Contextual Bandits

Input actions/arms $\mathcal{X}$

for $t = 1, 2, \ldots$

Nature reveals context $c_t \sim \nu$

Player chooses action $x_t \in \mathcal{X}$

Nature reveals reward $V_t(c_t, x_t) \in [0,1]$

where $\mathbb{E}[V_t(c_t, x_t) | c_t, x_t] = v(c_t, x_t)$

for some unknown $v : \mathcal{C} \times \mathcal{X} \to [0,1]$.

Suppose context space was finite $|\mathcal{C}| < \infty$.

Suppose we run $|\mathcal{C}|$ separate bandit algorithms in parallel, one for each $c \in \mathcal{C}$.

$$\max_x \sum_{t=1}^{T} \mathbb{1}\{c_t = c\}\big(v(c_t, x) - v(c_t, x_t)\big)$$

$$\lesssim \sqrt{|\mathcal{X}| T_c \log(T_c/\delta)}$$

$$\leq \sqrt{|\mathcal{X}| T_c \log(T/\delta)}$$

w.p. $\geq 1-\delta$ using elim. or UCB

$$T_c = \sum_{t=1}^{T} \mathbb{1}\{c_t = c\}$$

$$\sum_{t=1}^{T} \max_x \big(v(c_t, x) - v_t(c_t, x_t)\big)$$

$$= \sum_{t=1}^{T} \sum_{c \in \mathcal{C}} \max_x \mathbb{1}\{c_t = c\}\big(v(c_t, x) - v(c_t, x_t)\big)$$

$$= \sum_{c \in \mathcal{C}} \max_x \sum_{t=1}^{T} \mathbb{1}\{c_t = c\}\big(v(c_t, x) - v(c_t, x_t)\big)$$

$$\lesssim \sum_{c \in \mathcal{C}} \sqrt{|\mathcal{X}| \cdot T_c \cdot \log(T/\delta)}$$

Cauchy-Schwartz $\leq \sqrt{\left(\sum_{c\in\mathcal{C}} T_c\right)\left(\sum_{c\in\mathcal{C}} |\mathcal{X}| \log(T/\delta)\right)}$

$= \sqrt{|\mathcal{C}||\mathcal{X}| T \log(T/\delta)}$

What if we <u>ignored</u> context and played just a single bandit?

$$\max_x \sum_{t=1}^{T} V(c_t, x) - V(c_t, x) \leq \sqrt{|\mathcal{X}| T \log(T/\delta)}$$

Total Reward acts like

$$\sum_{t=1}^{T} V(c_t, x_t) \geq \max_x \sum_{t=1}^{T} V(c_t, x) - \sqrt{|\mathcal{X}| T \log(T/\delta)}.$$

But compare this to $|\mathcal{C}|$-bandit approach that achieves

$$\sum_{t=1}^{T} V(c_t, x_t) \geq \sum_{t=1}^{T} \max_x V(c_t, x) - \sqrt{|\mathcal{X}||\mathcal{C}| T \log(T/\delta)}$$

If $T \ll |\mathcal{C}|$ then may be far better to play just a single bandit then a bandit-per-context b/c latter will never learn.

Now suppose we have a set of policies $\Pi = \{\pi : \pi : \mathcal{C} \to \mathcal{X}\}$.

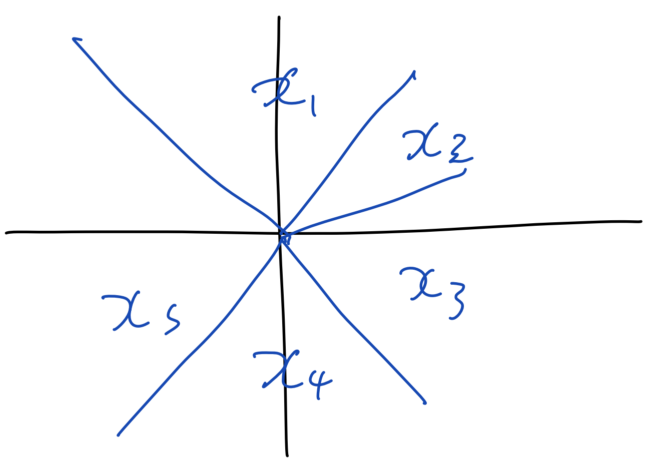
At each time t, choose $\pi_t \in \Pi$ and play $a_t = \pi(c_t)$.

