A Guided Tour of Chapter 14: Policy Gradient Algorithms

Ashwin Rao

ICME, Stanford University

Why do we care about Policy Gradient (PG)?

- Let us review how we got here
- We started with Markov Decision Processes and Bellman Equations
- Next we studied several variants of DP and RL algorithms
- We noted that the idea of *Generalized Policy Iteration* (GPI) is key
- Policy Improvement step: $\pi(s, a)$ derived from $\text{argmax}_a\, Q(s, a)$
- How do we do argmax when action space is large or continuous?
- Idea: Do Policy Improvement step with a Gradient Ascent instead

"Policy Improvement with a Gradient Ascent??"

- We want to find the Policy that fetches the "Best Expected Returns"
- Gradient Ascent on "Expected Returns" w.r.t params of Policy func
- So we need a func approx for (stochastic) Policy Func: $\pi(s, a; \boldsymbol{\theta})$
- In addition to the usual func approx for Action Value Func: $Q(s, a; \boldsymbol{w})$
- $\pi(s, a; \boldsymbol{\theta})$ called *Actor* and $Q(s, a; \boldsymbol{w})$ called *Critic*
- Critic parameters $\boldsymbol{w}$ are optimized w.r.t $Q(s, a; \boldsymbol{w})$ loss function min
- Actor parameters $\boldsymbol{\theta}$ are optimized w.r.t Expected Returns max
- We need to formally define "Expected Returns"
- But we already see that this idea is appealing for continuous actions
- GPI with Policy Improvement done as **Policy Gradient (Ascent)**

Value Function-based and Policy-based RL

- Value Function-based
  - Learn Value Function (with a function approximation)
  - Policy is implicit - readily derived from Value Function (eg: ϵ-greedy)
- Policy-based
  - Learn Policy (with a function approximation)
  - No need to learn a Value Function
- Actor-Critic
  - Learn Policy (Actor)
  - Learn Value Function (Critic)

Advantages and Disadvantages of Policy Gradient approach

Advantages:

- Finds the best *Stochastic* Policy (Optimal Deterministic Policy, produced by other RL algorithms, can be unsuitable for POMDPs)
- Naturally *explores* due to Stochastic Policy representation
- Effective in high-dimensional or continuous action spaces
- Small changes in $\boldsymbol{\theta} \Rightarrow$ small changes in π, and in state distribution
- This avoids the convergence issues seen in argmax-based algorithms

Disadvantages:

- Typically converge to a local optimum rather than a global optimum
- Policy Evaluation is typically inefficient and has high variance
- Policy Improvement happens in small steps $\Rightarrow$ slow convergence

Notation

- Assume episodic with $0 \leq \gamma \leq 1$ or non-episodic with $0 \leq \gamma < 1$
- Assume discrete-time, countable-spaces, time-homogeneous MDPs
- We lighten $\mathcal{P}(s, a, s')$ notation to $\mathcal{P}_{s,s'}^a$ and $\mathcal{R}(s, a)$ notation to $\mathcal{R}_s^a$
- Initial State Probability Distribution denoted as $p_0 : \mathcal{N} \rightarrow [0, 1]$
- Policy Function Approximation $\pi(s, a; \boldsymbol{\theta}) = \mathbb{P}[A_t = a | S_t = s; \boldsymbol{\theta}]$

PG coverage is quite similar for non-discounted non-episodic, by considering average-reward objective (we won't cover it)

"Expected Returns" Objective

Now we formalize the "Expected Returns" Objective $J(\boldsymbol{\theta})$

$$J(\boldsymbol{\theta}) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t \cdot R_{t+1}]$$

Value Function $V^\pi(s)$ and Action Value function $Q^\pi(s,a)$ defined as:

$$V^\pi(s) = \mathbb{E}_\pi[\sum_{k=t}^{\infty} \gamma^{k-t} \cdot R_{k+1} | S_t = s] \text{ for all } t = 0, 1, 2, \ldots$$

$$Q^\pi(s,a) = \mathbb{E}_\pi[\sum_{k=t}^{\infty} \gamma^{k-t} \cdot R_{k+1} | S_t = s, A_t = a] \text{ for all } t = 0, 1, 2, \ldots$$

$$\text{Advantage Function } A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$$

Also, $p(s \rightarrow s', t, \pi)$ will be a key function for us - it denotes the probability of going from state s to s' in t steps by following policy π

Discounted-Aggregate State-Visitation Measure

$$J(\boldsymbol{\theta}) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t \cdot R_{t+1}] = \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{E}_\pi[R_{t+1}]$$

$$= \sum_{t=0}^{\infty} \gamma^t \cdot \sum_{s \in \mathcal{N}} (\sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot p(S_0 \rightarrow s, t, \pi)) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot \mathcal{R}_s^a$$

$$= \sum_{s \in \mathcal{N}} (\sum_{S_0 \in \mathcal{N}} \sum_{t=0}^{\infty} \gamma^t \cdot p_0(S_0) \cdot p(S_0 \rightarrow s, t, \pi)) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot \mathcal{R}_s^a$$

Definition

$$J(\boldsymbol{\theta}) = \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot \mathcal{R}_s^a$$

where $\rho^\pi(s) = \sum_{S_0 \in \mathcal{N}} \sum_{t=0}^{\infty} \gamma^t \cdot p_0(S_0) \cdot p(S_0 \rightarrow s, t, \pi)$ is the key function (for PG) we'll refer to as *Discounted-Aggregate State-Visitation Measure*.

