_t = (X_t)_\sharp \rho_0$ (Ambrosio et al., 2008, Section 8.1). Hence, $\rho_t \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ and $\rho_t > 0$ a.e. (positivity clause) gives

$$\log \rho_t(X_t(x)) = \log \rho_0(x) - \int_0^t \nabla \cdot v(X_s(x), s) ds \quad \text{for a.e. } x,$$

such that $|\log \rho_t| \le |\log \rho_0| \circ X_t^{-1} + dLT \in L^1_{\text{loc}}$ by two-sided Lipschitz criterion above. Moreover, $\rho_t \in W^{1,1}_{\text{loc}}$ with $\nabla \log \rho_t \in L^2(\rho_t; \mathbb{R}^d)$, and by Assumption 5 and previously established momentum bound we obtain

$$\|\nabla \log \rho_{k+1}\|_{L^2(\rho_{k+1})} \le \|\nabla W_{\rho_{k+1}}\|_{L^2(\rho_{k+1})} + \|\nabla V\|_{L^2(\rho_{k+1})} \le C$$

uniformly in $k, \bar{\tau}$, with the relative Fisher information $\rho \mapsto \int |\nabla \log \rho + \nabla V|^2 d\rho$ is lower semi-continuous for narrow convergence, so the bound holds for ρ_t.

For the test field $\xi \in C^\infty_c(\mathbb{R}^d \times [0,T); \mathbb{R}^d)$ the scalar $\varphi := \langle \xi, \nabla \Phi \rangle$ is Lipschitz, compact supported, with $\varphi(\cdot, 0) = 0$; this is admissible in the velocity form continuity equation which gives

$$0 = \int_0^T \int \left[\langle \partial_t \xi, \nabla \Phi \rangle + \langle (D_x \xi) \nabla \Phi, \nabla \Phi \rangle \right] \mathrm{d}\rho_t \, \mathrm{d}t + \int_0^T \int \left[\langle \xi, \partial_t \nabla \Phi \rangle + \langle \xi, (\nabla^2 \Phi) \nabla \Phi \rangle \right] \mathrm{d}\rho_t \, \mathrm{d}t,$$

where $\nabla^2 \Phi \in L^\infty$ a.e. is symmetric (hessian of the potential). Subtracting this equation from the momentum (4) with $v = \nabla \Phi$ leaves

$$\int_0^T \int \langle \xi, \underbrace{\partial_t \nabla \Phi + (\nabla^2 \Phi) \nabla \Phi + \alpha_t \nabla \Phi + \nabla V}_{=: F \in L^\infty_{\text{loc}}} \rangle \, \mathrm{d}\rho_t \, \mathrm{d}t = \int_0^T \int \nabla \cdot \xi \, \mathrm{d}\rho_t \, \mathrm{d}t = - \int_0^T \int \langle \xi, \nabla \rho_t \rangle \, \mathrm{d}x \, \mathrm{d}t,$$

where the last equality is derived from the definition of weak gradient ($\rho_t \in W^{1,1}_{\text{loc}}$) from above. Since ξ was set arbitrary test function, $\rho_t F = -\nabla \rho_t$ a.e. on $\mathbb{R}^d \times (0,T)$. This is further simplified on the measure set $\{\rho_t > 0\}$ as $F = -\nabla \log \rho_t$ a.e. Now, setting

$$b := \partial_t \Phi + \alpha_t \Phi + \frac{1}{2} |\nabla \Phi|^2 + V + \log \rho_t \in L^1_{\text{loc}}(\mathbb{R}^d \times (0,T)).$$

We claim that $\nabla_x b = 0$ almost everywhere on $\mathbb{R}^d \times (0,T)$. While other terms are guaranteed on their differentiability, it has not been shown for $\log \rho_t$. Therefore, we replace it with $\log \max(\rho_t, \varepsilon)$ for $\varepsilon \in (0,1]$. $\rho_t \in W^{1,1}_{\text{loc}}$ and $s \mapsto \log \max(s, \varepsilon)$ is Lipschitz on $(0, \infty)$, the chain rule for Sobolev function gives $\log \max(\rho_t, \varepsilon) \in W^{1,1}_{\text{loc}}$ with natural gradient $\mathbf{1}_{\{\rho_t > \varepsilon\}} \nabla \rho_t / \rho_t$ which is equivalent to $-F \mathbf{1}_{\{\rho_t > \varepsilon\}}$ by the identity $\rho_t F = -\nabla \rho_t$. Leveraging the interchangeability $\nabla \partial_t \Phi = \partial_t \nabla \Phi$ and the natural identity $\frac{1}{2} \nabla |\nabla \Phi|^2 = (\nabla^2 \Phi) \nabla \Phi$ a.e. (from Wang & Li (2022, Appendix B.3)), we obtain the convergence of

$$\nabla_x b_\varepsilon = F - F \mathbf{1}_{\{\rho_t > \varepsilon\}} = F \mathbf{1}_{\{\rho_t \le \varepsilon\}} \longrightarrow 0 \quad \text{in } L^1_{\text{loc}}$$

with $\varepsilon \to 0$ by dominated convergence theorem (Rudin, 1974). Therefore, the gradient converge with $\nabla_x b = 0$. Hence, we have

