Lecture-36* (contd)

Policy gradient methods (PG methods)

The main idea behind PG methods is the "likelihood ratio" trick.

An illustration of this trick in a very simple setting.

Let X be a r.v. with mass function $p(\theta, \cdot)$ → parameter

i.e., $p(X = x)$ is parameterized by θ.

$$J(\theta) = E f(X)$$

Goal: $\min_\theta J(\theta)$

Want to find best θ using a gradient method

$$\theta_{t+1} = \theta_t - \beta_t \nabla J(\theta_t)$$

Need: "$\nabla J(\theta)$"

Use "likelihood ratio" trick, which is shown below.

$$J(\theta) = \sum_x f(x)\, p(\theta, x) \quad \leftarrow \text{LOTUS}$$

$$\nabla J(\theta) = \nabla \left(\sum_x f(x)\, p(\theta, x) \right)$$

need a few regularity conditions to justify interchange of $\sum$ & ∇

$$= \sum_x f(x)\, \nabla p(\theta, x)$$

mild regularity conditions → usually let one invoke "dominated

Convergence theorem

$$\nabla J(\theta) = \sum_x f(x) \nabla p(\theta, x)$$

$$= \sum_x \left(f(x) \frac{\nabla p(\theta,x)}{p(\theta,x)} \right) p(\theta,x)$$

$$\nabla J(\theta) = E\left(f \frac{\nabla p}{p} \right) = E(f \nabla \log p)$$

To get an estimate of $\nabla J(\theta)$, I can sample from

$\frac{\nabla p(\theta,x)}{p(\theta,x)}$ → likelihood ratio.

$\nabla \log p(\theta) = \frac{\nabla p(\theta)}{p(\theta)}$

Connecting likelihood ratio to RL:

Fix an SSP problem with state space $\mathcal{X}$, action space $\mathcal{A}$.

Π_{det} : class of admissible stationary deterministic policies

$\{ \pi : \pi : \mathcal{X} \to \mathcal{A}$ & it is time invariant$\}$

$x \to$ [unparameterized policy] $\to \pi(x) \to$ an action

π: states → actions

for PG methods, we consider stationary randomized policies, i.e.,

$\pi : \mathcal{X} \to \Delta(\mathcal{A})$ → set of all distributions over the actions.

For simplicity, assume all actions available in all states.

e.g. $$\pi_\theta(x,a) = \frac{\exp(h(\theta,x,a))}{\sum_b \exp(h(\theta,x,b))}$$ → randomized policy

$$\pi_\theta(x) = [\pi_\theta(x,a), \forall a \in \mathcal{A}]$$

π_θ → distribution over $\mathcal{A}$.

Simple example for h: $h(\theta,x,a) = \theta^T \phi(x,a)$

$$\pi_\theta(x,a) = \frac{\exp(\theta^T \phi(x,a))}{\sum_b \exp(\theta^T \phi(x,b))}$$ } Boltzmann distribution aka soft-max

Assumption A1: Policy π_θ is a continuously differentiable function of θ and $\nabla \log \pi_\theta$ exists.

Note: Every π_θ is identified by its parameter $\theta \in \mathbb{R}^d$

Goal: $\min_{\theta \in \Theta} J_{\pi_\theta}(x^0)$ → Find an approximately optimal policy in the class of parameterized policies

& we want to find the best parameter in a class

$$\{\pi_\theta \mid \theta \in \Theta\}$$ e.g. $\Theta \subset \mathbb{R}^d$

Want to find a $\theta^* \in \operatorname{argmin}_{\theta \in \Theta} J_{\pi_\theta}(x^0)$ ← fixed start state

Lecture 35*

Today: An expression for $\nabla_\theta J_{\pi_\theta}(x^0)$

So that we can do

$$\theta_{t+1} = \theta_t - \beta_t \hat{\nabla} J(\theta_t)$$

$\hat{\nabla} J(\theta_t)$ ↓ estimate of $\nabla_\theta J_{\pi_\theta}(x^0)$

Stochastic gradient algorithm

$\min_\theta f(\theta)$, $\theta_{t+1} = \theta_t - \beta_t f'(\theta_t)$

$\theta_t \to \theta^*$ as $t \to \infty$, $f'(\theta^*) = 0$

With policy gradients, we would want to find a θ^* s.t. $\nabla J_{\theta^*}(x^0) = 0$

Catch! We need to know $\nabla J_{\theta^*}(x^0)$

Usual RL settings, closed form of $\nabla_\theta J$ is not available & has to be estimated from samples".

