Online Learning

for $t = 1, 2, \ldots$

Nature reveals $z_t \in \mathcal{Z}$

Player predicts $\hat{y}_t$

Player receives loss $\ell(y_t, \hat{y}_t)$ (observes y_t)

Given $\mathcal{F}$ s.t. $f \in \mathcal{F}$ $f: \mathcal{Z} \to \mathbb{R}$

Regret online learner

$$R_T = \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (f(z_t) - y_t)^2$$

Intuition: If $z_t \in \mathcal{Z}$, $y_t = \langle \theta_*, z_t \rangle + \eta_t$ $\swarrow N(0,1)$ $\mathcal{F} = \{\langle z, \theta \rangle : \theta \in \mathbb{R}^d\}$

then $R_T \leq d \log(T)$. if

$$\hat{y}_t = \langle \hat{\theta}_t, z_t \rangle \quad w/ \quad \hat{\theta}_t = \left(\sum_s z_s z_s^T\right)^{-1} \sum_s z_s y_s$$

If $z_t \overset{iid}{\sim} \nu$, $y_t = f_*(z_t) + \eta_t$ for some $f_* \in \mathcal{F}$, it

is tempting to set $\hat{f}_t = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_s (y_s - f(z_s))^2$

and predict $\hat{y}_t = \hat{f}_t(z_t)$.

Works for linear functions but <u>not</u> in general.

Assume you have an online learner so that

$$\sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (f(z_t) - y_t)^2 \leq \underbrace{R_{Sq}(T)}_{\simeq\, C_{\mathcal{F}} \log(T)}.$$

Then $\exists$ algorithm for contextual bandits that feeds the online learner $z_t = (c_t, x_t)$ and has regret

$$\leq \sqrt{KT \cdot R_{sq}(T)} + \sqrt{KT \log(1/\delta)}$$

w.p. $\geq 1-\delta$, where $K = |\mathcal{X}|$.

Square CB Algorithm (Foster, Rakhlin 2020)

for $t = 1, 2, \ldots$

Nature reveals context c_t

For each $x \in \mathcal{X}$ ask oracle for $\hat{y}_t(c_t, x)$

Set $b_t = \underset{x}{\operatorname{argmin}}\ \hat{y}_t(c_t, x)$

For $x \neq b_t$ set $p_{t,x} = \dfrac{1}{\mu + \gamma(\hat{y}_{t,x} - \hat{y}_{t,b_t})}$

$$p_{t,b_t} = 1 - \sum_{x \neq b_t} y_{t,x}$$

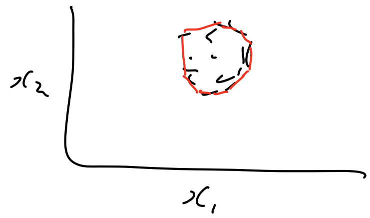
Sample and play $x_t \sim p_t$, receive y_t.

Feed (c_t, x_t, y_t) to online learner.

Opt choices of $\gamma = \sqrt{\dfrac{KT}{R_{sq}(T)}}$

Contextual Bandits Street fight: Practical methods

ε-Greedy

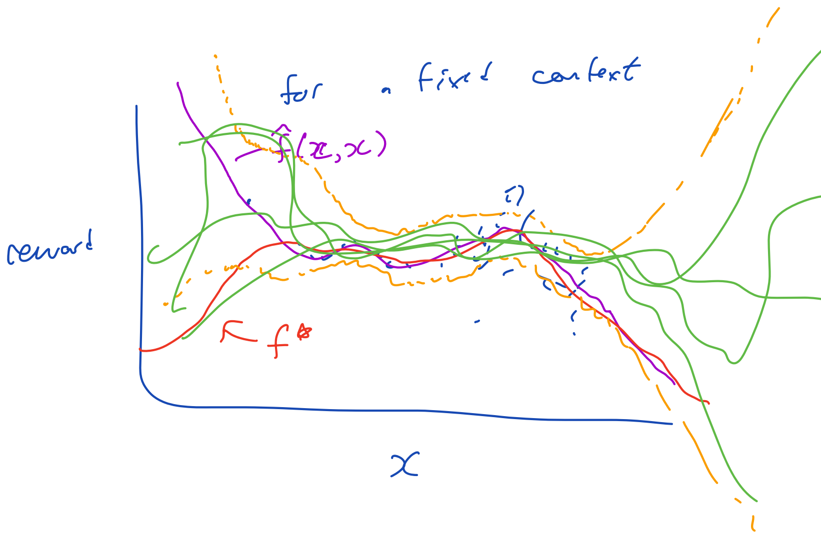


for $t=1,2,\dots$

$$\hat{f}_t = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{s=1}^{t-1} (y_s - f(c_s,x_s))^2$$

Nature reveals c_t

Play $x_t = \underset{x\in\mathcal{X}}{\operatorname{argmax}}\ \hat{f}_t(c_t, x)$ w.p. $1-\varepsilon$, and $x_t \sim \text{uniform}(\mathcal{X})$ w.p. ε



Posterior sampling

Input P_0 prior over $\mathcal{F}$

for $t=1,2,\dots$

$\tilde{f}_t \sim P_{t-1}$

Nature reveals c_t

Player plays $\underset{x\in\mathcal{X}}{\operatorname{argmax}}\ \tilde{f}_t(c_t, x)$, observes y_t