$$b(x,t) = c(t) \quad \text{a.e., and } c \in L^\infty(0,T).$$

This implies averaging $b(\cdot, t)$ over the unit ball, all other terms are bounded on $B_1 \times [0,T]$, while $\int_{B_1} |\log \rho_t| \, \mathrm{d}x \le e^{dLT} \int_{B_{R'}} |\log \rho_0| \, \mathrm{d}x + dLT |B_1|$ uniformly in t, by flow transport B.6 with $y = X_t(x)$ with Jacobian and and double-sided bound. Letting $g(t) = \int_0^t e^{-\int_s^t \alpha_r dr} c(s) ds$, we have the solution of $g' + \alpha_t g = c$ with $g(0) = 0$. Since c is bounded, g is Lipschitz on $[0,T]$, so the initial condition remains meaningful which results in $[\Phi] := \Phi - g$ being meaningful and satisfying 5 a.e. with $\nabla [\Phi] = \nabla \Phi$ and $[\Phi](\cdot, 0) \equiv 0$. This becomes the second half.

Finally, $t \mapsto \Phi(X_t(x), t)$ is Lipschitz and the set of (y,t) where $\nabla \Phi$ fails to be differentiable is Lebesgue-null, which combined with Fubini,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} V_t &= \left[\partial_t \nabla \Phi + (\nabla^2 \Phi) \nabla \Phi \right] \Big|_{(X_t, t)} \\ &= \left[F - \alpha_t \nabla \Phi - \nabla V \right] \Big|_{(X_t, t)} \\ &= -\alpha_t V_t - (\nabla V + \nabla \log \rho_t)(X_t) \\ &= -\alpha_t V_t - \nabla \left(\frac{\delta E}{\delta \rho_t} \right) (X_t), \end{aligned}$$

ρ_0-a.e. x and t along the trajectory. This is the exact particle version of W-AIG flow at 4 with $\dot{X}_t = V_t$ and $V_0 = \nabla \Phi(\cdot, 0) = 0$. This completes the proof. □

B.7 Weakening Assumption 7

In this section, we discuss the potential alleviation of the assumption 7 leveraging milder assumption. Consider the tangent space for $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ absolutely continuous, as

$$\mathrm{Tan}_\rho := \overline{\{\nabla\varphi : \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\rho;\mathbb{R}^d)}.$$

We now start with milder assumption:

Assumption 8. *The Brenier potential ϕ_k of $M_k = T_{\rho_k \to \rho_{k+1}}$ admits a condition $cI \preceq \nabla^2\phi_k \preceq \Lambda I$ on $\mathbb{R}^d$ such that $\phi_k - \frac{c}{2}|\cdot|^2$ and $\frac{\Lambda}{2}|\cdot|^2 - \phi_k$ are convex with $0 < c \le \Lambda < \infty$.*

Lemma 3 (IA-JKO Gauge)**.** *Assume 1-5 and Assumption 8. Then, the residual defined as*

$$r_k := M_k - \iota - u_k + s_k\,\nabla W_{\rho_{k+1}} \circ M_k \in L^2(\rho_k;\mathbb{R}^d)$$

satisfy $r_k \in \mathrm{Tan}_\rho^\perp := \{w \in L^2(\rho;\mathbb{R}^d) : \nabla\cdot(\rho w) = 0\}$.

Proof. Fix a small perturbation function $h \in C_c^\infty(\mathbb{R}^d)$ and $|\varepsilon| < c/\|\nabla^2 h\|_\infty$. Then, $\phi_k + \varepsilon h$ is convex so $M_\varepsilon := M_k + \varepsilon\nabla h$ is a Brenier map from ρ_k to $\rho_\varepsilon := (M_\varepsilon)_\sharp\rho_k$ and $\log_{\rho_k}\rho_\varepsilon = M_\varepsilon - \iota$. Hence, the inertial action becomes quadratic in ε:

$$\mathcal{A}(\rho_k,\rho_\varepsilon;u_k) = \|M_k - \iota - u_k\|_{\rho_k}^2 + 2\varepsilon\,\langle M_k - \iota - u_k, \nabla h\rangle_{\rho_k} + \varepsilon^2\|\nabla h\|_{\rho_k}^2.$$

For the energy, we write $\rho_\varepsilon = (\iota + \varepsilon\xi)_\sharp\rho_{k+1}$ with $\xi := \nabla h \circ M_k^{-1}$. By 8, $M_k^{-1} = \nabla\phi_k^*$ is c^{-1}-Lipschitz and M_k is Λ-Lipschitz which guarantees that ξ is Lipschitz with compact support. The expansion in the proof of proposition 6 gives

$$E(\rho_\varepsilon) - E(\rho_{k+1}) = \varepsilon\int [\langle\nabla V,\xi\rangle - \nabla\cdot\xi]\,d\rho_{k+1} + O(\varepsilon^2).$$

