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

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

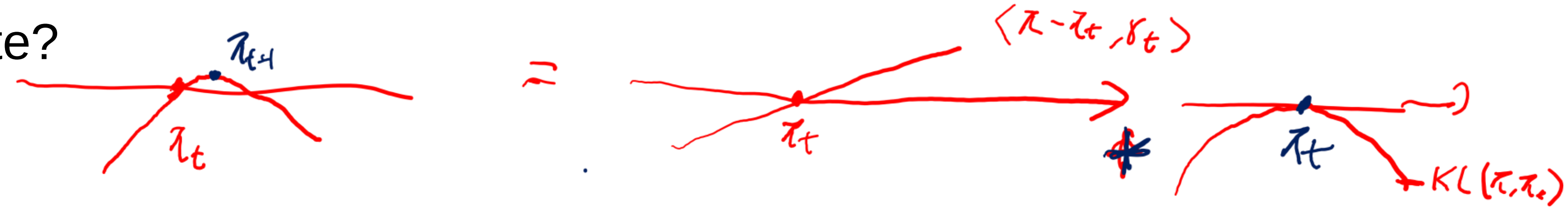
$\pi_t = \begin{bmatrix} \pi_t(1) \\ \vdots \\ \pi_t(A) \end{bmatrix}$

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

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

$\boxed{\langle \pi, r_t \rangle}$

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

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

The equivalence is shown in HW0

Regret Bound for Exponential Weight Updates

Theorem.

Assume that $\eta r_t(a) \le 1$ 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 $a^\star \in \mathcal{A}$,

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

≤ 1

$\sqrt{AT}$ ~ bandit

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

$\sqrt{\frac{\log A}{T}}$

Questions and Discussions

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

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

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

$$\lambda_t = \eta t$$

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

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.

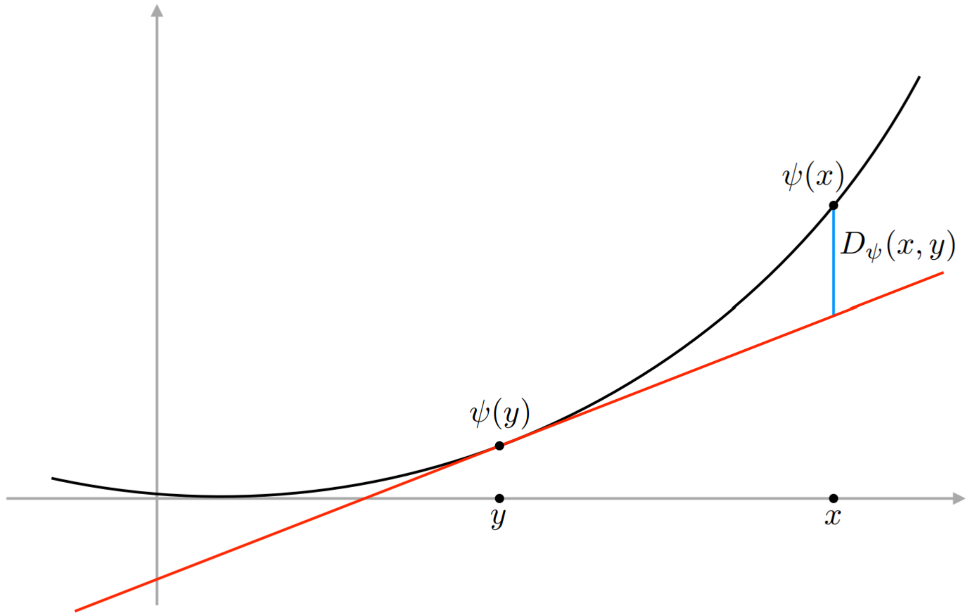
Exponential Weight Update ∈ Mirror Ascent

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**
$\frac{1}{2}\Vert x\Vert_2^2$	$\frac{1}{2}\Vert x-y\Vert_2^2$	$x_{t+1} = \mathcal{P}_\Omega(x_t + \eta r_t)$ **Gradient ascent**
$\sum_a x(a) \log x(a)$ **Negative entropy**	$\sum_a x(a) \log \frac{x(a)}{y(a)}$	$x_{t+1}(a) = \frac{x_t(a) e^{\eta r_t(a)}}{\sum_b x_t(b)\, e^{\eta r_t(b)}}$
$\sum_a \log \frac{1}{x(a)}$	$\sum_a \left(\frac{x(a)}{y(a)} - \log \frac{x(a)}{y(a)} - 1 \right)$	$\frac{1}{x_{t+1}(a)} = \frac{1}{x_t(a)} - \eta r_t(a) + \gamma_t$

(for distributions)

(for distributions)

Normalization factor

Regret Analysis for Exponential Weights

Theorem.

