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

Why Policy Gradients?

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 $\mathbf{s} \in \mathcal{S}$; actions are $\mathbf{a} \in \mathcal{A}$.
- f is the transition distribution. It satisfies the Markov property:

$$f(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = p(\mathbf{s}_{t+1} | \mathbf{s}_0, \mathbf{a}_0 ... \mathbf{s}_t, \mathbf{a}_t) = p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)$$

- $d_0(\mathbf{s}_0)$ is the initial distribution over states.
- $r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$ is the reward function, which may be deterministic or stochastic.
- Trajectories are sequences of state-action pairs: $\boldsymbol{\tau}_{0:t} = \{(\mathbf{s}_0, \mathbf{a}_0), ..., (\mathbf{s}_t, \mathbf{a}_t)\}$

We treat states $\mathbf{s}$ as fully observable.

Continuous State and Action Spaces

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

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

$$P(\mathbf{s}' \in \mathcal{S}' | \mathbf{s}, \mathbf{a}) = \int_{\mathcal{S}'} p(\mathbf{s}' | \mathbf{s}, \mathbf{a}) d\mathbf{s}'$$

Future states $\mathbf{s}'$ are only known probabilistically because our control and physical models are approximations.

Policies

Policies defines how an agent acts in the MDP:

- A *policy* $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, \infty)$ is the conditional density function:

$$\pi(\mathbf{a}|\mathbf{s}) := \text{probability of taking action } \mathbf{a} \text{ in state } \mathbf{s}$$

- The policy is deterministic when $\pi(\mathbf{a}|\mathbf{s})$ is a Dirac-delta function.
- Actions are chosen by sampling from the policy $\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})$.
- 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:

$$Q^{\pi}(\mathbf{s}, \mathbf{a}) := \int_{\mathcal{S}} f(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \left(r(\mathbf{s}, \mathbf{a}, \mathbf{s}') + \int_{\mathcal{A}} \pi(\mathbf{a}'|\mathbf{s}') \gamma Q^{\pi}(\mathbf{s}', \mathbf{a}') d\mathbf{a}' \right) d\mathbf{s}'$$

$$\begin{aligned}
V^{\pi}(\mathbf{s}) &:= \int_{\mathcal{A}} \pi(\mathbf{a}|\mathbf{s}) Q^{\pi}(\mathbf{s}, \mathbf{a}) d\mathbf{a} \\
&= \int_{\mathcal{A}} \pi(\mathbf{a}|\mathbf{s}) \int_{\mathcal{S}} f(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \left(r(\mathbf{s}, \mathbf{a}, \mathbf{s}') + \gamma V^{\pi}(\mathbf{s}') \right) d\mathbf{s}' d\mathbf{a} \\
&= \int_{\mathcal{A}} \pi(\mathbf{a}|\mathbf{s}) \int_{\mathcal{S}} f(\mathbf{s}'|\mathbf{s}, \mathbf{a}) r(\mathbf{s}, \mathbf{a}, \mathbf{s}') d\mathbf{s}' d\mathbf{a} \\
&+ \int_{\mathcal{A}} \pi(\mathbf{a}|\mathbf{s}) \int_{\mathcal{S}} f(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \gamma V^{\pi}(\mathbf{s}') d\mathbf{s}' d\mathbf{a}
\end{aligned}$$

Actor-Critic Methods

Three major flavors of reinforcement learning:

1. Critic-only methods: Learn an approximation of the state-action reward function: $R(\mathbf{s}, \mathbf{a}) \approx Q^{\pi}(\mathbf{s}, \mathbf{a})$.
2. Actor-only methods: Learn the policy π directly from observed rewards. A parametric policy π_θ 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 $R(\mathbf{s}, \mathbf{a})$ jointly with the policy $\pi(\mathbf{a}|\mathbf{s})$.