Since $\rho_{k+1} \in W^{1,1}_{\mathrm{loc}}(\mathbb{R}^d)$ and $\nabla\log\rho_{k+1} \in L^2(\rho_{k+1};\mathbb{R}^d)$, integrating by parts give

$$-\int \nabla\cdot\xi\, d\rho_{k+1} = \int\langle\xi,\nabla\log\rho_{k+1}\rangle d\rho_{k+1}.$$

Subsequently, changing the variables to $y = M_k x$ turns the first-order term into $\varepsilon\langle\nabla W_{\rho_{k+1}}\circ M_k, \nabla h\rangle_{\rho_k}$. Minimality of ρ_{k+1} on both sides give $\langle r_k,\nabla h\rangle_{\rho_k} = 0$ for every $h \in C_c^\infty$ resulting in $r_k \perp \mathrm{Tan}_{\rho_k}$. Finally, $r_k \in L^2(\rho_k)$ because $M_k - \iota$, $u_k \in L^2(\rho_k)$ and $\|\nabla W_{\rho_{k+1}}\circ M_k\|_{\rho_k} = \|\nabla W_{\rho_{k+1}}\|_{\rho_{k+1}} < \infty$ by 5. □

Lemma 4 (Tangential Facts)**.** *Assume 1-5 and 8. Then:*

1. *$v_{k+1} = -\tau_k^{-1}\log_{\rho_{k+1}}\rho_k \in \mathrm{Tan}_{\rho_{k+1}}$ for every $k \ge 0$; consequently $u_k \in \mathrm{Tan}_{\rho_k}$ and $Q_{\rho_k}u_k = 0$.*
2. *$\nabla W_{\rho_{k+1}} \in \mathrm{Tan}_{\rho_{k+1}}$*

Proof. 1) M_k is a Brenier map from ρ_k and both measures are absolutely continuous with the inverse as the reverse direction Brenier map $M_k^{-1} = T_{\rho_{k+1}\to\rho_k}$ holding ρ_{k+1}-a.e. (Santambrogio, 2015, Theorem 1.22). Therefore

$$v_{k+1} = \tau_k^{-1}(\iota - M_k^{-1}) = -\tau_k^{-1}\log_{\rho_{k+1}}\rho_k$$

which lies in $\mathrm{Tan}_{\rho_{k+1}}$ since $\log_\rho\nu = T_{\rho\to\nu} - \iota \in \mathrm{Tan}_\rho$ for all $\rho,\nu \in \mathcal{P}_2(\mathbb{R}^d)$ with ρ being absolutely continuous. $v_0 = 0$ by the 3. Since Tan_{ρ_k} is a linear subspace, we conclude $u_k = \beta_k\tau_k v_k \in \mathrm{Tan}_{\rho_k}$.

2) Any locally Lipschitz g with quadratic growth $|g(x)| \le C_g(1+|x|^2)$ and $\nabla g \in L^2(\rho)$ satisfies $\nabla g \in \mathrm{Tan}_\rho$, since C_c^∞ gradients are $L^2(\rho)$-dense among gradients of such functions (Ambrosio et al., 2008, Lemma 8.4.2). Subsequently, $\rho_{k+1} \in W^{1,1}_{\mathrm{loc}}(\mathbb{R}^d)$ and $\nabla\log\rho_{k+1} \in L^2(\rho_{k+1};\mathbb{R}^d)$ gives the Wasserstein subdifferential of Ent at ρ_{k+1} is nonempty and by Ambrosio et al. (2008, Theorem 10.4.6) the minimal-norm sub-differential with minimal selection $\nabla\log\rho_{k+1}$ and lies in $\mathrm{Tan}_{\rho_{k+1}}$. Summing the two, we obtain the desired result. □

Proposition 9. *Assume 1-5 and 8. Then, for the exact IA-JKO chain, the residual is* $r_k = s_k Q_{\rho_k}(\nabla W_{\rho_{k+1}} \circ M_k)$*, and*

$$\|r_k\|_{\rho_k} \le \kappa_2 \beta_k \|\nabla W_{\rho_{k+1}}\|_{\rho_{k+1}} \tau_k^3 \le C(T)\tau_k^3$$

where

$$C(T) := \kappa_2 C\big(1 + M_2(\rho_0)^{1/2} + C_0 T\big).$$

Proof. By lemma 3, $r_k = Q_{\rho_k} r_k$. Applying Q_{ρ_k} to the definition of r_k, $\log_\rho \nu = T_{\rho\to\nu} - \iota \in \mathrm{Tan}_\rho$ for all $\rho, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with ρ being absolutely continuous which gives $Q_{\rho_k}(M_k - \iota) = 0$ and $Q_{\rho_k} u_k = 0$ by the first part of lemma 4. This leaves the residual identity.

