Wasserstein Policy Optimization

Hu Hanyang
Instructor: Xiaochuan Tian

University of California San Diego
MATH 277A Course Project

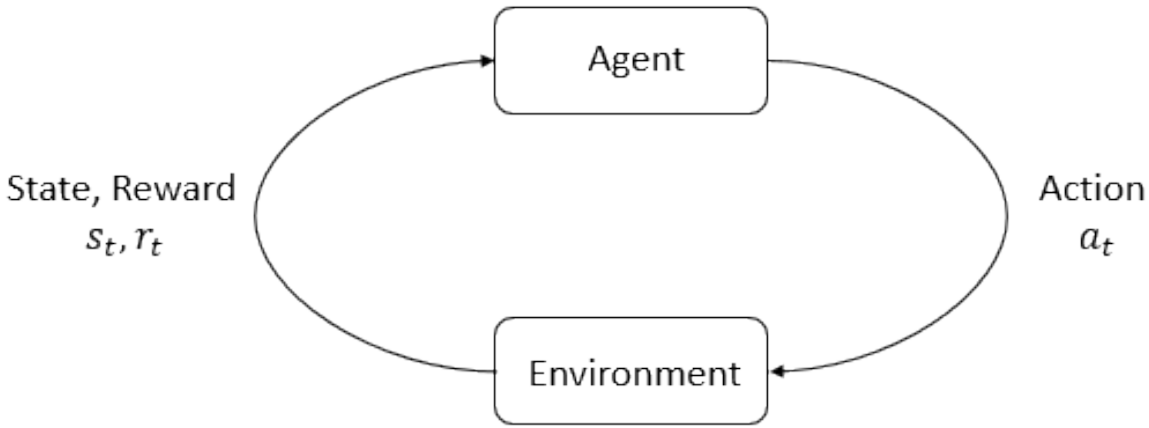
June 8, 2026

Outline

Problem Formulation: Markov Decision Processes

A **Markov decision process (MDP)** is defined by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, p_0, \gamma)$

- state space $\mathcal{S}$ and action space $\mathcal{A}$
- transition probability $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$
- reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- initial state distribution $s_0 \sim p_0(\cdot)$
- discount factor $\gamma \in (0, 1)$



Source: OpenAI Spinning Up (`spinningup.openai.com`).

Problem Formulation: Reinforcement Learning

Reinforcement learning studies sequential decision-making problems where an agent interacts with an environment to maximize long-term reward.

RL Objective

Given a stochastic policy $\pi(a|s)$, the goal is to maximize

$$\mathcal{J}[\pi] = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right], \tag{1}$$

where $s_0 \sim p_0$, $a_t \sim \pi(\cdot|s_t)$, and $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$.

Value Functions and Occupancy

Value Function

$$V^{\pi}(s) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}r(s_t,a_t)\,\middle|\, s_0 = s\right]. \tag{2}$$

Action-Value Function

$$Q^{\pi}(s,a) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}r(s_t,a_t)\,\middle|\, s_0 = s,\ a_0 = a\right]. \tag{3}$$

Discounted State Occupancy

$$d^{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t}\mathbb{P}^{\pi}(s_t = s). \tag{4}$$

Policy Gradient

A standard approach is to parameterize the policy by θ and optimize $\mathcal{J}[\pi_\theta]$ by gradient ascent $\theta \leftarrow \theta + \nabla_\theta \mathcal{J}[\pi_\theta]$.

Policy Gradient: Direct Policy-Derivative Form

$$\nabla_\theta \mathcal{J}[\pi_\theta] = \frac{1}{1-\gamma} \int Q^{\pi_\theta}(s,a) d^{\pi_\theta}(s) \nabla_\theta \pi_\theta(a|s)\, ds\, da. \tag{5}$$

Using the log-derivative trick $\nabla_\theta \pi_\theta(a|s) = \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)$, we obtain the more common score-function form.

Policy Gradient: Score-Function Form

$$\nabla_\theta \mathcal{J}[\pi_\theta] = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)} \left[Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) \right]. \tag{6}$$

Policy Gradient

Limitation

Policy gradient updates improve stochastic policies by reweighting the score function $\nabla_\theta \log \pi_\theta(a|s)$ according to scalar value estimates $Q^\pi(s, a)$.

Some extensions to address this issue:

- Replace $Q^\pi(s, a)$ with an advantage function $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$ to reduce variance.
- TRPO[1] constrains the update by a KL-divergence trust region.
- PPO[2] replaces the constrained problem with a clipped surrogate objective.

[1] John Schulman et al. *Trust Region Policy Optimization*. 2017. arXiv: 1502.05477 [cs.LG]. URL: https://arxiv.org/abs/1502.05477.

[2] John Schulman et al. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG]. URL: https://arxiv.org/abs/1707.06347.

Deterministic Policy Gradient

For continuous control, if we restrict to the class of deterministic policies $\pi_\theta : \mathcal{S} \to \mathcal{A}$, we can directly use the action-value gradient $\nabla_a Q^\pi(s, a)$ to update the policy (by the chain rule).

Deterministic Policy Gradient

$$\nabla_\theta \mathcal{J}[\pi_\theta] = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_\theta}} \left[\nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi_\theta}(s, a) \big|_{a=\pi_\theta(s)} \right]. \tag{7}$$

Limitation

DPG uses the action-space geometry of the critic through $\nabla_a Q^\pi(s, a)$, but deterministic policies usually require pre-defined exploration noise.