Value of a Policy

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

$$
\begin{aligned}
\frac{J(\pi)}{(1-\gamma)} &= \int_{\mathcal{S}} d_0(\mathbf{s}_0) V^{\pi}(\mathbf{s}_0) d\mathbf{s}_0 \\
&= \sum_{k=0}^{\infty} \int_{\mathcal{S}} p(\mathbf{s}_k = \bar{\mathbf{s}}) \int_{\mathcal{A}} \pi(\mathbf{a}_k | \bar{\mathbf{s}}) \int_{\mathcal{S}} f(\mathbf{s}_{k+1} | \bar{\mathbf{s}} \mathbf{a}_k) \gamma^k r(\bar{\mathbf{s}}, \mathbf{a}_k, \mathbf{s}_{k+1}) d\mathbf{s}_{t+1} d\mathbf{a} d\bar{\mathbf{s}} \\
&= \int_{\mathcal{S}} \sum_{k=0}^{\infty} \gamma^k p(\mathbf{s}_k = \bar{\mathbf{s}}) \int_{\mathcal{A}} \pi(\mathbf{a}_k | \bar{\mathbf{s}}) \int_{\mathcal{S}} f(\mathbf{s}_{k+1} | \bar{\mathbf{s}} \mathbf{a}_k) r(\bar{\mathbf{s}}, \mathbf{a}_k, \mathbf{s}_{k+1}) d\mathbf{s}_{t+1} d\mathbf{a} d\bar{\mathbf{s}}
\end{aligned}
$$

Define the "discounted state" distribution:

$$
d_{\gamma}^{\pi}(\bar{\mathbf{s}}) = (1-\gamma) \sum_{k=0}^{\infty} \gamma^k p(\mathbf{s}_k = \bar{\mathbf{s}})
$$

Value of Policy: Discounted Return

The final expression for the overall quality of the policy is the *discounted return*:

$$J(\pi) = \int_{\mathcal{S}} d_\gamma^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi(\mathbf{a}|\bar{\mathbf{s}}) \int_{\mathcal{S}} f(\mathbf{s}'|\bar{\mathbf{s}}, \mathbf{a}) r(\bar{\mathbf{s}}, \mathbf{a}, \mathbf{s}') d\mathbf{s}' d\mathbf{a} d\bar{\mathbf{s}}$$

Assuming that the policy is parameterized by θ, how can we compute the policy gradient $\nabla_\theta J(\pi_\theta)$?

The Policy Gradient Theorem

Policy Gradient Theorem: Statement

Theorem 1 - Policy Gradient: [5] The gradient of the discounted return is:

$$\nabla_\theta J(\pi_\theta) = \int_{\mathcal{S}} d_\gamma^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}_k|\bar{\mathbf{s}}) Q^\pi(\mathbf{s}, \mathbf{a}) d\mathbf{a} d\bar{\mathbf{s}}$$

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

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= (1-\gamma)\nabla_\theta \int_{\mathcal{S}} d_0(\mathbf{s}_0) V^\pi(\mathbf{s}_0) d\mathbf{s}_0 \\ &= (1-\gamma) \int_{\mathcal{S}} d_0(\mathbf{s}_0) \nabla_\theta V^\pi(\mathbf{s}_0) d\mathbf{s}_0\end{aligned}$$

Policy Gradient Theorem: Proof

Consider the gradient of the state value function:

$$
\begin{aligned}
\nabla_\theta V^\pi(\mathbf{s}) &= \nabla_\theta \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\mathbf{s}) Q^\pi(\mathbf{s}, \mathbf{a}) d\mathbf{a} \\
&= \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\mathbf{s}) Q^\pi(\mathbf{s}, \mathbf{a}) + \pi_\theta(\mathbf{a}|\mathbf{s}) \nabla_\theta Q^\pi(\mathbf{s}, \mathbf{a}) d\mathbf{a} \\
&= \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\mathbf{s}) Q^\pi(\mathbf{s}, \mathbf{a}) + \pi_\theta(\mathbf{a}|\mathbf{s}) \nabla_\theta \int_{\mathcal{S}} f(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \Big(r(\mathbf{s}, \mathbf{a}, \mathbf{s}') + \\
&\qquad\qquad \gamma V^\pi(\mathbf{s}') \Big) d\mathbf{s}' d\mathbf{a} \\
&= \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\mathbf{s}) Q^\pi(\mathbf{s}, \mathbf{a}) + \pi_\theta(\mathbf{a}|\mathbf{s}) \int_{\mathcal{S}} \gamma f(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \nabla_\theta V^\pi(\mathbf{s}') d\mathbf{s}' d\mathbf{a}
\end{aligned}
$$

This is recursive expression for the gradient that we can unroll!

