Monday, November 30, 2020

Lecture 27

CS59000-RL

Policy Gradient

Agenda

- Natural policy gradient
- Trust region policy gradient
- Linear dynamical systems.

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

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

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

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

$$\Delta\theta \approx 0 \rightarrow KL(\pi_\theta, \pi_{\theta+\Delta\theta}) \sim \|\Delta\theta\|_F^2$$

$\Delta\theta = 0 \rightarrow KL = 0$

$\Delta\theta = 0 \rightarrow \nabla KL = 0$

$\Delta\theta = 0 \rightarrow \nabla^2 KL = F$

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

Monday, November 30, 2020

$F(\theta)$ is huge and hard to invert.

Can we compute $\underbrace{\tilde{\nabla}\eta(\theta)}_{} = \underbrace{F(\theta)}_{}^{-1}\underbrace{\nabla\eta(\theta)}_{}$ directly

without computing $F(\theta)$ and $\underbrace{\nabla\eta}_{}(\theta)$ seperately.

Consider a feature vector $\underbrace{\psi_\pi(x,a) = \nabla \log \pi_\theta(a;x)}_{}$

we use this feature vector to approximate Q.

$$\Rightarrow \omega^T\psi \sim Q \quad \text{i.e. } \underbrace{(\omega^T\psi)(x,a)}_{} \sim \underbrace{Q_{\pi_\theta}(x,a)}_{}$$

Let's define

$$\varepsilon(\omega,\pi) = \sum_x \sum_a \underbrace{\mu_{\pi_\theta}(x)\pi_\theta(a;x)}_{} \underbrace{\left((\omega^T\psi)(x,a) - Q_{\pi_\theta}(x,a)\right)^2}_{}$$

Theorem (Kakade 2002 "A Natural Policy gradient")

Let $\tilde{\omega}$ be a minimizer of $\varepsilon(\omega,\pi)$

$\Rightarrow$ Then $\tilde{\omega} = \tilde{\nabla}\eta(\theta)$

Monday, November 30, 2020

Proof: $\nabla_\omega \, \mathcal{E}(\omega, \pi) = 0$

$$\Rightarrow \sum_{x,a} \mu_\theta(x) \pi_\theta(a;x) \psi_{\pi_\theta}(x,a) \psi_{\pi_\theta}(x,a)^T \tilde{\omega} = \sum_{x,a} \mu_\theta(x) \pi(a;x) \psi_{\pi_\theta}(x,a) Q_{\pi_\theta}(x,a)$$

Remember $\nabla_\theta \pi_\theta(a;x) = \pi_\theta(a;x) \, \nabla \log \pi_\theta(a;x)$

$$= \pi_\theta(a;x) \, \psi_\pi(x,a)$$

$$\Rightarrow \sum_{x,a} \mu_\theta(x) \pi_\theta(a;x) \psi_{\pi_\theta}(x,a) \psi_{\pi_\theta}(x,a)^T \tilde{\omega} = \nabla \eta(\theta)$$

$$\Rightarrow F(\theta) \tilde{\omega} = \nabla \eta(\theta) \Rightarrow \tilde{\omega} = F^{-1}(\theta) \nabla \eta(\theta)$$

$$F = \nabla^2_{\theta'} KL(\theta, \theta')\Big|_{\theta=\theta'}$$

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

If we follow policy π, we can estimate A_π
we have trajectories under π too

Can we optimize $\eta(\pi')$ to find an optimal policy?

If we have trajectories out of $\mu_{\pi'}$, then we could. We have trajectories induced by μ_π.

Monday, November 30, 2020

We can define and compute a surrogate function

$$\underline{L_\pi}(\pi') = \underset{\underline{\mu_\pi}}{E}\left[\underline{\underset{\underline{\pi'}}{E}\left[\underline{A}_\pi(x,A) \mid x\right]}\right] + \eta(\pi)$$

If we keep π' close to π, then $L_\pi(\pi')$ is a good surrogate of $\eta(\pi')$.

For $\pi' = \pi \Rightarrow L_\pi(\pi) = \eta(\pi)$

and $\nabla_{\pi'} L_\pi(\pi')\Big|_{\pi'=\pi} = \nabla \eta(\pi)$

$\Rightarrow L_\pi(\pi')$ is equal to $\eta(\pi')$ up to first order in the vicinity of π.

Great news:

Theorem (Schulman et al 2017) (Kakade & Langford 2002)

$$\eta(\pi') \geq L_\pi(\pi') - \frac{4 \epsilon_\pi \gamma}{(1-\gamma)^2} \max_x KL(\pi(\cdot;x), \pi'(\cdot;x))$$

where $\epsilon_\pi = \max_{x,a} |A_\pi(x,a)|$

[Monotonic Improvement Lemma]

Monday, November 30, 2020

optimize the $\overbrace{L_\pi(\pi') - \frac{4\epsilon_\pi \gamma}{(1-\gamma)^2} \max_s KL(\pi(\cdot;s), \pi'(\cdot;s))}$

when $\pi' = \pi \Rightarrow \max_s KL(\pi(\cdot;s), \pi'(\cdot;s)) = 0$

$L_\pi(\pi') = \eta(\pi)$ when $\pi' = \pi$

$\eta(\pi') \geq \max$ R.H.S. $\geq \eta(\pi)$

$\Rightarrow$ Conservative policy gradient (Kakade & Langford 2002)

$\Rightarrow$ Trust Region Policy optimization (Schulman 2017)
(TRPO)

e.g. $\max_{\pi'} L_\pi(\pi')$

$\text{s.t.} \max_s KL(\pi(\cdot;s), \pi'(\cdot;s)) < \delta$

You can extend it to POMDP.

