Policy Optimization as a Gradient Flow in Offline Reinforcement Learning

Anonymous Author
Anonymous Institution

Abstract

We present a geometric perspective on offline reinforcement learning (RL), viewing policy optimization as a gradient flow on probability metric spaces. In this view, the policy evolves continuously from the behavior policy along the steepest descent of the objective, with each policy update corresponding to a discrete Euler step, and conventional policy optimization can be interpreted as a one-step Euler discretization of the underlying gradient flow. Building on two geometries—the Fisher–Rao and Wasserstein-2 metrics—we propose new offline RL algorithms based on a single discrete Euler step. Empirically, our algorithms achieve performance comparable to or slightly exceeding deterministic counterparts. Leveraging the Euler-step interpretation, taking one additional policy update beyond the standard one-step improves performance, surpassing benchmarks. Our results highlight geometric gradient-flow formulations as a principled and effective unified framework, providing a clear path for advancing offline RL.

1 Introduction

Offline reinforcement learning (RL) enables policy learning from fixed datasets, eliminating the need for additional interactions with the environment. This is particularly valuable in domains where collecting new interaction data is costly or unsafe, such as healthcare, robotics, or autonomous driving. A central obstacle is *distributional shift*, where the learned policy selects actions outside the dataset support, resulting in unreliable value estimates and unstable policy improvement (Fujimoto et al., 2019). A large body of work addresses this problem by regularizing the learned policy toward dataset distribution.

Most prior methods address distributional shift via policy regularization. TD3+BC (Fujimoto and Gu, 2021) combines Q-value maximization with behavior cloning. IQL (Kostrikov et al., 2021) and ReBRAC (Tarasov et al., 2023) further improve empirical performance through advantage-weighted and conservative critic updates. Despite their empirical success, these approaches largely rely on heuristic regularization incorporated into the loss function. This raises a fundamental question:

Can we move beyond heuristic regularization to establish a theoretically principled loss function that drives the design of efficient RL algorithms?

In this paper, we introduce a geometric perspective on offline RL, casting policy optimization as a gradient flow on probability metric spaces. Unlike conventional heuristic loss minimization, our approach interprets a policy as the continuous evolution of a probability distribution, governed by an energy functional under a chosen geometry. Policy updates correspond to discrete Euler steps of this continuous variational dynamics, instantiated under either the Fisher-Rao metric from information geometry or the Wasserstein metric from optimal transport.

Building on this framework, we propose POGO, a novel geometric offline RL algorithm. Policy updates are formulated as the minimization of an energy functional that combines Q-function maximization with entropy regularization. In addition, the Euler discretization of the underlying gradient flow induces an implicit proximal regularization, which further stabilizes the updates. For efficient instantiation, policies are modeled as Gaussian distributions fitted via regression on offline data, enabling closed-form computations of distributional distances.

Leveraging the Euler step interpretation, we introduce

Preliminary work. Under review by AISTATS 2026. Do not distribute.

multi-step POGO, a multi-step policy update scheme that iteratively refines the policy by using each intermediate Euler update as the reference for the next. In our experiments, we implement a two-step update initialized from the behavior policy. POGO consistently matches or outperforms TD3+BC and IQL, and in several tasks attains performance comparable to or surpassing ReBRAC despite relying on a simpler TD3+BC backbone. Moreover, the two-step Euler update generally yields additional gains once the critic has stabilized, highlighting the effectiveness of the geometry-driven formulation. Investigating adaptive strategies for selecting the number of Euler steps remains an important direction for future research.

Our main contributions are summarized as:

- We cast offline RL as a gradient flow on probability measures, providing a unified perspective that establishes a theoretical foundation for existing heuristic regularization methods while generalizing them to a comprehensive framework.
- We propose POGO, whose policy updates integrate Q-function maximization, entropy regularization, and implicit proximal stabilization via Euler discretization. Leveraging the gradient flow perspective, we further extend it to multi-step POGO, yielding more refined policies.
- We implement the framework with Gaussian policies, enabling closed-form computations of distributional distances, and demonstrate its effectiveness through experiments across benchmark offline RL tasks.
- Despite its minimal TD3+BC backbone, POGO consistently matches or improves upon TD3+BC and IQL, and approaches the performance of ReBRAC on several tasks. These results indicate that gradient flow-based updates can provide competitive gains without relying on more complex architectures.

2 Related Work

Our work is related to the use of statistical manifolds in RL as a theoretical foundation, and to regularization techniques for mitigating distributional shift in offline RL as a technical objective.

Statistical Manifold in RL: Online policy optimization has been interpreted as a gradient flow in Wasserstein space (Zhang et al., 2018). Wasserstein Actor-Critic (Likmeta et al., 2023) uses Wasserstein distances in the critic to propagate Q-posterior uncertainty and guide exploration via optimistic Q-value bounds. Trust-region variants, including OT-TRPO (Terpin et al., 2022) and metric-aware TRPO (Song et al., 2023), redefine policy proximity under Wasserstein geometry, while Moskovitz et al. (2020) propose Wasserstein natural gradient methods for improved efficiency. W-BRAC (Wu et al., 2019) employs the Wasserstein-1 distance as a behavior regularizer, and Asadulaev et al. (2024) reinterpret it as a partial optimal transport objective that selectively matches high-value dataset portions.

