Policy Optimization as a Gradient Flow in Offline Reinforcement Learning: Supplementary Materials

1 Proofs

1.1 Convergence of Gradient Flows

Suppose $Q \in L^1_{\text{loc}}$ and that optimal policy $\pi^\star(a|s) = \exp(Q(s,a)/\lambda)/Z(s)$ is well-defined. Then we define:

$$\mathcal{L}[\pi] := \mathcal{E}_s[\pi] - \mathcal{E}_s[\pi^\star] = \lambda D_{\text{KL}}(\pi||\pi^\star) \geq 0. \tag{1}$$

1.1.1 Fisher-Rao gradient flow

We first have:

$$\frac{d}{dt}\mathcal{L}[\pi_t] = \int \frac{\delta \mathcal{E}_s}{\delta \pi}[\pi_t]\frac{\partial \pi_t}{\partial t} da = -\| \operatorname{grad}_{\text{FR}} \mathcal{E}_s[\pi_t]\|^2_{\text{FR}} \leq 0. \tag{2}$$

Since $\mathcal{L}$ is locally FR-strongly convex around the optimal policy $\pi^\star$, there exist constants $r > 0$ and $\mu > 0$ such that

$$\| \operatorname{grad}_{\text{FR}} \mathcal{L}(\pi)\|^2_{\text{FR}} = \| \operatorname{grad}_{\text{FR}} \mathcal{E}_s(\pi)\|^2_{\text{FR}} \;\geq\; 2\mu \left(\mathcal{L}[\pi] - \mathcal{L}[\pi^\star]\right) = 2\mu\, \mathcal{L}[\pi],$$

for all $\pi \in B_{\text{FR}}(\pi^\star, r)$, where $B_{\text{FR}}(\pi^\star, r)$ denotes the open ball of radius r centered at $\pi^\star$ under the Fisher-Rao metric. If the initial policy satisfies $\pi_0 \in B_{\text{FR}}(\pi^\star, r)$, the trajectory remains in the ball, and we obtain:

$$\frac{d}{dt} D_{\text{KL}}(\pi_t\|\pi^\star) = -\frac{1}{\lambda}\| \operatorname{grad}_{\text{FR}} \mathcal{E}(\pi_t)\|^2_{\text{FR}} \leq -\frac{2\mu}{\lambda}\, D_{\text{KL}}(\pi_t\|\pi^\star).$$

Therefore, $D_{\text{KL}}(\pi_t\|\pi^\star)$ decays exponentially over time, establishing the exponential convergence of the FR gradient flow toward the optimal policy.

1.1.2 Wasserstein gradient flow

Defining the relative Fisher information

$$\mathcal{I}(\mu\|\nu) := \int \left|\nabla \log \frac{d\mu}{d\nu}\right|^2 d\mu,$$

we have:

$$\frac{d}{dt}\mathcal{L}[\pi_t] = -\lambda \mathcal{I}(\pi_t\|\pi^\star) \leq 0.$$

Let $\mathcal{W}_2(\pi_\beta, \pi^\star) = R$ and $B_R(\pi^\star) := \{\pi : \mathcal{W}_2(\pi, \pi^\star) < R\}$. Assume that $\pi^\star$ satisfies the log-Sobolev inequality (LSI) locally in $B_R(\pi^\star)$, i.e.,

$$\mathcal{I}(\pi\|\pi^\star) \geq 2\kappa D_{\text{KL}}(\pi\|\pi^\star), \quad \forall \pi \in B_R(\pi^\star),$$

for some $\kappa > 0$. Choosing r so that the KL-sublevel set corresponding to $\text{KL}(\pi_\beta\|\pi^\star)$ is contained in $B_r(\pi^\star)$, the flow (π_t) remains in $B_r(\pi^\star)$, and we obtain:

$$\frac{d}{dt} D_{\text{KL}}(\pi_t\|\pi^\star) \leq -2\kappa\lambda\, D_{\text{KL}}(\pi_t\|\pi^\star) \;\Rightarrow\; D_{\text{KL}}(\pi_t\|\pi^\star) \leq e^{-2\kappa\lambda t} D_{\text{KL}}(\pi_\beta\|\pi^\star).$$

