Theoretical Foundations of Wasserstein Policy Optimization

Anonymous AI Agent (first author) **Anonymous Human Co-author(s)**
Affiliation
Address
email

Abstract

We revisit Wasserstein Policy Optimization (WPO) as policy transport on action densities followed by a projection onto a parametric manifold. Evolving policies by a 2-Wasserstein gradient flow and projecting in the Fisher/KL inner product yields a covariant natural step with a mixed-derivative cross-term. We make this projection-based view explicit, prove baseline invariance (via a constrained Gáteaux variation) and parameterization covariance (via Fisher pullbacks), and delineate when the step coincides with natural policy gradient (affine-in-action exponential families) versus when it departs (mixtures, squashings). For Gaussian policies we give mean and covariance updates, including a full-covariance Cholesky implementation that preserves SPD. We extend to c-Wasserstein dynamics to obtain principled stability via convex conjugates and state precise energy inequalities in the frozen-critic regime. Assumptions and weak-form conditions are spelled out, and connections to classic PG/DPG/NPG are established.

1 Introduction

Wasserstein Policy Optimization (WPO) offers a principled link between optimal transport and policy updates through Wasserstein gradient flows. Despite elegant foundations, theoretical questions remain: the precise projection from infinite-dimensional flows to parametric updates; invariance to baselines and parameterizations; and stability when action-value gradients are large. We present a self-contained treatment addressing these points with label-anchored derivations, and we identify the conditions under which the WPO update coincides with, or departs from, classic policy gradient/natural gradient directions.

2 Background

On the 2-Wasserstein manifold, steepest descent of a functional $\mathcal{J}[\pi]$ follows the continuity equation with velocity $v = -\nabla_{\mathbf{a}}(\delta\mathcal{J}/\delta\pi)$ [Ambrosio et al., 2008, Benamou and Brenier, 2000]. Taking $\mathcal{J}[\pi] = -\mathbb{E}[Q^\pi]$ yields $\partial_t\pi = -\nabla_{\mathbf{a}}\cdot(\pi\,\nabla_{\mathbf{a}}Q^\pi)$. Projecting this infinite-dimensional flow to a parametric family $\{\pi_\theta\}$ leads to the natural-gradient form $\Delta\theta = F_{\theta\theta}^{-1}\,\mathcal{F}_{t\theta}$, where the cross term is $\mathbb{E}[\nabla_\theta\nabla_{\mathbf{a}}\log\pi\,\nabla_{\mathbf{a}}Q^\pi]$ under mild regularity and boundary conditions. We relate to natural policy gradient [Kakade, 2001] and to neural ODE views of transport [Chen et al., 2018].

Weak form and boundary conditions. We interpret the continuity equation in weak form, with no-flux boundary condition $(\pi v)\cdot n = 0$ on bounded action domains or vanishing flux at infinity on $\mathbb{R}^d$. Under C^1/C^2 regularity and dominated convergence, integration by parts is valid and boundary terms vanish; all energy identities and projections below are stated in this weak-form sense.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

3 Preliminaries and Notation

Let $\mathcal{A} \subseteq \mathbb{R}^d$ denote the action domain and $\mathcal{S}$ the state space. Policies are absolutely continuous densities $\pi_\theta(\mathbf{a} \mid s)$ on $\mathcal{A}$ for each $s \in \mathcal{S}$. We write expectations as $\mathbb{E}_{s\sim d^\pi,\, \mathbf{a}\sim\pi_\theta(\cdot|s)}[\cdot]$ for a fixed reference weighting $d^\pi(s)$ (e.g., discounted occupancy). In line with the policy gradient theorem, our first-order variations operate in a semi-gradient setting where d^π and Q^π are treated as fixed.

Notation. $\mathcal{S}$ states, $\mathcal{A} \subseteq \mathbb{R}^d$ actions; d^π discounted occupancy; $\pi_\theta(\mathbf{a} \mid s)$ policy density; $Q^\pi(s, \mathbf{a})$ critic; $F_{\theta\theta}$ Fisher; $\mathrm{sym}(X) = \frac{1}{2}(X + X^\top)$; $\mathbb{S}^d_{++}$ SPD cone. We use bold $\mathbf{a}$ for vectors and a for 1D examples.

Throughout, expectations are with respect to the joint measure $s \sim d^\pi$, $\mathbf{a} \sim \pi_\theta(\cdot \mid s)$. When we abbreviate $\mathbb{E}_{s,\mathbf{a}}[\,\cdot\,]$, this is the understood measure unless stated otherwise.

The Fisher matrix is

$$F_{\theta\theta} \;=\; \mathbb{E}_{s\sim d^\pi,\, \mathbf{a}\sim\pi_\theta(\cdot|s)}\big[\nabla_\theta \log \pi_\theta(\mathbf{a} \mid s)\, \nabla_\theta \log \pi_\theta(\mathbf{a} \mid s)^\top\big]. \tag{1}$$

We equip density variations with the inner product

$$\langle f, g\rangle \;=\; \mathbb{E}_{s\sim d^\pi}\Big[\int_{\mathcal{A}} \frac{f(\mathbf{a}, s)\, g(\mathbf{a}, s)}{\pi_\theta(\mathbf{a} \mid s)}\, d\mathbf{a}\Big], \tag{2}$$

for which the parametric tangent directions are $\pi\, \nabla_\theta \log \pi$. Throughout, we assume sufficient smoothness (C^1/C^2), dominated convergence conditions to exchange expectation and differentiation, and vanishing boundary terms so that integration by parts is valid.

4 Derivation of WPO

4.1 Problem Setup

We consider continuous control with policy $\pi_\theta(\mathbf{a} \mid s)$ and critic $Q^\pi(s, \mathbf{a})$. We adopt a reference state distribution $d^\pi(s)$ (e.g., discounted occupancy) and study policy updates that arise as projections of Wasserstein gradient flows on action densities.

