

Bandits 2

Chen-Yu Wei

The Full-Information MAB

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

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

Environment decides the reward of all actions $R_t(1), R_t(2), \dots, R_t(A)$ 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

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

$$\pi_t = \begin{pmatrix} \pi_t(i) \\ \vdots \\ \pi_t(A) \end{pmatrix}$$

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

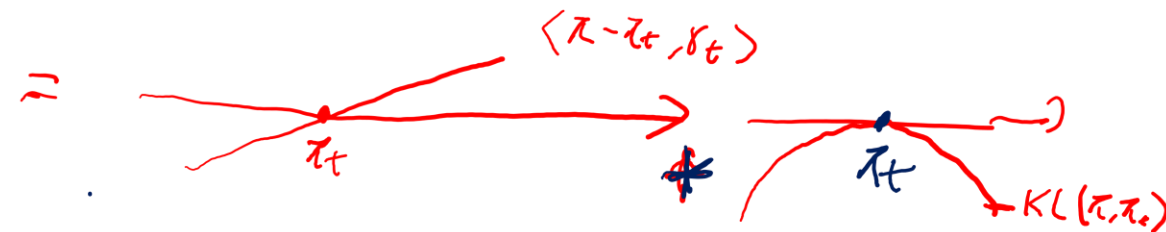
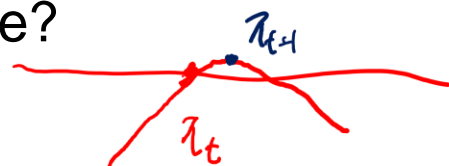
$$= \operatorname{argmax}_{\pi \in \Delta(\mathcal{A})} \left\{ \underbrace{\sum_a (\pi(a) - \pi_t(a)) r_t(a)}_{\text{The Improvement of } \pi \text{ over } \pi_t} - \underbrace{\frac{1}{\eta} \sum_a \pi(a) \log \frac{\pi(a)}{\pi_t(a)}}_{\text{Distance between } \pi \text{ and } \pi_t} \right\}$$

$$\langle \pi, r_t \rangle$$

The Improvement of π over π_t

Distance between π and π_t

Why regularize the update?



KL-Regularized Policy Updates

Maintaining stability for stochastic or adversarial environments

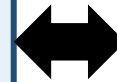
Time	1	2	3	4	5	6	...
$R_t(1)$	0.5	0	1	0	1	0	...
$R_t(2)$	0	1	0	1	0	1	...

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

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



$$\pi_{t+1}(a) = \frac{\pi_t(a) e^{\eta r_t(a)}}{\sum_{b \in \mathcal{A}} \pi_t(b) e^{\eta r_t(b)}}$$

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

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

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

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

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

Handwritten notes:
 \sqrt{AT} - bandit
 $\sqrt{\frac{\log A}{T}}$

Questions and Discussions

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

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

$$\lambda_t = \eta t$$

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

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

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 \in Mirror Ascent

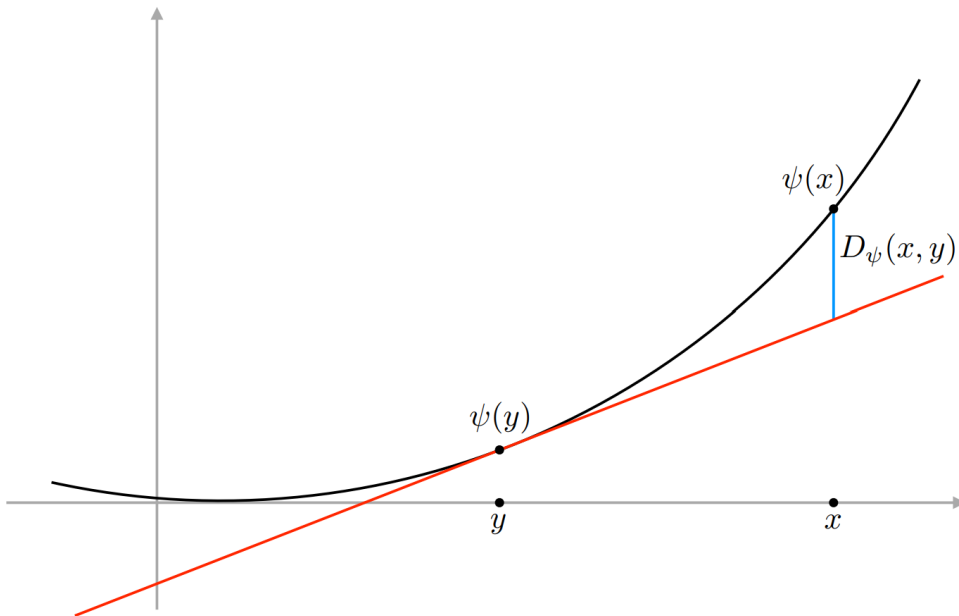
General form of **Mirror Ascent**:

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

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

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 \in Mirror Ascent

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

$\psi(x)$	$D_\psi(x, y)$	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

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

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

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

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

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

Regret Analysis for Exponential Weights

Useful Lemma

For fixed π_{ref} and v , define

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

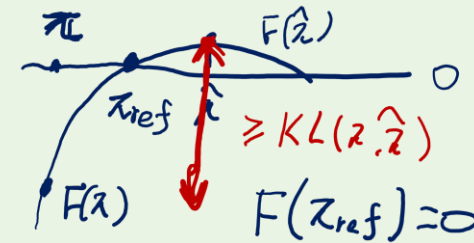
and let $\hat{\pi} = \max_{\pi} F(\pi)$

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

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

We will apply this lemma with

$$\pi_{\text{ref}} = \pi_t, \quad v = \eta r_t, \quad \hat{\pi} = \pi_{t+1}$$



(1) holds for all Bregman divergence

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

Regret Analysis for Exponential Weights

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

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

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

$$\begin{aligned} \textcircled{2} \quad \langle \pi^* - \pi_t, \eta r_t \rangle &\leq F(\pi_{t+1}) + KL(\pi^*, \pi_t) - KL(\pi^*, \pi_{t+1}) \\ &\leq \eta^2 \sum_a \pi_t(a) r_t(a)^2 \end{aligned}$$

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

Adversarial Multi-Armed Bandits

Adversarial MAB

Given: set of arms $\mathcal{A} = \{1, \dots, A\}$

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

Environment decides the reward vector $r_t = (r_t(1), \dots, r_t(A))$ (not revealing)

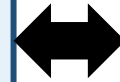
Learner chooses an arm $a_t \in \mathcal{A}$

Learner observes $r_t(a_t) + w_t$

$$\text{Regret} = \max_{a \in \mathcal{A}} \sum_{t=1}^T r_t(a) - \sum_{t=1}^T r_t(a_t)$$

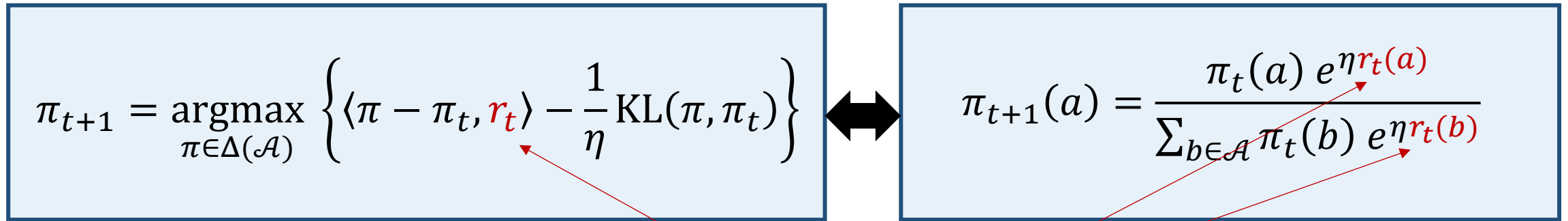
Recall: Exponential Weight Updates

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



$$\pi_{t+1}(a) = \frac{\pi_t(a) e^{\eta r_t(a)}}{\sum_{b \in \mathcal{A}} \pi_t(b) e^{\eta r_t(b)}}$$

Exponential Weight Updates for Bandits?

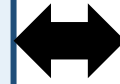
$$\pi_{t+1} = \operatorname{argmax}_{\pi \in \Delta(\mathcal{A})} \left\{ \langle \pi - \pi_t, \mathbf{r}_t \rangle - \frac{1}{\eta} \operatorname{KL}(\pi, \pi_t) \right\} \iff \pi_{t+1}(a) = \frac{\pi_t(a) e^{\eta \mathbf{r}_t(a)}}{\sum_{b \in \mathcal{A}} \pi_t(b) e^{\eta \mathbf{r}_t(b)}}$$


No longer observable

Only update the arm that we choose?

Exponential Weight Updates for Bandits?

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



$$\pi_{t+1}(a) = \frac{\pi_t(a) e^{\eta \hat{r}_t(a)}}{\sum_{b \in \mathcal{A}} \pi_t(b) e^{\eta \hat{r}_t(b)}}$$

- $\hat{r}_t(a)$ is an “**estimator**” for $r_t(a)$
- But we can only observe the reward of one arm
- Furthermore, $r_t(a)$ is different in every round (If we do not sample arm a in round t , we’ll never be able to estimate $r_t(a)$ in the future)

Unbiased Reward / Gradient Estimator

Fix arm a ,

$$\mathbb{E}[\hat{r}_t(a)] = \underbrace{p_r(a_t=a)}_{\pi_t(a)} \cdot \frac{r_t(a)}{\pi_t(a)} + p_r(a_t \neq a) \cdot 0 = \underline{r_t(a)} \quad \forall a$$