Policy Gradient Theorem (PGT)

Theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{s \in \mathcal{N}} \rho^{\pi}(s) \cdot \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) \cdot Q^{\pi}(s, a)$$

- Note: $\rho^{\pi}(s)$ depends on $\boldsymbol{\theta}$, but there's no $\nabla_{\boldsymbol{\theta}} \rho^{\pi}(s)$ term in $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
- Note: $\nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) = \pi(s, a; \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta})$
- So we can simply generate sampling traces, and at each time step, calculate $(\nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta})) \cdot Q^{\pi}(s, a)$ (probabilities implicit in paths)
- Note: $\nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta})$ is Score function (Gradient of log-likelihood)
- We will estimate $Q^{\pi}(s, a)$ with a function approximation $Q(s, a; \boldsymbol{w})$
- We will later show how to avoid the estimate bias of $Q(s, a; \boldsymbol{w})$
- This numerical estimate of $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ enables **Policy Gradient Ascent**
- Let us look at the score function of some canonical $\pi(s, a; \boldsymbol{\theta})$

Canonical $\pi(s, a; \boldsymbol{\theta})$ for finite action spaces

- For finite action spaces, we often use Softmax Policy
- $\boldsymbol{\theta}$ is an m-vector $(\theta_1, \ldots, \theta_m)$
- Features vector $\boldsymbol{\phi}(s, a) = (\phi_1(s, a), \ldots, \phi_m(s, a))$ for all $s \in \mathcal{N}, a \in \mathcal{A}$
- Weight actions using linear combinations of features: $\boldsymbol{\phi}(s, a)^T \cdot \boldsymbol{\theta}$
- Action probabilities proportional to exponentiated weights:

$$\pi(s, a; \boldsymbol{\theta}) = \frac{e^{\boldsymbol{\phi}(s,a)^T \cdot \boldsymbol{\theta}}}{\sum_{b \in \mathcal{A}} e^{\boldsymbol{\phi}(s,b)^T \cdot \boldsymbol{\theta}}} \text{ for all } s \in \mathcal{N}, a \in \mathcal{A}$$

- The score function is:

$$\nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta}) = \boldsymbol{\phi}(s, a) - \sum_{b \in \mathcal{A}} \pi(s, b; \boldsymbol{\theta}) \cdot \boldsymbol{\phi}(s, b) = \boldsymbol{\phi}(s, a) - \mathbb{E}_{\pi}[\boldsymbol{\phi}(s, \cdot)]$$

Canonical $\pi(s, a; \boldsymbol{\theta})$ for continuous action spaces

- For continuous action spaces, we often use Gaussian Policy
- $\boldsymbol{\theta}$ is an m-vector $(\theta_1, \ldots, \theta_m)$
- State features vector $\boldsymbol{\phi}(s) = (\phi_1(s), \ldots, \phi_m(s))$ for all $s \in \mathcal{N}$
- Gaussian Mean is a linear combination of state features $\boldsymbol{\phi}(s)^T \cdot \boldsymbol{\theta}$
- Variance may be fixed σ^2, or can also be parameterized
- Policy is Gaussian, $a \sim \mathcal{N}(\boldsymbol{\phi}(s)^T \cdot \boldsymbol{\theta}, \sigma^2)$ for all $s \in \mathcal{N}$
- The score function is:

$$\nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta}) = \frac{(a - \boldsymbol{\phi}(s)^T \cdot \boldsymbol{\theta}) \cdot \boldsymbol{\phi}(s)}{\sigma^2}$$

Proof of Policy Gradient Theorem

We begin the proof by noting that:

$$J(\boldsymbol{\theta}) = \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot V^{\pi}(S_0) = \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot Q^{\pi}(S_0, A_0)$$

Calculate $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ by parts $\pi(S_0, A_0; \boldsymbol{\theta})$ and $Q^{\pi}(S_0, A_0)$

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = & \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot Q^{\pi}(S_0, A_0) \\
& + \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} Q^{\pi}(S_0, A_0)
\end{aligned}$$

Proof of Policy Gradient Theorem

Now expand $Q^\pi(S_0, A_0)$ as:

$$\mathcal{R}_{S_0}^{A_0} + \sum_{S_1 \in \mathcal{N}} \gamma \cdot \mathcal{P}_{S_0,S_1}^{A_0} \cdot V^\pi(S_1) \text{ (Bellman Policy Equation)}$$

$$= \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot Q^\pi(S_0, A_0) +$$
$$\sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} (\mathcal{R}_{S_0}^{A_0} + \sum_{S_1 \in \mathcal{N}} \gamma \cdot \mathcal{P}_{S_0,S_1}^{A_0} \cdot V^\pi(S_1))$$

Note: $\nabla_\theta \mathcal{R}_{S_0}^{A_0} = 0$, so remove that term

$$= \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot Q^\pi(S_0, A_0) +$$
$$\sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} (\sum_{S_1 \in \mathcal{N}} \gamma \cdot \mathcal{P}_{S_0,S_1}^{A_0} \cdot V^\pi(S_1))$$

