

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

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

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

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

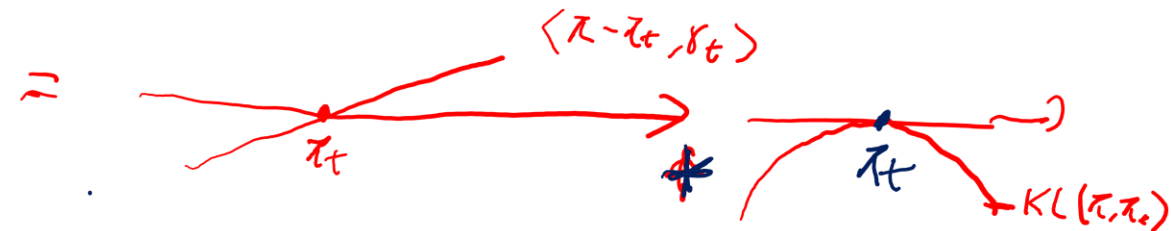
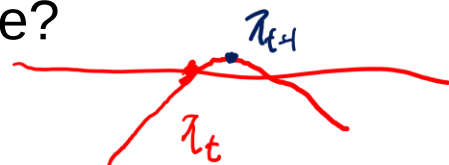
$$= \operatorname{argmax}_{\pi \in \Delta(\mathcal{A})} \left\{ \underbrace{\sum_a (\pi(a) - \pi_t(a)) r_t(a)}_{\text{The Improvement of } \pi \text{ over } \pi_t} - \underbrace{\frac{1}{\eta} \sum_a \pi(a) \log \frac{\pi(a)}{\pi_t(a)}}_{\text{Distance between } \pi \text{ and } \pi_t} \right\}$$

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

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

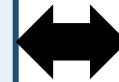
Time	1	2	3	4	5	6	...
$R_t(1)$	0.5	0	1	0	1	0	...
$R_t(2)$	0	1	0	1	0	1	...

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

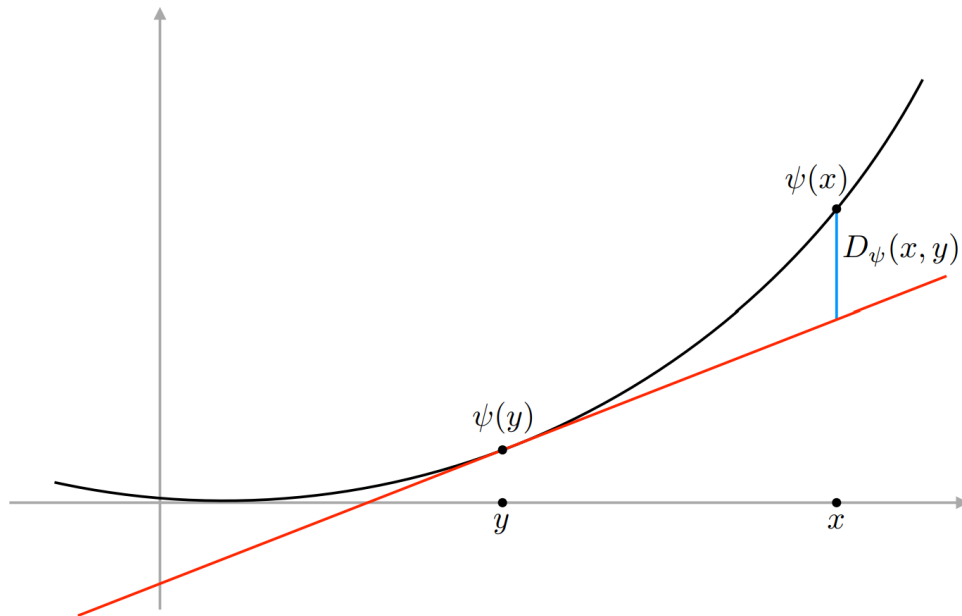
General form of **Mirror Ascent**:

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

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 $a^* \in \mathcal{A}$,

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

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

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

Regret Analysis for Exponential Weights

Useful Lemma

For fixed π_{ref} and v , define

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

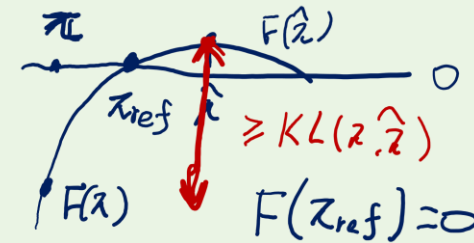
and let $\hat{\pi} = \max_{\pi} F(\pi)$

(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)

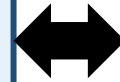
Learner chooses an arm $a_t \in \mathcal{A}$

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

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

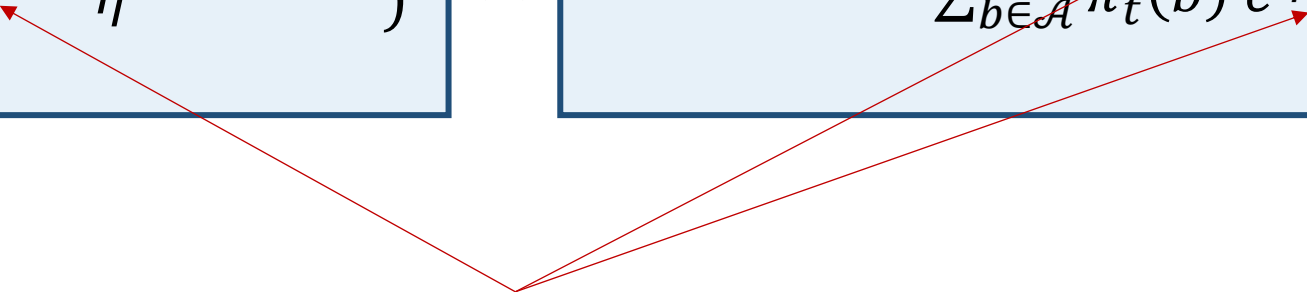
Recall: Exponential Weight Updates

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



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

Exponential Weight Updates for Bandits?

