Lecture 10: Natural Policy Gradient

In Lecture 9, we developed the policy gradient theorem, the REINFORCE algorithm, and several practical policy optimization methods: CPI, TRPO, and PPO. A common theme was the challenge of *state distribution shift*: the gradient ∇V^{π_θ} depends on d^{π_θ}, which changes as θ changes. CPI addresses this through mixture updates; TRPO and PPO through KL-constrained or clipped updates.

This lecture takes a deeper look at the *geometry* of policy optimization. We introduce the natural policy gradient (NPG), which replaces the Euclidean geometry of parameter space with the Fisher information geometry of policy space. This leads to remarkably clean theoretical results: NPG achieves dimension-free global convergence at rate $O(1/T)$, depending only on $\log|\mathcal{A}|$ and not on $|\mathcal{S}|$. We then extend the analysis to function approximation settings.

Setup. We consider a class of parametric policies $\Pi = \{\pi_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^d\}$, where $\theta \mapsto \pi_\theta(\cdot \mid s)$ is differentiable for each state s. The goal is $\max_{\theta\in\Theta} V^{\pi_\theta}(\rho)$. The primary example is the *tabular softmax* parameterization with $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$:

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}} \exp(\theta_{s,a'})}.$$

Note that (the closure of) this policy class contains all stationary stochastic policies. In the function approximation section we will also consider *log-linear* policies $\pi_\theta(a \mid s) \propto \exp(\theta^\mathsf{T}\phi_{s,a})$ with features $\phi_{s,a} \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$.

From Gradient Descent to Natural Gradient

The Reparameterization Problem

A fundamental issue with vanilla policy gradient is that gradient descent is not invariant to reparameterization of the policy. Recall that gradient descent can be viewed as solving a trust-region subproblem:

$$\theta' = \operatorname*{argmin}_{\theta} \langle \nabla \ell(\theta_0), \theta - \theta_0 \rangle \quad \text{s.t.} \quad \|\theta - \theta_0\|_2^2 \le \delta.$$

This formulation implicitly uses the *Euclidean distance* in θ-space as the proximity measure.

Now consider a linear reparameterization $\theta = A\phi$ for an invertible matrix A. Gradient descent in ϕ-space yields:

$$\phi' = \phi - \eta \nabla_\phi \ell(\phi) = \phi - \eta A^\intercal \nabla_\theta \ell(\theta),$$

which, mapped back to θ-space, gives:

$$\theta' = A\phi' = \theta - \eta A A^\intercal \nabla_\theta \ell(\theta).$$

The update depends on the arbitrary matrix A — different parameterizations of the *same* policy class lead to different optimization trajectories. This is problematic because the choice of parameterization is often arbitrary and should not affect the optimization behavior.

To see why this matters for RL, consider two softmax policies π_θ and $\pi_{\theta'}$ that are close in parameter space ($\|\theta - \theta'\| \leq \epsilon$) but correspond to very different distributions over actions. Conversely, two parameter vectors far apart in Euclidean distance may correspond to nearly identical policies (e.g., adding a large constant to all logits for a given state). The Euclidean geometry of θ-space does not reflect the *statistical geometry* of the induced policy distributions.

The Euclidean geometry of θ-space does not reflect the geometry of policy space. Different reparameterizations (scaling, translation) lead to different gradient descent trajectories. This motivates measuring distances in *policy space* (e.g., via KL divergence) rather than parameter space, leading to the natural gradient.

Achieving Parameterization Invariance

We saw above that GD in the reparameterized Ω-space (where $w = A^{-1}\theta$) with the Euclidean trust region $\|w - w_0\|_2^2 \leq \delta$ yields $\theta' = \theta_0 - \eta A A^\intercal \nabla \ell(\theta_0)$, which depends on A. What if we use a *different* distance metric in Ω-space that compensates for the distortion?

Consider using the metric $(w - w_0)^\intercal (A^\intercal A)(w - w_0) \leq \delta$ instead. Solving the trust-region subproblem gives:

$$w' = w_0 - \eta (A^\intercal A)^{-1} \nabla_w g(w_0),$$

where $g(w) = \ell(Aw)$. Since $\nabla_w g(w) = A^\intercal \nabla_\theta \ell(\theta)|_{\theta = Aw}$, mapping back to θ-space:

$$\theta' = Aw' = \theta_0 - \eta A \cdot (A^\intercal A)^{-1} \cdot A^\intercal \nabla_\theta \ell(\theta_0) = \theta_0 - \eta \nabla_\theta \ell(\theta_0).$$

The matrix A cancels completely — the update is *invariant* to the linear reparameterization. The key was choosing the distance metric to "undo" the distortion introduced by A.

For policy optimization, the natural analog is to measure distances in *policy space* rather than parameter space. Since policies are probability distributions, the KL divergence is the canonical choice from information geometry (Amari, 1998). This leads to the trust-region subproblem:

$$\max_{\theta} \langle \nabla V^{\pi_{\theta_t}}(\rho), \theta - \theta_t \rangle \tag{1}$$

$$\text{s.t.} \quad \mathrm{KL}\big(\mathbb{P}_\mu^{\pi_{\theta_t}} \| \mathbb{P}_\mu^{\pi_\theta}\big) \le \delta. \tag{2}$$

Linearizing the objective (first-order Taylor) and quadratizing the KL constraint (second-order Taylor at θ_t) gives:

$$\begin{aligned} \max_{\theta} \quad & \langle \nabla V^{\pi_{\theta_t}}(\rho), \theta - \theta_t \rangle \\ \text{s.t.} \quad & \tfrac{1}{2}(\theta - \theta_t)^\mathsf{T} \nabla_\theta^2 \mathrm{KL}|_{\theta=\theta_t} (\theta - \theta_t) \le \delta. \end{aligned}$$

The key question is: what is the Hessian of the KL divergence? The answer is the *Fisher information matrix*.

Fisher Information Matrix

Definition 1 (Fisher information matrix)**.** *The (average) Fisher information matrix on the family* $\{\pi_\theta(\cdot \mid s) \mid s \in \mathcal{S}\}$ *is:*

$$\mathcal{F}_\rho^\theta := \mathbb{E}_{s \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[(\nabla \log \pi_\theta(a \mid s))(\nabla \log \pi_\theta(a \mid s))^\mathsf{T} \right].$$

Lemma 1 (Hessian of KL equals Fisher)**.** *Consider a finite-horizon MDP with horizon H. We have:*

$$\nabla_\theta \mathrm{KL}\big(\mathbb{P}_\mu^{\pi_{\theta_t}} \| \mathbb{P}_\mu^{\pi_\theta}\big)\big|_{\theta=\theta_t} = 0, \qquad \nabla_\theta^2 \mathrm{KL}\big(\mathbb{P}_\mu^{\pi_{\theta_t}} \| \mathbb{P}_\mu^{\pi_\theta}\big)\big|_{\theta=\theta_t} = H \cdot F_{\theta_t},$$

where $F_{\theta_t} = \mathbb{E}_{s,a \sim d^{\pi_{\theta_t}}} \left[\nabla \log \pi_{\theta_t}(a \mid s) (\nabla \log \pi_{\theta_t}(a \mid s))^\mathsf{T} \right]$.