Proof of Policy Gradient Theorem

Now bring the $\nabla_{\boldsymbol{\theta}}$ inside the $\sum_{S_1 \in \mathcal{N}}$ to apply only on $V^{\pi}(S_1)$

$$= \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot Q^{\pi}(S_0, A_0) +$$
$$\sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot \sum_{S_1 \in \mathcal{N}} \gamma \cdot \mathcal{P}_{S_0, S_1}^{A_0} \cdot \nabla_{\boldsymbol{\theta}} V^{\pi}(S_1)$$

Now bring $\sum_{S_0 \in \mathcal{N}}$ and $\sum_{A_0 \in \mathcal{A}}$ inside the $\sum_{S_1 \in \mathcal{N}}$

$$= \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot Q^{\pi}(S_0, A_0) +$$
$$\sum_{S_1 \in \mathcal{N}} \sum_{S_0 \in \mathcal{N}} \gamma \cdot p_0(S_0) \cdot \Big(\sum_{A_0 \in \mathcal{A}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot \mathcal{P}_{S_0, S_1}^{A_0} \Big) \cdot \nabla_{\boldsymbol{\theta}} V^{\pi}(S_1)$$

Policy Gradient Theorem

$$\text{Note that } \sum_{A_0 \in \mathcal{A}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot \mathcal{P}_{S_0,S_1}^{A_0} = p(S_0 \rightarrow S_1, 1, \pi)$$

$$= \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot Q^\pi(S_0, A_0) +$$

$$\sum_{S_1 \in \mathcal{N}} \sum_{S_0 \in \mathcal{N}} \gamma \cdot p_0(S_0) \cdot p(S_0 \rightarrow S_1, 1, \pi) \cdot \nabla_{\boldsymbol{\theta}} V^\pi(S_1)$$

$$\text{Now expand } V^\pi(S_1) \text{ to } \sum_{A_1 \in \mathcal{A}} \pi(S_1, A_1; \boldsymbol{\theta}) \cdot Q^\pi(S_1, A_1)$$

$$= \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot Q^\pi(S_0, A_0) +$$

$$\sum_{S_1 \in \mathcal{N}} \sum_{S_0 \in \mathcal{N}} \gamma \cdot p_0(S_0) \cdot p(S_0 \rightarrow S_1, 1, \pi) \cdot \nabla_{\boldsymbol{\theta}} \Big(\sum_{A_1 \in \mathcal{A}} \pi(S_1, A_1; \boldsymbol{\theta}) \cdot Q^\pi(S_1, A_1) \Big)$$

We are now back to when we started calculating gradient of $\sum_a \pi \cdot Q^\pi$. Follow the same process of splitting $\pi \cdot Q^\pi$, then Bellman-expanding Q^π (to calculate its gradient), and iterate.

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{S_0 \in \mathcal{N}} p_0(S_0) \cdot \sum_{A_0 \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_0, A_0; \boldsymbol{\theta}) \cdot Q^\pi(S_0, A_0) +$$

$$\sum_{S_1 \in \mathcal{N}} \sum_{S_0 \in \mathcal{N}} \gamma \cdot p_0(S_0) \cdot p(S_0 \rightarrow S_1, 1, \pi) \cdot (\sum_{A_1 \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_1, A_1; \boldsymbol{\theta}) \cdot Q^\pi(S_1, A_1) + \ldots)$$

This iterative process leads us to:

$$= \sum_{t=0}^{\infty} \sum_{S_t \in \mathcal{N}} \sum_{S_0 \in \mathcal{N}} \gamma^t \cdot p_0(S_0) \cdot p(S_0 \rightarrow S_t, t, \pi) \cdot \sum_{A_t \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_t, A_t; \boldsymbol{\theta}) \cdot Q^\pi(S_t, A_t)$$

Proof of Policy Gradient Theorem

$$\text{Bring } \sum_{t=0}^{\infty} \text{ inside } \sum_{S_t \in \mathcal{N}} \sum_{S_0 \in \mathcal{N}} \text{ and note that}$$

$$\sum_{A_t \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(S_t, A_t; \boldsymbol{\theta}) \cdot Q^{\pi}(S_t, A_t) \text{ is independent of } t$$

$$= \sum_{s \in \mathcal{N}} \sum_{S_0 \in \mathcal{N}} \sum_{t=0}^{\infty} \gamma^t \cdot p_0(S_0) \cdot p(S_0 \rightarrow s, t, \pi) \cdot \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) \cdot Q^{\pi}(s, a)$$

$$\text{Reminder that } \sum_{S_0 \in \mathcal{N}} \sum_{t=0}^{\infty} \gamma^t \cdot p_0(S_0) \cdot p(S_0 \rightarrow s, t, \pi) \stackrel{\text{def}}{=} \rho^{\pi}(s). \text{ So,}$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{s \in \mathcal{N}} \rho^{\pi}(s) \cdot \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) \cdot Q^{\pi}(s, a)$$

$$\mathbb{Q}.\mathbb{E}.\mathbb{D}.$$

Monte-Carlo Policy Gradient (REINFORCE Algorithm)

- Update $\boldsymbol{\theta}$ by stochastic gradient ascent using PGT
- Using $G_t = \sum_{k=t}^{T} \gamma^{k-t} \cdot R_{k+1}$ as an unbiased sample of $Q^\pi(S_t, A_t)$

$$\Delta\boldsymbol{\theta} = \alpha \cdot \gamma^t \cdot \nabla_{\boldsymbol{\theta}} \log \pi(S_t, A_t; \boldsymbol{\theta}) \cdot G_t$$

