Bandits 2

Chen-Yu Wei

The Full-Information MAB

Given: set of actions $\mathcal{A} = \{1, \ldots, A\}$

For time $t = 1, 2, \ldots, T$:

Environment decides the reward of all actions $R_t(1), R_t(2), \ldots, R_t(A)$ without revealing

The learner chooses an action a_t

Environment reveals the noisy reward $r_t(a) = R_t(a) + w_t(a)$ **of all actions**

$$\text{Regret} = \max_a \sum_{t=1}^{T} R_t(a) - \sum_{t=1}^{T} R_t(a_t)$$

$$\sum_{t=1}^{T} \max_a R_t(a) \quad \text{(harder)}$$

KL-Regularized Policy Updates

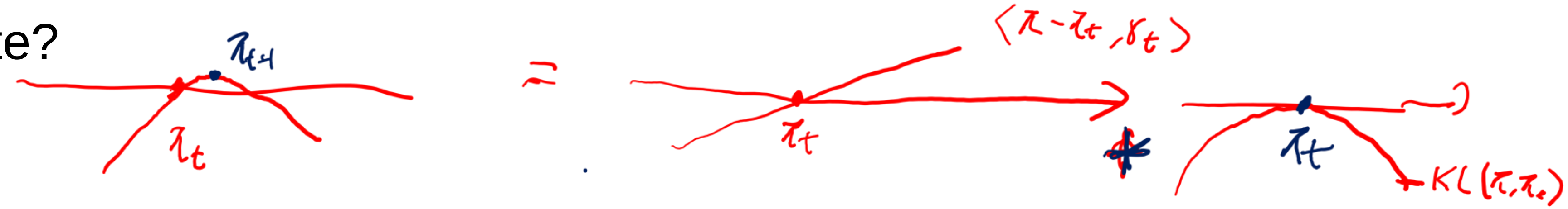
$$a_t \sim \pi_t \rightarrow r_t = \begin{bmatrix} r_t(1) \\ \vdots \\ r_t(A) \end{bmatrix}$$

$$\pi_t = \begin{bmatrix} \pi_t(1) \\ \vdots \\ \pi_t(A) \end{bmatrix}$$

$$\pi_{t+1} = \underset{\pi\in\Delta(\mathcal{A})}{\operatorname{argmax}} \left\{ \langle \pi - \pi_t, r_t \rangle - \frac{1}{\eta}\mathrm{KL}(\pi, \pi_t) \right\}$$

$$= \underset{\pi\in\Delta(\mathcal{A})}{\operatorname{argmax}} \left\{ \underbrace{\sum_a (\pi(a) - \pi_t(a)) r_t(a)}_{\text{The Improvement of } \pi \text{ over } \pi_t} \underbrace{- \frac{1}{\eta}\sum_a \pi(a)\log\frac{\pi(a)}{\pi_t(a)}}_{\text{Distance between } \pi \text{ and } \pi_t} \right\} \boxed{\langle \pi, r_t \rangle}$$

Why regularize the update?



KL-Regularized Policy Updates

Maintaining stability for stochastic or adversarial environments

Time	1	2	3	4	5	6	…
$R_t(1)$	0.5	0	1	0	1	0	…
$R_t(2)$	0	1	0	1	0	1	…

Follow the leader: $a_t = \max_{a \in \mathcal{A}} \left\{ \sum_{i=1}^{t-1} r_i(a) \right\}$

KL-Regularized Policy Updates

Exponential weight updates

$$\pi_{t+1} = \underset{\pi \in \Delta(\mathcal{A})}{\operatorname{argmax}} \left\{ \langle \pi - \pi_t, r_t \rangle - \frac{1}{\eta} \mathrm{KL}(\pi, \pi_t) \right\} \Leftrightarrow \pi_{t+1}(a) = \frac{\pi_t(a)\, e^{\eta r_t(a)}}{\sum_{b \in \mathcal{A}} \pi_t(b)\, e^{\eta r_t(b)}}$$

The equivalence is shown in HW0

Regret Bound for Exponential Weight Updates

Theorem.

