Lecture 27

CSS9GOG_RL Policy Gradient Avenda

- Natural policy gradient

- Trush region policy gradient

- Linear d'unamical systems.

=> F(e) = 5 (F(x;e)) => Fisher information

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

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

Δ0=0 → KL (Π0, 10+Δ0) ~ 11Δ0 11² Δ0=0 → KL =0

100 0 -5 KL=0

abzo -> 42 KlzF

F(0) is huge and hard to invert.
Can we compute Fn(0) = F(0) Tr(0) directly
without computing F(0) and Ty(0) separately.
Consider a feabure vector 4 (na)= 7 les 170 (a;n)
we we this feature vector to approximate Q.
=> Wy ~Q i.e. (Wy) (n,a) ~ Q(n,a)
Lets define
Let's define $ \Xi(\omega,\Pi) = \Xi \Xi \mu_{\Pi}(\lambda) \Pi_{\theta}(\alpha;\lambda) \left((\omega' \omega)(\mu_{1}\alpha) - Q_{\Pi}(\mu_{1}\alpha) \right) $ $ \Xi(\omega,\Pi) = \Xi \Sigma \mu_{\Pi}(\lambda) \Pi_{\theta}(\alpha;\lambda) \left((\omega' \omega)(\mu_{1}\alpha) - Q_{\Pi}(\mu_{1}\alpha) \right) $
X X
Theorem (Kakade 2002 "A Nahural Policy svadigal)
meorem (Markares a marking state (4)
Let 5 be a minimizer of E(W, M)
>> Then \(\overline{\pi} = \overline{\pi}\gamma(\theta)

Proof: 7 & (w,n)=0 > = M(n) [(4;4) 4 (ma) 4 (ma) 4 = = M(n) [(4;4) 4 (ma) Q(ma)

716 Remember 70 Po (a; h) = Mo (a; h) The Mo (a; h) => Z M (2) M (a; 2) Y (NG) Y (N/ 4) T W Z TY (8) >> F(0) \$\overline{\pi} = \text{Ty(0)} => \overline{\pi} = \text{F(0)} \text{Ty(0)} $= \sum_{n \in \mathbb{N}} \gamma(n') - \gamma(n) = E_{n'} \left[E_{n'} \left[A_n(x_{iA}) \mid x \right] \right]$ If we follow policy of we can estimate And we have trajectories under 11 too Can we optimize y(M) to find an optimal policy? If we have trajectories ont of My then we could. We have trajectories induced by My.

We can define and compute a surragable function

$$L_{\Pi}(\Pi') = E\left[E\left[A_{\Pi}(X_{t}A)IX\right] + \eta(\Pi)\right]$$

If we keep of close to on, then Lou(M) is agood supprogate of n(M).

amol
$$\Delta^{U, \Gamma}(U, J) = \Delta \Lambda(U)$$

=> 2, (Ti) is equal to y(Ti) up to first order in the vicinity of T.

Great nems:

Theorem (Schulmanet al 2017) (Kalcade Llangtord 2002 $\gamma(T) > L_{\Pi}(\Pi') - \frac{4 \Sigma_{\Pi} \chi}{(1-\chi)^2} \max_{n} KL(\Pi(\cdot;n),\Pi(\cdot;n))$

[monobonic Improvement lemm]

	Monday, November 30, 2020
	optimize the 2 (11)-4 En 8 man KL(11(1)), 11(-14)
	when T=T= max KL(T(:,sh),T(:,sh)=0
	$L_{\Pi}(\Pi) = \chi(\Pi)$ when $\Pi = \Pi$
Z(n)> w	at RHS. > 200)
	=> Consevative Policy gradient (Kckade & langtood 2002
	=> Trust Region Police Optinization (Schulman 2017)
	eg. max Ln(T)
	5.6. man 1<6 (17 (·; h), 11 (·; h)) (8
	You can extend it to POMPP.
_	

A = IRM, A= IRM, A, B, R, Q, P, E A = IRMXN, B = IRMXN, E, Q = IRM, R = IRMXN Where R is PD, Q, E are PSD X, N P; action, control input X + = A X + BV + W Mean zero nois The optimal palics T; U= KX; K \(\) KEIR P is the unique PD solution to the discrete	Linear anadratic Regulators (LAR)
X, N P, X ty = A X + B V + Wb Mean zero nois Ct = Xt 2 + Ut 2 with Cavaran. Policy to choose Ut to minimize undisconted in finite horizon asb. The optimal policy T; U = KX; KEIR MXM exists, and is deterministic. Kz - (R+BTPB) BPA P is the unique PD solution to the discrete	>(X=1R", A=1R", A,B,R,Q,P,, E)
X, N P, X ty = A X + B V + Wb Mean zero nois Ct = Xt 2 + Ut 2 with Cavaran. Policy to choose Ut to minimize undisconted in finite horizon asb. The optimal policy T; U = KX; KEIR MXM exists, and is deterministic. Kz - (R+BTPB) BPA P is the unique PD solution to the discrete	where R is PD, Q, E are PSD
X + = A X + B V + W to mean zero nois The optimal palice of the minimize undisconted exists, and is deterministic. R - (R + B P B) B P A P is the unique PD solution to the discrete	X, ~ P, action, control input
Policy to choose Ut to minimize undisconted in finite horizon ast. The optimal policy T; U=KX; KEIR mxn exists, and is deterministic. Kz - (R+BTPB) BPA P is the unique PD solution to the discrete	$X_{t+1} = A X_{t} + B V_{t} + W_{t}$
Policy to choose Ut to minimize undisconted in finite horizon ast. The optimal policy T; U=KX; KEIR mxn exists, and is deterministic. Kz - (R+BTPB) BPA P is the unique PD solution to the discrete	-> C+= X+ 2 + U+ R with Granan
R= (R+BTPB) BPA P is the unique PD solution to the discrete	
R= (R+BTPB) BPA P is the unique PD solution to the discrete	=> The optimal palics 17; U=KX; KEIR mxn exists, and is deterministic.
P is the unique PD solution to the discrete time algebraic Riccati equation (DARE) (similar to Poisson ex dynamic arcrass)	Kz-(R+BTPB) BPA
	P is the unique PD solution to the discrete time algebraic Riccati equation (DARE) (similar to Poisson or dynamic programin)

->	P= ATPA +Q-APB(R+BPB)BTPA
ij	>Does policy stadient converge to the optimi K?
	yes. Under some regulability condition, & beig PD
	> Powerfull.



