Recall:

Discounted state visitation:

$$d_{s_0}^{\pi}(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr[s_t = s \mid s_0]$$

- To sample from this distribution.

  Start at s_0 & simulate π; Accept a state with prob $1-\gamma$;

prob of accepting s at time t?

don't accept at $t=0, t=1, \ldots, t=t-1$, accept at t $= (1-\gamma)\gamma^t \Pr[s_t = s \mid s_0]$.

$\therefore$ the state s is distributed as $d_{s_0}^{\pi}(s)$!

If τ is a trajectory,

the unconditional distribution $\Pr_{\mu}^{\pi}(\tau)$

under π starting with initial distribution μ is

$$\mu(s_0)\pi(a_0|s_0)P(s_1|s_0 a_0)\ldots$$

Notation: $$d_{\mu}^{\pi}(s) = \mathbb{E}_{s_0 \sim \mu}\left[d_{s_0}^{\pi}(s)\right].$$

Given $f: S \times A \to \mathbb{R}$;

$$\mathbb{E}_{\tau \sim \Pr^{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right]$$

Expected discounted value of f along the trajectory

$$\overset{?}{=} \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} f(a,s).$$

In terms of sampling,

$$\mathbb{E}_{\tau \sim \Pr^{\pi}} \left[\sum \gamma^t f(s_t, a_t) \right] \text{ is :}$$

Sample a trajectory; compute the discounted value of f over the trajectory; Find the expected value;

- Run the markov chain. with prob $1-\gamma$ select state, and then an action $(\pi(\cdot|s))$ and compute $\dfrac{f(s,a)}{1-\gamma}$

prob of picking s in time t?

$$= (1-\gamma)\gamma^t \Pr[s_t = s | s_0]$$

Pro of picking action a $\pi(\cdot|s)$

$$\therefore \text{ get } \frac{\cancel{(1-\gamma)}\gamma^t \Pr[s_t = s|s_0] \cdot \pi(\cdot|s)}{\cancel{1-\gamma}} f(s,a)$$

Regroup this sum and get

$$\mathbb{E}_{\tau \sim \Pr_\mu^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t) \right]$$

$$\|$$

$$\frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_\mu}} \mathbb{E}_{a \sim \pi(\cdot|s)} f(s,a)$$

Lemma: For all policies π, π', and initial dist μ,

$$V^{\pi}(\mu) - V^{\pi'}(\mu) = \mathbb{E}_{\tau \sim \Pr_\mu^{\pi}} \left[\sum \gamma^t A^{\pi'}(s_t, a_t) \right]$$

advantage of a policy: $A^{\pi}(s,a) := \underbrace{Q^{\pi}(s,a) - V^{\pi}(s)}$

For a fixed state s,

$\mathbb{P}_s^{\pi}$ denote distribution over trajectories with $s_0 = s$;

$$V^{\pi}(s) - V^{\pi'}(s)$$

$$= (1-\gamma) \mathop{\mathbb{E}}_{\tau \sim \mathbb{P}_s^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] - V^{\pi'}(s)$$

$$= \mathop{\mathbb{E}}_{\tau \sim \mathbb{P}_s^{\pi}} \left[(1-\gamma) \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) + \sum \gamma^t V^{\pi'}(s_t) - \sum \gamma^t V^{\pi'}(s_t) \right]$$

$$- V^{\pi'}(s)$$

$$= \mathop{\mathbb{E}}_{\tau \sim \mathbb{P}_s^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \left[(1-\gamma) r(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t) \right] \right]$$

$$= \mathop{\mathbb{E}}_{\tau \sim \mathbb{P}_s^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \left[(1-\gamma) r(s_t, a_t) + \gamma \mathbb{E} \left[V^{\pi'}(s_{t+1}) \middle| s_t, a_t \right] - V^{\pi'}(s_t) \right] \right]$$

$$\left(\mathop{\mathbb{E}}_{s_t, a_t, s_{t+1}, \ldots} V^{\pi}(s_{t+1}) = \mathop{\mathbb{E}}_{s_t, a_t} \mathop{\mathbb{E}}_{s_{t+1}} \left[V^{\pi}(s_{t+1}) \middle| s_t, a_t \right] \right)$$

$$= \mathop{\mathbb{E}}_{\tau \sim \Pr^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \left(Q^{\pi'}(s_t, a_t) - V^{\pi'}(s_t) \right) \right]$$

$$= \mathop{\mathbb{E}}_{\tau \sim \Pr^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \, A^{\pi'}(s_t, a_t) \right].$$

$=$

Policy gradient:

Discounted total reward of a trajectory.

$$R(\tau) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

Now: $$V^{\pi_\theta}(\mu) = \mathop{\mathbb{E}}_{\tau \sim \Pr_{\mu}^{\pi_\theta}} [R(\tau)]$$

$=$

THM: POLICY GRADIENT'S:

$$\nabla V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}} \left[R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t | s_t) \right]$$

in terms of discounted rewards

REINFORCE:

$$\nabla V^{\pi_\theta}(\mu) = \nabla \mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}} (R(\tau))$$

Q: $\int$?

