Wednesday, November 18, 2020

Lecture 25

CS59000_RL

Model free policy learning

Agenda:

- policy gradient

Intuition: M, Fixed horizon

Given a policy π

$$\eta(\pi) = E_\pi\left[\sum_{h=1}^{H} r_h\right] \Rightarrow \psi = \sum_{h=1}^{H} r_h$$

$$\tau = (X_1, A_1, r_1, \dots X_H, A_H, r_H)$$

$$\eta(\pi) = \sum \Pr(\tau;\pi)\psi_\tau \quad ; \psi_\tau \Rightarrow \text{reward in trajectory } \tau.$$

Let assume reward is deterministic.

$\Rightarrow r(x,a)$

$$\Rightarrow \tau = (X_1, A_1, X_2, A_2 \sim -)$$

$$\Rightarrow \psi(\tau) = \sum_{h=1}^{A} r(X_h, A_h)$$

$$\rightarrow \Pr(\tau;\pi) = P_1(X_1)\ \pi(A_1;X_1)P(X_2|X_1,A_1) \dots \pi(A_H;X_H)$$

$$\eta(\Pi) = \sum_{\tau} P_r(\tau;\Pi)\,\Psi(\tau)$$

$$\Rightarrow \nabla_\Pi \eta(\Pi) = \nabla_\Pi \left(\sum_{\tau} P_r(\tau;\Pi)\,\Psi(\tau) \right)$$

$$= \sum_{\tau} \nabla_\Pi P_r(\tau;\Pi)\,\Psi(\tau)$$

$\nabla \log x = \frac{\nabla x}{x}$

$$= \sum_{\tau} P_r(\tau;\Pi)\,\nabla_\Pi \log P_r(\tau;\Pi)\,\Psi(\tau)$$

$$= E\left[\nabla_\Pi \log P_r(\tau;\Pi)\,\Psi(\tau) \right]$$

$$\log P_r(\tau;\Pi) = \log \rho_1(X_1) + \log \Pi(A_1;X_1) + \log P(X_2|X_1,A_1) + \cdots + \log \Pi(A_H;X_H)$$

$$\Rightarrow \nabla_\Pi \log P_r(\tau;\Pi) = \sum_{h=1}^{H} \log \Pi(A_h;X_h)$$

$$\Rightarrow \nabla_\Pi \eta(\Pi) = E\left[\nabla_\Pi \left(\sum \log \Pi(A_h;X_h) \right) \Psi(\tau) \right]$$

Monte Carlo sampling

$$\Rightarrow \nabla_\Pi \hat{\eta}(\Pi) = \frac{1}{N} \sum_{i=1}^{N} \left(\nabla_\Pi \left(\sum_{h=1}^{H} \log \Pi(A_h^i;X_h^i) \right) \Psi(\tau^i) \right)$$

Wednesday, November 18, 2020

Consider an MDP $M: (\mathcal{X}, A, P, P_1, R)$

- $\mathcal{X}$ as a Borel space (state space)
- A as a Borel space (Action space)
- P_1 a probability measure on the initial state.
- $\bar{P}$ a transition kernel
- R a reward kernel, $\bar{r}(x,a)$ is mean $\forall x,y \in \mathcal{X} \times A$

Consider a set of parameters Θ such that each $\theta \in \Theta$ parametrizes a policy kernel $\bar{\pi}_\theta$ from $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ to $(A, \mathcal{B}(A))$, such that $d\bar{\pi}_\theta = \pi_\theta \, da$.