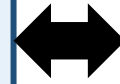
$$\pi_{t+1} = \operatorname{argmax}_{\pi \in \Delta(\mathcal{A})} \left\{ \langle \pi - \pi_t, \mathbf{r}_t \rangle - \frac{1}{\eta} \operatorname{KL}(\pi, \pi_t) \right\} \iff \pi_{t+1}(a) = \frac{\pi_t(a) e^{\eta \mathbf{r}_t(a)}}{\sum_{b \in \mathcal{A}} \pi_t(b) e^{\eta \mathbf{r}_t(b)}}$$


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**

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

Inverse Propensity Weighting / Inverse Probability Weighting / Importance Weighting

Directly Applying Exponential Weights

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

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

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

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

Define for all a :

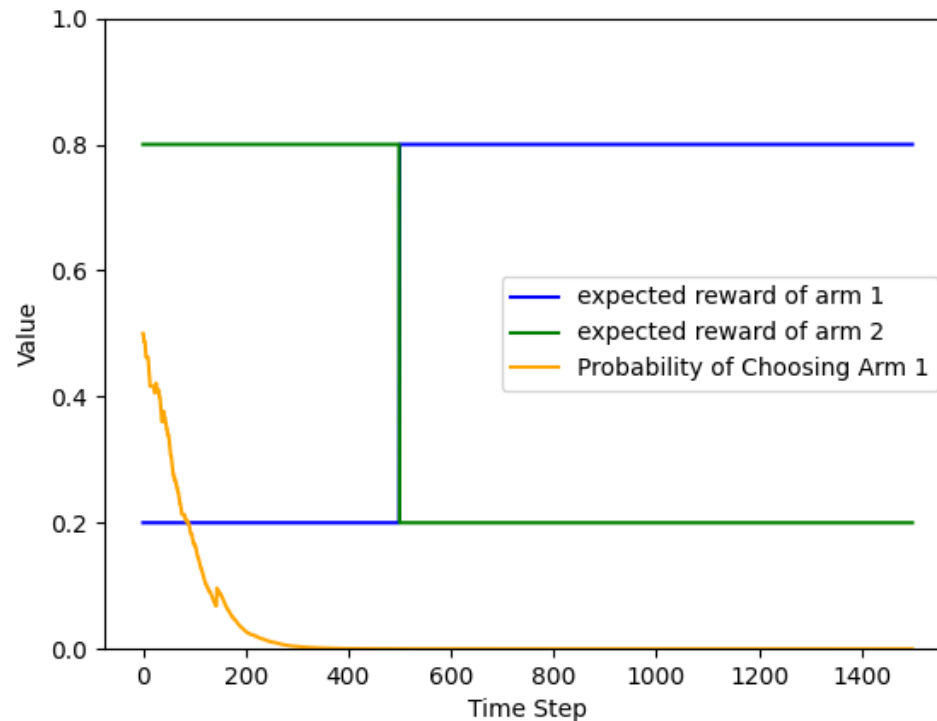
$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a)} \mathbb{I}\{a_t = a\}$$

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 $r_t(a_t)$

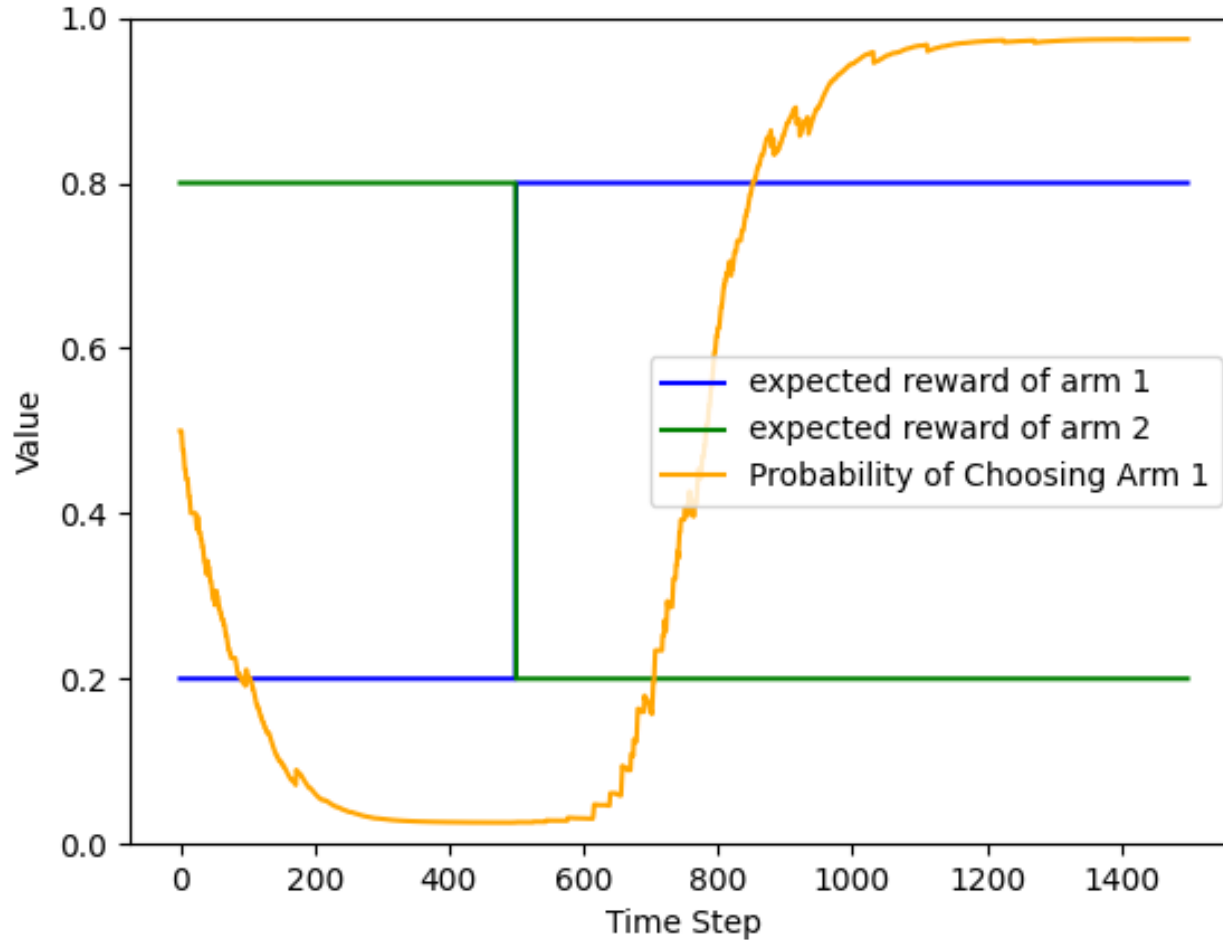
Define for all a :

