Policy Gradient

Problem Setup: • want to optimize $J(\pi) := \mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_t \mid \pi\right]$

• Large state space. Parametrized π_θ ←

example: $|A| = 2$. given state features $\phi(s)$.

$$\pi_\theta(s) = \begin{cases} -1, & \phi(s)^T\theta > 0 \\ 1, & \phi(s)^T\theta \le 0. \end{cases}$$

deterministic policy. below we mostly deal w/ stoch policy.

• $\max_\theta J(\pi_\theta)$. (S)GD: calculate/estimate $\nabla_\theta J(\pi_\theta)$

• Central Question: How to estimate $\nabla_\theta J(\pi_\theta)$.

w/o knowing MDP. but using data?

• YES! We can form an unbiased estimate of $\nabla_\theta J(\pi_\theta)\big|_{\theta=\theta_0}$

given an on-policy trajectory.

Data: $s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_H, a_H, r_H$.

where $\forall t, a_t \sim \pi_\theta$.

Goal: derive a func of this traj. whose expectation $= \nabla_\theta J(\pi_\theta)$.

Derivation. Notations & assumptions.

1. $\tau = (s_1, a_1, s_2, a_2, s_3, a_3, \cdots, s_H, a_H)$.
2. Assume wlog. deterministic reward.
3. Finite state/action spaces.
4. $R(\tau) = \sum_{t=1}^{H} \gamma^{t-1} R(s_t, a_t)$.

$$\nabla_\theta J(\pi_\theta) = \nabla_\theta \mathbb{E}[R(\tau) \mid \pi_\theta]$$

$$= \nabla_\theta \Big(\sum_\tau \underbrace{P^{\pi_\theta}(\tau)}_{\text{depends on } \theta} \cdot \underbrace{R(\tau)}_{\text{constant!}} \Big)$$

$\sum_\tau (\cdots)$ ↳ enumerates over all $|S \cdot A|^H$ trajectories.

$$= \sum_\tau \nabla_\theta \big(P^{\pi_\theta}(\tau) \cdot R(\tau) \big) = \sum_\tau R(\tau) \cdot \nabla_\theta P^{\pi_\theta}(\tau).$$

$$= \sum_\tau R(\tau) \cdot P^{\pi_\theta}(\tau) \cdot \nabla_\theta \log P^{\pi_\theta}(\tau).$$

$$(\log f(x))' = \frac{1}{f(x)} \cdot f'(x)$$

$$f'(x) = f(x) \cdot (\log f(x))'$$

$$P^{\pi_\theta}(\tau) = d_0(s_1) \cdot \pi_\theta(a_1 \mid s_1) \cdot P(s_2 \mid s_1, a_1) \cdot \pi_\theta(a_2 \mid s_2) \cdot \cdots \cdot \pi_\theta(a_H \mid s_H)$$

$$= \sum_\tau R(\tau) \cdot P^{\pi_\theta}(\tau) \cdot \nabla_\theta \Big(\cancel{\log d_0(s_1)} + \log \pi_\theta(a_1 \mid s_1) + \cancel{\log P(s_2 \mid s_1, a_1)} + \cdots + \log \pi_\theta(a_H \mid s_H) \Big).$$

$$= \sum_\tau P^{\pi_\theta}(\tau) \Big(R(\tau) \cdot \sum_{t=1}^{H} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \Big)$$

$$= \mathbb{E}\Big[\qquad \Big| \pi_\theta \Big].$$

Naive PG estimator (REINFORCE)

$$R(\tau) \cdot \sum_{t=1}^{H} \nabla_\theta \log \pi_\theta(a_t \mid s_t).$$

Remarks: (1) Similarity w/ ZS derivations — coincidence?

$$\nabla_\theta J(\pi_\theta) = \lim_{\Delta\theta \to 0} \frac{J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)}{\Delta\theta}$$

w/ traj. from π_θ. can estimate

- $J(\pi_\theta)$: on-policy MC.
- $J(\pi_{\theta+\Delta\theta})$: per-traj. importance sampling

} alternative proof of REINFORCE.

- step-wise IS?

Alternatively: $\nabla_\theta J(\pi_\theta) = \nabla_\theta \sum_\tau P^{\pi_\theta}(\tau) \cdot R(\tau)$.

improved ver: $\nabla_\theta J(\pi_\theta) = \nabla_\theta \sum_{t=1}^{H} \gamma^{t-1} \mathbb{E}[r_t \mid \pi_\theta]$.

$$= \sum_{t=1}^{H} \gamma^{t-1} \nabla_\theta \mathbb{E}[r_t \mid \pi_\theta].$$

$$= \sum_{t=1}^{H} \gamma^{t-1} \nabla_\theta \sum_{\tau_{1:t}} \underline{P^{\pi_\theta}(\tau_{1:t})} \cdot r_t$$

$\tau_{1:t}$ is the t-step prefix of τ.
$s_1, a_1, s_2, a_2, \dots, s_t, a_t$.

