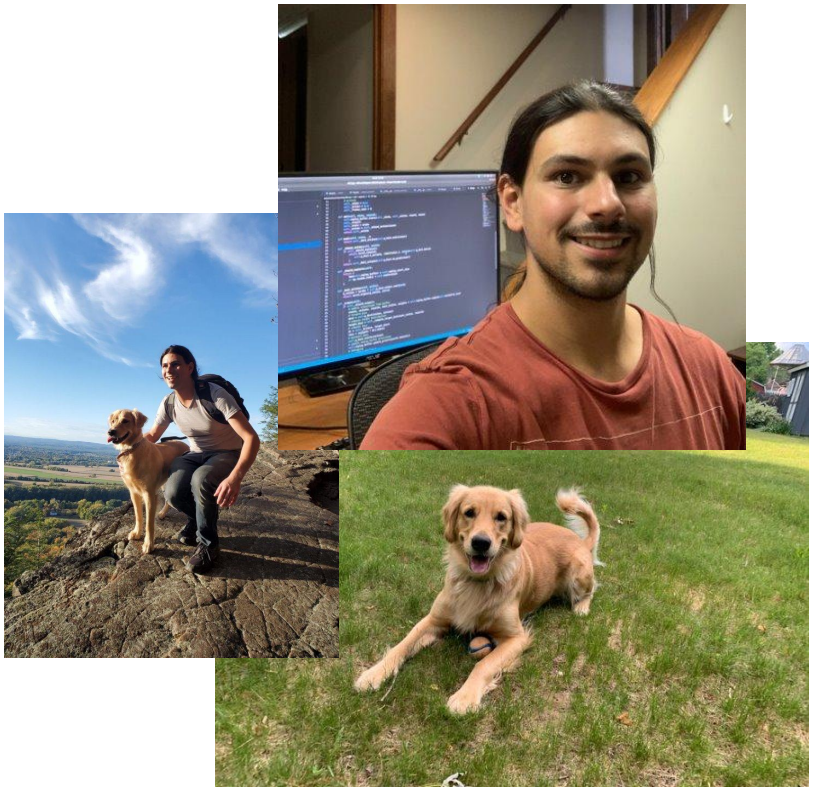
Is the Policy Gradient a Gradient?

Chris Nota, Philip S. Thomas
Autonomous Learning Lab
University of Massachusetts

Published in AAMAS 2020

About Me

- **Name**: Chris Nota
- **Position:** 4th Year PhD Candidate at University of Massachusetts
- **Advisor**: Prof. Philip Thomas
- **Lab:** Autonomous Learning Lab
- **Research Interests:** Reinforcement learning, policy gradient methods, cognitive architectures
- **History:**
  - B.S. in Computer Science from Worcester Polytechnic Institute (WPI)
  - Software Engineer at Cisco (2015-2017)
- **Website:** https://cpnota.github.io
- Author of Autonomous Learning Library



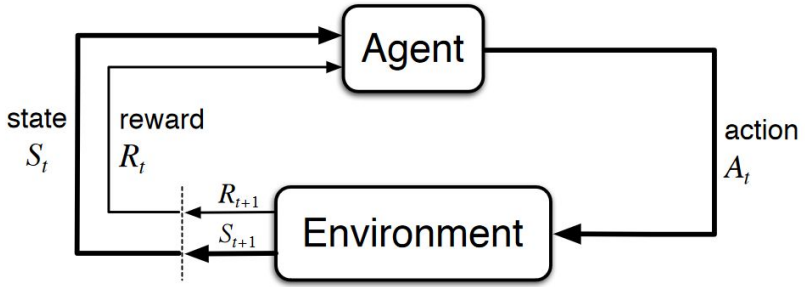
About the Autonomous Learning Lab

- **Founder**
  - Andrew Barto
- **Directors**
  - Philip S. Thomas
  - Bruno Castro da Silva
- **Notable Alumni (non-exhaustive list)**
  - Richard S. Sutton
  - Satinder Singh
  - Doina Precup
- **Website**: https://all.cs.umass.edu
- **Goals:** To develop more capable artificial agents, ensure that systems that use artificial intelligence methods are safe and well-behaved, improve our understanding of biological learning and its neural basis, and to forge stronger links between studies of learning by computer scientists, engineers, neuroscientists, and psychologists.

The Autonomous Learning Laboratory (ALL) conducts foundational artificial intelligence (AI) research, with emphases on AI safety and reinforcement learning (RL), and particularly the intersection of these two areas.