Weight a sample by **the inverse of the probability we observe it**

$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a)} \mathbb{I}\{a_t = a\} = \begin{cases} \frac{r_t(a)}{\pi_t(a)} & \text{if } a_t = a \\ 0 & \text{otherwise} \end{cases}$$

$= \begin{cases} 1 & \text{if } a_t = a \\ 0 & \text{if } a_t \neq a \end{cases}$

Inverse Propensity Weighting / Inverse Probability Weighting / Importance Weighting

Directly Applying Exponential Weights

$\pi_1(a) = 1/A$ for all a

$$\underline{r_t(a) \in [0, 1]}$$

For $t = 1, 2, \dots, T$:

Sample $a_t \sim \pi_t$, and observe $r_t(a_t)$

Define for all a :

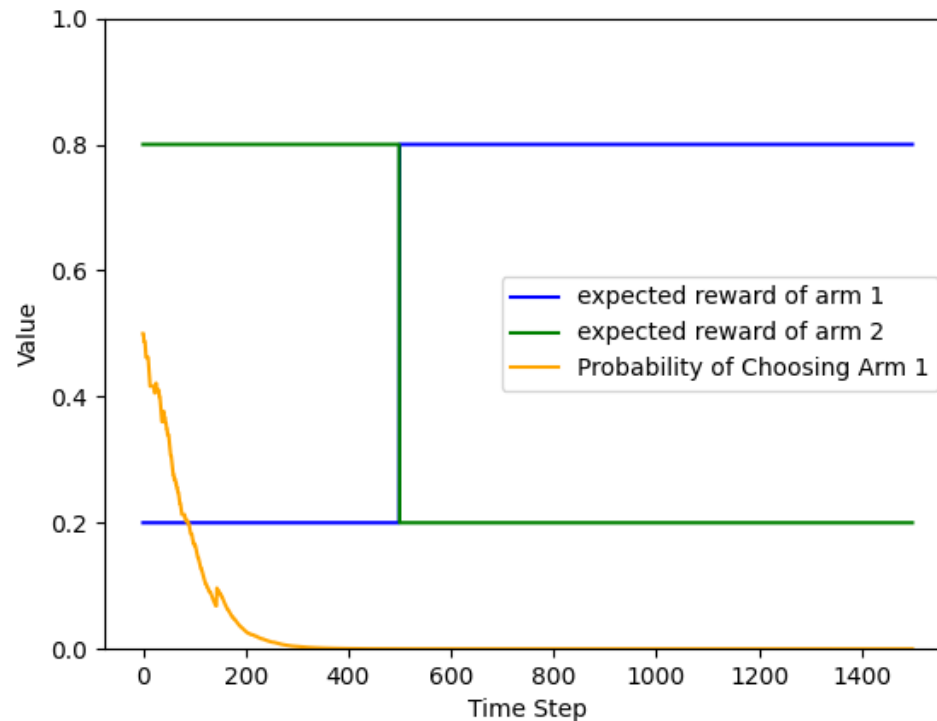
$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a)} \mathbb{I}\{a_t = a\}$$

Update policy:

$$\pi_{t+1}(a) = \frac{\pi_t(a) \exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} \pi_t(a') \exp(\eta \hat{r}_t(a'))}$$

Simple Experiment

- $A = 2$, $T = 1500$, $\eta = 1/\sqrt{T}$
- For $t \leq 500$, $r_t = [\text{Bernoulli}(0.2), \text{Bernoulli}(0.8)]$
- For $500 < t \leq 1500$, $r_t = [\text{Bernoulli}(0.8), \text{Bernoulli}(0.2)]$



Recall the Theorem

$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a)} \mathbb{1}_{\{a_t=a\}} \leq 1$$

↑
???

Theorem. Does this still hold?

Assume that $\eta \hat{r}_t(a) \leq 1$ for all t, a . Then EWU

$$\pi_{t+1}(a) = \frac{\pi_t(a) \exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} \pi_t(a') \exp(\eta \hat{r}_t(a'))}$$

ensures for any a^* ,

$$\mathbb{E} \left[\sum_{t=1}^T (\hat{r}_t(a^*) - \langle \pi_t, \hat{r}_t \rangle) \right] \leq \frac{\ln A}{\eta} + \eta \mathbb{E} \left[\sum_{t=1}^T \sum_{a=1}^A \pi_t(a) \hat{r}_t(a)^2 \right] \leq \frac{\ln A}{\eta} + \eta AT$$

How to relate the regret with this?