The second statement of 4, $g := \nabla W_{\rho_{k+1}} \in \mathrm{Tan}_{\rho_{k+1}}$ which from the transport-of-tangency bound we obtain

$$\|r_k\|_{\rho_k} \le s_k \kappa_2 \tau_k \|\nabla W_{\rho_{k+1}}\|_{\rho_{k+1}},$$

with $s_k = \beta_k \tau_k^2$ giving the first inequality. Furthermore, 5 gives

$$\|\nabla W_{\rho_{k+1}}\|_{\rho_{k+1}} \le C\big(1 + M_2(\rho_{k+1})^{1/2}\big)$$

and

$$M_2(\rho_{k+1})^{1/2} \le M_2(\rho_0)^{1/2} + C_0 T$$

was shown in the proof of 2. $\zeta_k \equiv 0$ for IA-JKO gives the desired result with $\beta_k \le 1$. □

C Matching Loss Derivation

The matching loss necessitates a matching target to minimze the MSE loss

$$\mathcal{L}_{\text{matching}} = \mathbb{E}\left[\| v_\theta(x_t, t) - \xi \|_2^2\right].$$

Yet, in the data-free setting the target field ξ is something not defined in the flow-matching style format. In this section, we show the derivation of the target ξ that correspond to the optimal velocity field that emulates a single (Implicitly) Accelerated-JKO step.

Lemma 5 (first-order optimality for the JKO step)**.** *Let* $\mu_k \in \mathcal{P}_2(\mathbb{R}^d)$*,* $s_k > 0$*,* $V \in C^1$ *with* $\mathcal{F}(\rho) = \int V \,\mathrm{d}\rho + \int \rho \log \rho$ *proper, lower semi-continuous, coercive on* $\mathcal{P}_2$*, and let*

$$\rho_{k+1} \in \operatorname*{arg\,min}_{\rho \in \mathcal{P}_2} \left\{ \tfrac{1}{2} W_2^2(\mu_k, \rho) + s_k \mathcal{F}(\rho) \right\}.$$

Let φ *be a Kantorovich potential on the* ρ_{k+1} *side for the cost* $c(y,z) = \frac{1}{2}|y-z|^2$ *in the pair* (ρ_{k+1}, μ_k)*. Then there is* $C \in \mathbb{R}$ *with*

$$\varphi(y) + s_k\big(V(y) + \log \rho_{k+1}(y)\big) = C \quad \text{for a.e. } y \in \{\rho_{k+1} > 0\}. \tag{23}$$

Proof. c.f. Proposition 8.7. of Santambrogio (2015). □

We then consider its analogue which is the first-order optimality of the IA-JKO step:

Proposition 10. *Assume 1, 3 and 8 and* $u_k \in L^2(\rho_k, \mathbb{R}^d)$*. Let* ρ_{k+1} *be the minimizer of the IA-JKO optimality Equation (13). Then,* $\rho_{k+1} > 0$ *and* $\log \rho_{k+1} \in L^1(\rho_{k+1})$ *and for every* $h \in C_c^\infty(\mathbb{R}^d)$ *with* $\xi_h := \nabla h \circ M_k^{-1}$*, we have*

$$\int \langle \nabla V, \xi_h \rangle \,\mathrm{d}\rho_{k+1} - \int \nabla \cdot \xi_h \,\mathrm{d}\rho_{k+1} + \frac{1}{s_k} \langle M_k - \iota - u_k,\, \nabla h \rangle_{\rho_k} = 0.$$

Furthermore, since $\rho_{k+1} \in W^{1,1}_{\text{loc}}$*,* $\nabla \log \rho_{k+1} \in L^2(\rho_{k+1})$*, there is a function* Ψ_k *with* $\nabla \Psi_k \in \mathrm{Tan}_{\rho_{k+1}}$*, determined by* u_k *and* M_k *alone and a constant* C *such that*

$$\log \rho_{k+1} + V + \frac{\varphi^{\mathrm{IA}} - \Psi_k}{s_k} = C \quad \text{a.e.}, \tag{24}$$

$$v_{k+1} := \frac{\iota - M_k^{-1}}{\tau_k} = \frac{\nabla \varphi^{\mathrm{IA}}}{\tau_k} = \frac{\nabla \Psi_k}{\tau_k} - \beta_k \tau_k \nabla W_{\rho_{k+1}} \quad \text{a.e.} \tag{25}$$

For $u_k = 0$ *one has* $\Psi_k = 0$*. For* $\tau = s_k$*, 24 and 25 gives equivalent form of 23*

We defer the proof to the last part of this section.

The only difference arises at the auxiliary function Ψ_k which is the potential of the tangential component. Differentiating Equation (23), we have the natural condition

$$\nabla\varphi(y) = \nabla\Psi_k - s_k \nabla W_{\rho_{k+1}}(y).$$

For the optimal coupling between ρ_{k+1} and μ_k, combined with the Kantorovich-duality relation $u = \frac{1}{2}|\cdot|^2 - \varphi$ for the c-concave potential, Brenier's theorem (Brenier, 1991) states the existence and uniqueness with the point in mu_k that pairs with $y \in \rho_{k+1}$ is given by

$$S_k(y) = y - \nabla\varphi(y).$$

Combining the two equations and rearranging with $S_k(y) = x + u_k(x)$, we obtain the proximal step

$$y = x + u_k(x) + \nabla\Psi - s_k \left[\underbrace{\nabla V(y) + \nabla \log \rho_{k+1}(y)}_{\nabla W_{\rho_{k+1}}(y)} \right].$$