Assume that $\eta r_t(a) \le 1$ 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 $a^\star \in \mathcal{A}$,

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

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

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

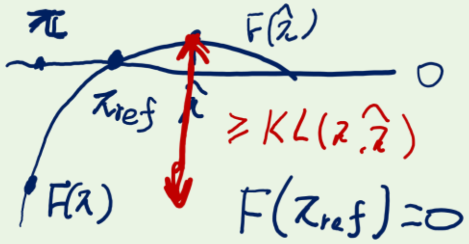
Regret Analysis for Exponential Weights

Useful Lemma

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

For fixed π_{ref} and v, define

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



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

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

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 $r_t(a_t) = R_t(a_t) + w_t(a_t)$

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

Recall: Exponential Weight Updates

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

Exponential Weight Updates for Bandits?

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

No longer observable

Only update the arm that we choose?

Exponential Weight Updates for Bandits?

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

- $\hat{r}_t(a)$ 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,

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

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

$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a)} \underbrace{\mathbb{I}\{a_t = a\}}_{= \begin{cases} 1 & \text{if } a_t = a \\ 0 & \text{if } a_t \neq a \end{cases}} = \begin{cases} \dfrac{r_t(a)}{\pi_t(a)} & \text{if } a_t = a \\ 0 & \text{otherwise} \end{cases}$$

Inverse Propensity Weighting / Inverse Probability Weighting / Importance Weighting

Directly Applying Exponential Weights

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

$r_t(a) \in [0,1)$

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

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

Define for all a:

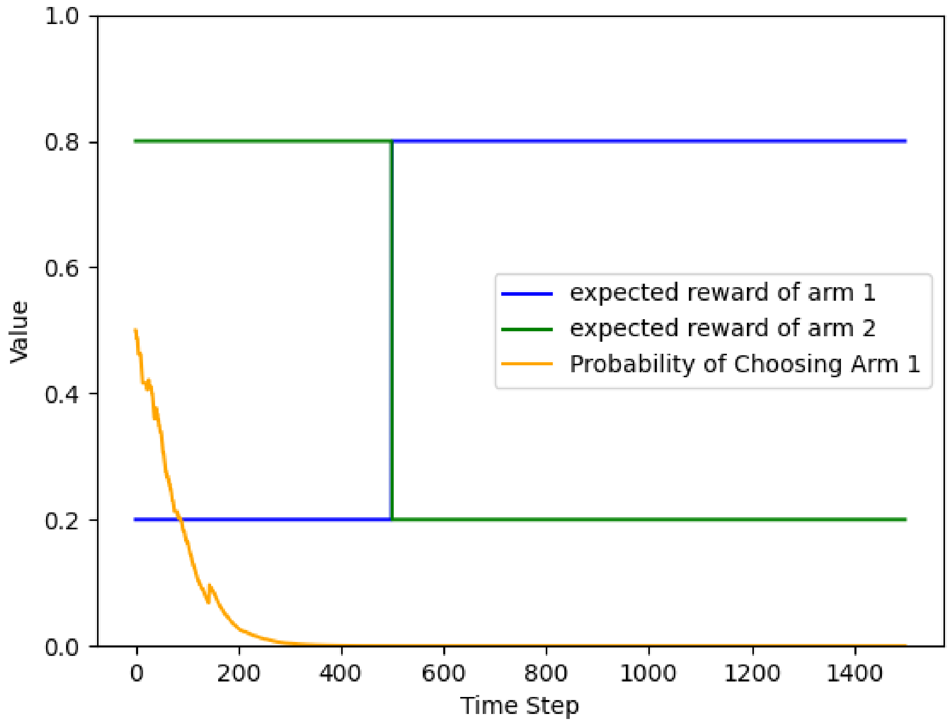
$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a)} \mathbb{I}\{a_t = a\}$$

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 $t \leq 500,\ r_t = [\mathrm{Bernoulli}(0.2), \mathrm{Bernoulli}(0.8)]$
- For $500 < t \leq 1500,\ r_t = [\mathrm{Bernoulli}(0.8), \mathrm{Bernoulli}(0.2)]$



Recall the Theorem

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

???

Theorem. Does this still hold?

Assume that $\boxed{\eta \hat{r}_t(a) \leq 1}$ 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 $a^\star$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \underbrace{\left(\hat{r}_t(a^\star) - \langle \pi_t, \hat{r}_t \rangle\right)}_{\text{How to relate the regret with this?}}\right] \leq \frac{\ln A}{\eta} + \eta\, \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \sum_{a=1}^{A} \pi_t(a) \hat{r}_t(a)^2\right]}_{\text{Is this still well-bounded?}} \leq \frac{\ln A}{\eta} + \eta A T$$