Notation

MDP $M = (\mathcal{S}, \mathcal{A}, P, R, d_0, \gamma)$
$\mathcal{S}$ = state set
$\mathcal{A}$ = action set
$P(s, a, s') := \Pr(S_{t+1} = s | S_t = s, A_t = a)$
$R(s, a) := \mathbb{E}[R_t | S_t = s, A_t = a]$
$d_0(s) := \Pr(S_0 = s)$
γ = discount factor



π^θ = policy parameterized by weights θ
$\pi^\theta(s, a) := \Pr(A_t = a | S_t = s, \theta)$
$Q^\theta_\gamma(s, a) := \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t, = s, A_t = a, \theta]$
$V^\theta_\gamma(s) := \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t, = s, \theta]$

Assume:
s_∞ = terminal absorbing state
$\sum_{t=0}^{\infty} \Pr(S_t \neq s_\infty) < \infty$
$R_t \in [-R_{\max}, R_{\max}]$

Policy Gradient Methods

Discounted objective:

$$J_\gamma(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| \theta \right]$$

Policy gradient theorem (Sutton et al., 2000):

$$\nabla J_\gamma(\theta) = \sum_s d^\pi(s) \sum_a Q_\gamma^\theta(s,a) \frac{\partial \pi^\theta(s,a)}{\partial \theta}$$

$$d^{\pi^\theta}(s) = \sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s | \theta)$$

Expected value form:

$$\nabla J_\gamma(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t Q_\gamma^\theta(s,a) \frac{\partial \ln \pi^\theta(s,a)}{\partial \theta} \middle| \theta \right]$$

Problem Statement

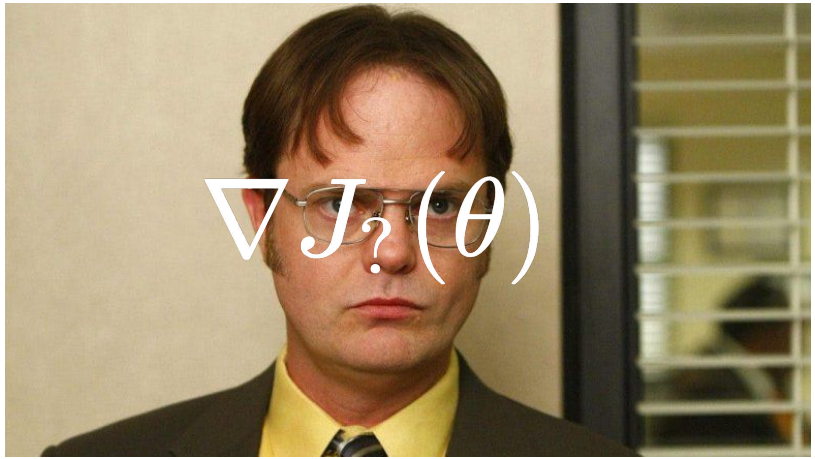
Discounted policy gradient:

$$\nabla J_\gamma(\theta) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t Q_\gamma^\theta(s,a) \frac{\partial \ln \pi^\theta(s,a)}{\partial \theta} \Big| \theta \Big]$$

What we actually use (“discounted approximation *to* the policy gradient”):

$$\nabla J_?(\theta) = \mathbb{E}\Big[\sum_{t=0}^{\infty} Q_\gamma^\theta(s,a) \frac{\partial \ln \pi^\theta(s,a)}{\partial \theta} \Big| \theta \Big]$$

$$\nabla J_?(\theta)$$



I’m the discounted policy gradient.

You’re the discounted approximation *to* the policy gradient.

Problem Statement

Discounted policy gradient:

$$\nabla J_\gamma(\theta) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t Q_\gamma^\theta(s,a) \frac{\partial \ln \pi^\theta(s,a)}{\partial \theta} \Big| \theta \Big]$$

What we actually use (“discounted approximation *to* the policy gradient”):

$$\nabla J_?(\theta) = \mathbb{E}\Big[\sum_{t=0}^{\infty} Q_\gamma^\theta(s,a) \frac{\partial \ln \pi^\theta(s,a)}{\partial \theta} \Big| \theta \Big]$$

- What is this?
- Is it the gradient of some other objective?
- Does it converge to something “reasonable?”
- How does it relate to the discounted and undiscounted policy gradients?

Main Result

THEOREM 4.4. *Let $\mathcal{M}$ be the set of all MDPs with rewards bounded by $[-R_{max}, R_{max}]$ satisfying $\forall \pi : \sum_{t=0}^{\infty} \Pr(S_t \neq s_\infty) < \infty$. Then, for all $\gamma < 1$:*

$$\neg \exists J_? : \forall M \in \mathcal{M} : \nabla J_?(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} Q_\gamma^\theta(S_t, A_t) \frac{\partial \ln \pi^\theta(S_t, A_t)}{\partial \theta} \middle| \theta \right].$$

It's not the gradient of any objective!