Proof. Note that $\mathrm{KL}(\mathbb{P}_\mu^{\pi_{\theta_t}} \| \mathbb{P}_\mu^{\pi_\theta}) = \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim \mathbb{P}_h^{\pi_{\theta_t}}} \ln \frac{\pi_{\theta_t}(a_h|s_h)}{\pi_\theta(a_h|s_h)}$.

Gradient is zero: $\nabla_\theta \mathrm{KL}|_{\theta=\theta_t} = -\sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim \mathbb{P}_h^{\pi_{\theta_t}}} \nabla \log \pi_{\theta_t}(a_h \mid s_h) = 0$, since $\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[\nabla \log \pi_\theta(a \mid s)] = 0$ (the score function has zero mean).

Hessian:

$$\begin{aligned} \nabla_\theta^2 \mathrm{KL}|_{\theta=\theta_t} &= -\sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim \mathbb{P}_h^{\pi_{\theta_t}}} \nabla_\theta^2 \ln \pi_\theta(a_h \mid s_h)\big|_{\theta=\theta_t} \\ &= -\sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim \mathbb{P}_h^{\pi_{\theta_t}}} \left(\frac{\nabla^2 \pi_\theta(a_h \mid s_h)}{\pi_\theta(a_h \mid s_h)} - \nabla \log \pi_\theta(a_h \mid s_h) (\nabla \log \pi_\theta(a_h \mid s_h))^\mathsf{T} \right) \Bigg|_{\theta=\theta_t}. \end{aligned}$$

The first term vanishes since $\mathbb{E}_{a\sim\pi_\theta}\frac{\nabla^2\pi_\theta(a|s)}{\pi_\theta(a|s)} = \nabla^2\sum_a \pi_\theta(a \mid s) = 0$. Hence:

$$\nabla^2_\theta \mathrm{KL}|_{\theta=\theta_t} = \sum_{h=0}^{H-1}\mathbb{E}_{s_h,a_h\sim\mathbb{P}_h^{\pi_{\theta_t}}}\left[\nabla\log\pi_{\theta_t}(a_h \mid s_h)(\nabla\log\pi_{\theta_t}(a_h \mid s_h))^\top\right] = H\cdot F_{\theta_t}.$$

□

The Natural Policy Gradient

Substituting the Fisher information into the approximate trust region problem (absorbing the horizon factor H into δ):

$$\begin{aligned}\max_\theta \quad & \nabla V^{\pi_{\theta_t}}(\rho)^\top(\theta-\theta_t)\\ \text{s.t.} \quad & \tfrac{1}{2}(\theta-\theta_t)^\top F_{\theta_t}(\theta-\theta_t)\le\delta.\end{aligned}$$

By the method of Lagrange multipliers, the solution satisfies $\nabla V^{\pi_{\theta_t}}(\rho) = \lambda F_{\theta_t}(\theta-\theta_t)$, giving update direction $\theta-\theta_t \propto F_{\theta_t}^{-1}\nabla V^{\pi_{\theta_t}}(\rho)$ when F_{θ_t} is invertible (more generally, the minimum-norm solution $F_{\theta_t}^\dagger\nabla V$). The trust region radius δ only affects the step size, not the direction. This yields the *natural policy gradient* (NPG) (Kakade, 2001; Amari, 1998):

Natural Policy Gradient update:

$$\theta_{k+1} = \theta_k + \eta\,(\mathcal{F}_\rho^{\theta_k})^\dagger\nabla V^{\pi_k}(\rho),$$

where $M^\dagger$ denotes the Moore-Penrose pseudoinverse ($M^\dagger = M^{-1}$ when M is invertible).

Connection to TRPO (Lecture 9). Solving the trust-region subproblem exactly gives an adaptive step size $\eta = \sqrt{2\delta/(g^\top F^{-1}g)}$ where $g=\nabla V$. This is precisely TRPO (Schulman et al., 2015). Historically, NPG (Kakade, 2001) and covariant policy search (Bagnell and Schneider, 2003) came first; TRPO (Schulman et al., 2015) is a practical refinement that replaces NPG's fixed step size with an adaptive one determined by the trust region radius δ.

Algorithm 1 Natural Policy Gradient (NPG) (Kakade, 2001)

```
Require: Initial parameters $\theta^{(0)}$, learning rate $\eta > 0$, iterations $T$
for $t = 0, 1, \ldots, T-1$ do
    Compute (or estimate) $\nabla V^{\pi_{\theta^{(t)}}}(\rho)$ and $\mathcal{F}_\rho^{\theta^{(t)}}$
    Update $\theta^{(t+1)} = \theta^{(t)} + \eta\,(\mathcal{F}_\rho^{\theta^{(t)}})^\dagger \nabla V^{\pi_{\theta^{(t)}}}(\rho)$
end for
```

Geometric intuition: why the Fisher metric helps near deterministic policies. Consider a 2-action softmax with a single parameter θ: $(p_1, p_2) = \left(\frac{e^\theta}{1+e^\theta}, \frac{1}{1+e^\theta}\right)$. The Fisher information is $F_\theta = p_1 p_2$. As $\theta \to \infty$, the policy approaches a deterministic one ($p_1 \to 1$) and $F_\theta \to 0^+$. The Fisher trust region $F_{\theta_0}(\theta - \theta_0)^2 \le \delta$ then allows $(\theta - \theta_0)^2 \le \delta / F_{\theta_0} \to \infty$: the step size in θ-space grows without bound.

In other words, plain GD in θ-space moves at constant speed toward $\theta = \infty$ (to reach a deterministic policy), while NPG *accelerates*: it takes increasingly large steps as the policy approaches the simplex boundary. This is precisely where vanilla PG suffers from vanishing gradients — the Fisher metric automatically compensates for the saturation of the softmax.

Natural Policy Gradient

The NPG algorithm performs gradient updates in the geometry induced by the Fisher information matrix (Kakade, 2001; Amari, 1998):

$$\theta^{(t+1)} = \theta^{(t)} + \eta\,(\mathcal{F}_\rho^{\theta^{(t)}})^\dagger \nabla V^{(t)}(\rho).$$

For the softmax parameterization, $\mathcal{F}_\rho^\theta$ is singular because the logits are invariant to per-state translations ($\theta_{s,a} \to \theta_{s,a} + c_s$ does not change π_θ), so the pseudoinverse is necessary. For log-linear policies, $\mathcal{F}_\rho^\theta$ is typically invertible and $(\mathcal{F}_\rho^\theta)^\dagger = (\mathcal{F}_\rho^\theta)^{-1}$.

