CSE 542: Statistical Reinforcement Learning

Lecture 16: Policy Gradient Methods, the NPG, CPI & TRPO

Kevin Jamieson

Paul G. Allen School of Computer Science & Engineering
University of Washington

Outline

A Different Kind of RL Algorithm

Everything so far has been value-based. Estimate $Q^\star$ (or Q^π), act greedily. Exploration via optimism; analysis via Bellman equations.

Today: direct policy optimization. Parameterize the policy π_θ and do gradient ascent on V^{π_θ}. This is what most of "deep RL" actually runs (REINFORCE, A2C, TRPO, PPO).

A Different Kind of RL Algorithm

Everything so far has been value-based. Estimate $Q^\star$ (or Q^π), act greedily. Exploration via optimism; analysis via Bellman equations.

Today: direct policy optimization. Parameterize the policy π_θ and do gradient ascent on V^{π_θ}. This is what most of "deep RL" actually runs (REINFORCE, A2C, TRPO, PPO).

The tension we will study (AJKS Ch. 9–12)

The objective V^{π_θ} is *non-concave* in θ, so first-order methods can only promise stationary points. Yet under the right geometry (natural gradient) and the right exploration measure, we recover *global* optimality — sometimes at a dimension-free rate.

A Different Kind of RL Algorithm

Everything so far has been value-based. Estimate $Q^\star$ (or Q^π), act greedily. Exploration via optimism; analysis via Bellman equations.

Today: direct policy optimization. Parameterize the policy π_θ and do gradient ascent on V^{π_θ}. This is what most of "deep RL" actually runs (REINFORCE, A2C, TRPO, PPO).

The tension we will study (AJKS Ch. 9–12)

The objective V^{π_θ} is *non-concave* in θ, so first-order methods can only promise stationary points. Yet under the right geometry (natural gradient) and the right exploration measure, we recover *global* optimality — sometimes at a dimension-free rate.

Roadmap for this lecture.

1. **Ch 9:** gradient expressions + non-convexity + sampling.
2. **Ch 10:** global convergence for tabular softmax; the NPG.
3. **Ch 11:** function approximation; compatible features; Q-NPG.
4. **Ch 12:** trust-region methods — TRPO and PPO — the "incremental update" viewpoint.

Today's Plan

1. Policy Gradients (Ch. 9)
2. Global Optimality & the NPG (Ch. 10)
3. Function Approximation & the NPG (Ch. 11)
4. Trust-Region Methods: TRPO & PPO (Ch. 12)
5. Summary

Setup: Parametric Policies

Discounted MDP, $\gamma \in [0, 1)$. Class of parametric policies $\{\pi_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$. For a start distribution ρ, the objective is

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\rho), \qquad V^{\pi_\theta}(\rho) := \mathbb{E}_{s_0 \sim \rho}[V^{\pi_\theta}(s_0)].$$

Why stochastic policies? Deterministic policies are not differentiable in θ. Stochastic, smoothly-parameterized classes give us gradients.

Setup: Parametric Policies

Discounted MDP, $\gamma \in [0, 1)$. Class of parametric policies $\{\pi_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$. For a start distribution ρ, the objective is

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\rho), \qquad V^{\pi_\theta}(\rho) := \mathbb{E}_{s_0 \sim \rho}[V^{\pi_\theta}(s_0)].$$

Why stochastic policies? Deterministic policies are not differentiable in θ. Stochastic, smoothly-parameterized classes give us gradients.

Three running examples.

$$\text{Softmax (tabular):} \quad \pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

$$\text{Log-linear:} \quad \pi_\theta(a|s) \propto \exp(\theta \cdot \phi_{s,a}), \quad \phi_{s,a} \in \mathbb{R}^d.$$

$$\text{Neural softmax:} \quad \pi_\theta(a|s) \propto \exp(f_\theta(s, a)), \quad f_\theta \text{ a network.}$$

The closure of the softmax class contains *all* stationary policies, including deterministic ones (push $\theta \to \infty$). This will matter later.

The Policy Gradient Theorem

Let $\tau = (s_0, a_0, s_1, a_1, \dots)$ be a trajectory with $\mathbb{P}_\mu^{\pi_\theta}(\tau) = \mu(s_0) \prod_t \pi_\theta(a_t|s_t) P(s_{t+1}|s_t, a_t)$, and $R(\tau) = \sum_t \gamma^t r(s_t, a_t)$, so $V^{\pi_\theta}(\mu) = \mathbb{E}_\tau[R(\tau)]$.

Recall the discounted state-visitation distribution $d_\mu^\pi(s) := (1-\gamma) \sum_{t \geq 0} \gamma^t \, \mathbb{P}^\pi(s_t = s \mid s_0 \sim \mu)$ — the normalized discounted fraction of time π spends in s.

Theorem 9.4 (Three faces of the policy gradient)

$$\text{REINFORCE:} \quad \nabla V^{\pi_\theta}(\mu) = \mathbb{E}_\tau\Big[R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t|s_t)\Big]$$

$$\text{Action-value:} \quad \nabla V^{\pi_\theta}(\mu) = \tfrac{1}{1-\gamma} \, \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}\big[Q^{\pi_\theta}(s,a) \nabla \log \pi_\theta(a|s)\big]$$

$$\text{Advantage:} \quad \nabla V^{\pi_\theta}(\mu) = \tfrac{1}{1-\gamma} \, \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}\big[A^{\pi_\theta}(s,a) \nabla \log \pi_\theta(a|s)\big]$$

The Policy Gradient Theorem

Let $\tau = (s_0, a_0, s_1, a_1, \dots)$ be a trajectory with $\mathbb{P}_\mu^{\pi_\theta}(\tau) = \mu(s_0) \prod_t \pi_\theta(a_t|s_t) P(s_{t+1}|s_t, a_t)$, and $R(\tau) = \sum_t \gamma^t r(s_t, a_t)$, so $V^{\pi_\theta}(\mu) = \mathbb{E}_\tau[R(\tau)]$.

Recall the discounted state-visitation distribution $d_\mu^\pi(s) := (1-\gamma) \sum_{t \geq 0} \gamma^t \, \mathbb{P}^\pi(s_t = s \mid s_0 \sim \mu)$ — the normalized discounted fraction of time π spends in s.

Theorem 9.4 (Three faces of the policy gradient)

$$\text{REINFORCE:} \quad \nabla V^{\pi_\theta}(\mu) = \mathbb{E}_\tau\Big[R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t|s_t)\Big]$$

$$\text{Action-value:} \quad \nabla V^{\pi_\theta}(\mu) = \tfrac{1}{1-\gamma} \, \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}\big[Q^{\pi_\theta}(s,a) \nabla \log \pi_\theta(a|s)\big]$$

$$\text{Advantage:} \quad \nabla V^{\pi_\theta}(\mu) = \tfrac{1}{1-\gamma} \, \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}\big[A^{\pi_\theta}(s,a) \nabla \log \pi_\theta(a|s)\big]$$

Why three? REINFORCE needs only a trajectory's return — no model. The Q/advantage forms localize the gradient to the *state-action visitation* $d_\mu^{\pi_\theta}$ and have far lower variance. All three have the same expectation.

The key trick (likelihood ratio / "log-derivative"). $\nabla p = p \, \nabla \log p$. The dynamics P and the start μ do not depend on θ, so they vanish from $\nabla \log \mathbb{P}^{\pi_\theta}(\tau)$.

Proof of the REINFORCE and Q Forms

REINFORCE. Differentiate $V^{\pi_\theta}(\mu) = \sum_\tau R(\tau)\mathbb{P}_\mu^{\pi_\theta}(\tau)$ and apply the log-derivative trick:

$$\nabla V^{\pi_\theta}(\mu) = \sum_\tau R(\tau)\,\mathbb{P}_\mu^{\pi_\theta}(\tau)\,\nabla \log \mathbb{P}_\mu^{\pi_\theta}(\tau) = \mathbb{E}_\tau\Big[R(\tau)\sum_{t=0}^{\infty}\nabla \log \pi_\theta(a_t|s_t)\Big],$$

since
$\nabla \log \mathbb{P}_\mu^{\pi_\theta}(\tau) = \nabla\big(\log \mu(s_0) + \sum_t \log \pi_\theta(a_t|s_t) + \sum_t \log P(s_{t+1}|s_t, a_t)\big) = \sum_t \nabla \log \pi_\theta(a_t|s_t)$.

Proof of the REINFORCE and Q Forms

REINFORCE. Differentiate $V^{\pi_\theta}(\mu) = \sum_\tau R(\tau)\mathbb{P}_\mu^{\pi_\theta}(\tau)$ and apply the log-derivative trick:

$$\nabla V^{\pi_\theta}(\mu) = \sum_\tau R(\tau)\,\mathbb{P}_\mu^{\pi_\theta}(\tau)\,\nabla \log \mathbb{P}_\mu^{\pi_\theta}(\tau) = \mathbb{E}_\tau\Big[R(\tau)\sum_{t=0}^{\infty}\nabla \log \pi_\theta(a_t|s_t)\Big],$$

since
$\nabla \log \mathbb{P}_\mu^{\pi_\theta}(\tau) = \nabla\big(\log \mu(s_0) + \sum_t \log \pi_\theta(a_t|s_t) + \sum_t \log P(s_{t+1}|s_t,a_t)\big) = \sum_t \nabla \log \pi_\theta(a_t|s_t)$.

Q-form (recursion). For any start s_0, using $V^{\pi_\theta}(s_0) = \sum_{a_0}\pi_\theta(a_0|s_0)Q^{\pi_\theta}(s_0,a_0)$ and the product rule:

$$\begin{aligned}\nabla V^{\pi_\theta}(s_0) &= \sum_{a_0}\Big(\nabla\pi_\theta(a_0|s_0)\Big)Q^{\pi_\theta}(s_0,a_0) + \sum_{a_0}\pi_\theta(a_0|s_0)\nabla Q^{\pi_\theta}(s_0,a_0)\\ &= \mathbb{E}\big[Q^{\pi_\theta}(s_0,a_0)\nabla\log\pi_\theta(a_0|s_0)\big] + \gamma\,\mathbb{E}\big[\nabla V^{\pi_\theta}(s_1)\big],\end{aligned}$$

because $\nabla Q^{\pi_\theta}(s_0,a_0) = \nabla\big(r(s_0,a_0) + \gamma\sum_{s_1}P(s_1|s_0,a_0)V^{\pi_\theta}(s_1)\big) = \gamma\sum_{s_1}P\,\nabla V^{\pi_\theta}(s_1)$.

Proof of the REINFORCE and Q Forms

REINFORCE. Differentiate $V^{\pi_\theta}(\mu) = \sum_\tau R(\tau)\mathbb{P}_\mu^{\pi_\theta}(\tau)$ and apply the log-derivative trick:

$$\nabla V^{\pi_\theta}(\mu) = \sum_\tau R(\tau)\,\mathbb{P}_\mu^{\pi_\theta}(\tau)\,\nabla \log \mathbb{P}_\mu^{\pi_\theta}(\tau) = \mathbb{E}_\tau\Big[R(\tau)\sum_{t=0}^{\infty}\nabla \log \pi_\theta(a_t|s_t)\Big],$$

since
$\nabla \log \mathbb{P}_\mu^{\pi_\theta}(\tau) = \nabla\big(\log \mu(s_0) + \sum_t \log \pi_\theta(a_t|s_t) + \sum_t \log P(s_{t+1}|s_t,a_t)\big) = \sum_t \nabla \log \pi_\theta(a_t|s_t)$.

Q-form (recursion). For any start s_0, using $V^{\pi_\theta}(s_0) = \sum_{a_0}\pi_\theta(a_0|s_0)Q^{\pi_\theta}(s_0,a_0)$ and the product rule:

$$\begin{aligned}\nabla V^{\pi_\theta}(s_0) &= \sum_{a_0}\Big(\nabla\pi_\theta(a_0|s_0)\Big)Q^{\pi_\theta}(s_0,a_0) + \sum_{a_0}\pi_\theta(a_0|s_0)\nabla Q^{\pi_\theta}(s_0,a_0)\\ &= \mathbb{E}\big[Q^{\pi_\theta}(s_0,a_0)\nabla\log\pi_\theta(a_0|s_0)\big] + \gamma\,\mathbb{E}\big[\nabla V^{\pi_\theta}(s_1)\big],\end{aligned}$$

because $\nabla Q^{\pi_\theta}(s_0,a_0) = \nabla\big(r(s_0,a_0) + \gamma\sum_{s_1}P(s_1|s_0,a_0)V^{\pi_\theta}(s_1)\big) = \gamma\sum_{s_1}P\,\nabla V^{\pi_\theta}(s_1)$.

Unrolling the recursion over t and weighting by γ^t collapses the discounted sum of visitations into $\frac{1}{1-\gamma}d_\mu^{\pi_\theta}$, giving the Q-form.

Proof of the REINFORCE and Q Forms

REINFORCE. Differentiate $V^{\pi_\theta}(\mu) = \sum_\tau R(\tau)\mathbb{P}_\mu^{\pi_\theta}(\tau)$ and apply the log-derivative trick:

$$\nabla V^{\pi_\theta}(\mu) = \sum_\tau R(\tau)\,\mathbb{P}_\mu^{\pi_\theta}(\tau)\,\nabla \log \mathbb{P}_\mu^{\pi_\theta}(\tau) = \mathbb{E}_\tau\Big[R(\tau)\sum_{t=0}^{\infty}\nabla \log \pi_\theta(a_t|s_t)\Big],$$

since
$\nabla \log \mathbb{P}_\mu^{\pi_\theta}(\tau) = \nabla\big(\log \mu(s_0) + \sum_t \log \pi_\theta(a_t|s_t) + \sum_t \log P(s_{t+1}|s_t,a_t)\big) = \sum_t \nabla \log \pi_\theta(a_t|s_t)$.

