Contextual Bandits w/ policies Π: $\pi \in \Pi$ is a map $\pi : \mathcal{C} \to \mathcal{X}$

Input Π

for $t = 1, 2, \ldots, T$

  Nature reveals context $c_t \in \mathcal{C}$ is iid from $\mathcal{D}$

  Player choose action $x_t \in \mathcal{X}$ [equiv. chooses $\pi_t \in \Pi$ and plays $\pi(c_t)$]

  Player receives $r_t \in [0,1]$ where $\mathbb{E}[r_t | c_t, x_t] = v(c_t, x_t)$

Define $V(\pi) = \mathbb{E}_c[v(c, \pi(c))]$

$$\text{Regret} = \max_{\pi} \sum_{t=1}^{T} V(\pi) - V(\pi_t) \qquad (1)$$

$$\max_{\pi \in \Pi} \sum_{t=1}^{T} v(c_t, \pi) - v(c_t, \pi_t) \qquad (2)$$

$$\max_{\pi} \sum r_t[\pi(c_t)] - r_t[\pi_t(c_t)] \qquad (3)$$

In our stochastic setting these are all within $O(\sqrt{T})$

Idea: play some logging policy μ: in response to context c_t, policy plays $x_t = x$ w.p. $\mu(x|c_t)$

Collect dataset $\{(c_t, x_t, r_t, p_t)\}_{t=1}$ $\quad p_t := \mu(x_t | c_t)$

Estimate $\hat{V}(\pi) = \frac{1}{S} \sum_{t=1}^{S} \frac{\mathbb{1}\{x_t = x\}}{p_t} r_t$

$$\mathbb{E}[\hat{V}(\pi)] = V(\pi)$$

$$\mathbb{E}[(\hat{V}(\pi) - V(\pi))^2] \leq \mathbb{E}_{c \sim \mathcal{D}}\left[\frac{1}{\mu(\pi(c)|c)}\right].$$

By Bernstein. w.p. $\geq 1-\delta$, $\forall \pi \in \Pi$ we have

$$|\hat{V}(\pi)-V(\pi)| \leq \sqrt{\mathbb{E}_{c\sim \mathcal{D}}\left[\frac{1}{\mu(\pi(c)|c)}\right] \cdot \frac{2\log(2|\Pi|/\delta)}{\tau}} + \frac{2 V_{max} \log(2|\Pi|/\delta)}{3\tau}$$

where $V_{max} = \max_{c,x} \frac{1}{\mu(x,c)}$.

τ-Greedy: Algorithm for "model the bias"/general policies

Input Π

for $t = 1, 2, \ldots, \tau$

  Nature reveals c_t

  Alg plays x_t uniformly at random

Construct $\hat{V}(\pi)$ as above

$\hat{\pi} = \arg\max_{\pi} \hat{V}(\pi)$

for $t = \tau+1, \ldots, T$

  Nature reveals c_t

  $x_t = \hat{\pi}(c_t)$

$\mu(x|c_t) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$.

$\Longrightarrow$ w.p. $\geq 1-\delta$ $\forall \pi \in \Pi$ $|\hat{V}(\pi)-V(\pi)| \leq \underbrace{\sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{\tau}}}_{=:\varepsilon_\tau}$

$\pi_* \in \arg\max_{\pi\in\Pi} V(\pi)$

$$V(\hat{\pi}) = \underbrace{V(\hat{\pi}) - \hat{V}(\hat{\pi})}_{\geq -\varepsilon_\tau} + \underbrace{\hat{V}(\hat{\pi}) - \hat{V}(\pi_*)}_{\geq 0} + \underbrace{\hat{V}(\pi_*) - V(\pi_*)}_{\geq -\varepsilon_\tau} + V(\pi_*)$$

$$\geq V(\pi_*) - 2\varepsilon_\tau$$

$$\text{Regret} = \sum_{t=1}^{T} V(\pi_*) - V(\pi_t)$$

$$= \sum_{t=1}^{\tau} V(\pi_*) - V(\pi_t) + \sum_{t=\tau+1}^{T} V(\pi_*) - V(\pi_t)$$