Policy Gradient Theorem: Proof Continued

Unrolling the expression from $\mathbf{s}_0$ gives:

$$\begin{aligned}\nabla_\theta V^\pi(\mathbf{s}_0) &= \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}_0|\mathbf{s}_0) Q^\pi(\mathbf{s}_0, \mathbf{a}_0) d\mathbf{a}_0 \\ &\quad + \int_{\mathcal{A}} \pi_\theta(\mathbf{a}_0|\mathbf{s}_0) \int_{\mathcal{S}} \gamma f(\mathbf{s}_1|\mathbf{s}_0, \mathbf{a}_0) \nabla_\theta V^\pi(\mathbf{s}_1) d\mathbf{s}_1 d\mathbf{a}_0 \\ &= \int_{\mathcal{S}} \sum_{k=0}^{\infty} \gamma^k p(\mathbf{s}_k = \bar{\mathbf{s}}|\mathbf{s}_0) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\bar{\mathbf{s}}\end{aligned}$$

So the policy gradient is given by:

$$\begin{aligned}\frac{\nabla_\theta J(\pi_\theta)}{(1-\gamma)} &= \int_{\mathcal{S}} d_0(s_0) \int_{\mathcal{S}} \sum_{k=0}^{\infty} \gamma^k p(\mathbf{s}_k = \bar{\mathbf{s}}|\mathbf{s}_0) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\bar{\mathbf{s}} \\ &= \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\bar{\mathbf{s}} \qquad \square\end{aligned}$$

Policy Gradient Theorem: Introducing Critics

- However, we generally don't know the state-action reward function $Q^\pi(\mathbf{s}, \mathbf{a})$.
- The Actor-Critic framework suggests learning an approximation $R_w(\mathbf{s}, \mathbf{a})$ with parameters w.
- Given a fixed policy π_θ, we want to minimize the expected least-squares error:

$$\mathbf{w} = \mathbf{argmin}_w \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \frac{1}{2} \left[Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) - R_w(\bar{\mathbf{s}}, \mathbf{a}) \right]^2 d\mathbf{a} d\bar{\mathbf{s}}$$

- Can we show that the policy gradient theorem holds for reward function learned this way?

Policy Gradient Theorem: The Way Forward

Let's rewrite the policy gradient theorem to use our approximate reward function:

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \left[R_w(\bar{\mathbf{s}}, \mathbf{a}) \right] d\mathbf{a} d\bar{\mathbf{s}} \\
&= \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \left[R_w(\bar{\mathbf{s}}, \mathbf{a}) - Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) + Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) \right] d\mathbf{a} d\bar{\mathbf{s}} \\
&= \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\bar{\mathbf{s}} - \\
&\qquad \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \left[Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) - R_w(\bar{\mathbf{s}}, \mathbf{a}) \right] d\mathbf{a} d\bar{\mathbf{s}}
\end{aligned}
$$

Intuition: We can impose technical conditions on $R_w(\bar{\mathbf{s}}, \mathbf{a})$ 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 π_θ in the sense:

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

- **w** has converged to a local minimum:

$$\nabla_w \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}})\frac{1}{2}\left[Q^\pi(\bar{\mathbf{s}},\mathbf{a}) - R_w(\bar{\mathbf{s}},\mathbf{a})\right]^2 d\mathbf{a}d\bar{\mathbf{s}} = 0$$
$$\int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}})\nabla_w R_w(\bar{\mathbf{s}},\mathbf{a})\left[Q^\pi(\bar{\mathbf{s}},\mathbf{a}) - R_w(\bar{\mathbf{s}},\mathbf{a})\right] d\mathbf{a}d\bar{\mathbf{s}} = 0$$
$$\int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}})\left[Q^\pi(\bar{\mathbf{s}},\mathbf{a}) - R_w(\bar{\mathbf{s}},\mathbf{a})\right] d\mathbf{a}d\bar{\mathbf{s}} = 0$$