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

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

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

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

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

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

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

$\text{Fix } a, \mathbb{E}[\hat{r}_t(a)]$
 $= \cancel{p_r(a=a)} \cdot \left(\frac{r_t(a) - 1}{\pi_t(a)} + 1 \right) + p_r(a \neq a) \cdot 1$
 $= \pi_t(a) \left(\frac{r_t(a) - 1}{\pi_t(a)} + 1 \right) + (1 - \pi_t(a))$
 $= r_t(a) - 1 + 1 = r_t(a)$

- Why this resolves the issue? ...

Applying Solution 2

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

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:

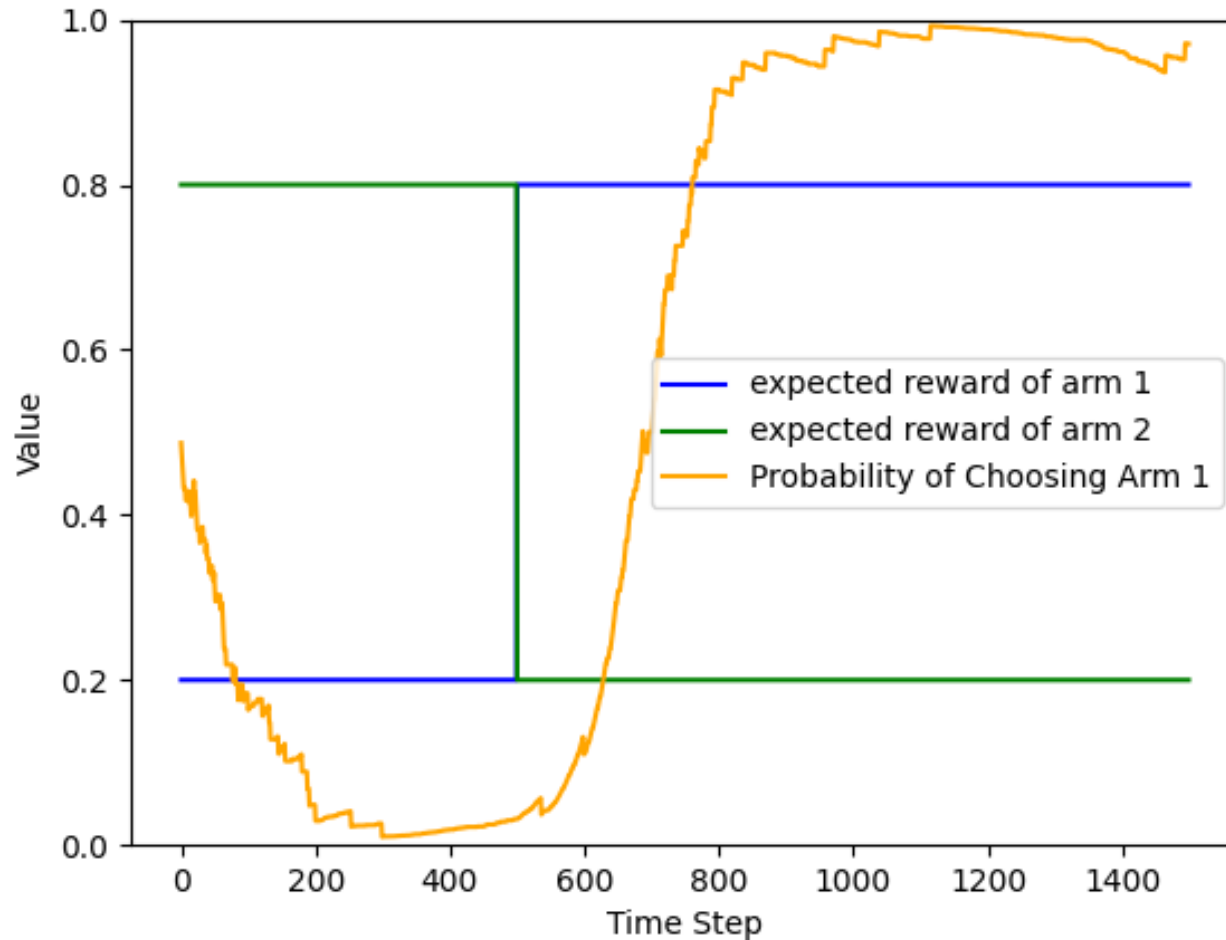
$$\pi_{t+1}(a) = \frac{\pi_t(a) \exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} \pi_t(a') \exp(\eta \hat{r}_t(a'))}$$

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

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

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

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

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

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

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

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

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

Small bias for often-picked arms

More negative bias for seldom-picked arms



Small bias for often-picked arms

More positive bias for seldom-picked arms



EXP3-IX

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

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

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

Define for all a :