Update posterior $P_{t-1} \to P_t$ w/ (c_t, x_t, y_t)

Follow the Perturbed leader

for $t=1,2,\dots$

Set $\hat{y}_s = y_s + \xi_s$, $\xi_s \leftarrow N(0,\sigma^2)$ $\forall s<t$

$$\hat{f}_t = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{s=1}^{t-1} (\hat{y}_s - f(c_s,x_s))^2$$

Nature reveals c_t

Play $x_t = \underset{x\in\mathcal{X}}{\operatorname{argmax}}\ \hat{f}_t(c_t, x)$

Given surrogate reward function $\hat{f} : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$

and context c_t, playing $x_t = \arg\max_{x \in \mathcal{X}} \hat{f}(c_t, x)$

is often intractable if $\dim(\mathcal{X})$ is high.

Learn a policy $\pi_\theta : \mathcal{C} \to \mathcal{X}$

neural network parameterized by $\theta \in \Theta$

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{t=1}^{T} \hat{f}(c_t, \pi_\theta(c_t))$$

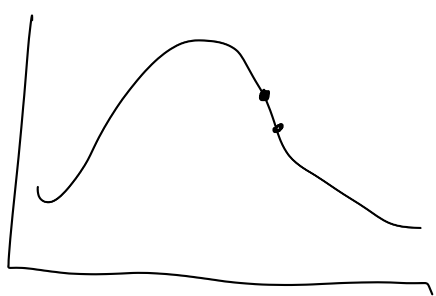
How!?

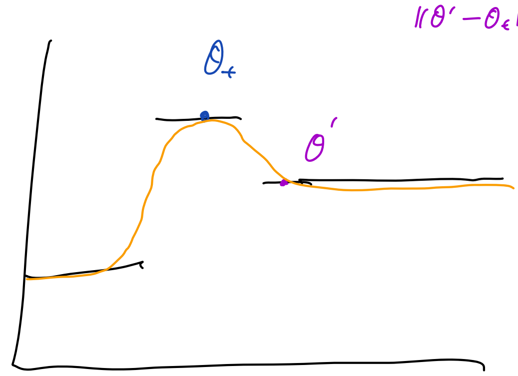
If $\hat{f}$ is differentiable and and $\mathcal{X}$ continuous/differentiable

then $\nabla_\theta \hat{f}(c_t, \pi_\theta(c_t)) = \nabla_z \hat{f}(c_t, z) \frac{\partial \pi_\theta(c_t)}{\partial \theta}$.

$$g_t = \frac{L(\theta') - L(\theta_t)}{\|\theta' - \theta_t\|}(\theta' - \theta_t)$$

$L(\theta)$





You can always maximize a fn using derivative free opt

but it can be <u>very</u> slow.

Consider a <u>stochastic</u> policy

$$\pi_\theta : \mathcal{C} \to \Delta_{\mathcal{X}}$$

$$\mathbb{E}_{x_t \sim \pi_\theta(x|c_t)}\left[\sum_t \hat{f}(c_t, x_t)\right] = J(\theta)$$

for $t = 1, 2, \ldots$ draw $x_t \sim \pi_\theta(\cdot | c_t)$

$$g := \nabla_\theta \sum_t \log\left(\underline{\pi_\theta(x_t | c_t)}\right) \cdot \hat{f}(c_t, x_t)$$

$$\mathbb{E}[g] = \mathbb{E}\left[\nabla_\theta \sum_t \log\left(\pi_\theta(x_t|c_t)\right) \cdot \hat{f}(c_t, x_t)\right]$$

$$= \mathbb{E}\left[\sum_t \frac{1}{\pi_\theta(x_t|c_t)} \cdot \nabla_\theta \pi_\theta(x_t|c_t) \cdot \hat{f}(c_t, x_t)\right]$$

$$= \sum_t \int_x \cancel{\pi_\theta(x|c_t)} \frac{1}{\cancel{\pi_\theta(x|c_t)}} \cdot \nabla_\theta \pi_\theta(x|c_t) \cdot \hat{f}(c_t, x)\, dx$$

$$= \sum_t \int_x \nabla_\theta \pi_\theta(x|c_t)\, \hat{f}(c_t, x)\, dx$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x_t \sim \pi_\theta(\cdot|c_t)}\left[\sum_t \hat{f}(c_t, x_t)\right]$$

$$= \nabla_\theta \sum_t \int \pi_\theta(x|c_t)\, \hat{f}(c_t, x)\, dx$$

$$= \sum_t \int \nabla_\theta \pi_\theta(x \mid c_t) \hat{f}(c_t, x)\, dx$$

$$\mathbb{E}[g] = \nabla_\theta J(\theta)$$

Ex. x is discrete then

$$\pi_\theta(x \mid c) = \frac{e^{h_\theta(x,c)}}{\sum_x e^{h_\theta(x,c)}}$$

Ex. x is continuous

$$\pi_\theta(x \mid c) = \mathcal{N}(x \,;\, h_\theta(c), I\sigma^2)$$

$$\nabla_\theta \log(\pi_\theta(x \mid c)) \propto \nabla_\theta \| h_\theta(c) - x \|^2$$

$$\propto (h_\theta(c) - x) \nabla_\theta h_\theta(c)$$