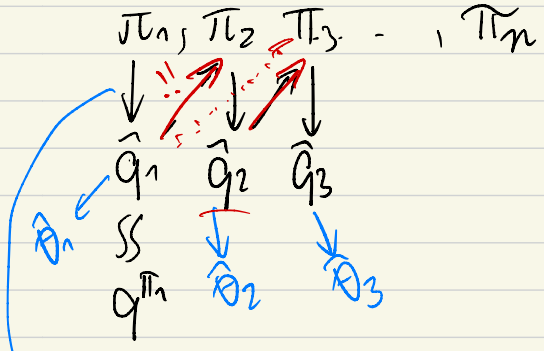
March 4

POLITEX — Policy Iteration with Expert Advice



$$\pi_k(a|s) \propto \exp\left(\eta \sum_{j=1}^{k-1} \hat{q}_j(s,a)\right) = E_k(s,a)$$

$\bar{\theta}_{k-1}^T \psi(s,a)$

$$\pi_k(a|s) = \frac{E_k(s,a)}{\sum_{a'} E_k(s,a')}$$

LSPE — G-optimal design

m rollouts, H length,

$$\bar{\theta}_{k-1} = \sum_{j=1}^{k-1} \hat{\theta}_j$$

$$\hat{q}_k = \Phi \hat{\theta}_k$$

$$\sum \hat{q}_j = \Phi \sum_j \hat{\theta}_j$$

Output: $\pi = \frac{1}{n}(\pi_1 + \dots + \pi_n)$.

$$\Rightarrow v^{\pi} = \frac{1}{n} \sum_{k=1}^{n} v^{\pi_k}$$

$$v^* - v^{\pi} = \frac{1}{n} \sum_{k=1}^{n} v^* - v^{\pi_k}$$

$$= \frac{1}{n}(I - \gamma P_{\pi^*})^{-1} \sum_{k=1}^{n} M_{\pi^*} \hat{q}_k - M_{\pi_k} \hat{q}_k + \frac{1}{n}(I - \gamma P_{\pi^*})^{-1} \sum_{k=1}^{n} (M_{\pi^*} - M_{\pi_k})(q^{\pi_k} - \hat{q}_k)$$

$\frac{1}{n}$ → $\frac{1}{(1-\gamma)\sqrt{n}}$

I. $\leq \sqrt{n}$

II.

Online Learning = Optimization

$$\|\Pi_\cdot\|_\infty \le \frac{2}{1-\gamma} \max_{1\le k\le n} \|q^{\pi_k} - \hat{q}_k\|_\infty \le \frac{2\,\varepsilon(1+\sqrt{d})}{1-\gamma} + \frac{2\sqrt{d}}{(1-\gamma)^2}\left(\gamma^H + \sqrt{\frac{\log(d(d+1)/\delta)}{2m}}\right)$$

Lemma (LSPE-G extrapolation error control): Fix any full-rank feature-map $\varphi : \mathcal{Z} \to \mathbb{R}^d$ and take the set $\mathcal{C} \subset \mathcal{Z}$ and the weighting function $\varrho \in \Delta_1(\mathcal{C})$ as in the Kiefer-Wolfowitz theorem. Fix an arbitrary policy π and let θ and ε_π such that $q^\pi = \Phi\theta + \varepsilon_\pi$ and assume that immediate rewards belong to the interval $[0, 1]$. Let $\hat{\theta}$ be as in Eq. (5). Then, for any $0 \le \delta \le 1$, with probability $1 - \delta$,

$$\left\|q^\pi - \Phi\hat{\theta}\right\|_\infty \le \|\varepsilon_\pi\|_\infty(1+\sqrt{d}) + \sqrt{d}\left(\frac{\gamma^H}{1-\gamma} + \frac{1}{1-\gamma}\sqrt{\frac{\log(2|C|/\delta)}{2m}}\right). \quad (6)$$

$\mathbb{E}[\|\Pi_\cdot\|] \le$ $\delta = \frac{\delta_0}{n}$, truncate $\hat{q}_k$ to $[0, \frac{1}{1-\gamma}]$