$\sqrt{AT \ln A}$

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

$\uparrow$ $\hat{r}_t$ is unbiased estimator

$\uparrow$ real regret we care about

$$\sum_a \pi_t(a)\, \hat{r}_t(a)^2 = \sum_a \pi_t(a) \left(\frac{r_t(a)}{\pi_t(a)} \mathbb{1}\{a_t = a\} \right)^2 = \sum_a \pi_t(a) \cdot \frac{r_t(a)^2}{\pi_t(a)^2} \mathbb{1}\{a_t = a\}$$

$$= \sum_a \frac{r_t(a)^2}{\pi_t(a)} \mathbb{1}\{a_t = a\}$$

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

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

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

Solution 1: Adding Extra Exploration

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

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

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

- Instead of sampling a_t according to π_t, use

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

w.p. $1-A\eta \Rightarrow$ use π_t

w.p. $A\eta \Rightarrow$ uniform exploration

ε

Then the unbiased reward estimator becomes

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

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

Applying Solution 1

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

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

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

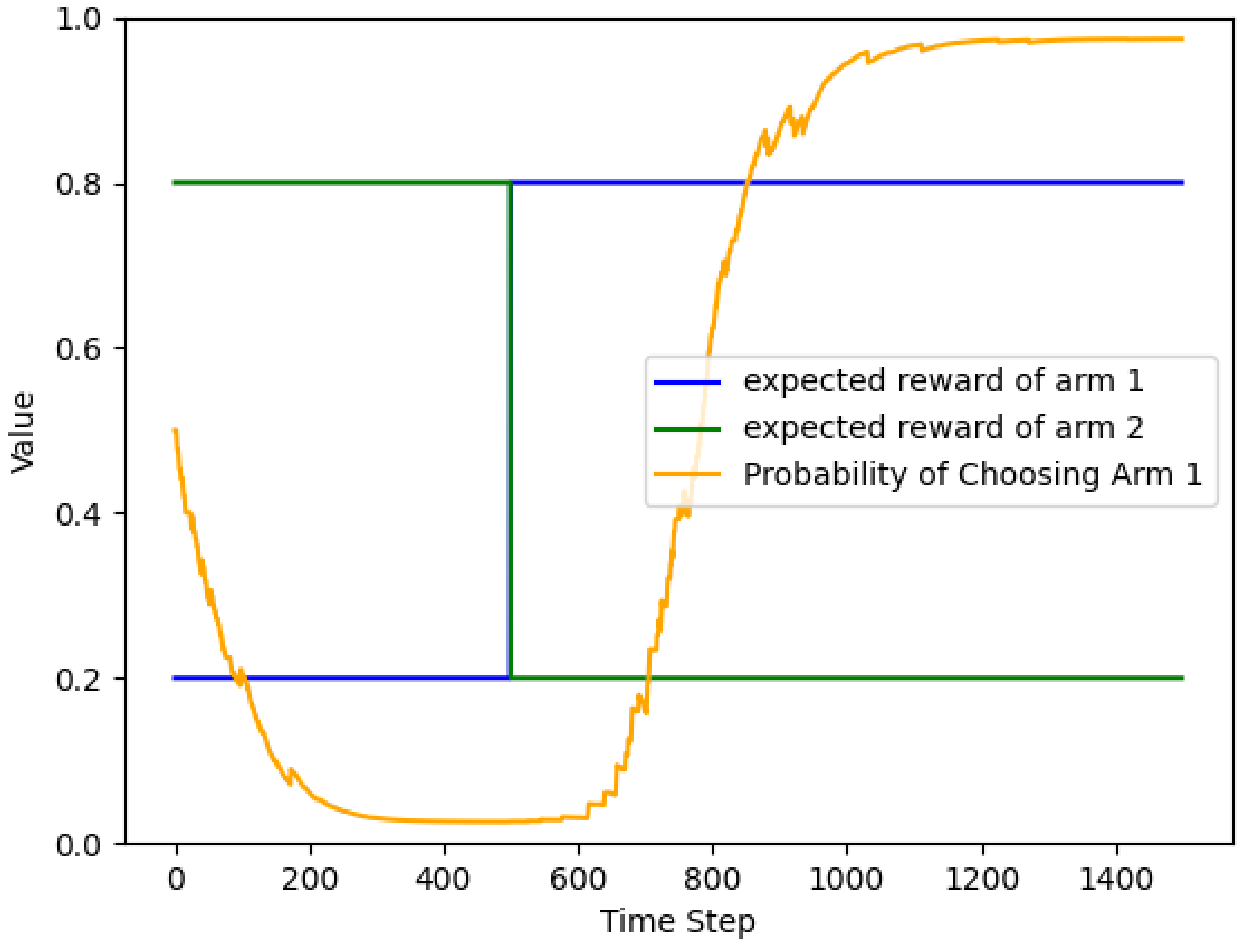
Define for all a:

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

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

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

Solution 2: Reward Estimator with a Baseline

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

- 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

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

Fix a, $\mathbb{E}[\hat{r}_t(a)]$

$= \cancel{Pr(a_t=a)} \cdot \left(\frac{r_t(a)-1}{\pi_t(a)} + 1 \right)$

$+ Pr(a_t \neq a) \cdot 1$

$= \pi_t(a) \left(\frac{r_t(a)-1}{\pi_t(a)} + 1 \right) + (1-\pi_t(a))$

$= r_t(a) - 1 + 1 = r_t(a)$

- Why this resolves the issue? ...