The proximal step is evaluated at the destination y instead of the current state x which is an implicit backward-Euler step for the gradient flow $\dot{y} = -\nabla(V + \log\rho)(y)$. In contrast, the explicit step evaluates with respect to the current state x as

$$y = x + u_k(x) + \nabla\Psi - s_k \nabla W_{\rho_k}.$$

From proposition 10, we perturb ϕ_k by a small ε such that $\phi_k + \varepsilon h$ stays convex. The action is then quadratic in ε. Setting the derivative to zero gives

$$\langle M_k - \iota - u_k + s_k \nabla W_{\rho_{k+1}} \circ M_k,\, \nabla h \rangle_{\rho_k} = 0 \qquad \forall h \in C_c^\infty.$$

We have $M_k - \iota = \log_{\rho_k} \rho_{k+1}$ and u_k concurrently lie in Tan_{ρ_k}, which gives

$$M_k - \iota = P_{\rho_k}[u_k - s_k \nabla W_{\rho_{k+1}} \circ M_k], \qquad r_k = s_k Q_{\rho_k}[\nabla W_{\rho_{k+1}} \circ M_k],$$

where P_ρ is the orthogonal projection of $L^2(\rho; \mathbb{R}^d)$.

This gives the target velocity field:

$$M_k - \iota = \arg\min_{v \in \mathrm{Tan}_{\rho_k}} \mathbb{E}_{x\sim\rho_k} \left[\left| v_\theta(x) - \underbrace{u_k(x) - s_k \nabla W_{\rho_{k+1}}(y)}_{\xi_k(x, M_k(x))} \right|^2 \right].$$

The stability of explicit Euler is bounded since the update matrix of the state is given by $I - s_k\lambda$. Representing in the residual form

$$\mathcal{R}(y) = y + s_k F(y) - (x + u_k(x)),$$

we need $\mathcal{R}(y) = 0$. To solve the implicit equation, we may consider using the Picard Iteration to solve for the y under relaxed η as

$$y \leftarrow \frac{\eta y_{\text{old}} + [x + u - s_k F(y_{\text{old}})]}{1 + \eta}, \qquad F = \nabla V + \nabla \log \rho_{k+1}.$$

Proof for Proposition 10. By the change of variables

$$\rho_{k+1}(M_k x) \det \nabla^2 \phi_k(x) = \rho_k(x),$$

the bi-Lipschitz map M_k preserves null sets, so $\rho_{k+1} > 0$ a.e. because $\rho_k > 0$ a.e. by 3, and $\log \rho_{k+1} \circ M_k = \log \rho_k - \log\det \nabla^2\phi_k$ with $\log\det\nabla^2\phi_k \in L^\infty$ by 8, so $\log\rho_{k+1} \in L^1(\rho_{k+1})$ because ρ_k has finite entropy. Lemma 8.6's competitor argument is not needed.

Fix $h \in C_c^\infty(\mathbb{R}^d)$ and a fix a small number $|\varepsilon| < c$. Then, $\phi_k + \varepsilon h$ is uniformly convex which

$$M_\varepsilon := M_k + \varepsilon \nabla h$$

is the optimal map from ρ_k to $\rho_\varepsilon := (M_\varepsilon)_\sharp \rho_k$. Since $M_\varepsilon = (\iota + \varepsilon\xi_h) \circ M_k$, also $\rho_\varepsilon = (\iota + \varepsilon\xi_h)_\sharp \rho_{k+1}$ with ξ_h being a Lipschitz with compact support and $\iota + \varepsilon\xi_h$ is injective.

Along this curve, the action is quadratic such that

$$\mathcal{A}(\rho_k, \rho_\varepsilon; u_k) = \|M_k - \iota - u_k + \varepsilon \nabla h\|^2_{\rho_k}$$

so its derivative at $\varepsilon = 0$ is naturally $2\langle M_k - \iota - u_k, \nabla h\rangle_{\rho_k}$.

Regarding the energy, the change of formula gives

$$\mathrm{Ent}(\rho_\varepsilon) = \mathrm{Ent}(\rho_{k+1}) - \int \log\det(I + \varepsilon D\xi_h)\,\mathrm{d}\rho_{k+1},$$

and Jacobi's formula with dominated convergence theorem gives

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} \left[\int V\,\mathrm{d}\rho_\varepsilon + \mathrm{Ent}(\rho_\varepsilon)\right] = \int \langle \nabla V, \xi_h\rangle\,\mathrm{d}\rho_{k+1} - \int \nabla\cdot\xi_h\,\mathrm{d}\rho_{k+1}.$$

Since $\varepsilon = 0$ is an minimum within the interior, the derivatives sum to zero.

