Policy Gradient - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

Policy gradient (PG) methods optimize a parameterized stochastic policy by *directly* ascending the expected return. Starting from the trajectory distribution induced by a Markov Decision Process (MDP) and a differentiable policy, we derive the *log-likelihood trick*, the *Policy Gradient Theorem*, unbiased Monte Carlo estimators (REINFORCE), variance-reduction via baselines and advantages, and actor–critic updates with temporal-difference bootstrapping. We also outline entropy regularization, natural gradients, and proximal surrogates, while declaring all variables and their dimensions.

1.2 Data and Notation

We consider a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$, where $\mathcal{S} \subseteq \mathbb{R}^{n_{\mathbf{s}}}$ is the state space. Each state $\mathbf{s} \in \mathbb{R}^{n_{\mathbf{s}}}$ is a column vector of dimension $n_{\mathbf{s}} \times 1$. $\mathcal{A} \subseteq \mathbb{R}^{n_{\mathbf{a}}}$ is the action space. Each action $\mathbf{a} \in \mathbb{R}^{n_{\mathbf{a}}}$ is a column vector of dimension $n_{\mathbf{a}} \times 1$. (For discrete actions, $\mathcal{A}$ is a finite set.) $p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$ is the state transition probability. $r(\mathbf{s}_t, \mathbf{a}_t) \in \mathbb{R}$ is the reward function. $\gamma \in [0, 1)$ is the discount factor (a scalar). A trajectory (or episode) is a sequence $\boldsymbol{\tau} = \{\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T\}$, with horizon $T \in \mathbb{N} \cup \{\infty\}$.

A stochastic policy $\pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s})$ is differentiable in $\boldsymbol{\theta} \in \mathbb{R}^p$ (column vector; p parameters). We denote (discounted) returns and value functions:

$$\begin{aligned} G_t &\triangleq \sum_{k=0}^{\infty} \gamma^k r(\mathbf{s}_{t+k}, \mathbf{a}_{t+k}), \\ V^\pi(\mathbf{s}) &\triangleq \mathbb{E}_\pi[G_0 \mid \mathbf{s}_0 = \mathbf{s}], \qquad Q^\pi(\mathbf{s}, \mathbf{a}) \triangleq \mathbb{E}_\pi[G_0 \mid \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a}], \\ A^\pi(\mathbf{s}, \mathbf{a}) &\triangleq Q^\pi(s, a) - V^\pi(\mathbf{s}). \end{aligned}$$

Dimensions: $\boldsymbol{\theta} \in \mathbb{R}^p$; $J(\boldsymbol{\theta}) \in \mathbb{R}$; $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \in \mathbb{R}^p$.

1.3 Policy Parameterization

We consider a parameterized policy $\pi(\mathbf{a} \mid \mathbf{s}; \boldsymbol{\theta})$, which is a probability density function (or mass function) over actions given state $\mathbf{s}$. Here: $\boldsymbol{\theta} \in \mathbb{R}^p$ is the parameter vector of the policy. For each state $\mathbf{s} \in \mathbb{R}^{n_s}$, $\pi(\mathbf{a} \mid \mathbf{s}; \boldsymbol{\theta}) \geq 0$ for all $\mathbf{a} \in \mathcal{A}$ and

$$\int_{\mathcal{A}} \pi(\mathbf{a} \mid \mathbf{s}; \boldsymbol{\theta})\, d\mathbf{a} = 1,$$

if $\mathcal{A}$ is continuous (or a sum to 1 if $\mathcal{A}$ is discrete).

1.4 Model Formulation: Trajectory Distribution and Objective

Let $\rho_0(\mathbf{s})$ be the initial-state distribution. Under $\pi_{\boldsymbol{\theta}}$, the trajectory density is

$$p_{\boldsymbol{\theta}}(\tau) = \rho_0(\mathbf{s}_0) \prod_{t=0}^{T-1} \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t)\, P(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t).$$

We optimize the expected return

$$J(\boldsymbol{\theta}) \;=\; \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}}\big[\, R(\tau) \,\big], \qquad R(\tau) = \sum_{t=0}^{T-1} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t). \tag{1}$$

Episodic: $T < \infty$; **Continuing:** either infinite horizon with $\gamma < 1$ or average reward.