NPG as Soft Policy Iteration

For the softmax parameterization, the NPG update takes a remarkably clean form.

Lemma 2 (Closed-form NPG for softmax)**.** *For the softmax policy parameterization, the NPG updates take the form:*

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}, \quad \textit{and} \quad \pi^{(t+1)}(a \mid s) = \pi^{(t)}(a \mid s) \frac{\exp\left(\eta A^{(t)}(s,a)/(1-\gamma)\right)}{Z_t(s)},$$

where $Z_t(s) = \sum_{a \in \mathcal{A}} \pi^{(t)}(a \mid s) \exp\left(\eta A^{(t)}(s,a)/(1-\gamma)\right)$ is the normalization constant.

Proof. By definition of the Moore-Penrose pseudoinverse, $(\mathcal{F}_\rho^\theta)^\dagger \nabla V^{\pi_\theta}(\rho) = w_\star$ if and only if $w_\star$ is the minimum norm solution of $\min_w \|\nabla V^{\pi_\theta}(\rho) - \mathcal{F}_\rho^\theta w\|^2$.

Expanding the definition of $\mathcal{F}_\rho^\theta$:

$$\begin{aligned}\mathcal{F}_\rho^\theta w &= \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla \log \pi_\theta(a \mid s) \cdot (\nabla \log \pi_\theta(a \mid s))^\intercal w\right] \\ &= \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla \log \pi_\theta(a \mid s) \cdot w^\intercal \nabla \log \pi_\theta(a \mid s)\right].\end{aligned}$$

For the softmax parameterization, $\frac{\partial}{\partial\theta_{s',a'}} \log \pi_\theta(a \mid s) = \mathbb{1}[s = s'](\mathbb{1}[a = a'] - \pi_\theta(a' \mid s))$, so $w^\intercal \nabla \log \pi_\theta(a \mid s) = w_{s,a} - \overline{w}_s$ where $\overline{w}_s = \sum_{a'} \pi_\theta(a' \mid s) w_{s,a'}$. This gives:

$$\mathcal{F}_\rho^\theta w = \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla \log \pi_\theta(a \mid s) \cdot (w_{s,a} - \overline{w}_s)\right].$$

Computing the residual: by the policy gradient theorem, $\nabla V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\nabla \log \pi_\theta(a \mid s) \cdot A^{\pi_\theta}(s,a)]$. Both ∇V and $\mathcal{F}_\rho^\theta w$ have the form $\mathbb{E}_{s,a}[\nabla \log \pi_\theta(a \mid s) \cdot f(s,a)]$, so their difference is:

$$\nabla V^{\pi_\theta}(\rho) - \mathcal{F}_\rho^\theta w = \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla \log \pi_\theta(a \mid s) \cdot \left(\frac{1}{1-\gamma} A^{\pi_\theta}(s,a) - w_{s,a} + \overline{w}_s\right)\right].$$

For the softmax parameterization, $\nabla \log \pi_\theta(a \mid s)$ is nonzero only in the block of coordinates corresponding to state s, so $\mathcal{F}_\rho^\theta$ is block-diagonal across states. The squared norm therefore decomposes as:

$$\|\nabla V^{\pi_\theta}(\rho) - \mathcal{F}_\rho^\theta w\|^2 = \sum_{s,a} d^{\pi_\theta}(s) \pi_\theta(a \mid s) \left(\frac{1}{1-\gamma} A^{\pi_\theta}(s,a) - w_{s,a} + \overline{w}_s\right)^2.$$

Setting $w = \frac{1}{1-\gamma} A^{\pi_\theta}$ achieves zero error (since A^{π_θ} already satisfies $\sum_a \pi(a \mid s) A^\pi(s,a) = 0$, the centering term vanishes), and this is the minimum norm solution. So $(\mathcal{F}_\rho^\theta)^\dagger \nabla V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} A^{\pi_\theta}$, and the NPG parameter update $\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla V$ becomes:

$$\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}(s,a).$$

Substituting into the softmax policy $\pi_\theta(a \mid s) = \exp(\theta_{s,a}) / \sum_{a'} \exp(\theta_{s,a'})$:

$$\pi^{(t+1)}(a \mid s) = \frac{\exp\left(\theta_{s,a}^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}(s,a)\right)}{\sum_{a'} \exp\left(\theta_{s,a'}^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}(s,a')\right)} = \frac{\pi^{(t)}(a \mid s) \exp\left(\frac{\eta A^{(t)}(s,a)}{1-\gamma}\right)}{\sum_{a'} \pi^{(t)}(a' \mid s) \exp\left(\frac{\eta A^{(t)}(s,a')}{1-\gamma}\right)},$$

which gives the result with $Z_t(s) = \sum_{a'} \pi^{(t)}(a' \mid s) \exp\left(\eta A^{(t)}(s,a') / (1-\gamma)\right)$. □

NPG for tabular softmax = soft policy iteration:

$$\pi^{(t+1)}(a \mid s) \propto \pi^{(t)}(a \mid s) \exp\left(\frac{\eta Q^{\pi_t}(s,a)}{1-\gamma}\right).$$

This equivalence is *specific to the tabular softmax* parameterization, where each (s,a) pair has an independent parameter. In this case, the compatible function approximation error is exactly zero (the score function features are rich enough to perfectly represent the advantage). For log-linear or neural policies, the NPG direction instead uses the *projected* advantage $\widehat{A}(s,a) = w^{*\mathsf{T}} \nabla \log \pi_\theta(a \mid s)$ from the compatible FA regression, which generally incurs approximation error. The resulting update is "soft PI with approximate advantages."
Compare with hard PI from Lecture 4: PI takes $\pi^{(t+1)}(s) = \operatorname{argmax}_a Q^{(t)}(s,a)$ (the $\eta \to \infty$ limit). The Fisher information cancels out the state distribution d^π, making the update distribution-free. This is why NPG achieves dimension-free rates.

Mirror Descent Interpretation

The NPG update can also be interpreted through the lens of *mirror descent* (Even-Dar et al., 2009; Beck, 2017), which provides a complementary perspective that is independent of the softmax parameterization.

Recall that in standard (projected) gradient ascent, we solve:

$$\theta^{(t+1)} = \operatorname*{argmin}_\theta \left\{ -\eta \langle \nabla f(\theta^{(t)}), \theta \rangle + \frac{1}{2}\|\theta - \theta^{(t)}\|_2^2 \right\}.$$

Mirror descent replaces the Euclidean distance $\|\cdot\|_2^2$ with a Bregman divergence $D_\psi(\cdot\|\cdot)$ that better reflects the geometry of the constraint set. For the probability simplex (policies are distributions over actions), the natural choice is the negative entropy $\psi(\pi) = \sum_a \pi(a) \log \pi(a)$, whose Bregman divergence is the KL divergence.

