A $\sqrt{T}$ regret algorithm for contextual Bandits

By defining $\mu(a|c)$ we can estimate $\hat{V}(\pi)$ like

$$\text{Variance}\left(\hat{V}(\pi)\right) = \mathbb{E}\left[\frac{1}{\mu(a|c)}\right]$$

Elimination alg.

Input: Π, $\delta \in (0,1)$

Init: $\Pi_1 = \Pi$, $\ell = 1$, $T_0 = 0$

while $|\Pi_\ell| > 1$ $\qquad T_\ell = T_{\ell-1} + \tau_\ell$

$$\varepsilon_\ell = 2^{-\ell}, \quad \tau_\ell = \lceil 32|A|\varepsilon_\ell^{-2}\log(2|\pi|/\delta)\rceil, \quad \gamma_\ell = \min\left\{\frac{1}{2|A|}, \sqrt{\frac{\log(2|\Pi|/\delta)}{4|A|\tau_\ell}}\right\},$$

$$\mu_\ell(a|c) = \gamma_\ell + (1-\gamma_\ell|A|)\textstyle\sum_{\pi\in\Pi_\ell}\lambda_\pi^\ell\, \mathbb{1}\{\pi(c)=a\}$$

where $\displaystyle \lambda^\ell = \underset{\lambda\in\Delta_{\Pi_\ell}}{\arg\min}\ \max_{\pi\in\Pi_\ell} \mathbb{E}_c\left[\frac{1}{\sum_{\pi'\in\Pi_\ell}\mathbb{1}\{\pi(c)=\pi'(c)\}}\right]$

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

Nature reveals context c_t

Algorithm plays $a_t \sim \mu_\ell(\cdot|c_t)$, set $p_t = \mu_\ell(a_t|c_t)$

and obtains r_t w/ $\mathbb{E}[r_t|c_t,a_t] = v(c_t,a_t)$

$$\hat{V}_\ell(\pi) = \frac{1}{T_\ell - T_{\ell-1}}\sum_{t=T_{\ell-1}+1}^{T_\ell}\frac{\mathbb{1}\{\pi(c_t)=a_t\}}{p_t}\, r_t \qquad \forall \pi$$

$$\Pi_{\ell+1} = \Pi_\ell \setminus \left\{\pi\in\Pi_\ell : \max_{\pi'\in\Pi_\ell}\hat{V}_\ell(\pi') - \hat{V}_\ell(\pi) > 2\varepsilon_\ell\right\}$$

$\ell \leftarrow \ell + 1$

Want to choose $\mu_\ell(a|c)$ and τ_ℓ s.t.

$|\hat{V}_\ell(\pi) - V(\pi)| \le \varepsilon_\ell \quad \forall\pi$ (with high prob)

and incure $\le \varepsilon_\ell$ regret per time step t.

What if $\mu_\ell(a|c) \propto \mathbb{1}\{\exists \pi \in \Pi_\ell : \pi(c) = a\}$

Very reasonable for exploration but we cannot (easily) bound the regret of the policy, even though we know (or will show) $|\hat{V}_\ell(\pi) - V(\pi)| \leq 8\varepsilon_\ell$.

Alternatively, If $|\hat{V}_\ell(\pi) - V(\pi)| \leq \varepsilon_\ell$ then by elim rule

$\max_{\pi \in \Pi_\ell} V(\pi_*) - V(\pi) \leq 8\varepsilon_\ell$. If I choose any $\pi_t \in \Pi_\ell$ randomly according to some distribution $\lambda \in \Delta_{\Pi_\ell}$ then

$$\mathbb{E}_{\pi_t \sim \lambda}\left[V(\pi_*) - V(\pi_t)\right]$$

$$= \sum_{\pi \in \Pi_\ell} \lambda_\pi \left(V(\pi_*) - V(\pi)\right)$$

$$\leq \sum_{\pi \in \Pi_\ell} \lambda_\pi \, 8\varepsilon_\ell = 8\varepsilon_\ell$$

Idea: draw $\pi_t \sim \lambda_\pi$ and play $a_t = \pi_t(c_t)$

$$\mu_\ell(a|c) = \sum_{\pi \in \Pi_\ell} \lambda_\pi \mathbb{1}\{\pi(c) = a\}$$

How do we choose λ? Well we want to minimize ξ_ℓ: estimate each π ε_ℓ accuracy.

