

Proximal Policy Optimization - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs	1
1.1 Introduction	1
1.2 Data and Notation	1
1.3 Policy Parameterization	2
1.4 Objective and the Policy Gradient Theorem	2
1.5 Surrogate with Importance Sampling	2
1.6 From Trust Regions to Proximal Updates	3
1.7 Advantage Estimation: $\text{GAE}(\gamma, \lambda)$	3
1.8 Value Learning and Entropy Regularization	3
1.9 Total Optimization Objective	4
1.10 Early Stopping via KL (Optional)	4
1.11 Algorithm (One PPO Iteration)	4
1.12 Proof Notes: Unbiasedness and Proximality	4
1.13 Summary of Variables and Their Dimensions	4
1.14 Summary	5

1 Mathematical Derivations & Proofs

1.1 Introduction

Proximal Policy Optimization (PPO) is a first-order policy-gradient method that performs multiple stochastic-gradient updates on batches of trajectories while *constraining* the policy update to remain *proximal* to a behavior policy. It is derived by (i) the Policy Gradient Theorem, (ii) an importance sampling (IS) surrogate objective to reuse on-policy data for several epochs, and (iii) a tractable trust-region surrogate that bounds the policy change, implemented either by a KL-penalty or the *clipped* objective. We derive PPO starting from the Markov Decision Process (MDP) control objective, establish the policy-gradient identity, introduce advantage baselines and generalized advantage estimation (GAE), prove properties of the clipped surrogate, and present the full learning objective including value learning and entropy regularization.

1.2 Data and Notation

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ be a discounted MDP with $\gamma \in [0, 1)$ is the discount factor (a scalar). A stochastic policy $\pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s})$ is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^p$; the state-value baseline is $V_{\boldsymbol{\phi}}(\mathbf{s})$ with $\boldsymbol{\phi} \in \mathbb{R}^q$. $\mathcal{S} \subseteq \mathbb{R}^{n_{\mathbf{s}}}$ 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} is the action space. In the continuous case, 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. The *discounted state visitation* measure under π from some initial distribution ρ_0 is

$$\rho^{\pi}(\mathbf{s}) \triangleq (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(\mathbf{s}_t = \mathbf{s} \mid \pi).$$

For a trajectory $\tau = (\mathbf{s}_0, \mathbf{a}_0, r_0, \mathbf{s}_1, \dots)$, define returns $G_t = \sum_{l=0}^{\infty} \gamma^l r_{t+l}$, the action-value $Q^\pi(\mathbf{s}, \mathbf{a}) = \mathbb{E}[G_0 \mid \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a}, \pi]$, and advantage $A^\pi(\mathbf{s}, \mathbf{a}) = Q^\pi(\mathbf{s}, \mathbf{a}) - V^\pi(\mathbf{s})$.

Dimensions: $\boldsymbol{\theta} \in \mathbb{R}^p$ (column), $\boldsymbol{\phi} \in \mathbb{R}^q$, $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) \in \mathbb{R}^p$.

1.3 Policy Parameterization

We consider a parameterized stochastic policy:

$$\pi(\mathbf{a} \mid \mathbf{s}; \boldsymbol{\theta}),$$

which is a probability density 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 Objective and the Policy Gradient Theorem

The control objective is to maximize the expected cumulative discounted reward:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}_0 \sim \rho_0} \left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right], \quad \mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(\cdot \mid \mathbf{s}_t), \quad \mathbf{s}_{t+1} \sim P(\cdot \mid \mathbf{s}_t, \mathbf{a}_t). \quad (1)$$

Directly optimizing $J(\boldsymbol{\theta})$ is challenging, so PPO maximizes a surrogate objective that uses importance sampling to compare the new policy $\pi(\mathbf{a} \mid \mathbf{s}; \boldsymbol{\theta})$ to the old policy $\pi(\mathbf{a} \mid \mathbf{s}; \boldsymbol{\theta}_{\text{old}})$.