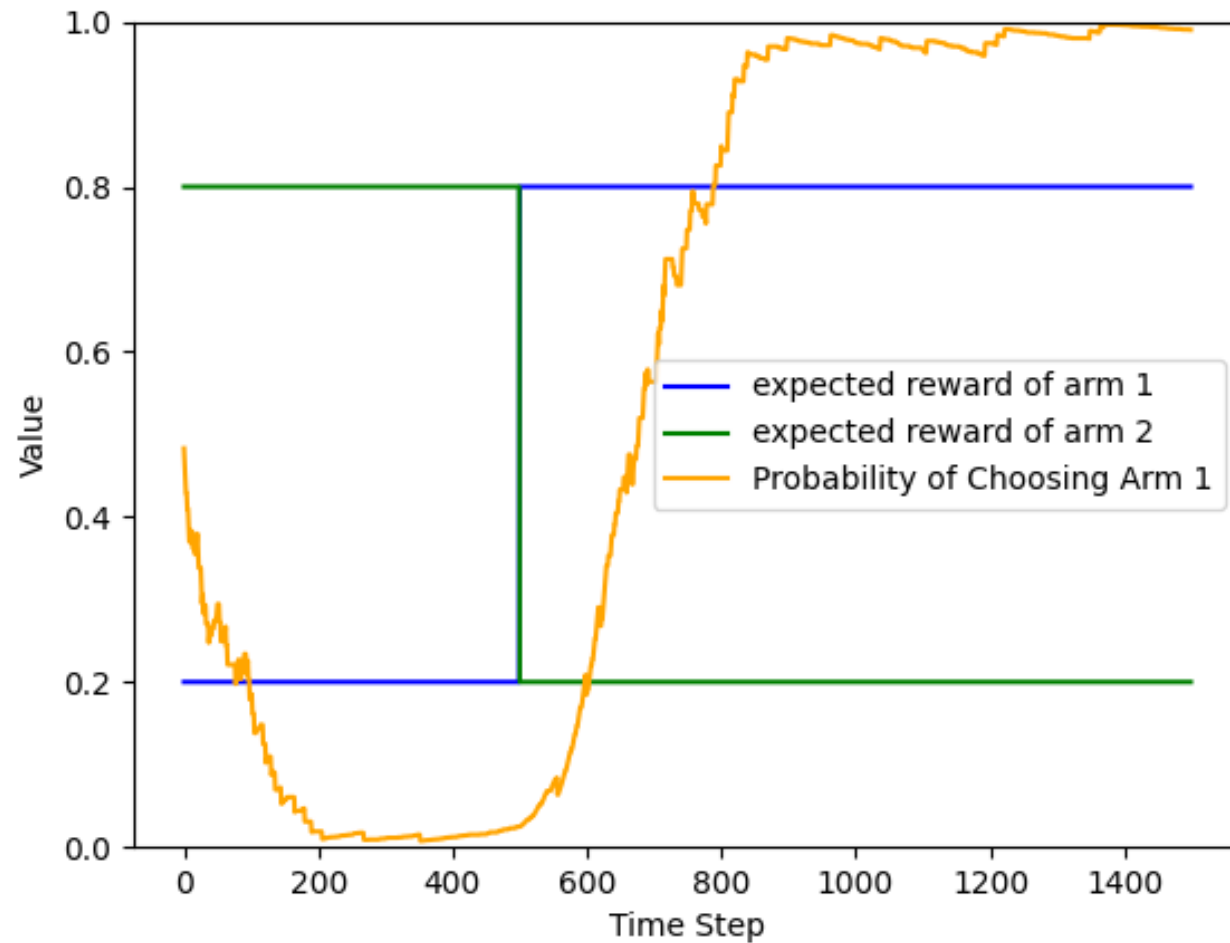
$$\hat{r}_t(a) = \frac{r_t(a) - 1}{\pi_t(a) + \eta} \mathbb{I}\{a_t = a\}$$

Update policy:

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

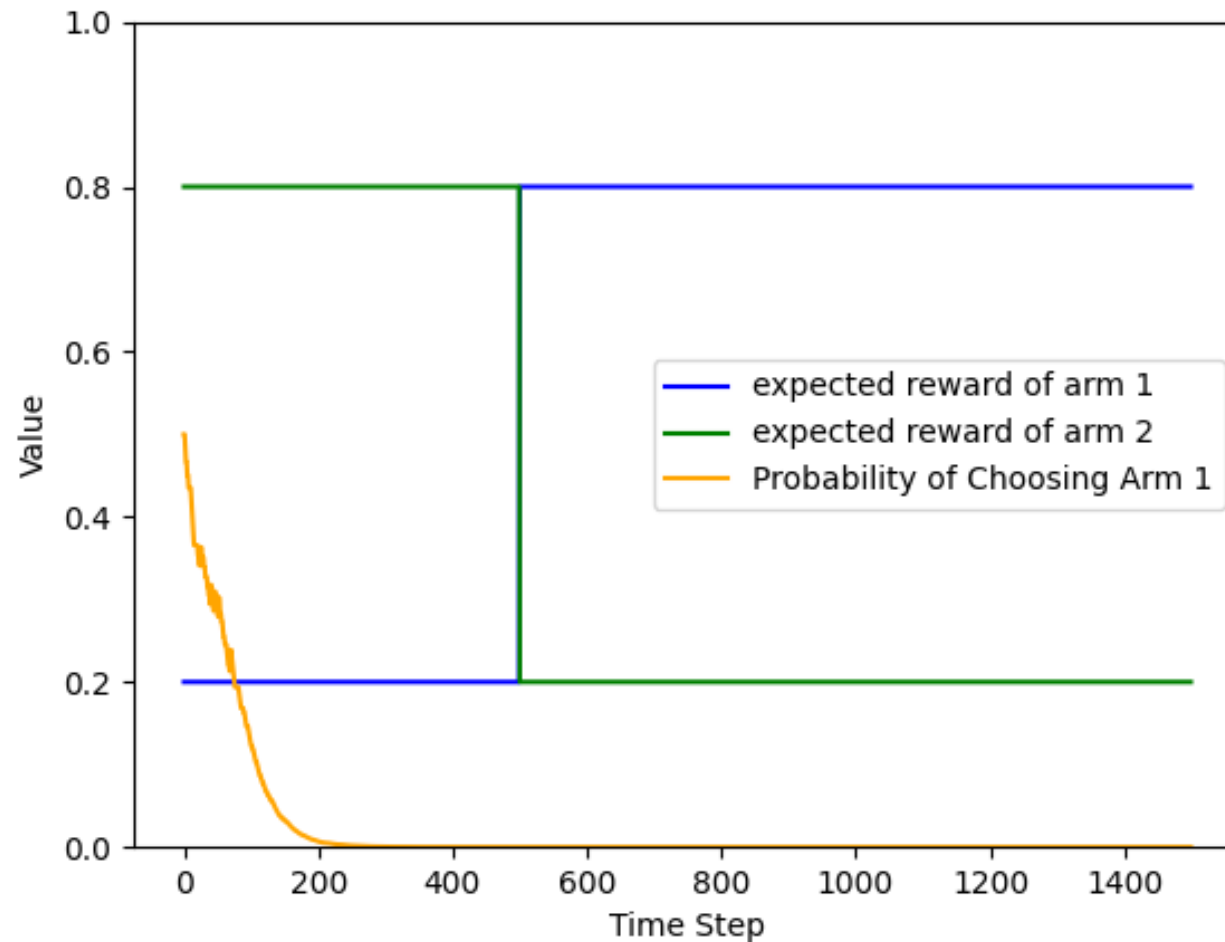
EXP3-IX

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



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 EXP3-IX

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.