Policy Gradient Theorem: Function Approximation Version

Theorem 2 - Policy Gradient with Function Approximation:
[5] If $R_w(\mathbf{s}, \mathbf{a})$ satisfies the conditions on the previous slide, the policy gradient using the learned reward function is:

$$\nabla_\theta J(\pi_\theta) = \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) R_w(\bar{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\bar{\mathbf{s}}.$$

Policy Gradient Theorem: Recap

- 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 $R_w(\mathbf{s}, \mathbf{a})$, since the compatibility condition requires

$$R_w(\mathbf{s}, \mathbf{a}) = \nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s})^\top \mathbf{w}$$

Policy Gradient Theorem: Actually Computing the Gradient

- We can estimate the policy gradient in practice using the score function estimator (aka REINFORCE):

$$\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) R_w(\bar{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\bar{\mathbf{s}} \\
&= \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \nabla_\theta \log \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) R_w(\bar{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\bar{\mathbf{s}} \\
&= \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \nabla_\theta \log \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s})^\top \mathbf{w} \; d\mathbf{a} d\bar{\mathbf{s}}
\end{aligned}$$

- We can approximate the necessary integrals using multiple trajectories $\boldsymbol{\tau}_{0:t}$ computed under the current policy π_θ.

An Algorithmic Template for Actor-Critic

1. Choose initial parameters $\mathbf{w}_0$, $\boldsymbol{\theta}_0$.
2. For $i = 0 ...$:
   2.1 Update the Critic:

$$\mathbf{w}_{i+1} = \mathbf{argmin}_w \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \frac{1}{2} \left[Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) - R_w(\bar{\mathbf{s}}, \mathbf{a}) \right]^2 d\mathbf{a} d\bar{\mathbf{s}}$$

   2.2 Take a policy gradient step:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \nabla_\theta \log \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) R_w(\bar{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\bar{\mathbf{s}}$$

This algorithm is guaranteed to converge when gradients and rewards are bounded and the α_t are chosen appropriately.

Natural Policy Gradients

Background on Natural Gradients: Motivation

- Consider optimizing a function with respect to parameters $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^* = \textbf{argmin}_\theta f(\boldsymbol{\theta})$$

- "Standard" gradient descent:

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \alpha_t \nabla_\theta f(\boldsymbol{\theta}) \\ &= \textbf{argmin}_\theta \{ f(\boldsymbol{\theta}_t) + \langle \nabla_\theta f(\boldsymbol{\theta}_t), \boldsymbol{\theta} - \boldsymbol{\theta}_t \rangle + \frac{1}{2\alpha} ||\boldsymbol{\theta} - \boldsymbol{\theta}_t||^2 \}\end{aligned}$$

- **Issues:**
  - the gradient is dependent on the parameterization/coordinate system (i.e. the choice of $\boldsymbol{\theta}$);
  - it implicitly assumes that the Eucledian distance reflects the true geometry of the problem.

Background on Natural Gradients: Definition

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

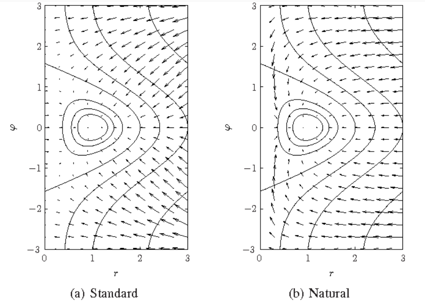
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{t+1} - \alpha_t \mathbf{G}(\boldsymbol{\theta})^{-1} \nabla_\theta f(\boldsymbol{\theta}),$$

  where $\mathbf{G}(\theta)$ is the Riemannian metric tensor for the manifold of θ.
- In Eucledian space: $\mathbf{G}(\boldsymbol{\theta}) = \mathbf{I}$.
- When the step size $\boldsymbol{\alpha}$ 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].

