Monday, November 23, 2020

Lecture 26

CS 59000 _RL Theory

Policy Gradient

Agenda

- Performance Difference Lemma
- Gradient Domination
- Global optimality :)

For discounted setting

$$\underline{\frac{\nabla \eta(\pi_\theta)}{1-\gamma}} = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \left(\int_A \underbrace{\mu^{\gamma}_{\pi_\theta}(x'; x)}_{\text{occupancy kernel}} \nabla \pi_\theta(a; x) \, Q_{\pi_\theta}(a, x) \, da \right) dx' \right) \underbrace{dP_1}$$

For undiscounted.

$$\nabla \eta(\pi_\theta) = \underline{\int_{\mathcal{X}} \mu_{\pi_\theta}(x) \left(\int_A \nabla_\theta \pi(a; x) \, \underline{Q_{\pi_\theta}(x, a)} \, da \right) dx}$$

$\Rightarrow$ Way to estimate: $\begin{cases} \text{Monte Carlo sampling} \nearrow \text{Alexandrov 1965} \\ \text{Bayesian Qudrature} \end{cases}$

$\hookrightarrow$ Bayesian Quadrature Policy Gradient Akella, 2020

Monday, November 23, 2020

Parametrization ; $\theta \in \Theta$; π_θ

Consider the finite state action MDP.

- Direct parametrization

$$\rightarrow \quad \pi_\theta(a;x) = \theta_{ax} \geq 0 \text{ and } \sum_a \theta_{ax} = 1 \ \forall x \in \mathcal{X}$$

- Softmax parametrization

$$\pi_\theta(a;x) = \frac{\exp(\theta_{ax})}{\sum_{a'} \exp(\theta_{a'x})}$$

(side note For direct parametrization) $\nabla_\theta \eta(\pi) \rightarrow \theta \leftarrow \theta + \alpha \nabla \eta(\pi) \Rightarrow$ need to project onto simplex)

General function Classes

$$\pi_\theta(a;x) = f_\theta(x,a) \ ; \ \sum_a f_\theta(x,a) = 1$$

Monday, November 23, 2020

Consider the discounted finite state action MDP.

Define : Advantage function $A_\pi(x,a) \; ; \forall x,a \in \mathcal{X} \times \mathcal{A}$

$$A_\pi(x,a) = Q_\pi(x,a) - V_\pi(x)$$

Lemma (Performance difference Lemma; Kakade & Langford 2002)

$$\eta(\pi) - \eta(\pi') = E_{\mu_\pi^\gamma}\left[E_\pi\left[A_{\pi'}(X,A) \mid X\right]\right]$$

proof in Final. (Azizzadenesheli et al 2018 POMDP)

---

what is policy gradient for direct parametrization

$$\mu_\pi^\gamma(x') = \sum_x \mu_\pi^\gamma(x'; x) \; P_1(x)$$

$$\nabla_\theta \eta(\pi_\theta) = (1-\gamma) \sum_{x'} \sum_a \nabla_\theta \pi_\theta(a; x) Q_\pi(x',a) \mu_{\pi_\theta}(x') da \, dx$$

$$\frac{\partial \eta(\pi_\theta)}{\partial \theta_{ax}} = (1-\gamma) Q_\theta(x,a) \mu(x) \longleftarrow$$

$$\hookrightarrow \nabla \eta(\pi_\theta) = (1-\gamma) \mu Q$$

Monday, November 23, 2020

Let π^* denote an optimal policy

Lemma: For direct parametrization

$$\Rightarrow \frac{\underbrace{\eta(\pi^*)} - \underbrace{\eta(\pi)}}{1-\gamma} \leq \left\| \frac{\mu^\gamma_{\pi^*}}{\mu^\gamma_\pi} \right\|_\infty \max_{\pi'} (\pi' - \pi)^T \nabla_\theta \eta(\pi_\theta)$$

$\Rightarrow$ Gradient dominance setting

proof:

$$\frac{\eta(\pi^*) - \eta(\pi)}{1-\gamma} = \sum_{x, a \in \mathcal{X} \times A} \mu^\gamma_{\pi^*}(x) \pi^*(a; x) A_\pi(x, a)$$

$$\leq \sum_{x \in \mathcal{X}} \mu^\gamma_{\pi^*}(x) \underbrace{\max_{a'} A_\pi(x, a)}_{\text{Non-negative}}$$

Empilicit assumption $\left\| \frac{\mu^\gamma_{\pi^*}}{\mu^\gamma_\pi} \right\|_\infty < \infty$

$$= \sum_{x \in \mathcal{X}} \underbrace{\frac{\mu^\gamma_{\pi^*}(x)}{\mu^\gamma_\pi(x)}} \underbrace{\mu^\gamma_\pi(x)} \underbrace{\max_{a'} A_\pi(x, a)}$$

$$\leq \left\| \frac{\mu^\gamma_{\pi^*}}{\mu^\gamma_\pi} \right\|_\infty \underbrace{\sum_{x \in \mathcal{X}} \mu^\gamma_\pi(x) \max_{a'} A_\pi(x, a')}$$

Monday, November 23, 2020

$$= \left\| \frac{\mu^{\gamma}_{\pi^*}}{\mu^{\gamma}_{\pi}} \right\|_{\infty} \max_{\pi'} \sum_{(x,a)\in \mathcal{X}\times\mathcal{A}} \mu^{\gamma}_{\pi}(x)\, \pi'(a;x)\, A_{\pi}(x,a)$$

Remember $\sum_a \pi(a;x)\, A_{\pi}(x,a) = 0$

$$= \left\| \frac{\mu^{\gamma}_{\pi^*}}{\mu^{\gamma}_{\pi}} \right\|_{\infty} \max_{\pi'} \sum_{(x,a)\in \mathcal{X}\times\mathcal{A}} \mu^{\gamma}_{\pi}(x) \left(\pi'(a;x) - \pi(a;x) \right) A_{\pi}(x,a)$$