1.5 Log-Likelihood Trick and the Basic Policy Gradient

Differentiate Eqn. (1) using the identity $\nabla_{\boldsymbol{\theta}} p = p\, \nabla_{\boldsymbol{\theta}} \log p$, we obtain

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \int R(\tau)\, \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau)\, d\tau = \int R(\tau)\, p_{\boldsymbol{\theta}}(\tau)\, \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\, d\tau \\ &= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}}\big[\, R(\tau)\, \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) \,\big]. \end{aligned} \tag{2}$$

Because environment dynamics P and ρ_0 do not depend on $\boldsymbol{\theta}$,

$$\log p_{\boldsymbol{\theta}}(\tau) = \log \rho_0(\mathbf{s}_0) + \sum_{t=0}^{T-1} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t) + \sum_{t=0}^{T-1} \log P(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t), \tag{3}$$

so $\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) = \sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t)$. Hence

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}\left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t) \right]. \tag{4}$$

Replacing $R(\tau)$ by the time-dependent return G_t yields an equivalent and final gradient estimator with lower variance:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{t=0}^{T-1} G_t \, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t)\right]. \tag{5}$$

1.6 Policy Gradient Theorem (State-Visitation Form)

Define the (discounted) state-visitation distribution

$$d^{\pi}(\mathbf{s}) \;\triangleq\; (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \, \Pr(\mathbf{s}_t = \mathbf{s} \mid \pi). \tag{6}$$

Theorem. For differentiable $\pi_{\boldsymbol{\theta}}$,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{1-\gamma}\, \mathbb{E}_{\substack{\mathbf{s}\sim d^{\pi}, \\ \mathbf{a}\sim\pi_{\boldsymbol{\theta}}}}\big[\, Q^{\pi}(\mathbf{s},\mathbf{a})\, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) \,\big]. \tag{7}$$

Proof (sketch). Start from Eqn. (5) and condition on $\mathbf{s}_t, \mathbf{a}_t$: $G_t = \mathbb{E}[G_t \mid \mathbf{s}_t, \mathbf{a}_t] = Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$. Unroll the expectation over time and normalize time weights into d^{π} to obtain Eqn. (7). ■

1.7 Baselines and Advantages (Variance Reduction)

Let $b : \mathcal{S} \to \mathbb{R}$ be any function independent of $\mathbf{a}$ given $\mathbf{s}$. Then

$$\mathbb{E}_{\mathbf{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|\mathbf{s})}\big[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s})\, b(\mathbf{s})\big] = b(\mathbf{s})\, \nabla_{\boldsymbol{\theta}} \sum_{\mathbf{a}} \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) = b(\mathbf{s})\, \nabla_{\boldsymbol{\theta}} 1 = \mathbf{0}, \tag{8}$$

so subtracting $b(\mathbf{s})$ leaves the gradient unbiased:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{1-\gamma}\, \mathbb{E}_{\mathbf{s}\sim d^{\pi},\, \mathbf{a}\sim\pi_{\boldsymbol{\theta}}}\big[\, \big(Q^{\pi}(\mathbf{s},\mathbf{a}) - b(\mathbf{s})\big)\, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) \,\big]. \tag{9}$$

Choosing $b(\mathbf{s}) = V^{\pi}(\mathbf{s})$ yields the *advantage* form with $A^{\pi}(\mathbf{s},\mathbf{a}) = Q^{\pi}(\mathbf{s},\mathbf{a}) - V^{\pi}(\mathbf{s})$:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{1-\gamma}\, \mathbb{E}\big[\, A^{\pi}(\mathbf{s},\mathbf{a})\, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) \,\big]. \tag{10}$$

1.8 Monte Carlo Estimators (REINFORCE)

From Eqn. (5), an unbiased sample-gradient from one episode is

$$\widehat{g} \;=\; \sum_{t=0}^{T-1} \Big(G_t - b(\mathbf{s}_t)\Big)\, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t). \tag{11}$$

With a mini-batch of N episodes, average $\widehat{g}$, then update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\, \widehat{g}$ (*ascent*) with step size $\alpha > 0$.