Applying mirror descent to the policy optimization problem *separately for each state s*:

$$\pi^{(t+1)}(\cdot \mid s) = \operatorname*{argmin}_{\pi(\cdot|s) \in \Delta_{\mathcal{A}}} \left\{ -\frac{\eta}{1-\gamma} \sum_a \pi(a \mid s) A^{\pi_t}(s,a) + \mathrm{KL}(\pi(\cdot \mid s) \| \pi^{(t)}(\cdot \mid s)) \right\}, \quad \forall s.$$

This is a convex optimization over the simplex $\Delta_{\mathcal{A}}$. Setting the KKT conditions and enforcing the normalization constraint $\sum_a \pi(a) = 1$ yields:

$$\pi^{(t+1)}(a \mid s) = \frac{\pi^{(t)}(a \mid s) \exp\left(\frac{\eta}{1-\gamma} A^{(t)}(s,a)\right)}{\sum_{a'} \pi^{(t)}(a' \mid s) \exp\left(\frac{\eta}{1-\gamma} A^{(t)}(s,a')\right)},$$

which is exactly the softmax NPG update from Lemma 2. This confirms that NPG in parameter space corresponds to mirror descent in policy space.

The mirror descent viewpoint is powerful because it: (1) applies to any policy class (not just softmax), (2) explains why KL divergence is the natural distance measure for policy optimization (it is the Bregman divergence of the entropy), and (3) provides access to the rich convergence theory of online convex optimization (Beck, 2017).

Global Convergence of Softmax Policy Gradient

Vanishing Gradients

Despite the favorable properties of the policy gradient theorem, the softmax parameterization creates challenges for gradient ascent. The core issue is that policy gradient methods rely on on-policy samples: the gradient is computed under the current policy's state-action distribution. When the current policy rarely visits high-reward regions of the state space, the gradient signal for improving in those regions becomes exponentially small.

Proposition 3 (Vanishing gradients at suboptimal parameters (Agarwal et al., 2021, Proposition 10.1))**.** *Consider the chain MDP of length $H + 2$ in Figure 1. With the direct policy parameterization $\pi_\theta(a \mid s) = \theta_{s,a}$ and $\gamma = H/(H+1)$, suppose $0 < \theta < 1$ componentwise and $\theta_{s,a_1} < 1/4$ for all states s. For all $k \le \frac{H}{40\log(2H)} - 1$, we have $\|\nabla^k V^{\pi_\theta}(s_0)\| \le (1/3)^{H/4}$, where ∇^k is the kth-order derivative tensor. Furthermore, $V^\star(s_0) - V^{\pi_\theta}(s_0) \ge (H+1)/8 - (H+1)^2/3^H$.*

This result is stated without proof. To build intuition, consider the chain MDP in Figure 1. The only reward is at the end of the chain: $r(s_{H+1}, a_1) = 1$. The optimal policy selects action a_1 at every state, traversing the chain to reach s_{H+1} and accumulating reward. But if $\theta_{s,a_1} < 1/4$ everywhere, the probability of traversing the entire chain is at most $(1/4)^{H+1}$ — exponentially small in H. By the policy gradient theorem, the gradient at state s_0 involves $d^{\pi_\theta}(s_{H+1})$, which is exponentially small.

The proposition is much stronger: not only the gradient, but all derivatives up to order $\Omega(H/\log H)$ are exponentially small. This means that even higher-order methods (Newton, natural gradient) cannot help — the objective function is exponentially flat around

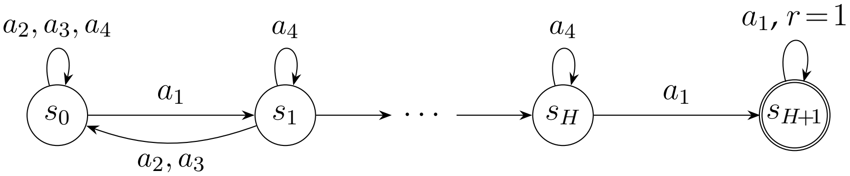

Figure 1: Chain MDP of length $H + 2$ illustrating vanishing gradients. At each state s_i: a_1 moves forward ($s_i \to s_{i+1}$), a_2 and a_3 move backward ($s_i \to s_{\max(i-1,0)}$), a_4 is a self-loop. Rewards are 0 everywhere except $r(s_{H+1}, a_1) = 1$.

suboptimal parameters. In other words, the optimization landscape has large, nearly flat plateaus where gradient-based methods make negligible progress.

This is fundamentally an *exploration* problem: the current policy does not explore enough to discover the high-reward region. It motivates two approaches:

(a) **Regularization** (Section below): log-barrier regularization prevents policies from becoming too deterministic, maintaining exploration.

(b) **Natural policy gradient**: NPG's Fisher information geometry cancels the state distribution factor, yielding updates that are insensitive to the exploration distribution.

Softmax Gradient Structure

Lemma 4 (Softmax policy gradient (Agarwal et al., 2021, Lemma 10.2))**.** *For the softmax policy class, we have:*

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

Proof. Using the advantage form of the policy gradient (Lecture 9) and the softmax score function:

$$\frac{\partial \log \pi_\theta(a \mid s)}{\partial \theta_{s',a'}} = \mathbb{1}[s = s'] \Big(\mathbb{1}[a = a'] - \pi_\theta(a' \mid s) \Big),$$

we compute:

$$\begin{aligned}
\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s',a'}} &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[A^{\pi_\theta}(s,a) \frac{\partial \log \pi_\theta(a \mid s)}{\partial \theta_{s',a'}} \right] \\
&= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s') \mathbb{E}_{a \sim \pi_\theta(\cdot|s')} \left[A^{\pi_\theta}(s',a) \big(\mathbb{1}[a = a'] - \pi_\theta(a' \mid s') \big) \right] \\
&= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s') \left(\pi_\theta(a' \mid s') A^{\pi_\theta}(s',a') - \pi_\theta(a' \mid s') \sum_a \pi_\theta(a \mid s') A^{\pi_\theta}(s',a) \right) \\
&= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s') \pi_\theta(a' \mid s') A^{\pi_\theta}(s',a'),
\end{aligned}$$

where the last step uses $\sum_a \pi(a \mid s) A^\pi(s,a) = 0$. □

Asymptotic Global Convergence (Without Regularization)

Despite the vanishing gradient issue, the softmax policy gradient with exact gradients does converge to a global optimum — under a coverage assumption on the optimization distribution.