The Role of Baseline

$$V_t \in [0, 1]$$

$$b_t \in [0, 1]$$

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

$$\pi_{t+1}(a) = \frac{\pi_t(a) \exp(\eta \hat{r}_t(a))}{\sum_{a' \in \mathcal{A}} \pi_t(a') \exp(\eta \hat{r}_t(a'))} \quad \text{or} \quad \pi_{t+1} = \operatorname{argmax}_{\pi \in \Delta(\mathcal{A})} \left\{ \langle \pi, \hat{r}_t \rangle - \frac{1}{\eta} \text{KL}(\pi, \pi_t) \right\}$$

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

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

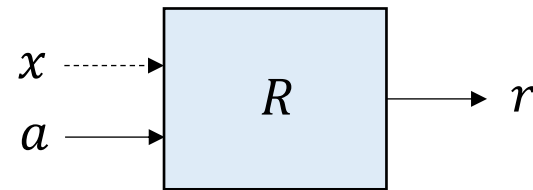
Summary

- 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

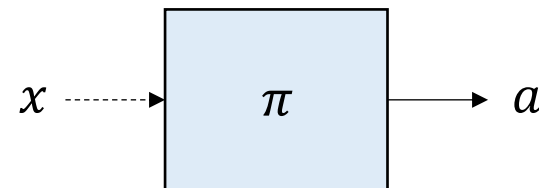
x : context, a : action, r : reward

Value-based



(context, action) to reward

Policy-based



context to action distribution

MAB

Mean estimation
+
EG, BE, IGW

CB

Regression
+
EG, BE, IGW

KL-regularized update
with reward estimators
(EXP3)

+
baseline, bias, or
uniform exploration

Next

Contextual Bandits

Contextual Bandits

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

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

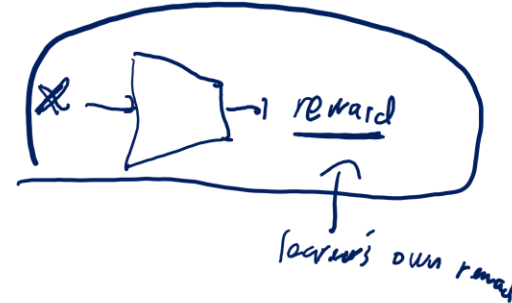
Learner chooses an action $a_t \in \mathcal{A}$

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

KL-Regularized Policy Updates



$$\pi_{\theta}(a|x)$$



MAB

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

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

$$\sum \pi(a) \cdot h_t = b_t$$

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

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

KL-Regularized Policy Updates

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

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

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

$\text{KL}(\underbrace{\pi_{\theta}(\cdot|x_t)}_{\text{KL}(\pi_{\theta}(\cdot|x_t), \pi_{\theta_t}(\cdot|x_t))}, \pi_{\theta_t}(\cdot|x_t))$

Proximal Policy Optimization (PPO) for CB

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

For $i = 1, \dots, N$: (2048)

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 $j = 1, \dots, M$: (10)

For minibatch $\mathcal{B} \subset \{1, 2, \dots, N\}$ of size B : (64)

one iteration of mirror ascent

$$\begin{aligned} \theta &\leftarrow \theta + \nabla_{\theta} \frac{1}{B} \sum_{i \in \mathcal{B}} \left(\sum_a \pi_{\theta}(a|x_i) \hat{r}_i(x_i, a) - \frac{1}{\eta} \sum_a \pi_{\theta}(a|x_i) \log \frac{\pi_{\theta}(a|x_i)}{\pi_{\theta_t}(a|x_i)} \right) \\ &= \theta + \nabla_{\theta} \frac{1}{B} \sum_{i \in \mathcal{B}} \left(\frac{\pi_{\theta}(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)} (r_i(x_i, a_i) - b_t(x_i)) - \frac{1}{\eta} \sum_a \pi_{\theta}(a|x_i) \log \frac{\pi_{\theta}(a|x_i)}{\pi_{\theta_t}(a|x_i)} \right) \end{aligned}$$

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

Proximal Policy Optimization (PPO) for CB

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

- May replace $\text{KL}(\pi_{\theta}(\cdot | x_i), \pi_{\theta_t}(\cdot | x_i))$ by $\text{KL}(\pi_{\theta_t}(\cdot | x_i), \pi_{\theta}(\cdot | x_i))$. The latter is easier to construct unbiased estimator.
- 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 $kl_i(\theta_t, \theta) = \frac{\pi_{\theta}(a_i | x_i)}{\pi_{\theta_t}(a_i | x_i)} - 1 - \log \frac{\pi_{\theta}(a_i | x_i)}{\pi_{\theta_t}(a_i | x_i)}$

Then $\mathbb{E}_{a_i \sim \pi_{\theta_t}(\cdot | x_i)}[kl_i(\theta_t, \theta)] = \text{KL}(\pi_{\theta_t}(\cdot | x_i), \pi_{\theta}(\cdot | x_i))$ Just need one sample of a_i

As we see before, the ways to construct an unbiased estimator are not unique.
This is a good one with low variance.