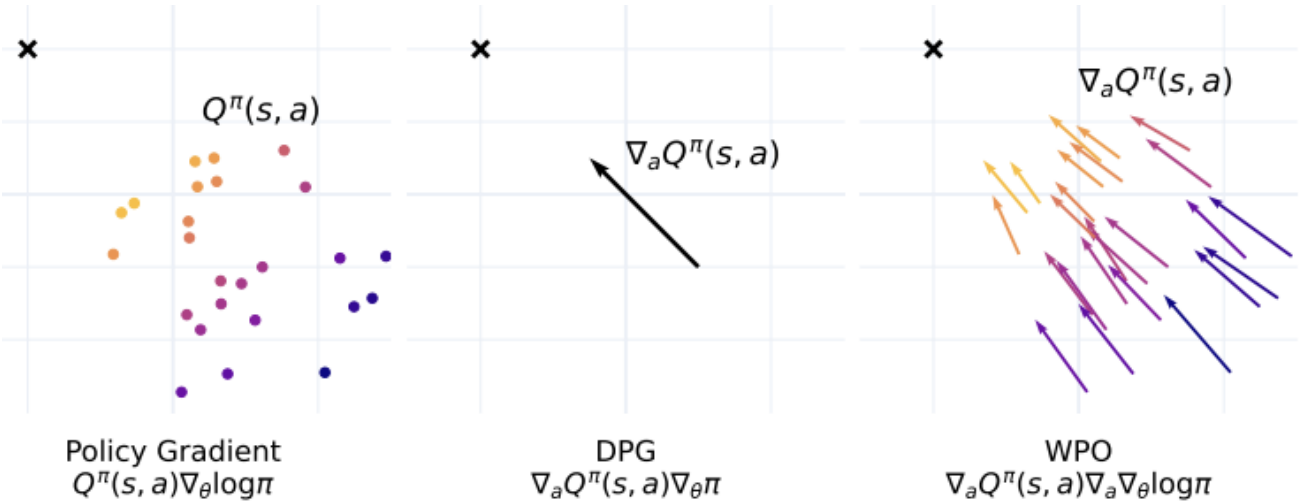
Motivation for WPO

Wasserstein policy optimization[3] combines two ideas:

- From policy gradient: optimize stochastic policies.
- From deterministic policy gradient: use action-value gradients.

Main Idea of WPO

View each conditional policy $\pi(\cdot|s)$ as a probability distribution over actions and update it by transporting probability mass in action space.



[3] David Pfau et al. *Wasserstein Policy Optimization*. 2025. arXiv: 2505.00663 [cs.LG]. URL: https://arxiv.org/abs/2505.00663.

Outline

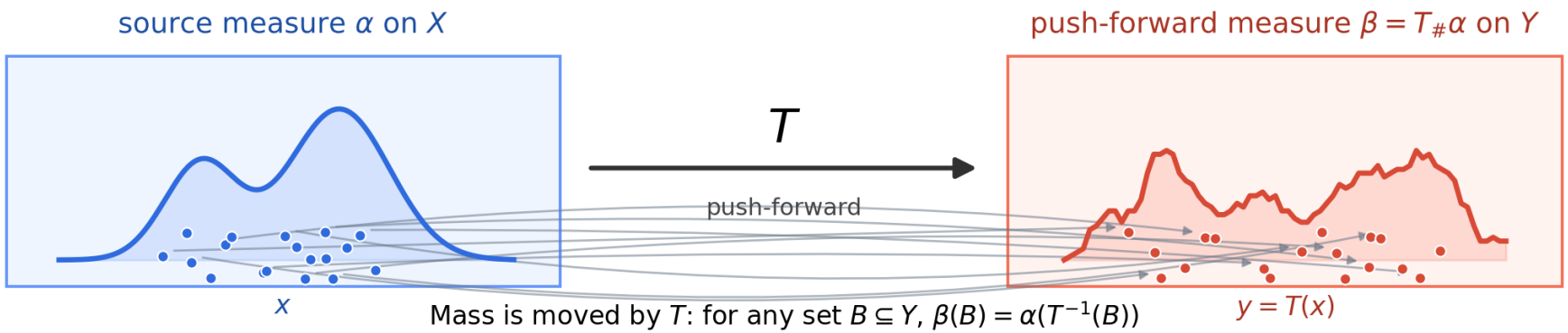
Optimal Transport

Push-Forward

For a measurable map $T : \mathcal{X} \to \mathcal{Y}$, the push-forward measure $\beta = T_{\#}\alpha$ satisfies

$$\int_{\mathcal{Y}} h(y)\, d\beta(y) = \int_{\mathcal{X}} h(T(x))\, d\alpha(x) \tag{8}$$

for all test functions h.



Monge Problem and Kantorovich Relaxation

Monge Problem

Given $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{Y})$, find a map T such that $T_{\#}\alpha = \beta$ and

$$\inf_{T_{\#}\alpha=\beta} \int_{\mathcal{X}} c(x, T(x))\, d\alpha(x). \tag{9}$$

Note. The Monge problem may not have a solution.

Kantorovich Problem

Instead of a map, optimize over couplings $\pi \in \Pi(\alpha, \beta)$:

$$\inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\pi(x, y), \tag{10}$$

where $\Pi(\alpha, \beta)$ is the set of joint distributions with marginals α and β.

Wasserstein Distance and Brenier's Theorem

p-Wasserstein Distance

For $c(x, y) = \|x - y\|^p$, the following defines the p-Wasserstein distance:

$$\mathcal{W}_p(\alpha, \beta) = \left(\inf_{\pi \in \Pi(\alpha, \beta)} \int_{\Omega \times \Omega} \|x - y\|^p \, d\pi(x, y) \right)^{1/p} \tag{11}$$

defines a metric for probability measures with finite p-th moment.