How to prove?

Clairaut-Schwarz Theorem

THEOREM 4.1. ***(Clairaut-Schwarz theorem)***: *If* $f : \mathbb{R}^n \to \mathbb{R}$ *exists and is continuously twice differentiable in some neighborhood of the point* $(a_1, a_2, \ldots, a_n)$*, then its second derivative is symmetric:*

$$\forall i, j : \frac{\partial^2 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j} = \frac{\partial^2 f(a_1, a_2, \ldots, a_n)}{\partial x_j \partial x_i}.$$

Clairaut-Schwarz Theorem

$$\frac{\partial^2 f(x,y)}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial f(x,y)}{\partial y}\right)$$

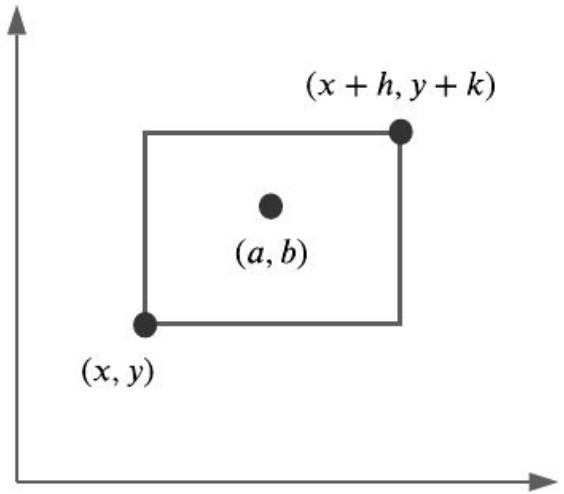
$$= \frac{\partial}{\partial x}\left(\lim_{k\to 0} \frac{f(x,y+k)-f(x,y)}{k}\right)$$

$$= \lim_{h\to 0} \frac{\lim_{k\to 0} \frac{f(x+h,y+k)-f(x+h,y)}{k} - \lim_{k\to 0} \frac{f(x,y+k)-f(x,y)}{k}}{h}$$

$$= \lim_{h\to 0} \lim_{k\to 0} \frac{f(x+h,y+k)-f(x+h,y)-f(x,y+k)+f(x,y)}{hk}$$

$$\frac{\partial^2 f(x,y)}{\partial y \partial x} = \lim_{k\to 0} \lim_{h\to 0} \frac{f(x+h,y+k)-f(x+h,y)-f(x,y+k)+f(x,y)}{hk}$$

Clairaut-Schwarz Theorem



$$\exists a, b: \quad \frac{\partial^2 f(a,b)}{\partial x \partial y} = \frac{f(x+h,y+k)-f(x+h,y)-f(x,y+k)+f(x,y)}{hk}$$

for any $\epsilon > 0$, for sufficiently small h and k:

$$\left| \frac{\partial^2 f(x,y)}{\partial x \partial y} - \frac{\partial^2 f(a,b)}{\partial x \partial y} \right| \leq \epsilon$$

$$\left| \frac{\partial^2 f(x,y)}{\partial x \partial y} - \frac{f(x+h,y+k)-f(x+h,y)-f(x,y+k)+f(x,y)}{hk} \right| \leq \epsilon$$

fix k, let $h \to 0$:

$$\left| \frac{\partial^2 f(x,y)}{\partial x \partial y} - \frac{\frac{\partial f(x,y+k)}{\partial x} - \frac{\partial f(x,y)}{\partial x}}{k} \right| \leq \epsilon$$

because ϵ is arbitrary, and above holds for all sufficiently small k, implies:

$$\frac{\partial^2 f(x,y)}{\partial x \partial y} = \frac{\partial^2 f(x,y)}{\partial y \partial x}$$

Clairaut-Schwarz Theorem

COROLLARY 4.2. ***(Contrapositive of Clairaut-Shwarz)***: *If at some point* $(a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ *there exists an* i *and* j *such that*

$$\frac{\partial^2 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j} \neq \frac{\partial^2 f(a_1, a_2, \ldots, a_n)}{\partial x_j \partial x_i},$$

then f *does not exist or is not continuously twice differentiable in any neighborhood of* $(a_1, a_2, \ldots, a_n)$.