Q-form (recursion). For any start s_0, using $V^{\pi_\theta}(s_0) = \sum_{a_0}\pi_\theta(a_0|s_0)Q^{\pi_\theta}(s_0,a_0)$ and the product rule:

$$\begin{aligned}\nabla V^{\pi_\theta}(s_0) &= \sum_{a_0}\Big(\nabla\pi_\theta(a_0|s_0)\Big)Q^{\pi_\theta}(s_0,a_0) + \sum_{a_0}\pi_\theta(a_0|s_0)\nabla Q^{\pi_\theta}(s_0,a_0)\\ &= \mathbb{E}\big[Q^{\pi_\theta}(s_0,a_0)\nabla \log \pi_\theta(a_0|s_0)\big] + \gamma\,\mathbb{E}\big[\nabla V^{\pi_\theta}(s_1)\big],\end{aligned}$$

because $\nabla Q^{\pi_\theta}(s_0,a_0) = \nabla\big(r(s_0,a_0) + \gamma\sum_{s_1}P(s_1|s_0,a_0)V^{\pi_\theta}(s_1)\big) = \gamma\sum_{s_1}P\,\nabla V^{\pi_\theta}(s_1)$.

Unrolling the recursion over t and weighting by γ^t collapses the discounted sum of visitations into $\frac{1}{1-\gamma}d_\mu^{\pi_\theta}$, giving the Q-form.

Advantage form — and why the baseline is arbitrary. For *any* action-independent $b(s)$, $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[b(s)\nabla \log \pi_\theta(a|s)] = b(s)\sum_a \nabla\pi_\theta(a|s) = b(s)\nabla 1 = 0$. So we may subtract any baseline $b(s)$ from Q^{π_θ} without changing the gradient:

$$\nabla V^{\pi_\theta}(\mu) = \tfrac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[(Q^{\pi_\theta}(s,a) - b(s))\,\nabla \log \pi_\theta(a|s)\big].$$

The advantage form is just the choice $b(s) = V^{\pi_\theta}(s)$, so $Q^{\pi_\theta} - V^{\pi_\theta} = A^{\pi_\theta}$. A^{π_θ} is not special — it is one (variance-reducing) baseline among infinitely many unbiased ones. □

The Objective is Non-Concave

Lemma 9.5 (Non-concavity)

There is an MDP for which $\theta \mapsto V^{\pi_\theta}(s)$ is *not* concave, for both the direct and softmax parameterizations.

The Objective is Non-Concave

Lemma 9.5 (Non-concavity)

There is an MDP for which $\theta \mapsto V^{\pi_\theta}(s)$ is *not* concave, for both the direct and softmax parameterizations.

Witness (a deterministic tree MDP). Two decision states; reaching the reward requires the "right" action in *both*, so $V^{\pi}(s_1) = \pi(a_2|s_1)\,\pi(a_2|s_2) \cdot r$ — a product of probabilities.

Pick $\theta^{(1)}, \theta^{(2)}$ symmetric about 0 and let $\theta^{(\mathrm{mid})} = \frac{1}{2}(\theta^{(1)} + \theta^{(2)})$. One computes

$$V^{(1)}(s_1) = \tfrac{9}{16} r, \quad V^{(2)}(s_1) = \tfrac{1}{16} r, \quad V^{(\mathrm{mid})}(s_1) = \tfrac{1}{4} r \;\Rightarrow\; V^{(1)} + V^{(2)} > 2V^{(\mathrm{mid})}.$$

So V violates concavity along this segment.

The Objective is Non-Concave

Lemma 9.5 (Non-concavity)

There is an MDP for which $\theta \mapsto V^{\pi_\theta}(s)$ is *not* concave, for both the direct and softmax parameterizations.

Witness (a deterministic tree MDP). Two decision states; reaching the reward requires the "right" action in *both*, so $V^{\pi}(s_1) = \pi(a_2|s_1)\,\pi(a_2|s_2) \cdot r$ — a product of probabilities.

Pick $\theta^{(1)}, \theta^{(2)}$ symmetric about 0 and let $\theta^{(\mathrm{mid})} = \frac{1}{2}(\theta^{(1)} + \theta^{(2)})$. One computes

$$V^{(1)}(s_1) = \tfrac{9}{16} r, \quad V^{(2)}(s_1) = \tfrac{1}{16} r, \quad V^{(\mathrm{mid})}(s_1) = \tfrac{1}{4} r \;\Rightarrow\; V^{(1)} + V^{(2)} > 2V^{(\mathrm{mid})}.$$

So V violates concavity along this segment.

Consequence

The standard convex-optimization toolkit *does not apply*. Gradient ascent, in general, only guarantees convergence to a stationary point ($\|\nabla V\| \to 0$). The rest of the lecture is about *when stationary* $\Rightarrow$ *global*.

Convergence to Stationary Points

We will treat these as black-box facts from non-convex optimization (Ghadimi–Lan 2013; Beck 2017). Call f β-smooth if $\|\nabla f(w) - \nabla f(w')\| \leq \beta \|w - w'\|$.

Lemma 9.6 (Exact gradients)

If V^{π_θ} is β-smooth and bounded by $V^\star$, gradient ascent with $\eta = 1/\beta$ satisfies

$$\min_{t \leq T} \left\| \nabla V^{(t)}(\mu) \right\|^2 \leq \frac{2\beta\left(V^\star(\mu) - V^{(0)}(\mu)\right)}{T}.$$

Convergence to Stationary Points

We will treat these as black-box facts from non-convex optimization (Ghadimi–Lan 2013; Beck 2017). Call f β-smooth if $\|\nabla f(w) - \nabla f(w')\| \leq \beta \|w - w'\|$.

Lemma 9.6 (Exact gradients)

If V^{π_θ} is β-smooth and bounded by $V^\star$, gradient ascent with $\eta = 1/\beta$ satisfies

$$\min_{t \leq T} \left\| \nabla V^{(t)}(\mu) \right\|^2 \leq \frac{2\beta \left(V^\star(\mu) - V^{(0)}(\mu) \right)}{T}.$$

Lemma 9.8 (Stochastic gradients)

If additionally $\mathbb{E} \left\| \widehat{\nabla V} - \nabla V \right\|^2 \leq \sigma^2$, an SGD schedule gives

$$\min_{t \leq T} \mathbb{E} \left\| \nabla V^{(t)}(\mu) \right\|^2 \leq \frac{2\beta (V^\star(\mu) - V^{(0)}(\mu))}{T} + \sqrt{\frac{2\sigma^2}{T}}.$$

Convergence to Stationary Points

We will treat these as black-box facts from non-convex optimization (Ghadimi–Lan 2013; Beck 2017). Call f β-smooth if $\|\nabla f(w) - \nabla f(w')\| \le \beta \|w - w'\|$.

Lemma 9.6 (Exact gradients)

If V^{π_θ} is β-smooth and bounded by $V^\star$, gradient ascent with $\eta = 1/\beta$ satisfies

$$\min_{t \le T} \left\| \nabla V^{(t)}(\mu) \right\|^2 \le \frac{2\beta\left(V^\star(\mu) - V^{(0)}(\mu)\right)}{T}.$$

Lemma 9.8 (Stochastic gradients)

If additionally $\mathbb{E}\left\| \widehat{\nabla V} - \nabla V \right\|^2 \le \sigma^2$, an SGD schedule gives

$$\min_{t \le T} \mathbb{E} \left\| \nabla V^{(t)}(\mu) \right\|^2 \le \frac{2\beta(V^\star(\mu) - V^{(0)}(\mu))}{T} + \sqrt{\frac{2\sigma^2}{T}}.$$

This is all unconstrained non-convexity gives us. A small gradient does *not* certify a good policy — the next chapter shows gradients can be exponentially small at terrible policies.

Proof of Lemma 9.6 (Smoothness $\Rightarrow$ Small Gradients)

The whole argument is the descent lemma plus a telescoping sum.

Step 1: one ascent step makes guaranteed progress. β-smoothness implies the quadratic bound $V(\theta') \geq V(\theta) + \nabla V(\theta) \cdot (\theta' - \theta) - \frac{\beta}{2} \|\theta' - \theta\|^2$. Plug in the gradient step $\theta' = \theta + \frac{1}{\beta} \nabla V(\theta)$:

$$V^{(t+1)}(\mu) \;\geq\; V^{(t)}(\mu) + \tfrac{1}{\beta} \left\|\nabla V^{(t)}(\mu)\right\|^2 - \tfrac{\beta}{2} \cdot \tfrac{1}{\beta^2} \left\|\nabla V^{(t)}(\mu)\right\|^2 \;=\; V^{(t)}(\mu) + \tfrac{1}{2\beta} \left\|\nabla V^{(t)}(\mu)\right\|^2.$$

Proof of Lemma 9.6 (Smoothness $\Rightarrow$ Small Gradients)

The whole argument is the descent lemma plus a telescoping sum.

Step 1: one ascent step makes guaranteed progress. β-smoothness implies the quadratic bound $V(\theta') \geq V(\theta) + \nabla V(\theta) \cdot (\theta' - \theta) - \frac{\beta}{2} \|\theta' - \theta\|^2$. Plug in the gradient step $\theta' = \theta + \frac{1}{\beta} \nabla V(\theta)$:

$$V^{(t+1)}(\mu) \;\geq\; V^{(t)}(\mu) + \tfrac{1}{\beta} \left\|\nabla V^{(t)}(\mu)\right\|^2 - \tfrac{\beta}{2} \cdot \tfrac{1}{\beta^2} \left\|\nabla V^{(t)}(\mu)\right\|^2 \;=\; V^{(t)}(\mu) + \tfrac{1}{2\beta} \left\|\nabla V^{(t)}(\mu)\right\|^2.$$

Step 2: telescope. Sum over $t = 0, \ldots, T-1$. The values telescope and are bounded by $V^\star$:

$$\tfrac{1}{2\beta} \sum_{t=0}^{T-1} \left\|\nabla V^{(t)}(\mu)\right\|^2 \;\leq\; V^{(T)}(\mu) - V^{(0)}(\mu) \;\leq\; V^\star(\mu) - V^{(0)}(\mu).$$

Proof of Lemma 9.6 (Smoothness $\Rightarrow$ Small Gradients)

The whole argument is the descent lemma plus a telescoping sum.

Step 1: one ascent step makes guaranteed progress. β-smoothness implies the quadratic bound $V(\theta') \geq V(\theta) + \nabla V(\theta) \cdot (\theta' - \theta) - \frac{\beta}{2} \|\theta' - \theta\|^2$. Plug in the gradient step $\theta' = \theta + \frac{1}{\beta} \nabla V(\theta)$:

$$V^{(t+1)}(\mu) \;\geq\; V^{(t)}(\mu) + \tfrac{1}{\beta} \left\|\nabla V^{(t)}(\mu)\right\|^2 - \tfrac{\beta}{2} \cdot \tfrac{1}{\beta^2} \left\|\nabla V^{(t)}(\mu)\right\|^2 \;=\; V^{(t)}(\mu) + \tfrac{1}{2\beta} \left\|\nabla V^{(t)}(\mu)\right\|^2.$$

Step 2: telescope. Sum over $t = 0, \ldots, T-1$. The values telescope and are bounded by $V^\star$:

$$\tfrac{1}{2\beta} \sum_{t=0}^{T-1} \left\|\nabla V^{(t)}(\mu)\right\|^2 \;\leq\; V^{(T)}(\mu) - V^{(0)}(\mu) \;\leq\; V^\star(\mu) - V^{(0)}(\mu).$$

Step 3: the min is below the average.

$$\min_{t \leq T} \left\|\nabla V^{(t)}(\mu)\right\|^2 \;\leq\; \tfrac{1}{T} \sum_{t=0}^{T-1} \left\|\nabla V^{(t)}(\mu)\right\|^2 \;\leq\; \frac{2\beta\left(V^\star(\mu) - V^{(0)}(\mu)\right)}{T}. \qquad \square$$

(Lemma 9.8 is the same argument carrying the variance σ^2 through the expectation of the descent lemma.)

Sampling the Gradient: REINFORCE in Practice

We rarely have exact gradients. From a single sampled trajectory $\tau \sim \mathbb{P}_\mu^{\pi_\theta}$, form the reward-to-go estimate

$$\widehat{Q}^{\pi_\theta}(s_t, a_t) := \sum_{t' \geq t} \gamma^{t'-t} r(s_{t'}, a_{t'}), \qquad \widehat{\nabla V}^{\pi_\theta}(\mu) := \sum_{t=0}^{\infty} \gamma^t \, \widehat{Q}^{\pi_\theta}(s_t, a_t) \, \nabla \log \pi_\theta(a_t | s_t).$$

Lemma 9.7 (Unbiasedness)

$\mathbb{E}_\tau \left[\widehat{\nabla V}^{\pi_\theta}(\mu) \right] = \nabla V^{\pi_\theta}(\mu).$

Sampling the Gradient: REINFORCE in Practice

We rarely have exact gradients. From a single sampled trajectory $\tau \sim \mathbb{P}_\mu^{\pi_\theta}$, form the reward-to-go estimate

$$\widehat{Q}^{\pi_\theta}(s_t, a_t) := \sum_{t' \geq t} \gamma^{t'-t} r(s_{t'}, a_{t'}), \qquad \widehat{\nabla V}^{\pi_\theta}(\mu) := \sum_{t=0}^{\infty} \gamma^t \, \widehat{Q}^{\pi_\theta}(s_t, a_t) \, \nabla \log \pi_\theta(a_t | s_t).$$

Lemma 9.7 (Unbiasedness)

$$\mathbb{E}_\tau \left[\widehat{\nabla V}^{\pi_\theta}(\mu) \right] = \nabla V^{\pi_\theta}(\mu).$$

Proof. By the tower property, $\mathbb{E}[\widehat{Q}^{\pi_\theta}(s_t, a_t) \mid s_t, a_t] = Q^{\pi_\theta}(s_t, a_t)$ (Markov), so the estimator matches the Q-form of Theorem 9.4 in expectation. □

Sampling the Gradient: REINFORCE in Practice

We rarely have exact gradients. From a single sampled trajectory $\tau \sim \mathbb{P}_\mu^{\pi_\theta}$, form the reward-to-go estimate

$$\widehat{Q}^{\pi_\theta}(s_t, a_t) := \sum_{t' \geq t} \gamma^{t'-t} r(s_{t'}, a_{t'}), \qquad \widehat{\nabla V}^{\pi_\theta}(\mu) := \sum_{t=0}^{\infty} \gamma^t \, \widehat{Q}^{\pi_\theta}(s_t, a_t) \, \nabla \log \pi_\theta(a_t | s_t).$$

Lemma 9.7 (Unbiasedness)

$$\mathbb{E}_\tau \left[\widehat{\nabla V}^{\pi_\theta}(\mu) \right] = \nabla V^{\pi_\theta}(\mu).$$

Proof. By the tower property, $\mathbb{E}[\widehat{Q}^{\pi_\theta}(s_t, a_t) \mid s_t, a_t] = Q^{\pi_\theta}(s_t, a_t)$ (Markov), so the estimator matches the Q-form of Theorem 9.4 in expectation. □

This yields a model-free stochastic gradient ascent: sample τ, take a step along $\widehat{\nabla V}$. (Truncate the infinite trajectory to control bias.)

