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

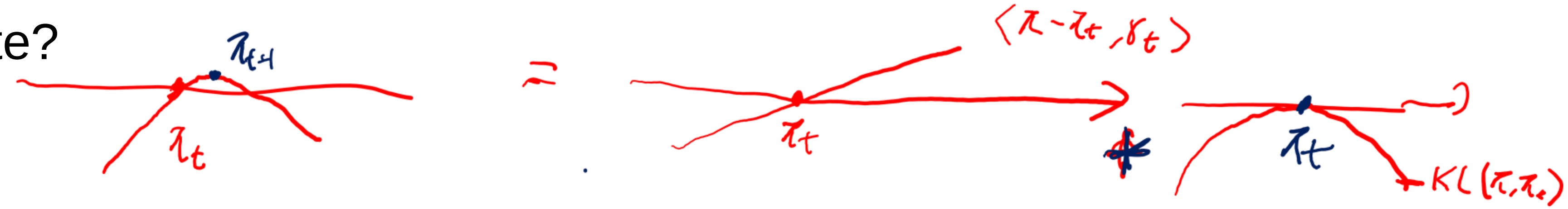
$$a_t \sim \pi_t \rightarrow r_t = \begin{bmatrix} r_t(1) \\ \vdots \\ r_t(A) \end{bmatrix}$$

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

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

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

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 $\eta r_t(a) \leq 1$ 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

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

$\sqrt{AT}$ ~ bandit

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

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

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.

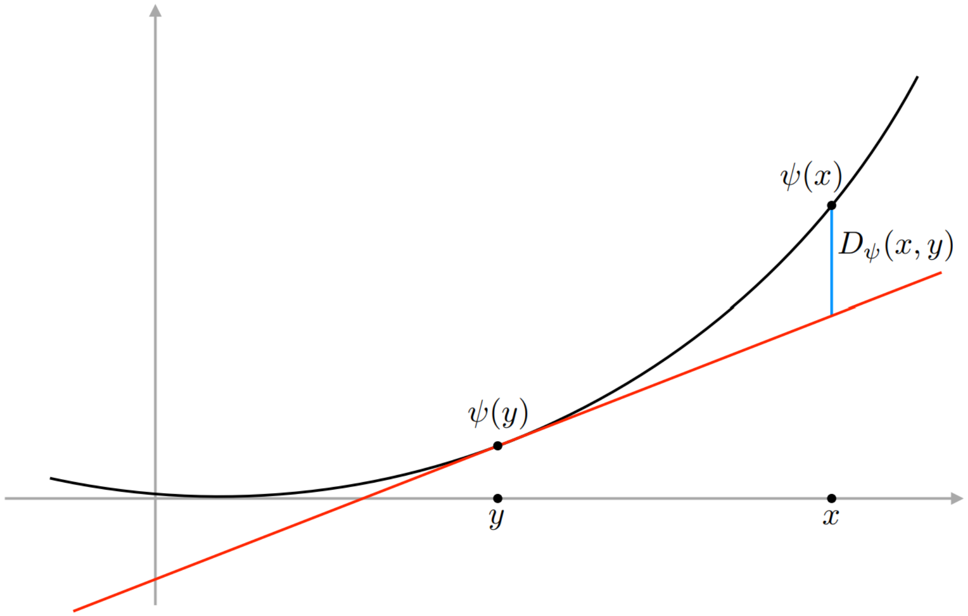
Exponential Weight Update ∈ Mirror Ascent

General form of **Mirror Ascent**:

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

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

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**	$\sum_a x(a) \log \frac{x(a)}{y(a)}$	$x_{t+1}(a) = \frac{x_t(a) e^{\eta r_t(a)}}{\sum_b x_t(b)\, e^{\eta r_t(b)}}$
$\sum_a \log \frac{1}{x(a)}$	$\sum_a \left(\frac{x(a)}{y(a)} - \log \frac{x(a)}{y(a)} - 1 \right)$	$\frac{1}{x_{t+1}(a)} = \frac{1}{x_t(a)} - \eta r_t(a) + \gamma_t$

(for distributions)

(for distributions)

Normalization factor

Regret Analysis for Exponential Weights

Theorem.