The natural policy gradient (Kakade, 2001), which underlies TRPO (Schulman et al., 2017), optimizes policies on the Fisher–Rao manifold via the Fisher information matrix. Lascu et al. (2025) replace KL-based trust regions with the Fisher–Rao metric in PPO, and Kerimkulov et al. (2025) link Fisher–Rao gradient flows with entropy-regularized MDPs by leveraging the intrinsic geometry of the policy space.

These studies highlight that manifold-based metrics can enrich RL by incorporating geometric structure to policy updates. While most prior work focuses on *online* RL, where interactions with the environment are available, such constructions in *offline* RL typically reduce to *a priori* regularization. In contrast, our work presents a principled geometric perspective on offline RL through gradient flows and introduces proximal updates grounded in both Wasserstein and Fisher–Rao geometries.

Regularization-based Offline RL: Overcoming distributional shift is a central challenge in offline RL, and various methods have been proposed to address it. One line of work constrains the learned policy to remain close to the dataset–i.e., the behavior policy–through policy or critic regularization. TD3+BC (Fujimoto and Gu, 2021) combines Q-value maximization with behavior cloning, while BRAC (Wu et al., 2019) enforces a KL constraint around the behavior policy to prevent large deviations. Its variant, ReBRAC (Tarasov et al., 2023), further adds conservative critic regularization to improve stability. IQL (Kostrikov et al., 2021) employs advantage-weighted regression, implicitly aligning the policy with the dataset without explicit cloning. In this work, we present a *unified* framework for policy optimization that not only grounds existing regularizations in theory but also facilitates enhanced, geometry-aware regularizations.

3 Preliminaries

3.1 MDP and Offline RL

A Markov decision process (MDP) is represented by the tuple $\langle \mathcal{S}, \mathcal{A}, T, r, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are Polish spaces for states and actions, respectively; $T : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the transition kernel with $\mathcal{P}(\mathcal{S})$ the set of probability measures on $\mathcal{S}$; $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function; and $\gamma \in [0, 1)$ is the discount factor. A (stochastic) policy is a Markov kernel $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ which assigns to each state $s \in \mathcal{S}$ a probability measure $\pi(\cdot|s)$ over actions. In RL, the objective is to find an optimal policy $\pi^\star$ that maximizes the expected discounted return $J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$, where the expectation is taken over trajectories generated from s_0 under policy π.

In offline RL, agents no longer interact with the environment and learn solely from a fixed dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ collected by one or more behavior policies (Sutton and Barto, 2018). While the optimization objective remains identical to the online setting, all value estimates and policy updates must be derived exclusively from $\mathcal{D}$. This restriction introduces the fundamental challenge of *distributional shift*: the learned policy may assign probability mass to state–action pairs that are rarely or never observed in $\mathcal{D}$ (Fujimoto et al., 2019), forcing the value function to extrapolate in out-of-distribution regions. This often leads to overestimation and unstable training. The primary objective of offline RL is to learn a policy that maximizes the expected return mitigating the distributional shift between the behavior policy and the learned policy.

3.2 Statistical Manifold

We introduce two representative statistical manifold structures: the Fisher–Rao manifold from information geometry and the Wasserstein geometry from optimal transport.

Fisher–Rao Manifold: Let Ω be a Polish space with Borel σ-field $\mathcal{B}$, and let $\mathcal{P}(\Omega)$ denote the set of all Borel probability measures on $(\Omega, \mathcal{B})$. We define the space of smooth strictly positive probability densities as $\mathcal{P}_+^\infty(\Omega) = \{\rho \in C^\infty(\Omega) : \rho(x) > 0,\ \int_\Omega \rho(x)\,dx = 1\} \subset \mathcal{P}(\Omega)$, where $C^\infty(\Omega)$ denotes the set of infinitely differentiable functions on Ω. This space forms an infinite-dimensional differentiable manifold, commonly called a *statistical manifold*, denoted by $\mathcal{M}$ (Amari and Nagaoka, 2000; Pistone and Sempi, 1995; Ay et al., 2007). Equipping it with the Fisher–Rao metric yields a genuine Riemannian manifold, whose tangent space at ρ, denoted by $T_\rho\mathcal{P}_+^\infty(\Omega)$, consists of smooth functions with zero integral. For $\xi, \eta \in T_\rho\mathcal{P}_+^\infty(\Omega)$, the Fisher–Rao inner product is given by

$$g_\rho^{\mathrm{FR}}(\xi, \eta) = \int_\Omega \frac{\xi(x)\,\eta(x)}{\rho(x)}\,dx. \tag{1}$$

The statistical manifold equipped with the Fisher-Rao inner product is called a *Fisher-Rao manifold*, denoted by $\mathcal{M}_{\mathrm{FR}} = (\mathcal{P}_+^\infty(\Omega), g_{\mathrm{FR}}(\cdot,\cdot))$. The Fisher–Rao distance induced by this inner product admits a closed-form expression. For $\rho_0, \rho_1 \in \mathcal{P}_+^\infty(\Omega)$, it is given by

$$d_{\mathrm{FR}}(\rho_0, \rho_1) = 2\arccos\left(\int_\Omega \sqrt{\rho_0(x)\rho_1(x)}\,dx\right), \tag{2}$$

where the integral is the *Bhattacharyya affinity*, measuring the similarity between the two distributions.