1.2 Wasserstein–2 distance for Gaussian distributions

For Gaussian distributions $\pi_i = \mathcal{N}(\mu_i, \Sigma_i)$, the squared Wasserstein-2 distance is

$$\mathcal{W}_2^2(\pi_1, \pi_2) = \|\mu_1 - \mu_2\|_2^2 \; + \; \mathrm{Tr}\Big(\Sigma_1 + \Sigma_2 - 2\big(\Sigma_2^{1/2}\,\Sigma_1\,\Sigma_2^{1/2}\big)^{1/2}\Big). \tag{3}$$

The second term is the Bures metric between the covariance matrices. Computing the matrix square root can be a computational bottleneck in high-dimensional settings. Since Σ_1 and Σ_2 commute in our modeling, the expression simplifies considerably:

$$\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2} = \mathrm{diag}(\sigma_2)\,\mathrm{diag}(\sigma_1^2)\,\mathrm{diag}(\sigma_2) = \mathrm{diag}(\sigma_1^2\sigma_2^2).$$

The trace term simplifies coordinate-wise:

$$\mathrm{Tr}\Big(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}\Big) = \sum_{j=1}^{d}\Big(\sigma_{1j}^2 + \sigma_{2j}^2 - 2\sigma_{1j}\sigma_{2j}\Big) = \|\sigma_1 - \sigma_2\|_2^2.$$

Plugging back into (3) yields

$$\mathcal{W}_2^2(\pi_1, \pi_2) = \|\mu_1 - \mu_2\|_2^2 \; + \; \|\sigma_1 - \sigma_2\|_2^2. \tag{4}$$

This measures displacement (means) and dispersion (standard deviations). It is $\mathcal{O}(d)$ to evaluate and remains finite under partial support mismatch. This completes the proof of Lemma 2.

1.3 Fisher–Rao distance for Gaussian distributions

The Fisher–Rao distance arises from the square-root density embedding $\psi : p \mapsto 2\sqrt{p}$, which maps probability densities to the unit sphere in L^2. The geodesic distance between two densities p and q is given by the angle between their embeddings:

$$d_{\mathrm{FR}}(p, q) \;=\; 2\,\arccos\Big(\,\langle\sqrt{p}, \sqrt{q}\rangle_{L^2}\,\Big) \;=\; 2\,\arccos\Big(\int \sqrt{p(a)\,q(a)}\,da\Big). \tag{5}$$

The inner product term is the *Bhattacharyya coefficient* (BC),

$$\mathrm{BC}(p, q) \;=\; \int \sqrt{p(a)\,q(a)}\,da, \tag{6}$$

which takes values in $[0, 1]$ and measures the affinity between distributions. The Bhattacharyya distance is defined as

$$D_B(p, q) \;=\; -\log \mathrm{BC}(p, q), \tag{7}$$

so that the Fisher–Rao distance can equivalently be expressed as

$$d_{\mathrm{FR}}(p, q) \;=\; 2\,\arccos\big(\mathrm{BC}(p, q)\big), \qquad \mathrm{BC}(p, q) = e^{-D_B(p,q)}. \tag{8}$$

For multivariate Gaussians $\pi_i = \mathcal{N}(\mu_i, \Sigma_i)$, $i \in \{1, 2\}$, the Bhattacharyya distance admits the well-known closed form:

$$D_B(\pi_1, \pi_2) = \tfrac{1}{8}(\mu_1 - \mu_2)^\top\Sigma^{-1}(\mu_1 - \mu_2) \; + \; \tfrac{1}{2}\log\frac{\det\Sigma}{\sqrt{\det\Sigma_1\,\det\Sigma_2}}, \tag{9}$$

$$\begin{aligned}\mathrm{BC}(\pi_1, \pi_2) &= \exp\big(-D_B(\pi_1, \pi_2)\big)\\ &= \frac{(\det\Sigma_1)^{1/4}(\det\Sigma_2)^{1/4}}{(\det\Sigma)^{1/2}}\;\exp\big(-\tfrac{1}{8}(\mu_1 - \mu_2)^\top\Sigma^{-1}(\mu_1 - \mu_2)\big)\,,\end{aligned} \tag{10}$$

where $\Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)$ is the average covariance. Substituting (10) into (8) yields the Fisher–Rao distance.