Ex. suppose $c_t \in \mathbb{R}^d$, then take Π to be linear multiclass classifiers

$$\pi \in \Pi \iff \pi(c_t) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \langle w_x, c_t \rangle$$

for some $\{w_x\}_{x \in \mathcal{X}}$

$|\mathcal{X}| = 5$



Goal is to minimize policy regret!

$$\max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} v(c_t, \pi(c_t)) - v(c_t, \underbrace{x_t}_{= \pi_t(c_t)})\right]$$

Ex. The bandit-per-context strategy is encoded w/ $|\Pi| = |\mathcal{X}|^{|C|}$

Ex. The single-bandit (ignore context) $|\Pi| = |\mathcal{X}|$

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

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

Could think of the policies themselves as "arms" so that I have $|\Pi|$ arms, and then play any MAB alg (UCB) for T times where at each time I choose a policy and recieve reward.

for $t = 1, 2, \ldots$

Player chooses $\pi_t \in \Pi$

Nature reveals reward $\underset{=}{r_t} = v_t(c_t, \pi(c_t))$

Regret $\max_{\pi} \sum_{t=1}^{T} v_t(c_t, \pi(c_t)) - r_t \leq \sqrt{|\Pi|\, T \log(T)}$

$\uparrow$

Regret bound is independent of # actions, # contexts

But horrible if $|\Pi|$ is large.

If Π is a class of multi-class classifiers (e.g. linear, neural nets) then we know for classification we can learn a good classifier from historical data. What if I have historical data for a deployed policy?

Ex. USPS observes destination location of a package (context) and then they choose a route to deliver it (action). Reward signal is time or cost.

Mail has been delivered for long time by some (implicit) policy $\hat{\pi}$ resulting in data

$$\{(c_t, x_t, r_t)\}_{t=1}^{T}$$

context: destination, weather, ⋮

action: path to destination among many

reward: $-$cost

Want to learn new policy from old data.

Idea: learn a function $f : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$ in $\mathcal{F}$

s.t. $\hat{f} = \underset{f \in \mathcal{F}}{\arg\min} \sum_{t=1}^{T} \left(f(c_t, a_t) - r_t \right)^2$

$\uparrow$ neural net, decision tree, whatever

To learn a new policy:

$$\pi_{new}(c) = \underset{x \in \mathcal{X}}{\arg\max} \ \hat{f}(c, x).$$

"Model the world"

What could go wrong?

USPS collected data under policy π_{old}, meaning in response to context c_t they took action $a_t = \pi_{old}(c_t)$.

By observing $v_t(c_t, a_t)$ that means we did not observe $v_t(c_t, a)$ for $a \neq a_t$.

$\Longrightarrow$ We may not be able to learn $\underset{a}{\arg\max}\, v(c_t, a)$!

Solution? Randomization will give you coverage of all actions and you can get consistent estimation of $v(c, a)$ $\forall c, a$.

In general is necessary.

But sometimes you can get lucky:

Suppose you defined a feature map

$$\phi : \mathcal{X} \times \mathcal{C} \to \mathbb{R}^d \quad \text{and}$$

$$\mathcal{F} = \{ f(c,a) = \langle \phi(c,a), \theta \rangle, \ \forall \theta \in \mathbb{R}^d \}$$

Assume $\exists\ \theta_* \in \mathbb{R}^d : V(c,a) = \langle \phi(c,a), \theta_* \rangle$.

Assume $r_t = V_t(c_t, a_t) = V(c_t, a_t) + \eta_t$

$\nwarrow \mathcal{N}(0,1)$

Then $\displaystyle \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(f(c_t, a_t) - r_t \right)^2$

is equiv. to

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{t=1}^{T} \left(\langle \phi(c_t, a_t) - r_t \rangle \right)^2$$

$$= \left(\sum_{t}^{T} \phi(c_t, a_t)\, \phi(c_t, a_t)^T \right)^{-1} \sum_{t=1}^{T} \phi(c_t, a_t)\, r_t$$

$$= \theta_* + \left(\sum_{t}^{T} \phi(c_t, a_t)\, \phi(c_t, a_t)^T \right)^{-1} \sum_{t=1}^{T} \phi(c_t, a_t)\, \eta_t$$

$$\|\hat{\theta} - \theta_*\|_{\left(\sum_t^3 \phi(c_t,a_t)\phi(c_t,a_t)^T\right)} \lesssim \text{const.}$$

vector martingale
self-normalized bound.

This linear example demonstrates that naively fitting a function to data <u>can</u> <u>work</u>.