PPO with KL Estimator

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 $j = 1, \dots, M$:

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

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

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

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

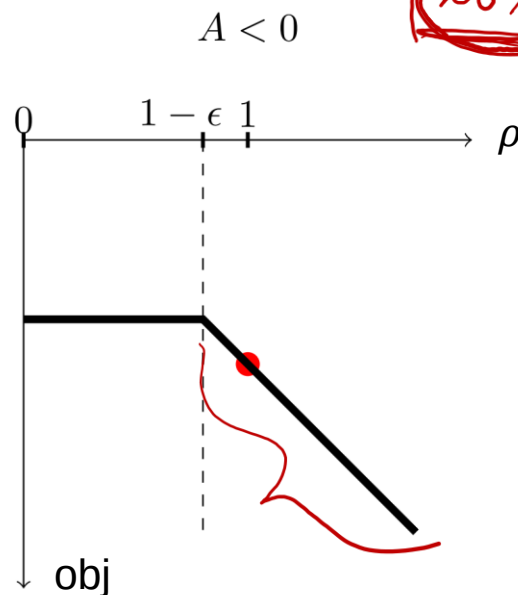
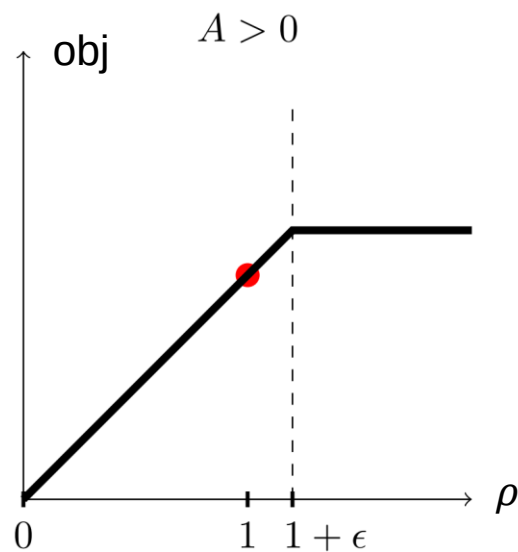
Additional Technique for PPO: Clipped Estimator

$$\rho = \frac{\pi_{\theta}(a|x)}{\pi_{\theta_k}(a|x)}$$

$$A = r(x, a) - b(x)$$

$$\min \{1+\epsilon, \max \{1-\epsilon, \rho\}\}$$

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



ρA

$(1+\epsilon)A$

$\rho = \infty$

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 $d_{\text{targ}} = 0.003$	0.68
Adaptive KL $d_{\text{targ}} = 0.01$	0.74
Adaptive KL $d_{\text{targ}} = 0.03$	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

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 additionally uses batching, reversed KL divergence, and KL estimators for computational efficiency

NPG and PG

Natural Policy Gradient

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

η close to zero

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

where $F_{\theta_t} = \mathbb{E}_x \mathbb{E}_{a \sim \pi_{\theta_t}(\cdot|x)} \left[(\nabla_{\theta} \log \pi_{\theta}(a|x)) (\nabla_{\theta} \log \pi_{\theta}(a|x))^{\top} \right] \Big|_{\theta=\theta_t}$ **Fisher information matrix**

Natural Policy Gradient (w/o context + full-info)

(PPO)
$$\theta_{t+1} = \operatorname{argmax}_{\theta} \sum_a \left(\pi_{\theta}(a) - \pi_{\theta_t}(a) \right) r_t(a) - \frac{1}{\eta} \sum_a \pi_{\theta}(a) \log \frac{\pi_{\theta}(a)}{\pi_{\theta_t}(a)}$$

η close to zero

(NPG)

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \sum_a \nabla_{\theta} \pi_{\theta}(a) r_t(a) \Big|_{\theta=\theta_t}$$

$\theta \in \mathbb{R}^d$
 $F_{\theta_t} \in \mathbb{R}^{d \times d}$

where $F_{\theta_t} = \mathbb{E}_{a \sim \pi_{\theta_t}} [(\nabla_{\theta} \log \pi_{\theta}(a)) (\nabla_{\theta} \log \pi_{\theta}(a))^{\top}] \Big|_{\theta=\theta_t}$ Fisher information matrix

Proof Sketch

$$f(\theta) \approx f(\theta_t) + (\theta - \theta_t)^\top [\nabla_\theta f(\theta)]_{\theta=\theta_t} + \frac{1}{2} (\theta - \theta_t)^\top [\nabla_\theta^2 f(\theta)]_{\theta=\theta_t} (\theta - \theta_t)$$

PPO

$$\theta_{t+1} = \operatorname{argmax}_\theta \left\{ \langle \pi_\theta - \pi_{\theta_t}, r_t \rangle - \frac{1}{\eta} \operatorname{KL}(\pi_\theta, \pi_{\theta_t}) \right\}$$

$$\left. \nabla_\theta \operatorname{KL}(\pi_\theta, \pi_{\theta_t}) \right|_{\theta=\theta_t} = 0$$

$$\begin{aligned} \langle \pi_\theta - \pi_{\theta_t}, r_t \rangle &= \sum_a (\pi_\theta(a) - \pi_{\theta_t}(a)) r_t(a) \\ &\approx (\theta - \theta_t)^\top \sum_a [\nabla_\theta \pi_\theta(a)]_{\theta=\theta_t} r_t(a) \end{aligned}$$

g_t

$$\begin{aligned} \theta_{t+1} &\approx \operatorname{argmax}_\theta \left\{ (\theta - \theta_t)^\top g_t - \frac{1}{2\eta} (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \right\} \\ &= \theta_t + \eta F_{\theta_t}^{-1} g_t \quad \textbf{NPG} \end{aligned}$$

$$F_{\theta_t} = [\nabla_\theta^2 \operatorname{KL}(\pi_\theta, \pi_{\theta_t})]_{\theta=\theta_t} \quad \textbf{(exercise)}$$

$$\operatorname{KL}(\pi_\theta, \pi_{\theta_t}) \approx \frac{1}{2} (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t)$$

NPG vs. PG

Expected reward of $\pi_\theta = \sum_a \pi_\theta(a) r_t(a)$

$\mathcal{J}_t = \nabla_\theta (\text{exped reward of } \pi_\theta)$

NPG

$$\theta_{t+1} = \theta_t + \eta F_t^{-1} \sum_a \nabla_\theta \pi_\theta(a) r_t(a) \Big|_{\theta=\theta_t}$$

$\mathcal{J}_t$

(Vanilla) PG

$$\theta_{t+1} = \theta_t + \eta \sum_a \nabla_\theta \pi_\theta(a) r_t(a) \Big|_{\theta=\theta_t}$$