When $\Sigma_i = \mathrm{diag}(\sigma_i^2)$, the expressions factorize dimension-wise. Let μ_{ij} and σ_{ij} denote the j-th components of the mean and standard deviation. Then

$$D_B(\pi_1, \pi_2) = \sum_{j=1}^{d} \left[\frac{(\mu_{1j} - \mu_{2j})^2}{4(\sigma_{1j}^2 + \sigma_{2j}^2)} \; + \; \frac{1}{2} \log\Big(\frac{\sigma_{1j}^2 + \sigma_{2j}^2}{2\sigma_{1j}\sigma_{2j}} \Big) \right], \tag{11}$$

$$\mathrm{BC}(\pi_1, \pi_2) = \prod_{j=1}^{d} \left(\frac{2\,\sigma_{1j}\sigma_{2j}}{\sigma_{1j}^2 + \sigma_{2j}^2} \right)^{1/2} \exp\left(-\frac{(\mu_{1j} - \mu_{2j})^2}{4(\sigma_{1j}^2 + \sigma_{2j}^2)} \right), \tag{12}$$

$$d_{\mathrm{FR}}(\pi_1, \pi_2) = 2 \arccos\big(\mathrm{BC}(\pi_1, \pi_2)\big). \tag{13}$$

This diagonal specialization is computationally efficient and aligns with Gaussian policies commonly used in reinforcement learning.

1.4 KL divergence approximation of FR distance

Let $\sigma \in T_{\pi^\star}\mathcal{P}_+^\infty$ and define

$$\pi = \mathrm{Exp}_{\pi^\star}(\sigma), \qquad \delta := \|\sigma\|_{\mathrm{FR},\pi^\star}.$$

Set $F(\pi) := 2\,\mathrm{KL}(\pi\|\pi^\star)$. Then

$$\nabla_{\mathrm{FR}} F(\pi^\star) = 0, \qquad \mathrm{Hess}_{\mathrm{FR}} F(\pi^\star) = 2\, g_{\mathrm{FR},\pi^\star},$$

where $\mathrm{Hess}_{\mathrm{FR}}$ denotes the Riemannian Hessian on FR manifold. By Taylor's theorem in FR normal coordinates,

$$F(\pi) = F(\pi^\star) + \tfrac{1}{2}\,\mathrm{Hess}_{\mathrm{FR}} F(\pi^\star)[\sigma, \sigma] + R_3 = \delta^2 + R_3,$$

where $|R_3| \le C\,\delta^3$ for some $C > 0$. Writing $\varepsilon = \delta^2$ gives

$$2\,D_{\mathrm{KL}}(\pi\|\pi^\star) = d_{\mathrm{FR}}^2(\pi, \pi^\star) + O\big(d_{\mathrm{FR}}^3(\pi, \pi^\star)\big) = \varepsilon + O(\varepsilon^{3/2}).$$

1.5 Natural Gradient on Parameter Space

We derive the natural gradient that results from one JKO step for the policy manifold under either the Wasserstein–2 or Fisher–Rao geometry, following the efficient Wasserstein–natural gradient viewpoint of and the information-geometric treatment of Fisher-Rao.