Applying Solution 2

$$\operatorname{argmax}\left\{\langle \pi - \pi_t, \hat{r}_t\rangle - \frac{1}{\eta} KL(\pi, \pi_t)\right\}$$

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

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

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

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

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

Define for all a:

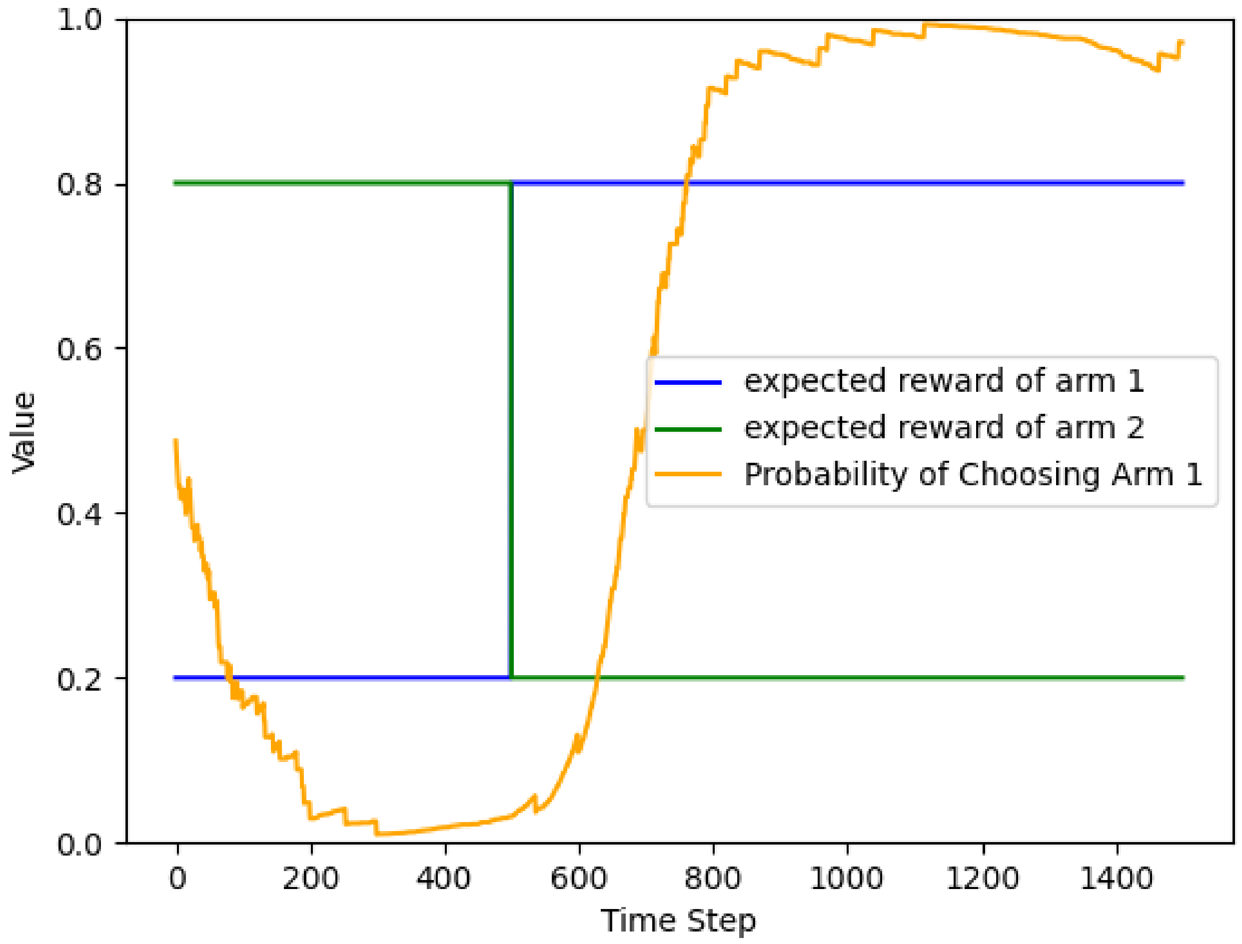
baseline

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

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 2: Reward Estimator with a Baseline



Regret Bound for Solution 2

Theorem. Exponential weights with Solution 2 ensures

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

EXP3 Algorithm

“**Exp**onential weight algorithm for **Exp**loration and **Exp**loitation”

- 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

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

✗ ✓

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

Biasing

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

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

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

Small bias for often-picked arms

More negative bias for seldom-picked arms

✗

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

Small bias for often-picked arms

More positive bias for seldom-picked arms

✓

EXP3-IX

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

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

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

Define for all a:

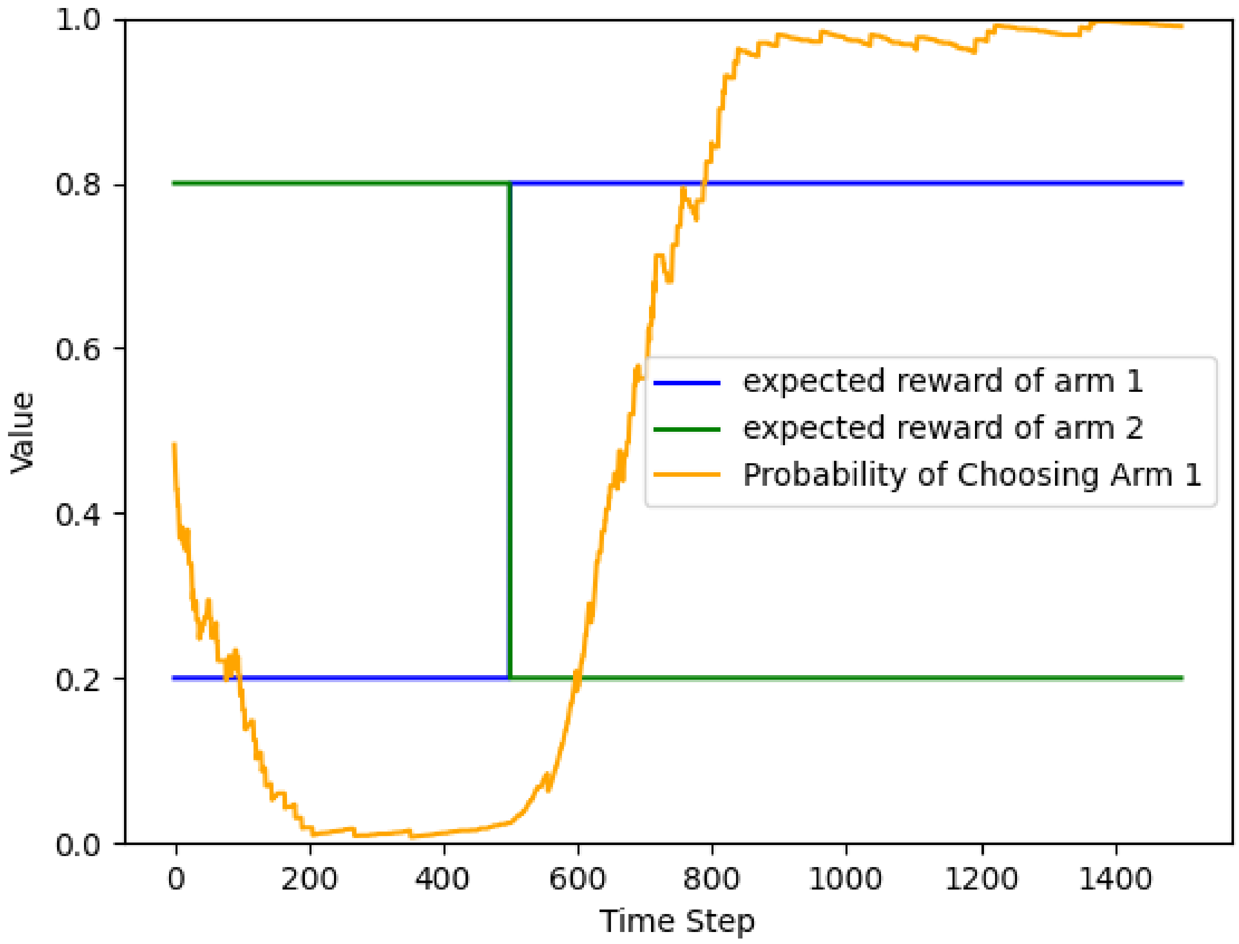
$$\hat{r}_t(a) = \frac{r_t(a) - 1}{\pi_t(a) + \eta} \mathbb{I}\{a_t = a\}$$

Update policy:

