Standard and Natural Policy Gradients for Discounted Rewards

Aaron Mishkin
August 8, 2020

UBC MLRG 2018W1
Motivating Example: Humanoid Robot Control

Consider learning a control model for a robotic arm that plays table tennis.

https://static.independent.co.uk/s3fs-public/thumbnails/image/2014/03/11/15/ping-pongv2.jpg?w968
Policy gradients have several advantages:

- Policy gradients permit explicit policies with complex parameterizations.
- Such policies are easily defined for continuous state and action spaces.
- Policy gradient approaches are guaranteed to converge under standard assumptions while greedy methods (SARSA, Q-learning, etc) are not.
Roadmap

Background and Notation

The Policy Gradient Theorem

Natural Policy Gradients
Background and Notation
Markov Decision Processes (MDPs)

A discrete-time MDP is specified by the tuple \( \{S, A, d_0, f, r\} \):

- States are \( s \in S \); actions are \( a \in A \).
- \( f \) is the transition distribution. It satisfies the Markov property:
  \[
  f(s_t, a_t, s_{t+1}) = p(s_{t+1}|s_0, a_0...s_t, a_t) = p(s_{t+1}|s_t, a_t)
  \]
- \( d_0(s_0) \) is the initial distribution over states.
- \( r(s_t, a_t, s_{t+1}) \) is the reward function, which may be deterministic or stochastic.
- Trajectories are sequences of state-action pairs:
  \[
  \tau_{0:t} = \{(s_0, a_0), ..., (s_t, a_t)\}
  \]

We treat states \( s \) as fully observable.
Continuous State and Action Spaces

We will consider MDPs with continuous state and action spaces. In the robot control example:

- \( s \in S \) is a real vector describing the configuration of the robotic arm’s movement system and the state of environment.
- \( a \in A \) real vector representing a motor command to the arm.
- Given action \( a \) in state \( s \), the probability of being in a region of state space \( S' \subseteq S \) is:

\[
P(s' \in S'|s, a) = \int_{S'} p(s'|s, a) ds'
\]

Future states \( s' \) are only known probabilistically because our control and physical models are approximations.
Policies defines how an agent acts in the MDP:

- A policy $\pi : S \times A \rightarrow [0, \infty)$ is the conditional density function:
  
  $$\pi(a|s) := \text{probability of taking action } a \text{ in state } s$$

- The policy is deterministic when $\pi(a|s)$ is a Dirac-delta function.
- Actions are chosen by sampling from the policy $a \sim \pi(a|s)$.
- The quality of a policy is given by an objective function $J(\pi)$. 
Bellman Equations

We consider discounted returns with factor $\gamma \in [0, 1]$. The Bellman equations describe the quality of a policy recursively:

$$Q^\pi(s, a) := \int_S f(s'|s, a) \left( r(s, a, s') + \int_A \pi(a'|s') \gamma Q^\pi(s', a') da' \right) ds'$$

$$V^\pi(s) := \int_A \pi(a|s) Q^\pi(s, a) da$$

$$= \int_A \pi(a|s) \int_S f(s'|s, a) (r(s, a, s') + \gamma V^\pi(s')) ds'da$$

$$= \int_A \pi(a|s) \int_S f(s'|s, a) r(s, a, s') ds'da$$

$$+ \int_A \pi(a|s) \int_S f(s'|s, a) \gamma V^\pi(s') ds'da$$
Three major flavors of reinforcement learning:

1. Critic-only methods: Learn an approximation of the state-action reward function: \( R(s, a) \approx Q^\pi(s, a) \).

2. Actor-only methods: Learn the policy \( \pi \) directly from observed rewards. A parametric policy \( \pi_\theta \) can be optimized by descending the policy gradient:

\[
\nabla_\theta J(\pi_\theta) = \frac{\partial J(\pi_\theta)}{\partial \pi_\theta} \frac{\partial \pi_\theta}{\partial \theta}
\]

3. Actor-Critic methods: Learn an approximation of the reward \( R(s, a) \) jointly with the policy \( \pi(a|s) \).
Value of a Policy