Policy Gradient Theorem. Assuming regularity (dominated convergence, differentiability),

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{1-\gamma} \mathbb{E}_{\mathbf{s} \sim \rho^{\pi_{\boldsymbol{\theta}}}, \mathbf{a} \sim \pi_{\boldsymbol{\theta}}(\cdot \mid \mathbf{s})} [\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) Q^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})]. \quad (2)$$

Moreover, for any baseline $b : \mathcal{S} \rightarrow \mathbb{R}$,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{1-\gamma} \mathbb{E}_{\mathbf{s}, \mathbf{a}} [\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{s}) (Q^{\pi}(\mathbf{s}, \mathbf{a}) - b(\mathbf{s}))], \quad (3)$$

and choosing $b(\mathbf{s}) = V^{\pi}(\mathbf{s})$ gives the *advantage form* with $A^{\pi}(\mathbf{s}, \mathbf{a})$.

Proof. Express $J(\boldsymbol{\theta}) = (1-\gamma)^{-1} \mathbb{E}_{\mathbf{s} \sim \rho^{\pi}} [\mathbb{E}_{\mathbf{a} \sim \pi} [r(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}' \sim P} V^{\pi}(\mathbf{s}')]]$, take $\nabla_{\boldsymbol{\theta}}$ and apply the log-derivative trick $\nabla_{\boldsymbol{\theta}} \pi = \pi \nabla_{\boldsymbol{\theta}} \log \pi$. A telescoping cancellation of terms involving $\nabla_{\boldsymbol{\theta}} \rho^{\pi}$ yields Eqn. (2). The baseline term integrates to zero since $\mathbb{E}_{\mathbf{a} \sim \pi} [\nabla_{\boldsymbol{\theta}} \log \pi(\mathbf{a} \mid \mathbf{s})] = \nabla_{\boldsymbol{\theta}} \int \pi(\mathbf{a} \mid \mathbf{s}) d\mathbf{a} = 0$, proving Eqn. (3). ■

1.5 Surrogate with Importance Sampling

In practice we optimize a *surrogate* using data generated by a reference (policy) $\pi_{\boldsymbol{\theta}_{\text{old}}}$. Define the importance ratio

$$r_t(\boldsymbol{\theta}) \triangleq \frac{\pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{a}_t \mid \mathbf{s}_t)} = \exp \left(\log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t) - \log \pi_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{a}_t \mid \mathbf{s}_t) \right). \quad (4)$$

Using Eqn. (3) and standard IS, an unbiased surrogate for a small step around $\boldsymbol{\theta}_{\text{old}}$ is

$$L^{\text{CPI}}(\boldsymbol{\theta}) \triangleq \mathbb{E}_t [r_t(\boldsymbol{\theta}) \hat{A}_t], \quad (5)$$

where \hat{A}_t is an estimator of $A^{\pi_{\boldsymbol{\theta}_{\text{old}}}}(\mathbf{s}_t, \mathbf{a}_t)$ and \mathbb{E}_t denotes the empirical average over samples collected by $\pi_{\boldsymbol{\theta}_{\text{old}}}$ in the current batch. The gradient $\nabla_{\boldsymbol{\theta}} L^{\text{CPI}}(\boldsymbol{\theta})$ equals the policy gradient at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$.

1.6 From Trust Regions to Proximal Updates

TRPO bound (motivation). A monotonic improvement surrogate maximizes L^{CPI} subject to a trust region on the average KL divergence:

$$\max_{\boldsymbol{\theta}} L^{\text{CPI}}(\boldsymbol{\theta}) \quad \text{s.t.} \quad \mathbb{E}_t[\text{KL}(\pi_{\boldsymbol{\theta}_{\text{old}}}(\cdot | \mathbf{s}_t) \parallel \pi_{\boldsymbol{\theta}}(\cdot | \mathbf{s}_t))] \leq \delta, \quad (6)$$

which yields (via a quadratic approximation and conjugate gradient) the natural gradient step. PPO replaces the hard constraint by a *first-order* proxy either a KL-penalized objective or the *clipped* surrogate.

Clipped surrogate (PPO-Clip). Let $\epsilon \in (0, 1)$ and define the elementwise clipping $\text{clip}(r, 1 - \epsilon, 1 + \epsilon) = \min\{\max\{r, 1 - \epsilon\}, 1 + \epsilon\}$. The PPO clipped objective is

$$L^{\text{CLIP}}(\boldsymbol{\theta}) = \mathbb{E}_t \left[\min \left(r_t(\boldsymbol{\theta}) \hat{A}_t, \text{clip}(r_t(\boldsymbol{\theta}), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right]. \quad (7)$$

The intuition behind the clipping is to prevent the new policy from deviating too much from the old policy, thereby ensuring a conservative update that improves stability. The gradient of this objective with respect to $\boldsymbol{\theta}$ is then used for updating the policy parameters via stochastic gradient ascent.