4.2 From Wasserstein Flow to Parametric Update

Let $\mathcal{J}[\pi] = -\mathbb{E}_{s\sim d^\pi,\mathbf{a}\sim\pi(\cdot|s)}[Q^\pi(s, \mathbf{a})]$. We adopt a per-state time rescaling and work with the rescaled flow whose velocity uses $v = \nabla_{\mathbf{a}} Q^\pi$ (absorbing the d^π factor). The 2-Wasserstein flow is

$$\partial_t \pi \;=\; -\nabla_{\mathbf{a}} \cdot \big(\pi\, \nabla_{\mathbf{a}} Q^\pi\big). \tag{3}$$

We note that embedding the discounted state weighting $d^\pi(s)$ into $\delta\mathcal{J}/\delta\pi$ scales the per-state velocity by $d^\pi(s)$. This is a per-state time reparameterization and preserves weak-form energy identities, but a Fisher-projected finite step may change unless additional conditions (e.g., state-wise collinearity of velocities or constant scaling) hold. To remove this ambiguity, we adopt the convention $\delta\mathcal{J}/\delta\pi = -Q^\pi$ and carry $d^\pi(s)$ only as an outer expectation in all cross terms and Fisher quantities. We project to $\{\pi_\theta\}$ by minimizing, in the Fisher metric, the discrepancy between $\partial_t \pi$ and $\delta\pi_\theta = \pi\, (\nabla_\theta \log \pi) \cdot \Delta\theta$, leading to $F_{\theta\theta}\, \Delta\theta = \mathcal{F}_{t\theta}$ with

$$\mathcal{F}_{t\theta} = \mathbb{E}_{s\sim d^\pi,\mathbf{a}\sim\pi}\big[\nabla_\theta \nabla_{\mathbf{a}} \log \pi_\theta(\mathbf{a} \mid s)\; \nabla_{\mathbf{a}} Q^\pi(s, \mathbf{a})\big]. \tag{4}$$

4.3 Energy Dissipation

Under suitable regularity, and in the frozen-critic regime, the rescaled gradient flow (3) dissipates $\mathcal{J}$:

Lemma 1 (Energy dissipation)**.** *Along solutions π_t of* (3)*, one has $\frac{d}{dt}\mathcal{J}[\pi_t] \le 0$ with $\frac{d}{dt}\mathcal{J}[\pi_t] = -\,\mathbb{E}_{s\sim d^\pi,\mathbf{a}\sim\pi_t}\big[\|\nabla_{\mathbf{a}} Q^\pi(s, \mathbf{a})\|^2\big]$. For the c-Wasserstein flow with velocity $\nabla c^*(\nabla_{\mathbf{a}}\delta\mathcal{J}/\delta\pi)$ one has $\frac{d}{dt}\mathcal{J}[\pi_t] = -\,\mathbb{E}_{s,\mathbf{a}}\big[\langle \nabla c^*(\nabla_{\mathbf{a}} Q^\pi),\, \nabla_{\mathbf{a}} Q^\pi\rangle\big] \le 0$ by monotonicity of ∇c^*. See Appendix (Energy Dissipation) for weak-form derivations and function-space assumptions.*

4.4 Functional Derivative of $\mathcal{J}$

For discounted RL, under the semi-gradient convention (holding d^π and Q^π fixed), we adopt the rescaled functional derivative compatible with (3):

$$\frac{\delta \mathcal{J}}{\delta \pi}(s, \mathbf{a}) \;=\; -Q^\pi(s, \mathbf{a}), \quad d^\pi(s) = (1-\gamma)\sum_t \gamma^t \, \Pr(s_t = s). \tag{5}$$

This corresponds to absorbing $d^\pi(s)$ via a per-state time reparameterization. Under the alternative convention that embeds $d^\pi(s)$ inside $\delta\mathcal{J}/\delta\pi$, the Fisher-projected finite step can differ unless state-wise collinearity or constant scaling holds; our adopted convention avoids this ambiguity while preserving the weak-form energy identities.

4.5 KL/Fisher Projection and Normal Equations

Equivalently, minimizing $\|\partial_t \pi - \delta\pi_\theta\|^2$ in the Fisher norm (2) yields the normal equations $F_{\theta\theta}\,\Delta\theta = \mathcal{F}_{t\theta}$. The cross term satisfies

$$\mathcal{F}_{t\theta} = \int \nabla_\theta \log \pi_\theta(\mathbf{a} \mid s)\, \partial_t \pi_\theta(\mathbf{a} \mid s)\, d\mathbf{a} \tag{6}$$

$$= -\int \nabla_\theta \log \pi_\theta \; \nabla_{\mathbf{a}} \cdot \big(\pi_\theta \, \nabla_{\mathbf{a}} Q^\pi\big)\, d\mathbf{a} \tag{7}$$

$$= \int \big(\nabla_{\mathbf{a}} \nabla_\theta \log \pi_\theta\big)^\top (\pi_\theta \, \nabla_{\mathbf{a}} Q^\pi)\, d\mathbf{a} \tag{8}$$

$$= \mathbb{E}_{\mathbf{a}\sim\pi_\theta}\big[\nabla_\theta \nabla_{\mathbf{a}} \log \pi_\theta \; \nabla_{\mathbf{a}} Q^\pi\big], \tag{9}$$

where boundary terms vanish and derivatives commute under the stated assumptions. Under C^2 regularity and vanishing boundary terms, this follows from integration by parts. The natural-gradient step is $\Delta\theta = F_{\theta\theta}^{-1}\,\mathcal{F}_{t\theta}$.

4.6 Projection Metric: Fisher vs. W_2

We project the W_2 flow in the Fisher/KL inner product on densities, which induces a covariant natural step on the parametric manifold and yields simple estimators. A W_2-parametric projection would instead align the parametric velocity with the W_2 tangent metric via a velocity potential on parameters.

Operator-level relation (per state). Let $\mathcal{T}_\theta$ denote the parametric tangent span $\{\pi \, \nabla_\theta \log \pi \, u : u \in \mathbb{R}^{\dim \theta}\}$. The Fisher projector and the W_2-tangent projector are two positive self-adjoint operators on $\mathcal{T}_\theta$ induced by different inner products. They coincide when $\mathcal{T}_\theta$ is invariant and the per-state velocity is collinear with the tangent directions (e.g., exponential families with sufficient statistics affine in $\mathbf{a}$ under a compatible parameterization, such as Gaussian means with fixed covariance). In general they differ by a positive operator on $\mathcal{T}_\theta$. We favor Fisher projection for statistical stability, parameterization covariance, and practicality.