Finally, for a functional $F : \mathcal{P}_+^\infty(\Omega) \to \mathbb{R}$, the associated Riemannian gradient under the Fisher–Rao geometry is given by

$$\mathrm{grad}_\rho^{\mathrm{FR}} F = \rho\left(\frac{\delta F}{\delta\rho} - \mathbb{E}_\rho\left[\frac{\delta F}{\delta\rho}\right]\right), \tag{3}$$

where $\delta F/\delta\rho$ denotes the first variation of F defined as:

$$\langle\frac{\delta F[\rho]}{\delta\rho}, \sigma\rangle = \lim_{\epsilon\to 0} \frac{F[\rho + \sigma\epsilon] - F[\rho]}{\epsilon} \quad \forall\sigma \in T_\rho\mathcal{M}.$$

see Carrillo et al. (2024) for details.

Wasserstein Geometry: Let $\mathcal{P}_2(\Omega)$ denote the space of probability measures on Ω with finite second moments. Following Otto's formal Riemannian calculus (Otto, 2001; Chen and Li, 2020), the tangent space at $\rho \in \mathcal{P}_2(\Omega)$ can be represented as

$$T_\rho\mathcal{P}_2(\Omega) \;=\; \overline{\{-\nabla\cdot(\rho\nabla\phi) : \phi \in C^\infty(\Omega)\}}^{L_2(\rho)}.$$

For tangent vectors $\xi = -\nabla \cdot (\rho\nabla\phi)$ and $\eta = -\nabla \cdot (\rho\nabla\psi)$, the Wasserstein inner product is defined by

$$g_\rho^{\mathcal{W}}(\xi, \eta) = \int_\Omega \nabla\phi(x) \cdot \nabla\psi(x)\,\rho(x)\,dx.$$

Equipped with this inner product, the statistical manifold $\mathcal{P}_2$ is referred to as *Wasserstein manifold*, denoted by $\mathcal{M}_{\mathcal{W}}(\mathcal{P}_2(\Omega), g_\rho^{\mathcal{W}}(\cdot,\cdot))$. Although not a classical Riemannian manifold, this structure admits a formal Riemannian interpretation. The quadratic Wasserstein distance between $\rho_0, \rho_1 \in \mathcal{P}_2(\Omega)$ is defined as

$$\mathcal{W}_2^2(\rho_0, \rho_1) = \inf_{\gamma\in\Gamma(\rho_0,\rho_1)} \int_{\Omega\times\Omega} |x - y|^2\,d\gamma(x, y), \tag{4}$$

where $\Gamma(\rho_0, \rho_1)$ denotes the set of joint distributions with marginals ρ_0 and ρ_1. It is known in (Villani, 2008, Thm. 4.1) that an optimal coupling $\gamma^\star$ always exists. Equivalently, the Benamou–Brenier dynamic formulation (Villani, 2008) expresses this distance as

$$\mathcal{W}_2^2(\rho_0, \rho_1) = \inf_{(\rho_t, v_t)} \Bigg\{ \int_0^1 \int_\Omega \rho_t(x)\,|v_t(x)|^2\,dx\,dt : \\ \partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0 \Bigg\},$$

where the continuity equation links the curve $(\rho_t)_{t\in[0,1]}$ with its velocity field v_t.

Finally, for a functional $F : \mathcal{P}_2(\Omega) \to \mathbb{R}$, the Wasserstein gradient is given by

$$\mathrm{grad}_\rho^{\mathcal{W}} F = -\nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho} \right). \tag{5}$$

4 POGO

In this section, we introduce **P**olicy **O**timization via **G**radient flow in **O**ffline RL (POGO).

4.1 Policy Gradient Flows on Statistical Manifolds

We formulate policy optimization from the perspective of gradient flows, interpreting policy updates as the evolution of a probability distribution along the steepest descent of an energy functional. Fix a state $s \in \mathcal{S}$ and let the policy space form a statistical manifold $\mathcal{M}$. For the state conditioned policy $\pi = \pi(\cdot|s) \in \mathcal{M}$, we define an energy functional $\mathcal{E}_s[\pi]$ on $\mathcal{M}$. Our objective is to identify a policy that maximizes the Q-function. Accordingly, we introduce the *energy functional* as:

$$\mathcal{E}_s[\pi] := -\,\mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q(s,a)\right] \; - \; \lambda\,\mathcal{H}\big(\pi(\cdot\,|\,s)\big), \tag{6}$$

where $\mathcal{H}$ denotes the information entropy and a single hyperparameter λ controls the strength of the entropy regularizer.

The *gradient flow* of $\mathcal{E}_s$ starting from an initial policy π_0 is defined as the continuous-time steepest descent dynamics:

$$\frac{d\pi_t}{dt} = -\mathrm{grad}_\pi\, \mathcal{E}_s[\pi_t], \tag{7}$$

where grad_π denotes the Riemannian gradient induced by the underlying metric. In particular, we have $\mathrm{grad}_\pi = \mathrm{grad}_\pi^{\mathrm{FR}}$ for the Fisher–Rao manifold and $\mathrm{grad}_\pi = \mathrm{grad}_\pi^{\mathcal{W}}$ for the Wasserstein manifold, as defined in (3) and (5), respectively.