Lower-bound property. For each sample t :

$$\hat{A}_t > 0 \Rightarrow \min(r_t \hat{A}_t, \text{clip}(r_t) \hat{A}_t) = \text{clip}(r_t) \hat{A}_t \leq r_t \hat{A}_t,$$

$$\hat{A}_t < 0 \Rightarrow \min(r_t \hat{A}_t, \text{clip}(r_t) \hat{A}_t) = \max(r_t, \text{clip}(r_t)) \hat{A}_t \leq r_t \hat{A}_t.$$

Hence pointwise $L^{\text{CLIP}}(\boldsymbol{\theta}) \leq L^{\text{CPI}}(\boldsymbol{\theta})$, and L^{CLIP} equals L^{CPI} while $r_t \in [1 - \epsilon, 1 + \epsilon]$. Maximizing Eqn. (7) thus discourages updates that move r_t outside the trust region, implementing a first-order trust-region step.

KL-penalty variant (PPO-KL). Alternatively,

$$L^{\text{KL PEN}}(\boldsymbol{\theta}) = \mathbb{E}_t \left[r_t(\boldsymbol{\theta}) \hat{A}_t - \beta \text{KL}(\pi_{\boldsymbol{\theta}_{\text{old}}}(\cdot | \mathbf{s}_t) \parallel \pi_{\boldsymbol{\theta}}(\cdot | \mathbf{s}_t)) \right], \quad (8)$$

with an adaptive penalty β adjusted to keep the measured KL near a target δ . In practice, PPO-Clip is more robust and simpler.

1.7 Advantage Estimation: GAE(γ, λ)

For temporal-difference residuals

$$\delta_t \triangleq r_t + \gamma V_{\phi}(\mathbf{s}_{t+1}) - V_{\phi}(\mathbf{s}_t), \quad (9)$$

the *generalized advantage estimator* is

$$\hat{A}_t^{\text{GAE}} = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l} \approx \sum_{l=0}^{T-t-1} (\gamma \lambda)^l \delta_{t+l}, \quad (10)$$

truncated to episode (or bootstrapped at cutoff). When $\lambda = 1$, this reduces to Monte Carlo advantages $G_t - V(\mathbf{s}_t)$; when $\lambda = 0$, it is one-step TD. The parameter $\lambda \in [0, 1]$ trades variance for bias.

1.8 Value Learning and Entropy Regularization

Alongside the policy, PPO fits V_{ϕ} by regression to return targets \hat{R}_t (e.g., G_t or $\sum_{l=0}^{L-1} \gamma^l r_{t+l} + \gamma^L V_{\phi}(\mathbf{s}_{t+L})$):

$$L^V(\phi) = \mathbb{E}_t \left[(V_{\phi}(\mathbf{s}_t) - \hat{R}_t)^2 \right]. \quad (11)$$

To encourage exploration and prevent premature collapse of π , include an entropy bonus

$$L^{\text{ENT}}(\boldsymbol{\theta}) = \mathbb{E}_t [\mathcal{H}(\pi_{\boldsymbol{\theta}}(\cdot | \mathbf{s}_t))] , \quad \mathcal{H}(p) = - \sum_{\mathbf{a}} p(\mathbf{a}) \log p(\mathbf{a}) \text{ (discrete) or } \mathcal{H} \text{ of the density (continuous).} \quad (12)$$

1.9 Total Optimization Objective

PPO maximizes the clipped surrogate while fitting the value function and adding entropy regularization. A common *minimization* loss is

$$\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\phi}) = -\mathbb{E}_t[L^{\text{CLIP}}(\boldsymbol{\theta})] + c_v \mathbb{E}_t[(V_\phi(\mathbf{s}_t) - \hat{R}_t)^2] - c_{\text{ent}} \mathbb{E}_t[\mathcal{H}(\pi_{\boldsymbol{\theta}}(\cdot | \mathbf{s}_t))], \quad (13)$$

with coefficients $c_v, c_{\text{ent}} > 0$. Gradients are estimated by minibatch SGD/Adam over several epochs on the same batch of trajectories collected by $\pi_{\boldsymbol{\theta}_{\text{old}}}$.