We can use the Bellman equations to write the overall quality of the policy:

\[
\frac{J(\pi)}{1 - \gamma} = \int_S d_0(s_0) V_\pi(s_0) ds_0
\]

\[
= \sum_{k=0}^{\infty} \int_S p(s_k = \bar{s}) \int_A \pi(a_k|\bar{s}) \int_S f(s_{k+1}|\bar{s}a_k) \gamma^k r(\bar{s}, a_k, s_{k+1}) ds_{t+1} da d\bar{s}
\]

\[
= \int_S \sum_{k=0}^{\infty} \gamma^k p(s_k = \bar{s}) \int_A \pi(a_k|\bar{s}) \int_S f(s_{k+1}|\bar{s}a_k) r(\bar{s}, a_k, s_{k+1}) ds_{t+1} da d\bar{s}
\]

Define the ”discounted state” distribution:

\[
d_\gamma^\pi(\bar{s}) = (1 - \gamma) \sum_{k=0}^{\infty} \gamma^k p(s_k = \bar{s})
\]
The final expression for the overall quality of the policy is the discounted return:

$$J(\pi) = \int_{S} d\gamma_{\pi}(\bar{s}) \int_{A} \pi(a|\bar{s}) \int_{S} f(s'|\bar{s}, a) r(\bar{s}, a, s') ds' da d\bar{s}$$

Assuming that the policy is parameterized by $\theta$, how can we compute the policy gradient $\nabla_{\theta} J(\pi_{\theta})$?
The Policy Gradient Theorem
Theorem 1 - Policy Gradient: [5] The gradient of the discounted return is:

$$\nabla_{\theta} J(\pi_{\theta}) = \int_{S} d_{\gamma}(\bar{s}) \int_{A} \nabla_{\theta} \pi_{\theta}(a_{k}|\bar{s}) Q_{\pi}(s, a) da d\bar{s}$$

**Proof:** The relationship between the discounted return and the state value function gives us our starting place:

$$\nabla_{\theta} J(\pi_{\theta}) = (1 - \gamma) \nabla_{\theta} \int_{S} d_{0}(s_{0}) V^{\pi}(s_{0}) ds_{0}$$

$$= (1 - \gamma) \int_{S} d_{0}(s_{0}) \nabla_{\theta} V^{\pi}(s_{0}) ds_{0}$$
Policy Gradient Theorem: Proof

Consider the gradient of the state value function:

\[ \nabla_\theta V^\pi(s) = \nabla_\theta \int_\mathcal{A} \pi_\theta(a|s) Q^\pi(s, a) da \]

\[ = \int_\mathcal{A} \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) + \pi_\theta(a|s) \nabla_\theta Q^\pi(s, a) da \]

\[ = \int_\mathcal{A} \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) + \pi_\theta(a|s) \nabla_\theta \left( \int_\mathcal{S} f(s'|s, a) \left( r(s, a, s') + \gamma V^\pi(s') \right) ds' da \right) \]

\[ = \int_\mathcal{A} \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) + \pi_\theta(a|s) \int_\mathcal{S} \gamma f(s'|s, a) \nabla_\theta V^\pi(s') ds' da \]

This is recursive expression for the gradient that we can unroll!
Unrolling the expression from $s_0$ gives:

$$
\nabla_\theta V^\pi(s_0) = \int_A \nabla_\theta \pi_\theta(a_0|s_0) Q^\pi(s_0, a_0) da_0 \\
+ \int_A \pi_\theta(a_0|s_0) \int_S \gamma f(s_1|s_0, a_0) \nabla_\theta V^\pi(s_1) ds_1 da_0 \\
= \int_S \sum_{k=0}^{\infty} \gamma^k p(s_k = \bar{s}|s_0) \int_A \nabla_\theta \pi_\theta(a|\bar{s}) Q^\pi(\bar{s}, a) da d\bar{s}
$$

So the policy gradient is given by:

$$
\frac{\nabla_\theta J(\pi_\theta)}{(1 - \gamma)} = \int_S d_0(s_0) \int_S \sum_{k=0}^{\infty} \gamma^k p(s_k = \bar{s}|s_0) \int_A \nabla_\theta \pi_\theta(a|\bar{s}) Q^\pi(\bar{s}, a) da d\bar{s} \\
= \int_S d^\pi(\bar{s}) \int_A \nabla_\theta \pi_\theta(a|\bar{s}) Q^\pi(\bar{s}, a) da d\bar{s} \quad \square
$$
• However, we generally don’t know the state-action reward function $Q^\pi(s, a)$.