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

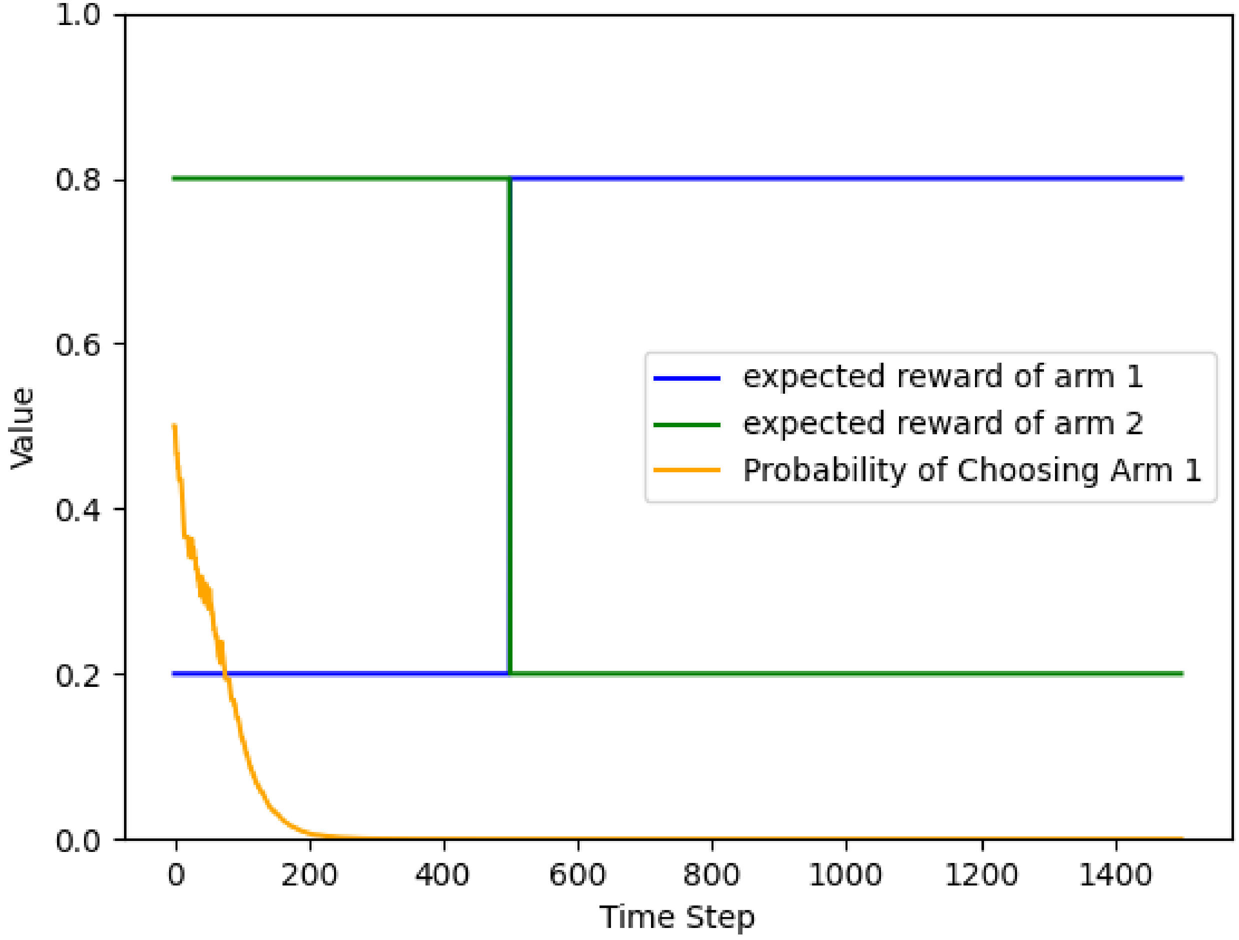
EXP3-IX

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



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 EXP3-IX

Theorem. EXP3-IX ensures with high probability,

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

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

The Role of Baseline

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

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

Larger b_t: More exploratory (tends to decrease the probability of the action just chosen)

– needed to detect changes in the environment.

Some moderate b_t: smaller variance and slight improvement in the regret bound

$$\sum_{a=1}^{A} \pi_t(a) \hat{r}_t(a)^2 = \sum_{a=1}^{A} \pi_t(a) \left(\frac{r_t(a) - b_t}{\pi_t(a)} \mathbb{I}\{a_t = a\} \right)^2$$

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

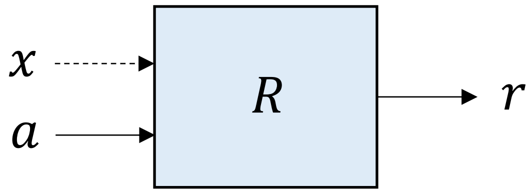
- High-probability bounds can be achieved by adding **baseline** and **bias** (EXP3-IX).

Review: Bandit Techniques

x: context, a: action, r: reward

MAB | **CB**

Value-based

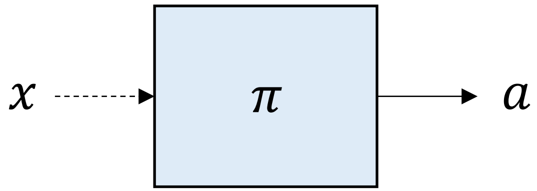


(context, action) to reward

MAB: Mean estimation + EG, BE, IGW

CB: Regression + EG, BE, IGW

Policy-based



context to action distribution

MAB: KL-regularized update with reward estimators (EXP3) + baseline, bias, or uniform exploration

CB: **Next**

Contextual Bandits

Contextual Bandits

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

Environment generates a context $x_t \in \mathcal{X}$

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

Learner observes $r_t(x_t, a_t) = R(x_t, a_t) + w_t$

KL-Regularized Policy Updates

$$\pi_{t+1} = \underset{\pi \in \Delta(\mathcal{A})}{\operatorname{argmax}} \left\{ \sum_a \pi(a) \hat{r}_t(a) - \frac{1}{\eta} \sum_a \pi(a) \log \frac{\pi(a)}{\pi_t(a)} \right\}$$