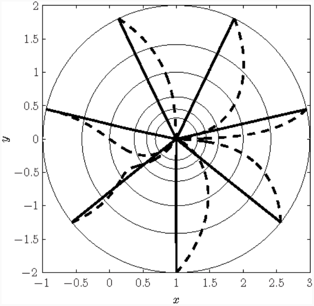
Consider an objective function defined in polar (r - radius, φ - angle) and Eucledian coordinates:

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

$$J(x,y) = (x-1)^2 + y^2$$



(a) Gradient Field



(b) Training Paths

Figures and example taken from [2].

Background on Natural Gradients: Fisher Information

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

$$\mathbf{F}(\theta) = \int p(\mathbf{x}|\boldsymbol{\theta}) \nabla_\theta \log p(\mathbf{x}|\boldsymbol{\theta}) \nabla_\theta \log p(\mathbf{x}|\boldsymbol{\theta})^\top d\mathbf{x}$$

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

$$\int p_\theta(\mathbf{x}) \nabla_\theta \log p_\theta(\mathbf{x}) \nabla_\theta \log p_\theta(\mathbf{x})^\top d\mathbf{x} = -\int p_\theta(\mathbf{x}) \nabla_\theta^2 \log p_\theta(\mathbf{x}) d\mathbf{x}$$

FI and the Policy Gradient Theorem

Let's return to policy gradients:

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \nabla_\theta \log \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s})^\top \mathbf{w} \; d\mathbf{a} d\bar{\mathbf{s}} \\ &= \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \mathbf{F}(\boldsymbol{\theta}) \mathbf{w} \; d\bar{\mathbf{s}}\end{aligned}$$

The policy gradient clearly contains the FI of the policy conditioned for state $\mathbf{s}$. Define the "average" FI:

$$\bar{\mathbf{F}}(\theta) := \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \mathbf{F}(\boldsymbol{\theta}) \; d\bar{\mathbf{s}}$$

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

$$\bar{\mathbf{F}}(\theta)^{-1} \nabla_\theta J(\pi_\theta) = \mathbf{w}$$

Natural Policy Gradients: Trajectories

- The probability of a trajectory $\boldsymbol{\tau}_{0:t}$ obtained when acting under the policy $\pi_\theta(\mathbf{a}|\mathbf{s})$ is:

$$p^\pi(\boldsymbol{\tau}_{0:t}) = d_0(\mathbf{s}_0) \prod_{i=0}^{t} f(\mathbf{s}_{i+1}|\mathbf{s}_i, \mathbf{a}_i)\pi_\theta(\mathbf{a}_i|\mathbf{s}_i)$$

- **Average reward:** it is straightforwad to show that $\bar{\mathbf{F}}(\theta)$ is the FI of $\lim_{t\to\infty} p^\pi(\boldsymbol{\tau}_{0:t})$.
- **Discounted reward:** Peters et al. [4] define a "discounted trajectory" distribution:

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

Natural Policy Gradients: Discounted Trajectory Distribution

Interpretations:

- **Probably Incorrect**: A single scaling factor on the distribution:

$$p_\gamma^\pi(\boldsymbol{\tau}_{0:t}) = p^\pi(\boldsymbol{\tau}_{0:t}) * \sum_{i=0}^{t} \gamma^i$$

- **Closer**: A set of equivalent probability distributions with different un-normalized density functions:

$$p_\gamma^\pi(\boldsymbol{\tau}_{0:t}) = p^\pi(\boldsymbol{\tau}_{0:t}) \sum_{i=0}^{t} \gamma^i \mathbb{1}_{s_i,a_i}(\boldsymbol{\tau}_{0:t})$$

Peters et al. [4] prove that $\bar{\mathbf{F}}(\theta)$ is the FI of the discounted trajectory distribution. Lets look carefully at their argument.