Let $\mathcal{E}_s[\pi]$ denote the energy functional, and consider a Gaussian policy

$$\pi \sim \mathcal{N}(\mu(s),\, \Sigma(s) = \mathrm{diag}(\sigma(s)))\,, \qquad \mu, \sigma \in \mathbb{R}^d,$$

where the covariance matrix is diagonal, i.e., $\Sigma(s) = \mathrm{diag}(\sigma_1^2(s), \ldots, \sigma_d^2(s))$. We define the coordinate mapping

$$\pi \;\mapsto\; \eta = (\mu,\, \sigma) \in \mathbb{R}^{2d},$$

where each $\sigma_i > 0$ is the standard deviation corresponding to the i-th dimension. For a reference policy $\bar{\pi}$ with coordinates

$$\bar{\eta} = (\mu + \Delta\mu,\, \sigma + \Delta\sigma),$$

the 2-Wasserstein distance between π and $\bar{\pi}$ is given by

$$\mathcal{W}_2^2(\pi, \bar{\pi}) = \|\Delta\mu\|_2^2 + \|\Delta\sigma\|_2^2.$$

Hence, the induced Riemannian metric tensor in the (μ, σ) coordinate system is

$$G_{\mathcal{W}}(\eta) = I_{2d}, \tag{14}$$

where I_{2d} denotes the $2d \times 2d$ identity matrix.

Alternatively, let us introduce the logarithmic coordinate

$$\rho_i = \log \sigma_i, \qquad H = (\mu, \rho),$$

so that $\sigma_i = e^{\rho_i}$ and $\Sigma = \mathrm{diag}(e^{2\rho_1}, \ldots, e^{2\rho_d})$. The coordinate transformation between $\eta = (\mu, \sigma)$ and $H = (\mu, \rho)$ is given by $d\sigma_i = e^{\rho_i}\, d\rho_i$, and hence the Jacobian matrix is

$$J = \frac{\partial(\mu, \sigma)}{\partial(\mu, \rho)} = \begin{bmatrix} I_d & 0 \\ 0 & \mathrm{diag}(e^{\rho_1}, \ldots, e^{\rho_d}) \end{bmatrix} = \begin{bmatrix} I_d & 0 \\ 0 & \Sigma^{1/2} \end{bmatrix}$$

The metric tensor transforms accordingly as

$$G_{\mathcal{W}}(H) = J^\top G_{\mathcal{W}}(\eta)\, J = \begin{bmatrix} I_d & 0 \\ 0 & \Sigma(s) \end{bmatrix}. \tag{15}$$

For the Fisher–Rao case, the metric tensor is defined by the Fisher information matrix:

$$G_{\mathrm{FR}}(\eta) = \mathbb{E}_{a\sim\pi}\big[\nabla_\eta \log \pi(a|s)\, \nabla_\eta \log \pi(a|s)^\top\big], \qquad \eta = (\mu, \sigma).$$

For a diagonal Gaussian policy $\pi(a|s) = \mathcal{N}\big(a; \mu(s), \Sigma(s)\big)$ with $\Sigma(s) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, the score functions are

$$\frac{\partial \log \pi}{\partial \mu_i} = \frac{a_i - \mu_i}{\sigma_i^2}, \qquad \frac{\partial \log \pi}{\partial \sigma_i} = -\frac{1}{\sigma_i} + \frac{(a_i - \mu_i)^2}{\sigma_i^3}.$$

Using the standardized variable $Z_i = (a_i - \mu_i)/\sigma_i \sim \mathcal{N}(0, 1)$, we have

$$\frac{\partial \log \pi}{\partial \mu_i} = \frac{Z_i}{\sigma_i}, \qquad \frac{\partial \log \pi}{\partial \sigma_i} = \frac{1}{\sigma_i}(Z_i^2 - 1).$$

Taking expectations over Z_i, and using $\mathbb{E}[Z_i] = 0$, $\mathbb{E}[Z_i^2] = 1$, $\mathbb{E}[Z_i^4] = 3$, the Fisher information matrix becomes diagonal:

$$G_{\mathrm{FR}}(\eta) = \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 2\,\Sigma^{-1} \end{bmatrix}. \tag{16}$$

Similarly, the standard deviation coordinate system can be transformed to the logarithmic coordinate system as:

$$G_{\mathrm{FR}}(H) = J^\top G_{\mathrm{FR}}(\eta)\, J = \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 2\, I_d \end{bmatrix}. \tag{17}$$