$\delta_0 \frac{1}{1-\gamma} + \dots$ assume $r_a(s) \in [0,1]$

$\|\Pi_\cdot\| = \sum_{k=1}^{n} M_{\pi^*}\hat{q}_k - M_{\pi_k}\hat{q}_k$

$I_\cdot(s) = \sum_{k=1}^{n} \hat{q}_k(s, \pi^*(s)) - \sum_a \pi_k(a|s)\hat{q}_k(s,a)$

$k = 1, 2, \dots, n$

1. Choose $\pi_k \propto ?? \exp\left(\eta \sum_{j=1}^{k-1} q_j\right)$

2. Compute $\hat{q}_k$ [depends on π_k]

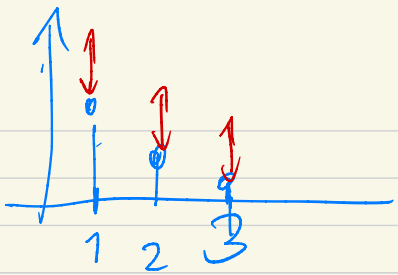
$\hat{q}_k \in [0, \frac{1}{1-\gamma}]$

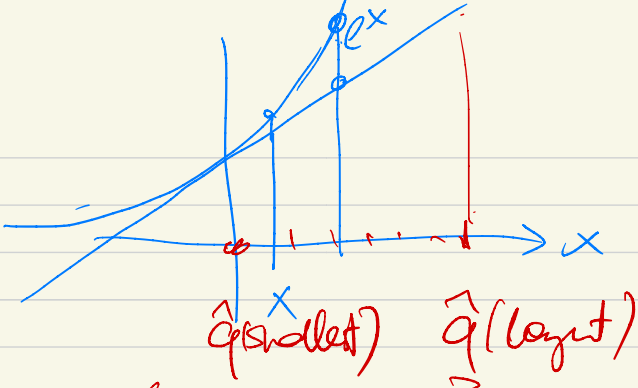
$\frac{1}{1-\gamma}\sqrt{2n\log A}$

$\pi_k \propto \exp(\eta \hat{q}_{k-1})$

$\pi_k(a|s) \propto \exp(\eta(\hat{q}_{k-1}(s,a)) + c)$, $a \in \mathcal{A}$

$\eta \to \infty$



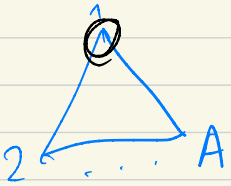


$$\frac{\exp(\eta \hat{q}_{t-1}(s,a))\, e^{\eta \varepsilon}}{\sum_{a} \exp(\eta \hat{q}_{t-1}(s,a))\, e^{\eta \varepsilon}}$$

Fixing $s \in S$.

$k = 1, \dots, n$:

(1.) Learner chooses $\pi_k \in \mathcal{M}_1(A)$



(2.) Adversary $\hat{q}_k : A \to [0, \frac{1}{1-\delta}]$ (reactive)

$\forall \pi^* \in \mathcal{M}_1(A)$

Need to control: $R_n = \sum_{k=1}^{n} \langle \pi^*, \hat{q}_k \rangle - \langle \pi_k, \hat{q}_k \rangle$

$R_n = \sigma(n)$ / $\frac{R_n}{n} \to 0$, $n \to \infty$

regret

"Online linear optimization"

$\subseteq$ "Online learning"

$\pi \longmapsto \langle \pi, \hat{q}_k \rangle$

$\ell_1, \dots, \ell_n : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subseteq \mathbb{R}^d$