Sampling the Gradient: REINFORCE in Practice

We rarely have exact gradients. From a single sampled trajectory $\tau \sim \mathbb{P}_\mu^{\pi_\theta}$, form the reward-to-go estimate

$$\widehat{Q}^{\pi_\theta}(s_t, a_t) := \sum_{t' \geq t} \gamma^{t'-t} r(s_{t'}, a_{t'}), \qquad \widehat{\nabla V}^{\pi_\theta}(\mu) := \sum_{t=0}^{\infty} \gamma^t \, \widehat{Q}^{\pi_\theta}(s_t, a_t) \, \nabla \log \pi_\theta(a_t | s_t).$$

Lemma 9.7 (Unbiasedness)

$\mathbb{E}_\tau \left[\widehat{\nabla V}^{\pi_\theta}(\mu) \right] = \nabla V^{\pi_\theta}(\mu).$

Proof. By the tower property, $\mathbb{E}[\widehat{Q}^{\pi_\theta}(s_t, a_t) \mid s_t, a_t] = Q^{\pi_\theta}(s_t, a_t)$ (Markov), so the estimator matches the Q-form of Theorem 9.4 in expectation. □

This yields a model-free stochastic gradient ascent: sample τ, take a step along $\widehat{\nabla V}$. (Truncate the infinite trajectory to control bias.)

The practical pain point: variance

σ^2 is typically *huge* — $\widehat{Q}$ is a Monte Carlo return over a whole trajectory. Variance reduction is essential in practice.

Baselines: Free Variance Reduction

Subtract a state-dependent baseline $f(s)$ from the return:

$$\widehat{\nabla V}^{\pi_\theta}(\mu) := \sum_{t=0}^{\infty} \gamma^t \left(\widehat{Q}^{\pi_\theta}(s_t, a_t) - f(s_t)\right) \nabla \log \pi_\theta(a_t|s_t).$$

Lemma 9.9 (Baselines preserve unbiasedness)

If the data used to build f is independent of τ, then for *any* $f : \mathcal{S} \to \mathbb{R}$,

$$\mathbb{E}\left[\sum_t \gamma^t \left(\widehat{Q}^{\pi_\theta}(s_t, a_t) - f(s_t)\right) \nabla \log \pi_\theta(a_t|s_t)\right] = \nabla V^{\pi_\theta}(\mu).$$

Baselines: Free Variance Reduction

Subtract a state-dependent baseline $f(s)$ from the return:

$$\widehat{\nabla V}^{\pi_\theta}(\mu) := \sum_{t=0}^{\infty} \gamma^t \big(\widehat{Q}^{\pi_\theta}(s_t, a_t) - f(s_t)\big) \nabla \log \pi_\theta(a_t|s_t).$$

Lemma 9.9 (Baselines preserve unbiasedness)

If the data used to build f is independent of τ, then for *any* $f : \mathcal{S} \to \mathbb{R}$,

$$\mathbb{E}\Big[\sum_t \gamma^t \big(\widehat{Q}^{\pi_\theta}(s_t, a_t) - f(s_t)\big) \nabla \log \pi_\theta(a_t|s_t)\Big] = \nabla V^{\pi_\theta}(\mu).$$

Proof. For any state-only g,
$\mathbb{E}_{a\sim\pi(\cdot|s)}[\nabla \log \pi(a|s)\, g(s)] = g(s)\nabla \sum_a \pi(a|s) = g(s)\nabla 1 = 0$. So subtracting $f(s_t)$ adds zero in expectation; combine with Lemma 9.7. □

Baselines: Free Variance Reduction

Subtract a state-dependent baseline $f(s)$ from the return:

$$\widehat{\nabla V}^{\pi_\theta}(\mu) := \sum_{t=0}^{\infty} \gamma^t \big(\widehat{Q}^{\pi_\theta}(s_t, a_t) - f(s_t)\big) \nabla \log \pi_\theta(a_t|s_t).$$

Lemma 9.9 (Baselines preserve unbiasedness)

If the data used to build f is independent of τ, then for *any* $f : \mathcal{S} \to \mathbb{R}$,

$$\mathbb{E}\Big[\sum_t \gamma^t \big(\widehat{Q}^{\pi_\theta}(s_t, a_t) - f(s_t)\big) \nabla \log \pi_\theta(a_t|s_t)\Big] = \nabla V^{\pi_\theta}(\mu).$$

Proof. For any state-only g,
$\mathbb{E}_{a\sim\pi(\cdot|s)}[\nabla \log \pi(a|s)\, g(s)] = g(s)\nabla \sum_a \pi(a|s) = g(s)\nabla 1 = 0$. So subtracting $f(s_t)$ adds zero in expectation; combine with Lemma 9.7. □

Best choice. Take $f(s) \approx V^{\pi_\theta}(s)$, so the multiplier becomes $\approx A^{\pi_\theta}(s, a)$ — "how much better than average is this action." This is exactly the actor–critic idea: learn a critic $f \approx V$ to reduce the actor's gradient variance.

Today's Plan

1. Policy Gradients (Ch. 9)
2. Global Optimality & the NPG (Ch. 10)
3. Function Approximation & the NPG (Ch. 11)
4. Trust-Region Methods: TRPO & PPO (Ch. 12)
5. Summary

The Setting: Tabular Softmax, Exact Gradients

Tabular. Finite $\mathcal{S}$ and $\mathcal{A}$. We give the policy one free parameter per state-action pair: $\theta = (\theta_{s,a})_{s \in \mathcal{S}, a \in \mathcal{A}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

Softmax (this setting). Each state's action distribution is an independent softmax over its own parameters:

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}.$$

There is *no sharing* across states — it is the tabular analogue of a lookup table of logits.

The Setting: Tabular Softmax, Exact Gradients

Tabular. Finite $\mathcal{S}$ and $\mathcal{A}$. We give the policy one free parameter per state-action pair: $\theta = (\theta_{s,a})_{s \in \mathcal{S}, a \in \mathcal{A}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

Softmax (this setting). Each state's action distribution is an independent softmax over its own parameters:

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}.$$

There is *no sharing* across states — it is the tabular analogue of a lookup table of logits.

Two consequences we will lean on.

- Fully expressive: the closure of this class contains *every* stationary policy (deterministic ones are reached as $\theta \to \infty$). So there is *no approximation error* — any suboptimality is purely an *optimization* failure.
- Still non-concave (Lemma 9.5), so this is a real question even with infinite data.

The Setting: Tabular Softmax, Exact Gradients

Tabular. Finite $\mathcal{S}$ and $\mathcal{A}$. We give the policy one free parameter per state-action pair: $\theta = (\theta_{s,a})_{s \in \mathcal{S}, a \in \mathcal{A}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

Softmax (this setting). Each state's action distribution is an independent softmax over its own parameters:

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}.$$

There is *no sharing* across states — it is the tabular analogue of a lookup table of logits.

Two consequences we will lean on.

- Fully expressive: the closure of this class contains *every* stationary policy (deterministic ones are reached as $\theta \to \infty$). So there is *no approximation error* — any suboptimality is purely an *optimization* failure.
- Still non-concave (Lemma 9.5), so this is a real question even with infinite data.

We also assume exact gradients this chapter: we isolate the optimization geometry from statistical estimation (that is Ch. 11).

Optimizing Under $\mu \neq \rho$ is a Restart Assumption

The recurring device. Although we ultimately care about $V^{\pi_\theta}(\rho)$, it helps to optimize a *surrogate* objective under a possibly different measure μ:

$$\max_\theta \; V^{\pi_\theta}(\mu), \qquad \text{but report } V^{\pi_\theta}(\rho),$$

bridged by the distribution mismatch coefficient $\left\| d_\rho^{\pi^\star} / \mu \right\|_\infty$.

Optimizing Under $\mu \neq \rho$ is a Restart Assumption

The recurring device. Although we ultimately care about $V^{\pi_\theta}(\rho)$, it helps to optimize a *surrogate* objective under a possibly different measure μ:

$$\max_\theta \; V^{\pi_\theta}(\mu), \qquad \text{but report } V^{\pi_\theta}(\rho),$$

bridged by the distribution mismatch coefficient $\left\| d_\rho^{\pi^\star} / \mu \right\|_\infty$.

What does "optimize under μ" physically require? (good question)

To ascend $V^{\pi_\theta}(\mu)$ we need its gradient, and every gradient expression (Thm 9.4) is an expectation over trajectories started at $s_0 \sim \mu$. So we must be able to *reset the system to* μ — an exploratory restart distribution, a mild generative-model-like ability.

If we cannot choose the start state, we are forced to take $\mu = \rho$. The bound then carries $\left\| d_\rho^{\pi^\star} / \rho \right\|_\infty$, which the learner *cannot* make small (e.g. ρ a point mass far from where $\pi^\star$ lives). The freedom to pick a spread-out μ is exactly what tames the mismatch coefficient.

Optimizing Under $\mu \neq \rho$ is a Restart Assumption

The recurring device. Although we ultimately care about $V^{\pi_\theta}(\rho)$, it helps to optimize a *surrogate* objective under a possibly different measure μ:

$$\max_\theta \; V^{\pi_\theta}(\mu), \qquad \text{but report } V^{\pi_\theta}(\rho),$$

bridged by the distribution mismatch coefficient $\left\| d_\rho^{\pi^\star} / \mu \right\|_\infty$.

What does "optimize under μ" physically require? (good question)

To ascend $V^{\pi_\theta}(\mu)$ we need its gradient, and every gradient expression (Thm 9.4) is an expectation over trajectories started at $s_0 \sim \mu$. So we must be able to *reset the system to* μ — an exploratory restart distribution, a mild generative-model-like ability.

If we cannot choose the start state, we are forced to take $\mu = \rho$. The bound then carries $\left\| d_\rho^{\pi^\star} / \rho \right\|_\infty$, which the learner *cannot* make small (e.g. ρ a point mass far from where $\pi^\star$ lives). The freedom to pick a spread-out μ is exactly what tames the mismatch coefficient.

Three algorithms, increasing power: (1) plain GA — asymptotic only; (2) log-barrier GA — polynomial, but $|\mathcal{S}|, |\mathcal{A}|$ factors; (3) NPG — $O(1/T)$, dimension-free.

Why Optimize Under $\mu \neq \rho$: Vanishing Gradients

Proposition 10.1 (Vanishing gradients at bad policies)

There is a length-$(H+2)$ chain MDP ($\gamma = H/(H+1)$) and a large set of policies π_θ where, for all $k \leq \frac{H}{40 \log(2H)} - 1$,

$$\left\|\nabla_\theta^k V^{\pi_\theta}(s_0)\right\| \leq (1/3)^{H/4}, \qquad \text{yet} \qquad V^\star(s_0) - V^{\pi_\theta}(s_0) \geq \frac{H+1}{8} - \frac{(H+1)^2}{3^H}.$$

Why Optimize Under $\mu \neq \rho$: Vanishing Gradients

Proposition 10.1 (Vanishing gradients at bad policies)

There is a length-$(H+2)$ chain MDP ($\gamma = H/(H+1)$) and a large set of policies π_θ where, for all $k \leq \frac{H}{40 \log(2H)} - 1$,

$$\left\|\nabla_\theta^k V^{\pi_\theta}(s_0)\right\| \leq (1/3)^{H/4}, \qquad \text{yet} \qquad V^\star(s_0) - V^{\pi_\theta}(s_0) \geq \frac{H+1}{8} - \frac{(H+1)^2}{3^H}.$$

Reading it. Not just the gradient — the first $\Omega(H/\log H)$ *derivatives* are exponentially small, at policies that are highly suboptimal. "Escape-from-saddle" results do not rescue us.

Intuition. Sparse reward at the end of a chain; a policy that doesn't reach the goal sees a near-flat objective. Exploration must be *built in* — here, via a good μ that puts mass where reward lives.

Why Optimize Under $\mu \neq \rho$: Vanishing Gradients

Proposition 10.1 (Vanishing gradients at bad policies)

There is a length-$(H+2)$ chain MDP ($\gamma = H/(H+1)$) and a large set of policies π_θ where, for all $k \leq \frac{H}{40 \log(2H)} - 1$,

$$\left\|\nabla_\theta^k V^{\pi_\theta}(s_0)\right\| \leq (1/3)^{H/4}, \qquad \text{yet} \qquad V^\star(s_0) - V^{\pi_\theta}(s_0) \geq \frac{H+1}{8} - \frac{(H+1)^2}{3^H}.$$

Reading it. Not just the gradient — the first $\Omega(H/\log H)$ *derivatives* are exponentially small, at policies that are highly suboptimal. "Escape-from-saddle" results do not rescue us.