$$\leq \tau \cdot 1 + (T-\tau)\underbrace{\left(V(\pi_*) - V(\hat{\pi})\right)}_{\leq 2\varepsilon_\tau}$$

$$= \tau + 2(T-\tau)\sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{\tau}}$$

$$= O\left(T^{2/3}\left(|\mathcal{X}|\log(2|\Pi|/\delta)\right)^{1/3}\right)$$

$$\tau = \left(|\mathcal{X}|T^2 \log(2|\Pi|/\delta)\right)^{1/3}$$

How do we find $\operatorname{argmax}_{\pi \in \Pi} \hat{V}(\pi)$ efficiently?

$$\hat{V}(\pi) = \frac{1}{3} \sum_{t=1}^{3} \frac{\mathbb{1}\{x_t = \pi(c_t)\}}{p_t} r_t$$

$$= \frac{1}{3} \sum_{t=1}^{3} \frac{(1 - \mathbb{1}\{x_t \neq \pi(c_t)\})}{p_t} r_t$$

$$= \frac{1}{3} \sum_{t=1}^{3} \frac{r_t}{p_t} - \frac{1}{3} \sum_{t=1}^{3} \mathbb{1}\{x_t \neq \pi(c_t)\} \frac{r_t}{p_t}$$

$$\operatorname{argmax}_{\pi} \hat{V}(\pi) = \operatorname{argmin}_{\pi} \frac{1}{3} \sum_{t=1}^{3} \mathbb{1}\{x_t \neq \pi(c_t)\} \frac{r_t}{p_t}$$

Think of each $\pi : \mathcal{C} \to \mathcal{X}$ as a classifier over classes $\{1, \ldots, |\mathcal{X}|\}$ and examples in $\mathcal{C}$.

$x_t \in \mathcal{X}$ is a "label". $c_t \in$ example, $\frac{r_t}{p_t}$ is weight

Minimize 0/1 loss on $\{(c_t, x_t, \frac{r_t}{p_t})\}_{t=1}^{3}$

$\uparrow$ weight of tth point

Ex. Let $f_\theta : \mathcal{C} \to \mathbb{R}^{\mathcal{X}}$ be a neural network parameterized by parameters $\theta \in \mathbb{R}^d$

$$\operatorname{loss}(\theta) = -\frac{1}{3} \sum_{t=1}^{3} \frac{r_t}{p_t} \log\left(\frac{\exp(f_\theta(c_t)[x_t])}{\sum_{x \in \mathcal{X}} \exp(f_\theta(c_t)[x])} \right)$$

Let $\hat{\theta}_3 = \arg\min_{\theta} \text{Loss}(\theta)$

When c_t arrives we play

$$x_t = \arg\max_{x \in \mathcal{X}} \left[f_{\hat{\theta}_3}(c_t) \right]_x$$

Zooming out: τ-greedy achieves $R_T \lesssim T^{2/3} \left(|\mathcal{X}| \log(\Pi/\delta) \right)^{1/3}$.

But $|C| = 1$ and $|\Pi| = |\mathcal{X}|$ in the standard MAB then we know $\sqrt{|\mathcal{X}| \cdot T}$ regret is possible.

$\longrightarrow$ For contextual Bandits, can we achieve $R_T \le \sqrt{|\mathcal{X}| T \cdot \log |\Pi|}$?

Answer: yes.