Theorem 5 (Asymptotic global convergence (Agarwal et al., 2021, Theorem 10.3); (Mei et al., 2020))**.** *Consider the softmax parameterization. Run gradient ascent $\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_{\theta^{(t)}}}(\mu)$, where μ is strictly positive, i.e., $\mu(s) > 0$ for all states s. For $\eta \le (1-\gamma)^3/8$:*

$$V^{(t)}(s) \to V^\star(s), \quad \text{as } t \to \infty, \quad \text{for all states } s.$$

This result is stated without proof. We highlight two important caveats:

(a) **The convergence rate can be exponentially slow** in $|\mathcal{S}|$ and $1/(1-\gamma)$. The vanishing gradient landscape (Proposition 3) is not just a theoretical curiosity — it directly manifests as exponentially slow progress.

(b) **The condition $\mu(s) > 0$ for all s is conjectured to be necessary.** Without full coverage, the on-policy gradient may never "discover" states that the current policy does not visit.

This theorem completes the landscape picture for vanilla softmax PG: it converges globally, but potentially at an exponential rate. The next two approaches improve this:

- Log-barrier regularization (below): polynomial rate $O(|\mathcal{S}|^2|\mathcal{A}|^2/\epsilon^2)$.
- NPG (next section): dimension-free rate $O(\log|\mathcal{A}|/\epsilon)$, with no dependence on $|\mathcal{S}|$.

Global Convergence via Regularization

The vanishing gradient problem arises because near-deterministic policies have $\pi_\theta(a \mid s) \approx 0$ for most actions, causing the gradient factor $d^\pi(s)\pi(a \mid s)$ to vanish. A natural fix is to add a *log-barrier* regularizer that penalizes $\pi_\theta(a \mid s) \to 0$, since $\log \pi_\theta(a \mid s) \to -\infty$. This keeps all action probabilities bounded away from zero, maintaining exploration. Define:

$$L_\lambda(\theta) := V^{\pi_\theta}(\mu) + \frac{\lambda}{|\mathcal{S}||\mathcal{A}|} \sum_{s,a} \log \pi_\theta(a \mid s) + \lambda \log |\mathcal{A}|,$$

where $\lambda > 0$ is a regularization parameter. The barrier term is equivalent (up to constants) to $-\frac{\lambda}{|\mathcal{S}|}\sum_s \mathrm{KL}(\mathrm{Unif}_\mathcal{A} \| \pi_\theta(\cdot \mid s))$, so the regularizer encourages the policy to stay close to the uniform distribution.

Theorem 6 (Global convergence with log barrier)**.** *Suppose θ is such that $\|\nabla L_\lambda(\theta)\|_2 \leq \epsilon_{\text{opt}}$ and $\epsilon_{\text{opt}} \leq \lambda/(2|\mathcal{S}|\,|\mathcal{A}|)$. Then for all starting state distributions ρ:*

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

Proof. It suffices to show that $\max_a A^{\pi_\theta}(s,a) \leq 2\lambda/(\mu(s)|\mathcal{S}|)$ for all states s. Indeed, by the performance difference lemma (Lecture 9):

$$\begin{aligned} V^\star(\rho) - V^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\sum_a \pi^\star(a \mid s)A^{\pi_\theta}(s,a) \\ &\leq \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\max_{a\in\mathcal{A}} A^{\pi_\theta}(s,a) \leq \frac{2\lambda}{1-\gamma}\max_s \frac{d_\rho^{\pi^\star}(s)}{\mu(s)}. \end{aligned}$$

We now show $\max_a A^{\pi_\theta}(s,a) \leq 2\lambda/(\mu(s)|\mathcal{S}|)$. Consider any (s,a) with $A^{\pi_\theta}(s,a) > 0$. Using the softmax gradient (Lemma 4):

$$\frac{\partial L_\lambda(\theta)}{\partial\theta_{s,a}} = \frac{1}{1-\gamma}d_\mu^{\pi_\theta}(s)\pi_\theta(a \mid s)A^{\pi_\theta}(s,a) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a \mid s)\right). \tag{3}$$

The gradient norm assumption $\epsilon_{\text{opt}} \geq \frac{\partial L_\lambda(\theta)}{\partial\theta_{s,a}}$ implies:

$$\epsilon_{\text{opt}} \geq \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a \mid s)\right),$$

since $A^{\pi_\theta}(s,a) \geq 0$. Using $\epsilon_{\text{opt}} \leq \lambda/(2|\mathcal{S}|\,|\mathcal{A}|)$, we get $\pi_\theta(a \mid s) \geq 1/(2|\mathcal{A}|)$.

Solving (3) for $A^{\pi_\theta}(s,a)$ (isolate the advantage term on the left, divide by its coefficient $\frac{1}{1-\gamma}d_\mu^{\pi_\theta}(s)\pi_\theta(a \mid s)$, then distribute $1/\pi_\theta(a \mid s)$):

$$\begin{aligned} A^{\pi_\theta}(s,a) &= \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a \mid s)}\frac{\partial L_\lambda(\theta)}{\partial\theta_{s,a}} + \frac{\lambda}{|\mathcal{S}|}\left(1 - \frac{1}{\pi_\theta(a \mid s)|\mathcal{A}|}\right)\right) \\ &\leq \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(2|\mathcal{A}|\epsilon_{\text{opt}} + \frac{\lambda}{|\mathcal{S}|}\right) \leq 2\frac{\lambda}{|\mathcal{S}|}\frac{1-\gamma}{d_\mu^{\pi_\theta}(s)} \leq \frac{2\lambda}{\mu(s)|\mathcal{S}|}, \end{aligned}$$

where we used $\frac{\partial L_\lambda}{\partial\theta_{s,a}} \leq \epsilon_{\text{opt}}$ with $\pi_\theta(a \mid s) \geq 1/(2|\mathcal{A}|)$ for the first term, $1 - \frac{1}{\pi|\mathcal{A}|} \leq 1$ for the second, $\epsilon_{\text{opt}} \leq \lambda/(2|\mathcal{S}|\,|\mathcal{A}|)$ for the penultimate step, and $d_\mu^{\pi_\theta}(s) \geq (1-\gamma)\mu(s)$ for the last. □

Corollary 7 (Iteration complexity)**.** *Let $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$. Starting from initial $\theta^{(0)}$, with*

$\lambda = \frac{\epsilon(1-\gamma)}{2\|d_\rho^{\pi^\star}/\mu\|_\infty}$ *and* $\eta = 1/\beta_\lambda$*, we have:*

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

Proof. Applying gradient ascent with exact gradients and stepsize $1/\beta_\lambda$ (cf. the deterministic convergence result in Lecture 9):

$$\min_{t\leq T} \|\nabla L_\lambda(\theta^{(t)})\|^2 \leq \frac{2\beta_\lambda}{T(1-\gamma)}.$$

We need $\epsilon_{\text{opt}} = \sqrt{2\beta_\lambda/((1-\gamma)T)} \leq \lambda/(2|\mathcal{S}|\,|\mathcal{A}|)$. Choosing $T \geq 8\beta_\lambda|\mathcal{S}|^2|\mathcal{A}|^2/((1-\gamma)\lambda^2)$ suffices. Substituting λ and bounding β_λ yields:

$$\frac{8\beta_\lambda|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)\lambda^2} \leq \frac{80|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^4\lambda^2} = \frac{320|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\epsilon^2} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty^2 ,$$

where we used $\lambda < 1$ in bounding β_λ. □

The distribution mismatch coefficient $\|d_\rho^{\pi^\star}/\mu\|_\infty$ is unavoidable — it measures how well the optimization distribution μ covers the states that matter for $\pi^\star$. The rate depends polynomially on $|\mathcal{S}|$, $|\mathcal{A}|$.