Monday, November 30, 2020

Linear Quadratic Regulators (LQR)

$\rightarrow (\mathcal{X} = \mathbb{R}^n, \mathcal{A} = \mathbb{R}^m, A, B, R, Q, P_1, \Sigma)$

$A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $\Sigma, Q \in \mathbb{R}^{n \times n}$, $R \in \mathbb{R}^{m \times m}$

where R is PD, Q, Σ are PSD

$X_1 \sim P_1$

action, control input

$$X_{t+1} = A X_t + B U_t + W_t$$

mean zero nois with Covariance Σ

$$\Rightarrow C_t = \|X_t\|_Q^2 + \|U_t\|_R^2$$

Policy to choose U_t to minimize undiscounted infinite horizon cost.

$\Rightarrow$ The optimal policy π ; $U = \underbrace{K}_{} X$; $K \in \mathbb{R}^{m \times n}$
exists, and is deterministic.

$$K = -\underbrace{(R + B^T P B)^{-1} B P A}_{}$$

P is the unique PD solution to the discrete time algebraic Riccati equation (DARE)
(similar to Poisson or dynamic programing)

Monday, November 30, 2020

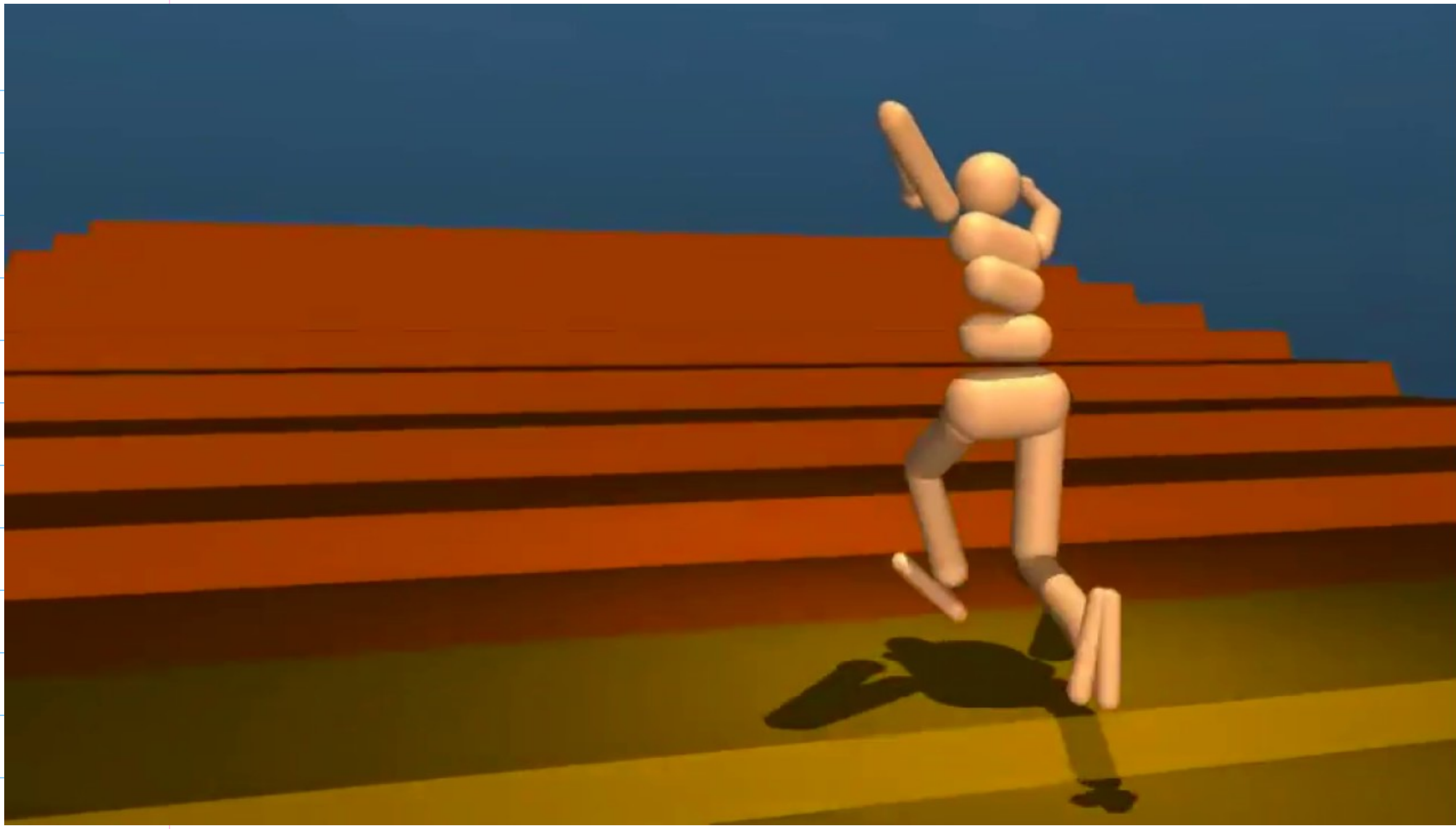
$\rightarrow$ $$P = A^T P A + Q - A P B (R + B P B)^{-1} B^T P A$$

$\Rightarrow$ Does policy gradient converge to the optimal K?

yes. Under some regularity condition, Σ being PD

$\Rightarrow$ PG powerfull.

Monday, November 30, 2020



Monday, November 30, 2020