We have $\rho_{k+1} \in W^{1,1}_{loc}(\mathbb{R}^d)$ and $\nabla W_{\rho_{k+1}} = \nabla V + \nabla \log \rho_{k+1} \in L^2(\rho_{k+1})$ from 10. Then, integrating C, and applying push-forward by M_k to the last term, using $(M_k - \iota) \circ M_k^{-1} = \nabla\varphi^{IA}$, we obtain:

$$\int \langle b_k,\, \nabla h \circ M_k^{-1}\rangle\,\mathrm{d}\rho_{k+1} = 0 \qquad \forall h \in C^\infty_c(\mathbb{R}^d),$$

where $b_k := \nabla\varphi^{\mathrm{IA}} - u_k \circ M_k^{-1} + s_k\nabla(\log\rho_{k+1} + V) \in L^2(\rho_{k+1}; \mathbb{R}^d)$. We have isometry between $L^2(\rho_k) \to L^2(\rho_{k+1})$ by the map $f \mapsto f \circ M_k^{-1}$ which allows C extend from ∇h to all measures in Tan_{ρ_k}. Particularly, for all $\chi \in C^\infty_c(\mathbb{R}^d)$, it extends to $\nabla(\chi \circ M_k)$, and as $\nabla(\chi\circ M_k)\circ M_k^{-1} = A_k\nabla\chi$ and A_k is symmetric, we have $\langle b_k, \nabla\chi\rangle_{A_k} = 0$ for all χ.

By definition of w_k, we also have $u_k \circ M_k^{-1} - w_k \perp_{A_k} \mathrm{Tan}_{\rho_{k+1}}$, hence we have

$$g := b_k + u_k \circ M_k^{-1} - w_k = \nabla\varphi^{\mathrm{IA}} + s_k\nabla W_{\rho_{k+1}} - w_k \perp_{A_k} \mathrm{Tan}_{\rho_{k+1}}.$$

On the other hand, we have $g \in \mathrm{Tan}_{\rho_{k+1}}$: $\nabla\varphi^{IA} = \iota - M_k^{-1}$ is minus the displacement of the optimal map from ρ_{k+1} to ρ_k by Ambrosio et al. (2008, Section 8.5), $\nabla(\log\rho_{k+1} + V)$ is tangent by 4 and w_k by construction. Therefore, we have $\langle g, g\rangle_{A_k} = 0$ and $A_k \succeq cI$ gives $g = 0$ ρ_{k+1}-a.e. Hence, by positivety we have

$$\nabla\varphi^{IA} + s_k\nabla W_{\rho_{k+1}} = w_k \qquad \text{a.e.}$$

The left hand side is a gradient, which gives $w_k = \nabla\Psi_k$ for a function Ψ_k. Integrating on $\mathbb{R}^d$ and dividing by τ_k with $s_k/\tau_k = \beta_k\tau_k$ gives (24) and (25). For $u_k = 0$, $w_k = P^{A_k}_{\rho_{k+1}}[0] = 0$, completing the proof. □

D A-JKO FOLDING

Continuous Normalizing Flows (CNFs) transport a prior density to a target through an invertible map $\Phi_t : x(0) \mapsto x(t)$ generated by an ODE, so that the marginal density ρ_t is tracked by the instantaneous change of variables

$$\frac{d\log\rho_t(x(t))}{dt} = -\nabla\cdot v_\theta(x(t), t) = -\mathrm{tr}\Big(\frac{\partial v_\theta}{\partial x}(x(t), t)\Big).$$

A-JKO is instead a composition of discrete maps, and log-densities must be propagated through the Jacobian determinant of each factor.

Transport map. For the explicit (extrapolating) A-JKO scheme,

$$\Phi_K = T_K \circ E_K \circ \cdots \circ T_1 \circ E_1, \qquad x_0 = z + m, \quad z \sim \mathcal{N}(0, \sigma^2 I_d),$$

where T_k is the JKO resolvent and E_k is the extrapolation step

$$y_k = E_k(x_{k-1}) = x_{k-1} + \beta\,(x_{k-1} - x_{k-2}) = (1+\beta)\,x_{k-1} - \beta\,x_{k-2}, \qquad x_k = T_k(y_k),$$

with $x_{-1} := x_0$ (zero initial momentum, so $E_1 = \mathrm{id}$). Since E_k depends on two iterates, it defines a map of x_{k-1} alone through $x_{k-2} = G_{k-1}^{-1}(x_{k-1})$, where $G_{k-1} := T_{k-1} \circ E_{k-1}$ (with $G_0 := \mathrm{id}$). Writing $M_{k-1} := \nabla G_{k-1}$, we obtain

$$\nabla E_k = (1+\beta)I - \beta\, M_{k-1}^{-1}, \qquad \det \nabla E_k = \frac{\det\big((1+\beta)M_{k-1} - \beta I\big)}{\det M_{k-1}},$$

and the log-density is updated by the discrete rule

$$\log \rho_{y_k}(y_k) = \log \rho_{k-1}(x_{k-1}) - \log \big|\det \nabla E_k(x_{k-1})\big|,$$

rather than by the trace formula above. Each T_k is an optimal transport map (Brenier), so ∇T_k is symmetric positive definite and $\det \nabla T_k > 0$. The extrapolation E_k—a single forward-Euler step of the momentum field and the only non-resolvent operation in the scheme—is therefore the only factor whose Jacobian can degenerate. This happens exactly when the preceding map has contracted by the factor $\beta/(1+\beta)$ in some direction, i.e. $\beta/(1+\beta) \in \mathrm{spec}(M_{k-1})$.

Folding. A vanishing Jacobian alone does not imply non-injectivity: in the linear case, $\det \nabla \Phi_K < 0$ is merely a reflection, and the map remains a valid pushforward. Genuine folding, where the transport ceases to be injective, arises when $\det \nabla \Phi_K$ *changes sign across* $\mathrm{supp}\, \rho_0$. Then distinct particles are mapped to the same point, and a caustic forms in ρ_K.