• The Actor-Critic framework suggests learning an approximation $R_w(s, a)$ with parameters $w$.

• Given a fixed policy $\pi_\theta$, we want to minimize the expected least-squares error:

$$w = \arg\min_w \int_S d^\pi(\bar{s}) \int_A \pi_\theta(a|\bar{s}) \frac{1}{2} [Q^\pi(\bar{s}, a) - R_w(\bar{s}, a)]^2 da d\bar{s}$$

• Can we show that the policy gradient theorem holds for reward function learned this way?
Let’s rewrite the policy gradient theorem to use our approximate reward function:

\[ \nabla_{\theta} J(\pi_{\theta}) = \int_{S} d^{\pi}(\bar{s}) \int_{A} \nabla_{\theta} \pi_{\theta}(a|\bar{s}) \left[ R_{w}(\bar{s}, a) \right] da d\bar{s} \]

\[ = \int_{S} d^{\pi}(\bar{s}) \int_{A} \nabla_{\theta} \pi_{\theta}(a|\bar{s}) \left[ R_{w}(\bar{s}, a) - Q^{\pi}(\bar{s}, a) + Q^{\pi}(\bar{s}, a) \right] da d\bar{s} \]

\[ = \int_{S} d^{\pi}(\bar{s}) \int_{A} \nabla_{\theta} \pi_{\theta}(a|\bar{s}) Q^{\pi}(\bar{s}, a) da d\bar{s} - \int_{S} d^{\pi}(\bar{s}) \int_{A} \nabla_{\theta} \pi_{\theta}(a|\bar{s}) \left[ Q^{\pi}(\bar{s}, a) - R_{w}(\bar{s}, a) \right] da d\bar{s} \]

Intuition: We can impose technical conditions on \( R_{w}(\bar{s}, a) \) to insure the second term is zero.
Policy Gradient Theorem: Restrictions on the Critic

The sufficient conditions on $R_w$ are:

- $R_w$ is compatible with the parameterization of the policy $\pi_\theta$ in the sense:

$$\nabla_w R_w(s, a) = \nabla_\theta \log \pi_\theta(a|s) = \frac{1}{\pi_\theta(a|s)} \nabla_\theta \pi_\theta(a|s)$$

- $w$ has converged to a local minimum:

$$\nabla_w \int_S d^\pi(\bar{s}) \int_A \pi_\theta(a|\bar{s}) \frac{1}{2} [Q^\pi(\bar{s}, a) - R_w(\bar{s}, a)]^2 \, da d\bar{s} = 0$$

$$\int_S d^\pi(\bar{s}) \int_A \pi_\theta(a|\bar{s}) \nabla_w R_w(\bar{s}, a) [Q^\pi(\bar{s}, a) - R_w(\bar{s}, a)] \, da d\bar{s} = 0$$

$$\int_S d^\pi(\bar{s}) \int_A \nabla_\theta \pi_\theta(a|\bar{s}) [Q^\pi(\bar{s}, a) - R_w(\bar{s}, a)] \, da d\bar{s} = 0$$
Theorem 2 - Policy Gradient with Function Approximation:

[5] If $R_w(s, a)$ satisfies the conditions on the previous slide, the policy gradient using the learned reward function is:

$$\nabla_\theta J(\pi_\theta) = \int_S d^\pi(\bar{s}) \int_A \nabla_\theta \pi_\theta(a|\bar{s}) R_w(\bar{s}, a) da d\bar{s}.$$
• We’ve shown that the gradient of the policy quality w.r.t the policy parameters has a simple form.
• We’ve derived sufficient conditions for an actor-critic algorithm to use the policy gradient theorem.
• We’ve obtained a necessary functional form for $R_w(s, a)$, since the compatibility condition requires

$$R_w(s, a) = \nabla_\theta \log \pi_\theta(a|s)^\top w$$
We can estimate the policy gradient in practice using the score function estimator (aka REINFORCE):