But not always.

Suppose the data was collected randomly s.t. $\mathbb{P}(a_t = a \mid c_t) = \mu(a \mid c_t)$ for some distribution μ.

Moreover the probability $P_t = \mu(a_t \mid c_t)$ was logged as well to give you

$$\{(c_t, a_t, P_t, r_t)\}_{t=1}^{3}$$

Consider the IPS (inverse propensity $\overset{\text{score}}{\wedge}$ estimator)

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

$$\mathbb{E}_{r_t, a_t}\left[\hat{V}_t(c_t, a) \mid c_t\right]$$

$$= \sum_{a' \in \mathcal{X}} \mu(a' | c_t) \mathbb{E}\left[\frac{\mathbb{1}\{a_t = a\} r_t}{p_t} \,\middle|\, c_t, a_t = a'\right]$$

$$= \sum_{a' \in \mathcal{X}} \mu(a' | c_t) \mathbb{E}\left[\frac{\mathbb{1}\{a_t = a\} r_t}{\mu(a_t | c_t)} \,\middle|\, c_t, a_t = a'\right]$$

$$= \sum_{a' \in \mathcal{X}} \cancel{\mu(a' | c_t)} \frac{\mathbb{1}\{a' = a\} v(c_t, a')}{\cancel{\mu(a' | c_t)}}$$

$$= \sum_{a' \in \mathcal{X}} \mathbb{1}\{a' = a\} v(c_t, a')$$

$$= v(c_t, a)$$

$$\hat{V}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \hat{V}_t(c_t, \pi(c_t))$$

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

$$= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{c_t \sim \gamma}[V(c_t, \pi(c_t))]$$

$$= V(\pi)$$

Great! Output $\hat{\pi} = \underset{\pi \in \Pi}{\operatorname{argmax}}\ \hat{V}(\pi)$.

What is $V(\pi_*) - V(\hat{\pi})$?

$$\mathbb{E}[(\hat{V}(\pi) - V(\pi))^2]$$

$$= \frac{1}{T^2}\sum_{t=1}^{T} \mathbb{E}_{c_t}\Big[\underbrace{\mathbb{E}[(\hat{V}_t(c_t, \pi(c_t)) - V(c_t, \pi(c_t)))^2 | c_t]}_{\le \frac{1}{\mu(\pi(c_t)|c_t)}}\Big]$$

Fact: for any R.V. X, $\underset{a}{\operatorname{argmin}}\ \mathbb{E}[(X-a)^2] = \mathbb{E}[X]$

$$\mathbb{E}\left[\left(\hat{v}_t(c_t, a) - v(c_t, a)\right)^2 \middle| c_t\right]$$

$$\leq \mathbb{E}\left[\hat{v}_t(c_t, a)^2 \middle| c_t\right]$$

$$= \mathbb{E}\left[\frac{\mathbb{1}\{a_t = a\}\, r_t^2}{p_t^2} \middle| c_t\right]$$

Assume $r_t \in [0,1]$

$$= \sum_{a' \in \mathcal{X}} \mu(a' | c_t)\, \mathbb{E}\left[\frac{\mathbb{1}\{a_t = a\}\, r_t^2}{p_t^2} \middle| c_t, a_t = a'\right]$$

$$= \sum_{a' \in \mathcal{X}} \mu(a' | c_t)\, \mathbb{E}\left[\frac{\mathbb{1}\{a_t = a\}\, r_t^2}{\mu(a_t | c_t)^2} \middle| c_t, a_t = a'\right]$$

$$\leq \sum_{a'} \frac{\mathbb{1}\{a' = a\}}{\mu(a' | c_t)} \qquad (r_t \in [0,1])$$

$$= \frac{1}{\mu(a | c_t)}$$

$$\mathrm{Var}\left[v_t(c_t, a) \middle| c_t\right] \leq \frac{1}{\mu(a | c_t)}$$

$$\hat{V}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \hat{V}_t(c_t, \pi(c_t))$$

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

$$\mathbb{E}[\hat{V}(\pi)] = V(\pi) \qquad \forall \pi \in \Pi$$

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

$$= \sum_c \nu_c \frac{1}{\mu(\pi(c)|c)}$$

$\mu(\pi(c)|c)$ can be arbitrarily small, which means

p_t can be arbitrarily small, which means

$\hat{V}(\pi)$ is heavy-tailed (not sub-Gaussian).