With the first variation of $\mathcal{E}_s$ with respect to π

$$\frac{\delta \mathcal{E}_s[\pi]}{\delta \pi}(a) = -Q(s,a) + \lambda(1 + \log \pi(a)), \tag{8}$$

the gradient flows induced by the Fisher–Rao and Wasserstein geometries are given as:

$$\frac{\partial \pi_t}{\partial t} = -\,\pi_t \left(\frac{\delta \mathcal{E}_s}{\delta \pi}\bigg|_{\pi=\pi_t} - \mathbb{E}_\pi \left[\frac{\delta \mathcal{E}_s}{\delta \pi}\bigg|_{\pi=\pi_t} \right] \right) \quad (\mathcal{M}_{\mathrm{FR}})$$

$$\frac{\partial \pi_t}{\partial t} = -\,\nabla \cdot (\pi_t \nabla_a Q(s,\cdot)) + \lambda \nabla_a^2 \pi_t \qquad (\mathcal{M}_{\mathcal{W}}).$$

Solving the gradient flow equation (7) in a closed-form is intractable in general. In practice, these continuous-time flows are typically discretized via the *explicit* Euler step on the manifold:

$$\pi_{k+1} \;=\; R_{\pi_k}\big(-\,\eta\,\mathrm{grad}_\pi\, \mathcal{E}_s[\pi_k]\big), \tag{9}$$

with a step size $\eta > 0$, where R_{π_k} denotes a retraction at π_k on $\mathcal{M}$ (Hu et al., 2019). This explicit scheme requires $\mathcal{E}_s$ to be differentiable, and its convergence rate depends on both the smoothness of $\mathcal{E}_s$ and the choice of the step size η (Salim et al., 2021; Garrigos and Gower, 2024).

In contrast, the *implicit* Euler step, also known as the Jordan–Kinderlehrer–Otto (JKO) scheme, defines the next iterate by minimizing a proximal functional. Given π_k, the *proximal energy* is

$$\mathrm{prox}_{\mathcal{M}}(\mathcal{E}_s \mid \pi_k) \;:=\; \mathcal{E}_s[\pi] \;+\; \frac{1}{2\tau}\, d^2_{\mathcal{M}}(\pi,\pi_k), \tag{10}$$

and the update is given by

$$\pi_{k+1} \in \operatorname*{arg\,min}_{\pi\in\mathcal{M}} \mathrm{prox}_{\mathcal{M}}(\mathcal{E}_s \mid \pi_k)\,, \tag{11}$$

where $d_{\mathcal{M}}(\cdot,\cdot)$ denotes the geodesic distance on the manifold $\mathcal{M}$. This recursion generates a discrete sequence $\{\pi_k\}_{k\geq 0}$. To relate it to the continuous-time flow, we define a piecewise-constant interpolation $\hat{\pi}_t = \pi_k$ for $t \in ((k-1)\tau, k\tau]$ with $\hat{\pi}_0 = \pi_0$, where $k = 1, 2, ..., N$. As $\tau \to 0$, $\hat{\pi}_t$ converges to the continuous solution π_t. Under standard assumptions, a minimizer exists for each $\tau > 0$, the energy decreases along the iterates, and differentiability of $\mathcal{E}_s$ is not required (Ambrosio et al., 2005). Owing to these advantages, we adopt the implicit Euler (JKO) scheme for policy updates in this work.

In (11), the geodesic distance can be taken as the Wasserstein-2 distance $\mathcal{W}_2^2(\cdot,\cdot)$ from (4) for transport-based flows, or as the Fisher–Rao distance $d_{\mathrm{FR}}(\cdot,\cdot)$ from (2) for information-geometric flows. Under standard regularity, the unique minimizer of $\mathcal{E}_s$ is the soft optimal policy

$$\pi^\star(a|s) = \frac{\exp(Q(s,a)/\lambda)}{Z(s)}, \tag{12}$$

where $Z(s) = \int_{\mathcal{A}} \exp(Q(s,a)/\lambda)\,da$ is the partition function and λ denotes the temperature. Convergence of gradient flow to the soft policy under these conditions is discussed in the Supplementary 1.

In the following, we focus on offline RL, where the gradient flow is initialized from the behavior policy and updated using a fixed dataset.

4.2 Offline RL as Gradient Flows

We identify that deterministic policies produced by offline RL algorithms with policy regularization–such as TD3+BC and BRAC–can be interpreted through the lens of gradient flows. Building on this perspective, we extend these algorithms to encompass stochastic policies and multi-step updates (see Figure 1).

Let $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$ be a fixed dataset collected by a behavior policy π_β. Consider the gradient flow of the energy functional in (6) initialized at π_β. Then its first implicit Euler step is given by

$$\pi_\tau = \arg\min_\pi \operatorname{prox}_{\mathcal{M}}(\mathcal{E}_s \mid \pi_\beta)\,. \tag{13}$$

The objective in (13) is closely related to the policy loss used in TD3+BC. The following proposition formalizes this connection.

Proposition 1. *Let $\pi_\tau(a|s) = \delta(a - \mu_\tau(s))$ and $\pi_\beta(a|s) = \delta(a-\mu_\beta(s))$ denote the deterministic policies over $a \in \mathbb{R}^d$, where δ denotes the Dirac delta function, and set $\lambda = 0$ in $\mathcal{E}_s$. Under this setting, the implicit Euler step on $\mathcal{M}_\mathcal{W}$ in (13) reduces to*

$$\mu_\tau = \arg\min_\mu \left\{ -Q\left(s, \mu(s)\right) + \frac{1}{2\tau}\|\mu(s) - \mu_\beta(s)\|_2^2 \right\},$$

which coincides with the TD3+BC policy loss with hyperparameter $\lambda_{TD3+BC} = 2\tau$.

Similarly, for $\pi_\tau, \pi_\beta \in \mathcal{P}_+^\infty(\Omega)$ with $d_{FR}(\pi_\tau, \pi_\beta) < \epsilon$, the Euler step on $\mathcal{M}$ is equivalent to the KL-BRAC policy loss with hyperparameter $\alpha_{BRAC} = \frac{1}{\tau}$ up to error $O(\epsilon^{3/2})$.