For a deterministic policy: $Q_\pi(x,a) = g(x,a) + \sum_{x'} P_{xa}(x') J_\pi(x')$

$J_\pi(x) = Q_\pi(x, \pi(x))$

With randomized policies: $J_\pi(x) = \sum_a \pi(x,a) Q_\pi(x,a)$

↳ need to average over actions since π is a distribution

$$Q_\pi(x,a) = g(x,a) + \sum_{x'} P_{x,a}(x')\, J_\pi(x')$$

Policy gradient theorem: (for SSPs)

Start with

$$J_{\pi_\theta}(x^0) = \sum_a \pi(x^0,a)\, Q_\pi(x^0,a)$$

$$\nabla_\theta J_{\pi_\theta}(x^0) = \nabla\left(\sum_a \pi_\theta(x^0,a)\, Q_{\pi_\theta}(x^0,a)\right)$$

$$= \sum_a \left(\nabla \pi_\theta(x^0,a)\, Q_{\pi_\theta}(x^0,a) + \pi_\theta(x^0,a)\nabla Q_{\pi_\theta}(x^0,a)\right)$$

$$= \sum_a \left(\nabla \pi_\theta(x^0,a)\, Q_{\pi_\theta}(x^0,a) + \pi_\theta(x^0,a)\nabla_\theta\left(g(x^0,a) + \sum_{x'} P_{x,a}(x')\, J_{\pi_\theta}(x')\right)\right)$$

$g(x^0,a)$: does not depend on θ

$P_{x,a}(x')$: does not depend on θ

$$= \sum_a \left[\nabla \pi_\theta(x^0,a)\, Q_{\pi_\theta}(x^0,a) + \pi_\theta(x^0,a) \sum_{x'} P_{x,a}(x')\, \nabla J_{\pi_\theta}(x')\right]$$

$$= \sum_a \Big[\nabla \pi_\theta(x^0,a)\, Q_{\pi_\theta}(x^0,a) +$$

$$\pi_\theta(x^0,a) \sum_{x'} P_{x,a}(x') \left(\sum_{a'} \nabla \pi_\theta(x',a')\, Q_{\pi_\theta}(x',a') + \pi(x',a') \sum_{x''} P_{x',a'}(x'')\, \nabla J_{\pi_\theta}(x'')\right)\Big]$$

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} P\left(x_k = x \mid x^0, \pi_\theta\right) \sum_a \nabla \pi_\theta(x,a)\, Q_{\pi_\theta}(x,a)$$

$P(x_k = x \mid x^0, \pi_\theta)$: 1

Prob of going from x^0 to x in k steps while following the policy π_θ

$$\nabla_\theta J_{\pi_\theta}(x^0) = \sum_{x \in \mathcal{X}} \sum_{k=0}^{\infty} P(x_k = x \mid x^0, \pi_\theta) \sum_a \nabla \pi_\theta(x,a)\, Q_{\pi_\theta}(x,a)$$

Policy gradient theorem.

occupancy measure (can be normalized)

With some abuse of notation, the policy gradient thm is written as

$$\nabla_\theta J_{\pi_\theta}(x^0) = \sum_x \underbrace{d^{\pi_\theta}(x)}_{= \sum_{k=0}^{\infty} P(x_k = x \mid x^0, \pi_\theta)} \sum_a \nabla \pi_\theta(x,a)\, Q_{\pi_\theta}(x,a)$$

$$= E_\pi \left(\sum_a \nabla \pi_\theta(X,a)\, Q_{\pi_\theta}(X,a) \right)$$

X is a r.v. governed by a distribution that uses d^{π_θ}

$\nabla \pi_\theta$ is available in closed form (we did the parametrization)

