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.

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

e.g. $\theta \in \mathbb{R}^d$

Goal: $\min_{\theta} J(\theta)$ → min among a class of parameterized r.v.s

Want to find best θ using a gradient method

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

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

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

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

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

need a few regularity conditions to justify interchange of Σ & ∇

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

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

$\frac{\nabla p(\theta,x)}{p(\theta,x)} \rightarrow$ 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

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

$x \rightarrow$ $\boxed{\text{Unparameterized policy}}$ $\rightarrow \pi(x) \rightarrow$ an action

π: states, actions

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

$$\pi : \mathcal{X} \rightarrow \Delta(\mathcal{A})$$

$\rightarrow$ set of all distributions over the actions.

For simplicity, assume all actions available in all states.

e.g. $\pi_\theta(x,a) = \dfrac{\exp(h(\theta,x,a))}{\sum_b \exp(h(\theta,x,b))}$ → randomized policy parameterized by "θ"

is the probability of choosing action a in state x

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

→ distribution over $\mathcal{A}$.

Simple example for h: $h(\theta,x,a) = \theta^T \phi(x,a)$ ($\theta \in \mathbb{R}^d$, $\phi(x,a) \in \mathbb{R}^d$)

state-action features

$$\pi_\theta(x,a) = \frac{\exp(\theta^T \phi(x,a))}{\sum_b \exp(\theta^T \phi(x,b))}$$

} Boltzmann distribution aka soft-max

<u>Assumption A1</u>: 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 $\theta \in \mathbb{R}^d$

$$\boxed{\text{Goal: } \min_{\theta \in \Theta} J_{\pi_\theta}(x^0)}$$

→ 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 $\theta^* \in \arg\min_{\theta \in \Theta} J_{\pi_\theta}(x^0)$

→ fix start state

"J_π" is not necessarily a convex function of parameter θ

So, a gradient algo in policy space converges to local optima.

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

So that we can do

$$\theta_{t+1} = \theta_t - \beta_t \underbrace{\hat{\nabla} J(\theta_t)}_{\text{estimate of } \nabla_\theta J_{\pi_\theta}(x^0)}$$

Policy gradient update iteration

Stochastic gradient algorithm

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

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

With policy gradients, we would not 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

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

<u>Policy gradient theorem</u>: (for SSPs)

Start with

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

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

$$= \sum_a \left(\nabla \pi_\theta(x^0,a)\, Q_{\pi_\theta}(x^0,a) + \pi_\theta(x^0,a)\, \nabla Q_{\pi_\theta}(x^0,a)\right) \quad \text{— } 1$$

expanded

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

does not depend on θ

$$\nabla_\theta J_{\pi_\theta}(x^0)$$

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

initial state

next state

$x_0 \to$ pos. prob $\pi_\theta(x_0,a)$ pos. prob $P_{x_0,a}(x')$ x'

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

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

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

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

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

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)

$Q_{\pi_\theta} \rightarrow$ estimated using a sample path

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

$\hat{Q}_{\pi_\theta} \rightarrow$ 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 $\pi(X,\cdot)$

$$\boxed{\sum_{i=1}^{100} p(i)\, h(i) = E(h(X))}$$

$p \rightarrow$ pmf $P(X=i)$

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

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

A more naive version:

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

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 $Q_\pi(x,a) = E\left(\sum g(x_t, a_t) \middle| x_0 = x, a_0 = a\right)$

total cost in SSP starting in x & taking action a

Estimation: $(\overset{x}{\overset{\shortparallel}{x_0}}, \overset{a}{\overset{\shortparallel}{a_0}}, \underline{\qquad\qquad})$

Follow π until termination.

Collect total cost sample, say C_t

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

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

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

Estimating

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

linear func. approx.

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

Suppose TD converges to r^*.

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

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

"Parametric approximation induce a bias"

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

Can do policy gradient with function approximation

$\downarrow$

Actor-Critic algorithms.

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

$$\downarrow$$

$$\underbrace{\hat{Q}_{\pi_\theta}(X_0, A_0)} \cdot \nabla \log \pi(X_0, A_0)$$

$$\downarrow$$

$$= r^{*T} \phi(X_0, A_0)$$

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.