PROOF. Contrapositive of Theorem 4.1. As a reminder, the contrapositive of a statement, $P \implies Q$, is $\neg Q \implies \neg P$. The contrapositive is always implied by the original statement. Additionally, recall that for any function g, $\neg \forall i : g(i) \implies \exists i : \neg g(i)$. □

Rewriting the Approximate Policy Gradient

LEMMA 4.3. *Let d_γ^θ be the unnormalized, weighted state distribution given by:*

$$d_\gamma^\theta(s) := d_0(s) + (1-\gamma)\sum_{t=1}^{\infty} \Pr(S_t = s|\theta).$$

Then:

$$\nabla J_?(\theta) = \sum_{s\in\mathcal{S}} d_\gamma^\theta(s)\frac{\partial}{\partial\theta}V_\gamma^\theta(s).$$

Rewriting the Approximate Policy Gradient

What's the significance of this distribution?

$$d_\gamma^\theta = d_0(s) + (1-\gamma)\sum_{t=1}^{\infty} \Pr(S_t = s|\theta)$$

It turns out that for any discount factor:

$$J_1(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} R_t \middle| \theta\right] = \sum_{s\in\mathcal{S}} d_\gamma^\theta(s) v_\gamma^\theta(s)$$

The gradient is therefore:

$$\nabla J_1(\theta) = \underbrace{\sum_{s\in\mathcal{S}} d_\gamma^\theta(s) \frac{\partial}{\partial\theta} v_\gamma^\theta(s)}_{\nabla J_\gamma(\theta)} + \sum_{s\in\mathcal{S}} v_\gamma^\theta(s) \frac{\partial}{\partial\theta} d_\gamma^\theta(s)$$

We can see that this is a "semi-gradient" of the undiscounted objective.

The discounted approximation is not a gradient!

THEOREM 4.4. *Let $\mathcal{M}$ be the set of all MDPs with rewards bounded by $[-R_{max}, R_{max}]$ satisfying $\forall \pi : \sum_{t=0}^{\infty} \Pr(S_t \neq s_\infty) < \infty$. Then, for all $\gamma < 1$:*

$$\neg \exists J_? : \forall M \in \mathcal{M} : \nabla J_?(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} Q_\gamma^\theta(S_t, A_t) \frac{\partial \ln \pi^\theta(S_t, A_t)}{\partial \theta} \middle| \theta \right].$$

PROOF. Theorem 4.1 states that if $J_?$ is a twice continuously differentiable function, then its second derivative is symmetric. It is easy to show that this is not true informally using Lemma 4.3:

$$\begin{aligned} \frac{\partial^2 J_?(\theta)}{\partial \theta_i \partial \theta_j} &= \frac{\partial}{\partial \theta_i} \left(\sum_{s \in \mathcal{S}} d_\gamma^\theta(s) \frac{\partial}{\partial \theta_j} V_\gamma^\theta(s) \right) \\ &= \underbrace{\sum_{s \in \mathcal{S}} d_\gamma^\theta(s) \frac{\partial^2}{\partial \theta_i \partial \theta_j} V_\gamma^\theta(s)}_{\text{symmetric}} + \underbrace{\sum_{s \in \mathcal{S}} \frac{\partial}{\partial \theta_i} d_\gamma^\theta(s) \frac{\partial}{\partial \theta_j} V_\gamma^\theta(s)}_{\text{asymmetric}}. \end{aligned}$$

The discounted approximation is not a gradient!

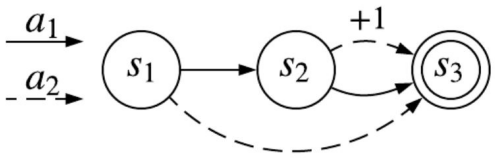


Figure 1: A counterexample wherein the derivative of $\nabla J_?(\theta)$ is asymmetric for all $\gamma < 1$, necessary for the proof of Theorem 4.4. The agent begins in s_1, and may choose between actions a_1 and a_2, each of which produces deterministic transitions. The rewards are 0 everywhere except when the agent chooses a_2 in state s_2, which produces a reward of +1. The policy is assumed to be tabular over states and actions, with one parameter, θ_1, determining the policy in s_1, and a second parameter, θ_2, determining the policy in s_2 (e.g., the policy may be determined by the standard logistic function, $\sigma(x) = \frac{1}{1+e^{-x}}$, such that $\pi(s_1, a_1) = \sigma(\theta_1)$).

$$\nabla J_?(\theta) = \sum_{s \in \mathcal{S}} d_\gamma^\theta(s) \frac{\partial}{\partial \theta} v_\gamma^\theta(s)$$

$$\frac{\partial}{\partial \theta_2}\left(\frac{\partial J_?(\theta)}{\partial \theta_1}\right) = \frac{\partial}{\partial \theta_2}\left(\gamma \sigma(\theta_2) \frac{\partial \sigma(\theta_1)}{\partial \theta_1}\right)$$