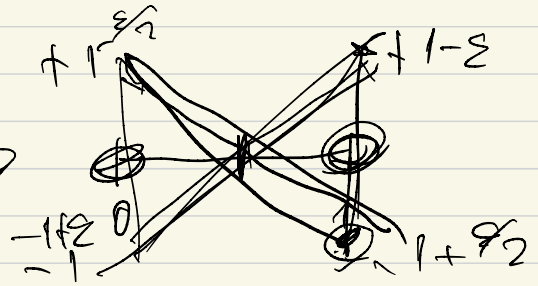
$x^* \in \mathcal{X}$ $\quad \mathcal{X} \neq \emptyset$ convex (closed)

$$R_n = \sum_{t=1}^{n} \ell_t(x) - \sum_{t=1}^{n} \ell_t(x^*)$$; ℓ_t convex

1. Convex opt.
2. Gradient
3. Nonconvex opt.
4. Learning

$\ell_1 = \ell_2 = \dots = \ell_n = \ell$

$\ell_t(x) = \langle x, y_t \rangle$



FTRL / MD

Theorem: $\pi_t(a) \propto \exp\left(\eta \sum_{j=1}^{t-1} \tilde{q}_j(\cdot)\right)$

$$R_n \leq \frac{1}{1-\gamma} \sqrt{2n \log(A)}$$

$\to \eta = ?!$

$\eta = \sqrt{\frac{2 \log A}{n}} \frac{1}{1-\gamma}$

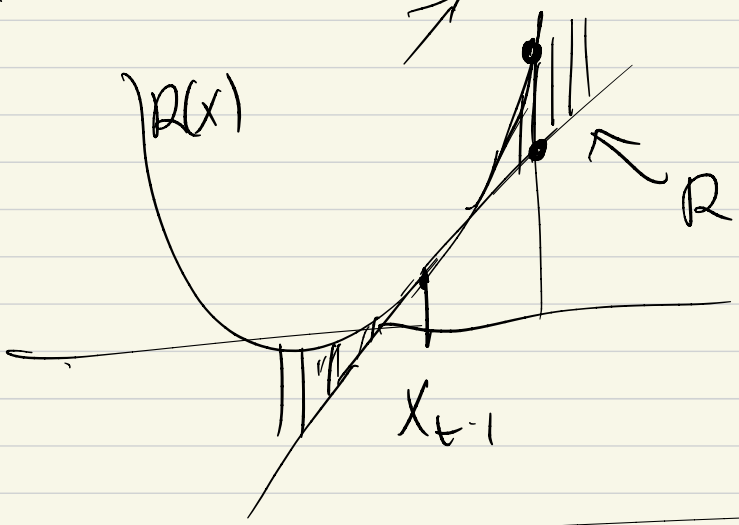
FTL : $x_t = \operatorname{argmin}_{x \in \mathcal{X}} \sum_{s=1}^{t-1} \ell_s(x)$

FTRL : $x_t = \operatorname{argmin}_{x \in \mathcal{X}} \sum_{s=1}^{t-1} \ell_s(x) + \eta R(x)$

$R : \mathcal{X} \to \mathbb{R}$ convex $\quad \eta > 0$

MD : $x_t = \operatorname{argmin}_{x \in X} \ell_{t-1}(x) + \eta D_R(x, x_{t-1})$

$0 \le D_R(x, x_{t-1}) = R(x) - [R(x_{t-1}) + \langle \nabla R(x_{t-1}), x - x_{t-1} \rangle]$



$X = \Delta_1(A) = \{ p \in [0,1]^A \mid \sum_{i=1}^{A} p_i = 1 \}$

$R(p) = -\sum_{i=1}^{A} p_i \log p_i$ | $R(p) = \frac{1}{2}\|p\|_2^2$

Tsallis entropy

$D_R(p, q) = KL(p, q)$

$\log(A)$

$\sqrt{A}$

$p_t = \operatorname{argmin}_{p \in \Delta_1(A)} \langle p, y_{t-1} \rangle + \eta KL(p, p_{t-1})$

+ Pot(w.x)

$\Downarrow$

$p_t(a) \propto p_{t-1}(a) e^{-\eta y_{t-1}(a)}, \quad a \in A$

$$p_1(a) = \frac{1}{A}$$

$$p_t(a) \propto e^{-\eta(g_{t-1}(a) + \cdots + g_1(a))}$$