Assume that $\eta r_t(a) \leq 1$ for all t, a. Then EWU

$$\pi_{t+1} = \operatorname*{argmax}_{\pi \in \Delta(\mathcal{A})} \left\{ \langle \pi - \pi_t, r_t \rangle - \frac{1}{\eta} \mathrm{KL}(\pi, \pi_t) \right\}$$

ensures for any $a^\star \in \mathcal{A}$,

≤ 1

$$\sum_{t=1}^{T} (r_t(a^\star) - \langle \pi_t, r_t \rangle) \leq \frac{\log A}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{A} \pi_t(a) r_t(a)^2$$

$\sqrt{AT}$ ~ bandit

If $|r_t(a)| \leq 1$ and $\eta \leq 1 \Rightarrow \mathbb{E}\left[\sum_{t=1}^{T} (R_t(a^\star) - R_t(a_t))\right] \leq \frac{\log A}{\eta} + \eta T \approx \sqrt{(\log A) T}$

Questions and Discussions

- How is exponential weight update related to Boltzmann's exploration?

$$\pi_{t+1}(a) \propto \pi_t(a)\, e^{\eta r_t(a)} \propto \pi_{t-1}(a)\, e^{\eta r_{t-1}(a)} \cdot e^{\eta r_t(a)} \cdots \propto e^{\eta \sum_{s=1}^{t} r_s(a)} = e^{\eta t \cdot \hat{R}_t(a)}$$

$$\boxed{\hat{R}_t(a) = \frac{1}{t} \sum_{s=1}^{t} r_s(a)}$$

$$\lambda_t = \eta t$$

$$\pi_{t+1}(a) \propto e^{\lambda_t \hat{R}_t(a)}$$

Questions and Discussions

- Why do we care about regret against a **fixed** action when the reward function is changing?
  - Environments where reward function is mostly stationary, but occasionally being changed adversarially
  - When we discuss about MDP, we will re-use this theorem but with R_t replaced by the "Q-function" of the policy used by the learner (and the policy of the learner changes over time)
  - This framework is suitable for a lot of other applications: game theory, constrained optimization, boosting, etc.

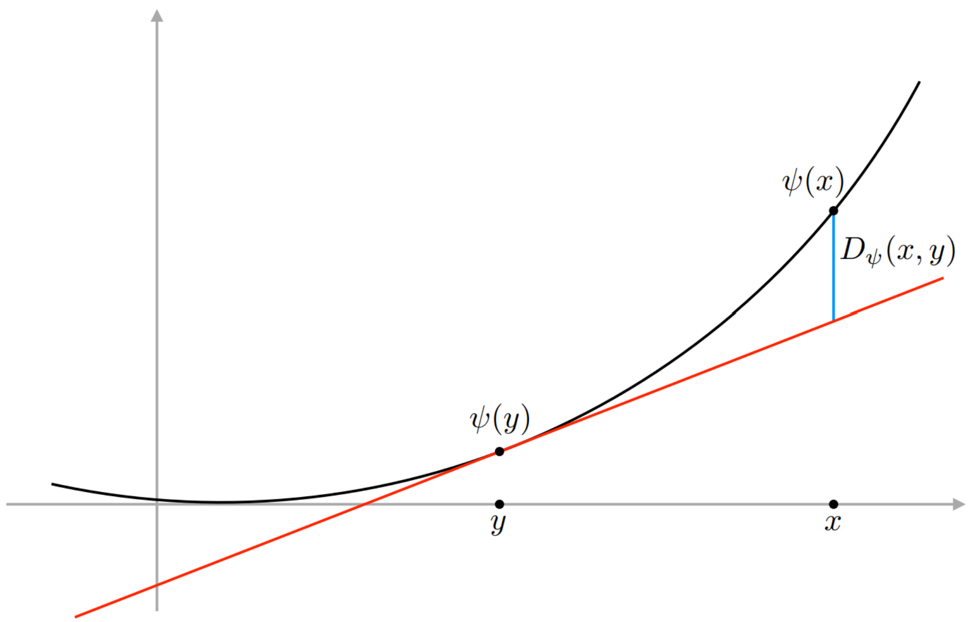
Exponential Weight Update ∈ Mirror Ascent

General form of **Mirror Ascent**:

Usually, $r_t = \nabla f_t(x_t)$ for some function f_t that we want to maximize

$$x_{t+1} = \underset{x\in\Omega}{\operatorname{argmax}} \left\{ \langle x - x_t, r_t \rangle - \frac{1}{\eta} \underbrace{D_\psi(x, x_t)} \right\}$$

Bregman divergence with respect to a convex function ψ



$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$$

Exponential Weight Update ∈ Mirror Ascent

Special cases of **Mirror Ascent**: $$x_{t+1} = \underset{x \in \Omega}{\operatorname{argmax}} \left\{ \langle x - x_t, r_t \rangle - \frac{1}{\eta} D_\psi(x, x_t) \right\}$$