$\Rightarrow$ (we consider $\mathcal{X}$ and A as subsets of Euclidean spaces)

- Consider discounted setting. (Ignoring a.s.)

$$V_{\pi_\theta} = E_{\pi_\theta}\left[\sum_{t \geq 1} \gamma^{t-1} r_t \,\middle|\, \alpha(x_1)\right]$$

$$Q_{\pi_\theta} = E_{\pi_\theta}\left[\sum_{t \geq 1} \gamma^{t-1} r_t \,\middle|\, \alpha(x_1, A_1)\right]$$

For a given x

$$\Rightarrow V_{\pi_\theta}(x) = E_{\pi_\theta}\left[Q(x_1, A_1) \,|\, \alpha(x_1)\right](x)$$

$$= \int_A \pi_\theta(a; x) \, Q_{\pi_\theta}(a; x) \, da$$

Therefore,

$$\nabla_\theta V_{\pi_\theta}(x) = \nabla_\theta \int_A \pi_\theta(a;x) Q_{\pi_\theta}(a;x)\, da$$

$$= \int_A \nabla_\theta \pi_\theta(a;x) Q_{\pi_\theta}(a;x)\, da + \int_A \pi_\theta(a;x) \nabla_\theta Q_{\pi_\theta}(a;x)\, da$$

$$\nabla_\theta Q_{\pi_\theta}(a;x) = \nabla_\theta \left[\bar{r}(x,a) + \gamma \int P(x';x,a) V_{\pi_\theta}(x')\, dx' \right]$$

(we used the fact that $d\bar{P}(x';x,a) = P(x';x,a)\, dx', \forall x,a,x' \in \mathcal{X}\times\mathcal{A}\times\mathcal{X}$)

$$\Rightarrow \nabla_\theta V_{\pi_\theta}(x) = \int_A \nabla_\theta \pi_\theta(a;x) Q_{\pi_\theta}(a;x)\, da + \gamma \int_A \pi_\theta(a;x) \left(\int_X P(x';x,a) \nabla_\theta V_{\pi_\theta}(x')\, dx' \right) da$$

Let's define occupancy kernel ; for $S \in \mathcal{B}(\mathcal{X})$

$$\bar{\mu}^{\gamma}_{\pi_\theta}(S;x) = \lim_{T\to\infty} \frac{1}{1-\gamma} \mathbb{E}_{\pi_\theta}\left(\sum_{t=1}^{T} \gamma^{t-1} I(X \in S) \,\middle|\, \sigma(X_1) \right)(x)$$

$$\Rightarrow \mu^{\gamma}_{\pi_\theta} \text{ such that } d\bar{\mu}^{\gamma}_{\pi_\theta}(\cdot;x) = \mu^{\gamma}_{\pi_\theta}(x';x)\, dx'$$

Wednesday, November 18, 2020

Side note $V_{\pi_\theta}(x) = (1-\gamma)\int_{\mathcal{X}}\int_{\mathcal{A}} r(x,a)\,\pi(a;x)\,\mu^{\gamma}_{\pi_\theta}(x';x)\,da\,dx'$

$$\nabla_\theta V_{\pi_\theta}(x) = \int_{\mathcal{X}}\int_{\mathcal{A}} \nabla \pi_\theta(a;x)\, Q_{\pi_\theta}(a;x')\,\mu_{\pi_\theta}(x';x)\,dx'$$

$$\Rightarrow \nabla\rho(\pi_\theta) = \mathbb{E}_{\rho_1}\left[\nabla_\theta V_{\pi_\theta}\right]$$

Undiscounted setting.

$$\rho(\pi_\theta) = \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} r_t\right]$$

Let $\bar{\mu}_{\pi_\theta}(S) = \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}_{\pi_\theta}\left[\sum I(X\in S)\right]$

$S \in \mathcal{B}(\mathcal{X})$

↳ occupancy kernel.