Therefore, we identify that standard policy-regularization methods in offline RL can be interpreted as single-step Euler updates initialized at π_β. This perspective naturally motivates a generalized policy optimization framework grounded in implicit Euler steps.

4.3 Algorithm and Implementation

We now present POGO, which follows an actor-critic framework based on function approximation. In this framework, the actor and critic correspond to the policy and Q-functions, respectively, and are represented by parameterized functions, denoted as π_θ and Q_{ψ_i} for $i = 1, 2$. The soft-updated target networks are denoted by θ^- and ψ_i^-.

To estimate the critics from a fixed dataset $\mathcal{D}$, we adopt the *clipped double Q-learning* objective, as used in TD3 (Fujimoto et al., 2018) and TD3-BC (Fujimoto and Gu, 2021). The two critics are updated by minimizing the squared Bellman error:

$$L_{\psi_i} = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[(Q_{\psi_i}(s,a) - y)^2\right],\ i = 1, 2, \tag{14}$$

where the bootstrapped target is defined as

$$y \ :=\ r + \gamma \min_{j\in\{1,2\}} Q_{\psi_j^-}(s',\, a')\,, \quad a' \sim \pi_{\theta^-}(\cdot|s').$$

The behavior policy π_β is estimated from $\mathcal{D}$ via maximum likelihood:

$$L_\beta = -\mathbb{E}_{(s,a)\sim\mathcal{D}}[\log \pi_\beta(a|s)]. \tag{15}$$

Having established the critics and the behavior policy, the actor is updated by approximating the implicit Euler step in (13):

$$L_\theta = \mathbb{E}_{s\sim\mathcal{D}}\Big[\operatorname{prox}_{\mathcal{M}}\Big(\hat{\mathcal{E}}_s \mid \pi_\beta(\cdot|s)\Big)\Big], \tag{16}$$

with the estimated energy functional

$$\hat{\mathcal{E}}_s[\pi_\theta] := \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[-\min_{i\in\{1,2\}} Q_{\psi_i}(s,a) - \lambda\,\mathcal{H}\big(\pi_\theta(\cdot|s)\big)\right].$$

To implement the actor update described above, we assume a Gaussian policy. Specifically, all policies are modeled as diagonal Gaussian distributions on Wasserstein manifold $\mathcal{M}_\mathcal{W}$ (See Section 1 of the Supplementary Material for Fisher-Rao case). For each state $s \in \mathcal{S}$, the behavior policy is defined as

$$\pi_\beta(\cdot|s) = \mathcal{N}\left(\mu_\beta(s), \operatorname{diag}(\sigma_\beta^2(s))\right),$$

and the learned policy is defined as

$$\pi_\theta(\cdot|s) = \mathcal{N}\left(\mu_\theta(s), \operatorname{diag}(\sigma_\theta^2(s))\right).$$

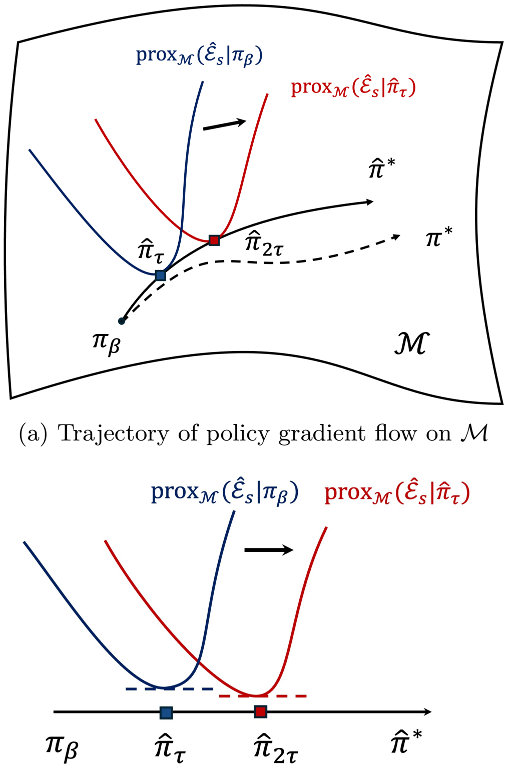
Under this Gaussian assumption, the policy loss in (16) becomes tractable because a closed-form expression for the Wasserstein distance between diagonal Gaussian policies can be derived:

Lemma 2. *Let $\pi_i = \mathcal{N}(\mu_i, \operatorname{diag}(\sigma_i^2)) \in \mathcal{P}_2(\Omega)$, $i \in \{1, 2\}$, with $\mu_i \in \mathbb{R}^d$, $\sigma_i \in \mathbb{R}^d$. Then, the squared Wasserstein distance is given by*

$$\mathcal{W}_2^2(\pi_1, \pi_2) = \|\mu_1 - \mu_2\|_2^2 \ +\ \|\sigma_1 - \sigma_2\|_2^2. \tag{17}$$

From Lemma 2, the actor loss in (16) can be explicitly written as

$$\begin{aligned} L_{\theta:=(\mu_\theta,\sigma_\theta)} = \hat{\mathcal{E}}_s[\pi_\theta] + \frac{1}{2\tau}\Big(&\|\mu_\beta(s) - \mu_\theta(s)\|_2^2 \\ &+ \|\sigma_\beta(s) - \sigma_\theta(s)\|_2^2\Big). \end{aligned} \tag{18}$$