$Q(x,a) - V(x)$

$$= \left\| \frac{\mu^{\gamma}_{\pi^*}}{\mu^{\gamma}_{\pi}} \right\|_{\infty} \max_{\pi'} \sum_{(x,a)\in \mathcal{X}\times\mathcal{A}} \mu^{\gamma}_{\pi}(x) \left(\pi'(a;x) - \pi(a;x) \right) Q_{\pi}(x,a)$$

$$= \left\| \frac{\mu^{\gamma}_{\pi^*}}{\mu^{\gamma}_{\pi}} \right\|_{\infty} \max_{\pi'} \nabla_{\pi} \eta(\pi)^{T} (\pi' - \pi)$$

$\Rightarrow$ Projected gradient ascent:

$t = 0$; $\pi_0 = \pi_{\text{init}}$

$$\pi_{t+1} = \mathrm{Proj}_{\Delta(\mathcal{A})^{|\mathcal{X}|}} \left(\pi_t + \alpha \nabla_{\pi} \eta(\pi) \big|_{\pi = \pi_t} \right)$$

Monday, November 23, 2020

Theorem: (Projected gradient ascent);

Agarwal 2019

The projected gradient ascent algorithm on $\eta(\pi)$ with stepsize $\alpha = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}|}$, ran for

T iterations satisfies:

$$\min_{t<T} \eta(\pi^*) - \eta(\pi_t) \leq \varepsilon$$

whenever $T > \frac{64\,\gamma\,|\mathcal{X}|\,|\mathcal{A}|}{(1-\gamma)^6\,\varepsilon^2} \sup_\pi \left\| \frac{\mu_\pi^*}{\mu_\pi} \right\|$

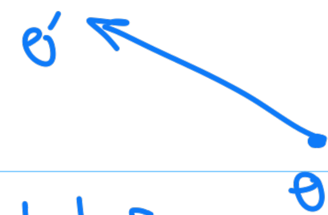
Proof Final.

Natural policy gradient

Let's imagine our policy kernel is parametrized with θ; π_θ from $(\mathcal{X}, \sigma(\mathcal{X}))$ to $(\mathcal{A}, \sigma(\mathcal{A}))$

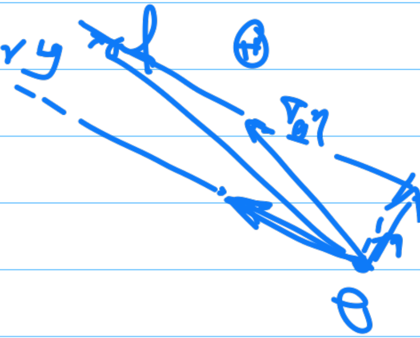
remember $\underline{\eta(\pi_\theta)} = \int_{\mathcal{X}} \mu_\pi(x) \left(\int \underline{\pi_\theta}(a; x)\, \bar{r}(x,a)\, da \right) dx$

Monday, November 23, 2020



- what if we change θ, but π_θ would not change!
- we want to change θ such that π_θ changes sufficiently

⇒ we change the geometry of



Consider Fisher information (state-dependent)

$$F(x;\theta) = E_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta \; \nabla_\theta \log \pi_\theta^T \,\middle|\, x\right](x)$$

and Fisher information matrix

$$F(\theta) = E_{\mu_{\pi_\theta}}\left[F(x;\theta)\right]$$

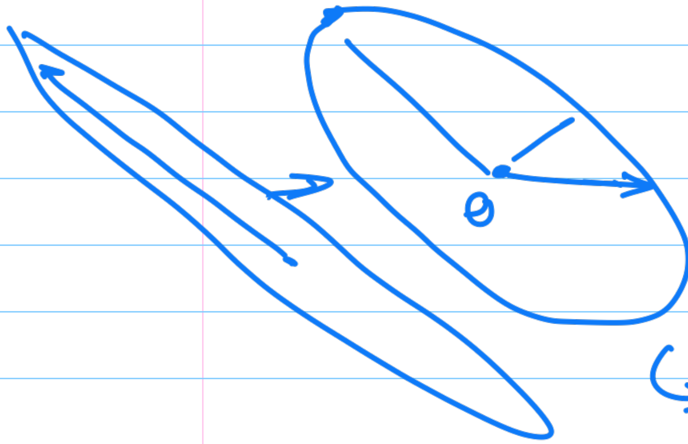
Fisher information matrix induces a Remanian geometry such that locally the parameters of π are metric invariant

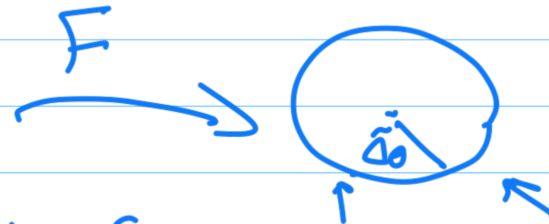
(look Amari's book) information Geometry and application 1995 et al.)

Monday, November 23, 2020

⇒ Natural policy gradient:

$$\tilde{\nabla}_\theta \eta(\pi_\theta) = \underline{F(\theta)^{-1}} \nabla \eta(\theta)$$





$$\hat{\Delta\theta} = F^{1/2} \Delta\theta$$

$$\|\Delta\theta\|_F = \delta$$

$$(\Delta\theta^T F \Delta\theta) = \delta^2$$

$$KL(\underline{\pi_\theta, \pi_{\theta+\Delta\theta}}) \cong \|\Delta\theta\|_F$$

Natural gardients are essential in probabilistic optimization and the design of geometry-based optimizer.