Intuition. Sparse reward at the end of a chain; a policy that doesn't reach the goal sees a near-flat objective. Exploration must be *built in* — here, via a good μ that puts mass where reward lives.

Moral: some condition on the state distribution (equivalently, on μ) is *necessary* for stationarity to imply optimality.

The Softmax Policy Gradient

For the softmax parameterization $\pi_\theta(a|s) \propto \exp(\theta_{s,a})$,

$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_{s',a'}} = \mathbf{1}[s = s'](\mathbf{1}[a = a'] - \pi_\theta(a'|s)).$$

Lemma 10.2 (Softmax gradient)

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\, \pi_\theta(a|s)\, A^{\pi_\theta}(s,a).$$

The Softmax Policy Gradient

For the softmax parameterization $\pi_\theta(a|s) \propto \exp(\theta_{s,a})$,

$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_{s',a'}} = \mathbf{1}[s = s'](\mathbf{1}[a = a'] - \pi_\theta(a'|s)).$$

Lemma 10.2 (Softmax gradient)

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\, \pi_\theta(a|s)\, A^{\pi_\theta}(s,a).$$

Proof. Plug the partial of $\log \pi_\theta$ into the advantage form of the policy gradient; the indicator collapses the state sum to s, and the $-\pi_\theta(a'|s)$ term contributes $\mathbb{E}_{a\sim\pi}[A^{\pi_\theta}(s,a)] = 0$. □

The Softmax Policy Gradient

For the softmax parameterization $\pi_\theta(a|s) \propto \exp(\theta_{s,a})$,

$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_{s',a'}} = \mathbf{1}[s = s'](\mathbf{1}[a = a'] - \pi_\theta(a'|s)).$$

Lemma 10.2 (Softmax gradient)

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\, \pi_\theta(a|s)\, A^{\pi_\theta}(s,a).$$

Proof. Plug the partial of $\log \pi_\theta$ into the advantage form of the policy gradient; the indicator collapses the state sum to s, and the $-\pi_\theta(a'|s)$ term contributes $\mathbb{E}_{a\sim\pi}[A^{\pi_\theta}(s,a)] = 0$. □

Two structural warnings hidden in this formula

- The factor $d_\mu^{\pi_\theta}(s)$: states the current policy rarely visits get a *tiny* effective learning rate. (This is the vanishing-gradient mechanism.)
- The factor $\pi_\theta(a|s)$: as the policy becomes near-deterministic, gradients $\to 0$. Convergence can be exponentially slow.

Plain Gradient Ascent: Asymptotic Global Convergence

Theorem 10.3 (Asymptotic convergence, softmax)

Run $\theta^{(t+1)} = \theta^{(t)} + \eta \nabla V^{(t)}(\mu)$ with $\mu(s) > 0$ for all s and $\eta \leq \frac{(1-\gamma)^3}{8}$. Then $V^{(t)}(s) \to V^\star(s)$ for all s as $t \to \infty$.

(Proof is technical; omitted — see AJKS §10.5.)

Plain Gradient Ascent: Asymptotic Global Convergence

Theorem 10.3 (Asymptotic convergence, softmax)

Run $\theta^{(t+1)} = \theta^{(t)} + \eta \nabla V^{(t)}(\mu)$ with $\mu(s) > 0$ for all s and $\eta \leq \frac{(1-\gamma)^3}{8}$. Then $V^{(t)}(s) \to V^\star(s)$ for all s as $t \to \infty$.

(Proof is technical; omitted — see AJKS §10.5.)

What it does not give.

- No rate. There is strong evidence the worst-case rate is exponentially slow in $|\mathcal{S}|$ and the mismatch coefficient.
- Needs $\mu(s) > 0$ everywhere — consistent with the vanishing-gradient story: $d_\mu^{\pi_\theta}(s)$ scales down updates at rarely-visited states.

Plain Gradient Ascent: Asymptotic Global Convergence

Theorem 10.3 (Asymptotic convergence, softmax)

Run $\theta^{(t+1)} = \theta^{(t)} + \eta \nabla V^{(t)}(\mu)$ with $\mu(s) > 0$ for all s and $\eta \leq \frac{(1-\gamma)^3}{8}$. Then $V^{(t)}(s) \to V^\star(s)$ for all s as $t \to \infty$.

(Proof is technical; omitted — see AJKS §10.5.)

What it does not give.

- No rate. There is strong evidence the worst-case rate is exponentially slow in $|\mathcal{S}|$ and the mismatch coefficient.
- Needs $\mu(s) > 0$ everywhere — consistent with the vanishing-gradient story: $d_\mu^{\pi_\theta}(s)$ scales down updates at rarely-visited states.

So: plain GA is globally convergent but possibly glacial. Regularization fixes the rate.

Log-Barrier Regularization

Add a relative-entropy barrier against the uniform policy (this is *not* the entropy regularizer — it is far more aggressive at small probabilities):

$$L_\lambda(\theta) := V^{\pi_\theta}(\mu) - \lambda\, \mathbb{E}_{s\sim \mathrm{Unif}_{\mathcal{S}}} \mathrm{KL}\big(\mathrm{Unif}_{\mathcal{A}}, \pi_\theta(\cdot|s)\big) = V^{\pi_\theta}(\mu) + \frac{\lambda}{|\mathcal{S}||\mathcal{A}|} \sum_{s,a} \log \pi_\theta(a|s) + \text{const.}$$

Theorem 10.4 (Stationary points of L_λ are near-optimal)

If $\|\nabla_\theta L_\lambda(\theta)\|_2 \le \varepsilon_{\mathrm{opt}}$ with $\varepsilon_{\mathrm{opt}} \le \lambda/(2|\mathcal{S}||\mathcal{A}|)$, then for *every* ρ:

$$V^{\pi_\theta}(\rho) \;\ge\; V^{\star}(\rho) - \frac{2\lambda}{1-\gamma} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty .$$

Log-Barrier Regularization

Add a relative-entropy barrier against the uniform policy (this is *not* the entropy regularizer — it is far more aggressive at small probabilities):

$$L_\lambda(\theta) := V^{\pi_\theta}(\mu) - \lambda\, \mathbb{E}_{s\sim \mathrm{Unif}_{\mathcal{S}}} \mathrm{KL}\big(\mathrm{Unif}_{\mathcal{A}}, \pi_\theta(\cdot|s)\big) = V^{\pi_\theta}(\mu) + \frac{\lambda}{|\mathcal{S}||\mathcal{A}|} \sum_{s,a} \log \pi_\theta(a|s) + \text{const.}$$

Theorem 10.4 (Stationary points of L_λ are near-optimal)

If $\|\nabla_\theta L_\lambda(\theta)\|_2 \le \varepsilon_{\mathrm{opt}}$ with $\varepsilon_{\mathrm{opt}} \le \lambda/(2|\mathcal{S}||\mathcal{A}|)$, then for *every* ρ:

$$V^{\pi_\theta}(\rho) \;\ge\; V^{\star}(\rho) - \frac{2\lambda}{1-\gamma} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty .$$

The barrier keeps every $\pi_\theta(a|s) \gtrsim \frac{1}{2|\mathcal{A}|}$, so no action probability collapses — exactly what kills the $\pi_\theta(a|s)$ factor that stalls plain GA. The price is the distribution mismatch coefficient $\|d_\rho^{\pi^\star}/\mu\|_\infty$.

Proof of Theorem 10.4

Goal. Show $\max_a A^{\pi_\theta}(s,a) \leq \frac{2\lambda}{\mu(s)|\mathcal{S}|}$ for all s. Given this, the performance-difference lemma finishes it:

$$V^\star(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\sum_a \pi^\star(a|s)A^{\pi_\theta}(s,a) \leq \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\max_a A^{\pi_\theta}(s,a)$$

$$\leq \frac{2\lambda}{1-\gamma}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty.$$

Proof of Theorem 10.4

Goal. Show $\max_a A^{\pi_\theta}(s,a) \leq \frac{2\lambda}{\mu(s)|\mathcal{S}|}$ for all s. Given this, the performance-difference lemma finishes it:

$$V^\star(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\sum_a \pi^\star(a|s)A^{\pi_\theta}(s,a) \leq \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\max_a A^{\pi_\theta}(s,a)$$

$$\leq \frac{2\lambda}{1-\gamma}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty.$$

Now the claim. Fix (s,a) with $A^{\pi_\theta}(s,a) \geq 0$ (else trivial). From Lemma 10.2, the regularized gradient is

$$\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right).$$

Proof of Theorem 10.4

Goal. Show $\max_a A^{\pi_\theta}(s,a) \le \frac{2\lambda}{\mu(s)|\mathcal{S}|}$ for all s. Given this, the performance-difference lemma finishes it:

$$V^\star(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\sum_a \pi^\star(a|s)A^{\pi_\theta}(s,a) \;\le\; \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\max_a A^{\pi_\theta}(s,a)$$

$$\le \frac{2\lambda}{1-\gamma}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty.$$

Now the claim. Fix (s,a) with $A^{\pi_\theta}(s,a) \ge 0$ (else trivial). From Lemma 10.2, the regularized gradient is

$$\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a) + \frac{\lambda}{|\mathcal{S}|}\Big(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\Big).$$

Step 1: probabilities stay bounded below. Since $A^{\pi_\theta}(s,a) \ge 0$, the first term is ≥ 0, so $\varepsilon_{\text{opt}} \ge \left|\partial L_\lambda/\partial\theta_{s,a}\right| \ge \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right)$, hence $\pi_\theta(a|s) \ge \frac{1}{|\mathcal{A}|} - \frac{\varepsilon_{\text{opt}}|\mathcal{S}|}{\lambda} \ge \frac{1}{2|\mathcal{A}|}$ (using $\varepsilon_{\text{opt}} \le \lambda/(2|\mathcal{S}||\mathcal{A}|)$).

Proof of Theorem 10.4

Goal. Show $\max_a A^{\pi_\theta}(s,a) \le \frac{2\lambda}{\mu(s)|\mathcal{S}|}$ for all s. Given this, the performance-difference lemma finishes it:

$$V^\star(\rho) - V^{\pi_\theta}(\rho) = \tfrac{1}{1-\gamma}\textstyle\sum_s d_\rho^{\pi^\star}(s)\sum_a \pi^\star(a|s)A^{\pi_\theta}(s,a) \;\le\; \tfrac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\max_a A^{\pi_\theta}(s,a)$$
$$\le \tfrac{2\lambda}{1-\gamma}\left\|\tfrac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty.$$

Now the claim. Fix (s,a) with $A^{\pi_\theta}(s,a) \ge 0$ (else trivial). From Lemma 10.2, the regularized gradient is

$$\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a) + \frac{\lambda}{|\mathcal{S}|}\Big(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\Big).$$

Step 1: probabilities stay bounded below. Since $A^{\pi_\theta}(s,a) \ge 0$, the first term is ≥ 0, so $\varepsilon_{\text{opt}} \ge \left|\partial L_\lambda/\partial\theta_{s,a}\right| \ge \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right)$, hence $\pi_\theta(a|s) \ge \frac{1}{|\mathcal{A}|} - \frac{\varepsilon_{\text{opt}}|\mathcal{S}|}{\lambda} \ge \frac{1}{2|\mathcal{A}|}$ (using $\varepsilon_{\text{opt}} \le \lambda/(2|\mathcal{S}||\mathcal{A}|)$).

Step 2: solve for the advantage. Rearranging the gradient identity and using ε_{opt}, $\pi_\theta(a|s) \ge \frac{1}{2|\mathcal{A}|}$, and $d_\mu^{\pi_\theta}(s) \ge (1-\gamma)\mu(s)$:

$$A^{\pi_\theta}(s,a) = \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\Big(\frac{1}{\pi_\theta(a|s)}\frac{\partial L_\lambda}{\partial\theta_{s,a}} + \frac{\lambda}{|\mathcal{S}|}\big(1 - \tfrac{1}{\pi_\theta(a|s)|\mathcal{A}|}\big)\Big)$$
$$\le \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\cdot\frac{2\lambda}{|\mathcal{S}|} \;\le\; \frac{2\lambda}{\mu(s)|\mathcal{S}|}. \qquad \square$$

Log-Barrier: Iteration Complexity

Combining Theorem 10.4 with Lemma 9.6 (GA to stationary points). We assume L_λ is β_λ-smooth with $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$ (a valid bound one can verify by computing $\nabla^2 L_\lambda$; we take it as given here).

Corollary 10.5

Under that smoothness, with $\lambda = \frac{\varepsilon(1-\gamma)}{2\|d_\rho^{\pi^\star}/\mu\|_\infty}$ and step size $\eta = 1/\beta_\lambda$,

$$\min_{t<T} \left\{ V^\star(\rho) - V^{(t)}(\rho) \right\} \leq \varepsilon \quad \text{whenever} \quad T \geq \frac{320\,|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\varepsilon^2} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty^2.$$

Log-Barrier: Iteration Complexity

Combining Theorem 10.4 with Lemma 9.6 (GA to stationary points). We assume L_λ is β_λ-smooth with $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$ (a valid bound one can verify by computing $\nabla^2 L_\lambda$; we take it as given here).

Corollary 10.5

Under that smoothness, with $\lambda = \frac{\varepsilon(1-\gamma)}{2\|d_\rho^{\pi^\star}/\mu\|_\infty}$ and step size $\eta = 1/\beta_\lambda$,

$$\min_{t<T} \left\{ V^\star(\rho) - V^{(t)}(\rho) \right\} \leq \varepsilon \quad \text{whenever} \quad T \geq \frac{320\,|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\varepsilon^2} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty^2.$$

The good and the bad.

- Good: polynomial in *everything*, including the mismatch coefficient. Plain GA had no such guarantee.
- Bad: explicit $|\mathcal{S}|^2|\mathcal{A}|^2$. Useless when the state space is large — i.e. exactly the function-approximation regime we care about.

Log-Barrier: Iteration Complexity

Combining Theorem 10.4 with Lemma 9.6 (GA to stationary points). We assume L_λ is β_λ-smooth with $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$ (a valid bound one can verify by computing $\nabla^2 L_\lambda$; we take it as given here).