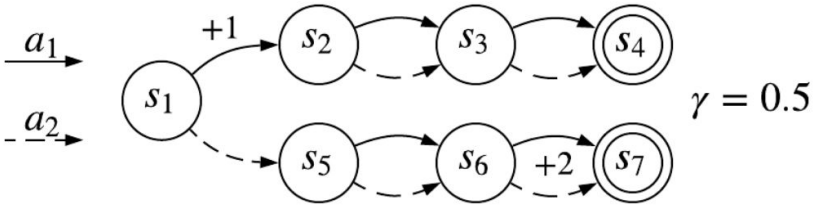
$$= \gamma \frac{\partial \sigma(\theta_1)}{\partial \theta_1} \frac{\partial \sigma(\theta_2)}{\partial \theta_2}$$

$$\frac{\partial}{\partial \theta_1}\left(\frac{\partial J_?(\theta)}{\partial \theta_2}\right) = \frac{\partial}{\partial \theta_1}\left(\sigma(\theta_1) \frac{\partial \sigma(\theta_2)}{\partial \theta_2}\right)$$

$$= \frac{\partial \sigma(\theta_1)}{\partial \theta_1} \frac{\partial \sigma(\theta_2)}{\partial \theta_2}$$

Does the discounted approximation converge to something reasonable?

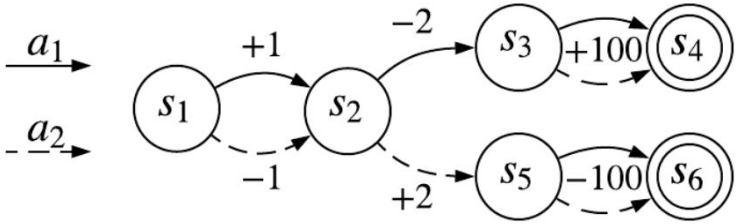
Normal problem with discounting:



Well-understood tradeoff. Can we construct a pathological case?

Does the discounted approximation converge to something reasonable?

Weirder case:



$\gamma = 0$

$\forall s \in S :$

$\pi(s, a_1) = \sigma(\theta)$

$\pi(s, a_2) = 1 - \sigma(\theta)$

Converges to “always choose a_2”

Worst possible policy under both the discounted and undiscounted objective!

Halfway between the gradient of the discounted and undiscounted objectives?

Reminder:

$$\nabla J_\gamma(\theta) = \sum_{s\in\mathcal{S}} d_0(s) \frac{\partial}{\partial\theta} v_\gamma^\theta(s)$$

$$\nabla J_?(\theta) = \sum_{s\in\mathcal{S}} d_\gamma^\theta(s) \frac{\partial}{\partial\theta} v_\gamma^\theta(s)$$

$$\nabla J_1(\theta) = \sum_{s\in\mathcal{S}} d_0(s) \frac{\partial}{\partial\theta} v_\gamma^\theta(s) + \sum_{s\in\mathcal{S}} v_\gamma^\theta(s) \frac{\partial}{\partial\theta} d_\gamma^\theta(s)$$

$$d_\gamma^\theta(s) = d_0(s) + (1-\gamma) \sum_{t=1}^{\infty} \Pr(S_t = s|\theta)$$

Expanded view:

$$\nabla J_\gamma(\theta) = \sum_{s\in\mathcal{S}} d_0(s) \frac{\partial}{\partial\theta} v_\gamma^\theta(s)$$

$$\nabla J_?(\theta) = \sum_{s\in\mathcal{S}} d_0(s) \frac{\partial}{\partial\theta} v_\gamma^\theta(s) + (1-\gamma) \sum_{s\in\mathcal{S}} \sum_{t=1}^{\infty} \Pr(S_t = s|\theta) \frac{\partial}{\partial\theta} v_\gamma^\theta(s)$$

$$\nabla J_1(\theta) = \sum_{s\in\mathcal{S}} d_0(s) \frac{\partial}{\partial\theta} v_\gamma^\theta(\theta) + (1-\gamma) \sum_{s\in\mathcal{S}} \sum_{t=1}^{\infty} \Pr(S_t = s|\theta) \frac{\partial}{\partial\theta} v_\gamma^\theta(\theta) + \sum_{s\in\mathcal{S}} v_\gamma^\theta(\theta) \frac{\partial}{\partial\theta} d_\gamma^\theta(s)$$

Future Work

- Try to better understand the effectiveness of the discounted approximation
- Quantifying the bias/variance tradeoff
- Establishing convergence when the discount factor is increased over time

Questions?