$\mathrm{prox}_{\mathcal{M}}(\hat{\mathcal{E}}_s|\pi_\beta)$
$\mathrm{prox}_{\mathcal{M}}(\hat{\mathcal{E}}_s|\hat{\pi}_\tau)$
$\hat{\pi}^*$
π^*
$\hat{\pi}_\tau$
$\hat{\pi}_{2\tau}$
π_β
$\mathcal{M}$
$\mathrm{prox}_{\mathcal{M}}(\hat{\mathcal{E}}_s|\pi_\beta)$
$\mathrm{prox}_{\mathcal{M}}(\hat{\mathcal{E}}_s|\hat{\pi}_\tau)$
π_β
$\hat{\pi}_\tau$
$\hat{\pi}_{2\tau}$
$\hat{\pi}^*$

(a) Trajectory of policy gradient flow on $\mathcal{M}$

(b) Proximal energy shift along with timesteps

Figure 1: Policy evolution on the manifold $\mathcal{M}$. Starting from the behavior policy π_β, discrete Euler steps $(\pi_\tau, \pi_{2\tau}, \ldots)$ approximate the continuous gradient flow trajectory toward the estimated soft optimum $\hat{\pi}^\star$ in (12), as induced by the learned Q-network. The dashed curve denotes the true gradient flow path toward the actual soft optimum $\pi^\star$.

This formulation enables efficient and tractable computation of the actor update under the Wasserstein geometry.

Based on the above formulation, we summarize the overall training procedure in Algorithm 1.

The gradients $\tilde{\nabla}$ of the actor and behavior loss functions can be computed using either Euclidean or natural gradient. For identity matrix $I \in \mathbb{R}^{d \times d}$, the metric tensor on parameter space is given as:

$$G_{\mathcal{W}} = \begin{bmatrix} I & 0 \\ 0 & 2I \end{bmatrix},$$

and the corresponding Wasserstein natural gradient is

$$\nabla^{\mathcal{W}} = G_{\mathcal{W}_2}^{-1} \nabla.$$

Algorithm 1 POGO

```
Require: Offline dataset $\mathcal{D}$, geometry $d_\mathcal{M}$
Ensure: Optimized policy $\pi_\theta$
Initialize critics $\psi_1, \psi_2$, actor $\theta$, behavior $\beta$
Set targets $\psi_i^- \leftarrow \psi_i$, $\theta^- \leftarrow \theta$
for each gradient step do
    Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
    Update critics: $\psi_i \leftarrow \psi_i - \lambda_Q \nabla_{\psi_i} L_{\psi_i}$
    Update behavior policy: $\beta \leftarrow \beta - \lambda_\beta \tilde{\nabla}_\beta L_\beta$
    Update actor (geometric step):
        $\theta \leftarrow \theta - \lambda_\pi \tilde{\nabla}_\theta L_\theta$
    if target update frequency then
        $\psi_i^- \leftarrow (1-\rho)\psi_i^- + \rho\psi_i, \quad i = 1, 2$
        $\theta^- \leftarrow (1-\rho)\theta^- + \rho\theta$
    end if
end for
```

Therefore, the gradient respect to each parameter is formulated as

$$\nabla_\mu^{\mathcal{W}} = \nabla_\mu \tag{19}$$

$$\nabla_{\sigma^2}^{\mathcal{W}} = \frac{1}{2} \nabla_{\sigma^2}. \tag{20}$$

More details are provided in the Supplementary Material.

Multi-step POGO: A single implicit Euler step recovers the standard proximal update. By incorporating an additional re-centered step, the update is naturally extended along the gradient-flow trajectory (see Fig. 1), yielding a closer approximation of the surrogate optimum and potentially approaching the true optimal policy.

After completing the first step training in Algorithm 1, an additional implicit Euler step can be performed using the learned actor and critic networks. The second Euler step is formulated as:

$$\pi_{2\tau} = \arg\min_{\pi} \mathrm{prox}_{\mathcal{M}} \left(\hat{\mathcal{E}}_s \mid \pi_\tau \right),$$

where π_τ is the learned policy in Algorithm 1.

From (14), mimicking π_τ as the behavior policy yields the critic loss:

$$L_{\psi_i} = \mathbb{E}_{s \sim \mathcal{D}, a \sim \pi_\tau(\cdot|s)} \left[(Q_{\psi_i}(s,a) - y)^2 \right],$$

which is intractable in offline RL, as the bootstrapped target y cannot be computed. To address this, the critic networks learned in the first step are kept fixed via the stop-gradient operator, sg. Geometrically, this corresponds to performing the subsequent proximal step along the fixed solid line in Figure 1a. In other words, the estimated energy functional $\hat{\mathcal{E}}_s$ obtained from the first step remains unchanged. However, since

Algorithm 2 Multi-step POGO

```
Require: Offline dataset $\mathcal{D}$, geometry $d_{\mathcal{M}}$, number of Euler steps $K$, initial policy $\pi_\tau$ from Algorithm 1
Ensure: Optimized policy $\pi_{K\tau}$
for $k = 1, \ldots, K$ do
    function (Multistep–POGO)($\psi_1, \psi_2, \pi_{k\tau}$)
        Fix $\psi_i \leftarrow \mathrm{sg}(\psi_i)$, $i \in [1, 2]$
        Set $\pi_\beta, \pi_\theta \leftarrow \pi_{\theta,k\tau}$
        for each gradient step do
            Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
            Update actor (geometric step):
                $\theta \leftarrow \theta - \lambda_\pi \tilde{\nabla}_\theta L_\theta$
        end for
        Set $\pi_{(k+1)\tau} \leftarrow \pi_\theta$
        return $\pi_{(k+1)\tau}$
    end function
end for
```

the proximal energy functional is shifted, it admits a new minimizer as depicted in Figure 1b.