$\boldsymbol{\psi(x)}$	$\boldsymbol{D_\psi(x, y)}$	**Update Rule**	
$\frac{1}{2}\|x\|_2^2$	$\frac{1}{2}\|x - y\|_2^2$	$x_{t+1} = \mathcal{P}_\Omega(x_t + \eta r_t)$ **Gradient ascent**	
$\sum_a x(a) \log x(a)$ **Negative entropy**	$\sum_a x(a) \log \frac{x(a)}{y(a)}$	$x_{t+1}(a) = \frac{x_t(a) e^{\eta r_t(a)}}{\sum_b x_t(b)\, e^{\eta r_t(b)}}$	(for distributions)
$\sum_a \log \frac{1}{x(a)}$	$\sum_a \left(\frac{x(a)}{y(a)} - \log \frac{x(a)}{y(a)} - 1 \right)$	$\frac{1}{x_{t+1}(a)} = \frac{1}{x_t(a)} - \eta r_t(a) + \gamma_t$	(for distributions)

Normalization factor

Regret Analysis for Exponential Weights

Theorem.
Assume that $\eta r_t(a) \leq 1$ for all t, a. Then EWU

$$\pi_{t+1} = \underset{\pi \in \Delta(\mathcal{A})}{\operatorname{argmax}} \left\{ \langle \pi - \pi_t, r_t \rangle - \frac{1}{\eta} \mathrm{KL}(\pi, \pi_t) \right\}$$

ensures for any $a^\star \in \mathcal{A}$,

$$\sum_{t=1}^{T} \left(\underbrace{r_t(a^\star)} - \langle \pi_t, r_t \rangle \right) \leq \frac{\log A}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{A} \pi_t(a) r_t(a)^2$$

$$\pi^* = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{at the } a^*\text{'s arm}$$

$$\langle \pi^*, r_t \rangle = r_t(a^*)$$

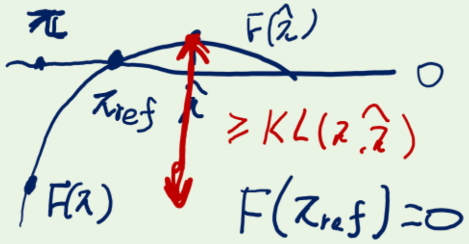
Regret Analysis for Exponential Weights

Useful Lemma

For fixed π_{ref} and v, define

We will apply this lemma with
$\pi_{\mathrm{ref}} = \pi_t, \qquad v = \eta r_t, \qquad \hat{\pi} = \pi_{t+1}$

$$F(\pi) = \langle \pi - \pi_{\mathrm{ref}}, v \rangle - \mathrm{KL}(\pi, \pi_{\mathrm{ref}})$$



and let $\hat{\pi} = \max_{\pi} F(\pi)$

(1) $F(\hat{\pi}) \geq F(\pi) + \mathrm{KL}(\pi, \hat{\pi})$ for any π

(2) If $v(a) \leq 1$ for all a, then $F(\hat{\pi}) \leq \langle \pi_{\mathrm{ref}}, v^2 \rangle = \sum_a \pi_{\mathrm{ref}}(a) v(a)^2$

(1) holds for all Bregman divergence

(2) is specific to KL divergence (but has counterpart for other divergence)

Regret Analysis for Exponential Weights

$$F(\pi) = \langle \pi - \pi_t , \eta r_t \rangle - KL(\pi, \pi_t)$$

$$\pi_{t+1} = \underset{\pi}{\operatorname{argmax}}\, F(\pi)$$

① $$F(\pi_{t+1}) = \langle \pi_{t+1} - \pi_t , \eta r_t \rangle - KL(\pi_{t+1}, \pi_t)$$

$$\geq \underbrace{\langle \pi^* - \pi_t , \eta r_t \rangle}_{\text{regret at time } t} - KL(\pi^*, \pi_t) + KL(\pi^*, \pi_{t+1}) = F(\pi^*) + KL(\pi^*, \pi_{t+1})$$

② $$\langle \pi^* - \pi_t , \eta r_t \rangle \leq \underbrace{F(\pi_{t+1})}_{\boxed{\leq \eta^2 \sum_a \pi_t(a) r_t(a)^2}} + KL(\pi^*, \pi_t) - KL(\pi^*, \pi_{t+1})$$

$$\sum_{t=1}^{T} \langle \pi^* - \pi_t , r_t \rangle \leq \eta \sum_t \sum_a \pi_t(a) r_t(a)^2 + \frac{1}{\eta} \underbrace{KL(\pi^*, \pi_1)}_{\log A} \cancel{- KL(\pi^*, \pi_{T+1})}$$