$$\max_{i=1,\dots,n} \sum_{t=1}^{T} x_{t,i} - x_{t,I_t} \leq \sqrt{nT}$$

$$\mathbb{E}[x_{t,i}] = \mu_i$$

Online-to-batch conversion.

$$\hat{i} \sim \text{uniform}(\{I_t\}_{t=1}^{T})$$

$$\max_{j=1,\dots,n} \mathbb{E}[\mu_j - \mu_{\hat{i}}] = \max_{j=1,\dots,n} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x_{t,j} - x_{t,I_t}\right]$$

$$= \frac{\text{Regret}(T)}{T}$$

Contextual Bandits (Stochastic)

Input: Π, $\pi \in \Pi$, $\pi: \mathcal{C} \to [n]$

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

Nature reveals $c_t \overset{iid}{\sim} \mathcal{V}$

Player chooses $\pi_t \in \Pi$, $a_t := \pi_t(c_t)$

Receives reward $r_t \in [0,1]$: $\mathbb{E}[r_t \mid c_t] = r(c_t, a_t)$

minimize regret $\max_{\pi \in \Pi} \sum_{t=1}^{T} r(c_t, \pi(c_t)) - r(c_t, a_t)$

$V(\pi) = \mathbb{E}_{c, a \sim \pi(c)}[r(c, a)]$

$= \mathbb{E}_{c}[r(c, \pi(c))]$

Logging policy $\mu(\cdot \mid c_t) \in \Delta_n$, $p_t = \mu(a_t \mid c_t)$. Assume $\mu(a|c) > 0$ $\forall a, c$

Collect dataset $\{(c_t, a_t, r_t, p_t)\}_t$. Always log sampling probs p_t!

Forget context. Given $\{(a_t, r_t, p_t)\}_t$. Idea: learn some function

f: $f(a_t) \simeq r_t$ (e.g. $\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_t (f(a_t) - r_t)^2$)

recommend $\arg\max_a \hat{f}(a)$

Model the World

Given $\{(c_t, a_t, r_t, p_t)\}$

Learn f: $f(c_t, a_t) \simeq r_t$ (e.g. $\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_t (f(c_t, a_t) - r_t)^2$

recommend $\arg\max_a \hat{f}(c_t, a)$

Model the Bias

$$\hat{r}(c_t, a) := r_t \frac{\mathbb{1}\{a_t = a\}}{p_t}$$

$$\mathbb{E}[\hat{r}(c_t, a)] = \sum_{a'} \underbrace{\mathbb{P}(a_t = a')}_{=\mu(a'|c_t)} r(c_t, a') \frac{\mathbb{1}\{a' = a\}}{\mu(a'|c_t)}$$

$$= r(c_t, a)$$

$$\hat{V}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \underline{\hat{r}(c_t, \pi(c_t))}, \quad \mathbb{E}[\hat{V}(\pi)] = V(\pi)$$

$$\mathbb{E}[(\hat{V}(\pi) - V(\pi))^2] = \frac{1}{T^2}\mathbb{E}\sum_{t=1}^{T}(\hat{r}(c_t, \pi(c_t)) - r(c_t, \pi(c_t)))^2$$

$$\leq \frac{1}{T^2}\mathbb{E}\left[\sum_t \hat{r}(c_t, \pi(c_t))^2\right]$$

$$= \frac{1}{T^2}\sum_t \mathbb{E}_{c_t}\left[r_t^2 \frac{\mathbb{1}(a_t = \pi(c_t))}{p_t^2}\right]$$

$$\leq \frac{1}{T^2}\sum_t \mathbb{E}_{a_t, c_t}\left[\frac{\mathbb{1}\{a_t = \pi(c_t)\}}{\mu(a_t|c_t)^2}\right]$$

$$= \frac{1}{T}\mathbb{E}_{c\sim\nu}\left[\frac{1}{\mu(\pi(c)|c)}\right]$$

I want high prob bound on $|\hat{V}(\pi) - V(\pi)|$.