Corollary 10.5

Under that smoothness, with $\lambda = \frac{\varepsilon(1-\gamma)}{2\|d_\rho^{\pi^\star}/\mu\|_\infty}$ and step size $\eta = 1/\beta_\lambda$,

$$\min_{t<T} \left\{ V^\star(\rho) - V^{(t)}(\rho) \right\} \leq \varepsilon \quad \text{whenever} \quad T \geq \frac{320\,|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\varepsilon^2} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty^2.$$

The good and the bad.

- Good: polynomial in *everything*, including the mismatch coefficient. Plain GA had no such guarantee.
- Bad: explicit $|\mathcal{S}|^2|\mathcal{A}|^2$. Useless when the state space is large — i.e. exactly the function-approximation regime we care about.

Why log-barrier and not entropy? Entropy is bounded ($\leq \log|\mathcal{A}|$); the relative entropy $\to \infty$ as probabilities $\to 0$. Only the aggressive barrier yields a polynomial rate.

The Natural Policy Gradient: Steepest Ascent Under KL

Idea. Measure a step's size not by Euclidean parameter distance, but by how much the policy *distribution* moves. The local distance between π_θ and $\pi_{\theta+\delta}$ is the KL, whose second-order Taylor expansion is

$$\mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathrm{KL}\big(\pi_\theta(\cdot|s) \,\|\, \pi_{\theta+\delta}(\cdot|s)\big) \approx \tfrac{1}{2}\,\delta^\top F_\rho(\theta)\,\delta, \quad F_\rho(\theta) := \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta}\left[\nabla \log \pi_\theta \, \nabla \log \pi_\theta^\top\right],$$

the (average) Fisher information matrix.

The Natural Policy Gradient: Steepest Ascent Under KL

Idea. Measure a step's size not by Euclidean parameter distance, but by how much the policy *distribution* moves. The local distance between π_θ and $\pi_{\theta+\delta}$ is the KL, whose second-order Taylor expansion is

$$\mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathrm{KL}\big(\pi_\theta(\cdot|s) \,\|\, \pi_{\theta+\delta}(\cdot|s)\big) \approx \tfrac{1}{2}\,\delta^\top F_\rho(\theta)\,\delta, \quad F_\rho(\theta) := \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta}\left[\nabla \log \pi_\theta \, \nabla \log \pi_\theta^\top\right],$$

the (average) Fisher information matrix.

Steepest-ascent step. Maximize the *linearized* objective subject to a small distribution change:

$$\max_{\delta} \ \langle \nabla_\theta V^{\pi_\theta}(\rho), \delta\rangle \quad \text{s.t.} \quad \tfrac{1}{2}\delta^\top F_\rho(\theta)\delta \le \epsilon.$$

The Lagrangian solution is $\delta \propto F_\rho(\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\rho)$ — the NPG update:

$$\boxed{\theta^{(t+1)} = \theta^{(t)} + \eta\, F_\rho(\theta^{(t)})^\dagger\, \nabla_\theta V^{(t)}(\rho)} \qquad (\dagger = \text{pseudoinverse}).$$

The Natural Policy Gradient: Steepest Ascent Under KL

Idea. Measure a step's size not by Euclidean parameter distance, but by how much the policy *distribution* moves. The local distance between π_θ and $\pi_{\theta+\delta}$ is the KL, whose second-order Taylor expansion is

$$\mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathrm{KL}\big(\pi_\theta(\cdot|s) \,\|\, \pi_{\theta+\delta}(\cdot|s)\big) \approx \tfrac{1}{2}\,\delta^\top F_\rho(\theta)\,\delta, \quad F_\rho(\theta) := \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta}\left[\nabla \log \pi_\theta \, \nabla \log \pi_\theta^\top\right],$$

the (average) Fisher information matrix.

Steepest-ascent step. Maximize the *linearized* objective subject to a small distribution change:

$$\max_\delta \ \langle \nabla_\theta V^{\pi_\theta}(\rho), \delta\rangle \quad \text{s.t.} \quad \tfrac{1}{2}\delta^\top F_\rho(\theta)\delta \le \epsilon.$$

The Lagrangian solution is $\delta \propto F_\rho(\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\rho)$ — the NPG update:

$$\boxed{\theta^{(t+1)} = \theta^{(t)} + \eta\, F_\rho(\theta^{(t)})^\dagger\, \nabla_\theta V^{(t)}(\rho)} \qquad (\dagger = \text{pseudoinverse}).$$

This is *not* a Newton step

Newton would precondition by $\nabla^2_\theta V$ — the Hessian of the *objective*. NPG preconditions by F = Fisher = Hessian of the *KL between policies*: a property of the parameterization's geometry, *independent of the rewards*. $F \succeq 0$ always (it is a covariance); $\nabla^2 V$ need not be, and V is not even concave. NPG is a metric / mirror-descent method, not a second-order method.

The Natural Policy Gradient: Steepest Ascent Under KL

Idea. Measure a step's size not by Euclidean parameter distance, but by how much the policy *distribution* moves. The local distance between π_θ and $\pi_{\theta+\delta}$ is the KL, whose second-order Taylor expansion is

$$\mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathrm{KL}\big(\pi_\theta(\cdot|s) \,\|\, \pi_{\theta+\delta}(\cdot|s)\big) \approx \tfrac{1}{2}\,\delta^\top F_\rho(\theta)\,\delta, \quad F_\rho(\theta) := \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta}\left[\nabla \log \pi_\theta \, \nabla \log \pi_\theta^\top\right],$$

the (average) Fisher information matrix.

Steepest-ascent step. Maximize the *linearized* objective subject to a small distribution change:

$$\max_\delta \ \langle \nabla_\theta V^{\pi_\theta}(\rho),\, \delta\rangle \quad \text{s.t.} \quad \tfrac{1}{2}\delta^\top F_\rho(\theta)\delta \le \epsilon.$$

The Lagrangian solution is $\delta \propto F_\rho(\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\rho)$ — the NPG update:

$$\boxed{\theta^{(t+1)} = \theta^{(t)} + \eta\, F_\rho(\theta^{(t)})^\dagger\, \nabla_\theta V^{(t)}(\rho)} \qquad (\dagger = \text{pseudoinverse}).$$

This is *not* a Newton step

Newton would precondition by $\nabla_\theta^2 V$ — the Hessian of the *objective*. NPG preconditions by F = Fisher = Hessian of the *KL between policies*: a property of the parameterization's geometry, *independent of the rewards*. $F \succeq 0$ always (it is a covariance); $\nabla^2 V$ need not be, and V is not even concave. NPG is a metric / mirror-descent method, not a second-order method.

Payoff. Reparameterization-invariant; the $F^\dagger$ cancels the troublesome $d_\rho^{\pi_\theta}(s)$, $\pi_\theta(a|s)$ factors of Lemma 10.2; and (next slide) the softmax update ends up with *no ρ-dependence at all*.

Softmax NPG = Soft Policy Iteration

Lemma 10.6 (Closed form of the softmax NPG step)

With $A^{(t)} := A^{\pi^{(t)}}$, the NPG update on logits / on the policy is

$$\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}(s,a) \; (+ c_s), \qquad \pi^{(t+1)}(a|s) = \pi^{(t)}(a|s) \, \frac{\exp(\frac{\eta}{1-\gamma} A^{(t)}(s,a))}{Z_t(s)},$$

$Z_t(s) = \sum_a \pi^{(t)}(a|s) \exp(\frac{\eta}{1-\gamma} A^{(t)}(s,a))$, and c_s an arbitrary per-state constant.

Softmax NPG = Soft Policy Iteration

Lemma 10.6 (Closed form of the softmax NPG step)

With $A^{(t)} := A^{\pi^{(t)}}$, the NPG update on logits / on the policy is

$$\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}(s,a) \; (+ c_s), \qquad \pi^{(t+1)}(a|s) = \pi^{(t)}(a|s) \, \frac{\exp(\frac{\eta}{1-\gamma} A^{(t)}(s,a))}{Z_t(s)},$$

$Z_t(s) = \sum_a \pi^{(t)}(a|s) \exp(\frac{\eta}{1-\gamma} A^{(t)}(s,a))$, and c_s an arbitrary per-state constant.

Step 1: characterize the direction. By definition of the pseudoinverse, $w_\star := F_\rho^\dagger \nabla_\theta V^{\pi_\theta}(\rho)$ is the minimum-norm solution of the normal equations $F_\rho(\theta)\, w = \nabla_\theta V^{\pi_\theta}(\rho)$. We solve for w.

Softmax NPG = Soft Policy Iteration

Lemma 10.6 (Closed form of the softmax NPG step)

With $A^{(t)} := A^{\pi^{(t)}}$, the NPG update on logits / on the policy is

$$\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \tfrac{\eta}{1-\gamma} A^{(t)}(s,a)\ (+ c_s), \qquad \pi^{(t+1)}(a|s) = \pi^{(t)}(a|s)\ \frac{\exp(\frac{\eta}{1-\gamma} A^{(t)}(s,a))}{Z_t(s)},$$

$Z_t(s) = \sum_a \pi^{(t)}(a|s) \exp(\frac{\eta}{1-\gamma} A^{(t)}(s,a))$, and c_s an arbitrary per-state constant.

Step 1: characterize the direction. By definition of the pseudoinverse, $w_\star := F_\rho^\dagger \nabla_\theta V^{\pi_\theta}(\rho)$ is the minimum-norm solution of the normal equations $F_\rho(\theta)\, w = \nabla_\theta V^{\pi_\theta}(\rho)$. We solve for w.

Step 2: compute $F_\rho w$. The softmax score is $\nabla_{\theta_{s',a'}} \log \pi_\theta(a|s) = \mathbf{1}[s{=}s'](\mathbf{1}[a{=}a'] - \pi_\theta(a'|s))$, so $w \cdot \nabla \log \pi_\theta(a|s) = w_{s,a} - \bar{w}_s$ with $\bar{w}_s := \sum_{a'} \pi_\theta(a'|s) w_{s,a'}$. Since $\mathbb{E}_{a\sim\pi_\theta}[\nabla \log \pi_\theta] = 0$, this gives $[F_\rho w]_{s,a} = d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)(w_{s,a} - \bar{w}_s)$.

Softmax NPG = Soft Policy Iteration

Lemma 10.6 (Closed form of the softmax NPG step)

With $A^{(t)} := A^{\pi^{(t)}}$, the NPG update on logits / on the policy is

$$\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \tfrac{\eta}{1-\gamma} A^{(t)}(s,a)\ (+ c_s), \qquad \pi^{(t+1)}(a|s) = \pi^{(t)}(a|s)\ \frac{\exp(\frac{\eta}{1-\gamma} A^{(t)}(s,a))}{Z_t(s)},$$

$Z_t(s) = \sum_a \pi^{(t)}(a|s) \exp(\frac{\eta}{1-\gamma} A^{(t)}(s,a))$, and c_s an arbitrary per-state constant.

Step 1: characterize the direction. By definition of the pseudoinverse, $w_\star := F_\rho^\dagger \nabla_\theta V^{\pi_\theta}(\rho)$ is the minimum-norm solution of the normal equations $F_\rho(\theta)\, w = \nabla_\theta V^{\pi_\theta}(\rho)$. We solve for w.

Step 2: compute $F_\rho w$. The softmax score is $\nabla_{\theta_{s',a'}} \log \pi_\theta(a|s) = \mathbf{1}[s{=}s'](\mathbf{1}[a{=}a'] - \pi_\theta(a'|s))$, so $w \cdot \nabla \log \pi_\theta(a|s) = w_{s,a} - \bar{w}_s$ with $\bar{w}_s := \sum_{a'} \pi_\theta(a'|s) w_{s,a'}$. Since $\mathbb{E}_{a\sim\pi_\theta}[\nabla \log \pi_\theta] = 0$, this gives $[F_\rho w]_{s,a} = d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)(w_{s,a} - \bar{w}_s)$.

Step 3: match to ∇V. Lemma 10.2 says $[\nabla_\theta V^{\pi_\theta}(\rho)]_{s,a} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a)$. So $F_\rho w = \nabla V$ iff $w_{s,a} - \bar{w}_s = \frac{1}{1-\gamma} A^{\pi_\theta}(s,a)$. Thus $w = \frac{1}{1-\gamma} A^{\pi_\theta}$ works; adding a state-only offset v ($v_{s,a} = c_s$) leaves $w_{s,a} - \bar{w}_s$ unchanged, so all solutions are $w = \frac{1}{1-\gamma} A^{\pi_\theta} + v$.

Softmax NPG = Soft Policy Iteration

Lemma 10.6 (Closed form of the softmax NPG step)

With $A^{(t)} := A^{\pi^{(t)}}$, the NPG update on logits / on the policy is

$$\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}(s,a)\ (+\,c_s), \qquad \pi^{(t+1)}(a|s) = \pi^{(t)}(a|s)\,\frac{\exp(\frac{\eta}{1-\gamma}A^{(t)}(s,a))}{Z_t(s)},$$

$Z_t(s) = \sum_a \pi^{(t)}(a|s)\exp(\frac{\eta}{1-\gamma}A^{(t)}(s,a))$, and c_s an arbitrary per-state constant.

Step 1: characterize the direction. By definition of the pseudoinverse, $w_\star := F_\rho^\dagger \nabla_\theta V^{\pi_\theta}(\rho)$ is the minimum-norm solution of the normal equations $F_\rho(\theta)\, w = \nabla_\theta V^{\pi_\theta}(\rho)$. We solve for w.