Q_{π_θ} → estimated using a sample path

PG update: $\theta_{t+1} = \theta_t - \beta_t \sum_a \nabla \pi_{\theta_t}(X,a)\, \hat{Q}_{\pi_\theta}(X,a)$

$\hat{Q}_{\pi_\theta}$ → estimate of Q_{π_θ}

$$\nabla_\theta J_{\pi_\theta}(x^0) = E_{\pi_\theta}\left(\sum_a \nabla \pi(X,a)\, Q_\pi(X,a) \right)$$

$$= E_{\pi_\theta}\left(\sum_a \pi(X,a)\, Q_\pi(X,a)\, \frac{\nabla \pi(X,a)}{\pi(X,a)} \right)$$

$$= E_{\pi_\theta}\left(\sum_a \pi(X,a)\, Q_\pi(X,a)\, \nabla \log \pi(X,a) \right)$$

Suppose A is a r.v. chosen using the distribution π

$$= E_{\pi_\theta}\left(Q_\pi(X,A)\, \nabla \log \pi(X,A) \right)$$

both X, A are random

X is chosen from a distribution over the occupancy measure

& A is chosen using $\pi(X,\cdot)$

$\sum_{i=1}^{100} p(i)\, h(i)$ → pmf $P(X=i)$

$= E(h(X))$

Suppose $E(\hat{Q}) = Q_\pi$. Then,

$$\theta_{t+1} = \theta_t - \beta_t \hat{Q}_t \nabla \log \pi_{\theta_t}(X_t, A_t)$$

REINFORCE algorithm

A more naive version:

Fix π_{θ_t}, Simulate an episode $(X_0, A_0, \dots, X_T)$ (terminal)

Collect a sample of the total cost from this episode, call it $\hat{Q}_t$ → Check $\hat{Q}_t$ is an unbiased estimate of Q_π

$$\theta_{t+1} = \theta_t - \beta_t \hat{Q}_t \nabla \log \pi_{\theta_t}(X_0, A_0)$$

$A_0 \sim \pi(X_0, \cdot)$

Aside! Want to estimate $Q_\pi(x,a) = \mathbb{E}\left(\sum g(x_t, a_t) \,\middle|\, x_0 = x, a_0 = a\right)$

total cost in SSP starting in x & taking action a

Estimation: $(x_0, a_0, \_\_\_\_\_)$

Follow π until termination.

Collect total cost sample, say C_t

$$\mathbb{E}(C_t) = Q_\pi(x,a)$$

Can be extended to cover an action A that is chosen from a random policy $\pi(x, \cdot)$

Remark: In REINFORCE, policy evaluation (= estimating $Q^{\pi_\theta}(\cdot,\cdot)$) is "Monte Carlo".

Instead, if we use a parametric approximation for Q_π, say $Q_{\pi_\theta}(x,a) \simeq r^T \phi(x,a)$

↑ linear func. approx.

Can we TD with LFA to obtain an estimate of $Q_{\pi_\theta}(\cdot,\cdot)$.

Suppose TD converges to r^*.

$$Q_\pi(x,a) \simeq r^{*T} \phi(x,a)$$

$$\mathbb{E}\left(r^{*T}\phi(x,a)\right) \neq Q_\pi(x,a) \;??$$

"Parametric approximation induce a bias"

On the other hand, MC methods usually have large variance.

Can do policy gradient with function approximation

↓

Actor-Critic algorithms.

$$\theta_{t+1} = \theta_t - \beta_t \underbrace{\hat{\nabla} J_{\theta_t}(x^0)}_{\downarrow}$$

$$\underbrace{\hat{Q}_{\pi_\theta}(x_0, A_0)}_{\downarrow} \cdot \nabla \log \pi(x_0, A_0)$$

$$= r^{*T} \phi(x_0, A_0)$$

How to get r^*?

Fix policy π_{θ_t}

Obtain a sample path using π_{θ_t}

Run TD using this sample path $\rightarrow$ $\hat{r}$ $\rightarrow$ estimate of r^*

Use $\hat{r}$ & do gradient descent in policy parameter.