\[
\nabla_\theta J(\pi_\theta) = \int_S d^\pi(\bar{s}) \int_A \nabla_\theta \pi_\theta(a|\bar{s}) R_w(\bar{s}, a) da d\bar{s} \\
= \int_S d^\pi(\bar{s}) \int_A \pi_\theta(a|\bar{s}) \nabla_\theta \log \pi_\theta(a|\bar{s}) R_w(\bar{s}, a) da d\bar{s} \\
= \int_S d^\pi(\bar{s}) \int_A \pi_\theta(a|\bar{s}) \nabla_\theta \log \pi_\theta(a|\bar{s}) \nabla_\theta \log \pi_\theta(a|s)^T w da d\bar{s}
\]

We can approximate the necessary integrals using multiple trajectories \(\tau_{0:t}\) computed under the current policy \(\pi_\theta\).
An Algorithmic Template for Actor-Critic

1. Choose initial parameters $w_0$, $\theta_0$.

2. For $i = 0, \ldots$:
   
   2.1 Update the Critic:
   
   $$w_{i+1} = \arg\min_w \int_S d^\pi(\bar{s}) \int_A \pi_\theta(a|\bar{s}) \frac{1}{2} [Q^\pi(\bar{s}, a) - R_w(\bar{s}, a)]^2 da d\bar{s}$$

   2.2 Take a policy gradient step:
   
   $$\theta_{t+1} = \theta_t + \alpha_t \int_S d^\pi(\bar{s}) \int_A \pi_\theta(a|\bar{s}) \nabla_\theta \log \pi_\theta(a|\bar{s}) R_w(\bar{s}, a) da d\bar{s}$$

This algorithm is guaranteed to converge when gradients and rewards are bounded and the $\alpha_t$ are chosen appropriately.
Natural Policy Gradients
Background on Natural Gradients: Motivation

• Consider optimizing a function with respect to parameters $\theta$:

$$\theta^* = \underset{\theta}{\text{argmin}} f(\theta)$$

• "Standard" gradient descent:

$$\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta f(\theta)$$

$$= \underset{\theta}{\text{argmin}} \{ f(\theta_t) + \langle \nabla_\theta f(\theta_t), \theta - \theta_t \rangle + \frac{1}{2\alpha} \| \theta - \theta_t \|^2 \}$$

• Issues:
  • the gradient is dependent on the parameterization/coordinate system (i.e. the choice of $\theta$);
  • it implicitly assumes that the Euclidean distance reflects the true geometry of the problem.
Background on Natural Gradients: Definition

- What can we do when $\theta$ "lives" on a manifold (e.g. the unit sphere)?
- An alternative is Amari’s "Natural" gradient descent [1]:

$$\theta_{t+1} = \theta_t + 1 - \alpha_t G(\theta)^{-1} \nabla_\theta f(\theta),$$

where $G(\theta)$ is the Riemannian metric tensor for the manifold of $\theta$.

- In Euclidean space: $G(\theta) = I$.
- When the step size $\alpha$ is arbitrarily small:
  - the natural gradient is invariant to smooth, invertible reparameterizations;
  - the natural gradient performs "steepest descent in the space of realizable [functions]" [3].
Consider an objective function defined in polar ($r$ - radius, $\varphi$ - angle) and Euclidean coordinates:

$$J(r, \varphi) = \frac{1}{2} [(r\cos\varphi - 1)^2 + r^2\sin^2\varphi]$$

$$J(x, y) = (x - 1)^2 + y^2$$
Background on Natural Gradients: Fisher Information

- Consider the case where $f$ is a probability distribution parameterized by $\theta$: $(f(\theta) = p(x|\theta))$. Then the correct metric tensor is the Fisher Information (FI) matrix:

$$F(\theta) = \int p(x|\theta) \nabla_\theta \log p(x|\theta) \nabla_\theta \log p(x|\theta)^\top dx$$

- **Interpretation**: FI is the expected (centered) second moment of the score function $\nabla_\theta \log p(x|\theta)$ and measures the information about parameters $\theta$ in the random variable $x$.