This additional step can be applied recursively. Set $\pi_0 := \pi_\beta$ and, with the critic frozen, consider the following iteration from $m = 0, 1, ..., K-1$:

$$\pi_{(m+1)\tau} = \arg\min_{\pi} \operatorname{prox}_{\mathcal{M}}\left(\hat{\mathcal{E}}_s \mid \pi_{m\tau}\right).$$

This yields a K-step chain of re-centered proximal updates that monotonically decrease the surrogate energy along the fixed gradient flow trajectory. Each iteration is summarized in Algorithm 2.

5 Experimental Results

We evaluate our methods on D4RL (Fu et al., 2020) Gym-MuJoCo environments (HalfCheetah, Hopper, Walker2d), reporting normalized returns. We compare widely recognized offline RL baselines, including TD3+BC (Fujimoto and Gu, 2021), IQL (Kostrikov et al., 2021), and ReBRAC (Tarasov et al., 2023), using the recommended hyperparameters and network architectures provided by CORL (Tarasov et al., 2022). To decouple the effect of geometric choices from network design, we assess two variants of our approach—POGO–W (Wasserstein) and POGO–FR (Fisher–Rao)—both instantiated with a TD3+BC-inspired backbone. Additional implementation details are provided in Section 2 of the Supplementary Material, while the results for the Fisher–Rao geometry and AntMaze environments are deferred to Section 3 of the Supplementary Material.

Two-step vs. One-step POGO: Figure 2 demonstrates that our proposed method improves performance through two-step Euler updates. Importantly, this approach is fundamentally different from merely increasing the rollout horizon in the one-step approach. As illustrated in Figure 1, the two-step scheme optimizes a distinct objective function, yielding a superior policy under the estimated energy functional. This highlights that the performance gains stem not from longer rollouts, but from the variational structure of the two-step update, which more effectively aligns policy updates with the underlying geometry.

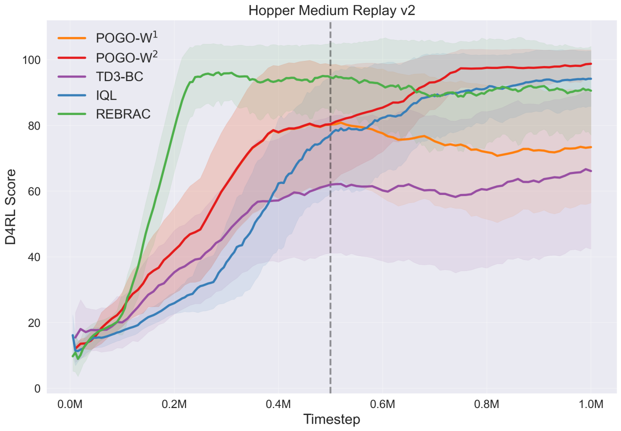


Figure 2: Shaded regions show one standard deviation over five seeds on `hopper-medium-replay-v2`. POGO–W^1 and POGO–W^2 denotes the one step POGO and re-centered two-step POGO, respectively. The black dashed vertical line marks the switching point to the POGO-W^2.

In some cases, however, the two-step variant underperforms the one-step approach—for example, in `halfcheetah-medium-expert`. In this setting, we allocated half of the training steps to the one-step update and the remaining half to the two-step update, without explicitly accounting for the stabilization of the critic. Under such a coarse schedule, prematurely applying the second Euler step tends to propagate value-estimation errors rather than refining the policy update, leading to the observed performance degradation. Nevertheless, across the majority of environments, the two-step update yielded higher final performance, highlighting its effectiveness.

Comparison with baselines: Table 1 summarizes the results. On MuJoCo tasks, POGO–W (with a TD3+BC base) matches or outperforms TD3+BC and IQL. The two-step Euler update effectively extends the implicit Euler integration along the estimated gradient-flow field (Fig. 1) and generally provides performance gains once the critic has stabilized. Unlike ReBRAC, which adopts a more elaborate architecture with an ensemble of ten critics, larger batch sizes, deeper networks, and normalization techniques

Table 1: D4RL Evaluation of Offline RL Algorithms in MuJoCo Environments.

Task Name	**IQL**	**TD3+BC**	**ReBRAC**	**POGO–W**	**POGO–W(2step)**
halfcheetah-medium	48.33 ± 0.26	48.33 ± 0.48	**65.86** ± 0.73	58.57 ± 0.66	61.91 ± 1.25
halfcheetah-medium-replay	42.95 ± 1.87	44.51 ± 0.36	49.78 ± 1.29	52.33 ± 1.34	**53.75** ± 1.32
halfcheetah-medium-expert	94.05 ± 0.53	88.76 ± 7.75	**98.32** ± 11.02	96.72 ± 1.21	50.52 ± 5.37
hopper-medium	67.23 ± 5.49	54.91 ± 3.46	**102.43** ± 0.28	64.71 ± 4.87	74.78 ± 11.19
hopper-medium-replay	93.51 ± 9.12	60.10 ± 33.17	89.62 ± 19.78	85.68 ± 15.56	**99.34** ± 1.69
hopper-medium-expert	102.66 ± 8.34	100.30 ± 11.19	**109.04** ± 5.66	100.65 ± 5.20	93.67 ± 11.98
walker2d-medium	76.27 ± 6.69	84.33 ± 0.18	84.39 ± 1.60	85.69 ± 1.61	**86.70** ± 1.65
walker2d-medium-replay	66.14 ± 12.00	78.57 ± 10.20	80.11 ± 9.88	**93.78** ± 1.74	91.61 ± 1.72
walker2d-medium-expert	109.02 ± 2.38	110.55 ± 0.35	111.78 ± 0.42	110.93 ± 0.37	**111.92** ± 0.51
MuJoCo Average	77.80	74.48	87.93	83.23	80.47