4.7 Stability via c-Wasserstein

For convex c, the c-Wasserstein flow uses the velocity $v = \nabla c^*\big(\nabla_{\mathbf{a}}(\delta\mathcal{J}/\delta\pi)\big) = \nabla c^*(-\nabla_{\mathbf{a}} Q^\pi)$. The corresponding flow is $\partial_t \pi = -\nabla_{\mathbf{a}} \cdot \big(\pi \, \nabla c^*(\nabla_{\mathbf{a}}(\delta\mathcal{J}/\delta\pi))\big)$. If c^* is even, then $\nabla c^*(-x) = -\nabla c^*(x)$ and one may equivalently write $\partial_t \pi = -\nabla_{\mathbf{a}} \cdot \big(\pi \, \nabla c^*(\nabla_{\mathbf{a}} Q^\pi)\big)$. For a separable choice $c(u) = \frac{1}{4}\sum_i |u_i|^4$, one has $c^*(x) = \frac{3}{4}\sum_i |x_i|^{4/3}$ and the elementwise cube-root map $\nabla c^*(x)_i = \mathrm{sign}(x_i)|x_i|^{1/3}$, controlling large action-gradients while preserving descent/ascent directions. A rotationally invariant alternative $c(u) = \frac{1}{4}\|u\|_2^4$ yields $\nabla c^*(x) = \|x\|_2^{-2/3} x$ (radial shrinkage). These shrinkage maps are direction-preserving, monotone, and Hölder-continuous, ensuring the energy identity holds and often allowing larger stable steps than heuristic gradient clipping under the same frozen-critic assumptions.

Theorem 1 (Energy decay for c-Wasserstein flows)**.** *Under Assumption 1, let c be proper, convex, l.s.c., with differentiable c^* and monotone ∇c^*. The flow*

$$\partial_t \pi \;=\; -\nabla_{\mathbf{a}} \cdot \Big(\pi \, \nabla c^*\big(\nabla_{\mathbf{a}}(\delta\mathcal{J}/\delta\pi)\big)\Big)$$

satisfies the intrinsic energy identity

$$\frac{d}{dt}\mathcal{J}[\pi_t] = -\,\mathbb{E}_{s,\mathbf{a}}\Big[\big\langle \nabla c^*\big(\nabla_{\mathbf{a}}(\delta\mathcal{J}/\delta\pi)\big),\, \nabla_{\mathbf{a}}(\delta\mathcal{J}/\delta\pi)\big\rangle\Big] \le 0,$$

with equality iff $\nabla_{\mathbf{a}}(\delta\mathcal{J}/\delta\pi) = 0$ *almost everywhere. Specializing to* $\delta\mathcal{J}/\delta\pi = -Q^\pi$ *gives* $-\,\mathbb{E}[\langle \nabla c^*(-\nabla_{\mathbf{a}}Q^\pi),\, -\nabla_{\mathbf{a}}Q^\pi\rangle] \le 0$*; if in addition* c^* *is even (hence* ∇c^* *odd), this equals* $-\,\mathbb{E}[\langle \nabla c^*(\nabla_{\mathbf{a}}Q^\pi),\, \nabla_{\mathbf{a}}Q^\pi\rangle] \le 0$*. For* $c(u) = \frac{1}{4}\sum_i |u_i|^4$ *one has elementwise shrinkage* $\nabla c^*(x)_i = \mathrm{sign}(x_i)|x_i|^{1/3}$*; for* $c(u) = \frac{1}{4}\|u\|_2^4$ *one has radial shrinkage* $\nabla c^*(x) = \|x\|_2^{-2/3}x$.

Practical note: the induced map is direction-preserving and monotone, which yields energy decay under frozen critics. Unlike heuristic gradient clipping, it retains a variational interpretation and often permits larger stable steps in ill-conditioned regimes. See Theorem 1 and Appendix A.2.

4.8 Estimators and variance reduction

With a reparameterization $\mathbf{a} = f_\theta(\epsilon, s)$ and a frozen critic, an unbiased single-sample estimator of the cross term is

$$\widehat{g}_\theta(\epsilon, s) = \nabla_\theta \nabla_{\mathbf{a}} \log \pi_\theta\big(f_\theta(\epsilon, s) \mid s\big)\, \widehat{\nabla_{\mathbf{a}} Q}(s, f_\theta(\epsilon, s)).$$

Mixed derivatives are computed efficiently via JVP/VJP (one forward- and one reverse-mode pass suffice in most autodiff systems). The critic action-gradient $\nabla_{\mathbf{a}}Q$ comes from differentiating the critic wrt its action input. A score-aligned control variate subtracts $c(s)\,\nabla_\theta \log \pi_\theta(\mathbf{a} \mid s)$, with $c(s)$ chosen by least-squares to minimize variance, leaving $\mathbb{E}[\widehat{g}_\theta]$ unchanged. The optimal per-state coefficient is

$$c^*(s) \;=\; \frac{\mathrm{Cov}\left(\nabla_\theta \log \pi_\theta,\, \widehat{g}_\theta \,\middle|\, s\right)}{\mathrm{Var}\left(\nabla_\theta \log \pi_\theta \,\middle|\, s\right)}, \tag{10}$$

estimated online via running moments.

4.9 Gaussian Parameter Updates

For Gaussian policies, 1D mean/variance updates are

$$\Delta\mu = \mathbb{E}[\nabla_a Q], \tag{11}$$

$$\Delta(\sigma^2) = 2\,\mathbb{E}[(a-\mu)\,\nabla_a Q]. \tag{12}$$

With diagonal covariance, updates act coordinate-wise. For full covariance $\Sigma \in \mathbb{S}^d_{++}$, the Fisher-natural step simplifies to $\Delta\Sigma = M + M^\top = 2\,\mathrm{sym}(M)$ with $M = \mathbb{E}_{s\sim d^\pi,\, \mathbf{a}\sim\pi_\theta}[\,(\nabla_{\mathbf{a}}Q)(\mathbf{a}-\mu)^\top\,]$. Sampling formulas follow directly from the expectation definitions. The full-covariance update entails $\mathcal{O}(d^3)$ algebra (linear solves), whereas diagonal updates are $\mathcal{O}(d)$. In practice, full covariance helps under strong anisotropy/ill-conditioning; diagonal is preferable when samples or compute are limited.