<u>An ellimination alg</u> for <u>stochastic Contextual Bandits</u>

Input Π, δ

$\hat{\Pi}_1 = \Pi$

for $\ell = 1, 2, \ldots$

$\varepsilon_\ell = 2^{-\ell}$, # measurements τ_ℓ TBD

regularization γ_ℓ TBD

$$Q_\ell = \arg\min_{Q \in \Delta_{\hat{\Pi}_\ell}} \max_{\pi \in \hat{\Pi}_\ell} \mathbb{E}_c \left[\frac{1}{Q^\gamma(\pi(c) | c)} \right]$$

where $Q^\gamma(x|c) := \gamma + (1 - |\mathcal{X}|\gamma) Q(x|c)$

$Q(x|c) := \sum_{\pi : \pi(c) = x} Q(\pi)$ $\longleftarrow$ prob dist over actions

$$T_\ell = \sum_{i=1}^{\ell} \tau_i$$

for $t = T_{\ell-1}+1, \ldots, T_\ell$ $\quad (T_\ell - T_{\ell-1} = \tau_\ell)$

Nature reveals c_t

Player plays $x_t \sim Q_\ell^\gamma(\cdot \mid c_t)$ to get reward r_t

$$\hat{V}_\ell(\pi) = \frac{1}{\tau_\ell} \sum_{t=1}^{\tau_\ell} \frac{\mathbb{1}\{\pi(c_t) = x_t\}}{p_t} r_t$$

$$\hat{\Pi}_{\ell+1} = \hat{\Pi}_\ell \setminus \{\pi \in \hat{\Pi}_\ell : \max_{\pi'} \hat{V}(\pi') - \hat{V}(\pi) \geq 2\varepsilon_\ell\}.$$

// Q_ℓ is chosen to minimize maximum variance of $\hat{V}(\pi)$:

$$Q_\ell \approx \arg\min_{Q \in \Delta_{\hat{\Pi}_\ell}} \max_{\pi \in \hat{\Pi}_\ell} \mathbb{E}\left[\left(\hat{V}_\ell(\pi) - V(\pi)\right)^2\right]$$

// Any $\pi \in \hat{\Pi}_\ell$ we have $V(\pi^*) - V(\pi) \leq 8\varepsilon_\ell$

$\Longrightarrow$ average regret incurred @ stage ℓ is $\uparrow$

<u>Lemma</u> For any policy Π and dist. over contexts

$$\min_{Q \in \Delta_\Pi} \max_{\pi \in \Pi} \mathbb{E}\left[\frac{1}{Q(\pi(c) \mid c)}\right] \leq |\mathcal{X}|.$$

Furthermore, if $\gamma < \frac{1}{2K}$ then

$$\min_{Q \in \Delta_\Pi} \max_{\pi \in \Pi} \mathbb{E}\left[\frac{1}{Q^\gamma(\pi(c) \mid c)}\right] \leq 2|\mathcal{X}|.$$

This $\nearrow$ lemma is a corollary of Kiefer-Wolfowitz which we will show next time.

Consequently we have by Bernstein

$$|\hat{V}_\ell(\pi) - V(\pi)| \leq \sqrt{\frac{(2|\mathcal{X}|)\cdot 2\log(2|\Pi|/\delta)}{\tau_\ell}} + \frac{2\log(2|\Pi|/\delta)}{\gamma_\ell 3 \tau_\ell}$$

$$\leq \sqrt{\frac{16|\mathcal{X}|\log(2|\Pi|/\delta)}{\tau_\ell}} =: \varepsilon_\ell$$

for $\gamma_\ell = \min\left\{\frac{1}{2|\mathcal{X}|}, \sqrt{\frac{2\log(|\Pi|/\delta)}{9|\mathcal{X}|\tau_\ell}}\right\}$

$$\tau_\ell = 16|\mathcal{X}|\log(2|\Pi|/\delta)\,\varepsilon_\ell^{-2}$$

<u>Lemma</u> For all $\ell = 1, 2, \ldots$ we have w.p. $\geq 1-\delta$, $\pi^* \in \hat{\Pi}_\ell$ and $\max_{\pi \in \hat{\Pi}_\ell} V(\pi^*) - V(\pi) \leq 8\varepsilon_\ell$.