Hoeffding says that if $Z_t \in [a,b]$ and

$\mathbb{E}[Z_t] = 0$ then $\mathbb{P}\left(\frac{1}{T} \sum_{t=1}^{T} Z_t \geq |b-a| \sqrt{\frac{\log(1/\delta)}{2T}} \right) \leq \delta.$

$\hat{r}(c_t, \pi(c_t)) \in [0, \frac{1}{\min_{c,a} \mu(a|c)}]$

Bernstein's inequality says that if $Z_t \leq B$ and

$\mathbb{E}[Z_t] = 0$, $\mathbb{E}[Z_t^2] \leq \sigma^2$ then

$$\mathbb{P}\left(\frac{1}{T} \sum_{t=1}^{T} Z_t \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{T}} + \frac{2B \log(1/\delta)}{3T} \right) \leq \delta.$$

$\Longrightarrow$ w.p. $\geq 1-\delta$

$$|\hat{V}(\pi) - V(\pi)| \leq \sqrt{\underbrace{\mathbb{E}_c\left[\frac{1}{\mu(\pi(c)|c)}\right]}_{\leq n \text{ if uniform}} \cdot \frac{2\log(2/\delta)}{T}} + \underbrace{\max_{c,a} \frac{1}{\mu(a|c)}}_{\leq n \text{ if uniform}} \cdot \frac{2\log(2/\delta)}{3T}$$

Side note: There exists an estimator $\hat{\mu}: \mathbb{R}^T \to \mathbb{R}$

s.t. if $\mathbb{E}[Z_t] = 0$, $\mathbb{E}[Z_t^2] \leq \sigma^2$ then

$$\mathbb{P}\left(\hat{\mu}(\{Z_t\}_{t=1}^{T}) \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{T}} \right) \leq \delta.$$

See Catoni's estimator or median of means.

If $\mu(a|c) = \frac{1}{n}$ for c, a then for a $\pi \in \Pi$

$$|\hat{V}(\pi) - V(\pi)| \leq \sqrt{\frac{2n \log(2/\delta)}{T}} + \frac{2n \log(2/\delta)}{3T}$$

$$\leq \sqrt{\frac{4n \log(2/\delta)}{T}}$$

and $\forall$ π w.p. $\geq 1-\delta$

$$|\hat{V}(\pi) - V(\pi)| \leq \sqrt{\frac{4n \log(2|\Pi|/\delta)}{T}} = C$$

You collected a data set computed $\hat{V}(\pi)$ for all $\pi \in \Pi$. Which one do you choose?

Natural to choose $\hat{\pi}_{MLE} = \operatorname{argmax}_{\pi \in \Pi} \hat{V}(\pi)$

Define $\underbrace{C(\pi) = \sqrt{\mathbb{E}\left[\frac{1}{\mu(\pi(c)|c)}\right] \frac{2 \log(2|\Pi|/\delta)}{T}} + 2 \max_{c,a} \frac{1}{\mu(\pi(c)|c)} \log(}$

$$V(\hat{\pi}_{MLE}) \geq \hat{V}(\hat{\pi}_{MLE}) - C(\hat{\pi}_{MLE})$$

$$\geq \hat{V}(\pi_{*}) - C(\hat{\pi})$$

$$\geq V(\pi_*) - C(\pi_*) - C(\hat{\pi})$$

$$\geq V(\pi_*) - 2 \max_{\pi \in \Pi} C(\pi)$$

Pessimism

Define $\hat{\pi}_{pess} = \arg\max_{\pi \in \Pi} \hat{V}(\pi) - C(\pi)$

$$V(\hat{\pi}_{pess}) \geq \hat{V}(\hat{\pi}_{pess}) - C(\hat{\pi}_{pess})$$

$$\geq \hat{V}(\pi_*) - C(\pi_*)$$

$$\geq V(\pi_*) - 2C(\pi_*)$$

Doubly Robust Estimator

$$\hat{r}_{DR}(c_t, a) = \hat{f}(c_t, a) + (r_t - \hat{f}(c_t, a)) \frac{\mathbb{1}\{a_t = a\}}{p_t}$$

$$\mathbb{E}[\hat{r}_{DR}(c_t, a)] = r(c_t, a)$$

$$\hat{\pi}_{MLE} = \underset{\pi \in \Pi}{\arg\max}\ \hat{V}(\pi)$$

$$= \underset{\pi \in \Pi}{\arg\max}\ \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{1}\{a_t = \pi(c_t)\}}{p_t} r_t$$

$$= \underset{\pi}{\arg\max}\ \frac{1}{T} \sum_{t=1}^{T} \frac{(1 - \mathbb{1}\{a_t \neq \pi(c_t)\})}{p_t} r_t$$

$$= \underset{\pi}{\arg\max}\ \frac{1}{T} \sum_{t=1}^{T} - \frac{\mathbb{1}\{a_t \neq \pi(c_t)\}}{p_t} r_t$$