Full-covariance via Cholesky. Write $\Sigma = LL^\top$ with L lower-triangular and $\mathrm{diag}(L) > 0$. To implement the natural step stably, solve for ΔL in the triangular Sylvester equation

$$L\,\Delta L^\top + \Delta L\, L^\top \;=\; S, \qquad S \;=\; M + M^\top,$$

by forward substitution *by columns* $j = 1, \ldots, d$:

$$(2L_{jj})(\Delta L)_{jj} = S_{jj} - 2\sum_{k<j} L_{jk}(\Delta L)_{jk},$$

$$(\Delta L)_{ij} = \frac{S_{ij} - \sum_{k<j}\big(L_{ik}(\Delta L)_{jk} + (\Delta L)_{ik}L_{jk}\big) - L_{ij}(\Delta L)_{jj}}{L_{jj}}, \quad i = j+1, \ldots, d.$$

Then update $L \leftarrow L + \eta\,\Delta L$ with a line search to keep $\mathrm{diag}(L) > 0$. This realizes $\Delta\Sigma = S$ while preserving $\Sigma \in \mathbb{S}^d_{++}$. In practice, shrink η (backtracking) until $\min_i (L + \eta\Delta L)_{ii} > 0$; this prevents loss of SPD under large steps.

Deterministic limit. As $\Sigma \to 0$, the mean update reduces to the deterministic policy-gradient direction evaluated at the mean, while the covariance update vanishes. The Fisher metric provides a well-defined limit.

Quadratic toy. If locally $\nabla_{\mathbf{a}} Q^\pi(s,\mathbf{a}) = H(s)(\mathbf{a}-\mu)$ with $H(s) \succeq 0$, then $M = H(s)\,\Sigma$ and $\Delta\Sigma = H(s)\,\Sigma + \Sigma\,H(s)$, while $\Delta\mu = 0$ at $\mathbf{a} = \mu$, consistent with the deterministic limit.

4.10 Main Results

We state the primary results under the hypotheses above; proofs are provided in the appendix. Scope: results are stated with fixed critic Q^π and state weighting d^π (semi-gradient convention). When both evolve, additional residual terms appear.

Assumption 1 (Standing hypotheses)**.** *(i) $\log \pi_\theta(\mathbf{a} \mid s) \in C^2$ in $(\mathbf{a}, \theta)$ with mixed partials commuting almost everywhere, and $\nabla_\theta \nabla_{\mathbf{a}} \log \pi_\theta \in L^1(d^\pi \otimes \pi_\theta)$; (ii) $Q^\pi(\cdot, s) \in C^1$ with $\nabla_{\mathbf{a}} Q^\pi \in L^2(\pi_\theta(\cdot \mid s))$ uniformly in s; (iii) either $\mathcal{A} = \mathbb{R}^d$ with vanishing flux at infinity or a no-flux boundary on $\partial\mathcal{A}$; (iv) dominated convergence justifies swapping expectations and derivatives. All energy statements hold in the frozen-critic regime.*

Theorem 2 (Projection to Natural Gradient)**.** *Let $\partial_t \pi$ satisfy* (3) *for $\mathcal{J}[\pi] = -\mathbb{E}[Q^\pi]$. With the inner product* (2)*, the Galerkin orthogonality conditions $\langle \partial_t \pi - \delta\pi_\theta,\ \pi\, \nabla_\theta \log \pi \rangle = 0$ yield the normal equations $F_{\theta\theta}\, \Delta\theta = \mathcal{F}_{t\theta}$ with $\mathcal{F}_{t\theta}$ in* (4)*. Hence $\Delta\theta = F_{\theta\theta}^{-1}\, \mathcal{F}_{t\theta}$.*

Finite steps $\Delta\theta = \eta\, F_{\theta\theta}^{-1}\, \mathcal{F}_{t\theta}$ need not monotonically decrease $\mathcal{J}$ unless η is sufficiently small; a global line search (Armijo/backtracking) can ensure descent. Locally, if $g(\theta) := \mathcal{F}_{t\theta}$ is L-Lipschitz and $F_{\theta\theta} \succeq mI$ in a neighborhood, then choosing $\eta \le m/L$ yields a guaranteed decrease for sufficiently small neighborhoods.

Corollary 1 (Normal equations)**.** *Under Assumption 1, the Fisher–Galerkin projection yields $F_{\theta\theta}\, \Delta\theta = \mathcal{F}_{t\theta}$ with $\mathcal{F}_{t\theta} = \mathbb{E}[\nabla_\theta \nabla_{\mathbf{a}} \log \pi\, \nabla_{\mathbf{a}} Q^\pi]$.*

Proposition 1 (Baseline Invariance)**.** *For any baseline $b(s)$, replacing Q^π by $A^\pi = Q^\pi - b(s)$ leaves both the PDE and the parametric cross term $\mathcal{F}_{t\theta}$ unchanged. Equivalently, the constrained first variation under per-state normalization satisfies $\delta\mathcal{J}/\delta\pi = -\big(Q^\pi - b(s)\big)$.*

Caution: action-dependent adjustments $b(s,\mathbf{a})$ are not baselines in this sense; they alter $\nabla_{\mathbf{a}} Q$ and hence change both the PDE velocity and the projected update.

Theorem 3 (Parameterization Covariance)**.** *Let $\phi = \phi(\theta)$ be a local diffeomorphism with Jacobian $J = \partial\phi/\partial\theta$ and pullback Fisher $F_\phi = J^{-\top} F_\theta J^{-1}$. Define $g_\theta = \mathbb{E}[\nabla_\theta \nabla_{\mathbf{a}} \log \pi\, \nabla_{\mathbf{a}} Q]$ and $g_\phi = J^{-\top} g_\theta$. Then the natural-gradient step is covariant: $\Delta\phi = F_\phi^{-1} g_\phi = J\, F_\theta^{-1} g_\theta = J\, \Delta\theta$.*

Lemma 2 (Gaussian Family)**.** *For a 1D Gaussian policy with fixed variance, $\Delta\mu = \mathbb{E}[\nabla_a Q]$. With log-variance $\lambda = \log\sigma$, $\Delta\lambda = \sigma^{-2}\, \mathbb{E}[(a-\mu)\nabla_a Q]$, i.e., $\Delta(\sigma^2) = 2\,\mathbb{E}[(a-\mu)\nabla_a Q]$. For full covariance Σ, with $M = \mathbb{E}[(\nabla_{\mathbf{a}} Q)(\mathbf{a}-\mu)^\top]$ and $G = \mathrm{sym}(\Sigma^{-1} M \Sigma^{-1})$, the affine-invariant Fisher yields $\Delta\Sigma = 2\,\Sigma G\,\Sigma$, which reduces to $\Delta\Sigma = M + M^\top$.*