- A useful identity for the FI:

$$\int p_\theta(x) \nabla_\theta \log p_\theta(x) \nabla_\theta \log p_\theta(x)^\top dx = -\int p_\theta(x) \nabla_\theta^2 \log p_\theta(x) dx$$
Let's return to policy gradients:

\[
\nabla_\theta J(\pi_\theta) = \int_S d^{\pi}(\bar{s}) \int_\mathcal{A} \pi_\theta(a|\bar{s}) \nabla_\theta \log \pi_\theta(a|\bar{s}) \nabla_\theta \log \pi_\theta(a|s)^\top w \ da d\bar{s} \\
= \int_S d^{\pi}(\bar{s}) F(\theta) w \ d\bar{s}
\]

The policy gradient clearly contains the FI of the policy conditioned for state \( s \). Define the "average" FI:

\[
\bar{F}(\theta) := \int_S d^{\pi}(\bar{s}) F(\theta) \ d\bar{s}
\]

If \( \bar{F}(\theta) \) is the FI of an "appropriate" distribution, the natural gradient is:

\[
\bar{F}(\theta)^{-1} \nabla_\theta J(\pi_\theta) = w
\]
The probability of a trajectory $\tau_{0:t}$ obtained when acting under the policy $\pi_\theta(a|s)$ is:

$$p^\pi(\tau_{0:t}) = d_0(s_0) \prod_{i=0}^{t} f(s_{i+1}|s_i, a_i) \pi_\theta(a_i|s_i)$$

**Average reward:** it is straightforward to show that $\bar{F}(\theta)$ is the FI of $\lim_{t\to\infty} p^\pi(\tau_{0:t})$.

**Discounted reward:** Peters et al. [4] define a "discounted trajectory" distribution:

$$p^\pi_\gamma(\tau_{0:t}) = p^\pi(\tau_{0:t}) \left( \sum_{i=0}^{n} \gamma^i \mathbb{1}_{s_i,a_i} \right)$$
Interpretations:

- **Probably Incorrect**: A single scaling factor on the distribution:
  \[
  p^\pi_\gamma(\tau_{0:t}) = p^\pi(\tau_{0:t}) \times \sum_{i=0}^{t} \gamma^i
  \]

- **Closer**: A set of equivalent probability distributions with different un-normalized density functions:
  \[
  p^\pi_\gamma(\tau_{0:t}) = p^\pi(\tau_{0:t}) \sum_{i=0}^{t} \gamma^i \mathbb{1}_{s_i,a_i(\tau_{0:t})}
  \]

Peters et al. [4] prove that \( \bar{F}(\theta) \) is the FI of the discounted trajectory distribution. Let's look carefully at their argument.
Theorem 3 - Natural Policy Gradient: [4] The average FI information

\[ \bar{F}(\theta) = \int_S d^{\pi}(\bar{s})F(\theta) \, d\bar{s} \]

is the FI of the discounted trajectory distribution \( p^{\pi}_{\gamma}(\tau_{0:t}) \).

Proof:

Recall the definition of the trace distribution:

\[ p^{\pi}(\tau_{0:t}) = d_0(s_0) \prod_{i=0}^{t} f(s_{i+1}|s_i, a_i) \pi_{\theta}(a_i|s_i) \]

The Hessian of the log probability is

\[ \nabla^2_{\theta} \log p^{\pi}_{\gamma}(\tau_{0:t}) = \sum_{i=0}^{t} \nabla^2_{\theta} \log \pi_{\theta}(a_i|s_i) \]
Natural Policy Gradients: Starting the Derivation

**Approach:** transform the expression for the FI of $p_{\pi}(\tau_{0:t})$ to match that for $\bar{F}(\theta)$:

$$F(\theta) = \lim_{t \to \infty} \int p_{\pi}(\tau_{0:t}) \nabla_{\theta} \log p_{\pi}(\tau_{0:t}) \nabla_{\theta} p_{\pi}(\tau_{0:t})^\top d\tau_{0:t}$$

$$= - \lim_{t \to \infty} \int p_{\pi}(\tau_{0:t}) \nabla_{\theta}^2 \log p_{\pi}(\tau_{0:t}) d\tau_{0:t}$$