<u>Theorem</u> W.p. $\geq 1-\delta$ we have

$$\sum_{t=1}^{T} V(\pi^*) - V(\pi_t) \leq c\sqrt{|\mathcal{X}|T\log(|\Pi|T/\delta)}.$$

<u>Proof</u> Fix $\nu \in [0,1)$ $\quad \gamma_\ell \approx \frac{\varepsilon_\ell}{|\mathcal{X}|}$

$$\sum_{t=1}^{T} V(\pi^*) - V(\pi_t) \leq \nu T + \sum_{\ell=1}^{\log_2 8\nu^{-1}} \left(\gamma_\ell |\mathcal{X}| + (1-\gamma_\ell |\mathcal{X}|) \cdot 8\varepsilon_\ell\right) T_\ell$$

$$\lesssim \nu T + \sum_{\ell=1}^{\log_2 \nu^{-1}} \varepsilon_\ell \cdot T_\ell$$

$$\lesssim \nu T + \sum_{\ell}^{\log_2 \nu^{-1}} \varepsilon_\ell^{-1} |\mathcal{X}| \log(|\Pi| T/\delta)$$

$$\leq \nu T + \frac{1}{\nu} |\mathcal{X}| \log(|\Pi| \cdot T/\delta)$$

minimize over ν yields the result.

Proof of Kiefer-Wolfowitz generalization.

Recall Kiefer-Wolfowitz says for $\mathcal{X} \subset \mathbb{R}^d$

$$\inf_{\lambda \in \Delta_\mathcal{X}} \max_{x \in \mathcal{X}} x^T \Big(\sum_{x'} \lambda_{x'} x' x'^T\Big)^{-1} x \leq d.$$

<u>Lemma</u> Let $\eta \in S$ be a R.V. and let $\phi(x, \eta) \in \mathbb{R}^d$ be a feature map.

Then

$$\min_{\lambda \in \Delta_{\mathcal{X}}} \max_{x \in \mathcal{X}} \mathbb{E}_z \left[\phi(x,z)^T \left(\sum_{x'} \lambda_{x'} \phi(x',z)\phi(x',z)^T \right)^{-1} \phi(x,z) \right] \le d$$

(Recovers KW w/ $\phi(x,z) = x$).

Key idea proof: $\frac{\partial}{\partial t} \log\det(A(t))$

$$= \mathrm{Trace}\left(A(0)^{-1} \frac{\partial}{\partial t} A(t) \right)$$

$$f(\lambda) = \mathbb{E}_z \left[\log\det\left(\sum_{x'} \lambda_{x'} \phi(x',z)\phi(x',z)^T \right) \right]$$

$\lambda^* = \mathrm{amax}\, f(\lambda)$

$$0 \ge \langle \nabla f(\lambda^*), e_x - \lambda^* \rangle$$

$$\ge \mathbb{E}_z \left[\phi(x,z)^T \left(\sum_{x'} \lambda_{x'} \phi(x',z)\phi(x',z)^T \right)^{-1} \phi(x,z) \right] - d$$

$\langle \nabla f(\lambda), \lambda \rangle \le d$ for all λ.

To prove $\underline{\min_{Q \in \Delta_\pi} \max_{\pi} \mathbb{E}_c \left[\frac{1}{Q(a(c)|c)} \right]} \le |\mathcal{X}|$

Consider actions $i = 1, 2, \ldots, |\mathcal{X}|$ and define

$\pi_c = e_{\pi(c)} \in \{0,1\}^{|\mathcal{X}|}$

(i.e. $\phi(\pi, c) = e_{\pi(c)}$)

$$\mathbb{E}_c\left[\pi_c^T \underbrace{\left(\sum_{\pi'} Q(\pi')\, \pi'_c \pi'^T_c\right)}_{\text{diagonal matrix w/ entries only summed over the } \pi}^{-1} \pi_c\right]$$

$\le |\mathcal{X}|$

Okay, $\exists Q \in \Delta_\Pi$ s.t. variance bounded by $|\mathcal{X}|$.

In linear bandits we found a sparse Q.

W/ contextual bandits $|\text{support}(Q^\gamma)| \simeq \min\{\frac{1}{\gamma}, |\mathcal{X}|\cdot|C|\}$

Computational Effic. Algorithm for context bandits.