Brenier's Theorem and Kantorovich Potential

If $p = 2$ and α is absolutely continuous, the optimal transport map is given by the gradient of a convex function ϕ:

$$T(x) = \nabla \phi(x). \tag{12}$$

For **Kantorovich potential** $\varphi(x) = \frac{1}{2}\|x\|^2 - \phi(x)$, $T(x) = x - \nabla \varphi(x)$.

Outline

Gradient Flow on Probability Measures

In Euclidean space, gradient flow can be viewed as the continuous-time limit of gradient descent:

$$\dot{x}_t = -\nabla F(x_t). \tag{13}$$

For probability measures, we replace the Euclidean distance by $\mathcal{W}_2$.

Minimizing Movement Scheme

Given μ_k^τ, define

$$\mu_{k+1}^\tau \in \arg\min_{\mu} \left\{ F(\mu) + \frac{\mathcal{W}_2^2(\mu, \mu_k^\tau)}{2\tau} \right\}. \tag{14}$$

- $F(\mu)$ is the objective functional.
- The Wasserstein term penalizes moving too far from μ_k^τ.
- Letting $\tau \to 0$ gives a continuous-time flow of measures.

First Variation

To derive a gradient flow equation, we need a derivative of a functional with respect to a measure.

First Variation

The first variation of $F : \mathcal{P}_2(\Omega) \to \mathbb{R}$ is a function $\frac{\delta F}{\delta \mu}(\mu)$ such that

$$\frac{d}{d\epsilon} F(\mu + \epsilon\xi)\Big|_{\epsilon=0} = \int_\Omega \frac{\delta F}{\delta \mu}(\mu)(x)\, d\xi(x) \tag{15}$$

for zero-mass perturbations ξ, i.e. $\int_\Omega d\xi = 0$.

Interpretation

Equivalently, the first variation gives the first-order expansion

$$F(\mu + \epsilon\xi) = F(\mu) + \epsilon \int_\Omega \frac{\delta F}{\delta \mu}(\mu) d\xi(x) + o(\epsilon). \tag{16}$$

First Variation of the Wasserstein Term

Key Fact

Let φ be the Kantorovich potential for transporting μ to ν under the quadratic cost. Then

$$\frac{\delta}{\delta \mu} \frac{1}{2} \mathcal{W}_2^2(\mu, \nu) = \varphi. \tag{17}$$

Therefore, the optimality condition for the MMS step gives

$$\frac{\delta F}{\delta \mu} + \frac{\varphi}{\tau} = \text{const}. \tag{18}$$

Taking gradients,

$$\nabla \varphi = -\tau \nabla \frac{\delta F}{\delta \mu}. \tag{19}$$

Wasserstein Steepest-Descent Velocity

Discrete Transport Velocity

In the MMS step, the displacement is controlled by $\nabla\varphi$, i.e. the discrete velocity moving a point x from μ_k^τ to μ_{k+1}^τ is

$$v^\tau(x) = \frac{x - T(x)}{\tau} = \frac{\nabla\varphi(x)}{\tau}. \tag{20}$$

Using

$$\nabla\varphi = -\tau\nabla\frac{\delta F}{\delta\mu},$$

we obtain the limiting velocity field:

Wasserstein Steepest-Descent Velocity

$$v_t = -\nabla\frac{\delta F}{\delta\mu}(\mu_t). \tag{21}$$

Continuity Equation

Wasserstein gradient flow moves probability mass by the velocity field

$$v_t = -\nabla \frac{\delta F}{\delta \mu}(\mu_t).$$

If ρ_t is the density of μ_t, then mass conservation gives

Continuity Equation

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0. \tag{22}$$

Substituting the Wasserstein velocity field gives:

Wasserstein Gradient Flow

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho} \right). \tag{23}$$

Descent Property

Along sufficiently regular solutions, the functional decreases over time.

$$\begin{aligned}\frac{d}{dt}F(\rho_t) &= \int_\Omega \frac{\delta F}{\delta \rho}(\rho_t)\partial_t \rho_t \, dx \\ &= \int_\Omega \frac{\delta F}{\delta \rho}(\rho_t)\nabla \cdot \left(\rho_t \nabla \frac{\delta F}{\delta \rho}(\rho_t)\right) dx \\ &= -\int_\Omega \rho_t \left\|\nabla \frac{\delta F}{\delta \rho}(\rho_t)\right\|^2 dx \leq 0.\end{aligned} \tag{24}$$

Takeaway

Wasserstein gradient flow is a steepest-descent dynamics under transport geometry.

Example: Fokker–Planck Equation

Consider the entropy-plus-potential functional

$$F(\rho) = \int_\Omega \rho(x) \log \rho(x)\, dx + \int_\Omega V(x)\rho(x)\, dx. \tag{25}$$

Its first variation is

$$\frac{\delta F}{\delta \rho} = \log \rho + 1 + V. \tag{26}$$

Substituting into Wasserstein gradient flow:

$$\partial_t \rho = \nabla \cdot (\rho \nabla (\log \rho + 1 + V)) \tag{27}$$

$$= \Delta \rho + \nabla \cdot (\rho \nabla V). \tag{28}$$

Result

This is the Fokker–Planck equation.