Is this still well-bounded?

$$\sqrt{AT \ln A}$$

$$\mathbb{E} \left[\sum_{t=1}^T \left(\hat{r}_t(a^*) - \langle \pi_t, \hat{r}_t \rangle \right) \right] = \mathbb{E} \left[\sum_{t=1}^T \left(r_t(a^*) - \langle \pi_t, r_t \rangle \right) \right]$$

\uparrow \hat{r}_t is unbiased estimator \uparrow real regret we care about

$$\begin{aligned} \sum_a \pi_t(a) \hat{r}_t(a)^2 &= \sum_a \pi_t(a) \left(\frac{r_t(a)}{\pi_t(a)} \mathbb{1}_{\{a_t=a\}} \right)^2 = \sum_a \pi_t(a) \cdot \frac{r_t(a)^2}{\pi_t(a)^2} \mathbb{1}_{\{a_t=a\}} \\ &= \sum_a \frac{r_t(a)^2}{\pi_t(a)} \mathbb{1}_{\{a_t=a\}} \end{aligned}$$

$$\mathbb{E} \left[\sum_a \pi_t(a) \hat{r}_t(a)^2 \right] = \mathbb{E} \left[\sum_a \frac{r_t(a)^2}{\pi_t(a)} \mathbb{1}_{\{a_t=a\}} \right] = \sum_a r_t(a)^2 \leq A$$

$$\sum_{t=1}^T \left(\hat{r}_t(a^*) - \underbrace{\langle \pi_t, \hat{r}_t \rangle}_{\downarrow} \right)$$

$$\sum_a \pi_t(a) \hat{r}_t(a) = \sum_a \pi_t(a) \cdot \frac{r_t(a)}{\pi_t(a)} \mathbb{1}\{a_t=a\} = r_t(a_t)$$

Solution 1: Adding Extra Exploration

- **Idea:** use at least η probability to choose each arm

$$\frac{r_t(a) \in [0, 1]}{r_t(a) \in [-1, 1]}$$

- Instead of sampling a_t according to π_t , use

$$\pi'_t(a) = (1 - A\eta)\pi_t(a) + \eta$$

w.p. $1 - A\eta \Rightarrow \underline{\text{use } \pi_t}$
w.p. $A\eta \Rightarrow \text{uniform exploration}$
 \uparrow
 ϵ

Then the unbiased reward estimator becomes

$$\hat{r}_t(a) = \frac{r_t(a)}{\pi'_t(a)} \mathbb{I}\{a_t = a\} = \frac{r_t(a)}{(1 - A\eta)\pi_t(a) + \eta} \mathbb{I}\{a_t = a\}$$

$$\Rightarrow \hat{r}_t(a) = \frac{r_t(a)}{(1 - A\eta)\pi_t(a) + \eta} \mathbb{I}(\dots) \leq r_t(a) \mathbb{I} \leq 1$$

Applying Solution 1

$\pi_1(a) = 1/A$ for all a

For $t = 1, 2, \dots, T$:

Sample a_t from $\pi'_t = (1 - A\eta)\pi_t + A\eta \text{ uniform}(\mathcal{A})$, and observe $r_t(a_t)$

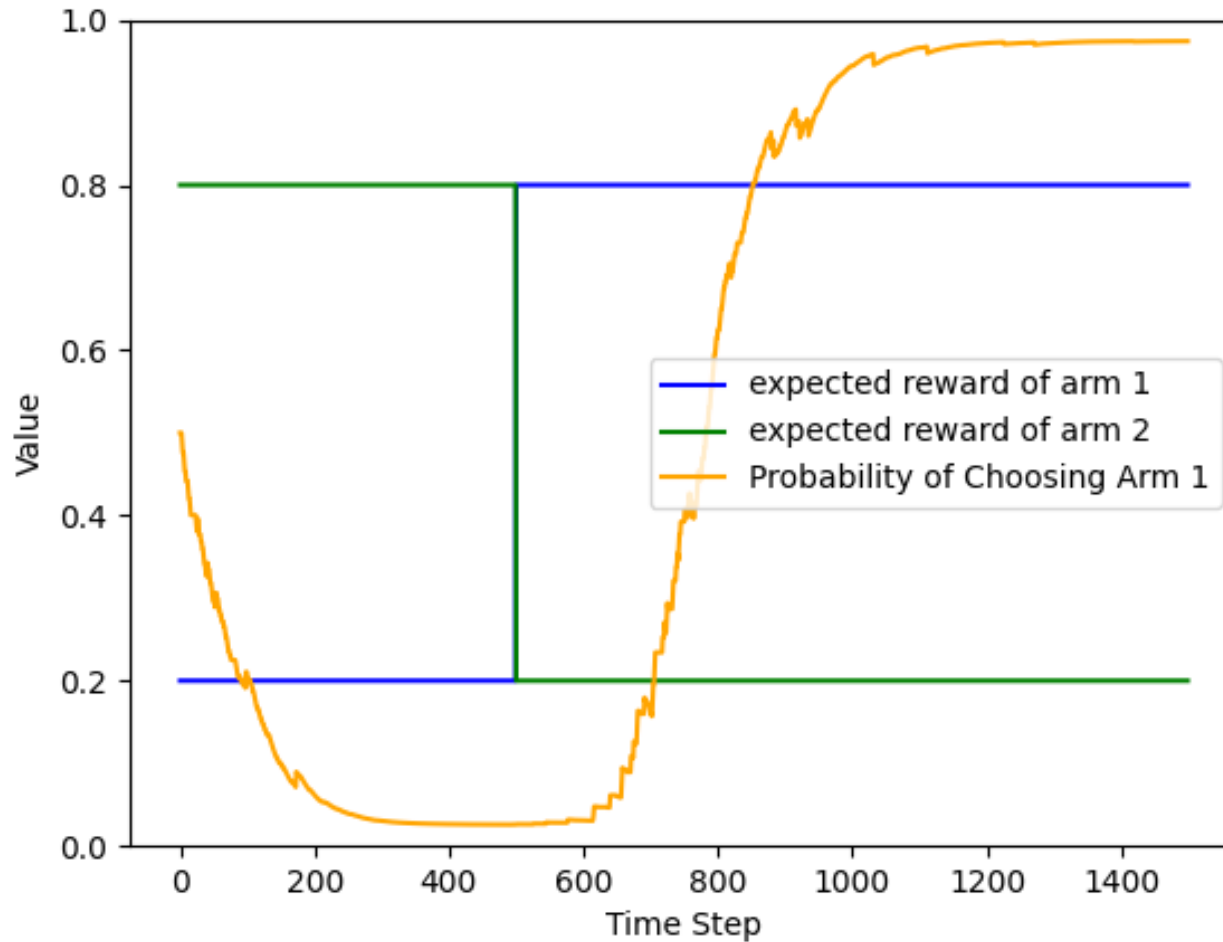
Define for all a :