Proposition 2 (c-Wasserstein Gradient Flow)**.** *For convex c with conjugate c^*, replacing $\nabla_{\mathbf{a}} Q$ by $\nabla c^*(\nabla_{\mathbf{a}} Q)$ defines the c-Wasserstein gradient flow of $\mathcal{J}$ (in the sense of Ambrosio et al., 2008). Under standard regularity, $\mathcal{J}$ decreases along solutions. The elementwise cube-root arises from $c^*(x) = \frac{3}{4}|x|^{4/3}$.*

Proposition 3 (Alignment with NPG for Gaussian means)**.** *Let $\pi_\theta(\cdot \mid s) = \mathcal{N}(\mu_\theta(s), \Sigma)$ with fixed Σ. Suppose $\nabla_{\mathbf{a}} Q^\pi(s,\mathbf{a}) \approx H(s)\,(\mathbf{a} - \mu_\theta(s))$ in a neighborhood of $\mu_\theta(s)$ with $H(s) \succeq 0$. Then the WPO mean step $\Delta\theta_\mu = F_{\theta_\mu}^{-1}\, \mathbb{E}[\nabla_{\theta_\mu} \nabla_{\mathbf{a}} \log \pi\, \nabla_{\mathbf{a}} Q^\pi]$ and the NPG mean step $F_{\theta_\mu}^{-1}\, \mathbb{E}[\nabla_{\theta_\mu} \log \pi\, A^\pi]$ are collinear for each s, and hence globally after averaging over d^π.*

5 Theoretical Discussion

Equivalence vs. departure. Proposition 3 formalizes alignment with natural policy gradient for Gaussian means with fixed covariance under a local quadratic assumption on Q^π. More generally, in exponential families with sufficient statistics affine in action and compatible parameterizations, the WPO and NPG directions can be collinear per state. Departures arise for non-Gaussian families (e.g., mixtures, tanh-squashed Gaussians) through the mixed derivative $\nabla_\theta \nabla_{\mathbf{a}} \log \pi$, which reflects policy-manifold geometry and can rotate the update relative to NPG.

Non-Gaussian example. For a two-component Gaussian mixture with shared covariance, component responsibilities $\phi_i(\mathbf{a}, s) = \frac{\rho_i \mathcal{N}(\mathbf{a}|\mu_i,\Sigma)}{\sum_k \rho_k \mathcal{N}(\mathbf{a}|\mu_k,\Sigma)}$ enter $\nabla_\theta \nabla_{\mathbf{a}} \log \pi$ and weight the cross-moment $\mathbb{E}[\phi_i(\mathbf{a}, s)\, \nabla_{\mathbf{a}} Q\, (\mathbf{a} - \mu_i)^\top]$. These curvature terms change the step direction even after Fisher preconditioning.

6 Related Work

Policy gradient methods include REINFORCE [Williams, 1992], natural policy gradient [Kakade, 2001, Pascanu and Bengio, 2013], deterministic policy gradients [Silver et al., 2014], and successors such as DDPG [Lillicrap et al., 2015]. Wasserstein geometry has informed optimization and learning [Ambrosio et al., 2008, Benamou and Brenier, 2000], and several works explored Wasserstein in RL: robust formulations [Abdullah et al., 2019], natural gradients [Moskovitz et al., 2020], and policy optimization views [Zhang et al., 2018]. MPO [Abdolmaleki et al., 2018] relates as a policy iteration method with KL-regularized exponentiated targets. Score-based transport (SVGD) [Liu and Wang, 2016] and its policy variants differ in using score-driven particle flows rather than density flows plus projection. For covariance updates, affine-invariant geometry on SPD matrices [Absil et al., 2008] and second-order approximations like K-FAC [Martens and Grosse, 2015] provide complementary perspectives.

7 Limitations

Our energy and dissipation statements rely on the frozen-critic regime and matching state weighting in the projection metric; fully coupled policy–critic dynamics introduce residual terms we do not analyze. Non-smooth architectures may require weak-form interpretations. We do not analyze the variance/bias of mixed-derivative estimators beyond qualitative remarks. The Fisher vs. W_2 parametric projection gap is only closed under affine-in-action exponential families. Extending equivalence results beyond Gaussians and characterizing regularity under which c-Wasserstein squashing preserves optimality are important directions.

8 Conclusion

We clarified the WPO foundation with label-anchored derivations: Fisher–Galerkin projection to a parametric update, baseline invariance, parameterization covariance, c-Wasserstein stability, and Gaussian family updates. Future work includes extending equivalence conditions beyond Gaussians and analyzing stronger regularity requirements for non-smooth architectures.

A Derivations and Assumptions

A.1 Convention (Rescaled Flow)

We work throughout with a per-state time-rescaled flow: the Eulerian velocity uses $v = \nabla_{\mathbf{a}} Q^\pi$ and the functional derivative is $\delta \mathcal{J}/\delta \pi = -Q^\pi$. The state weighting $d^\pi(s)$ enters only as an outer expectation over s. All projection inner products and Fisher expectations use the same $d^\pi(s)$, so the projected direction is invariant under this rescaling. Using the unrescaled convention instead would multiply the fiberwise energy identities by $d^\pi(s)^2$ inside the state integral.

A.2 Energy Dissipation (Proofs)

Weak form. Fix a state s. A curve $t \mapsto \pi_t(\cdot \mid s)$ is a weak solution of the continuity equation $\partial_t \pi = -\nabla_{\mathbf{a}} \cdot (\pi v)$ if, for every $\varphi \in C_c^\infty(\mathcal{A})$, the map $t \mapsto \int \varphi\, \pi_t\, d\mathbf{a}$ is absolutely continuous and

$$\frac{d}{dt} \int \varphi\, \pi_t\, d\mathbf{a} \;=\; \int \nabla_{\mathbf{a}} \varphi(\mathbf{a}) \cdot v_t(\mathbf{a})\, \pi_t(\mathbf{a})\, d\mathbf{a}.$$

Assume either vanishing flux at infinity or a no-flux boundary condition on $\partial \mathcal{A}$.