$\mathcal{J}_t$

NPG vs. PG

NPG

PG

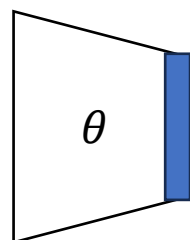
$$\theta_{t+1} = \theta_t + \eta F_t^{-1} \sum_a \nabla_{\theta} \pi_{\theta}(a) r_t(a) \Big|_{\theta=\theta_t}$$

$$\theta_{t+1} = \theta_t + \eta \sum_a \nabla_{\theta} \pi_{\theta}(a) r_t(a) \Big|_{\theta=\theta_t}$$



$$\theta_{t+1} = \operatorname{argmax}_{\theta} \langle \pi_{\theta} - \pi_{\theta_t}, r_t \rangle - \frac{1}{\eta} \text{KL}(\pi_{\theta}, \pi_{\theta_t})$$

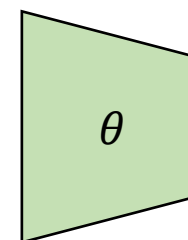
$$\theta_{t+1} = \operatorname{argmax}_{\theta} \langle \pi_{\theta} - \pi_{\theta_t}, r_t \rangle - \frac{1}{2\eta} \|\theta - \theta_t\|^2$$



$\pi_{\theta}(\cdot)$

$\|\pi_{\theta_1} - \pi_{\theta_2}\| \text{ small}$

$\|\theta_1 - \theta_2\| \text{ big}$



$\pi_{\theta}(\cdot)$

Example: NPG vs. PG with softmax policy

Consider multi-armed bandits with **softmax policy** $\pi_{\theta}(a) = \frac{e^{\theta(a)}}{\sum_{a'} e^{\theta(a')}}$
parameterized by $\theta(1), \theta(2), \dots, \theta(A)$

NPG (= Exponential Weight, without requiring $\eta \approx 0$ assumption)

For $t = 1, 2, \dots$

$$\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta r_t(a)$$

Check the equivalence (exercise)

NPG can also be written as

$$\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \tilde{r}_t(a)$$

$$\tilde{r}_t(a) = r_t(a) - \sum_{a'} \pi_{\theta_t}(a') r_t(a')$$

PG

For $k = 1, 2, \dots$

$$\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \pi_{\theta_t}(a) \tilde{r}_t(a)$$

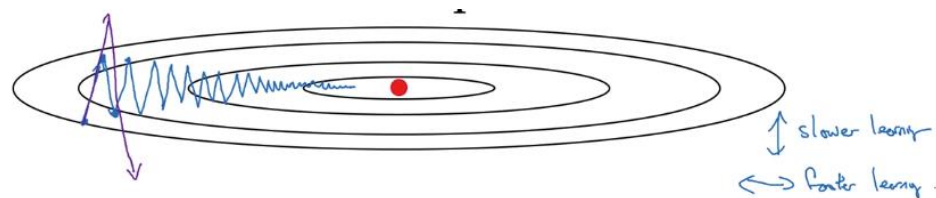
NPG (EW) vs. PG

600 total reward

Reward = [Ber(0.6), Ber(0.4)]

Initial policy $\pi = [0.0001, 0.9999]$

Plot total reward in 1000 rounds



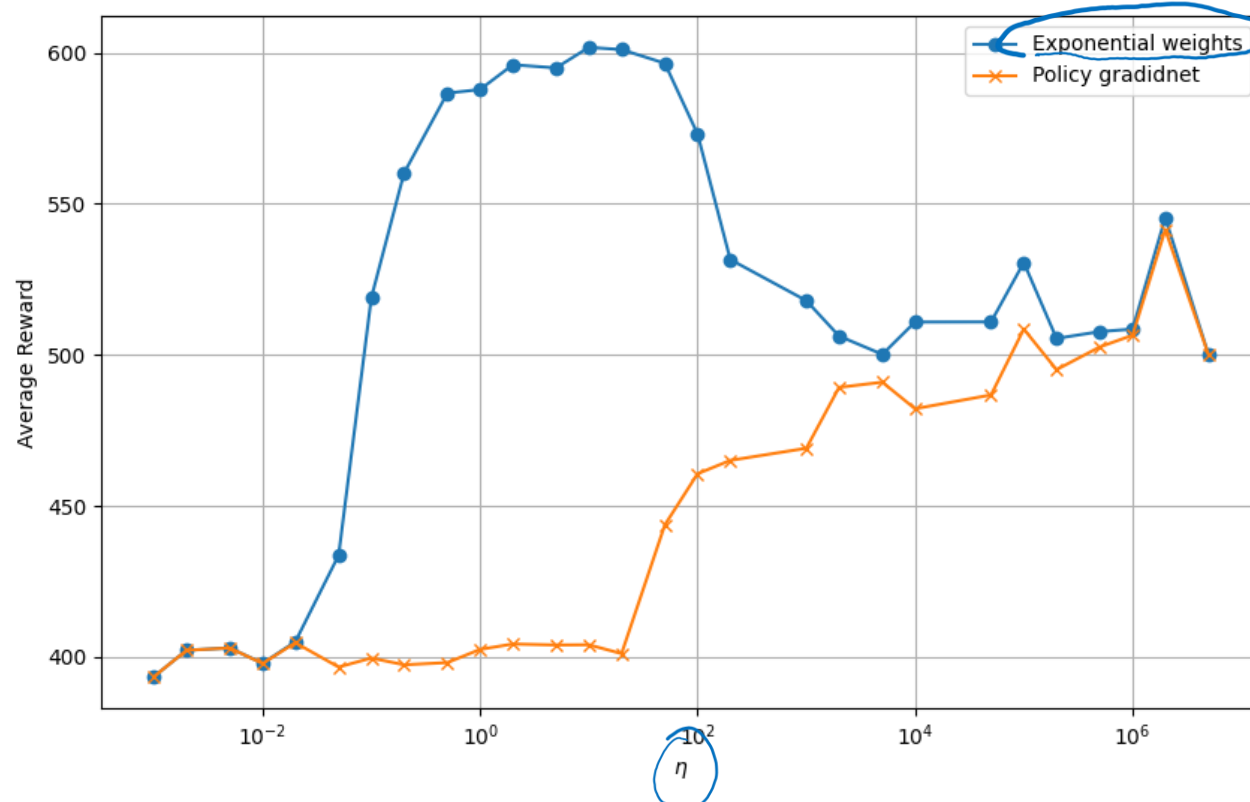
<https://math.stackexchange.com/questions/2285282/relating-condition-number-of-hessian-to-the-rate-of-convergence>