$= \dots$ (leave as exercise)

"vanilla PG".

$$\boxed{= \mathbb{E}\left[\sum_{t=1}^{H} \left(\sum_{t'=t}^{H} \gamma^{t'-1} r_{t'}\right) \cdot \nabla \log \pi_\theta(a_t \mid s_t) \,\middle|\, \pi_\theta\right].}$$

- "Actor-critic" form of PG

$$\sum_{t=1}^{H} \gamma^{t-1} \mathbb{E}\left[\left(\sum_{t'=t}^{H} \gamma^{t'-t} r_{t'}\right) \cdot \nabla \log \pi_\theta(a_t \mid s_t) \,\middle|\, \pi_\theta\right]$$

$$= \sum_{t=1}^{H} \gamma^{t-1} \mathbb{E}\Big[\underbrace{\mathbb{E}[\,\text{- - - - - - -}\mid \pi_\theta, s_t, a_t]}_{?}\Big]$$

function of s_t, a_t.

$$\nabla \log \pi_\theta(a_t|s_t) \underbrace{\mathbb{E}\left[\sum_{t'=t}^{H} \gamma^{t'-t} r_{t'} \,\middle|\, s_t, a_t \right]}_{Q^\pi(s_t, a_t).}$$

therefore: $$\nabla_\theta J(\pi_\theta) = \sum_{t=1}^{H} \gamma^{t-1} \mathbb{E}_{s_t, a_t \sim \pi_\theta} \left[Q^\pi(s_t, a_t) \cdot \nabla \log \pi_\theta(a_t|s_t) \right]$$

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

- REINFORCE / Vanilla PG is on-policy.

  Actor-Critic: $\nabla_\theta J(\pi_\theta) \approx \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim d^{\pi_\theta}} \left[\hat{Q}^\pi(s,a) \cdot \nabla \log \pi_\theta(a|s) \right]$

  - $\hat{Q}^\pi$ can be estimated separately via e.g. TD. (and can be off-policy.
  - still require on-policy data to approximate $\mathbb{E}_{s,a \sim d^{\pi_\theta}}[\cdot]$

- In general π_θ has to be stochastic.

Parametrized Stochastic policy.

A popular example: linear + softmax.

- Assume we know $\phi: S \times A \to \mathbb{R}^d$.
- $\pi_\theta(a|s) \propto e^{\theta^T \phi(s,a)}$ $\Longleftrightarrow$ $\pi(a|s) = \boxed{\dfrac{e^{\theta^T \phi(s,a)}}{\sum_{a'} e^{\theta^T \phi(s,a')}}}$.

  "proportional to"

$$\nabla_\theta \log \pi_\theta(a|s) = \nabla \log \frac{e^{\theta^T \phi(s,a)}}{\sum_{a'} e^{\theta^T \phi(s,a')}}$$

$$= \nabla \left(\underline{\log} \, \underline{e}^{\theta^T \phi(s,a)} - \log \sum_{a'} (\cdot\cdot) \right) = \nabla(\theta^T \phi(s,a)) - \nabla\left(\log \sum_{a'} (\cdot\cdot)\right)$$

$$= \phi(s,a) - \nabla \log \left(\sum_{a'} e^{\theta^T \phi(s,a')} \right)$$

$$= \underline{\phi(s,a)} - \underbrace{\frac{\sum_{a'} e^{\theta^T \phi(s,a')} \cdot \phi(s,a')}{\sum_{a'} e^{\theta^T \phi(s,a')}}}_{\mathbb{E}_{a' \sim \pi_\theta(\cdot|s)}[\phi(s,a')]}$$

Remark:

- Compare to softmax policy in on-policy control w/ SARSA.
  w/ linear func. approx of $Q^\pi \approx \theta^T \phi(s,a)$
  softmax policy: $\pi(a|s) \propto e^{\theta^T \phi(s,a)/T}$;
- Why T in SARSA and not here?
  → in SARSA: if $\theta^{*T} \phi(s,a) = Q^\pi(s,a)$
  then can't use $2\theta^*$
  → in PG: can arbitrarily rescale θ.
- When linear softmax policies can represent π^*?
  $\pi_\theta(a|s) \propto e^{\boxed{\phi^T(s,a)\,\theta}}$ → doesn't have to be a value func.

to represent π^* $\begin{cases} ① \text{ Assume } Q^* = \phi^T(s,a) \cdot \theta^* . \\ ② \text{ Take } \underline{\theta = c \cdot \theta^*} \text{ w/ large } c. \end{cases}$

in the limit of $c \to +\infty \Rightarrow \pi_\theta \to \pi^*$.

· Variance reduction in PG. (w/o introducing bias).

$$\nabla J(\pi_\theta) = \frac{1}{1-\gamma} \mathbb{E}_{\underline{\underline{(s,a)\sim d^{\pi_\theta}}}} \left[\left(Q^{\pi_\theta}(s,a) - f(s) \right) \nabla \log \pi_\theta(a|s) \right].$$

"state baseline".

Unbiasedness: fixing arbitrary state s.

$$\underline{\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}} \left[\underline{f(s)} \cdot \nabla \log \pi_\theta(a|s) \right].$$

$$= f(s) \cdot \underbrace{\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla \log \pi_\theta(a|s) \right]}$$

$$= \sum_a \cancel{\pi_\theta(a|s)} \frac{\nabla \pi_\theta(a|s)}{\cancel{\pi_\theta(a|s)}}$$

$$= \nabla \sum_a \pi_\theta(a|s) = \nabla(1) = \vec{0}.$$

A special case: $f(s) := V^\pi(s)$.

"advantage".

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