Natural Policy Gradients: Statement

Theorem 3 - Natural Policy Gradient: [4] The average FI information

$$\bar{\mathbf{F}}(\theta) = \int_{\mathcal{S}} d^{\pi}(\bar{\mathbf{s}})\mathbf{F}(\boldsymbol{\theta})\ d\bar{\mathbf{s}}$$

is the FI of the discounted trajectory distribution $p_{\gamma}^{\pi}(\boldsymbol{\tau}_{0:t})$.

Proof:

Recall the defintion of the trace distribution:

$$p^{\pi}(\boldsymbol{\tau}_{0:t}) = d_0(\mathbf{s}_0)\prod_{i=0}^{t} f(\mathbf{s}_{i+1}|\mathbf{s}_i, \mathbf{a}_i)\pi_{\theta}(\mathbf{a}_i|\mathbf{s}_i)$$

The Hessian of the log probability is

$$\nabla_{\theta}^2 \log p_{\gamma}^{\pi}(\boldsymbol{\tau}_{0:t}) = \sum_{i=0}^{t} \nabla_{\theta}^2 \log \pi_{\theta}(\mathbf{a}_i|\mathbf{s}_i)$$

Natural Policy Gradients: Starting the Derivation

Approach: transform the expression for the FI of $p_\gamma^\pi(\boldsymbol{\tau}_{0:t})$ to match that for $\bar{\mathbf{F}}(\theta)$:

$$\begin{aligned}
\mathbf{F}(\theta) &= \lim_{t\to\infty} \int p_\gamma^\pi(\boldsymbol{\tau}_{0:t}) \nabla_\theta \log p_\gamma^\pi(\boldsymbol{\tau}_{0:t}) \nabla_\theta p_\gamma^\pi(\boldsymbol{\tau}_{0:t})^\top d\boldsymbol{\tau}_{0:t} \\
&= -\lim_{t\to\infty} \int p_\gamma^\pi(\boldsymbol{\tau}_{0:t}) \nabla_\theta^2 \log p_\gamma^\pi(\boldsymbol{\tau}_{0:t}) d\boldsymbol{\tau}_{0:t} \\
&= -\lim_{t\to\infty} \int p_\gamma^\pi(\boldsymbol{\tau}_{0:t}) \sum_{i=0}^{t} \nabla_\theta^2 \log \pi(\mathbf{a}_i|\mathbf{s}_i) d\boldsymbol{\tau}_{0:t} \\
&= -\lim_{t\to\infty} \int \sum_{i=0}^{t} p_\gamma^\pi(\boldsymbol{\tau}_{0:t}) \nabla_\theta^2 \log \pi(\mathbf{a}_i|\mathbf{s}_i) d\boldsymbol{\tau}_{0:t}
\end{aligned}$$

Natural Policy Gradients: Following Peters et al.

They appear to evaluate the indicator functions and then normalize the **sum** of density functions:

$$
\begin{aligned}
\mathbf{F}(\theta) &= -\lim_{t\to\infty} \int (1-\gamma) \sum_{i=0}^{t} \gamma^i p^\pi(\boldsymbol{\tau}_{0:t}) \nabla_\theta^2 \log \pi(\mathbf{a}_i|\mathbf{s}_i) d\boldsymbol{\tau}_{0:t} \\
&= -\lim_{t\to\infty} \int (1-\gamma) \sum_{i=0}^{t} \gamma^i p^\pi(\boldsymbol{\tau}_{0:i}) \nabla_\theta^2 \log \pi(\mathbf{a}_i|\mathbf{s}_i) d\boldsymbol{\tau}_{0:i} \\
&= -\lim_{t\to\infty} \int_{\mathcal{S}} (1-\gamma) \sum_{i=0}^{t} \gamma^i p^\pi(\mathbf{s}_i = \bar{s}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}_i|\bar{\mathbf{s}}) \nabla_\theta^2 \log \pi(\mathbf{a}_i|\bar{\mathbf{s}}) d\mathbf{a}_i d\bar{\mathbf{s}} \\
&= -\int_{\mathcal{S}} \gamma^i d^\pi(\mathbf{s} = \bar{s}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \nabla_\theta^2 \log \pi(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\bar{\mathbf{s}} \\
&= \int_{\mathcal{S}} \gamma^i d^\pi(\mathbf{s} = \bar{s}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \nabla_\theta \log \pi(\mathbf{a}|\mathbf{s}) \nabla_\theta \log \pi(\mathbf{a}|\mathbf{s})^\top d\mathbf{a} d\bar{\mathbf{s}}
\end{aligned}
$$

Is this still defined w.r.t the correct distribution?