Algorithm 0.1: $\mathrm{REINFORCE}(\cdot)$

Initialize $\boldsymbol{\theta}$ arbitrarily
for each episode $\{S_0, A_0, R_1, S_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T\} \sim \pi(\cdot, \cdot; \boldsymbol{\theta})$

$$\textbf{do} \begin{cases} \textbf{for } t \leftarrow 0 \textbf{ to } T \\ \quad \textbf{do} \begin{cases} G \leftarrow \sum_{k=t}^{T} \gamma^{k-t} \cdot R_{k+1} \\ \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \cdot \gamma^t \cdot \nabla_{\boldsymbol{\theta}} \log \pi(S_t, A_t; \boldsymbol{\theta}) \cdot G \end{cases} \end{cases}$$

Reducing Variance using a Critic

- Monte Carlo Policy Gradient has high variance
- We use a Critic $Q(s, a; \boldsymbol{w})$ to estimate $Q^\pi(s, a)$
- Actor-Critic algorithms maintain two sets of parameters:
    - Critic updates parameters $\boldsymbol{w}$ to approximate Q-function for policy π
    - Critic could use any of the algorithms we learnt earlier:
        - Monte Carlo policy evaluation
        - Temporal-Difference Learning
        - $TD(\lambda)$ based on Eligibility Traces
        - Could even use LSTD (if critic function approximation is linear)
    - Actor updates policy parameters $\boldsymbol{\theta}$ in direction suggested by Critic
    - This is Approximate Policy Gradient due to *Bias* of Critic

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) \cdot Q(s, a; \boldsymbol{w})$$

So what does the algorithm look like?

- Generate a sufficient set of sampling traces $S_0, A_0, R_1, S_1, A_1, R_2, S_2 \ldots$
- S_0 is sampled from the distribution $p_0(\cdot)$
- A_t is sampled from $\pi(S_t, \cdot; \boldsymbol{\theta})$
- Receive atomic experience (R_{t+1}, S_{t+1}) from the environment
- At each time step t, update $\boldsymbol{w}$ proportional to gradient of appropriate (MC or TD-based) loss function of $Q(s, a; \boldsymbol{w})$
- Sum $\gamma^t \cdot (\nabla_{\boldsymbol{\theta}} \log \pi(S_t, A_t; \boldsymbol{\theta})) \cdot Q(S_t, A_t; \boldsymbol{w})$ over t and over paths
- Update $\boldsymbol{\theta}$ using this (biased) estimate of $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
- Iterate with a new set of sampling traces ...

Reducing Variance with a Baseline

- We can reduce variance by subtracting a baseline function $B(s)$ from $Q(s, a; \boldsymbol{w})$ in the Policy Gradient estimate
- This means at each time step, we replace $\gamma^t \cdot \nabla_{\boldsymbol{\theta}} \log \pi(S_t, A_t; \boldsymbol{\theta}) \cdot Q(S_t, A_t; \boldsymbol{w})$ with $\gamma^t \cdot \nabla_{\boldsymbol{\theta}} \log \pi(S_t, A_t; \boldsymbol{\theta}) \cdot (Q(S_t, A_t; \boldsymbol{w}) - B(S_t))$
- Note that Baseline function $B(s)$ is only a function of s (and not a)
- This ensures that subtracting Baseline $B(s)$ does not add bias

$$
\begin{aligned}
&\sum_{s \in \mathcal{N}} \rho^\pi(s) \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) \cdot B(s) \\
=& \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot B(s) \cdot \nabla_{\boldsymbol{\theta}} (\sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta})) \\
=& \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot B(s) \cdot \nabla_{\boldsymbol{\theta}} 1 \\
=& 0
\end{aligned}
$$

Using State Value function as Baseline

- A good baseline $B(s)$ is state value function $V(s; \boldsymbol{v})$
- Rewrite Policy Gradient algorithm using advantage function estimate

$$A(s, a; \boldsymbol{w}, \boldsymbol{v}) = Q(s, a; \boldsymbol{w}) - V(s; \boldsymbol{v})$$

- Now the estimate of $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is given by:

$$\sum_{s \in \mathcal{N}} \rho^{\pi}(s) \cdot \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) \cdot A(s, a; \boldsymbol{w}, \boldsymbol{v})$$

- At each time step, we update both sets of parameters $\boldsymbol{w}$ and $\boldsymbol{v}$

TD Error as estimate of Advantage Function

- Consider TD error δ^π for the *true* Value Function $V^\pi(s)$

$$\delta^\pi = r + \gamma \cdot V^\pi(s') - V^\pi(s)$$

- δ^π is an unbiased estimate of Advantage function $A^\pi(s, a)$

$$\mathbb{E}_\pi[\delta^\pi | s, a] = \mathbb{E}_\pi[r + \gamma \cdot V^\pi(s') | s, a] - V^\pi(s) = Q^\pi(s, a) - V^\pi(s) = A^\pi(s, a)$$

- So we can write Policy Gradient in terms of $\mathbb{E}_\pi[\delta^\pi | s, a]$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) \cdot \mathbb{E}_\pi[\delta^\pi | s, a]$$

- In practice, we can use func approx for TD error (and sample):

$$\delta(s, r, s'; \boldsymbol{v}) = r + \gamma \cdot V(s'; \boldsymbol{v}) - V(s; \boldsymbol{v})$$

- This approach requires only one set of critic parameters $\boldsymbol{v}$

TD Error can be used by both Actor and Critic

Algorithm 0.2: ACTOR-CRITIC-TD-ERROR($\cdot$)