If $\gamma_\ell = \min_{a,c} \mu_\ell(a|c)$ then

$$|\hat{V}_\ell(\pi) - V(\pi)| \le \sqrt{\mathbb{E}_c\left[\frac{1}{\mu_\ell(\pi(c)|c)}\right]\frac{2\log(2|\Pi|/\delta)}{\tau_\ell}} + \frac{2\log(2|\Pi|/\delta)}{2\tau_\ell\gamma_\ell}$$

$$\le \sqrt{\mathbb{E}_c\left[\frac{1}{\mu_\ell(\pi(c)|c)}\right]\frac{16\log(2|\Pi|/\delta)}{\tau_\ell}}$$ Want $\le \varepsilon_\ell$ w/ minimum τ_ℓ

Holds for conditions on γ_ℓ.

Choose $\lambda \in \Delta_{\Pi_\ell}$ to minimize, where $\mu_\ell(a|c) = \sum_\pi \lambda_\pi \mathbb{1}\{\pi(c)=a\}$

Claim for any set of policies $\Pi : \pi : \mathcal{C} \to \mathcal{A}$, $\gamma \le \frac{1}{2|\mathcal{A}|}$

$$\min_{\lambda \in \Delta_\Pi} \max_{\pi \in \Pi} \mathbb{E}_c\left[\frac{1}{\sum_{\pi'} \lambda_{\pi'} \mathbb{1}\{\pi'(c) = \pi(c)\}}\right] \le |\mathcal{A}|$$

Moreover if $\mu(a|c) = \gamma + (1 - |\mathcal{A}|\gamma)\sum_{\pi'} \lambda_{\pi'} \mathbb{1}\{\pi'(c) = a\}$

$$\max_{\pi \in \Pi} \mathbb{E}_c\left[\frac{1}{\mu(a|c)}\right] \le \max_{\pi \in \Pi} \mathbb{E}_c\left[\frac{2}{\sum_{\pi'} \lambda_{\pi'} \mathbb{1}\{\pi'(c) = \pi(c)\}}\right] \le 2|\mathcal{A}|$$

Suppose it is. Then

$$\max_\pi |\hat{V}_\ell(\pi) - V(\pi)| \le \max_\pi \sqrt{\mathbb{E}_c\left[\frac{1}{\mu_\ell(\pi(c)|c)}\right]\frac{2\log(2|\Pi|/\delta)}{\tau_\ell}} + \frac{2\log(2|\Pi|/\delta)}{2\tau_\ell\gamma_\ell}$$

For our choice of γ_ℓ $\to$

$$\le \max_\pi \sqrt{\max\left\{\mathbb{E}_c\left[\frac{1}{\mu_\ell(\pi(c)|c)}\right], |\mathcal{A}|\right\}\frac{16\log(2|\Pi|/\delta)}{\tau_\ell}}$$

$$\le \sqrt{|\mathcal{A}|\frac{32\log(2|\Pi|/\delta)}{\tau_\ell}}$$

$\leq \varepsilon_\ell$ by our choice of $\mathfrak{Z}_\ell$.

Assume this holds for all ℓ

1) Show $\pi_A \in \Pi_\ell$ for all ℓ

2) Show $\max_{\pi \in \Pi_\ell} V(\pi_A) - V(\pi) \leq 8\varepsilon_\ell$

3) Prove regret bound

$$\text{Regret } \nu T + \sum_{\ell=1}^{L \wedge \log_2 \frac{1}{\nu}} \mathfrak{Z}_\ell \left(|A|\gamma_\ell \cdot 1 + (1-|A|\gamma_\ell)\cdot 8\varepsilon_\ell \right)$$

$$\leq \nu T + \sum_{\ell=1}^{L \wedge \log_2 \frac{1}{\nu}} \underbrace{\mathfrak{Z}_\ell |A| \gamma_\ell}_{\text{Way these are choosen}} + 8\varepsilon_\ell \mathfrak{Z}_\ell$$

$$\leq \nu T + c \sum_{\ell=1}^{L \wedge \log_2 \frac{1}{\nu}} \varepsilon_\ell \mathfrak{Z}_\ell$$

$$\leq \nu T + c|A|\log(|\Pi|/\delta) \sum_{\ell=1}^{L \wedge \log \frac{1}{\nu}} 2^\ell$$

$$\leq \nu T + |A|\log(|\Pi|/\delta)/\nu$$

Optimize over $\nu \Rightarrow \sqrt{|A| T \log(|\Pi|/\delta)}$