such as LayerNorm, POGO–W is instantiated on the minimal TD3+BC backbone. Despite this streamlined design, POGO–W attains performance that is comparable to, and in several cases surpasses, that of ReBRAC. This demonstrates that the gradient-flow formulation provides substantial benefits without relying on heavy architectural modifications.

6 Conclusion, Limitation, and Future Work

We presented a geometric framework for offline RL, formulating policy optimization as gradient flow over probability spaces and deriving POGO, a stable and efficient algorithm instantiated under Fisher–Rao and Wasserstein geometries. Our Euler-step interpretation naturally induces proximal regularization and motivates multi-step POGO, which achieves performance gains across diverse tasks. While our study assumes Gaussian policies and fixed step numbers, it lays the groundwork for geometry-driven approaches to offline RL.

Our work is limited to diagonal Gaussian policies in order to leverage closed-form probability metrics. Our next research contains gradient flow-based policy optimization for general stochastic policies. We also aim to develop the theory and practical criteria that guarantee monotone energy decrease for multi-step re-centered updates.

Future work includes extending gradient flow-based policy optimization to more general stochastic policies, developing theoretical guarantees for multi-step re-centered updates, and exploring adaptive discretization as well as broader probability geometries. Another promising direction is to investigate adaptive scheduling of the second Euler step, for example by delaying it until the critic has sufficiently stabilized.

Bibliography

Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry.* Translations of mathematical monographs. American Mathematical Society.

Ambrosio, L., Gigli, N., and Savare, G. (2005). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* Lectures in Mathematics. ETH Zürich. Birkhäuser Basel.

Asadulaev, A., Korst, R., Korotin, A., Egiazarian, V., Filchenkov, A., and Burnaev, E. (2024). Rethinking optimal transport in offline reinforcement learning.

Ay, N., Jost, J., Vân Lê, H., and Schwachhöfer, L. (2007). *Information Geometry.* Springer Cham.

Carrillo, J. A., Chen, Y., Huang, D. Z., Huang, J., and Wei, D. (2024). Fisher-rao gradient flow: Geodesic convexity and functional inequalities.

Chen, Y. and Li, W. (2020). Optimal transport natural gradient for statistical manifolds with continuous sample space.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020). D4rl: Datasets for deep data-driven reinforcement learning.

Fujimoto, S. and Gu, S. S. (2021). A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145.

Fujimoto, S., Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR.

Fujimoto, S., Meger, D., and Precup, D. (2019). Off-policy deep reinforcement learning without exploration.

Garrigos, G. and Gower, R. M. (2024). Handbook of convergence theorems for (stochastic) gradient methods.

Hu, J., Liu, X., Wen, Z., and Yuan, Y. (2019). A brief introduction to manifold optimization.

Kakade, S. M. (2001). A natural policy gradient. In *NIPS*.

Kerimkulov, B., Leahy, J.-M., Siska, D., Szpruch, L., and Zhang, Y. (2025). A fisher–rao gradient flow for entropy-regularised markov decision processes in polish spaces: B. kerimkulov et al. *Foundations of Computational Mathematics*, pages 1–75.

Kostrikov, I., Nair, A., and Levine, S. (2021). Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*.

Lascu, R.-A., Šiška, D., and Łukasz Szpruch (2025). Ppo in the fisher-rao geometry.

Likmeta, A., Sacco, M., Metelli, A. M., and Restelli, M. (2023). Wasserstein actor-critic: Directed exploration via optimism for continuous-actions control.

Moskovitz, T., Arbel, M., Huszar, F., and Gretton, A. (2020). Efficient wasserstein natural gradients for reinforcement learning. *arXiv preprint arXiv:2010.05380*.

Otto, F. (2001). The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174.

Pistone, G. and Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The annals of statistics*, pages 1543–1561.

Salim, A., Korba, A., and Luise, G. (2021). The wasserstein proximal gradient algorithm.

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. (2017). Trust region policy optimization.

Song, J., He, N., Ding, L., and Zhao, C. (2023). Provably convergent policy optimization via metric-aware trust region methods.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Tarasov, D., Kurenkov, V., Nikulin, A., and Kolesnikov, S. (2023). Revisiting the minimalist approach to offline reinforcement learning.

Tarasov, D., Nikulin, A., Akimov, D., Kurenkov, V., and Kolesnikov, S. (2022). CORL: Research-oriented deep offline reinforcement learning library. In *3rd Offline RL Workshop: Offline RL as a "Launchpad"*.

Terpin, A., Lanzetti, N., Yardim, B., Dörfler, F., and Ramponi, G. (2022). Trust region policy optimization with optimal transport discrepancies: Duality and algorithm for continuous actions.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Wu, Y., Tucker, G., and Nachum, O. (2019). Behavior regularized offline reinforcement learning.

Zhang, R., Chen, C., Li, C., and Carin, L. (2018). Policy optimization as wasserstein gradient flows. *CoRR*, abs/1808.03030.