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

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}} \left\{ \sum_a \pi_\theta(a|x_t) \, \hat{r}_t(x_t, a) - \frac{1}{\eta} \sum_a \pi_\theta(a|x_t) \log \frac{\pi_\theta(a|x_t)}{\pi_{\theta_t}(a|x_t)} \right\}$$

$$\hat{r}_t(x_t, a) = \frac{r_t(x_t, a) - b_t(x_t)}{\pi_{\theta_t}(a|x_t)} \, \mathbb{I}\{a_t = a\}$$

KL-Regularized Policy Updates

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

Receive context x_t

Take action $a_t \sim \pi_{\theta_t}(\cdot|x_t)$ and receive reward $r_t(x_t, a_t)$

Create reward estimator $\hat{r}_t(x_t, a) = \frac{r_t(x_t,a) - b_t(x_t)}{\pi_{\theta_t}(a|x_t)} \mathbb{I}\{a_t = a\}$

Update

$$\theta_{t+1} = \operatorname*{argmax}_{\theta} \left\{ \sum_a \pi_\theta(a|x_t)\, \hat{r}_t(x_t, a) - \frac{1}{\eta} \sum_a \pi_\theta(a|x_t) \log \frac{\pi_\theta(a|x_t)}{\pi_{\theta_t}(a|x_t)} \right\}$$

Proximal Policy Optimization (PPO) for CB

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

For $i = 1, \dots, N$:

Receive context x_i

Take action $a_i \sim \pi_{\theta_t}(\cdot|x_i)$ and receive reward $r_i(x_i, a_i)$

Create reward estimator $\hat{r}_i(x_i, a) = \frac{r_i(x_i,a) - b_t(x_i)}{\pi_{\theta_t}(a|x_i)} \mathbb{I}\{a_i = a\}$

For $j = 1, \dots, M$:

one iteration of mirror ascent

For minibatch $\mathcal{B} \subset \{1, 2, \dots, N\}$ of size B:

$$\theta \leftarrow \theta - \nabla_\theta \frac{1}{B} \sum_{i \in \mathcal{B}} \left(\sum_a \pi_\theta(a|x_i)\, \hat{r}_i(x_i, a) - \frac{1}{\eta} \sum_a \pi_\theta(a|x_i) \log \frac{\pi_\theta(a|x_i)}{\pi_{\theta_t}(a|x_i)} \right)$$

$$= \theta - \nabla_\theta \frac{1}{B} \sum_{i \in \mathcal{B}} \left(\frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)} (r_i(x_i, a_i) - b_t(x_i)) - \frac{1}{\eta} \sum_a \pi_\theta(a|x_i) \log \frac{\pi_\theta(a|x_i)}{\pi_{\theta_t}(a|x_i)} \right)$$

$\theta_{t+1} \leftarrow \theta$

Proximal Policy Optimization (PPO) for CB

$$\theta \leftarrow \theta - \nabla_\theta \frac{1}{B} \sum_{i \in \mathcal{B}} \left(\frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)} (r_i(x_i, a_i) - b_t(x_i)) - \underbrace{\frac{1}{\eta} \sum_a \pi_\theta(a|x_i) \log \frac{\pi_\theta(a|x_i)}{\pi_{\theta_t}(a|x_i)}}_{\mathrm{KL}\left(\pi_\theta(\cdot|x_i), \pi_{\theta_t}(\cdot|x_i)\right)} \right)$$

- May replace $\mathrm{KL}\left(\pi_\theta(\cdot|x_i), \pi_{\theta_t}(\cdot|x_i)\right)$ by $\mathrm{KL}\left(\pi_{\theta_t}(\cdot|x_i), \pi_\theta(\cdot|x_i)\right)$
- Although this term can be calculated exactly, we often use samples to estimate it (so we do not need to sum over a)

Estimating KL by Samples

http://joschu.net/blog/kl-approx.html

Sample $a_i \sim \pi_{\theta_t}(\cdot\,|x_i)$ and define $kl_i(\theta_t, \theta) = \boxed{\frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)} - 1 - \log\frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)}}$

Then $\mathbb{E}_{a_i \sim \pi_{\theta_t}(\cdot|x_i)}[kl_i(\theta_t, \theta)] = \mathrm{KL}\left(\pi_{\theta_t}(\cdot\,|x_i), \pi_\theta(\cdot\,|x_i)\right)$

Just need one sample of a_i

As we see before, there are many ways to construct an unbiased estimator.

This is a good one with low variance.