Example 1. Quadratic model *Consider the potential energy $\mathcal{F}(\rho) = \int V\, d\rho$ with $V(x) = \frac{1}{2}(x-m)^\top A(x-m)$, where $A \succ 0$ has eigenvalues $\lambda_i \le L$. The resolvent is then the proximal map $T(y) = m + (I+\tau A)^{-1}(y-m)$, and all maps share the eigenbasis of A. Along the i-th eigendirection, $x_k - m = p_k\,(x_0 - m)$, where*

$$p_{k+1} = s_i\big((1+\beta)p_k - \beta p_{k-1}\big), \qquad s_i = \frac{1}{1+\tau\lambda_i}, \qquad p_{-1} = p_0 = 1.$$

Consider the characteristic roots of this recursion. If they are real, both lie in $(0,1)$ and $p_k > 0$ for all k. If they are complex, i.e. $s_i(1+\beta)^2 < 4\beta$, then p_k oscillates with modulus $(s_i\beta)^{k/2}$ and changes sign. The worst direction is $s_{\min} = 1/(1+\tau L)$, so sign changes are excluded whenever

$$\tau L \;\le\; \frac{(1-\beta)^2}{4\beta}.$$

This is the discrete analogue of critical damping: folding is the footprint of the underdamped regime of the accelerated flow. In this homogeneous setting, a sign change is a global reflection. For non-quadratic V, heterogeneous curvature makes different regions oscillate out of phase, the sign of $\det \nabla \Phi_K$ varies over the support, and genuine folding occurs. Folding can be avoided by choosing (τ, β) within the condition above, by adaptively restarting the momentum, or by treating the momentum implicitly (ID-JKO), which replaces E_k by a resolvent with positive definite Jacobian.

E Experiment Details

We provide the experimental details here.

Architecture. All proximal maps are continuous normalizing flows whose velocity field is an $S_n \times O(d)$-equivariant EGNN (3 layers, 32 radial basis functions); hidden/message widths are 16/12 on DW-4 and 32/16 on LJ-13 and LJ-55. The step size τ_k of block k follows either a constant schedule or a linear ramp $\tau_k = \tau_0 + (\tau_{\min} - \tau_0)\, k/(n_{\text{ramp}} - 1)$ for $k < n_{\text{ramp}}$ and $\tau_k = \tau_{\min}$ afterwards.

Remark 5 (Comparison with the flow sampling)**.** *Havens et al. (2026) uses significantly larger number of EGNN layers: 3 for DW-4 and 5 for LJ-13 and 55. Furthermore, the size of hidden layers differ such that it becomes 128 layers for all three densities in the Flow-Sampling. While we use stacked blocks, the number of parameters used in the training are smaller than FS at all times.*

DW-4. We use a constant step $\tau_k = 3$ for 10 blocks with Nesterov momentum $\beta_k = k/(k+3)$. Each block minimizes the simulated proximal objective with 3000 Adam steps (learning rate 10^{-3}). A-JKO extrapolates the particles along a learned equivariant momentum field before the proximal step, while IA-JKO absorbs the momentum into a single proximal step through the shifted kinetic term.

LJ-13. For A-JKO we set $\tau_0 = 1.0$ and decrease it linearly to $\tau_{\min} = 0.35$ over $n_{\text{ramp}} = 6$ blocks, with constant momentum $\beta = 0.5$; we train 10 blocks of 6000 Adam steps on the simulated proximal objective and report the final block. For IA-JKO we use the same $\tau_0 = 1.0 \to 0.35$ ramp over $n_{\text{ramp}} = 12$ blocks and $\beta = 0.5$, and train 10 blocks with the matching objective (30 exploration rounds, 6000 steps, 5000 exploration samples per block). The proximal target is solved by a Newton iteration, the entropy term uses a denoising-score-matching EGNN score network, and the step is capped by $s_k \le 2/q_{0.9}\big(-\lambda_{\min}(\nabla^2 V)\big)$. A 1000-step simulated-loss corrector (batch 1000) is applied after every block.

LJ-55. Both A-JKO and IA-JKO uses the same configuration with LJ-13 configuration ($\tau_0 = 1.0 \to 0.35$, $n_{\text{ramp}} = 12$, $\beta = 0.5$, 20 blocks, 6000 steps per block) with three changes for the 165-dimensional system: the curvature cap is taken at the median ($q_{0.5}$), the exploration batch is 2048, and the 1000-step corrector is applied only from block 11 onward.

JKO-IC. On DW-4 we run the public implementation with its default configuration (4 flow and 6 importance-corrected blocks, $\tau_k = 0.01 \cdot 4^k$, 5000 steps per block). On LJ-13 we use 12 flow and 8 importance-corrected blocks with $\tau_k = 0.02 \cdot 2^{3k/5}$ and 6000 steps per block following their exponentially increasing block and scaling the step size accordingly.

F Evaluation protocol

We follow the evaluation protocol and reference data of Flow Sampling (FS) (Havens et al., 2026); the baseline results in Tables 1 and 2 are taken from that work. Every metric compares 2000 model samples against 2000 reference samples.