(remember $\bar{P} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} P^{t-1}$)

(we assume all the limits exist and is stationary)

$$\rho(\pi_\theta) = \int_{\mathcal{X}}\left(\int_{\mathcal{A}} \pi(a;x)\,\bar{r}(x,a)\,da\right)\mu_{\pi_\theta}(x)\,dx$$

where $d\bar{\mu}_{\pi_\theta} = \mu_{\pi_\theta}\,dx$

Wednesday, November 18, 2020

what is $\nabla_\theta \rho(\pi_\theta)$

remember

$$V_{\pi_\theta} = \lim_{T\to\infty} \frac{1}{T} E_{\pi_\theta}\left[\sum_{t=1}^{T} r_t - \rho(\pi_\theta) \,\middle|\, \sigma(X_1)\right]$$

and $$Q_{\pi_\theta} = \lim_{T\to\infty} \frac{1}{T} E_{\pi_\theta}\left[\sum_{t=1}^{T} r_t - \rho(\pi_\theta) \,\middle|\, \sigma(X_1, A_1)\right]$$

$$\Rightarrow V_{\pi_\theta}(x) = E_{\pi_\theta}\left[Q(X_1, A_1) \,\middle|\, \sigma(X_1)\right](x)$$

$$= \int_A Q_{\pi_\theta}(x, a)\, \pi_\theta(a; x)\, da$$

$$\Rightarrow \nabla_\theta V_{\pi_\theta}(x) = \int_A \nabla_\theta \pi_\theta(a; x)\, Q_\pi(x, a)\, da$$

$$+ \int_A \pi(a; x)\, \nabla_\theta Q_{\pi_\theta}(x, a)\, da$$

$$\to \bar{r}(x, a) - \rho(\pi_\theta) + \int_X p(x'; x, a) V_{\pi_\theta}(x')\, dx'$$

$$\nabla_\theta V_{\pi_\theta}(x) = \int_A \nabla_\theta \pi_\theta(a;x) Q_{\pi_\theta}(x,a)\, da$$

$$+ \int_A \pi_\theta(a;x) \left(-\nabla_\theta \rho(\pi_\theta) + \int_X p(x';x,a)\, \nabla_\theta V_{\pi_\theta}(x')\, dx' \right)$$

$$\Rightarrow \nabla_\theta \rho(\pi_\theta) = \int_A \nabla_\theta \pi_\theta(a;x) Q_{\pi_\theta}(x,a)$$

$$+ \int_A \pi(a;x) \int_X p(x';x,a)\, V_{\pi_\theta}(x')\, dx - \nabla_\theta V_{\pi_\theta}(x)$$

$$\Rightarrow \int \nabla_\theta \rho(\pi_\theta)\, \mu_{\pi_\theta}\, dx = \int_X \mu_{\pi_\theta} \left(\int_A \nabla_\theta \pi_\theta(a;x)\, Q_{\pi_\theta}(x,a)\, da \right) dx$$

$$+ \int_X \mu_{\pi_\theta} \left(\int_A \pi(a;x) \left(\int_X p(x';x,a)\, \nabla_\theta V_{\pi_\theta}(x')\, dx' \right) da \right) dx$$

$$- \int \mu_{\pi_\theta}\, \nabla_\theta V_{\pi_\theta}\, dx$$

$$\Rightarrow \nabla_\theta \rho(\pi_\theta) = \int_X \mu_{\pi_\theta} \left(\int_A \nabla_\theta \pi(a;x)\, Q_{\pi_\theta}(x,a)\, da \right) dx$$

Wednesday, November 18, 2020

Function, Measure, kernel, functional, Operator

→ $\mathcal{X}$ $\sigma(\mathcal{X})$ $\quad f: \mathcal{X} \to \mathcal{Y}$ ←

$\mathcal{Y}$ $\sigma(\mathcal{Y})$ $\quad \mu; \sigma(\mathcal{X}) \to \mathbb{R}$

new Markov kernel ⟸ $\begin{cases} K(\,; B) \to \text{measure on } \sigma(\mathcal{Y}) \\ K(y;\,) \text{ measurable function} \end{cases}$

$\mathcal{H}$ is a Banach space e.g. space of function

functional $F: \mathcal{H} \to \mathbb{R}$

operator $\mathcal{H}, \mathcal{G}$

$T: \mathcal{H} \to \mathcal{G}$