PPO with KL Estimator

$$kl_i(\theta_t, \theta) = \frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)} - 1 - \log \frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)}$$

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

For $i = 1, \ldots, N$:

Receive context x_i

Take action $a_i \sim \pi_{\theta_t}(\cdot|x_i)$ and receive reward $r_i(x_i, a_i)$

Create reward estimator $\hat{r}_i(x_i, a) = \frac{r_i(x_i,a) - b_t(x_i)}{\pi_{\theta_t}(a|x_i)} \mathbb{I}\{a_i = a\}$

For $j = 1, \ldots, M$:

For minibatch $\mathcal{B} \subset \{1, 2, \ldots, N\}$ of size B:

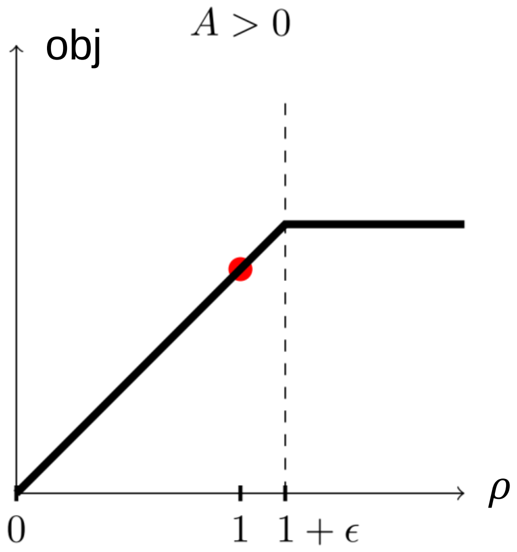
$$\theta \leftarrow \theta - \nabla_\theta \frac{1}{B} \sum_{i \in \mathcal{B}} \left(\frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)} \left(r_i(x_i, a_i) - b_t(x_i)\right) - \frac{1}{\eta} kl_i(\theta_t, \theta) \right)$$

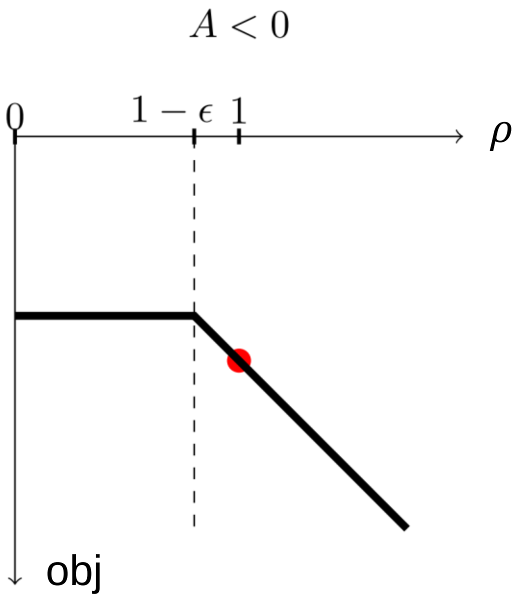
$\theta_{t+1} \leftarrow \theta$

Additional Technique for PPO: Clipped Estimator

$$\rho = \frac{\pi_\theta(a|x)}{\pi_{\theta_k}(a|x)} \qquad\qquad A = r(x,a) - b(x)$$

Instead of using ρA as the estimator, use $\min\left\{\rho A,\ \mathrm{clip}_{[1-\epsilon, 1+\epsilon]}(\rho) A\right\}$





algorithm	avg. normalized score
No clipping or penalty	-0.39
Clipping, $\epsilon = 0.1$	0.76
Clipping, $\epsilon = 0.2$	**0.82**
Clipping, $\epsilon = 0.3$	0.70
Adaptive KL $d_{\mathrm{targ}} = 0.003$	0.68
Adaptive KL $d_{\mathrm{targ}} = 0.01$	0.74
Adaptive KL $d_{\mathrm{targ}} = 0.03$	0.71
Fixed KL, $\beta = 0.3$	0.62
Fixed KL, $\beta = 1.$	0.71
Fixed KL, $\beta = 3.$	0.72
Fixed KL, $\beta = 10.$	0.69

Schulman et al., Proximal Policy Optimization Algorithms. 2017.

Summary: PPO

- PPO-CB can be viewed as an extension of EXP3 to contextual bandits. The central idea is KL-regularized policy updates
- Common techniques: baselines, avoiding **overly positive** reward estimator. These techniques prevent over exploitation
- PPO additional uses batching and KL estimators for computational efficiency

Natural Policy Gradient

(PPO)

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\ \mathbb{E}_x \left[\sum_a \pi_\theta(a|x)\, \hat{r}_t(x,a) - \frac{1}{\eta} \sum_a \pi_\theta(a|x) \log \frac{\pi_\theta(a|x)}{\pi_{\theta_t}(a|x)} \right]$$

$$\eta \to 0$$

(NPG)

$$\theta_{t+1} = \theta_t + \eta F_t^{-1} \left(\mathbb{E}_x \left[\sum_a \nabla_\theta \pi_\theta(a|x)\, \hat{r}_t(x,a) \right] \right) \Bigg|_{\theta=\theta_t}$$

where $F_t = \mathbb{E}_x \left[\left(\nabla_\theta \log \pi_\theta(a|x) \right) \left(\nabla_\theta \log \pi_\theta(a|x) \right)^\top \right] \Big|_{\theta=\theta_t}$