$$\hat{r}_t(a) = \frac{r_t(a)}{\pi'_t(a)} \mathbb{I}\{a_t = a\}$$

Update policy:

$$\pi_{t+1}(a) = \frac{\pi_t(a) \exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} \pi_t(a') \exp(\eta \hat{r}_t(a'))}$$

Solution 1: Adding Extra Exploration



Regret Bound for Solution 1

Theorem. Exponential weights with Solution 1 ensures

$$\eta \approx \sqrt{\frac{\ln A}{T}}$$

$$\max_{a^*} \mathbb{E} \left[\sum_{t=1}^T (r_t(a^*) - r_t(a_t)) \right] \leq O \left(\frac{\ln A}{\eta} + \eta AT \right) \quad \sqrt{AT \ln A}$$

Solution 2: Reward Estimator with a Baseline

$$r_t(a) \in [-1, 1]$$

- Notice that the condition is only $\eta \hat{r}_t(a) \leq 1$. The reward estimator is allowed to be **very negative**! (Check our proof)

- Still sample a_t from π_t , but construct the reward estimator as

$$\hat{r}_t(a) = \frac{\overset{<0}{r_t(a) - 1}}{\pi_t(a)} \mathbb{I}\{a_t = a\} + 1$$

$$\begin{aligned} \text{Fix } a, \mathbb{E}[\hat{r}_t(a)] &= \cancel{P(a_t=a)} \cdot \left(\frac{r_t(a)-1}{\pi_t(a)} + 1 \right) \\ &\quad + P(a_t \neq a) \cdot 1 \end{aligned}$$

- Why this resolves the issue? ...

$$\begin{aligned} &= \pi_t(a) \left(\left(\frac{r_t(a)-1}{\pi_t(a)} + 1 \right) + \underbrace{(1 - \pi_t(a))}_{\text{wavy line}} \right) \\ &= r_t(a) - 1 + 1 = r_t(a) \end{aligned}$$

Applying Solution 2

$$\pi_1(a) = 1/A \text{ for all } a$$

For $t = 1, 2, \dots, T$:

Sample a_t from π_t , and observe $r_t(a_t)$

Define for all a :

$$\hat{r}_t(a) = \frac{r_t(a) - 1}{\pi_t(a)} \mathbb{I}\{a_t = a\} + 1 \text{ or equivalently } \hat{r}_t(a) = \frac{r_t(a) - \text{baseline}}{\pi_t(a)} \mathbb{I}\{a_t = a\}$$

Update policy:

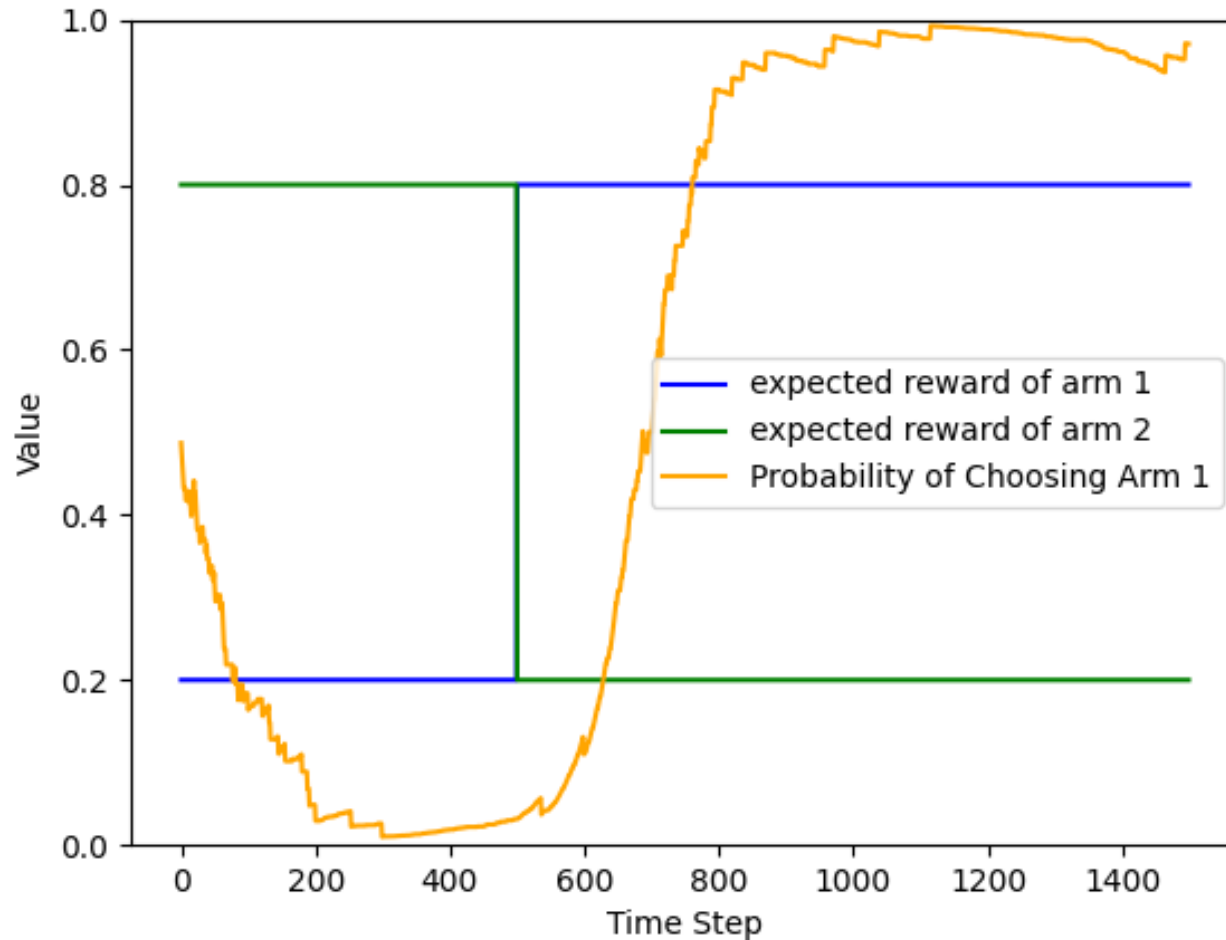
$$\pi_{t+1}(a) = \frac{\pi_t(a) \exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} \pi_t(a') \exp(\eta \hat{r}_t(a'))}$$

$$\arg \max \left\{ \langle \pi - \pi_t, \hat{r}_t \rangle - \frac{1}{2} \text{KL}(\pi, \pi_t) \right\}$$

$$\hat{r}_t + c \begin{bmatrix} 1 \\ \vdots \end{bmatrix}$$

$$\langle \pi - \pi_t, \begin{bmatrix} 1 \\ \vdots \end{bmatrix} \rangle = 0$$

Solution 2: Reward Estimator with a Baseline



Regret Bound for Solution 2

Theorem. Exponential weights with Solution 2 ensures

$$\max_{a^*} \mathbb{E} \left[\sum_{t=1}^T (r_t(a^*) - r_t(a_t)) \right] \leq O \left(\frac{\ln A}{\eta} + \eta AT \right)$$

EXP3 Algorithm

“**Ex**ponential weight algorithm for **Ex**ploration and **Ex**ploitation”

- Exponential weights + either of the two solutions

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, Robert Schapire.
The Nonstochastic Multiarmed Bandit Problem. 2002.

Biasing

To keep $\eta \hat{r}_t(a) \leq 1$, we may also use “biased” reward estimator

$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a) + \eta} \mathbb{I}\{a_t = a\} \quad \text{or} \quad \hat{r}_t(a) = \frac{r_t(a) - 1}{\pi_t(a) + \eta} \mathbb{I}\{a_t = a\}$$



Different from Solution 1 (adding an extra uniform exploration), here we do not add exploration. Therefore, the reward estimator is **biased**.