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

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

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

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

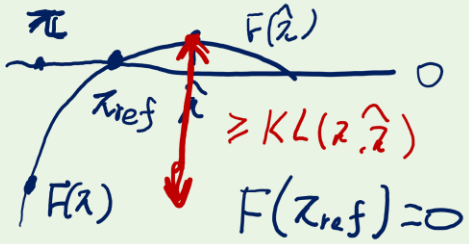
Regret Analysis for Exponential Weights

Useful Lemma

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

For fixed π_{ref} and v, define

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



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

(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?

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

No longer observable

Only update the arm that we choose?

Exponential Weight Updates for Bandits?

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

- $\hat{r}_t(a)$ 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 $t = 1,2,\ldots,T$:

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

Define for all a:

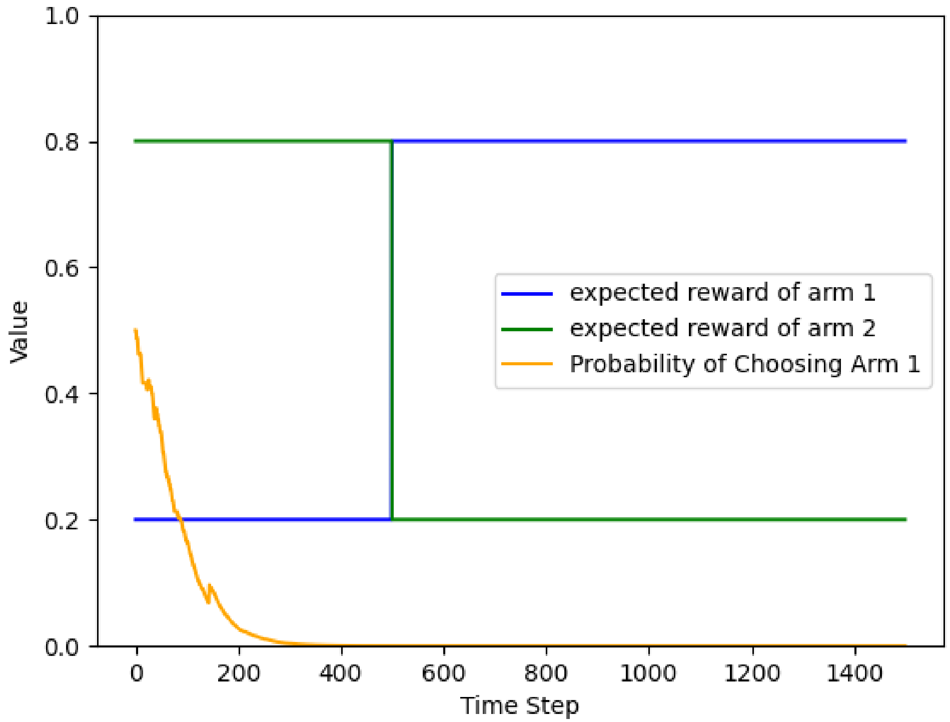
$$\hat{r}_t(a) = \frac{r_t(a)}{\pi_t(a)} \mathbb{I}\{a_t = a\}$$

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 $a^\star$,

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

How to relate the regret with this? Is this still well-bounded?