Outline

Expected Return as a Policy Functional

Policy Functional

The expected return can be viewed as a functional of the policy distribution:

$$\mathcal{J}[\pi] = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right]. \tag{29}$$

More explicitly,

$$\mathcal{J}[\pi] = \sum_{t=0}^{\infty}\gamma^t \int r_t p_0(s_0)\prod_{t'=0}^{t}\pi(a_{t'}|s_{t'})\prod_{t'=0}^{t-1}\mathbb{P}(s_{t'+1}|s_{t'}, a_{t'})d\tau_{0:t}. \tag{30}$$

Goal

Compute how $\mathcal{J}[\pi]$ changes when we perturb the policy distribution $\pi(a|s)$ itself.

Functional Policy Gradient

The first variation of the return gives the marginal effect of increasing probability mass at (s, a).

First Variation of the Return

$$\frac{\delta \mathcal{J}[\pi]}{\delta \pi}(s,a) = Q^\pi(s,a) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s) = \frac{1}{1-\gamma} d^\pi(s) Q^\pi(s,a). \quad (31)$$

- $Q^\pi(s,a)$ measures the value of taking action a at state s.
- $d^\pi(s)$ measures how often state s is visited under policy π.

Intuition

Increasing $\pi(a|s)$ is most useful when state s is frequently visited and action a has high value.

Connection to Classical Policy Gradient

For a parametric policy π_θ, applying the chain rule to the policy functional gives

$$\begin{aligned}\nabla_\theta \mathcal{J}[\pi_\theta] &= \int \frac{\delta \mathcal{J}[\pi_\theta]}{\delta \pi}(s,a)\nabla_\theta \pi_\theta(a|s)\, ds\, da \\ &= \frac{1}{1-\gamma}\int d^{\pi_\theta}(s) Q^{\pi_\theta}(s,a)\nabla_\theta \pi_\theta(a|s)\, ds\, da.\end{aligned} \tag{32}$$

Using the log-derivative identity

$$\nabla_\theta \pi_\theta(a|s) = \pi_\theta(a|s)\nabla_\theta \log \pi_\theta(a|s),$$

we recover the standard score-function policy gradient:

$$\nabla_\theta \mathcal{J}[\pi_\theta] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[Q^{\pi_\theta}(s,a)\nabla_\theta \log \pi_\theta(a|s)\right]. \tag{33}$$

From Functional PG to Wasserstein Flow

For each fixed state s, apply Wasserstein gradient ascent to $\pi(\cdot|s)$:

Continuity Equation in Action Space

$$\partial_t \pi_t(a|s) + \nabla_a \cdot (\pi_t(a|s) v_t(s,a)) = 0. \tag{34}$$

Since we maximize $\mathcal{J}$, the Wasserstein ascent velocity is

$$v_t(s,a) = \nabla_a \frac{\delta \mathcal{J}}{\delta \pi}(s,a). \tag{35}$$

Using the functional policy gradient,

$$v_t(s,a) = \frac{d^{\pi_t}(s)}{1-\gamma} \nabla_a Q^{\pi_t}(s,a), \tag{36}$$

i.e., WPO transports action probability mass along $\nabla_a Q^{\pi}(s,a)$.

Comparison with Policy Gradient

Policy gradient and WPO both optimize stochastic policies, but they use different geometric information.

Policy Gradient

$$\Delta\theta_{\mathrm{PG}} \sim \mathbb{E}\left[Q^{\pi}(s,a)\nabla_{\theta}\log\pi_{\theta}(a|s)\right]. \tag{37}$$

- Samples actions from the current policy.
- Uses scalar value estimates $Q^{\pi}(s,a)$ or $A^{\pi}(s,a)$.
- Increases or decreases action likelihoods through the score function.

Limitation

Policy gradient mainly reweights sampled actions; it does not explicitly use the local action-space direction in which $Q^{\pi}(s,a)$ increases.

Comparison with Deterministic Policy Gradient

DPG directly uses the action-space gradient of the critic.

Deterministic Policy Gradient

For a deterministic policy $a = \pi_\theta(s)$,

$$\Delta\theta_{\mathrm{DPG}} \sim \mathbb{E}_{s\sim d^\pi}\left[\nabla_\theta \pi_\theta(s) \nabla_a Q^\pi(s,a)\big|_{a=\pi_\theta(s)}\right]. \tag{38}$$

WPO

For a stochastic policy, WPO uses

$$v(s,a) \sim \nabla_a Q^\pi(s,a) \tag{39}$$

as a velocity field for transporting the whole action distribution.

Projection onto a Parametric Policy Class

The Wasserstein flow gives a desired distribution-level update

$$\pi_\theta \quad \longrightarrow \quad \pi_\theta + \frac{\partial \pi_\theta}{\partial t} dt.$$

However, the policy is restricted to a parametric class, so a parameter update gives only

$$\pi_{\theta+\Delta\theta} \approx \pi_\theta + \nabla_\theta \pi_\theta \, \Delta\theta.$$

KL Projection

Choose $\Delta\theta$ so that the parametric update best matches the Wasserstein-flow update:

$$\Delta\theta = \arg\min_{\Delta\theta} D_{\mathrm{KL}} \left[\pi_\theta \,\middle\|\, \pi_\theta + \frac{\partial \pi_\theta}{\partial t} dt - \nabla_\theta \pi_\theta \Delta\theta \right]. \tag{40}$$