Step 2: compute $F_\rho w$. The softmax score is $\nabla_{\theta_{s',a'}} \log \pi_\theta(a|s) = \mathbf{1}[s{=}s'](\mathbf{1}[a{=}a'] - \pi_\theta(a'|s))$, so $w \cdot \nabla \log \pi_\theta(a|s) = w_{s,a} - \bar{w}_s$ with $\bar{w}_s := \sum_{a'} \pi_\theta(a'|s) w_{s,a'}$. Since $\mathbb{E}_{a\sim\pi_\theta}[\nabla \log \pi_\theta] = 0$, this gives $[F_\rho w]_{s,a} = d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)(w_{s,a} - \bar{w}_s)$.

Step 3: match to ∇V. Lemma 10.2 says $[\nabla_\theta V^{\pi_\theta}(\rho)]_{s,a} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a)$. So $F_\rho w = \nabla V$ iff $w_{s,a} - \bar{w}_s = \frac{1}{1-\gamma} A^{\pi_\theta}(s,a)$. Thus $w = \frac{1}{1-\gamma} A^{\pi_\theta}$ works; adding a state-only offset v ($v_{s,a} = c_s$) leaves $w_{s,a} - \bar{w}_s$ unchanged, so all solutions are $w = \frac{1}{1-\gamma} A^{\pi_\theta} + v$.

Step 4: exponentiate. Adding $\eta w = \frac{\eta}{1-\gamma} A^{(t)} (+\, c_s)$ to the logits and renormalizing gives the multiplicative update (c_s cancels in Z_t). □ Multiplicative weights on actions — no $d_\rho^{\pi_\theta}$ factor, no ρ dependence.

NPG: Dimension-Free Global Convergence

Theorem 10.7 (Global convergence for NPG)

Run softmax NPG from $\theta^{(0)} = 0$ with any fixed $\eta > 0$. For all T,

$$V^{(T)}(\rho) \geq V^{\star}(\rho) - \frac{\log|\mathcal{A}|}{\eta T} - \frac{1}{(1-\gamma)^2 T}.$$

NPG: Dimension-Free Global Convergence

Theorem 10.7 (Global convergence for NPG)

Run softmax NPG from $\theta^{(0)} = 0$ with any fixed $\eta > 0$. For all T,

$$V^{(T)}(\rho) \;\geq\; V^{\star}(\rho) - \frac{\log|\mathcal{A}|}{\eta T} - \frac{1}{(1-\gamma)^2 T}.$$

Setting $\eta \geq (1-\gamma)^2 \log|\mathcal{A}|$ gives an ε-optimal policy in

$$T \leq \frac{2}{(1-\gamma)^2 \varepsilon}$$

iterations — no dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$, despite non-concavity, and no mismatch coefficient.

NPG: Dimension-Free Global Convergence

Theorem 10.7 (Global convergence for NPG)

Run softmax NPG from $\theta^{(0)} = 0$ with any fixed $\eta > 0$. For all T,

$$V^{(T)}(\rho) \;\geq\; V^{\star}(\rho) - \frac{\log|\mathcal{A}|}{\eta T} - \frac{1}{(1-\gamma)^2 T}.$$

Setting $\eta \geq (1-\gamma)^2 \log|\mathcal{A}|$ gives an ε-optimal policy in

$$T \leq \frac{2}{(1-\gamma)^2 \varepsilon}$$

iterations — no dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$, despite non-concavity, and no mismatch coefficient.

Contrast. Log-barrier GA: $\widetilde{O}\big(|\mathcal{S}|^2|\mathcal{A}|^2/\varepsilon^2\big)$ with mismatch. NPG: $O(1/\varepsilon)$, dimension-free. The geometry did all the work.

Proof: a multiplicative-weights–style telescoping, via the improvement lemma next.

The NPG Improvement Lemma

Lemma 10.8 (Per-step improvement)

For the NPG iterates and *any* distribution μ, $V^{(t+1)}(\mu) - V^{(t)}(\mu) \geq \frac{1-\gamma}{\eta} \mathbb{E}_{s\sim\mu} \log Z_t(s) \geq 0.$

The NPG Improvement Lemma

Lemma 10.8 (Per-step improvement)

For the NPG iterates and *any* distribution μ, $V^{(t+1)}(\mu) - V^{(t)}(\mu) \geq \frac{1-\gamma}{\eta} \mathbb{E}_{s\sim\mu} \log Z_t(s) \geq 0$.

Step 1: $\log Z_t(s) \geq 0$. By Jensen on the concave log and $\sum_a \pi^{(t)}(a|s)A^{(t)}(s,a) = 0$,

$$\log Z_t(s) = \log \sum_a \pi^{(t)}(a|s) e^{\eta A^{(t)}(s,a)/(1-\gamma)} \geq \frac{\eta}{1-\gamma} \sum_a \pi^{(t)}(a|s) A^{(t)}(s,a) = 0.$$

The NPG Improvement Lemma

Lemma 10.8 (Per-step improvement)

For the NPG iterates and *any* distribution μ, $V^{(t+1)}(\mu) - V^{(t)}(\mu) \geq \frac{1-\gamma}{\eta} \mathbb{E}_{s\sim\mu} \log Z_t(s) \geq 0$.

Step 1: $\log Z_t(s) \geq 0$. By Jensen on the concave log and $\sum_a \pi^{(t)}(a|s)A^{(t)}(s,a) = 0$,

$$\log Z_t(s) = \log \sum_a \pi^{(t)}(a|s) e^{\eta A^{(t)}(s,a)/(1-\gamma)} \geq \frac{\eta}{1-\gamma} \sum_a \pi^{(t)}(a|s) A^{(t)}(s,a) = 0.$$

Step 2: PDL, written out. The performance-difference lemma says, for the move $\pi^{(t)} \to \pi^{(t+1)}$,

$$V^{(t+1)}(\mu) - V^{(t)}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{(t+1)}} \sum_a \pi^{(t+1)}(a|s)\, A^{(t)}(s,a).$$

The NPG Improvement Lemma

Lemma 10.8 (Per-step improvement)

For the NPG iterates and *any* distribution μ, $V^{(t+1)}(\mu) - V^{(t)}(\mu) \geq \frac{1-\gamma}{\eta} \mathbb{E}_{s\sim\mu} \log Z_t(s) \geq 0$.

Step 1: $\log Z_t(s) \geq 0$. By Jensen on the concave log and $\sum_a \pi^{(t)}(a|s)A^{(t)}(s,a) = 0$,

$$\log Z_t(s) = \log \sum_a \pi^{(t)}(a|s) e^{\eta A^{(t)}(s,a)/(1-\gamma)} \geq \frac{\eta}{1-\gamma} \sum_a \pi^{(t)}(a|s) A^{(t)}(s,a) = 0.$$

Step 2: PDL, written out. The performance-difference lemma says, for the move $\pi^{(t)} \to \pi^{(t+1)}$,

$$V^{(t+1)}(\mu) - V^{(t)}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{(t+1)}} \sum_a \pi^{(t+1)}(a|s)\, A^{(t)}(s,a).$$

Step 3: substitute the closed form. From Lemma 10.6, $\frac{\eta}{1-\gamma} A^{(t)}(s,a) = \log \frac{\pi^{(t+1)}(a|s) Z_t(s)}{\pi^{(t)}(a|s)}$, so $A^{(t)}(s,a) = \frac{1-\gamma}{\eta} \log \frac{\pi^{(t+1)}(a|s) Z_t(s)}{\pi^{(t)}(a|s)}$. Hence for each s,

$$\sum_a \pi^{(t+1)}(a|s) A^{(t)}(s,a) = \frac{1-\gamma}{\eta} \Big[\underbrace{\textstyle\sum_a \pi^{(t+1)} \log \frac{\pi^{(t+1)}}{\pi^{(t)}}}_{= \mathrm{KL}(\pi_s^{(t+1)} \| \pi_s^{(t)})} + \underbrace{\textstyle\sum_a \pi^{(t+1)} \log Z_t(s)}_{= \log Z_t(s)} \Big].$$

The NPG Improvement Lemma

Lemma 10.8 (Per-step improvement)

For the NPG iterates and *any* distribution μ, $V^{(t+1)}(\mu) - V^{(t)}(\mu) \geq \frac{1-\gamma}{\eta}\,\mathbb{E}_{s\sim\mu} \log Z_t(s) \geq 0$.

Step 1: $\log Z_t(s) \geq 0$. By Jensen on the concave log and $\sum_a \pi^{(t)}(a|s)A^{(t)}(s,a) = 0$,

$$\log Z_t(s) = \log \sum_a \pi^{(t)}(a|s) e^{\eta A^{(t)}(s,a)/(1-\gamma)} \geq \frac{\eta}{1-\gamma} \sum_a \pi^{(t)}(a|s) A^{(t)}(s,a) = 0.$$

Step 2: PDL, written out. The performance-difference lemma says, for the move $\pi^{(t)} \to \pi^{(t+1)}$,

$$V^{(t+1)}(\mu) - V^{(t)}(\mu) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\mu^{(t+1)}} \sum_a \pi^{(t+1)}(a|s)\, A^{(t)}(s,a).$$

Step 3: substitute the closed form. From Lemma 10.6, $\frac{\eta}{1-\gamma} A^{(t)}(s,a) = \log \frac{\pi^{(t+1)}(a|s) Z_t(s)}{\pi^{(t)}(a|s)}$, so $A^{(t)}(s,a) = \frac{1-\gamma}{\eta} \log \frac{\pi^{(t+1)}(a|s) Z_t(s)}{\pi^{(t)}(a|s)}$. Hence for each s,

$$\sum_a \pi^{(t+1)}(a|s) A^{(t)}(s,a) = \frac{1-\gamma}{\eta} \Big[\underbrace{\textstyle\sum_a \pi^{(t+1)} \log \frac{\pi^{(t+1)}}{\pi^{(t)}}}_{= \mathrm{KL}(\pi_s^{(t+1)} \| \pi_s^{(t)})} + \underbrace{\textstyle\sum_a \pi^{(t+1)} \log Z_t(s)}_{= \log Z_t(s)} \Big].$$

Step 4: assemble. Plug back: $V^{(t+1)}(\mu) - V^{(t)}(\mu) = \frac{1}{\eta} \mathbb{E}_{s\sim d_\mu^{(t+1)}} [\mathrm{KL}(\pi_s^{(t+1)} \| \pi_s^{(t)}) + \log Z_t(s)]$.

Drop the nonneg. KL, then use $d_\mu^{(t+1)} \geq (1-\gamma)\mu$ and $\log Z_t \geq 0$. □

Proof of Theorem 10.7

Write $d^\star = d_\rho^{\pi^\star}$. By the performance-difference lemma and the closed-form update (Lemma 10.6), *telescoping the KL*:

$$V^{\pi^\star}(\rho) - V^{(t)}(\rho) = \frac{1}{\eta}\,\mathbb{E}_{s\sim d^\star}\left[\mathrm{KL}(\pi_s^\star \| \pi_s^{(t)}) - \mathrm{KL}(\pi_s^\star \| \pi_s^{(t+1)}) + \log Z_t(s)\right].$$

Proof of Theorem 10.7

Write $d^\star = d_\rho^{\pi^\star}$. By the performance-difference lemma and the closed-form update (Lemma 10.6), *telescoping the KL*:

$$V^{\pi^\star}(\rho) - V^{(t)}(\rho) = \frac{1}{\eta}\,\mathbb{E}_{s\sim d^\star}\left[\mathrm{KL}(\pi_s^\star \| \pi_s^{(t)}) - \mathrm{KL}(\pi_s^\star \| \pi_s^{(t+1)}) + \log Z_t(s)\right].$$

Bound the $\log Z_t$ term. Apply Lemma 10.8 with starting distribution $d^\star$:

$$\mathbb{E}_{s\sim d^\star} \log Z_t(s) \le \frac{\eta}{1-\gamma}\left(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\right).$$

Proof of Theorem 10.7

Write $d^\star = d_\rho^{\pi^\star}$. By the performance-difference lemma and the closed-form update (Lemma 10.6), *telescoping the KL*:

$$V^{\pi^\star}(\rho) - V^{(t)}(\rho) = \frac{1}{\eta}\,\mathbb{E}_{s\sim d^\star}\left[\mathrm{KL}(\pi_s^\star \| \pi_s^{(t)}) - \mathrm{KL}(\pi_s^\star \| \pi_s^{(t+1)}) + \log Z_t(s)\right].$$

Bound the $\log Z_t$ term. Apply Lemma 10.8 with starting distribution $d^\star$:

$$\mathbb{E}_{s\sim d^\star} \log Z_t(s) \le \frac{\eta}{1-\gamma}\left(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\right).$$

Average over $t = 0, \ldots, T-1$. The KL terms telescope and $\log Z_t$ terms telescope through the values:

$$V^{\pi^\star}(\rho) - V^{(T-1)}(\rho) \le \frac{1}{T}\sum_{t=0}^{T-1}\left(V^{\pi^\star}(\rho) - V^{(t)}(\rho)\right) \le \frac{\mathbb{E}_{s\sim d^\star}\mathrm{KL}(\pi_s^\star \| \pi^{(0)})}{\eta T} + \frac{V^{(T)}(d^\star) - V^{(0)}(d^\star)}{(1-\gamma)T}.$$

Proof of Theorem 10.7

Write $d^\star = d_\rho^{\pi^\star}$. By the performance-difference lemma and the closed-form update (Lemma 10.6), *telescoping the KL*:

$$V^{\pi^\star}(\rho) - V^{(t)}(\rho) = \frac{1}{\eta}\,\mathbb{E}_{s\sim d^\star}\left[\mathrm{KL}(\pi_s^\star \| \pi_s^{(t)}) - \mathrm{KL}(\pi_s^\star \| \pi_s^{(t+1)}) + \log Z_t(s)\right].$$

Bound the $\log Z_t$ term. Apply Lemma 10.8 with starting distribution $d^\star$:

$$\mathbb{E}_{s\sim d^\star} \log Z_t(s) \le \frac{\eta}{1-\gamma}\left(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\right).$$

Average over $t = 0, \ldots, T-1$. The KL terms telescope and $\log Z_t$ terms telescope through the values:

$$V^{\pi^\star}(\rho) - V^{(T-1)}(\rho) \le \frac{1}{T}\sum_{t=0}^{T-1}\left(V^{\pi^\star}(\rho) - V^{(t)}(\rho)\right) \le \frac{\mathbb{E}_{s\sim d^\star}\mathrm{KL}(\pi_s^\star \| \pi^{(0)})}{\eta T} + \frac{V^{(T)}(d^\star) - V^{(0)}(d^\star)}{(1-\gamma)T}.$$

With $\pi^{(0)}$ uniform, $\mathrm{KL}(\pi_s^\star \| \pi^{(0)}) \le \log|\mathcal{A}|$; and $V^{(T)}(d^\star) - V^{(0)}(d^\star) \le \frac{1}{1-\gamma}$. Hence

$$V^{\pi^\star}(\rho) - V^{(T-1)}(\rho) \le \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T},$$

and $V^{(T)}(\rho) \ge V^{(T-1)}(\rho)$ by monotonicity. □

Where dimension-freeness comes from: the only "size" that survives is $\mathrm{KL}(\pi^\star \| \pi^{(0)}) \le \log|\mathcal{A}|$ — the mismatch coefficient cancels because NPG steers in distribution space.