<u>Assume</u> $\mu(a|c) \geq \gamma \quad \forall c, a.$ (said another way $\gamma := \min_{c,a} \mu(a|c)$)

Now $\underbrace{\frac{\mathbb{1}\{a_t = \pi(c_t)\}}{p_t} r_t}_{= X_t}$ are unbiased, have variance $\leq \mathbb{E}\left[\frac{1}{\mu(\pi(c)|c)}\right]$ (which could be small)

$r_t \in [0,1]$ and have support in $[0, \frac{1}{\gamma}]$.

Bernstein's Inequality. Let $X_1, \ldots, X_n$ be independent R.V.s w/ $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \le \sigma^2$ and $|X_i| \le b$. Then

$$\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \le \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} + \frac{2b\log(1/\delta)}{3n}$$

w.p. $\ge 1-\delta$.

Compare to Hoeffding. All we have is that $|X_i| \le b$ so Hoeffding says

$$\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \le b\sqrt{\frac{2\log(1/\delta)}{n}}$$

w.p. $\ge 1-\delta$.

$$\hat{V}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \underbrace{\frac{\mathbb{1}\{a_t = \pi(c_t)\}}{p_t} r_t}_{= X_t}$$

"model the bias"

$$\mathbb{E}[X_t] = \mathbb{E}_{c_t}\left[\mathbb{E}\left[\frac{\mathbb{1}\{a_t = \pi(c_t)\}}{p_t} r_t \,\middle|\, c_t\right]\right]$$

$$= \mathbb{E}_{c_t}[v(c_t, \pi(c_t))] = V(\pi)$$

$$\mathbb{E}\left[(X_t - \mathbb{E}[X_t])^2\right]$$

$$= \mathbb{E}_{c_t}\left[\mathbb{E}\left[\left(\frac{\mathbb{1}\{a_t = \pi(c_t)\}}{p_t} r_t - V(\pi)\right)^2 \Big| c_t\right]\right]$$

$$\leq \mathbb{E}\left[\frac{1}{\mu(\pi(c)|c)}\right] = \sigma^2$$

$$X_t \in [0, \tfrac{1}{\gamma}) \qquad b = \frac{1}{\gamma}$$

$\Longrightarrow$ Bernstein applied: For any $\pi \in \Pi$

$$|\hat{V}(\pi) - V(\pi)| \leq \sqrt{\mathbb{E}\left[\frac{1}{\mu(\pi(c)|c)}\right]\frac{2\log(2/\delta)}{T}} + \frac{2\log(2/\delta)}{3T\gamma}$$

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

To hold for all $\pi \in \Pi$ take union bound $\delta \mapsto \frac{\delta}{|\Pi|}$.

We said uniform was a good idea

for exploration: $\mu(a|c) = \frac{1}{|\mathcal{X}|}$ ($\mathcal{X}$ actions)

That implies that for all $\pi \in \Pi$ simultaneously

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

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

$$\leq \sqrt{\frac{4|\mathcal{X}| \log(2|\Pi|/\delta)}{3}}$$

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

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

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

$$\leq 2 \max_{\pi \in \Pi} \left| \hat{V}(\pi) - V(\pi) \right|$$

$$\leq 2 \sqrt{\frac{4|\mathcal{X}| \log(2|\Pi|/\delta)}{3}}.$$

Great: you will show on hw a explore-then-commit strat that exploits this to achieve $(T)^{2/3}$ regret.

Can we do better? yes.

But first how do you solve

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

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

$$= \underset{\theta \in \Theta}{\arg\max} \; \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{1}\{a_t = \pi_\theta(c_t)\} r_t}{p_t}$$

$$= \underset{\theta \in \Theta}{\arg\min} \; \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{1}\{a_t \neq \pi_\theta(c_t)\}}{p_t} r_t$$

$$\hat{f} = \underset{\theta \in \Theta}{\arg\min} \; \frac{1}{T} \sum_{t=1}^{T} \left(\frac{r_t}{p_t}\right) \text{softmax}\left(f_\theta(c_t), a_t \right)$$

$$\hat{\pi}(c) = \hat{f}(c)$$

Doubly robust estimator

Given $\{(c_t, a_t, p_t, r_t)\}_{t=1}^{T}$ and $\hat{f} : \mathcal{C} \times \mathcal{X} \to [0,1]$

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

Unbiased $\mathbb{E}[\hat{V}_t^{(DR)}(c_t, a) | c_t] = V(c_t, a) \quad \forall \hat{f}$

(IPS estimator takes $\hat{f}(c,a)=0$)