Initialize Policy params $\boldsymbol{\theta}$ and State VF params $\boldsymbol{v}$ arbitrarily
for each episode

$$
\textbf{do}
\begin{cases}
\text{Initialize } s \text{ (first state of episode)} \\
P \leftarrow 1 \\
\textbf{while } s \text{ is not terminal} \\
\quad \textbf{do}
\begin{cases}
a \sim \pi(s, \cdot; \boldsymbol{\theta}) \\
\text{Take action } a, \text{ receive } r, s' \text{ from the environment} \\
\delta \leftarrow r + \gamma \cdot V(s'; \boldsymbol{v}) - V(s; \boldsymbol{v}) \\
\boldsymbol{v} \leftarrow \boldsymbol{v} + \alpha_{\boldsymbol{v}} \cdot \delta \cdot \nabla_{\boldsymbol{v}} V(s; \boldsymbol{v}) \\
\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \cdot P \cdot \delta \cdot \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta}) \\
P \leftarrow \gamma \cdot P \\
s \leftarrow s'
\end{cases}
\end{cases}
$$

Using Eligibility Traces for both Actor and Critic

Algorithm 0.3: ACTOR-CRITIC-ELIGIBILITY-TRACES$(\cdot)$

Initialize Policy params $\boldsymbol{\theta}$ and State VF params $\boldsymbol{v}$ arbitrarily

for each episode

do $\begin{cases} \text{Initialize } s \text{ (first state of episode)} \\ \boldsymbol{z_\theta}, \boldsymbol{z_v} \leftarrow 0 \text{ (eligibility traces for } \boldsymbol{\theta} \text{ and } \boldsymbol{v}) \\ P \leftarrow 1 \\ \textbf{while } s \text{ is not terminal} \\ \quad \textbf{do} \begin{cases} a \sim \pi(s, \cdot; \theta) \\ \text{Take action } a, \text{ observe } r, s' \\ \delta \leftarrow r + \gamma \cdot V(s'; \boldsymbol{v}) - V(s; \boldsymbol{v}) \\ \boldsymbol{z_v} \leftarrow \gamma \cdot \lambda_{\boldsymbol{v}} \cdot \boldsymbol{z_v} + \nabla_{\boldsymbol{v}} V(s; \boldsymbol{v}) \\ \boldsymbol{z_\theta} \leftarrow \gamma \cdot \lambda_{\boldsymbol{\theta}} \cdot \boldsymbol{z_\theta} + P \cdot \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta}) \\ \boldsymbol{v} \leftarrow \boldsymbol{v} + \alpha_{\boldsymbol{v}} \cdot \delta \cdot \boldsymbol{z_v} \\ \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \cdot \delta \cdot \boldsymbol{z_\theta} \\ P \leftarrow \gamma \cdot P, s \leftarrow s' \end{cases} \end{cases}$

Overcoming Bias

- We've learnt a few ways of how to reduce variance
- But we haven't discussed how to overcome bias
- All of the following substitutes for $Q^\pi(s,a)$ in PG have bias:
  - $Q(s,a;\boldsymbol{w})$
  - $A(s,a;\boldsymbol{w},\boldsymbol{v})$
  - $\delta(s,s',r;\boldsymbol{v})$
- Turns out there is indeed a way to overcome bias
- It is called the *Compatible Function Approximation Theorem*

Compatible Function Approximation Theorem

Theorem

Let $\boldsymbol{w}_{\boldsymbol{\theta}}^*$ *denote the Critic parameters* $\boldsymbol{w}$ *that minimize the following mean-squared-error for given policy parameters* $\boldsymbol{\theta}$*:*

$$\sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot (Q^\pi(s, a) - Q(s, a; \boldsymbol{w}))^2$$

Assume that the data type of $\boldsymbol{\theta}$ *is the same as the data type of* $\boldsymbol{w}$ *and furthermore, assume that for any policy parameters* $\boldsymbol{\theta}$*, the Critic gradient at* $\boldsymbol{w}_{\boldsymbol{\theta}}^*$ *is* compatible *with the Actor score function, i.e.,*

$$\nabla_{\boldsymbol{w}} Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*) = \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta}) \textit{ for all } s \in \mathcal{N}, \textit{ for all } a \in \mathcal{A}$$

Then the Policy Gradient using critic $Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*)$ *is exact:*

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) \cdot Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*)$$

Proof of Compatible Function Approximation Theorem

For a given $\boldsymbol{\theta}$, since $\boldsymbol{w}_{\boldsymbol{\theta}}^*$ minimizes the mean-squared-error as defined above, we have:

$$\sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot (Q^\pi(s, a) - Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*)) \cdot \nabla_{\boldsymbol{w}} Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*) = 0$$

But since $\nabla_{\boldsymbol{w}} Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*) = \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta})$, we have:

$$\sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot (Q^\pi(s, a) - Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*)) \cdot \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta}) = 0$$

Therefore,

$$\begin{aligned} & \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot Q^\pi(s, a) \cdot \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta}) \\ = & \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*) \cdot \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta}) \end{aligned}$$

Proof of Compatible Function Approximation Theorem

$$\text{But } \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot Q^\pi(s, a) \cdot \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta})$$

$$\begin{aligned}
\text{So, } \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*) \cdot \nabla_{\boldsymbol{\theta}} \log \pi(s, a; \boldsymbol{\theta}) \\
&= \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi(s, a; \boldsymbol{\theta}) \cdot Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*)
\end{aligned}$$

This means with conditions of Compatible Function Approximation Theorem, we can use the critic func approx $Q(s, a; \boldsymbol{w}_{\boldsymbol{\theta}}^*)$ and still have the exact Policy Gradient.