Quadratic Approximation of KL Divergence

For a parametric distribution p_θ, the KL divergence can be approximated by its second-order Taylor expansion[4]

$$
\begin{aligned}
& D_{\mathrm{KL}}[p_\theta \,\|\, p_{\theta+\Delta\theta}] \\
& = \mathbb{E}_{z\sim p_\theta}\left[\log p_\theta(z) - \log p_{\theta+\Delta\theta}(z)\right] \\
& \approx -\mathbb{E}_{z\sim p_\theta}\left[\nabla_\theta \log p_\theta(z)\right]^T \Delta\theta - \frac{1}{2}\Delta\theta^T \mathbb{E}_{z\sim p_\theta}\left[\nabla_\theta^2 \log p_\theta(z)\right]\Delta\theta \\
& = \frac{1}{2}\Delta\theta^T \mathbb{E}_{z\sim p_\theta}\left[\nabla_\theta \log p_\theta(z) \nabla_\theta \log p_\theta(z)^\top\right]\Delta\theta \\
& = \frac{1}{2}\Delta\theta^T \mathcal{F}_\theta \Delta\theta.
\end{aligned}
\tag{41}
$$

[4] Razvan Pascanu and Yoshua Bengio. "Revisiting natural gradient for deep networks". In: *arXiv preprint arXiv:1301.3584* (2013).

KL Projection Objective (Quadratic Form)

Using the same KL projection objective from the previous slide,

$$\begin{aligned} & D_{\mathrm{KL}}\left[\pi_\theta \,\middle\|\, \pi_\theta + \frac{\partial \pi_\theta}{\partial t} dt - \nabla_\theta \pi_\theta \Delta\theta\right] \\ & \quad \approx \frac{1}{2} \Delta\theta^\top \mathcal{F}_{\theta\theta} \Delta\theta - dt\, \Delta\theta^\top \mathcal{F}_{t\theta} + \frac{1}{2} dt^2 \mathcal{F}_{tt}. \end{aligned} \tag{42}$$

Here

$$\mathcal{F}_{\theta\theta} = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top\right], \tag{43}$$

$$\mathcal{F}_{t\theta} = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\partial_t \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)\right], \tag{44}$$

$$\mathcal{F}_{tt} = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\left(\partial_t \log \pi_\theta(a|s)\right)^2\right] \tag{45}$$

which gives the closed-form solution

$$\Delta\theta = \mathcal{F}_{\theta\theta}^{-1} \mathcal{F}_{t\theta} dt. \tag{46}$$

Computing the Cross Term

It remains to compute $\mathcal{F}_{t\theta}$:

$$\mathcal{F}_{t\theta} = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\partial_t \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)\right] \tag{47}$$

$$= \int \partial_t \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)\, da. \tag{48}$$

The policy density evolves by the continuity equation in action space:

$$\partial_t \pi_\theta(a|s) = -\nabla_a \cdot \left(\pi_\theta(a|s) v(s,a)\right). \tag{49}$$

For return maximization, the Wasserstein velocity is

$$v(s,a) = \nabla_a \frac{\delta \mathcal{J}}{\delta \pi}(s,a) \propto \nabla_a Q^\pi(s,a), \tag{50}$$

where the positive factor $\frac{1}{1-\gamma} d^\pi(s)$ is suppressed for a fixed state s.

Computing the Cross Term (Cont.)

Substituting the WPO continuity equation gives

$$\mathcal{F}_{t\theta} \propto -\int \nabla_a \cdot (\pi_\theta(a|s)\nabla_a Q^\pi(s,a))\, \nabla_\theta \log \pi_\theta(a|s)\, da. \tag{51}$$

Assuming the boundary term vanishes, integration by parts gives

$$\begin{aligned} \mathcal{F}_{t\theta} &= \int \pi_\theta(a|s) \left[\nabla_a \nabla_\theta \log \pi_\theta(a|s)\right] \nabla_a Q^\pi(s,a)\, da \\ &= \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \nabla_a \log \pi_\theta(a|s) \nabla_a Q^\pi(s,a)\right]. \end{aligned} \tag{52}$$

WPO Update Direction

$$\Delta\theta = \mathcal{F}_{\theta\theta}^{-1} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \nabla_a \log \pi_\theta(a|s) \nabla_a Q^\pi(s,a)\right] dt. \tag{53}$$

Deterministic Limit of WPO

Consider a Gaussian stochastic policy with fixed covariance:

$$\pi_\theta(a|s) = \mathcal{N}(a; m_\theta(s), \Sigma), \tag{54}$$

where $m_\theta(s)$ is the mean action. As $\Sigma \to 0$, the density $\pi_\theta(\cdot|s)$ concentrates at $a = m_\theta(s)$, so the stochastic policy approaches the deterministic policy $a = m_\theta(s)$.

For a fixed state s,

$$\nabla_\theta \log \pi_\theta(a|s) = \nabla_\theta m_\theta(s) \Sigma^{-1}(a - m_\theta(s)). \tag{55}$$

Therefore, the per-state Fisher matrix becomes

$$\begin{aligned} \mathcal{F}_{\theta\theta}(s) &= \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top \right] \\ &= \nabla_\theta m_\theta(s) \Sigma^{-1} \nabla_\theta m_\theta(s)^\top. \end{aligned} \tag{56}$$

Deterministic Limit: Cross Term

For the same Gaussian policy,

$$\nabla_a \log \pi_\theta(a|s) = -\Sigma^{-1}(a - m_\theta(s)). \tag{57}$$

Hence the mixed derivative appearing in WPO is

$$\nabla_\theta \nabla_a \log \pi_\theta(a|s) = \nabla_\theta m_\theta(s) \Sigma^{-1}. \tag{58}$$