1.9 Actor–Critic: Bootstrapped Advantages and TD Error

Introduce a differentiable critic $V_{\boldsymbol{\phi}}(\mathbf{s}) \approx V^{\pi}(\mathbf{s})$ with parameters $\boldsymbol{\phi} \in \mathbb{R}^q$. Define one-step TD error

$$\delta_t \;\triangleq\; r_t + \gamma V_{\boldsymbol{\phi}}(\mathbf{s}_{t+1}) - V_{\boldsymbol{\phi}}(\mathbf{s}_t), \tag{12}$$

a biased but lower-variance estimator of $A^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$. The actor update is

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \sum_{t} \delta_t\, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t), \tag{13}$$

and the critic minimizes the squared TD error or the one-step Bellman residual, e.g., by stochastic gradient descent on $\frac{1}{2}\big(r_t + \gamma V_{\boldsymbol{\phi}}(\mathbf{s}_{t+1}) - V_{\boldsymbol{\phi}}(\mathbf{s}_t)\big)^2$.

1.10 Generalized Advantage Estimation (GAE)

For $\lambda \in [0,1]$, the λ-return advantage is

$$\widehat{A}_t^{\lambda} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \, \delta_{t+l}, \qquad \delta_t = r_t + \gamma V_{\boldsymbol{\phi}}(\mathbf{s}_{t+1}) - V_{\boldsymbol{\phi}}(\mathbf{s}_t). \tag{14}$$

Setting $\lambda = 1$ recovers Monte Carlo advantages; $\lambda = 0$ uses one-step TD.

1.11 Entropy Regularization

Encourage exploration with entropy bonus $\mathcal{H}(\pi_{\boldsymbol{\theta}}(\cdot \mid \mathbf{s}))$:

$$J_\beta(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \beta \, \mathbb{E}_{s\sim d^\pi}\big[\mathcal{H}(\pi_{\boldsymbol{\theta}}(\cdot \mid \mathbf{s}))\big], \quad \nabla_{\boldsymbol{\theta}} J_\beta = \nabla_{\boldsymbol{\theta}} J + \beta \, \mathbb{E}\big[\nabla_{\boldsymbol{\theta}} \mathcal{H}\big]. \tag{15}$$

1.12 Natural Policy Gradient (NPG) and Trust Regions

The steepest-ascent direction under the Fisher–Rao metric uses the policy Fisher information matrix

$$F(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}\sim d^\pi,\, \mathbf{a}\sim\pi_{\boldsymbol{\theta}}}\Big[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) \, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s})^\top\Big]. \tag{16}$$

Natural gradient takes $\Delta\boldsymbol{\theta} = F(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. Equivalently, solve

$$\max_{\Delta\boldsymbol{\theta}} \; g^\top \Delta\boldsymbol{\theta} \quad \text{s.t.} \quad \Delta\boldsymbol{\theta}^\top F(\boldsymbol{\theta}) \, \Delta\boldsymbol{\theta} \le \epsilon, \tag{17}$$

with $g = \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$; the optimizer is $\Delta\boldsymbol{\theta} = \sqrt{\epsilon}\, F^{-1} g / \sqrt{g^\top F^{-1} g}$. Trust-region PG (TRPO) enforces a small expected KL divergence between old/new policies; PPO uses a clipped surrogate as a first-order approximation.