How to enable Compatible Function Approximation

A simple way to enable Compatible Function Approximation $\frac{\partial Q(s,a;\boldsymbol{w}^*_{\boldsymbol{\theta}})}{\partial w_i} = \frac{\partial \log \pi(s,a;\boldsymbol{\theta})}{\partial \theta_i}, \forall i$ is to set $Q(s,a;\boldsymbol{w})$ to be linear in its features.

$$Q(s,a;\boldsymbol{w}) = \sum_{i=1}^{m} \phi_i(s,a) \cdot w_i = \sum_{i=1}^{m} \frac{\partial \log \pi(s,a;\boldsymbol{\theta})}{\partial \theta_i} \cdot w_i$$

We note below that a compatible $Q(s,a;\boldsymbol{w})$ serves as an approximation of the advantage function.

$$\sum_{a \in \mathcal{A}} \pi(s,a;\boldsymbol{\theta}) \cdot Q(s,a;\boldsymbol{w}) = \sum_{a \in \mathcal{A}} \pi(s,a;\boldsymbol{\theta}) \cdot (\sum_{i=1}^{m} \frac{\partial \log \pi(s,a;\boldsymbol{\theta})}{\partial \theta_i} \cdot w_i)$$

$$= \sum_{a \in \mathcal{A}} (\sum_{i=1}^{m} \frac{\partial \pi(s,a;\boldsymbol{\theta})}{\partial \theta_i} \cdot w_i) = \sum_{i=1}^{m} (\sum_{a \in \mathcal{A}} \frac{\partial \pi(s,a;\boldsymbol{\theta})}{\partial \theta_i}) \cdot w_i$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \theta_i} (\sum_{a \in \mathcal{A}} \pi(s,a;\boldsymbol{\theta})) \cdot w_i = \sum_{i=1}^{m} \frac{\partial 1}{\partial \theta_i} \cdot w_i = 0$$

Fisher Information Matrix

Denoting $[\frac{\partial \log \pi(s,a;\boldsymbol{\theta})}{\partial \theta_i}], i = 1, \ldots, m$ as the score column vector $\boldsymbol{SC}(s,a;\boldsymbol{\theta})$ and assuming compatible linear-approximation critic:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s,a;\boldsymbol{\theta}) \cdot \boldsymbol{SC}(s,a;\boldsymbol{\theta}) \cdot (\boldsymbol{SC}(s,a;\boldsymbol{\theta})^T \cdot \boldsymbol{w}_{\boldsymbol{\theta}}^*) \\
&= \sum_{s \in \mathcal{N}} \rho^\pi(s) \cdot \sum_{a \in \mathcal{A}} \pi(s,a;\boldsymbol{\theta}) \cdot (\boldsymbol{SC}(s,a;\boldsymbol{\theta}) \cdot \boldsymbol{SC}(s,a;\boldsymbol{\theta})^T) \cdot \boldsymbol{w}_{\boldsymbol{\theta}}^* \\
&= \mathbb{E}_{s \sim \rho^\pi, a \sim \pi}[\boldsymbol{SC}(s,a;\boldsymbol{\theta}) \cdot \boldsymbol{SC}(s,a;\boldsymbol{\theta})^T] \cdot \boldsymbol{w}_{\boldsymbol{\theta}}^* \\
&= FIM_{\rho^\pi,\pi}(\boldsymbol{\theta}) \cdot \boldsymbol{w}_{\boldsymbol{\theta}}^*
\end{aligned}$$

where $FIM_{\rho_\pi,\pi}(\boldsymbol{\theta})$ is the Fisher Information Matrix w.r.t. $s \sim \rho^\pi, a \sim \pi$.
Hence, updates after each atomic experience are as follows:

$$\Delta \boldsymbol{\theta} = \alpha_{\boldsymbol{\theta}} \cdot \gamma^t \cdot \boldsymbol{SC}(S_t, A_t; \boldsymbol{w}) \cdot \boldsymbol{SC}(S_t, A_t; \boldsymbol{w})^T \cdot \boldsymbol{w}$$

$$\Delta \boldsymbol{w} = \alpha_{\boldsymbol{w}} \cdot (R_{t+1} + \gamma \cdot \boldsymbol{SC}(S_{t+1}, A_{t+1}; \boldsymbol{\theta})^T \cdot \boldsymbol{w} - \boldsymbol{SC}(S_t, A_t; \boldsymbol{\theta})^T \cdot \boldsymbol{w}) \cdot \boldsymbol{SC}(S_t, A_t; \boldsymbol{\theta}$$

Natural Policy Gradient (NPG)

- Natural gradient $\nabla_{\boldsymbol{\theta}}^{nat} J(\boldsymbol{\theta})$ is the direction of optimal $\boldsymbol{\theta}$ movement
- In terms of the KL-divergence metric (versus plain Euclidean norm)
- Formally defined as:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = FIM_{\rho_\pi,\pi}(\boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}}^{nat} J(\boldsymbol{\theta})$$

