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

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

Online-to-batch conversion.

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

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

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

Contextual Bandits (Stochastic)

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

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

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

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

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

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. $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_t (f(a_t) - r_t)^2$)

recommend $\operatorname{argmax}_a \hat{f}(a)$

Model the World

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

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

recommend $\operatorname{argmax}_a \hat{f}(c_t, a)$

Model the Bias

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

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

$$= r(c_t, a)$$

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

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

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

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

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

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

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

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

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

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. $\geq 1-\delta$

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

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

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 $\hat{\pi}_{pess} = \underset{\pi \in \Pi}{\operatorname{argmax}} \hat{V}(\pi) - C(\pi)$

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

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

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

Doubly Robust Estimator

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

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

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

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

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

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

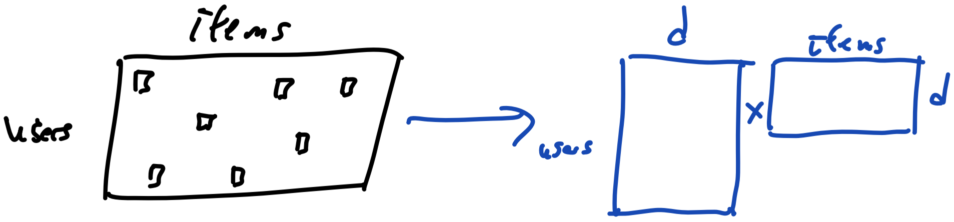
$$= \arg\min_{\pi} \frac{1}{T}\sum_{t=1}^{T} \frac{r_t}{p_t} \mathbb{1}\{a_t \neq \pi(c_t)\}$$

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

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

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

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



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)

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

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

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

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

$\theta_0 = 0$

for $t = 1, 2, \ldots$

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

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

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