For a fixed state s, the WPO cross term is

$$\mathcal{F}_{t\theta}(s) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \nabla_a \log \pi_\theta(a|s) \nabla_a Q^{\pi_\theta}(s,a) \right]. \tag{59}$$

As $\Sigma \to 0$, the policy concentrates at $a = m_\theta(s)$, so

$$\mathcal{F}_{t\theta}(s) \approx \nabla_\theta m_\theta(s) \Sigma^{-1} \nabla_a Q^{m_\theta}(s,a)\big|_{a=m_\theta(s)}. \tag{60}$$

WPO as Preconditioned DPG

For a fixed state s, the WPO parameter update is

$$\Delta\theta_{\mathrm{WPO}}(s) = \mathcal{F}_{\theta\theta}(s)^{-1}\mathcal{F}_{t\theta}(s). \tag{61}$$

For isotropic covariance $\Sigma = \sigma^2 I$, the factors σ^{-2} cancel. Define

$$G_\theta(s) = \nabla_\theta m_\theta(s) \nabla_\theta m_\theta(s)^\top. \tag{62}$$

The fixed-state deterministic policy-gradient direction is

$$g_{\mathrm{DPG}}(s) = \nabla_\theta m_\theta(s) \nabla_a Q^{m_\theta}(s,a)\big|_{a=m_\theta(s)}. \tag{63}$$

Therefore, in the deterministic Gaussian limit,

WPO as Preconditioned DPG

$$\Delta\theta_{\mathrm{WPO}}(s) = G_\theta(s)^{-1} g_{\mathrm{DPG}}(s). \tag{64}$$

Key Takeaway: WPO Update Rule

For a fixed state s, the ideal Wasserstein policy improvement transports the action distribution by

$$\partial_t \pi_\theta(a|s) = -\nabla_a \cdot (\pi_\theta(a|s) \nabla_a Q^\pi(s,a)), \tag{65}$$

where the state-dependent factor $\frac{1}{1-\gamma} d^\pi(s)$ is only implicitly incorporated.

Projected WPO Update

After projecting this distribution-level flow onto the parametric policy class, the parameter update is

$$\theta \leftarrow \theta + \eta \mathcal{F}_{\theta\theta}^{-1} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} [\nabla_\theta \nabla_a \log \pi_\theta(a|s) \nabla_a Q^\pi(s,a)], \tag{66}$$

where $\eta > 0$ is the learning rate.

Outline

Toy Example: Setup

We reproduce the 1-D mixture-of-Gaussians example from the original WPO paper.

State-Independent Action-Value Function

$$Q(a) = -\frac{1}{100}a^4 + a^2.$$

This objective has two symmetric global maxima $a^\star = \pm\sqrt{50}$.

Mixture Policy

$$\pi_\theta(a) = \sum_{i=1}^{K} \rho_i \mathcal{N}(a \mid \mu_i, \sigma_i^2), \qquad \sum_{i=1}^{K} \rho_i = 1.$$

We use $K = 2$, initial means $\mu_1 = -1$ and $\mu_2 = 1$, initial standard deviation $\sigma_1 = \sigma_2 = 10$, and balanced initial weights $\rho_1 = \rho_2 = 0.5$.

Toy Example: Update Rules

Policy Gradient

$$\Delta_{\mathrm{PG}} \approx \frac{1}{M} \sum_{m=1}^{M} Q(a_m) \nabla_\theta \log \pi_\theta(a_m).$$

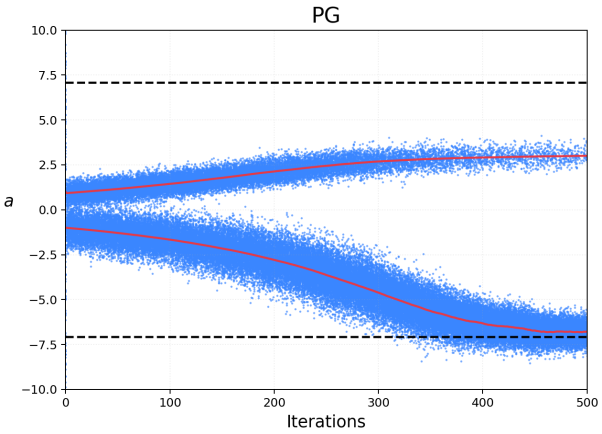
Natural Policy Gradient

$$\Delta_{\mathrm{NPG}} \approx \mathcal{F}_{\theta\theta}^{-1} \frac{1}{M} \sum_{m=1}^{M} Q(a_m) \nabla_\theta \log \pi_\theta(a_m).$$

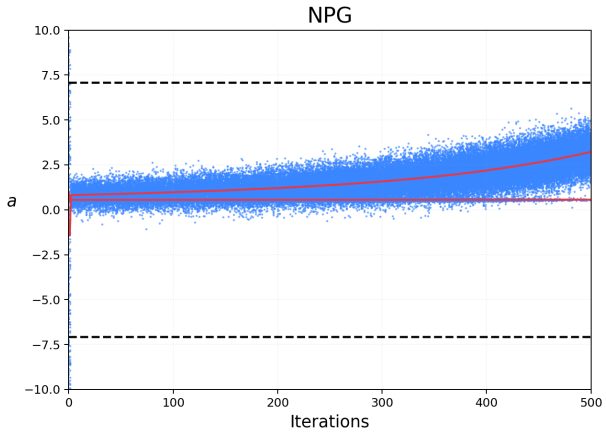
Wasserstein Policy Optimization

$$\Delta_{\mathrm{WPO}} \approx \mathcal{F}_{\theta\theta}^{-1} \frac{1}{M} \sum_{m=1}^{M} \nabla_\theta \nabla_a \log \pi_\theta(a_m) \nabla_a Q(a_m).$$