Biasing

To keep $\eta \hat{r}_t(a) \leq 1$, we may also use “biased” reward estimator

$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a) + \eta} \mathbb{I}\{a_t = a\} \quad \text{or} \quad \hat{r}_t(a) = \frac{r_t(a) - 1}{\pi_t(a) + \eta} \mathbb{I}\{a_t = a\}$$

$$\mathbb{E}[\hat{r}_t(a)] - r_t(a) = r_t(a) \left(\frac{-\eta}{\pi_t(a) + \eta} \right)$$

Negative bias for seldom-picked arms



$$\mathbb{E}[\hat{r}_t(a)] - r_t(a) = (r_t(a) - 1) \left(\frac{-\eta}{\pi_t(a) + \eta} \right)$$

Positive bias for seldom-picked arms



EXP3-IX

$\pi_1(a) = 1/A$ for all a

For $t = 1, 2, \dots, T$:

Sample a_t from π_t and observe $r_t(a_t)$

Define for all a :

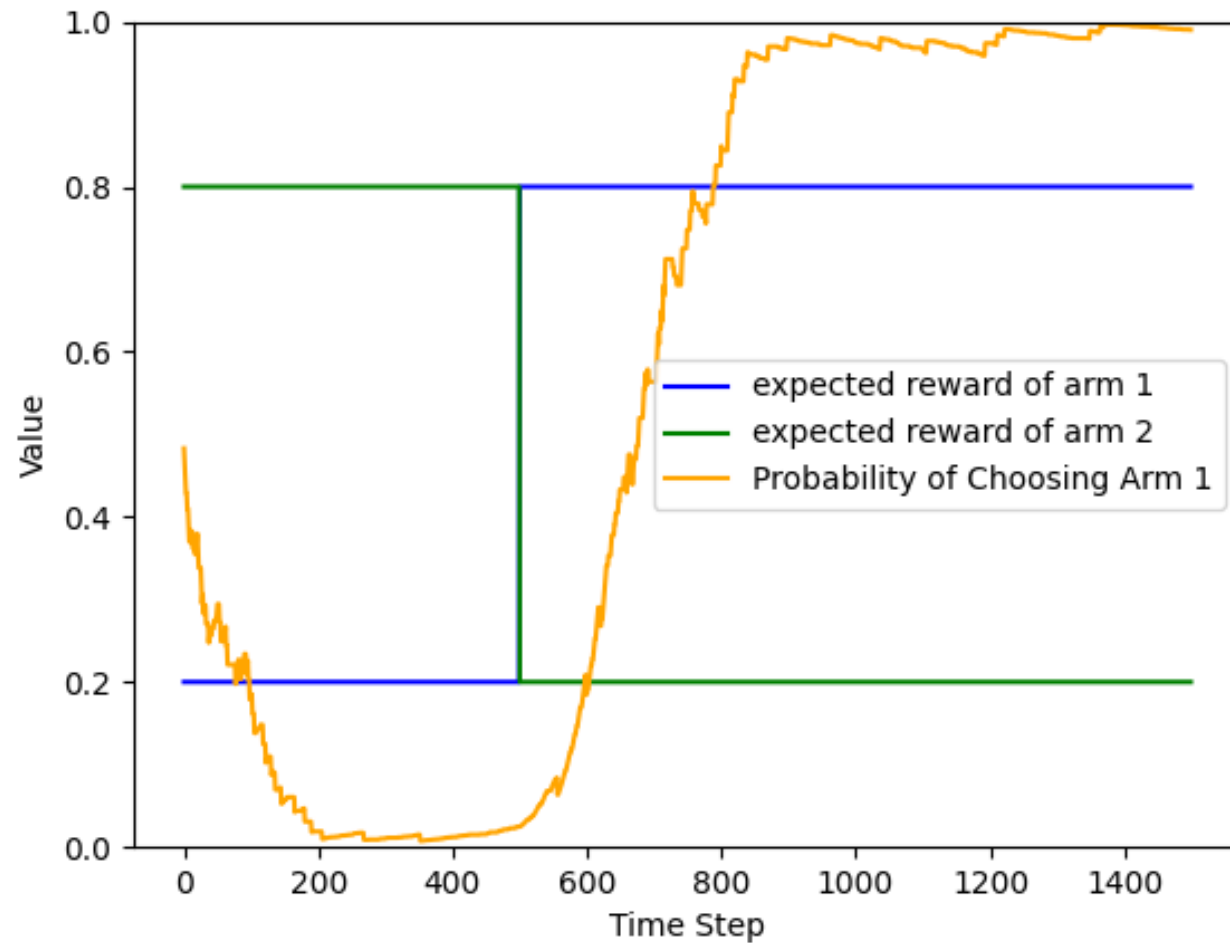
$$\hat{r}_t(a) = \frac{r_t(a) - 1}{\pi_t(a) + \eta} \mathbb{I}\{a_t = a\}$$

Update policy:

$$\pi_{t+1}(a) = \frac{\pi_t(a) \exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} \pi_t(a') \exp(\eta \hat{r}_t(a'))}$$