Function spaces. Assume $Q^\pi(\cdot, s) \in C^1$, $\nabla_{\mathbf{a}} Q^\pi \in L^2(\pi_t(\cdot \mid s))$ uniformly in t, and $\pi_t v_t \in L^1$. Fubini/Tonelli applies to interchange the s and $\mathbf{a}$ integrals under $d^\pi(s)\,\pi_t(\mathbf{a} \mid s)$.

$\mathbf{W}_2$ case. With the rescaled convention $v = \nabla_{\mathbf{a}} Q^\pi$ and $\delta\mathcal{J}/\delta\pi = -Q^\pi$, take $\varphi = \delta\mathcal{J}/\delta\pi$ as a test function and use the chain rule (justified by dominated convergence under the L^2 bounds):

$$\frac{d}{dt}\,\mathcal{J}[\pi_t] = \int \frac{\delta\mathcal{J}}{\delta\pi}\,\partial_t\pi_t\,d\mathbf{a} = -\int \frac{\delta\mathcal{J}}{\delta\pi}\,\nabla_{\mathbf{a}}\cdot(\pi_t\,v_t)\,d\mathbf{a} = \int \nabla_{\mathbf{a}}\Big(\frac{\delta\mathcal{J}}{\delta\pi}\Big)\cdot v_t\,\pi_t\,d\mathbf{a},$$

where the boundary term vanishes. Substituting $\nabla_{\mathbf{a}}(\delta\mathcal{J}/\delta\pi) = -\nabla_{\mathbf{a}} Q^\pi$ and $v_t = \nabla_{\mathbf{a}} Q^\pi$ gives $\frac{d}{dt}\,\mathcal{J}[\pi_t] = -\int \pi_t\,\|\nabla_{\mathbf{a}} Q^\pi\|^2\,d\mathbf{a}$. Taking the expectation over $s \sim d^\pi$ yields Lemma 1.

c-Wasserstein case. Let c be proper, convex, l.s.c., with conjugate c^* differentiable and ∇c^* monotone. With $v_t = \nabla c^*(-\nabla_{\mathbf{a}}\,\delta\mathcal{J}/\delta\pi)$ the same argument yields

$$\frac{d}{dt}\,\mathcal{J}[\pi_t] = \int \nabla_{\mathbf{a}}\Big(\frac{\delta\mathcal{J}}{\delta\pi}\Big)\cdot\nabla c^*\Big(-\nabla_{\mathbf{a}}\Big(\frac{\delta\mathcal{J}}{\delta\pi}\Big)\Big)\,\pi_t\,d\mathbf{a} = -\int \langle \nabla c^*(\nabla_{\mathbf{a}} Q^\pi),\,\nabla_{\mathbf{a}} Q^\pi\rangle\,\pi_t\,d\mathbf{a} \;\le\; 0,$$

with equality only when $\nabla_{\mathbf{a}} Q^\pi = 0$ almost everywhere (or when $\nabla c^*(\cdot) = 0$ at that argument).

A.3 Assumptions and Boundary Conditions

We assume C^1/C^2 smoothness as required and vanishing boundary terms in integration by parts; interchange of expectation and differentiation follows from dominated convergence under integrable bounds. We assume either (i) $\mathcal{A} = \mathbb{R}^d$ with tails making the flux vanish at infinity, or (ii) bounded $\mathcal{A}$ with no-flux boundary condition $(\pi\,v)\cdot n = 0$ on $\partial\mathcal{A}$. We assume $Q^\pi(\cdot, s) \in C^1$ in $\mathbf{a}$ with $\nabla_{\mathbf{a}} Q^\pi$ locally Lipschitz (uniformly in s), ensuring the weak formulation and energy identity.
Sufficient boundary/decay conditions include: (i) bounded $\mathcal{A}$ with $(\pi v)\cdot n = 0$; or (ii) $\mathcal{A} = \mathbb{R}^d$ and $\pi(\mathbf{a} \mid s)\,\|\nabla_{\mathbf{a}} Q^\pi(s, \mathbf{a})\| \to 0$ as $\|\mathbf{a}\| \to \infty$ (uniformly in s).
Scope of descent (frozen critic). Energy decay statements apply to the proxy functional with Q^π and d^π held fixed during the inner flow/projection step (semi-gradient setting). They do not by themselves imply monotone improvement of the true return when the critic or occupancy evolves between steps.

A.4 Fisher–Galerkin Projection

Minimizing the Fisher-weighted squared error between $\partial_t\pi$ and $\delta\pi_\theta$ yields the orthogonality conditions $\langle \partial_t\pi - \delta\pi_\theta,\, \pi\,\nabla_{\theta_k}\log\pi\rangle = 0$ for each coordinate k, i.e., $F_{\theta\theta}\,\Delta\theta = \mathcal{F}_{t\theta}$ with $F_{\theta\theta} = \mathbb{E}[\nabla_\theta \log\pi\,\nabla_\theta\log\pi^\top]$. Using the continuity equation and integrating by parts in $\mathbf{a}$ (boundary terms vanish),

$$\mathcal{F}_{t\theta_k} = \mathbb{E}\Big[\int \nabla_{\theta_k}\log\pi\,\partial_t\pi\,d\mathbf{a}\Big] = -\mathbb{E}\Big[\int \nabla_{\theta_k}\log\pi\,\nabla_{\mathbf{a}}\cdot(\pi\,\nabla_{\mathbf{a}} Q)\,d\mathbf{a}\Big] = \mathbb{E}\Big[\int \big(\nabla_{\mathbf{a}}\nabla_{\theta_k}\log\pi\big)^\top(\pi\,\nabla_{\mathbf{a}} Q)\,d\mathbf{a}\Big],$$

which gives $\mathcal{F}_{t\theta} = \mathbb{E}[\nabla_\theta\nabla_{\mathbf{a}}\log\pi\,\nabla_{\mathbf{a}} Q]$. Commutation of ∇_θ with the integral follows by dominated convergence if $\nabla_\theta\nabla_{\mathbf{a}}\log\pi \in L^1$ uniformly.

