Policy Gradient Method

$J(\theta)$, $\pi : S \to a$ instead we will use

$\pi_\theta : S \to p(a|s)$ (stochastic policy)

$$\nabla_\theta J = \nabla_\theta E_{p(\xi|\theta)} [R(\xi)]$$

$$\left[\xi_e := (s_0, a_0, s_1, a_1 \text{ ----- })\right]$$

$$= \nabla_\theta \sum_J p(\xi_e|\theta) R(\xi_e)$$

[where J is the space of all possible trajectories]

[Note that $p(\xi_e|\theta) = p(s_0) \cdot \pi(a_0|s_0) \cdot p(s_1|a_0,s_0)$
$\cdot \pi(a_1|s_1) \cdot p(s_2|s_1,a_1)$ ---]

$$= \sum_J \nabla_\theta p(\xi_e|\theta) \cdot R(\xi_e) \quad (\text{as } \nabla_\theta R(\xi) = 0)$$

$$= \sum_J \frac{p(\xi_e|\theta) \nabla_\theta p(\xi_e|\theta) \cdot R(\xi_e)}{p(\xi_e|\theta)}$$

$$= E_{p(\xi_e|\theta)} \left[\frac{\nabla_\theta p(\xi_e|\theta) \cdot R(\xi_e)}{p(\xi_e|\theta)} \right]$$

$$= \mathbb{E}_{p(\xi_e|\theta)} \nabla_\theta \log p(\xi_e|\theta) \cdot R(\xi_e)$$

$$= \mathbb{E}_{p(\xi|\theta)} \Big[\nabla_\theta \Big[\sum_t \log p(s_t|s_{t-1}, a_{t-1}) \Big]$$

($\log p(s_t|s_{t-1}, a_{t-1})$: 0 (NOT dependent on θ))

$$+ \sum_t \log \pi(a_t|s_t) \Big] R(\xi_e) \Big]$$

($\log \pi(a_t|s_t)$: ↑ (Dependent on θ))

$$\nabla_\theta J(\theta) = \mathbb{E}_{p(\xi_e|\theta)} \Big[\sum_t \nabla_\theta \log \pi(a_t|s_t) \cdot R(\xi_e) \Big]$$

Policy Gradient Theorem.

Reinforce (Version 1)

(1) Run simulator n times with π_θ

(2) $$\tilde{\nabla} J = \frac{1}{N} \sum_{i=1}^{N} \left[\left[\sum_{t}^{T} \nabla \log \pi(a_t^i | s_t^i) \right] R(\xi_e^i) \right]$$

(3) $\theta_{new} \leftarrow \theta_{old} + \alpha \tilde{\nabla} J.$

Reinforce (Version 2)

Rewriting step (2). what happened to other terms?

$$\tilde{\nabla} J = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\nabla_\theta \log(\pi(a_t|s_t)) \sum_{t=t+1}^{T} r(s_t) \right]$$

$$\Downarrow$$

$$Q^\pi(s_t, a_t)$$

(in expectation)

$\Downarrow$

can get this by regression on rollout states and hence lower variance.

(better than a bunch of rollouts as lower variance and leverage what they've

So $\tilde{\nabla} J = \frac{1}{N} \sum_{i=1}^{N} \left[\sum_t \nabla_\theta \log \left(\pi_\theta(a_t^i | s_t^i) \right) \cdot \tilde{Q}^\pi(s_t^i, a_t^i) \right]$

Advantages over discrete gradient.

Here we understand policy as function of θ.

It has extra causal information.

Downsides

1. Have to have stochastic policy.
2. Has to be differentiable.
3. Be able to differentiate.
4. Step-size parameter.

You can rewrite (2) in REINFORCE as

$$\nabla J = \mathbb{E}_{P(s|\pi_\theta)} \mathbb{E}_{\pi(a|s)} \left[\nabla \log \pi(a|s) \cdot Q^\pi(s,a) \right]$$