Lec 6: Policy gradient methods

(Based on Chi Jin's lecture notes)

Framework: So far, we've focused only on $\underbrace{\text{value-based methods}}$

↳ estimating $Q^*(s,a)$

Now we focus on policy-based method

Θ convex set

policy class $\quad \Pi = \{\pi_\theta \mid \theta \in \Theta\}$

policy optimization: $\quad \max_{\theta \in \Theta} J(\theta) = V_1^{\pi_\theta}(s_1)$

Typical algorithm: Projected Gradient Ascend (PGA)

$$\theta_{t+1} = \mathrm{Proj}_\Theta\left[\theta_t + \eta \nabla J(\theta_t)\right]$$

$$\Longleftrightarrow \quad \theta_{t+1} = \operatorname*{argmax}_{\theta \in \Theta} \underbrace{\langle \Delta J(\theta_t), \theta - \theta_t \rangle}_{\text{linear approx}} - \underbrace{\frac{1}{2\eta}\|\theta - \theta_t\|^2}_{\text{quadratic regularization}}$$

Existing theory for (P)GA when J is concave

In general, we can replace quadratic regularization by any Bregman divergence $D_{\Phi}(\theta, \theta_t)$

<u>Mirror Ascent</u>: $$\theta_{t+1} = \underset{\theta \in \Theta}{\operatorname{argmax}} \langle \nabla J(\theta_t), \theta - \theta_t \rangle - \frac{1}{\eta} D_{\Phi}(\theta, \theta_t)$$

e.g. $D_{\Phi}(\cdot, \cdot) = KL[\cdot, \cdot]$

<u>Parametrization of policies in tabular MDP</u>:

$$\pi = (\pi_1, \dots, \pi_H) \quad \text{where} \quad \pi_h(\cdot | s) \in \Delta(A)$$

a policy is represented by SHA-dimensional vector

(a $|A|$-dimensional vector on simplex for each (s, h))

Case 1 : MAB with known mean rewards

$(S = H = 1)$ a policy θ : one single vector on A- dim simplex

$$J(\theta) = \langle \theta, r \rangle \ , \qquad \nabla J(\theta) = r$$

$$\theta_{t+1} = \underset{\theta \in \Delta(A)}{\mathrm{argmax}} \ \langle r, \theta - \theta_t \rangle - \frac{1}{\eta} KL[\theta \| \theta_t]$$

simplex constraints

Lagrange multiplier

$$L(\theta, \lambda) = \langle r, \theta - \theta_t \rangle - \frac{1}{\eta} KL(\theta, \theta_t) - \lambda \left(\sum_a \theta(a) - 1 \right)$$

$$\sum_a \theta(a) \log \frac{\theta(a)}{\theta_t(a)}$$

- $$\frac{\partial L}{\partial \theta(a)} = r(a) - \frac{1}{\eta} \left(\log \frac{\theta(a)}{\theta_t(a)} + 1 \right) - \lambda = 0$$

$$\Rightarrow \theta(a) \propto \theta_t(a) \, e^{\eta r(a)}$$

- $$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow \sum_a \theta(a) = 1 \qquad \Rightarrow \theta(a) = \frac{\theta_t(a) e^{\eta r(a)}}{Z_t}$$

Case 2: MAB w/ unknown reward (estimate the reward)

Exponential weight algorithm for exploration and exploitation (EXP3)

- initialize θ_1 to be uniform
- For $t = 1, 2, .., T$:

where $|R_t| \le B$

pull arm a_t according to distribution θ_t and observe reward $R_t(a_t)$;

$$\hat{r}_t(a) \leftarrow B - \overbrace{\frac{B - R_t(a)}{\theta_t(a)} 1\{a = a_t\}}^{\in [0, \infty]}$$

$$\hat{r}_t(a) \in [-\infty, B]$$

$$\theta_{t+1}(a) \leftarrow \theta_t(a)\, e^{\eta \hat{r}_t(a)} / Z_t \quad \forall a$$

Analysis:

$$\theta_{t+1}(a) = \theta_t(a)\, e^{\eta \hat{r}_t(a)} / Z_t$$

$$\frac{1}{\eta} \log \frac{\theta_{t+1}(a)}{\theta_t(a)} = \hat{r}_t(a) - \frac{1}{\eta} \log Z_t$$

$$\langle \hat{r}_t, \theta_{t+1} \rangle = \frac{1}{\eta} \sum_a \left(\log \frac{\theta_{t+1}(a)}{\theta_t(a)} + \log Z_t \right) \theta_{t+1}(a)$$

$$- \quad \langle \hat{r}_t, \theta^* \rangle = \frac{1}{\eta} \sum_a \left(\log \frac{\theta_{t+1}(a)}{\theta_t(a)} + \log Z_t \right) \theta^*(a)$$