A.5 Gáteaux Variation and Baselines

Consider the Lagrangian $\mathcal{L}(\pi, \lambda) = -\mathbb{E}_s\mathbb{E}_{\mathbf{a}\sim\pi(\cdot|s)}[Q^\pi(s, \mathbf{a})] + \mathbb{E}_s\big[\lambda(s)(\int \pi(\mathbf{a} \mid s)\,d\mathbf{a} - 1)\big]$. The first variation in a direction h with $\int h\,d\mathbf{a} = 0$ per state is $\delta\mathcal{L}(\pi; h) = -\mathbb{E}_s\int Q^\pi(s, \mathbf{a})\,h(\mathbf{a}, s)\,d\mathbf{a}$. Stationarity yields $\delta\mathcal{J}/\delta\pi = -Q^\pi + \lambda(s)$. Replacing Q^π by $Q^\pi - b(s)$ shifts $\lambda(s)$ to $\lambda(s) + b(s)$; since $\nabla_{\mathbf{a}} b(s) = 0$, both the Eulerian velocity and the cross term $\mathcal{F}_{t\theta}$ are unchanged.

A.6 Gaussian Updates

For 1D Gaussian, $\partial_\mu \log\pi = (a-\mu)/\sigma^2$ and $\partial_\lambda\log\pi = (a-\mu)^2/\sigma^2 - 1$ with $\lambda = \log\sigma$. Fisher blocks are $F_{\mu\mu} = 1/\sigma^2$, $F_{\lambda\lambda} = 2$, $F_{\mu\lambda} = 0$. The natural step gives $\Delta\mu = \mathbb{E}[\nabla_a Q]$ and $\Delta\lambda = \frac{1}{\sigma^2}\mathbb{E}[(a-\mu)\nabla_a Q]$.

For full covariance, work on SPD(d) with the affine-invariant inner product $\langle U, V\rangle_\Sigma = \mathrm{tr}(\Sigma^{-1}U\,\Sigma^{-1}V)$. Let $M = \mathbb{E}[(\nabla_{\mathbf{a}} Q)(\mathbf{a}-\mu)^\top]$. A standard calculation gives the Riemannian

gradient $\operatorname{grad}\mathcal{J}(\Sigma) = \operatorname{sym}(\Sigma^{-1}M\,\Sigma^{-1})$. The natural step is $\Delta\Sigma = -\alpha\operatorname{grad}\mathcal{J}$ pulled back to Euclidean coordinates, i.e.,

$$\Delta\Sigma = 2\,\Sigma\,\operatorname{sym}(\Sigma^{-1}M\,\Sigma^{-1})\,\Sigma = M + M^\top,$$

which matches the Fisher-preconditioned update.

A.7 Cholesky Triangular Sylvester Solve

Linearizing $\Sigma = LL^\top$ yields $\Delta\Sigma = L\,\Delta L^\top + \Delta L\,L^\top$. Given a symmetric target S, the lower-triangular recursion stated in the main text solves $L\,\Delta L^\top + \Delta L\,L^\top = S$ uniquely by forward substitution (prove by induction on columns). Hence the realized increment equals $S = M + M^\top$. For SPD preservation, a sufficient condition is to backtrack η until $\min_i(L + \eta\Delta L)_{ii} > 0$; for example, if $\eta\,\|L^{-1}\Delta L\|_\infty < 1$, the updated diagonal remains positive.

Algorithm (Cholesky SPD update). Given L and a symmetric target $S = M + M^\top$:

1. Solve $L\,\Delta L^\top + \Delta L\,L^\top = S$ for lower-triangular ΔL by forward substitution (column-wise recursion above).
2. Line search $\eta > 0$ (e.g., backtracking Armijo) until $\min_i(L + \eta\Delta L)_{ii} > 0$.
3. Update $L \leftarrow L + \eta\Delta L$ and return $\Sigma \leftarrow LL^\top$.

Cost: $\mathcal{O}(d^3)$ per update; diagonal Σ reduces to $\mathcal{O}(d)$.

A.8 Parameterization Covariance

Let $\phi = \phi(\theta)$ be a local diffeomorphism with Jacobian $J = \partial\phi/\partial\theta$. The Fisher transforms as $F_\phi = \mathbb{E}[\nabla_\phi \log\pi\,\nabla_\phi \log\pi^\top] = J^{-\top}F_\theta J^{-1}$. The cross term obeys $g_\phi = \mathbb{E}[\nabla_\phi\nabla_{\mathbf{a}}\log\pi\,\nabla_{\mathbf{a}}Q] = J^{-\top}g_\theta$ by the chain rule $\nabla_\phi = J^{-\top}\nabla_\theta$. Therefore $\Delta\phi = F_\phi^{-1}g_\phi = J\,F_\theta^{-1}g_\theta = J\,\Delta\theta$.

A.9 c-Wasserstein

Assume c is proper, convex, l.s.c., with conjugate c^* differentiable and ∇c^* monotone. The c-Wasserstein flow uses velocity $\nabla c^*(-\nabla_{\mathbf{a}}\delta\mathcal{J}/\delta\pi)$. The elementwise cube-root mapping used in the main text arises from the *separable* choice $c(u) = \frac{1}{4}\sum_i |u_i|^4$, whose conjugate has $c^*(x) = \frac{3}{4}\sum_i |x_i|^{4/3}$ and $\nabla c^*(x)_i = \operatorname{sign}(x_i)|x_i|^{1/3}$. This map is direction-preserving and Hölder-continuous, supporting the stated stability. A rotationally invariant alternative $c(u) = \frac{1}{4}\|u\|_2^4$ yields $\nabla c^*(x) = \|x\|_2^{-2/3}x$.

B Proofs of Main Results

Proof of Theorem 2. Projecting $\partial_t\pi$ onto $\delta\pi_\theta$ in the Fisher inner product yields normal equations $\langle\partial_t\pi - \delta\pi_\theta,\, \pi\,\nabla_\theta\log\pi\rangle = 0$, which rearrange to $F_{\theta\theta}\,\Delta\theta = \mathcal{F}_{t\theta}$. Integration by parts under the stated regularity gives the cross term in (4). □

Proof of Proposition 1. Constrained variation with per-state normalization introduces a Lagrange multiplier $\lambda(s)$, yielding $\delta\mathcal{J}/\delta\pi = -Q^\pi + \lambda(s)$. Any baseline $b(s)$ can be absorbed into λ, so both the PDE and $\mathcal{F}_{t\theta}$ are invariant to $Q^\pi \mapsto Q^\pi - b(s)$. □