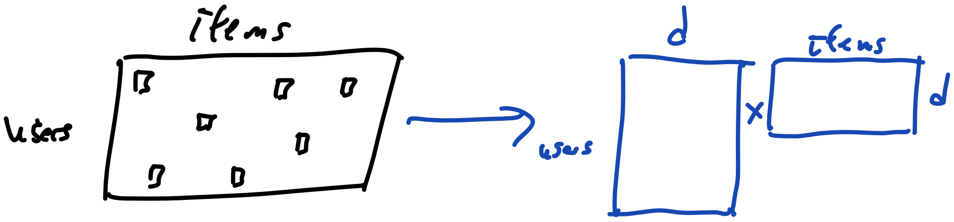
$$= \underset{\pi}{\arg\min}\ \frac{1}{T} \sum_{t=1}^{T} \frac{r_t}{p_t} \mathbb{1}\{a_t \neq \pi(c_t)\}$$

<u>Example</u> Let $\phi : \mathcal{C} \times [n] \to \mathbb{R}^d$, for each $\theta \in \mathbb{R}^d$, there exists

$\pi_\theta \in \Pi$: $\pi_\theta(c_t) = \underset{i=1,\dots,n}{\arg\max} \langle \phi(c_t, i), \theta \rangle$

$$\hat{\theta} = \underset{\theta}{\arg\min} \sum_{t} \frac{r_t}{p_t} - \log\left(\frac{\exp(\langle \phi(c_t, a_t), \theta \rangle)}{\sum_i \exp(\langle \phi(c_t, i), \theta \rangle)} \right)$$

$$\theta_{k+1} = \theta_k + \eta_k \sum_{t=1}^{T} \frac{r_t}{p_t} \nabla_\theta \log\left(\frac{\exp(\langle \phi(c_t, a_t), \theta \rangle)}{\sum_i \exp(\langle \phi(c_t, i), \theta \rangle)} \right)$$



Policy Gradient (REINFORCE)

$$V(\pi) = \mathbb{E}_{\substack{c \sim \nu \\ a \sim \pi(\cdot|c)}} \left[r(c, a) \right]$$

Parameterize our policies : $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^d\}$

$$\nabla_\theta V(\pi_\theta) = \nabla_\theta \sum_c \nu_c \sum_a \pi_\theta(a|c)\, r_{c,a}$$

$$= \sum_c \nu_c \sum_a r_{c,a} \nabla_\theta \pi_\theta(a|c)$$

$$= \sum_c \nu_c \sum_a r_{c,a} \pi_\theta(a|c) \cdot \underbrace{\frac{\nabla_\theta \pi_\theta(a|c)}{\pi_\theta(a|c)}}_{\nabla_\theta \log(\pi_\theta(a|c))}$$

$$= \sum_c \sum_a \nu_c \pi_\theta(a|c)\, r_{c,a} \cdot \nabla_\theta \log(\pi_\theta(a|c))$$

$$= \mathbb{E}\left[r_t \cdot \nabla_\theta \log(\pi_{\theta_t}(a_t|c_t)) \right]$$

At time t, $a_t \sim \pi_{\theta_t}(c_t)$

$$\theta_{t+1} = \theta_t + \eta_t \nabla_\theta \log(\pi_{\theta_t}(a_t|c_t))$$

n arms (no context)

$$[\pi_\theta]_i = \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)}$$

$$\left[\nabla_{\theta_i} \log(\pi_\theta(i))\right]_k = \nabla_\theta \left(\theta_i - \log \sum_j e^{\theta_j}\right)$$