Therefore, the natural output gradients under logarithmic coordinate system on each manifold are expressed with the Euclidean gradient ∇ as:

$$\text{Fisher–Rao:} \qquad \tilde{\nabla}_\mu^{\mathrm{FR}} = \Sigma \nabla_\mu, \qquad \tilde{\nabla}_\rho^{\mathrm{FR}} = \tfrac{1}{2}\nabla_\rho, \tag{18}$$

$$\text{Wasserstein–2:} \qquad \tilde{\nabla}_\mu^{\mathcal{W}} = \nabla_\mu, \qquad \tilde{\nabla}_\rho^{\mathcal{W}} = \Sigma^{-1}\, \nabla_\rho. \tag{19}$$

Equations (18)–(19) are realized in autodiff by scaling the backpropagated gradients at the policy outputs (statewise): for FR, multiply the gradient with respect to μ by σ^2 and with respect to ρ by $1/2$; for $\mathcal{W}_2$, leave the μ gradient unchanged and multiply the ρ gradient by $1/\sigma^2$.

2 Implementation Details

Our implementation builds upon the TD3+BC backbone with minimal modifications. The main differences are: (i) a *stochastic policy* as a diagonal Gaussian with action-space clipping, instead of the deterministic actor used in TD3+BC, and (ii) the addition of a proximal term and an entropy regularizer in the actor objective. The actor network consists of two fully connected layers with 256 hidden units and ReLU activations, while each critic employs three layers of the same width. The actor outputs the action parameters: a mean passed through $\tanh(\cdot)$ to enforce bounds, and a log-standard-deviation vector. Target networks are maintained for both the actor and the critics, and are updated softly with a rate of $\tau = 0.005$. All models are optimized using Adam with a batch size of 256.

Table 1: Implementation details of POGO.

Parameter	Value
backbone	TD3+BC framework
policy type	Stochastic actor (diagonal Gaussian)
optimizer	Adam (Kingma & Ba, 2014)
batch size	256
learning rate	environment-specific tuning in Table 2
tau (τ)	0.005
hidden dim (all networks)	256
number of hidden layers	actor: 2, critic: 3
nonlinearity	ReLU
update schedule	actor updated once every two critic updates (TD3 rule)
training steps	1M gradient steps
discount factor (γ)	0.99
dataset	D4RL v2 (Gym-MuJoCo, AntMaze)
two-phase training	Phase 1: joint actor–critic; Phase 2: critic frozen, actor continues

Following the TD3 update rule, the actor is updated once every two critic updates, and each run is trained for one million gradient steps. The discount factor is set to $\gamma = 0.99$. All experiments are conducted on the D4RL v2 offline datasets. When the two-phase training schedule is used, the process is evenly split with a ratio of 0.5: in Phase 1, both the actor and critics are jointly updated; in Phase 2, the critics are frozen while the actor continues to optimize under the same configuration as in Phase 1. Environment-specific coefficients for the Wasserstein-2 proximal weight, entropy weight, and learning rates are summarized in Table 2.

Table 2: Hyperparameter settings for POGO-W

Task Name	w2 weight	entropy weight	learning rate
halfcheetah-medium	0.05	5e-3	3e-4
halfcheetah-medium-replay	0.06	1e-5	3e-4
halfcheetah-medium-expert	2.0	1e-4	3e-4
hopper-medium	0.2	1e-5	3e-4
hopper-medium-replay	0.1	1e-5	3e-4
hopper-medium-expert	0.5	1e-5	3e-4
walker2d-medium	0.3	5e-3	3e-4
walker2d-medium-replay	0.2	1e-3	3e-4
walker2d-medium-expert	0.3	1e-2	3e-4
antmaze-umaze-v2	0.9	0.0	1e-4
antmaze-umaze-diverse-v2	0.5	1e-3	1e-4
antmaze-medium-play-v2	0.2	3e-3	1e-4
antmaze-medium-diverse-v2	0.06	1e-3	1e-4
antmaze-large-play-v2	0.1	6e-3	1e-4
antmaze-large-diverse-v2	0.06	1e-4	1e-4