1.10 Early Stopping via KL (Optional)

Even with clipping, one may monitor the empirical $\overline{\text{KL}} = \mathbb{E}_t[\text{KL}(\pi_{\boldsymbol{\theta}_{\text{old}}} || \pi_{\boldsymbol{\theta}})]$ and stop epochs early if $\overline{\text{KL}} > \kappa$; alternatively adapt ϵ or the learning rate to keep updates proximal.

1.11 Algorithm (One PPO Iteration)

1. **Rollout.** Using $\pi_{\boldsymbol{\theta}_{\text{old}}}$, collect N transitions (or T -step segments) $\{(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})\}$; record $\log \pi_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{a}_t | \mathbf{s}_t)$.
2. **Compute targets.**
 - Evaluate $V_\phi(\mathbf{s}_t)$ along trajectories.
 - Compute δ_t by Eqn. (9); compute \hat{A}_t by Eqn. (10); normalize \hat{A}_t (zero mean, unit std).
 - Set returns $\hat{R}_t = \hat{A}_t + V_\phi(\mathbf{s}_t)$ (or MC returns).
3. **Optimization.** For K epochs, shuffle data and iterate minibatches \mathcal{B} :
 - (a) Compute ratios $r_t(\boldsymbol{\theta})$ by Eqn. (4).
 - (b) Evaluate $L^{\text{CLIP}}(\boldsymbol{\theta})$ in Eqn. (7), the value loss Eqn. (11), and entropy Eqn. (12).
 - (c) Take an optimizer step on $\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\phi})$ in Eqn. (13).
 - (d) (Optional) If empirical $\text{KL} > \kappa$, early-stop remaining epochs.
4. **Policy update.** Set $\boldsymbol{\theta}_{\text{old}} \leftarrow \boldsymbol{\theta}$; repeat.

1.12 Proof Notes: Unbiasedness and Proximality

Unbiased gradient at the trust boundary. At $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$, $r_t = 1$ and L^{CLIP} reduces to $\mathbb{E}_t[\hat{A}_t]$, with gradient equal to the on-policy policy gradient (by Eqn. (3) and $\mathbb{E}_t[\hat{A}_t] = 0$ when \hat{A} is computed w.r.t. $V^{\pi_{\boldsymbol{\theta}_{\text{old}}}}$).

Proximal lower bound. Because $L^{\text{CLIP}} \leq L^{\text{CPI}}$ pointwise and equals it inside the clip interval, maximizing L^{CLIP} is a conservative ascent on L^{CPI} that naturally limits $|r_t - 1|$; this mimics the TRPO trust region Eqn. (6) without second-order computation.

1.13 Summary of Variables and Their Dimensions

- $\boldsymbol{\theta} \in \mathbb{R}^p$: policy parameters; $\pi_{\boldsymbol{\theta}}(\mathbf{a} | \mathbf{s})$ density or probability mass.
- $\boldsymbol{\phi} \in \mathbb{R}^q$: value-function parameters; $V_\phi : \mathcal{S} \rightarrow \mathbb{R}$.
- $r_t(\boldsymbol{\theta}) \in \mathbb{R}_+$: IS ratio (scalar).

- $\hat{A}_t \in \mathbb{R}$: advantage estimator (scalar), typically $\text{GAE}(\gamma, \lambda)$.
- $\delta_t \in \mathbb{R}$: TD residual Eqn. (9); $\hat{R}_t \in \mathbb{R}$: return target.
- $\epsilon \in (0, 1)$: clip parameter; $\beta > 0$: KL-penalty weight (if used).
- $c_v, c_{\text{ent}} > 0$: coefficients for value loss and entropy.
- Batch size N , epochs K , minibatch size M (positive integers).

1.14 Summary

Starting from the policy-gradient objective Eqn. (2), PPO constructs the IS surrogate Eqn. (5) to reuse on-policy data, then enforces proximity by a first-order trust region via the clipped objective Eqn. (7) (or the KL-penalized Eqn. (8)). Advantages are estimated by GAE Eqn. (10); the value function Eqn. (11) and entropy Eqn. (12) complete the loss Eqn. (13). The resulting procedure performs robust, sample-efficient policy improvement with only first-order optimization and minimal tuning.