Countability issues.

$$= \nabla \sum_\tau R(\tau) \Pr_\mu^{\pi_\theta}(\tau)$$

$$= \sum_\tau R(\tau) \nabla \Pr_\mu^{\pi_\theta}(\tau)$$

$$= \sum_\tau R(\tau) \Pr_\mu^{\pi_\theta}(\tau) \nabla \log \Pr_\mu^{\pi_\theta}(\tau)$$

$$= \sum_\tau R(\tau) \Pr_\mu^{\pi_\theta}(\tau) \nabla \log \left[\mu(s_0) \pi_\theta(a_0|s_0) P(s_0, a_0, s_1) \dots \right]$$

$$= \sum_\tau R(\tau) \Pr_\mu^{\pi_\theta}(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t|s_t)$$

$$= \mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}} R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t|s_t)$$

$$\nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s \sim d^{\pi_\theta}} \mathop{\mathbb{E}}_{a \sim \pi(\cdot|s)} \left[Q^{\pi_\theta}(s,a) \nabla \log \pi_\theta(a|s) \right]$$

$$\|$$

$$\mathop{\mathbb{E}}_{\tau \sim \mathbb{P}r_\mu^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \underline{\underline{\nabla \log \pi_\theta}}(s_t | a_t) \right]$$

In terms of action value

An unbiased estimate!

- $$\nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s \sim d^{\pi_\theta}} \mathop{\mathbb{E}}_{a \sim \pi_\theta(\cdot|s)} \left[A^{\pi_\theta}(s,a) \nabla \log \pi_\theta(a|s) \right]$$

In terms of advantage

= **Gradient ascent:**

$f : \mathbb{R}^d \to \mathbb{R}$ is β smooth if

$$\|\nabla f(w) - \nabla f(w')\| \leq \beta \|w - w'\|$$

<u>Update:</u> $\theta_{t+1} = \theta_t + \eta \nabla V^{\pi_\theta}(\mu)$.

Notation

$V^{(t)} = V^{\pi_{\theta_t}}$

$\pi^{(t)} := \pi_{\theta_t}$.

<u>Lemma:</u> Assume V^{π_θ} is β smooth $\forall \theta$, Assume V^{π_θ} is bounded below by V_*, Using $\eta = 1/\beta$, $\forall T$

$$\min_{t \le T} \| \nabla V^{(t)}(\mu) \| \le \frac{2\beta \left(V^*(\mu) - V^0(\mu) \right)}{T}.$$

When

$$\boxed{T \ge \frac{1}{\varepsilon} 2\beta \left[V^*(\mu) - V^0(\mu) \right],}$$

one of the gradients is $\le \varepsilon$, (ie) close to a stationary point then!

Unbiased estimate of gradients:

Assume we have samples, $\tau \sim \mathbb{P}_\mu^{\pi_\theta}$;

For a trajectory τ define:

$$\widehat{Q}^{\pi_\theta}(s_t, a_t) := (1-\gamma) \sum_{t'=t} \gamma^{t'-t} r(s_{t'}, a_{t'})$$

Discounted Gain from state t onwards

$$\widehat{\nabla V^{\pi_\theta}(\mu)} := \sum_{t=0}^{\infty} \gamma^t \widehat{Q}^{\pi_\theta}(s_t, a_t) \nabla \log \pi_\theta(a_t | s_t)$$

Claim: $\mathbb{E}_{\tau \sim \mathbb{P}_\gamma^{\pi_\theta}} \left[\widehat{\nabla V^{\pi_\theta}(\mu)} \right] = \nabla V^{\pi_\theta}(\mu)$

unbiased estimator;

Proof:

$$\mathbb{E}_{\tau \sim \mathbb{P}_\gamma^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t \widehat{Q}^{\pi_\theta}(s_t, a_t) \nabla \log \pi_\theta(a_t | s_t) \right]$$

$$= \mathbb{E}_{\tau \sim \mathbb{P}_\theta^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t \underbrace{\mathbb{E}\left[\widehat{Q}^{\pi_\theta}(s_t, a_t) \,\middle|\, s_t, a_t \right]}_{= Q(s_t, a_t)} \nabla \log (\ ) \right]$$

$$= \mathbb{E}_{\tau \sim \mathbb{P}_\theta^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t Q(s_t, a_t) \nabla \log (a_t | s_t) \right] = \nabla V^{\pi_\theta}(\mu)$$

Note $\widehat{Q^{\pi_\theta}(s_t, a_t)}$ is an unbiased estimator

of $Q(s_t, a_t)$.

Gives us:

Initialize θ_0:

2 For $t = 0, 1, \ldots$

(a) Sample $\tau \sim \Pr_\mu^{\pi_\theta}$

(b) $\theta_{t+1} = \theta_t + \eta_t \widehat{\nabla V^{\pi_\theta}(\mu)}$

- To get $\widehat{\nabla V^{\pi_\theta}(\mu)}$ from the sample

compute $\widehat{Q^{\pi_\theta}(s_t, a_t)}$ for t, and use in

$\widehat{\nabla^{\pi_\theta}(\mu)}$.