1.13 Continuous Actions: Gaussian Policies

For $\mathcal{A} = \mathbb{R}^m$ with $\pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) = \mathcal{N}\big(\mathbf{a} \mid \mu_{\boldsymbol{\theta}}(\mathbf{s}), \Sigma_{\boldsymbol{\theta}}(\mathbf{s})\big)$,

$$\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) = \big(\nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(\mathbf{s})\big)^\top \Sigma_{\boldsymbol{\theta}}(\mathbf{s})^{-1} \big(\mathbf{a} - \mu_{\boldsymbol{\theta}}(\mathbf{s})\big) + \frac{1}{2} \nabla_{\boldsymbol{\theta}} \Big[\log |\Sigma_{\boldsymbol{\theta}}(\mathbf{s})^{-1}| - (\mathbf{a} - \mu)^\top \Sigma^{-1} (\mathbf{a} - \mu)\Big], \tag{18}$$

substituted into Eqn. (10). Often $\Sigma_{\boldsymbol{\theta}}(\mathbf{s})$ is diagonal or state-independent.

1.14 Off-Policy Correction (Importance Sampling)

With data from a behavior policy $\mu(\mathbf{a} \mid \mathbf{s})$,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{1-\gamma} \, \mathbb{E}_{\mathbf{s}\sim d^\mu,\, \mathbf{a}\sim\mu}\left[\underbrace{\frac{\pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s})}{\mu(\mathbf{a} \mid \mathbf{s})}}_{\text{IS ratio}} \, Q^\pi(\mathbf{s}, \mathbf{a}) \, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s})\right], \tag{19}$$

possibly using per-decision ratios and variance-control (e.g., truncation).

1.15 Algorithm (REINFORCE / Actor–Critic)

1. **Input:** step sizes $\alpha_{\text{act}}, \alpha_{\text{crt}}$, discount γ, (optional) entropy weight β, GAE parameter λ.
2. **Initialize** $\boldsymbol{\theta}$, critic parameters $\boldsymbol{\phi}$.
3. **Repeat** for iterations:
   (a) Roll out trajectories (or n-step fragments) using $\pi_{\boldsymbol{\theta}}$.
   (b) Compute returns G_t or advantages $\widehat{A}_t$ (e.g., GAE Eqn. (14)).
   (c) **Actor step:** $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\text{act}} \sum_t \big(\widehat{A}_t \, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t) + \beta \, \nabla_{\boldsymbol{\theta}} \mathcal{H}(\pi_{\boldsymbol{\theta}}(\cdot \mid \mathbf{s}_t))\big)$.
   (d) **Critic step:** update $\boldsymbol{\phi}$ by minimizing $\sum_t \delta_t^2$ or a value regression loss.

1.16 Summary of Variables and Their Dimensions

- $\boldsymbol{\theta} \in \mathbb{R}^p$: policy parameters (vector).
- $J(\boldsymbol{\theta}) \in \mathbb{R}$: expected return (scalar).
- $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \in \mathbb{R}^p$: policy gradient.
- $V^\pi : \mathcal{S} \to \mathbb{R}$, $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.
- $d^\pi : \mathcal{S} \to \mathbb{R}_{\geq 0}$: discounted state visitation distribution.
- $\boldsymbol{\phi} \in \mathbb{R}^q$: critic parameters.
- $\delta_t \in \mathbb{R}$: TD error Eqn. (12); $\widehat{A}_t \in \mathbb{R}$: advantage estimate.

1.17 Summary

Starting from the trajectory distribution and the score-function identity, we derived the policy gradient in several equivalent forms, culminating in the Policy Gradient Theorem Eqn. (7). Subtracting any state-dependent baseline (especially V^π) preserves unbiasedness Eqn. (8) and yields advantage-weighted updates Eqn. (10). Actor–critic algorithms replace Monte Carlo returns by bootstrapped TD errors to reduce variance, with GAE trading bias and variance. Natural gradients and trust regions respect the information geometry of policies, and entropy regularization encourages exploration. These ingredients constitute a principled, scalable framework for optimizing stochastic policies from first principles.