Natural Policy Gradients: Getting Stuck

Normalizing the sum of density functions reweights the terms in the sum. Consider the same expression with pre-normalized densities:

$$\begin{aligned}
\mathbf{F}(\theta) &= -\lim_{t\to\infty} \int \sum_{i=0}^{t} \frac{\gamma^i}{\gamma^i} p^\pi(\boldsymbol{\tau}_{0:t}) \nabla_\theta^2 \log \pi(\mathbf{a}_i|\mathbf{s}_i) d\boldsymbol{\tau}_{0:t} \\
&= -\lim_{t\to\infty} \int \sum_{i=0}^{t} \frac{\gamma^i}{\gamma^i} p^\pi(\boldsymbol{\tau}_{0:i}) \nabla_\theta^2 \log \pi(\mathbf{a}_i|\mathbf{s}_i) d\boldsymbol{\tau}_{0:i} \\
&= -\lim_{t\to\infty} \int_{\mathcal{S}} \sum_{i=0}^{t} p^\pi(\mathbf{s}_i = \bar{s}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}_i|\bar{\mathbf{s}}) \nabla_\theta^2 \log \pi(\mathbf{a}_i|\bar{\mathbf{s}}) d\mathbf{a} d\bar{\mathbf{s}} \\
&= -\lim_{t\to\infty} \int_{\mathcal{S}} \sum_{i=0}^{t} p^\pi(\mathbf{s}_i = \bar{s}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}_i|\bar{\mathbf{s}}) \nabla_\theta \log \pi(\mathbf{a}_i|\bar{\mathbf{s}}) \nabla_\theta \log \pi(\mathbf{a}_i|\bar{\mathbf{s}})^\top d\mathbf{a} d\bar{\mathbf{s}}
\end{aligned}$$

Crux of the Issue: the discounted trajectory distribution $p_\gamma^\pi(\boldsymbol{\tau}_{0:t})$.

An Algorithmic Template for Natural Actor-Critic

1. Choose initial parameters $\mathbf{w}_0$, $\boldsymbol{\theta}_0$.
2. For $i = 0 \ldots$:
   2.1 Update the Critic:

   $$\mathbf{w}_{i+1} = \mathbf{argmin}_w \int_{\mathcal{S}} d^\pi(\bar{\mathbf{s}}) \int_{\mathcal{A}} \pi_\theta(\mathbf{a}|\bar{\mathbf{s}}) \frac{1}{2} \left[Q^\pi(\bar{\mathbf{s}}, \mathbf{a}) - R_w(\bar{\mathbf{s}}, \mathbf{a}) \right]^2 d\mathbf{a} d\bar{\mathbf{s}}$$

   2.2 Take a policy gradient step:

   $$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t \mathbf{w}_{i+1}$$

Convergence results for natural actor-critic algorithms depend on how the critic is updated. Convergence with probability 1 is guaranteed for some schemes.

References i

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

Ivo Grondman, Lucian Busoniu, Gabriel AD Lopes, and Robert Babuska.
A survey of actor-critic reinforcement learning: Standard and natural policy gradients.
IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(6):1291–1307, 2012.

James Martens.
New insights and perspectives on the natural gradient method.
arXiv preprint arXiv:1412.1193, 2014.

References ii

Jan Peters, Sethu Vijayakumar, and Stefan Schaal.
Reinforcement learning for humanoid robotics.
In *Proceedings of the third IEEE-RAS international conference on humanoid robots*, pages 1–20, 2003.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour.
Policy gradient methods for reinforcement learning with function approximation.
In *Advances in neural information processing systems*, pages 1057–1063, 2000.