$$\langle \hat{r}_t, \theta_{t+1} - \theta^* \rangle = \frac{1}{\eta} \left\langle \theta_{t+1} - \theta_*, \log \frac{\theta_{t+1}}{\theta_t} \right\rangle$$

$$\geq -\frac{1}{\eta} \left\langle \theta_*, \log \frac{\theta_{t+1}}{\theta_t} \right\rangle = -\frac{1}{\eta} \left(KL(\theta^*, \theta_t) - KL(\theta^*, \theta_{t+1}) \right)$$

$$\Rightarrow \langle \hat{r}_t, \theta^* - \theta_t \rangle \leq \langle \hat{r}_t, \theta_{t+1} - \theta_t \rangle + \frac{1}{\eta} \left(KL(\theta^*, \theta_t) - KL(\theta^*, \theta_{t+1}) \right) \qquad (\mathrm{I})$$

Let F_t: σ-algebra generated by RVs up to (inclusive) time step t

$$\Rightarrow \mathbb{E}\left[\langle \hat{r}_t, \theta^* - \theta_t \rangle \mid F_{t-1} \right] = \langle r, \theta^* - \theta_t \rangle \qquad (\mathrm{II})$$

However, $\mathbb{E}\left[\langle \hat{r}_t, \theta_{t+1} - \theta_t \rangle \mid F_{t-1}\right] \neq \langle r, \theta_{t+1} - \theta_t \rangle$

since θ_{t+1} depends on $\hat{r}_t$

Note: $\hat{r}_t(a) \in [-\infty, B]$

$$\hat{l}_t(a) = B - \hat{r}_t(a) = \frac{B - R_t(a)}{\theta_t(a)} \mathbb{1}\{a = a_t\} \in [0, \infty], \quad \theta_{t+1}(a) \propto \theta_t(a) e^{\eta \hat{r}_t(a)} \propto \theta_t(a) e^{-\eta \hat{l}_t(a)}$$

$$\langle \hat{r}_t, \theta_{t+1} - \theta_t \rangle = \langle \hat{r}_t - B, \theta_{t+1} - \theta_t \rangle = -\langle \hat{l}_t, \theta_{t+1} - \theta_t \rangle$$

$$= -\sum_a \underbrace{\hat{l}_t(a)}_{\geq 0} \left(\frac{\theta_t(a) e^{-\eta \hat{l}_t(a)}}{\underbrace{\sum_{a'} \theta_t(a) e^{-\eta \hat{l}_t(a')}}_{\leq 1}} - \theta_t \right)$$

$$\leq -\sum_a \hat{l}_t(a)\, \theta_t(a) \left(e^{-\eta \hat{l}_t(a)} - 1 \right) \qquad e^x \geq x + 1 \ \forall x$$

$$\leq -\sum_a \hat{l}_t(a)\, \theta_t(a) \left(-\eta \hat{l}_t(a) \right)$$

$$= \eta \sum_a \hat{l}_t(a)^2 \theta_t(a)$$

$$= \eta \sum_a \frac{(B - R_t(a))^2}{\theta_t(a)} \mathbb{1}\{a = a_t\} \leq \eta B^2 \sum_a \frac{\mathbb{1}\{a = a_t\}}{\theta_t(a)}$$

$$= \frac{\eta}{\theta_t(a_t)}$$

$$\Rightarrow \mathbb{E}\left[\langle \hat{r}_t, \theta_{t+1} - \theta_t \rangle \mid F_{t-1}\right] \leq \eta B^2 \mathbb{E}\left[\frac{1}{\theta_t(a_t)} \mid F_{t-1}\right]$$

$$= \eta B^2 \sum_a \theta_t(a) \frac{1}{\theta_t(a)} = \eta B^2 A \quad \text{(III)}$$

Combine (I), (II), (III):

$$\text{regret}(T) = \sum_{t=1}^{T} \langle r, \theta^* - \theta_t \rangle \leq \eta T A B^2 + \underbrace{\frac{1}{\eta} KL(\theta^*, \theta_0)}_{\frac{1}{\eta} \sum_a \theta^*(a) \log(\theta^*(a) A)}$$

$$\leq \frac{1}{\eta} \log A$$

$$\leq \eta T A B^2 + \frac{1}{\eta} \log A \lesssim \boxed{B\sqrt{TA \log A}}$$

by setting η as follows:

$$B^2 \eta T A = \frac{1}{\eta} \log A \iff \eta = \sqrt{\frac{\log A}{TAB^2}} \Rightarrow B^2 \eta T A = B \sqrt{TA \log A}$$