3 ADDITIONAL EXPERIMENTS

Table 3 summarizes the normalized returns on the D4RL AntMaze benchmark. Despite being built upon the simpler TD3+BC backbone, POGO–W achieves competitive performance compared to more sophisticated algorithms such as IQL and ReBRAC. Notably, on smaller navigation tasks such as `antmaze-umaze-v2`, POGO–W not only outperforms IQL but also attains scores comparable to ReBRAC, demonstrating that the proposed Wasserstein proximal formulation can substantially improve exploration and stability even without the architectural or regularization enhancements of recent state-of-the-art methods. This represents a notable improvement, given that vanilla TD3+BC typically collapses to near-zero performance in most AntMaze domains except the easiest `umaze` tasks. In some cases—particularly in the more complex AntMaze environments—the two-step

update yielded limited gains, which can be attributed to the structural limitations of the TD3+BC critic, where imperfect bootstrapping can lead to unstable or biased value targets.

Table 3: Performance comparisons on the D4RL AntMaze benchmark

Task Name	**IQL**	**TD3+BC**	**POGO–W**^{1}	**ReBRAC**	**POGO–W**^{2}
antmaze-umaze-v2	68.00 ± 16.00	72.00 ± 37.09	92.40 ± 4.16	97.80 ± 1.17	83.60 ± 8.60
antmaze-umaze-diverse-v2	60.00 ± 8.94	30.0 ± 14.97	51.60 ± 31.66	87.00 ± 2.15	35.60 ± 28.07
antmaze-medium-play-v2	70.00 ± 12.65	0.00 ± 0.00	26.00 ± 14.55	90.20 ± 2.48	33.60 ± 21.53
antmaze-medium-diverse-v2	68.00 ± 17.21	4.00 ± 4.92	48.40 ± 11.89	77.37 ± 17.69	25.20 ± 14.89
antmaze-large-play-v2	42.00 ± 4.00	0.00 ± 0.00	30.40 ± 9.87	43.78 ± 0.07	15.20 ± 12.70
antmaze-large-diverse-v2	22.00 ± 7.48	0.00 ± 0.00	22.80 ± 11.71	43.81 ± 0.06	6.40 ± 6.65
AntMaze Average	55.00	17.67	35.80	73.66	33.27

Finally, we applied the same first-order formulation to incorporate the Wasserstein natural gradient, using the metric tensor derived in Supplementary Section 1. Empirically, this variant (POGO–WNG) achieves performance comparable to or slightly higher than the standard Wasserstein formulation across most environments (Table 4).

Table 4: Evaluation of POGO–FR and POGO–W with natural gradients on Mujoco environments

Task Name	**POGO–FR**^{1}	**POGO–W**^{1}	**POGO–W**^{2}	**POGO–WNG**^{1}	**POGO–WNG**^{2}
halfcheetah-medium	42.34 ± 0.66	58.57 ± 0.66	61.91 ± 1.25	59.48 ± 1.14	62.42 ± 1.54
halfcheetah-medium-replay	40.84 ± 6.35	52.33 ± 1.34	53.75 ± 1.32	52.02 ± 0.86	52.32 ± 2.93
halfcheetah-medium-expert	46.13 ± 1.79	87.00 ± 4.44	79.27 ± 8.41	78.50 ± 2.90	81.63 ± 7.09
hopper-medium	52.25 ± 3.64	64.71 ± 4.87	74.78 ± 11.19	61.98 ± 2.99	75.81 ± 7.47
hopper-medium-replay	49.62 ± 4.84	85.68 ± 15.56	99.34 ± 1.69	88.15 ± 10.15	100.31 ± 2.45
hopper-medium-expert	76.96 ± 20.04	100.65 ± 5.20	93.67 ± 11.98	100.39 ± 3.52	92.82 ± 13.18
walker2d-medium	72.99 ± 2.45	85.69 ± 1.61	86.70 ± 1.65	85.74 ± 0.63	86.31 ± 1.89
walker2d-medium-replay	12.31 ± 1.62	91.78 ± 1.74	91.61 ± 1.72	91.22 ± 1.87	91.97 ± 2.76
walker2d-medium-expert	96.05 ± 10.74	110.93 ± 0.37	111.92 ± 0.51	110.78 ± 0.43	112.10 ± 0.62
MuJoCo Average	54.39	81.96	83.66	80.94	83.95