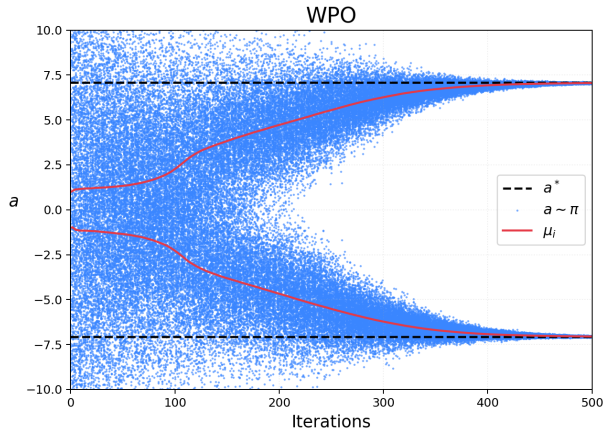
Toy Example: Results



(a) PG



(b) NPG



(c) WPO

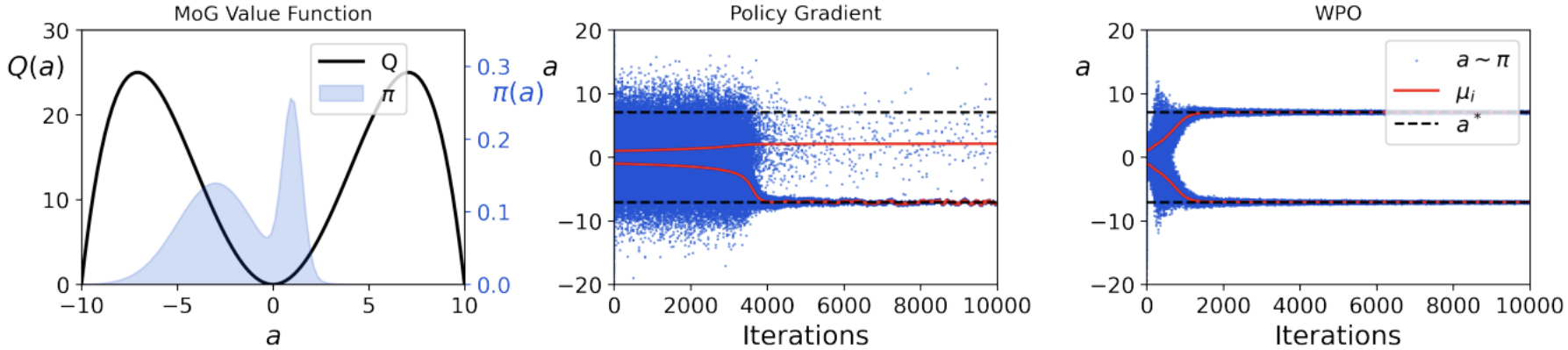


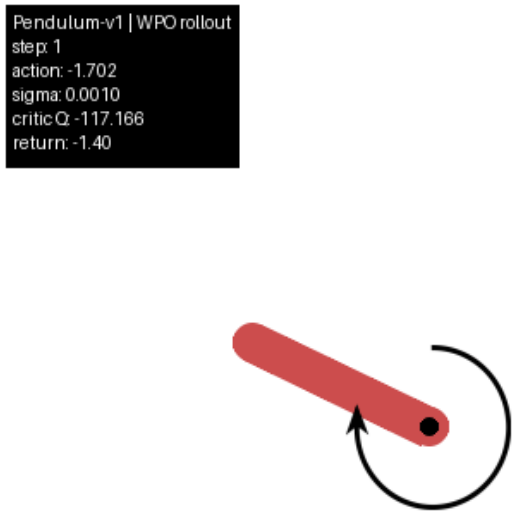
Figure: Original WPO paper results

Inverted Pendulum: Setup

We test WPO on the inverted pendulum environment from Gymnasium, a standard continuous-control task.

Environment:

- The pendulum starts from a non-upright configuration.
- The goal is to apply torques to swing up and stabilize the pendulum.
- This is a more representative test of WPO than the toy example, since the Q-function is learned from data and the policy depends on the state.



Unified Actor-Critic Framework

Initialize actor π_θ, critic Q_ψ, target networks, and replay buffer $\mathcal{D}$.
for each environment step **do**
    Sample action from the current behavior policy.
    Execute action and store transition (s, a, r, s') in $\mathcal{D}$.
    Sample a mini-batch of transitions from $\mathcal{D}$.
    Compute TD target:

$$y = r + \gamma Q_{\bar{\psi}}(s', a')$$

using the target critic and target actor.
    Update critic Q_ψ by minimizing the TD error.
    Compute the actor update direction using PG, DPG, or WPO.
    Update actor parameters θ.
    Update target networks by Polyak averaging.
end for

Actor Update Rules

The critic $Q_\psi(s, a)$ is learned from replay-buffer samples. For a mini-batch $\{s_i\}_{i=1}^B$, PG and WPO sample M actions $a_i^{(k)} \sim \pi_\theta(\cdot|s_i)$.

Policy Gradient

$$\Delta_{\mathrm{PG}} = \frac{1}{NM} \sum_i \sum_k (Q_\psi(s_i, \tilde{a}_i^k) - b_i) \nabla_\theta \log \pi_\theta(\tilde{a}_i^k \mid s_i),$$

where the sample-based baseline is $b_i = \frac{1}{M} \sum_{k=1}^M Q_\psi(s_i, a_i^{(k)}) \approx V^\pi(s_i)$.