Input Π, δ

Play uniform at random for $\log(\Pi)$ steps to get $\hat{\Delta}_0(\pi)$

for $\ell = 1, 2, \ldots$

$\varepsilon_\ell = 2^{-\ell}$, # measurements τ_ℓ TBD

regularization γ_ℓ TBD

Q_ℓ is any $Q \in \Delta_\Pi$ such that

1) $$\sum_\pi \hat{\Delta}_{\ell-1}(\pi) Q(\pi) \le C'' \varepsilon_\ell$$

2) $$\sqrt{\mathbb{E}_c\left[\frac{1}{Q^\gamma(\pi(c)|c)}\right] \frac{c \log(\Pi/\delta)}{\tau_\ell}} \le \varepsilon_\ell + \hat{\Delta}_{\ell-1}(\pi)$$

where $Q^\gamma(x|c) := \gamma + (1 - |\mathcal{X}|\gamma) Q(x|c)$

$Q(x|c) := \sum_{\pi:\pi(c)=x} Q(\pi)$ ← prob dist over actions

$T_\ell = \sum_{i=1}^{\ell} \tau_i$

for $t = T_{\ell-1}+1, \ldots, T_\ell$ $(T_\ell - T_{\ell-1} = \tau_\ell)$

Nature reveals c_t

Player plays $x_t \sim Q_\ell^\gamma(\cdot | c_t)$ to get reward r_t

$\hat{V}_\ell(\pi) = \frac{1}{\tau_\ell} \sum_{t=1}^{\tau_\ell} \frac{\mathbb{1}\{\pi(c_t) = x_t\}}{p_t} r_t$, $\hat{\Delta}_\ell(\pi) = \max_{\pi'} \hat{V}(\pi') - \hat{V}(\pi)$

By induction one can show

$$\hat{V}_\ell(\pi) - V(\pi) \le c' \max\{\varepsilon_\ell, \Delta(\pi)\}$$

$$\hat{\Delta}_\ell(\pi) \ge \hat{V}_\ell(\pi^*) - \hat{V}_\ell(\pi)$$

$$\ge \Delta(\pi) - 2 \max\{\varepsilon_\ell, \Delta(\pi)\}$$

which is $\ge \Delta(\pi)/2$ when $\varepsilon_\ell < \dfrac{\Delta(\pi)}{8}$

If $\varepsilon_\ell < \Delta(\pi)$ then $\hat{\Delta}_\ell(\pi) \approx \Delta(\pi)$

otherwise $\hat{\Delta}_\ell(\pi) \le \varepsilon_\ell$

$\Rightarrow$ If Q_ℓ feasible is found then average regret at stage ℓ is just

$$|X| \gamma_\ell + c'' \varepsilon_\ell \le (1 + c'') \varepsilon_\ell$$

$\Rightarrow$ Regret analysis is <u>identical</u> to the comp. inef. alg.

So we just need to find a feasible Q_ℓ efficiently.

Idea: Let $\Pi_j = \{\pi : \Delta(\pi) \le \varepsilon_j\}$. If P_j is

$$P_j = \operatorname*{arg\,min}_{P \in \Delta_\Pi : \text{support}(P) \in \Pi_j} \max_{\pi \in \Pi_j} \mathbb{E}\left[\frac{1}{P^\gamma(\pi(c) \mid c)}\right]$$

and

$$\overline{P}_\ell = \frac{1}{\sum_j \gamma_j} \sum_{j=1}^{\ell} \gamma_\ell P_j .$$

Some algebra shows

$$\sum_\pi \Delta_\pi \overline{P}_\ell(\pi) \le c\, \varepsilon_\ell \qquad (1)$$

and $\max_{\pi \in \Pi} \mathbb{E}\left[\frac{1}{\overline{P}^\ell(\pi(c) \mid c)}\right]$ satisfies (2).

Thus $\overline{P}_\ell$ is feasible for Q_ℓ

Find a π that violates (2)

using cost-sensitive/weighted

classification.

iterations per opt problem

$$\leq \frac{1}{\gamma \epsilon}.$$