Today's Plan

1. Policy Gradients (Ch. 9)
2. Global Optimality & the NPG (Ch. 10)
3. Function Approximation & the NPG (Ch. 11)
4. Trust-Region Methods: TRPO & PPO (Ch. 12)
5. Summary

Now the Policy Class is Not Expressive

We move to parametric classes $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^d\}$ with $d \ll |\mathcal{S}||\mathcal{A}|$ (indeed $|\mathcal{S}|, |\mathcal{A}|$ may be infinite). Π may *not* contain an optimal policy.

Agnostic goal. Compete with the best (or any chosen comparator) policy $\pi^\star$, possibly outside reach of exact optimality.

Now the Policy Class is Not Expressive

We move to parametric classes $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^d\}$ with $d \ll |\mathcal{S}||\mathcal{A}|$ (indeed $|\mathcal{S}|, |\mathcal{A}|$ may be infinite). Π may *not* contain an optimal policy.

Agnostic goal. Compete with the best (or any chosen comparator) policy $\pi^\star$, possibly outside reach of exact optimality.

The question this chapter answers. What is the right notion of "approximation error" for the NPG when Π cannot represent every policy?

Now the Policy Class is Not Expressive

We move to parametric classes $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^d\}$ with $d \ll |\mathcal{S}||\mathcal{A}|$ (indeed $|\mathcal{S}|, |\mathcal{A}|$ may be infinite). Π may *not* contain an optimal policy.

Agnostic goal. Compete with the best (or any chosen comparator) policy $\pi^\star$, possibly outside reach of exact optimality.

The question this chapter answers. What is the right notion of "approximation error" for the NPG when Π cannot represent every policy?

The answer: compatible function approximation

The relevant error is how well the score features $\nabla_\theta \log \pi_\theta$ can linearly fit the advantage A^{π_θ}. This single quantity governs the NPG's expressivity — and it is exactly the object the NPG update implicitly regresses. We develop just this idea; the full Q-NPG sample-complexity analysis (log-linear policies, relative condition number) is in AJKS §11.4.

Compatible Function Approximation

Lemma 11.1 (Gradient = best linear fit of the advantage)

Let $w^\star$ minimize the compatible function approximation error

$$w^\star \in \arg\min_w \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left(A^{\pi_\theta}(s,a) - w \cdot \nabla_\theta \log \pi_\theta(a|s)\right)^2.$$

With $\widehat{A}^{\pi_\theta} := w^\star \cdot \nabla_\theta \log \pi_\theta$, we then have $\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\mathbb{E}[\nabla \log \pi_\theta \widehat{A}^{\pi_\theta}]$.

Compatible Function Approximation

Lemma 11.1 (Gradient = best linear fit of the advantage)

Let $w^\star$ minimize the compatible function approximation error

$$w^\star \in \arg\min_w \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left(A^{\pi_\theta}(s,a) - w \cdot \nabla_\theta \log \pi_\theta(a|s)\right)^2.$$

With $\widehat{A}^{\pi_\theta} := w^\star \cdot \nabla_\theta \log \pi_\theta$, we then have $\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\mathbb{E}[\nabla \log \pi_\theta \widehat{A}^{\pi_\theta}]$.

Proof. First-order optimality for $w^\star$ gives $\mathbb{E}\left[(A^{\pi_\theta} - w^\star \cdot \nabla \log \pi_\theta)\nabla \log \pi_\theta\right] = 0$; substitute into the advantage form of the gradient. □

Compatible Function Approximation

Lemma 11.1 (Gradient = best linear fit of the advantage)

Let $w^\star$ minimize the compatible function approximation error

$$w^\star \in \arg\min_w \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left(A^{\pi_\theta}(s,a) - w \cdot \nabla_\theta \log \pi_\theta(a|s)\right)^2.$$

With $\widehat{A}^{\pi_\theta} := w^\star \cdot \nabla_\theta \log \pi_\theta$, we then have $\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\mathbb{E}[\nabla \log \pi_\theta \widehat{A}^{\pi_\theta}]$.

Proof. First-order optimality for $w^\star$ gives $\mathbb{E}\left[(A^{\pi_\theta} - w^\star \cdot \nabla \log \pi_\theta)\nabla \log \pi_\theta\right] = 0$; substitute into the advantage form of the gradient. □

Lemma 11.2 (The NPG direction *is* $w^\star$)

$$F_\rho(\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} w^\star, \quad w^\star \in \arg\min_w \mathbb{E}_{s\sim d_\rho^{\pi_\theta}, a\sim\pi_\theta} \left(w^\top \nabla \log \pi_\theta - A^{\pi_\theta}\right)^2.$$

Compatible Function Approximation

Lemma 11.1 (Gradient = best linear fit of the advantage)

Let $w^\star$ minimize the compatible function approximation error

$$w^\star \in \arg\min_w \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left(A^{\pi_\theta}(s,a) - w \cdot \nabla_\theta \log \pi_\theta(a|s)\right)^2.$$

With $\widehat{A}^{\pi_\theta} := w^\star \cdot \nabla_\theta \log \pi_\theta$, we then have $\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\mathbb{E}[\nabla \log \pi_\theta \widehat{A}^{\pi_\theta}]$.

Proof. First-order optimality for $w^\star$ gives $\mathbb{E}\left[(A^{\pi_\theta} - w^\star \cdot \nabla \log \pi_\theta)\nabla \log \pi_\theta\right] = 0$; substitute into the advantage form of the gradient. □

Lemma 11.2 (The NPG direction *is* $w^\star$)

$$F_\rho(\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} w^\star, \quad w^\star \in \arg\min_w \mathbb{E}_{s\sim d_\rho^{\pi_\theta}, a\sim\pi_\theta} \left(w^\top \nabla \log \pi_\theta - A^{\pi_\theta}\right)^2.$$

Proof. The first-order condition of Lemma 11.1 reads $\frac{1}{1-\gamma}\mathbb{E}[\nabla \log \pi_\theta A^{\pi_\theta}] = \frac{1}{1-\gamma} F_\rho(\theta) w^\star$. The left side is exactly $\nabla_\theta V^{\pi_\theta}(\rho)$ (advantage form); apply $F_\rho^\dagger$. □

Yes — this is exactly the general-parameterization version of the softmax computation in Lemma 10.6, the same first-order condition with arbitrary scores instead of centered logits. So NPG = "fit the advantage with the scores, then step"; that fitting error is the expressivity that matters.

Today's Plan

1. Policy Gradients (Ch. 9)
2. Global Optimality & the NPG (Ch. 10)
3. Function Approximation & the NPG (Ch. 11)
4. Trust-Region Methods: TRPO & PPO (Ch. 12)
5. Summary

A Common Theme: Incremental Updates

NPG, TRPO, and PPO are all variations on one idea:

improve the policy, but don't let its state-action distribution move too far in one step.

A Common Theme: Incremental Updates

NPG, TRPO, and PPO are all variations on one idea:

improve the policy, but don't let its state-action distribution move too far in one step.

Why? The performance-difference lemma evaluates the advantage under the *new* policy's visitation $d^{\pi_{t+1}}$, which we cannot see before committing to π_{t+1}. If $d^{\pi_{t+1}} \approx d^{\pi_t}$, we may safely use the *measurable* d^{π_t} as a proxy — the surrogate objective

$$\mathbb{E}_{s \sim d_\mu^{\pi_t}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} A^{\pi_t}(s, a)$$

then genuinely predicts improvement.

A Common Theme: Incremental Updates

NPG, TRPO, and PPO are all variations on one idea:

improve the policy, but don't let its state-action distribution move too far in one step.

Why? The performance-difference lemma evaluates the advantage under the *new* policy's visitation $d^{\pi_{t+1}}$, which we cannot see before committing to π_{t+1}. If $d^{\pi_{t+1}} \approx d^{\pi_t}$, we may safely use the *measurable* d^{π_t} as a proxy — the surrogate objective

$$\mathbb{E}_{s \sim d_\mu^{\pi_t}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} A^{\pi_t}(s, a)$$

then genuinely predicts improvement.

How each enforces "closeness".

- **NPG / TRPO:** an explicit *KL trust region* on the policy / trajectory distribution.
- **PPO:** a *clipped* surrogate that removes any incentive to move the density ratio far.

A Common Theme: Incremental Updates

NPG, TRPO, and PPO are all variations on one idea:

improve the policy, but don't let its state-action distribution move too far in one step.

Why? The performance-difference lemma evaluates the advantage under the *new* policy's visitation $d^{\pi_{t+1}}$, which we cannot see before committing to π_{t+1}. If $d^{\pi_{t+1}} \approx d^{\pi_t}$, we may safely use the *measurable* d^{π_t} as a proxy — the surrogate objective

$$\mathbb{E}_{s\sim d_\mu^{\pi_t}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_t}(s,a)$$

then genuinely predicts improvement.

How each enforces "closeness".

- **NPG / TRPO:** an explicit *KL trust region* on the policy / trajectory distribution.
- **PPO:** a *clipped* surrogate that removes any incentive to move the density ratio far.

Contrast with full greedy policy iteration: a full argmax step can abruptly change d^π, breaking the proxy, and provably can *fail* to improve without strong (all-policy) concentrability. "Move a little" is what makes the proxy trustworthy.

Trust Region Policy Optimization (TRPO)

At the current π_{θ_t}, TRPO maximizes the surrogate advantage objective under an explicit KL trust region on the *trajectory* distributions:

$$\max_{\theta} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta_t}}} \mathbb{E}_{a \sim \pi_{\theta}(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \quad \text{s.t.} \quad \mathrm{KL}\left(\mathbb{P}_{\mu}^{\pi_{\theta_t}} \,\|\, \mathbb{P}_{\mu}^{\pi_{\theta}}\right) \leq \delta.$$

Trust Region Policy Optimization (TRPO)

At the current π_{θ_t}, TRPO maximizes the surrogate advantage objective under an explicit KL trust region on the *trajectory* distributions:

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \quad \text{s.t.} \quad \mathrm{KL}\big(\mathbb{P}_\mu^{\pi_{\theta_t}} \,\|\, \mathbb{P}_\mu^{\pi_\theta}\big) \le \delta.$$

Make it a local quadratic program. Since we only want a small step, do sequential quadratic programming: *linearize* the objective and *quadratize* the KL constraint about θ_t (with $\Delta := \theta - \theta_t$):

$$\max_{\Delta} \left\langle \nabla_\theta V^{\pi_{\theta_t}}, \Delta \right\rangle \text{ s.t. } \left\langle \nabla_\theta \mathrm{KL}|_{\theta_t}, \Delta \right\rangle + \tfrac{1}{2}\Delta^\top (\nabla_\theta^2 \mathrm{KL}|_{\theta_t})\Delta \le \delta.$$

(The linearized objective is the policy gradient, since $\mathbb{E}_{a\sim\pi_{\theta_t}} A^{\pi_{\theta_t}} = 0$ at θ_t.)

Trust Region Policy Optimization (TRPO)

At the current π_{θ_t}, TRPO maximizes the surrogate advantage objective under an explicit KL trust region on the *trajectory* distributions:

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \quad \text{s.t.} \quad \mathrm{KL}\big(\mathbb{P}_\mu^{\pi_{\theta_t}} \,\|\, \mathbb{P}_\mu^{\pi_\theta}\big) \le \delta.$$

Make it a local quadratic program. Since we only want a small step, do sequential quadratic programming: *linearize* the objective and *quadratize* the KL constraint about θ_t (with $\Delta := \theta - \theta_t$):

$$\max_{\Delta} \left\langle \nabla_\theta V^{\pi_{\theta_t}}, \Delta \right\rangle \text{ s.t. } \left\langle \nabla_\theta \mathrm{KL}|_{\theta_t}, \Delta \right\rangle + \tfrac{1}{2}\Delta^\top (\nabla_\theta^2 \mathrm{KL}|_{\theta_t})\Delta \le \delta.$$

(The linearized objective is the policy gradient, since $\mathbb{E}_{a\sim\pi_{\theta_t}} A^{\pi_{\theta_t}} = 0$ at θ_t.)

So we need the gradient and Hessian of the trajectory KL at θ_t. Both have clean closed forms — and the Hessian turns out to be exactly the Fisher information. That is Claim 12.5 (next slide), and it is what makes this QP a *Fisher* trust region.