Deterministic Policy Gradient

$$\Delta_{\mathrm{DPG}} = \frac{1}{N} \sum_i \nabla_a Q_\psi(s_i, a)\big|_{a=\pi_\theta(s_i)} \nabla_\theta \pi_\theta(s_i).$$

Actor Update Rules (Cont.)

Wasserstein Policy Optimization

$$\Delta_{\mathrm{WPO}} = \frac{1}{NM} \sum_i \sum_k \overline{\nabla_\theta} \nabla_a \log \pi_\theta(\tilde{a}_i^{(k)} \mid s_i) \nabla_a Q_\psi(s_i, \tilde{a}_i^{(k)}).$$

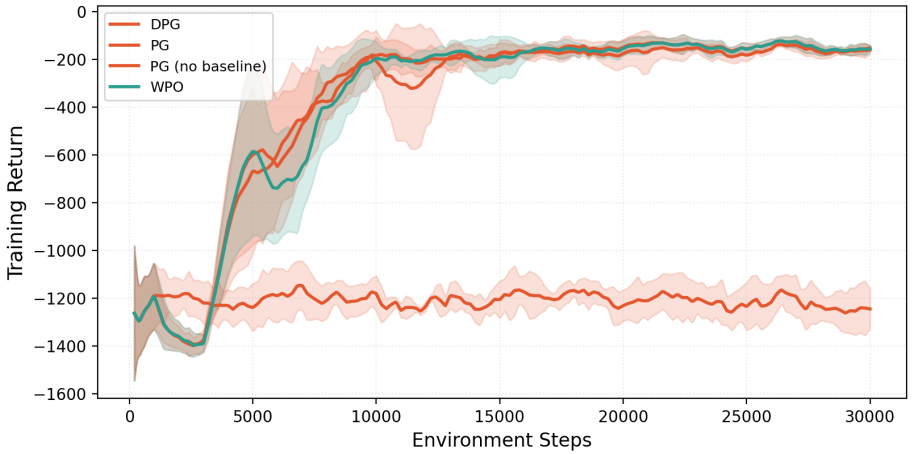
Here the gradient is modified to replace the full Fisher-matrix inversion.

Modified Gradient

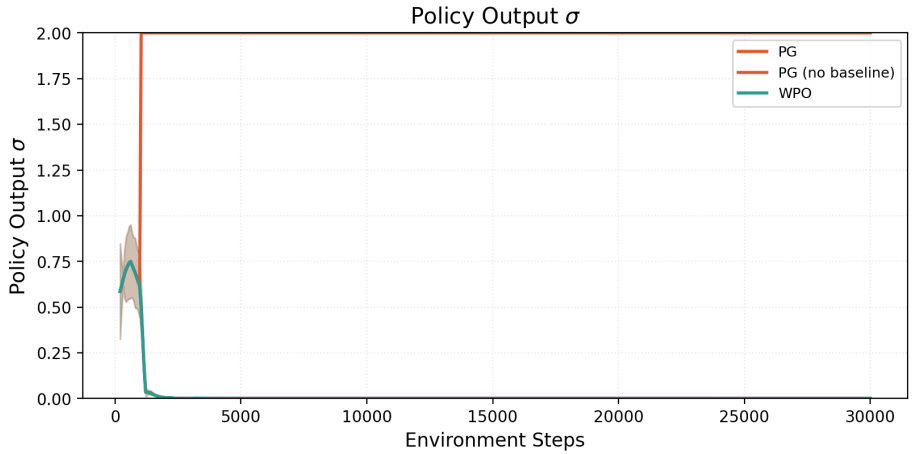
In practice, WPO ignores the contribution of the gradients of μ and Σ w.r.t. the policy parameter θ:

$$\overline{\nabla_\mu} \log \mathcal{N}(a \mid \mu, \sigma^2) = \sigma^2 \nabla_\mu \log \mathcal{N}(a \mid \mu, \sigma^2),$$
$$\overline{\nabla_\sigma} \log \mathcal{N}(a \mid \mu, \sigma^2) = \frac{1}{2} \sigma^2 \nabla_\sigma \log \mathcal{N}(a \mid \mu, \sigma^2).$$

Inverted Pendulum: Results



(a) Training return



(b) Policy standard deviation

Observations

- WPO reaches performance comparable to DPG and the stabilized PG baseline, the unstabilized PG performs poorly.
- The policy standard deviation decreases quickly, suggesting that WPO tends toward a nearly deterministic policy on this task.

Outline

Conclusion

Summary

WPO views $\pi_\theta(\cdot|s)$ as a distribution over actions and improves it by transporting probability mass along the critic action-gradient $\nabla_a Q^\pi(s,a)$.

Core Update Rule

$$\theta \leftarrow \theta + \eta \mathcal{F}_{\theta\theta}^{-1} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \nabla_a \log \pi_\theta(a|s) \nabla_a Q^\pi(s,a)\right].$$

Key Takeaway:

- Compared to PG, WPO is more stable and incorporates local action-space geometry of the critic.
- Compared to DPG, WPO works for any class of stochastic policies.
- Recent theoretical work has also begun to analyze convergence of WPO under additional assumptions.[5]

[5] David Šiška and Yufei Zhang. *A note on convergence of Wasserstein policy optimization*. 2026. arXiv: 2605.22622 [cs.LG]. URL: https://arxiv.org/abs/2605.22622.