3.1 First-order analysis of one step updates

We derive the first-order approximation of the one-step updates for the gradient flows under the Wasserstein and Fisher–Rao (FR) geometries. Fix a state s and expand Q at μ_β up to second order. The local approximation of the energy functional becomes

$$\mathcal{E}_s[\pi] = -\mathrm{G}^\mathsf{T}(\mu - \mu_\beta) - \frac{1}{2}\mathrm{tr}(\mathrm{H}\Sigma) - \frac{\lambda}{2}\log\det\Sigma + \text{const.}$$

where $\mathrm{G} = \nabla_a Q(s,\cdot)|_{a=\mu_\beta}$ and $\mathrm{H} = \nabla_a^2 Q(s,\cdot)|_{a=\mu_\beta}$.

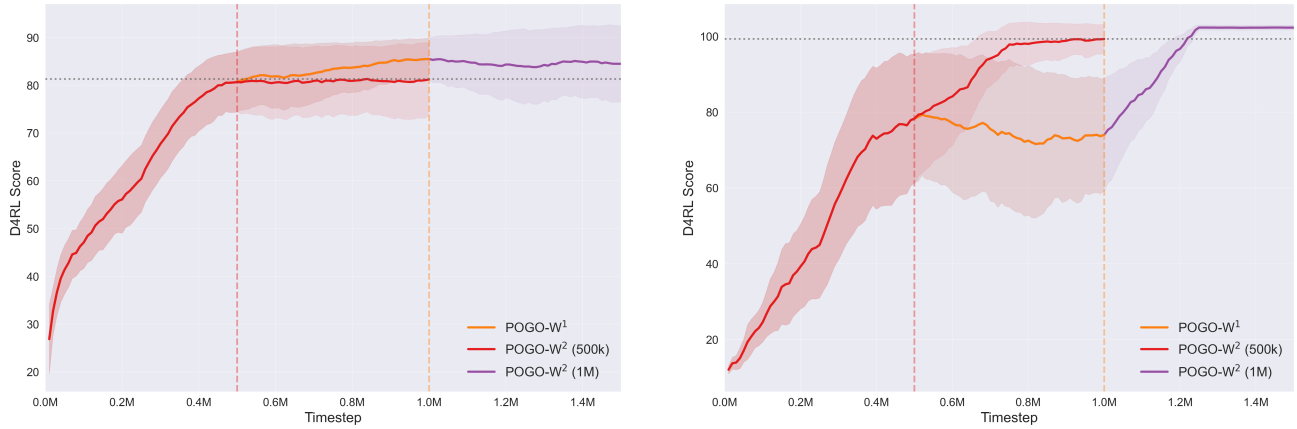
The proximal energy on each manifold becomes

$$\text{Wasserstein:}\quad \mathrm{prox}_{\mathcal{W}_2}(\mathcal{E}_s \mid \pi_\beta) = -\mathrm{G}^\mathsf{T}(\mu - \mu_\beta) - \frac{1}{2}\sum_{i=1}^d \mathrm{H}_{ii}\sigma_i^2 - \lambda\sum_{i=1}^d \log\sigma_i + \frac{1}{2\tau}\mathcal{W}_2^2(\pi,\pi_\beta).$$

$$\text{Fisher-Rao:}\quad \mathrm{prox}_{\mathcal{W}_2}(\mathcal{E}_s \mid \pi_\beta) = -\mathrm{G}^\mathsf{T}(\mu - \mu_\beta) - \frac{1}{2}\sum_{i=1}^d \mathrm{H}_{ii}\sigma_i^2 - \lambda\sum_{i=1}^d \log\sigma_i + \frac{1}{2\tau}d_{\mathrm{FR}}^2(\pi,\pi_\beta).$$