- Enabling Compatible Function Approximation implies:

$$\nabla_{\boldsymbol{\theta}}^{nat} J(\boldsymbol{\theta}) = \boldsymbol{w}_{\boldsymbol{\theta}}^*$$

- **This compact result is great for our algorithm:**
  - Update Critic params $\boldsymbol{w}$ with the critic loss gradient (at step t) as:

$$(R_{t+1} + \gamma \cdot \boldsymbol{SC}(S_{t+1}, A_{t+1}; \boldsymbol{\theta})^T \cdot \boldsymbol{w} - \boldsymbol{SC}(S_t, A_t; \boldsymbol{\theta})^T \cdot \boldsymbol{w}) \cdot \boldsymbol{SC}(S_t, A_t; \boldsymbol{\theta})$$

  - Update Actor params $\boldsymbol{\theta}$ in the direction of $\boldsymbol{w}$

Deterministic Policy Gradient (DPG)

- Function approximation for deterministic policy for continuous actions
- DPG expressed as Expected Gradient of Q-Value
- Integrates only over state space, so efficient for high-dim action spaces
- Usual machinery of PG is applicable to DPG
- Intuition: Instead of greedy policy improvement for continuous action spaces, move policy in the direction of gradient of Q-Value Function
- Policy parameters $\boldsymbol{\theta}$ are updated in proportion to $\nabla_{\boldsymbol{\theta}} Q(s, \pi_D(s; \boldsymbol{\theta}))$
- Average direction of policy improvements is given by:

$$\mathbb{E}_{s \sim \rho^{\pi_D}}[\nabla_{\boldsymbol{\theta}} Q(s, \pi_D(s; \boldsymbol{\theta}))] = \mathbb{E}_{s \sim \rho^{\pi_D}}[\nabla_{\boldsymbol{\theta}} \pi_D(s; \boldsymbol{\theta}) \cdot \nabla_a Q^{\pi_D}(s, a)\Big|_{a=\pi_D(s;\boldsymbol{\theta})}]$$

$$\rho^{\pi_D}(s) = \sum_{S_0 \in \mathcal{N}} \sum_{t=0}^{\infty} \gamma^t \cdot p_0(S_0) \cdot p(S_0 \rightarrow s, t, \pi_D)$$

- For multi-dimensional a, $\nabla_{\boldsymbol{\theta}} \pi(s; \boldsymbol{\theta})$ is a Jacobian matrix

DPG Theorem

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_D}[\sum_{t=0}^{\infty} \gamma^t \cdot R_{t+1}] = \sum_{s \in \mathcal{N}} \rho^{\pi_D}(s) \cdot \mathcal{R}_s^{\pi_D(s;\boldsymbol{\theta})} = \mathbb{E}_{s \sim \rho^{\pi_D}}[\mathcal{R}_s^{\pi_D(s;\boldsymbol{\theta})}]$$

Theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{s \in \mathcal{N}} \rho^{\pi_D}(s) \cdot \nabla_{\boldsymbol{\theta}} \pi_D(s; \boldsymbol{\theta}) \cdot \nabla_a Q^{\pi_D}(s, a) \Big|_{a=\pi_D(s;\boldsymbol{\theta})}$$

$$= \mathbb{E}_{s \sim \rho^{\pi_D}}[\nabla_{\boldsymbol{\theta}} \pi_D(s; \boldsymbol{\theta}) \cdot \nabla_a Q^{\pi_D}(s, a) \Big|_{a=\pi_D(s;\boldsymbol{\theta})}]$$

- Since π_D (target policy) is deterministic, explore with behavior policy
- Actor and Critic parameters are updated after each atomic experience:

$$\Delta \boldsymbol{w} \propto (R_{t+1} + \gamma \cdot Q(S_{t+1}, \pi_D(S_{t+1}; \boldsymbol{\theta}); \boldsymbol{w}) - Q(S_t, A_t; \boldsymbol{w})) \cdot \nabla_{\boldsymbol{w}} Q(S_t, A_t; \boldsymbol{w}$$

$$\Delta \boldsymbol{\theta} \propto \nabla_{\boldsymbol{\theta}} \pi_D(S_t; \boldsymbol{\theta}) \cdot \nabla_a Q(S_t, a; \boldsymbol{w}) \Big|_{a=\pi_D(S_t;\boldsymbol{\theta})}$$

Introduction to Evolutionary Strategies

- Evolutionary Strategies (ES) are a type of Black-Box Optimization
- Popularized in the 1970s as *Heuristic Search Methods*
- Loosely inspired by natural evolution of living beings
- We focus on a subclass called Natural Evolution Strategies (NES)
- The original setting was generic and nothing to do with MDPs or RL
- Given an objective function $F(\psi)$, where ψ refers to parameters
- We consider a probability distribution $p_\theta(\psi)$ over ψ
- Where θ refers to the parameters of the probability distribution
- We want to maximize the average objective $\mathbb{E}_{\psi \sim p_\theta}[F(\psi)]$
- We search for optimal θ with stochastic gradient ascent as follows:

$$\nabla_\theta(\mathbb{E}_{\psi \sim p_\theta}[F(\psi)]) = \nabla_\theta(\int_\psi p_\theta(\psi) \cdot F(\psi) \cdot d\psi)$$

$$= \int_\psi \nabla_\theta(p_\theta(\psi)) \cdot F(\psi) \cdot d\psi = \int_\psi p_\theta(\psi) \cdot \nabla_\theta(\log p_\theta(\psi)) \cdot F(\psi) \cdot d\psi$$

$$= \mathbb{E}_{\psi \sim p_\theta}[\nabla_\theta(\log p_\theta(\psi)) \cdot F(\psi)]$$