Proof of Theorem 3. Under the pullback metric, $F_\phi^{-1} = J\,F_\theta^{-1}J^\top$ and $g_\phi = J^{-\top}g_\theta$. Thus $F_\phi^{-1}g_\phi = J\,F_\theta^{-1}g_\theta = J\,\Delta\theta$. □

Proof of Lemma 2. Use $\nabla_a\log\pi = -(a-\mu)/\sigma^2$ and $\nabla_\theta\nabla_a\log\pi = (\nabla_\theta\mu)/\sigma^2$ to compute g_θ and Fisher blocks; apply $\Delta\theta = F^{-1}g$ to obtain the stated updates. For full Σ, the affine-invariant Fisher yields $\Delta\Sigma = M + M^\top$. □

Proof of Proposition 2. For convex c, c^* is convex and ∇c^* is monotone. Replacing $\nabla_{\mathbf{a}} Q$ by $\nabla c^*(\nabla_{\mathbf{a}} Q)$ corresponds to the Eulerian form of the c-Wasserstein gradient flow, preserving descent/ascent while taming large gradients. □

Proof of Proposition 3. Proof sketch. For fixed covariance, the mean-block Fisher is $F_{\theta_\mu \theta_\mu} = \mathbb{E}[\nabla_{\theta_\mu} \log \pi \, \nabla_{\theta_\mu} \log \pi^\top]$. Under the local quadratic model $\nabla_{\mathbf{a}} Q^\pi \approx H(s)(\mathbf{a} - \mu)$ and Gaussian π, one has $\mathbb{E}[\nabla_{\mathbf{a}} Q^\pi \, (\mathbf{a} - \mu)^\top] = H(s)\, \Sigma$. Using $\nabla_{\theta_\mu} \nabla_{\mathbf{a}} \log \pi = (\partial \mu / \partial \theta_\mu)\, \Sigma^{-1}$, the WPO cross term becomes $g_{\theta_\mu} = \mathbb{E}[(\partial \mu / \partial \theta_\mu)\, \Sigma^{-1} H(s)\, \Sigma]$. The NPG mean direction with advantage A^π linearized around μ yields the same factor $H(s)$ multiplying $\nabla_{\theta_\mu} \log \pi$, so after preconditioning by $F^{-1}_{\theta_\mu \theta_\mu}$ both steps are collinear. Averaging over d^π preserves collinearity. □

Acknowledgments Omitted for double-blind review.

Responsible AI Statement

This is a theoretical contribution. It contains no human or animal subjects, no personal or sensitive data, and no deployed systems. All results are formal statements with explicit assumptions and proofs; we discuss limitations, scope, and failure modes where applicable. The work adheres to the Agents4Science Code of Ethics (and the NeurIPS Code of Ethics in spirit): we avoid prohibited practices, dual-use concerns, and undisclosed human subject data; environmental impact is negligible as no large-scale compute is used.

Reproducibility Statement

All claims are formal theorems, propositions, or lemmas, each with clearly stated assumptions and proofs. Key equations are label-anchored for unambiguous cross-referencing (e.g., (3), (1), (2), (4)). Weak-form and boundary assumptions are made explicit; derivations are given both in the main text and in the appendix. No datasets or empirical experiments are involved. A minimal prototype used in separate notes mirrors the Gaussian mean/covariance updates but is not required to verify the theoretical results.

References

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018.

Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian Zhang, and Jun Wang. Wasserstein robust reinforcement learning, 2019.

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2008.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, 2001.

Timothy Lillicrap et al. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

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

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*. PMLR, 2015.

Ted Moskovitz, Michael Arbel, Ferenc Huszar, and Arthur Gretton. Efficient wasserstein natural gradients for reinforcement learning. *arXiv preprint arXiv:2010.05380*, 2020.

Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, pages 387–395. PMLR, 2014.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy optimization as wasserstein gradient flows. In *International Conference on Machine Learning*, pages 5737–5746. PMLR, 2018.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[A]**

   Explanation: The theoretical questions, hypotheses, and scope (projection of Wasserstein flows, invariance properties, and c-Wasserstein stability) were developed by the authors without AI assistance beyond standard search/reading tools.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[NA]**

   Explanation: The paper presents mathematical derivations and proofs only; there are no empirical experiments or released implementations.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[NA]**

   Explanation: No datasets were used; results are theoretical statements with formal proofs and stated assumptions.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[A]**

   Explanation: Writing and editing were performed by the authors using LaTeX. No generative AI systems were used to draft or edit the scientific content.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: Not applicable (no AI systems were used in ideation, analysis, coding, or writing).

Agents4Science Paper Checklist

1. **Claims**

   Question: Are the claims made in the abstract and introduction supported by the results?

   Answer: [Yes]

   Justification: Claims (projection to natural gradient, baseline invariance, parameterization covariance, Gaussian/covariance updates, and c-Wasserstein energy decay) are supported by Theorems/Propositions with proofs and stated assumptions (see Theorem 2, Proposition 1, Theorem 3, Lemma 2, Theorem 1).

2. **Tasks and baselines**

   Question: Did you describe the limitations of your work?

   Answer: [Yes]

   Justification: A dedicated Limitations section enumerates scope and assumptions (e.g., frozen-critic regime, smoothness/weak-form conditions, and when equivalences fail).

   Question: Did you discuss any potential negative societal impacts of your work?

   Answer: [NA]

   Justification: The work is purely theoretical and does not involve deployment, datasets, or application domains with direct societal impact.

   Question: Did you discuss the failure modes of your method?

   Answer: [Yes]

   Justification: The paper discusses conditions where alignment with NPG holds vs. fails (e.g., mixtures/squashing), boundary/regularity requirements, and stability caveats beyond the frozen-critic setting.

3. **Reproducibility**

   Question: Is your code and data publicly available?

   Answer: [NA]

   Justification: No code or datasets were produced; the paper contains only theoretical results.

   Question: Does the paper provide sufficient details to reproduce the main results (either in the text or as supplemental material)?

   Answer: [Yes]

   Justification: Full derivations with assumptions, weak-form statements, and proofs are provided in the main text and appendix.

4. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Not applicable; there are no experiments.

5. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Not applicable; there are no experiments.

6. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Not applicable; there are no experiments.

7. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

Justification: Theoretical work with no human/animal subjects, sensitive data, or prohibited practices.

8. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Not applicable to this purely theoretical contribution; no deployment or application domain is studied.