Global Convergence of NPG (Softmax Case)

Throughout this section, we use the shorthand $\pi^{(t)} := \pi_{\theta^{(t)}}$, $V^{(t)} := V^{\pi^{(t)}}$, $A^{(t)} := A^{\pi^{(t)}}$, and $d_\mu^{(t)} := d_\mu^{\pi^{(t)}}$. The NPG iterates follow the soft PI update (Lemma 2):

$$\pi^{(t+1)}(a \mid s) = \pi^{(t)}(a \mid s) \frac{\exp\bigl(\eta A^{(t)}(s,a)/(1-\gamma)\bigr)}{Z_t(s)},$$

where $Z_t(s) := \sum_{a'} \pi^{(t)}(a' \mid s) \exp\bigl(\eta A^{(t)}(s,a')/(1-\gamma)\bigr)$ is the partition function (normalization constant).

Monotonic Improvement

Lemma 8 (Improvement lower bound (Agarwal et al., 2021, Lemma 10.8)). *For the iterates $\pi^{(t)}$ generated by the NPG updates, we have for all starting state distributions μ:*

$$V^{(t+1)}(\mu) - V^{(t)}(\mu) \geq \frac{(1-\gamma)}{\eta}\mathbb{E}_{s\sim\mu}\left[\log Z_t(s)\right] \geq 0.$$

Proof. First, $\log Z_t(s) \geq 0$ by Jensen's inequality:

$$\log Z_t(s) = \log \sum_a \pi^{(t)}(a \mid s) \exp\left(\frac{\eta A^{(t)}(s,a)}{1-\gamma}\right) \geq \sum_a \pi^{(t)}(a \mid s) \cdot \frac{\eta A^{(t)}(s,a)}{1-\gamma} = 0,$$

since $\sum_a \pi^{(t)}(a \mid s) A^{(t)}(s,a) = 0$.

By the performance difference lemma:

$$\begin{aligned}
V^{(t+1)}(\mu) - V^{(t)}(\mu) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{(t+1)}} \sum_a \pi^{(t+1)}(a \mid s) A^{(t)}(s,a) \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d_\mu^{(t+1)}} \sum_a \pi^{(t+1)}(a \mid s) \log \frac{\pi^{(t+1)}(a \mid s) Z_t(s)}{\pi^{(t)}(a \mid s)} \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d_\mu^{(t+1)}} \left[\mathrm{KL}(\pi_s^{(t+1)} \| \pi_s^{(t)}) + \log Z_t(s)\right] \\
&\geq \frac{1}{\eta}\mathbb{E}_{s\sim d_\mu^{(t+1)}} \log Z_t(s) \geq \frac{1-\gamma}{\eta}\mathbb{E}_{s\sim\mu} \log Z_t(s),
\end{aligned}$$

where the last inequality uses $d_\mu^{(t+1)} \geq (1-\gamma)\mu$ and $\log Z_t(s) \geq 0$. □

Dimension-Free Convergence Rate

Theorem 9 (NPG global convergence (Agarwal et al., 2021, Theorem 10.7)). *Suppose we run the NPG updates with $\rho \in \Delta(\mathcal{S})$ and $\theta^{(0)} = 0$. Fix $\eta > 0$. For all $T > 0$:*

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

Setting $\eta = (1-\gamma)^2 \log|\mathcal{A}|$: to achieve ϵ-optimality, it suffices to take $T = \frac{2}{(1-\gamma)^2\epsilon}$.

Proof. Since ρ is fixed, write $d^\star = d_\rho^{\pi^\star}$, $\pi_s^\star = \pi^\star(\cdot \mid s)$. By the performance difference lemma:

$$V^{\pi^\star}(\rho) - V^{(t)}(\rho) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star} \sum_a \pi^\star(a \mid s) A^{(t)}(s,a)$$

$$
\begin{aligned}
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\sum_a \pi^\star(a \mid s)\log\frac{\pi^{(t+1)}(a \mid s)Z_t(s)}{\pi^{(t)}(a \mid s)} \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\Big(\mathrm{KL}(\pi_s^\star\|\pi_s^{(t)}) - \mathrm{KL}(\pi_s^\star\|\pi_s^{(t+1)}) + \sum_a \pi^\star(a \mid s)\log Z_t(s)\Big).
\end{aligned}
$$

By Lemma 8 applied with $d^\star$ as the starting state distribution:

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

Averaging over $t = 0, \dots, T-1$ and using monotonicity $V^{(T)}(\rho) \ge V^{(T-1)}(\rho)$:

$$
\begin{aligned}
V^{\pi^\star}(\rho) - V^{(T)}(\rho) &\le \frac{1}{T}\sum_{t=0}^{T-1}\big(V^{\pi^\star}(\rho) - V^{(t)}(\rho)\big) \\
&\le \frac{1}{\eta T}\mathbb{E}_{s\sim d^\star}\mathrm{KL}(\pi_s^\star\|\pi_s^{(0)}) + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\big(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\big) \\
&= \frac{\mathbb{E}_{s\sim d^\star}\mathrm{KL}(\pi_s^\star\|\pi_s^{(0)})}{\eta T} + \frac{V^{(T)}(d^\star) - V^{(0)}(d^\star)}{(1-\gamma)T} \\
&\le \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T}.
\end{aligned}
$$

The last step uses: (i) $\theta^{(0)} = 0$ implies $\pi^{(0)}$ is uniform, so $\mathrm{KL}(\pi_s^\star\|\pi_s^{(0)}) \le \log|\mathcal{A}|$; and (ii) $V^{(T)}(d^\star) - V^{(0)}(d^\star) \le 1/(1-\gamma)$ since all values are bounded in $[0, 1/(1-\gamma)]$. □

NPG convergence — dimension-free rate:

$$
V^{\pi^\star}(\rho) - V^{(T)}(\rho) \le \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T}. \qquad \text{Complexity: } T = O\left(\frac{1}{(1-\gamma)^2\epsilon}\right).
$$

No dependence on $|\mathcal{S}|$!