$$= \mathbb{1}\{k=i\} - \frac{e^{\theta_k}}{\sum_j e^{\theta_j}}$$

$$\theta_0 = 0 \qquad\qquad = \mathbb{1}\{k=i\} - \pi_\theta(k)$$

for $t=1,2,\dots$

Player draws $I_t \sim \dfrac{e^{\theta_{t,i}}}{\sum_j e^{\theta_{t,j}}}$

Nature reveals r_t, $\mathbb{E}[r_t] = \mu_{I_t}$

$$\theta_{t+1} = \theta_t + \eta\left(e_{I_t} - \pi_\theta(\cdot)\right) r_t$$

Input: Policy set Π such that $\pi : \mathcal{X} \to [n]$ for all $\pi \in \Pi$, confidence level $\delta \in (0,1)$.
Let $\widehat{\Pi}_1 \leftarrow \Pi, \ell \leftarrow 1, T_0 \leftarrow 0$
while $|\widehat{\Pi}_\ell| > 1$ **do**
$\epsilon_\ell = 2^{-\ell}$, $\tau_\ell = \lceil 16n\epsilon_\ell^{-2}\log(2|\Pi|T/\delta)\rceil$, $\gamma_\ell = \min\{\frac{1}{2n}, \sqrt{\frac{\log(2|\Pi|T/\delta)}{9n\tau_\ell}}\}$, $T_\ell = T_{\ell-1} + \tau_\ell$
$Q_\ell = \arg\min_{Q \in \triangle_{\widehat{\Pi}_\ell}} \max_{\pi \in \widehat{\Pi}_\ell} \mathbb{E}_C\left[-\frac{1}{\mu(\pi(C)|C)}\right]$
s.t. $\mu_\ell(x|c) = \gamma + (1-\gamma n)\sum_{\pi \in \widehat{\Pi}_\ell : \pi(c) = x} Q(\pi)$
for $t = T_{\ell-1}+1, \ldots, T_\ell$
Observe context c_t
Play $x_t \sim \mu_\ell(\cdot|c_t)$, set $p_t = \mu_\ell(x_t|c_t)$ and observe reward $r_t = v(c_t, x_t) + \eta_t$
Set $\widehat{V}_\ell(\pi) = \frac{1}{T_\ell - T_{\ell-1}} \sum_{t \in (T_{\ell-1}, T_\ell]} r_t \frac{\mathbf{1}\{\pi(c_t) = c_t\}}{p_t}$
$\widehat{\Pi}_{\ell+1} \leftarrow \widehat{\Pi}_\ell \setminus \{\pi \in \widehat{\Pi}_\ell | \max_{\pi' \in \widehat{\Pi}_\ell} \widehat{V}_\ell(\pi') - \widehat{V}_\ell(\pi) \geq 2\epsilon_\ell\}$
$t \leftarrow t + 1$
Output: Π_{t+1}

The following lemma is somewhat of a generalization of Kiefer-Wolfowitz.

Lemma 19. *Let $\xi \in \Xi$ be a random variable and let $\phi : \mathcal{X} \times \Xi \to \mathbb{R}^d$. Then*

$$\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \in \mathcal{X}} \mathbb{E}_\xi\left[\phi(x,\xi)^\top \left(\sum_{x' \in \mathcal{X}} \lambda_{x'}\phi(x',\xi)\phi(x',\xi)^\top\right)^\dagger \phi(x,\xi)\right] \leq d,$$

with equality if $\operatorname{dimspan}(\{\phi(x,\xi) : x \in \mathcal{X}\}) = d$ *for all* $\mathcal{X} \in \mathcal{X}$.

$$\min_{Q \in \Delta_\Pi} \max_{\pi \in \Pi} \mathbb{E}_c\left[\underbrace{e_{\pi(c)}^T \left(\sum_{\pi' \in \Pi} Q(\pi')\, e_{\pi'(c)} e_{\pi'(c)}^T\right)^{-1} e_{\pi(c)}}_{} \right] \leq n$$

$$[e_{\pi(c)}]_i = \begin{cases} 1 & \text{if } \pi(c) = i \\ 0 & \text{o.w.} \end{cases} \qquad = \frac{1}{\sum_{\pi' : \pi'(c) = \pi(c)} Q(\pi')}$$

$$e_{\pi(c)} \in \{0,1\}^n$$

$$\mu_\ell(a|c) = \gamma + (1-\gamma n)\sum_{\pi' : \pi'(c) = a} Q(\pi') \qquad \text{Assume } \gamma < \frac{1}{2n}$$

$$\mathbb{E}_c\left[\frac{1}{\mu_\ell(\pi(c)|c)}\right] = \mathbb{E}_c\left[\frac{1}{\gamma + (1-\gamma n)\sum_{\pi' : \pi'(c) = \pi(c)} Q(\pi')}\right]$$

$$\leq \mathbb{E}_c\left[\frac{1}{\frac{1}{2}\sum_{\pi':\pi'(c)=\pi(c)} Q(\pi')}\right]$$

$$\leq 2n$$

Let $\hat{Q} = \operatorname{argmin}_{Q\in\Delta_\Pi} \max_{\pi\in\Pi} \mathbb{E}_c\left[\frac{1}{\sum_{\pi':\pi'(c)=\pi(c)} Q(\pi')(1-\gamma n)+\gamma}\right]$, $\exists$ dist. over contexts and a policy set Π s.t. any minimizer has $\text{support}(\hat{Q}) = |C|n \approx \frac{1}{\gamma} \wedge |\Pi|$.