$$= - \lim_{t \to \infty} \int p_{\pi}(\tau_{0:t}) \sum_{i=0}^{t} \nabla_{\theta}^2 \log \pi(a_i|s_i) d\tau_{0:t}$$

$$= - \lim_{t \to \infty} \sum_{i=0}^{t} p_{\pi}(\tau_{0:t}) \nabla_{\theta}^2 \log \pi(a_i|s_i) d\tau_{0:t}$$
They appear to evaluate the indicator functions and then normalize the sum of density functions:

$$F(\theta) = - \lim_{t \to \infty} \int (1 - \gamma) \sum_{i=0}^{t} \gamma^i p^\pi(\tau_{0:t}) \nabla^2 \theta \log \pi(a_i | s_i) d\tau_{0:t}$$

$$= - \lim_{t \to \infty} \int (1 - \gamma) \sum_{i=0}^{t} \gamma^i p^\pi(\tau_{0:i}) \nabla^2 \theta \log \pi(a_i | s_i) d\tau_{0:i}$$

$$= - \lim_{t \to \infty} \int_{S} (1 - \gamma) \sum_{i=0}^{t} \gamma^i p^\pi(s_i = \bar{s}) \int_{A} \pi_\theta(a_i | \bar{s}) \nabla^2 \theta \log \pi(a_i | \bar{s}) da_i d\bar{s}$$

$$= - \int_{S} \gamma^i d^\pi(s = \bar{s}) \int_{A} \pi_\theta(a | \bar{s}) \nabla^2 \theta \log \pi(a | s) da d\bar{s}$$

$$= \int_{S} \gamma^i d^\pi(s = \bar{s}) \int_{A} \pi_\theta(a | \bar{s}) \nabla \theta \log \pi(a | s) \nabla \theta \log \pi(a | s)^T da d\bar{s}$$

Is this still defined w.r.t the correct distribution?
Normalizing the sum of density functions reweights the terms in the sum. Consider the same expression with pre-normalized densities:

\[
F(\theta) = - \lim_{t \to \infty} \int \sum_{i=0}^{t} \gamma_i \gamma_i p^{\pi}(\tau_{0:t}) \nabla_{\theta}^{2} \log \pi(a_i|s_i) d\tau_{0:t}
\]

\[
= - \lim_{t \to \infty} \int \sum_{i=0}^{t} \gamma_i \gamma_i p^{\pi}(\tau_{0:i}) \nabla_{\theta}^{2} \log \pi(a_i|s_i) d\tau_{0:i}
\]

\[
= - \lim_{t \to \infty} \int \sum_{i=0}^{t} p^{\pi}(s_i = \bar{s}) \int_{A} \nabla_{\theta}^{2} \log \pi(a_i|\bar{s}) d\theta d\bar{s}
\]

\[
= - \lim_{t \to \infty} \int \sum_{i=0}^{t} p^{\pi}(s_i = \bar{s}) \int_{A} \nabla_{\theta} \log \pi(a_i|\bar{s}) \nabla_{\theta} \log \pi(a_i|\bar{s})^{\top} d\theta d\bar{s}
\]

**Crux of the Issue:** the discounted trajectory distribution \( p^{\pi}_{\gamma}(\tau_{0:t}) \).
1. Choose initial parameters $w_0, \theta_0$.

2. For $i = 0...$:
   
   2.1 Update the Critic:
   
   $$w_{i+1} = \text{argmin}_w \int_S d^\pi(\bar{s}) \int_A \pi_\theta(a|\bar{s}) \frac{1}{2} [Q^\pi(\bar{s}, a) - R_w(\bar{s}, a)]^2 da d\bar{s}$$

   2.2 Take a policy gradient step:
   
   $$\theta_{t+1} = \theta_t + \alpha_t w_{i+1}$$

Convergence results for natural actor-critic algorithms depend on how the critic is updated. Convergence with probability 1 is guaranteed for some schemes.
Shun-Ichi Amari.  
**Natural gradient works efficiently in learning.**  

Ivo Grondman, Lucian Busoniu, Gabriel AD Lopes, and Robert Babuska.  
**A survey of actor-critic reinforcement learning: Standard and natural policy gradients.**  

James Martens.  
**New insights and perspectives on the natural gradient method.**  