Comparison of policy gradient methods:

Method	Convergence	Rate	Key dependencies
Vanilla PG	Stationary points	$O(1/\sqrt{T})$	β, σ^2
Softmax PG + log barrier	Global	$O(1/\sqrt{T})$	$\lvert\mathcal{S}\rvert^2$, $\lvert\mathcal{A}\rvert^2$, $\|d^{\pi^\star}/\mu\|_\infty^2$
NPG (softmax)	Global	$O(1/T)$	$\log\lvert\mathcal{A}\rvert$ only

The dimension-free convergence of NPG is remarkable. To appreciate why, recall that the vanilla softmax PG with log-barrier (Corollary 7) requires $T = \widetilde{O}(|\mathcal{S}|^2|\mathcal{A}|^2/\epsilon^2)$ iterations, with polynomial dependence on both $|\mathcal{S}|$ and $|\mathcal{A}|$. The NPG result eliminates all dependence on $|\mathcal{S}|$ and replaces the polynomial $|\mathcal{A}|$ dependence with $\log|\mathcal{A}|$. Moreover, the rate is $O(1/T)$ rather than $O(1/\sqrt{T})$.

The key mechanism is that the Fisher information matrix cancels the state distribution $d^{\pi_\theta}(s)$ from the gradient, yielding an update that acts *independently at each state*. In contrast, vanilla PG mixes the per-state updates through $d^{\pi_\theta}(s)$, creating coupling that slows convergence. This "distribution-free" property is what enables NPG to avoid the dependence on the exploration distribution that plagues vanilla PG.

NPG with Function Approximation

Compatible Function Approximation

In the function approximation regime, the policy class $\Pi = \{\pi_\theta \mid \theta \in \mathbb{R}^d\}$ may not contain all stochastic policies. The NPG update leverages the notion of *compatible function approximation* (Sutton et al., 1999).

Lemma 10 (Gradients and compatible FA). *Let $w^\star$ denote the minimizer of the compatible function approximation error:*

$$w^\star \in \operatorname*{argmin}_w \mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)} \left[\Big(A^{\pi_\theta}(s,a) - w \cdot \nabla \log \pi_\theta(a \mid s) \Big)^2 \right].$$

Define the best linear predictor $\widehat{A}^{\pi_\theta}(s,a) := w^\star \cdot \nabla \log \pi_\theta(a \mid s)$. Then:

$$\nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \Big[\nabla \log \pi_\theta(a \mid s) \cdot \widehat{A}^{\pi_\theta}(s,a) \Big].$$

Proof. The first-order optimality conditions for $w^\star$ imply:

$$\mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)} \left[\big(A^{\pi_\theta}(s,a) - w^\star \cdot \nabla \log \pi_\theta(a \mid s) \big) \nabla \log \pi_\theta(a \mid s) \right] = 0.$$

Rearranging and using the advantage form of the policy gradient (Lecture 9):

$$\nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s,a} \left[\nabla \log \pi_\theta(a \mid s) \cdot A^{\pi_\theta}(s,a) \right] = \frac{1}{1-\gamma} \mathbb{E}_{s,a} \Big[\nabla \log \pi_\theta(a \mid s) \cdot \widehat{A}^{\pi_\theta}(s,a) \Big].$$

□

Lemma 11 (NPG direction as regression)**.** *We have:*

$$(\mathcal{F}_\rho^\theta)^\dagger \nabla V^\theta(\rho) = \frac{1}{1-\gamma} w^\star,$$

where $w^\star$ is a minimizer of $\mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}[(w^\intercal \nabla \log \pi_\theta(a \mid s) - A^{\pi_\theta}(s,a))^2]$.

Proof. By the first-order optimality conditions (Lemma 10) and the advantage form of the policy gradient: $\nabla V^\theta(\rho) = \frac{1}{1-\gamma}\mathcal{F}_\rho^\theta w^\star$, so $w^\star = (1-\gamma)(\mathcal{F}_\rho^\theta)^{-1}\nabla V^\theta(\rho)$, i.e., $(\mathcal{F}_\rho^\theta)^\dagger \nabla V^\theta(\rho) = \frac{1}{1-\gamma} w^\star$. □

The NPG direction = solving a regression problem: predict $A^{\pi_\theta}(s,a)$ using score function features $\nabla \log \pi_\theta(a \mid s)$. This gives a practical recipe: fit a linear model, and use its weights as the update direction. This is independent of the dimension of the state space.

Q-NPG and the Regret Lemma

We consider a variant of NPG called *Q-NPG*, where instead of regressing advantages, we regress Q-values: $w^\star = \operatorname{argmin}_w \mathbb{E}_{s,a\sim d^{(t)}}[(Q^{\pi_t}(s,a) - w \cdot \phi_{s,a})^2]$. For log-linear policies with features $\phi_{s,a}$, the centered features $\overline{\phi}^{\theta}_{s,a} = \phi_{s,a} - \mathbb{E}_{a'\sim\pi_\theta}[\phi_{s,a'}]$ serve as the score function $\nabla \log \pi_\theta(a \mid s)$.

The following "regret lemma" is the key tool for analyzing NPG with function approximation.

Lemma 12 (NPG regret lemma)**.** *Fix a comparison policy $\widetilde{\pi}$ and a state distribution ρ. Assume $\log \pi_\theta(a \mid s)$ is β-smooth in θ. Consider the update rule $\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}$ where $\pi^{(0)}$ is the uniform distribution, $\|w^{(t)}\|_2 \le W$, and $\eta = \sqrt{2\log|\mathcal{A}|/(\beta W^2 T)}$. Define:*

$$\mathrm{err}_t = \mathbb{E}_{s\sim\widetilde{d}, a\sim\widetilde{\pi}(\cdot|s)}\Big[A^{(t)}(s,a) - w^{(t)} \cdot \nabla \log \pi^{(t)}(a \mid s)\Big].$$

Then:

$$\min_{t<T}\left\{V^{\widetilde{\pi}}(\rho) - V^{(t)}(\rho)\right\} \le \frac{1}{1-\gamma}\left(W\sqrt{\frac{2\beta\log|\mathcal{A}|}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\mathrm{err}_t\right).$$

Proof sketch. By β-smoothness: $\log\frac{\pi^{(t+1)}(a|s)}{\pi^{(t)}(a|s)} \ge \nabla \log \pi^{(t)}(a \mid s) \cdot \eta w^{(t)} - \frac{\beta}{2}\|\eta w^{(t)}\|_2^2$.