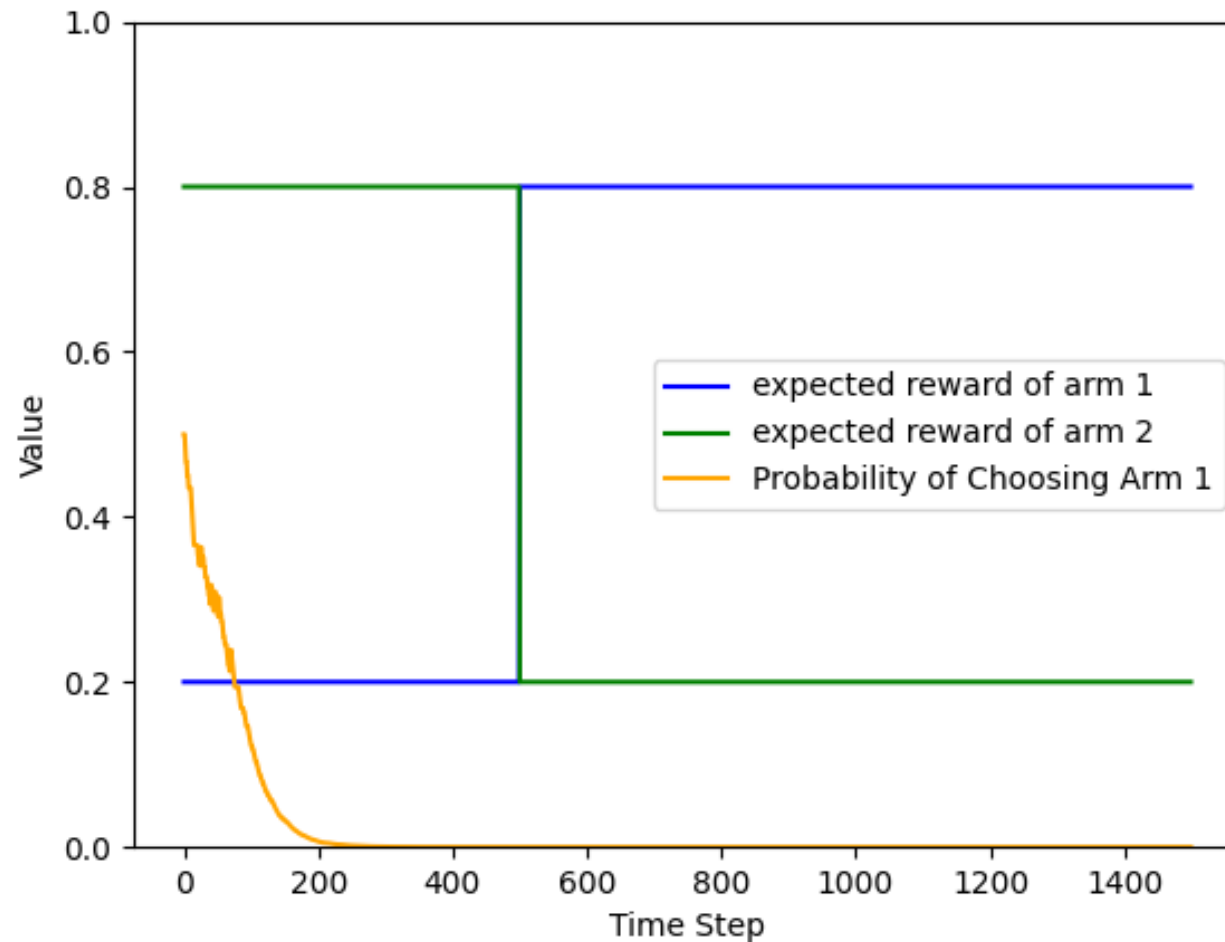
EXP3-IX

$$\hat{r}_t(a) = \frac{r_t(a) - 1}{\pi_t(a) + \eta} \mathbb{I}\{a_t = a\}$$



If Biasing in a Wrong Way

$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a) + \eta} \mathbb{I}\{a_t = a\}$$



Regret Bound for Solution 3

Theorem. EXP3-IX ensures **with high probability**,

$$\max_{a^*} \sum_{t=1}^T (r_t(a^*) - r_t(a_t)) \leq \tilde{O} \left(\frac{\ln A}{\eta} + \eta AT \right)$$

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. 2015.

Summary

- Exponential weight update (EWU) is an effective algorithm for full-information setting. It guarantees sublinear regret even when the environment changes over time.
- Extending EWU to bandit with naïve unbiased reward estimator does not work (lack of exploration). Two ways to fix it:
 - Adding **extra uniform exploration** with probability $\geq A\eta$
 - Adding a **baseline** in the reward estimator to encourage exploration
- High-probability bounds can be achieved by adding **baseline** and **bias** (EXP3-IX).