NES applied to solving Markov Decision Processes (MDPs)

- We set $F(\cdot)$ to be the (stochastic) *Return* of an MDP
- ψ refers to the parameters of a policy $\pi_\psi : \mathcal{S} \rightarrow \mathcal{A}$
- ψ will be drawn from an isotropic multivariate Gaussian distribution
- Gaussian with mean vector θ and fixed diagonal covariance matrix $\sigma^2 I$
- The average objective (*Expected Return*) can then be written as:

$$\mathbb{E}_{\psi \sim p_\theta}[F(\psi)] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)}[F(\theta + \sigma \cdot \epsilon)]$$

- The gradient (∇_θ) of *Expected Return* can be written as:

$$\mathbb{E}_{\psi \sim p_\theta}[\nabla_\theta(\log p_\theta(\psi)) \cdot F(\psi)]$$

$$= \mathbb{E}_{\psi \sim \mathcal{N}(\theta, \sigma^2 I)}[\nabla_\theta(\frac{-(\psi - \theta)^T \cdot (\psi - \theta)}{2\sigma^2}) \cdot F(\psi)]$$

$$= \frac{1}{\sigma} \cdot \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)}[\epsilon \cdot F(\theta + \sigma \cdot \epsilon)]$$

A sampling-based algorithm to solve the MDP

- The above formula helps estimate gradient of *Expected Return*
- By sampling several ϵ (each ϵ represents a *Policy* $\pi_{\theta+\sigma\cdot\epsilon}$)
- And averaging $\epsilon \cdot F(\theta + \sigma \cdot \epsilon)$ across a large set (n) of ϵ samples
- Note $F(\theta + \sigma \cdot \epsilon)$ involves playing an episode for a given sampled ϵ, and obtaining that episode's *Return* $F(\theta + \sigma \cdot \epsilon)$
- Hence, n values of ϵ, n *Policies* $\pi_{\theta+\sigma\cdot\epsilon}$, and n *Returns* $F(\theta + \sigma \cdot \epsilon)$
- Given gradient estimate, we update θ in this gradient direction
- Which in turn leads to new samples of ϵ (new set of *Policies* $\pi_{\theta+\sigma\cdot\epsilon}$)
- And the process repeats until $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[F(\theta + \sigma \cdot \epsilon)]$ is maximized
- The key inputs to the algorithm will be:
  - Learning rate (SGD Step Size) α
  - Standard Deviation σ
  - Initial value of parameter vector θ_0

The Algorithm

Algorithm 0.4: NATURAL EVOLUTION STRATEGIES$(\alpha, \sigma, \theta_0)$

for $t \leftarrow 0, 1, 2, \ldots$

do $\begin{cases} \text{Sample } \epsilon_1, \epsilon_2, \ldots \epsilon_n \sim \mathcal{N}(0, I) \\ \text{Compute Returns } F_i \leftarrow F(\theta_t + \sigma \cdot \epsilon_i) \text{ for } i = 1, 2, \ldots, n \\ \theta_{t+1} \leftarrow \theta_t + \frac{\alpha}{n\sigma} \sum_{i=1}^n \epsilon_i \cdot F_i \end{cases}$

Resemblance to Policy Gradient?

- On the surface, this NES algorithm looks like Policy Gradient (PG)
- Because it's not Value Function-based (it's Policy-based, like PG)
- Also, similar to PG, it uses a gradient to move towards optimality
- But, ES does not interact with the environment (like PG/RL does)
- ES operates at a high-level, ignoring (state,action,reward) interplay
- Specifically, does not aim to assign credit to actions in specific states
- Hence, ES doesn't have the core essence of RL: *Estimating the Q-Value Function of a Policy and using it to Improve the Policy*
- Therefore, we don't classify ES as Reinforcement Learning
- We consider ES to be an alternative approach to RL Algorithms

ES versus RL

- Traditional view has been that ES won't work on high-dim problems
- Specifically, ES has been shown to be data-inefficient relative to RL
- Because ES resembles simple hill-climbing based only on finite differences along a few random directions at each step
- However, ES is very simple to implement (no Value Function approx. or back-propagation needed), and is highly parallelizable
- ES has the benefits of being indifferent to distribution of rewards and to action frequency, and is tolerant of long horizons
- This paper from OpenAI Researchers shows techniques to make NES more robust and more data-efficient, and they demonstrate that NES has more exploratory behavior than advanced PG algorithms
- I'd always recommend trying NES before attempting to solve with RL

Key Takeaways from this Chapter

- PG Algorithms are based on GPI with Policy Improvement as a Stochastic Gradient Ascent for "Expected Returns" Objective $J(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are parameters of the function approximation for the Policy
- Policy Gradient Theorem gives us a simple formula for $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ in terms of the score of the policy function approximation
- We can reduce variance in PG algorithms by using a critic and by using an estimate of the advantage function for the Q-Value Function
- Compatible Function Approximation Theorem enables us to overcome bias in PG Algorithms
- Natural Policy Gradient and Deterministic Policy Gradient are specialized PG algorithms that have worked well in practice
- Evolutionary Strategies are technically not RL, but they resemble PG Algorithms and can sometimes be quite effective for MDP Control