Using the performance difference lemma and the closed-form relationship between KL telescoping and the log-ratio:

$$\mathbb{E}_{s\sim\widetilde{d}}\Big(\mathrm{KL}(\widetilde{\pi}_s\|\pi_s^{(t)}) - \mathrm{KL}(\widetilde{\pi}_s\|\pi_s^{(t+1)})\Big)$$

$$\geq \eta \mathbb{E}_{s\sim \widetilde{d}, a\sim\widetilde{\pi}} \left[\nabla \log \pi^{(t)}(a \mid s) \cdot w^{(t)}\right] - \eta^2 \frac{\beta}{2}\|w^{(t)}\|_2^2$$
$$= (1-\gamma)\eta\Big(V^{\widetilde{\pi}}(\rho) - V^{(t)}(\rho)\Big) - \eta^2\frac{\beta}{2}W^2 - \eta\,\mathrm{err}_t.$$

Averaging over t, the KL terms telescope. Using $\mathrm{KL}(\widetilde{\pi}_s \| \pi_s^{(0)}) \leq \log|\mathcal{A}|$ and optimizing η gives the result. □

The regret lemma — key tool separating optimization from approximation:

$$\min_{t<T}\left\{V^{\widetilde{\pi}}(\rho) - V^{(t)}(\rho)\right\} \leq \frac{1}{1-\gamma}\left(W\sqrt{\frac{2\beta\log|\mathcal{A}|}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\mathrm{err}_t\right).$$

Performance Bound for Q-NPG

Assumption 1 (Approximation/estimation errors (Agarwal et al., 2021, Assumption 11.4)). *Let $w^{(0)}, w^{(1)}, \ldots, w^{(T-1)}$ be the iterates used by the Q-NPG algorithm. Suppose for all $t < T$:*

*(1) **Excess risk:** $L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_\star^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}}$.*

*(2) **Approximation error:** $L(w_\star^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{approx}}$.*

Here $L(w;\theta,v) := \mathbb{E}_{s,a\sim v}[(Q^{\pi_\theta}(s,a) - w\cdot\phi_{s,a})^2]$.

Assumption 2 (Relative condition number (Agarwal et al., 2021, Assumption 11.5)). *Fix a state distribution ρ and a comparator policy $\pi^\star$. Define $d^\star(s,a) = d_\rho^{\pi^\star}(s)\cdot \mathrm{Unif}_{\mathcal{A}}(a)$ and*

$$\kappa := \sup_{w\in\mathbb{R}^d} \frac{w^\top \Sigma_{d^\star} w}{w^\top \Sigma_\nu w},$$

where $\Sigma_v = \mathbb{E}_{s,a\sim v}[\phi_{s,a}\phi_{s,a}^\top]$, and assume $\kappa < \infty$.

Theorem 13 (Q-NPG convergence (Agarwal et al., 2021, Theorem 11.6)). *Fix a state distribution ρ, state-action distribution ν, and an arbitrary comparator $\pi^\star$. Suppose Assumptions 1 and 2 hold with $\|\phi_{s,a}\|_2 \leq B$. Starting with $\theta^{(0)} = 0$ and $\eta = \sqrt{2\log|\mathcal{A}|/(B^2W^2T)}$:*

$$\mathbb{E}\left[\min_{t<T}\left\{V^{\pi^\star}(\rho) - V^{(t)}(\rho)\right\}\right] \leq \frac{BW}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{\sqrt{4|\mathcal{A}|}}{(1-\gamma)^3}\left(\sqrt{\kappa\cdot\epsilon_{\text{stat}}} + \left\|\frac{d^\star}{\nu}\right\|_\infty \cdot \sqrt{\epsilon_{\text{approx}}}\right).$$

Proof sketch. We decompose err_t from the regret lemma into two terms:

$$\mathrm{err}_t = \mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)}\left[A^{(t)}(s,a) - w_\star^{(t)}\cdot\nabla\log\pi^{(t)}(a\mid s)\right]$$
$$+ \mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)}\left[(w_\star^{(t)} - w^{(t)})\cdot\nabla\log\pi^{(t)}(a\mid s)\right].$$

First term (approximation): Using $\nabla \log \pi_\theta(a \mid s) = \overline{\phi}^{\theta}_{s,a}$ for log-linear policies:

$$\mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star}\left[A^{(t)}(s,a) - w_\star^{(t)} \cdot \nabla \log \pi^{(t)}\right] \le 2\sqrt{|\mathcal{A}| \left\| \frac{d^\star}{\nu} \right\|_\infty L(w_\star^{(t)}; \theta^{(t)}, d^\star)}.$$

A distribution-shift argument then relates $L(w_\star^{(t)}; \theta^{(t)}, d^\star) \le \frac{1}{1-\gamma}\|d^\star/\nu\|_\infty \cdot \epsilon_{\text{approx}}$.

Second term (estimation): Similarly:

$$\mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star}\left[(w_\star^{(t)} - w^{(t)}) \cdot \nabla \log \pi^{(t)}\right] \le 2\sqrt{\frac{|\mathcal{A}|\kappa}{1-\gamma}\big(L(w^{(t)}) - L(w_\star^{(t)})\big)} \le 2\sqrt{\frac{|\mathcal{A}|\kappa}{1-\gamma}\epsilon_{\text{stat}}}.$$

Substituting into the regret lemma (Lemma 12) gives the result. □

Two striking features: (1) Estimation error enters via $\sqrt{\epsilon_{\text{stat}}}$ — a benign square-root dependence. With N samples, $\epsilon_{\text{stat}} = O(1/\sqrt{N})$, so the contribution is $O(N^{-1/4})$. (2) Approximation error is measured against the *fixed* comparator distribution $d^\star$, not all policies — a transfer learning interpretation. The relative condition number κ can be dimension-independent.

Summary

Result	Statement	Key Technique
Softmax PG (vanilla)	Global (asymptotic)	Exponentially slow in worst case
Softmax PG + log barrier	Global, $O(1/\sqrt{T})$	Gradient domination via regularization
NPG = Soft PI	$\pi^{(t+1)} \propto \pi^{(t)} \exp(\eta Q^{(t)})$	Fisher cancels d^π
NPG convergence	$V^\star - V^{(T)} \le O(\log|\mathcal{A}|/T)$	KL telescoping, dimension-free
Q-NPG with FA	$\sqrt{1/T} + \sqrt{\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{approx}}}$	Regret lemma + compatible FA

Key takeaways.

- The reparameterization problem motivates measuring distances in policy space (KL divergence) rather than parameter space, leading to the natural gradient.
- Vanilla softmax PG converges globally under $\mu > 0$ (Theorem 5), but potentially at an exponentially slow rate. Log-barrier regularization achieves polynomial rates but depends on $|\mathcal{S}|$, $|\mathcal{A}|$.
- NPG achieves *dimension-free* global convergence by using Fisher information geometry — the rate depends only on $\log |\mathcal{A}|$, not $|\mathcal{S}|$.
- NPG with function approximation: compatible FA gives a practical recipe; the regret lemma cleanly separates optimization from approximation error.
- TRPO (Lecture 9) $\approx$ NPG with adaptive step size determined by trust region radius.

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. Self-published, 2021. Available at https://rltheorybook.github.io/.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998.

J Andrew Bagnell and Jeff Schneider. Covariant policy search. In *International Joint Conference on Artificial Intelligence*, pages 1019–1024, 2003.

Amir Beck. *First-Order Methods in Optimization*. SIAM, 2017.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Sham Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pages 1057–1063, 1999.