Differentiating the proximal energies and evaluating to first order in τ yields:

$$\text{Wasserstein:}\quad \begin{aligned} \mu_\tau &= \mu_\beta - \tau\,\mathrm{G} + O(\tau^2), \\ \Sigma_\tau &= \Sigma_\beta - 2\tau\big(\mathrm{H}\,\Sigma_\beta - \lambda I_d\big) + O(\tau^2). \end{aligned}$$



(a) `halfCheetah-medium-expert-v2` (b) `hopper-medium-replay-v2`

Figure 1: Effect of re-centering in POGO–W on two MuJoCo tasks. For each task we compare the one-step update (POGO–W^1) with the re-centered two-step variant started at either 0.5M or 1.0M timesteps (POGO–W^2 (500k) / (1M)). Shaded bands denote mean ± one standard deviation over seeds; vertical dashed lines mark the switching points.

$$\text{Fisher–Rao:} \quad \begin{aligned} \mu_\tau &= \mu_\beta - \tau\,\Sigma_\beta\,\mathrm{G} + O(\tau^2), \\ \Sigma_\tau &= \Sigma_\beta - \tau\big(\mathrm{H}\,\Sigma_\beta^2 - \lambda\,\Sigma_\beta\big) + O(\tau^2). \end{aligned}$$

These relations describe the first-order evolution of the mean and log-variance with respect to the small step size τ under each geometry.

Following the first-order update equations, our formulation that models the actor as a diagonal Gaussian distribution introduces a structural limitation when combined with the FR geometry. Since the FR metric rescales gradients by the current variance, the mean update depends multiplicatively on Σ_β. When the behavioral policy exhibits small variance, the effective step size of the update becomes severely damped, leading to overly conservative dynamics and minimal policy improvement after early convergence. This phenomenon can be partly attributed to the nature of offline datasets in continuous-control benchmarks, where the recorded behavior policy is often nearly deterministic. In such cases, the empirical variance estimated from the dataset tends to be very small, limiting the ability of the FR geometry to propagate meaningful updates. Moreover, the FR metric is inherently more sensitive to the estimation error of Σ_β, as it directly scales both the gradient and curvature terms by the covariance. Consequently, imperfect covariance estimation—inevitable when inferring the behavior policy from finite samples—can further degrade the stability of FR-based updates.

Nevertheless, when the behavior policy possesses a moderate level of stochasticity, the FR geometry can provide a principled natural-gradient scaling that better aligns the update direction with the underlying statistical manifold. This suggests that FR-based formulations may be more suitable in settings where the data-generating policy maintains non-negligible variance, while Wasserstein-based updates remain more robust when variance estimates are small or uncertain.

3.2 Effect of step-switching schedule

Although two-step integration generally improves the convergence stability of POGO, its benefit depends on the accuracy of the critic and the maturity of the learned Q-function. When the critic remains unstable or the value landscape has not yet converged, an early transition to two-step updates can propagate inaccurate gradient information, occasionally degrading performance—as observed in environments such as `halfcheetah-medium-expert` and `antmaze-umaze` in Table 4.

To examine this dependency, we conducted an ablation in which the one-step update was extended to 1M iterations before switching to the two-step phase for an additional 500k steps. Representative learning curves for `hopper-medium-replay` and `halfcheetah-medium-expert` are shown in Figure 1.

Although two-step integration generally improves the convergence stability of POGO, its benefit is conditional on the maturity of the value function. In particular, Fig. 1(a) shows that switching *too early* can degrade performance—because the two-step phase does not update the critic, premature recentering propagates inaccurate gradients. Conversely, Fig. 1(b) indicates that switching *too late* slows convergence; in such cases a mid-horizon schedule (e.g., the 0.5M switch) is preferable. These results reinforce that the switching time is focal: two-step Euler integration is most helpful once the critic has converged to a locally smooth, reliable Q-landscape, whereas an overly delayed switch causes additional training cost. Motivated by this, we view *adaptive scheduling*—based on stability criteria such as critic-loss variance or Q-value consistency—as an important on-going direction.