Target densities. Each target is a Boltzmann density $\pi(x) \propto \exp(-E(x))$; the energy E can be evaluated pointwise, and the normalizing constant is unknown. For the synthetic systems, $x = (x_1, \dots, x_N) \in \mathbb{R}^{N\times n}$ collects the positions of N particles in $\mathbb{R}^n$, and we write $d_{ij} = \|x_i - x_j\|$ and $d = Nn$. DW-4 ($N = 4$, $n = 2$, $d = 8$) uses the pairwise double-well potential of Köhler et al. (2020),

$$E_{\text{DW}}(x) = \sum_{i<j} \left[b\,(d_{ij} - r_0)^2 + c\,(d_{ij} - r_0)^4 \right], \qquad b = -4,\ c = 0.9,\ r_0 = 4.$$

LJ-13 and LJ-55 ($n = 3$, $d = 39$ and $d = 165$) use the Lennard-Jones potential with a harmonic confinement

$$E_{\text{LJ}}(x) = \varepsilon \sum_{i\neq j} \left[\left(\frac{r_m}{d_{ij}}\right)^{12} - 2\left(\frac{r_m}{d_{ij}}\right)^{6} \right] + \frac{c}{2} \sum_{i=1}^{N} \|x_i - \bar{x}\|^2, \qquad \varepsilon = r_m = c = 1,$$

where $\bar{x}$ is the centre of mass. Both energies are invariant under translations, rotations, reflections and permutations of the particles, so all configurations are centred before comparison. Reference samples are the long-run MCMC samples of Akhound-Sadegh et al. (2024). Alanine dipeptide (Ala2) has 22 atoms in Cartesian coordinates ($d = 66$). Its energy is a classical force field with implicit solvent, evaluated with OpenMM (Eastman et al., 2017) using the settings of Nam et al. (2026), and the target is $\pi \propto \exp(-E/k_B T)$ at $T = 300$ K. Reference samples are the 10^7 replica-exchange MD samples of Midgley et al. (2023).

Metrics for the synthetic systems. W_2 denotes the geometric Wasserstein-2 distance, i.e. the Wasserstein-2 distance for the symmetry-reduced cost

$$d(x, y) = \min_{P\in S_N,\ R\in \mathrm{O}(n)} \big\| x - P\, y\, R^\top \big\|_F,$$

where x and y are centred. Following FS, the minimization is approximated sequentially: P is obtained by the Hungarian algorithm and then R by Kabsch alignment. The resulting pair (P, R) is feasible, so the computed cost is an upper bound on d. The transport problem between the two empirical measures is then solved exactly with POT (Flamary et al., 2021). $E(\cdot)W_2$ is the Wasserstein-2 distance between the one-dimensional distributions of $E(x)$ under the model and under

the reference. For equal sample sizes it equals $\left(M^{-1}\sum_i (e_{(i)} - e'_{(i)})^2\right)^{1/2}$, where $e_{(i)}$ and $e'_{(i)}$ are the sorted model and reference energies. The two metrics are complementary. W_2 measures the match in configuration space. $E(\cdot)W_2$ measures whether the sampler reproduces the energy distribution, including its concentration in low-energy regions. It is sensitive to errors near the repulsive core of the Lennard-Jones potential, which W_2 barely registers.

Metrics for alanine dipeptide. The backbone dihedral angles (ϕ, ψ) of each configuration are computed following the same procedure as (Havens et al., 2026). We report three quantities: (i) the KL divergence between the model and reference histograms of ϕ and of ψ separately; (ii) the Jensen–Shannon divergence between the joint (ϕ, ψ) histograms, with 16×16 bins on $[-\pi, \pi)^2$ (logarithms base 2, so JSD $\in [0, 1]$), defined as $\mathrm{JSD}(\hat{p} \,\|\, \hat{q}) = \frac{1}{2} D_{\mathrm{KL}}(\hat{p} \,\|\, m) + \frac{1}{2} D_{\mathrm{KL}}(\hat{q} \,\|\, m)$ with $m = \frac{1}{2}(\hat{p} + \hat{q})$—the bin count trades finite-sample bias against resolution: with N=2,000 evaluation samples the reference-vs-reference floor is 0.018 ± 0.002 at 16×16 but 0.62 at 200×200, where the metric is dominated by sampling noise, while method rankings remain unchanged across 8^2–200^2 bins; (iii) the energy $\mathcal{W}_2$, defined as $E(\cdot)W_2$ above with the classical force-field energy in reduced units of $k_{\mathrm{B}}T$ at $T = 300\,\mathrm{K}$.

Reporting. The MCMC (GT) row of Table 1 evaluates each metric between two disjoint random subsets of the reference samples. It is the finite-sample floor, i.e. the best value attainable with M samples. Synthetic results are the mean $\pm$ standard error over 5 runs, as in FS; Ala2 results are the mean $\pm$ standard deviation over 3 runs. All rows are computed on raw samples. The only exception is IA-JKO (w/ rej.), which is evaluated after the importance-based rejection step of Hertrich & Gruhlke (2025).