Ignore that τ is ∞; Truncate

Algorithm: Stochastic Gradient Ascent on J.

REINFORCE.

- Initialize θ arbitrarily.
- for each episode do

Generate $S_0 A_0 R_0 S_1 A_1 \cdots S_{L-1} A_{L-1} R_{L-1}$

using θ.

For each t compute $G_t = \overbrace{Q^{\pi_\theta}(s_t, a_t)}$

$$\widehat{\nabla J(\theta)} = \sum_{t=0}^{L-1} \gamma^t G_t \frac{\partial \ln \pi(A_t \mid S_t ; \theta)}{\partial \theta}$$

$$\theta = \theta + \alpha \widehat{\nabla J(\theta)}$$

end

- Evidently γ^t is ignored in practice
- Maybe this is stochastic gradient for a different obj

The estimate of $\overbrace{Q^{\pi_\theta}(s_t, a_t)}$ has high variance.

Aside:

Estimate $\mathbb{E}(X)$.

- a single sample $x_0 \to$ estimate: x_0.

Problem: high variance;

Suppose we take a sample of another Y, whose expectation $\mathbb{E}(Y)$ we know.

Try $\hat{\mu} = X - Y + E(Y)$;

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - \mathrm{Cov}(X, Y)$$

If $\mathrm{Var}(Y) < \mathrm{Cov}(X, Y)$ — in business

- Y is called a control variate

Want: X, Y positively correlated;

FOR REINFORCE:

$$\theta \leftarrow \theta + \alpha \sum_{t=0}^{L-1} \gamma^t \left(\overbrace{Q^{\pi_\theta}(s_t, a_t)} - \underline{f(s_t)} \right) \nabla \log \pi_\theta (a_t | s_t)$$

Here: $f: S \to \mathbb{R}$ is a function indep of τ a traj

For any function $g(s)$,

$$\mathbb{E}\left[\nabla \log \pi(a|s)\, g(s) \right]$$

$$= \sum_a \pi(a|s) \nabla \log(\pi(a|s))\, g(s)$$

$$= \sum_a \frac{\cancel{\pi(a|s)}}{\cancel{\pi(a|s)}} \nabla \pi(a|s)\, g(s)$$

$$= g(s) \sum_a \nabla \pi(a|s) = g(s) \nabla \sum_a \pi(a|s)$$

$$= g(s) \nabla (1) = \underline{\underline{0}}.$$

Now if $f(\cdot)$ is independent of τ, $\forall t$.

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t) \nabla \log \pi_\theta(a_t|s_t)\right] = 0$$

$$\parallel$$

$$\sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}\Big[\underbrace{f(s_t) \nabla \log \pi_\theta (a_t|s_t)}_{\substack{\parallel \\ 0}}\Big]$$

- $f(s)$ called baseline at state s.

<u>What do we use for f?</u>

Common choice is v^θ;

· More MC algorithms for $v^{\pi}(\cdot)$

- $\mathbb{P}_{s}^{\pi} \leftarrow$ the distribution of trajectories starting at s_0.

Recall: Generate episodes using π;

for every state s, t_s be the first time s appears in episode;

$$G_s \leftarrow \sum_{k=0}^{\infty} \gamma^k r(s_{t_s+k}, a_{t_s+k})$$

discounted reward from then on

Append G_s to returns(s);

For each s, return avg$($returns$(s))$.

Gradient based MC: update at time t.

$$\boxed{v(S_t) \leftarrow v(S_t) + \alpha\left(G_t - V(S_t)\right)}$$

- This is minimizing mean squared value error;

$$\frac{1}{2}\mathbb{E}_s\left[\left(v^{\pi}(s) - v(s)\right)^2\right].$$

- Move against the gradient of the above loss.

$$v \leftarrow v - \alpha \cdot \frac{\partial}{\partial v}\mathbb{E}_s\left[(\ )^2\right]$$

$$\leftarrow v - \alpha \cdot \mathbb{E}\left(v^{\pi}(s) - v(s)\right)(-1)\frac{\partial v(s)}{\partial v}$$

$$= v + \alpha\ \mathbb{E}\left[\left(v(s) - v^{\pi}(s)\right)\frac{\partial v(s)}{\partial v}\right]$$

- Don't know $v^{\pi}(s)$.

So use the discounted reward from first visit to state s in the sample trajectory.

Unbiased estimator of $v^{\pi}(s)$;

- <u>Temporal difference learning:</u>

- policy evaluation algorithm
- we don't know P & R;

- Get samples & learn from experience (as in MC algorithms)

- In TD we estimate V^{π};

<u>TD update:</u>

- if in state s, we take action a, go to s' & get reward r

$$v(s) \leftarrow v(s) + \alpha\left(r + v(s') - v(s)\right)$$