TRPO's Local Constraint is a Fisher Trust Region

Claim 12.5 (KL gradient and Hessian at θ_t)

For a horizon-H MDP and fixed θ_t,

$$\nabla_\theta \mathrm{KL}(\mathbb{P}_\mu^{\pi_{\theta_t}} \| \mathbb{P}_\mu^{\pi_\theta})\big|_{\theta_t} = 0, \quad \nabla_\theta^2 \mathrm{KL}\big|_{\theta_t} = H\,\mathbb{E}_{s,a\sim d^{\pi_{\theta_t}}}\left[\nabla \log \pi_{\theta_t}(a|s) \nabla \log \pi_{\theta_t}(a|s)^\top\right] =: F_{\theta_t}.$$

TRPO's Local Constraint is a Fisher Trust Region

Claim 12.5 (KL gradient and Hessian at θ_t)

For a horizon-H MDP and fixed θ_t,

$$\nabla_\theta \mathrm{KL}(\mathbb{P}_\mu^{\pi_{\theta_t}} \| \mathbb{P}_\mu^{\pi_\theta})\big|_{\theta_t} = 0, \quad \nabla_\theta^2 \mathrm{KL}\big|_{\theta_t} = H\,\mathbb{E}_{s,a\sim d^{\pi_{\theta_t}}}\left[\nabla \log \pi_{\theta_t}(a|s) \nabla \log \pi_{\theta_t}(a|s)^\top\right] =: F_{\theta_t}.$$

Proof. The trajectory KL telescopes over steps: $\mathrm{KL}(\mathbb{P}_\mu^{\pi_{\theta_t}} \| \mathbb{P}_\mu^{\pi_\theta}) = \sum_h \mathbb{E}_{s_h,a_h\sim\pi_{\theta_t}} \log \frac{\pi_{\theta_t}(a_h|s_h)}{\pi_\theta(a_h|s_h)}$ (μ, P cancel). Differentiating at $\theta = \theta_t$: $-\sum_h \mathbb{E}_{s_h}\mathbb{E}_{a_h\sim\pi_{\theta_t}} \nabla \log \pi_{\theta_t} = 0$ (score has mean zero). Hessian: $-\mathbb{E}\nabla_\theta^2 \log \pi_\theta|_{\theta_t} = \mathbb{E}[\nabla \log \pi_{\theta_t} \nabla \log \pi_{\theta_t}^\top]$, since $\mathbb{E}[\nabla_\theta^2 \pi_\theta / \pi_\theta] = 0$. □

TRPO's Local Constraint is a Fisher Trust Region

Claim 12.5 (KL gradient and Hessian at θ_t)

For a horizon-H MDP and fixed θ_t,

$$\nabla_\theta \mathrm{KL}(\mathbb{P}_\mu^{\pi_{\theta_t}} \| \mathbb{P}_\mu^{\pi_\theta})\big|_{\theta_t} = 0, \quad \nabla_\theta^2 \mathrm{KL}\big|_{\theta_t} = H\, \mathbb{E}_{s,a\sim d^{\pi_{\theta_t}}}\left[\nabla \log \pi_{\theta_t}(a|s) \nabla \log \pi_{\theta_t}(a|s)^\top\right] =: F_{\theta_t}.$$

Proof. The trajectory KL telescopes over steps: $\mathrm{KL}(\mathbb{P}_\mu^{\pi_{\theta_t}} \| \mathbb{P}_\mu^{\pi_\theta}) = \sum_h \mathbb{E}_{s_h,a_h\sim\pi_{\theta_t}} \log \frac{\pi_{\theta_t}(a_h|s_h)}{\pi_\theta(a_h|s_h)}$ (μ, P cancel). Differentiating at $\theta = \theta_t$: $-\sum_h \mathbb{E}_{s_h}\mathbb{E}_{a_h\sim\pi_{\theta_t}} \nabla \log \pi_{\theta_t} = 0$ (score has mean zero). Hessian: $-\mathbb{E}\nabla_\theta^2 \log \pi_\theta|_{\theta_t} = \mathbb{E}[\nabla \log \pi_{\theta_t} \nabla \log \pi_{\theta_t}^\top]$, since $\mathbb{E}[\nabla_\theta^2 \pi_\theta / \pi_\theta] = 0$. □

Plug back in. The linear KL term vanishes, so TRPO's local QP is

$$\max_\Delta \left\langle \nabla_\theta V^{\pi_{\theta_t}}, \Delta \right\rangle \quad \text{s.t.} \quad \tfrac{1}{2}\Delta^\top F_{\theta_t} \Delta \le \delta.$$

Exactly the steepest-ascent-under-KL program we used to define the NPG (Ch. 10) — now derived from TRPO's trajectory-KL constraint.

TRPO *is* Natural Policy Gradient

Solving that Fisher trust-region QP (a Lagrangian / KKT computation) gives the natural gradient direction with a self-normalized step size:

$$\theta_{t+1} = \theta_t + \sqrt{\frac{\delta}{(\nabla V^{\pi_{\theta_t}})^\top F_{\theta_t}^{-1} \nabla V^{\pi_{\theta_t}}}} \, F_{\theta_t}^{-1} \nabla V^{\pi_{\theta_t}}.$$

TRPO *is* Natural Policy Gradient

Solving that Fisher trust-region QP (a Lagrangian / KKT computation) gives the natural gradient direction with a self-normalized step size:

$$\theta_{t+1} = \theta_t + \sqrt{\frac{\delta}{(\nabla V^{\pi_{\theta_t}})^\top F_{\theta_t}^{-1} \nabla V^{\pi_{\theta_t}}}}\, F_{\theta_t}^{-1} \nabla V^{\pi_{\theta_t}}.$$

The unification

TRPO = NPG with a trust-region-determined learning rate. So *all of Chapter 11's generalization/approximation analysis applies to TRPO*. The Fisher metric makes the update covariant — invariant to reparameterization ("covariant policy search").

TRPO *is* Natural Policy Gradient

Solving that Fisher trust-region QP (a Lagrangian / KKT computation) gives the natural gradient direction with a self-normalized step size:

$$\theta_{t+1} = \theta_t + \sqrt{\frac{\delta}{(\nabla V^{\pi_{\theta_t}})^\top F_{\theta_t}^{-1} \nabla V^{\pi_{\theta_t}}}}\, F_{\theta_t}^{-1} \nabla V^{\pi_{\theta_t}}.$$

The unification

TRPO = NPG with a trust-region-determined learning rate. So *all of Chapter 11's generalization/approximation analysis applies to TRPO*. The Fisher metric makes the update covariant — invariant to reparameterization ("covariant policy search").

This is why we spent so long on the NPG: the most-used deep-RL trust-region method is the NPG in disguise.

Proximal Policy Optimization (PPO)

Start from the same surrogate, importance-weighted. Using samples from the *old* policy π_{θ_t}, the improvement objective is, with the density ratio $r_\theta(s,a) := \frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}$,

$$\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}},\, a\sim\pi_{\theta_t}}\left[r_\theta(s,a)\, A^{\pi_{\theta_t}}(s,a)\right].$$

The danger. Maximizing this directly pushes $r_\theta \to \infty$ on good actions ($A > 0$) and $r_\theta \to 0$ on bad ones — driving π_θ far from π_{θ_t}, where the surrogate is no longer valid. TRPO fixed this with a KL constraint; PPO fixes it by clipping.

Proximal Policy Optimization (PPO)

Start from the same surrogate, importance-weighted. Using samples from the *old* policy π_{θ_t}, the improvement objective is, with the density ratio $r_\theta(s,a) := \frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}$,

$$\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}},\, a\sim\pi_{\theta_t}}\left[r_\theta(s,a)\, A^{\pi_{\theta_t}}(s,a)\right].$$

The danger. Maximizing this directly pushes $r_\theta \to \infty$ on good actions ($A > 0$) and $r_\theta \to 0$ on bad ones — driving π_θ far from π_{θ_t}, where the surrogate is no longer valid. TRPO fixed this with a KL constraint; PPO fixes it by clipping.

PPO clipped objective

$$L(\theta) = \mathbb{E}_{s,a}\Big[\min\big(\, r_\theta(s,a)\, A^{\pi_{\theta_t}}(s,a),\;\; \mathrm{clip}(r_\theta(s,a);\, 1-\varepsilon,\, 1+\varepsilon)\, A^{\pi_{\theta_t}}(s,a)\,\big)\Big].$$

Proximal Policy Optimization (PPO)

Start from the same surrogate, importance-weighted. Using samples from the *old* policy π_{θ_t}, the improvement objective is, with the density ratio $r_\theta(s,a) := \frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}$,

$$\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}},\, a\sim\pi_{\theta_t}}\left[r_\theta(s,a)\, A^{\pi_{\theta_t}}(s,a)\right].$$

The danger. Maximizing this directly pushes $r_\theta \to \infty$ on good actions ($A > 0$) and $r_\theta \to 0$ on bad ones — driving π_θ far from π_{θ_t}, where the surrogate is no longer valid. TRPO fixed this with a KL constraint; PPO fixes it by clipping.

PPO clipped objective

$$L(\theta) = \mathbb{E}_{s,a}\Big[\min\big(\, r_\theta(s,a)\, A^{\pi_{\theta_t}}(s,a),\;\; \mathrm{clip}(r_\theta(s,a);\, 1-\varepsilon,\, 1+\varepsilon)\, A^{\pi_{\theta_t}}(s,a)\,\big)\Big].$$

Read it case by case (the min keeps the *pessimistic* branch):

- $A > 0$ (good action): reward grows with r_θ, but is *capped* at $r_\theta = 1+\varepsilon$. Past that, the objective is flat — zero gradient, no reason to increase $\pi_\theta(a|s)$ further.
- $A < 0$ (bad action): reward grows as r_θ shrinks, but the min stops crediting improvement once $r_\theta = 1-\varepsilon$ — again flat beyond the clip.

So once the ratio leaves $[1-\varepsilon, 1+\varepsilon]$, the gradient vanishes: PPO *removes the incentive* to move the policy far, without any matrix inverse or explicit KL.

Proximal Policy Optimization (PPO)

Start from the same surrogate, importance-weighted. Using samples from the *old* policy π_{θ_t}, the improvement objective is, with the density ratio $r_\theta(s,a) := \frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}$,

$$\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}},\, a\sim\pi_{\theta_t}}\left[r_\theta(s,a)\, A^{\pi_{\theta_t}}(s,a)\right].$$

The danger. Maximizing this directly pushes $r_\theta \to \infty$ on good actions ($A > 0$) and $r_\theta \to 0$ on bad ones — driving π_θ far from π_{θ_t}, where the surrogate is no longer valid. TRPO fixed this with a KL constraint; PPO fixes it by clipping.

PPO clipped objective

$$L(\theta) = \mathbb{E}_{s,a}\Big[\min\big(\, r_\theta(s,a)\, A^{\pi_{\theta_t}}(s,a),\;\; \mathrm{clip}(r_\theta(s,a);\, 1-\varepsilon,\, 1+\varepsilon)\, A^{\pi_{\theta_t}}(s,a)\,\big)\Big].$$

Read it case by case (the min keeps the *pessimistic* branch):

- $A > 0$ (good action): reward grows with r_θ, but is *capped* at $r_\theta = 1+\varepsilon$. Past that, the objective is flat — zero gradient, no reason to increase $\pi_\theta(a|s)$ further.
- $A < 0$ (bad action): reward grows as r_θ shrinks, but the min stops crediting improvement once $r_\theta = 1-\varepsilon$ — again flat beyond the clip.

So once the ratio leaves $[1-\varepsilon, 1+\varepsilon]$, the gradient vanishes: PPO *removes the incentive* to move the policy far, without any matrix inverse or explicit KL.

Why it's popular. PPO runs *several* cheap mini-batch SGD steps on $L(\theta)$ per data batch (vs. NPG/TRPO's single Fisher-preconditioned step). Same "don't move too far" instinct, fully first-order.

The Through-Line

Method	**Guarantee**	**Key cost**
GA (softmax, exact)	global, *asymptotic*	no rate; needs $\mu > 0$
Log-barrier GA	$\widetilde{O}(\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2/\varepsilon^2)$	mismatch coeff. $\lVert d_\rho^{\pi^\star}/\mu\rVert_\infty$
NPG (softmax)	$O(1/\varepsilon)$, dimension-free	none (geometry cancels it)
NPG + func. approx.	competes with $\pi^\star$ up to fit error	compatible-FA error (AJKS §11.4)
TRPO	= NPG	trust-region δ
PPO	heuristic surrogate	clipping (first-order, cheap)

The Through-Line

Method	Guarantee	Key cost
GA (softmax, exact)	global, *asymptotic*	no rate; needs $\mu > 0$
Log-barrier GA	$\widetilde{O}(\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2/\varepsilon^2)$	mismatch coeff. $\lVert d_\rho^{\pi^\star}/\mu\rVert_\infty$
NPG (softmax)	$O(1/\varepsilon)$, dimension-free	none (geometry cancels it)
NPG + func. approx.	competes with $\pi^\star$ up to fit error	compatible-FA error (AJKS §11.4)
TRPO	= NPG	trust-region δ
PPO	heuristic surrogate	clipping (first-order, cheap)

Two ideas to remember

1. **Geometry matters.** Preconditioning by the Fisher information (NPG) turns a non-concave, $|\mathcal{S}|$-dependent problem into a dimension-free, mirror-descent one.
2. **Distribution control matters.** Every global guarantee is bought either with a mismatch coefficient (GA, approximation error) or by explicitly limiting how fast the visitation moves (TRPO/PPO trust regions).

Reading & References

Primary reference:

ABJKS Agarwal, Brantley, Jiang, Kakade, Sun, *Reinforcement Learning: Theory and Algorithms*, Chapters 9–12.

Foundational papers:

- Williams, *Simple statistical gradient-following algorithms* (REINFORCE), 1992.
- Sutton, McAllester, Singh, Mansour, *Policy gradient methods* (compatible FA), NeurIPS 1999.
- Kakade, *A natural policy gradient*, NeurIPS 2001.
- Kakade, Langford, *Approximately optimal approximate RL* (CPI), ICML 2002.
- Agarwal, Kakade, Lee, Mahajan, *On the theory of policy gradient methods*, JMLR 2021.

Trust-region / practical:

- Schulman et al., *Trust Region Policy Optimization* (TRPO), ICML 2015.
- Schulman et al., *Proximal Policy Optimization* (PPO), 2017.

Questions?