EW: $\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \tilde{r}_t(a)$

PG: $\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \pi_{\theta_t}(a) \tilde{r}_t(a)$

$\theta_t(1)$

$\pi_{\theta_t}(2) \tilde{r}_t(a)$



NPG and PG with bandit feedback

$$\theta_{t+1} = \theta_t + \eta F_t^{-1} \sum_a \nabla_{\theta} \pi_{\theta}(a) \hat{r}_t(a) \Big|_{\theta=\theta_t}$$

g_t

$$\theta_{t+1} = \theta_t + \eta \sum_a \nabla_{\theta} \pi_{\theta}(a) \hat{r}_t(a) \Big|_{\theta=\theta_t}$$

g_t

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

$$g_t \Big|_{\theta=\theta_t} = \sum_a \nabla_{\theta} \pi_{\theta}(a) \cdot \frac{r_t(a) - b}{\pi_t(a)} \mathbb{I}\{a_t = a\} = \frac{\nabla_{\theta} \pi_{\theta}(a_t)}{\pi_t(a_t)} (r_t(a_t) - b) \Big|_{\theta=\theta_t}$$

$$= \left(\nabla_{\theta} \log \pi_{\theta}(a_t) \right) \Big|_{\theta=\theta_t} (r_t(a_t) - b)$$

PG (REINFORCE) for contextual bandits

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

Receive context x_t

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

Update

$$\theta_{t+1} \leftarrow \theta_t + \eta [\nabla_{\theta} \log \pi_{\theta}(a_t|x_t)]_{\theta=\theta_t} (r_t(x_t, a_t) - b_t(x_t))$$


g_t

Or simply written as

$$\theta \leftarrow \theta + \eta \underbrace{\nabla_{\theta} \log \pi_{\theta}(a_t|x_t)}_{\text{Coming from inverse propensity weighting / importance weighting}} (r_t(x_t, a_t) - b_t(x_t))$$

Coming from inverse propensity weighting / importance weighting

Verify (again) that reward offset does not affect the algorithm

$$g_t = \sum_a \nabla_{\theta} \pi_{\theta}(a) \left(\hat{r}(a) + b \right)$$

$$= \sum_a \nabla_{\theta} \pi_{\theta}(a) \hat{r}(a) + b \sum_a \nabla_{\theta} \pi_{\theta}(a)$$

$$\hat{r}(a) = \frac{r(a) - b}{\pi_t(a)} \mathbb{1}\{a_t = a\} + b$$

$$r(a) - b$$

$$\nabla_{\theta} \left(\sum_a \pi_{\theta}(a) \right) = 0$$

Natural Policy Gradient

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

Receive context x_t

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

Update

$$\theta_{t+1} \leftarrow \theta_t + \eta \mathbf{F}_{\theta_t}^{-1} [\nabla_{\theta} \log \pi_{\theta}(a_t|x_t)]_{\theta=\theta_t} (r_t(x_t, a_t) - b_t(x_t))$$

A naïve calculation of $\mathbf{F}_{\theta_t}^{-1}$ will take $O(d^3)$ time

Sample-Based NPG*

A naïve calculation of $F_{\theta_t}^{-1}$ will take $O(d^3)$ time

But we can actually view $h_t := F_{\theta_t}^{-1} g_t$ as a solution of a linear regression problem

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \mathbb{E}_{a \sim \pi_{\theta_t}} [(\nabla_{\theta} \log \pi_{\theta_t}(a)) r_t(a)]$$

$$\text{where } F_{\theta_t} = \mathbb{E}_{a \sim \pi_{\theta_t}} [(\nabla_{\theta} \log \pi_{\theta_t}(a)) (\nabla_{\theta} \log \pi_{\theta_t}(a))^{\top}]$$

$$h_t = \left(\mathbb{E}_{a \sim \pi_{\theta_t}} [\phi_t(a) \phi_t(a)^{\top}] \right)^{-1} \mathbb{E}_{a \sim \pi_{\theta_t}} [\phi_t(a) r_t(a)]$$

$$= \underset{h}{\operatorname{argmin}} \mathbb{E}_{a \sim \pi_{\theta_t}} [(\phi_t(a)^{\top} h - r_t(a))^2]$$

$$\phi_t(a) = \nabla_{\theta} \log \pi_{\theta_t}(a)$$

Summary: Policy Learning in Bandits

PG	PPO / NPG
$\theta_{t+1} = \operatorname{argmax}_{\theta} \underbrace{\langle \pi_{\theta} - \pi_{\theta_t}, \hat{r}_t \rangle} - \frac{1}{2\eta} \ \theta - \theta_t\ ^2$	$\theta_{t+1} = \operatorname{argmax}_{\theta} \underbrace{\langle \pi_{\theta} - \pi_{\theta_t}, \hat{r}_t \rangle} - \frac{1}{\eta} \operatorname{KL}(\pi_{\theta}, \pi_{\theta_t})$
$\theta \leftarrow \theta + \eta \underbrace{\nabla_{\theta} \langle \pi_{\theta}, \hat{r}_t \rangle}$	$\theta \leftarrow \theta + \eta \underbrace{F_{\theta}^{-1}} \nabla_{\theta} \langle \pi_{\theta}, \hat{r}_t \rangle$
$\hat{r}_t(a) = \frac{r_t(a) - b_t}{\pi_{\theta_t}(a)} \mathbb{I}\{a = a_t\}$	
$\theta \leftarrow \theta + \eta \underbrace{\nabla_{\theta} \log \pi_{\theta}(a_t)} (r_t(a_t) - b_t)$	$\theta \leftarrow \theta + \eta \underbrace{F_{\theta}^{-1}} \nabla_{\theta} \log \pi_{\theta}(a_t) (r_t(a_t) - b_t)$

$$\underbrace{F_{\theta} = \mathbb{E}_{a \sim \pi_{\theta}} [(\nabla_{\theta} \log \pi_{\theta}(a)) (\nabla_{\theta} \log \pi_{\theta}(a))^{\top}]}_{d \times d} \quad \theta \in \mathbb{R}^d$$