Moreover for any Π and context dist. $\exists$ a minimizer $\hat{Q}$ s.t. $\text{Support}(\hat{Q}) \leq \frac{1}{\gamma}$

$$|\hat{V}_\ell(\pi) - V(\pi)| \leq \sqrt{\underbrace{\mathbb{E}\left[\frac{1}{\mu(\pi(c)|c)}\right]}_{\leq 2n} \cdot \frac{2\log(2|\Pi|T/\delta)}{T_\ell}} + \frac{2\log(2|\Pi|T/\delta)}{\gamma_\ell T_\ell}$$

$$= \sqrt{\frac{4n\log(2|\Pi|T/\delta)}{T_\ell}} + \sqrt{\frac{4n\log(2|\Pi|T/\delta)}{T_\ell}}$$

$$\leq \varepsilon_\ell$$

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

$$\leq 2\varepsilon_\ell \implies \pi_* \in \hat{\Pi}_\ell \text{ for all } \ell \geq 1$$

$$\max_{\pi \in \hat{\Pi}_\ell} V(\pi_*) - V(\pi) \leq 8\varepsilon_\ell$$

$$\pi_t(c) = \begin{cases} \text{uniform w/ prob } \gamma_n \\ Q(\cdot \mid c) \text{ w/ prob } (1-\gamma_n) \end{cases}$$

$$\sum_{t=1}^{T} r(c_t, \pi_*(c_t)) - r(c_t, \pi_t(c_t))$$

$$= \sum_{\ell=1} \sum_{t=T_{\ell-1}+1}^{T_\ell} r(c_t, \pi_*(c_t)) - r(c_t, \pi_t(c_t))$$

$$\leq \sum_\ell \tau_\ell (\gamma_\ell n \cdot 1 + (1-\gamma_n) \cdot \varepsilon_\ell)$$

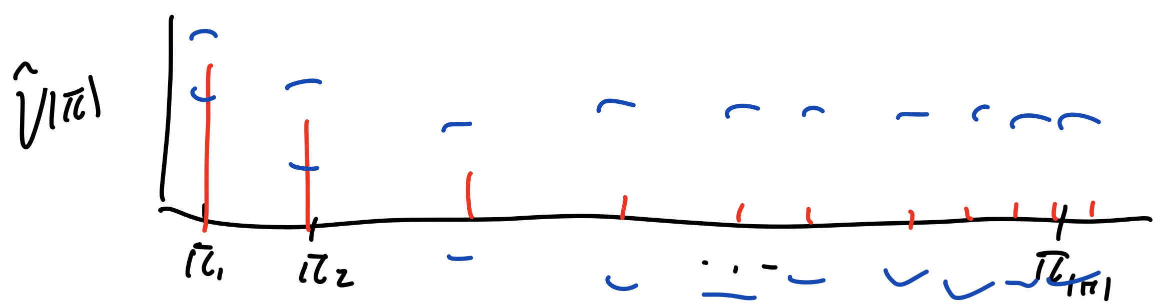
$$\vdots$$

$$\lesssim \nu T + c\,(\Delta \vee \nu)^{-1} n \log(|\Pi| \cdot T / \delta)$$

for any $\nu \geq 0$. Minimize over ν we get

$$\text{Regret} \lesssim \sqrt{n T \log(|\Pi| T / \delta)}$$

Practical Implementation



$$\min_{Q \in \mathcal{S}_\pi} \max_{\pi} \sqrt{\frac{\mathbb{E}\left[\frac{1}{\mu(\pi(c)|c)}\right]}{3_e}} - \left(\Delta(\pi) \vee \varepsilon_e\right)$$

$$\Delta(\pi) = V(\pi_*) - V(\pi)$$

$$2\sqrt{\frac{c}{g}} = \max_{p \geq 0} \; cp + \frac{1}{gp}$$

$$c = \frac{1}{gp^2} \qquad p = \sqrt{\frac{1}{gc}}$$

$$\frac{c}{g} = \min_{\alpha : \alpha \leq g} \frac{c}{\alpha} \qquad \mathcal{L}(\alpha, \lambda) = \frac{c}{\alpha} + \lambda(\alpha - g)$$

$$\mathbb{E}\left[\frac{1}{\mu(\pi(c)|c)}\right] \Longrightarrow \frac{1}{\alpha} - \lambda \mathbb{E}\left[\mu(\pi(c)|c)\right]$$