$\sqrt{AT \ln A}$

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

$\hat{r}_t$ is unbiased estimator

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

w.p. $1 - A\eta \Rightarrow$ use π_t

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

ε

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

Then the unbiased reward estimator becomes

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

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

Applying Solution 1

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

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

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

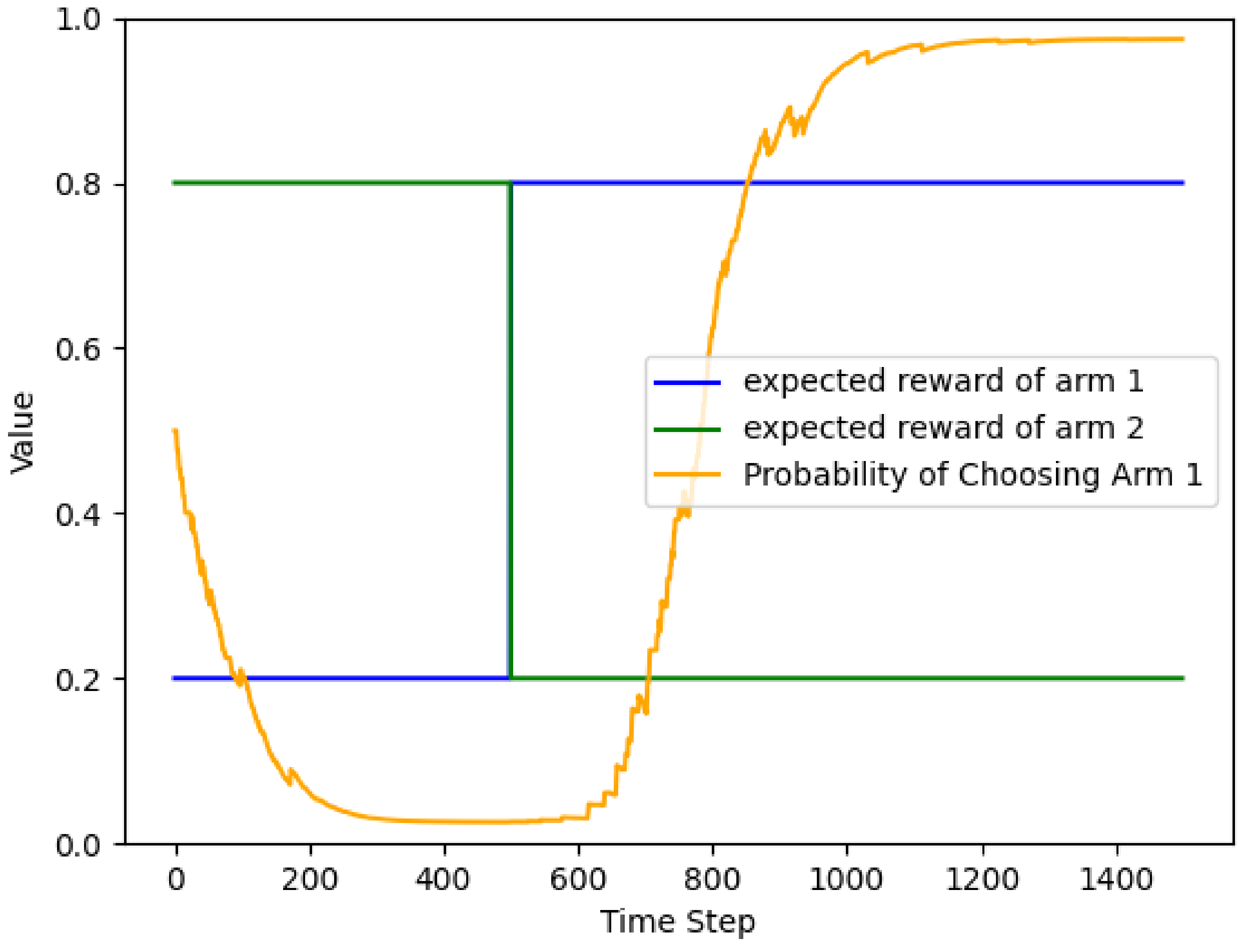
Define for all a:

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

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

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

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

$\sqrt{AT \ln A}$

Solution 2: Reward Estimator with a Baseline

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

- Notice that the condition is only $\eta \hat{r}_t(a) \le 1$. 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

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

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

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

$\pi_1(a) = 1/A$ 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:

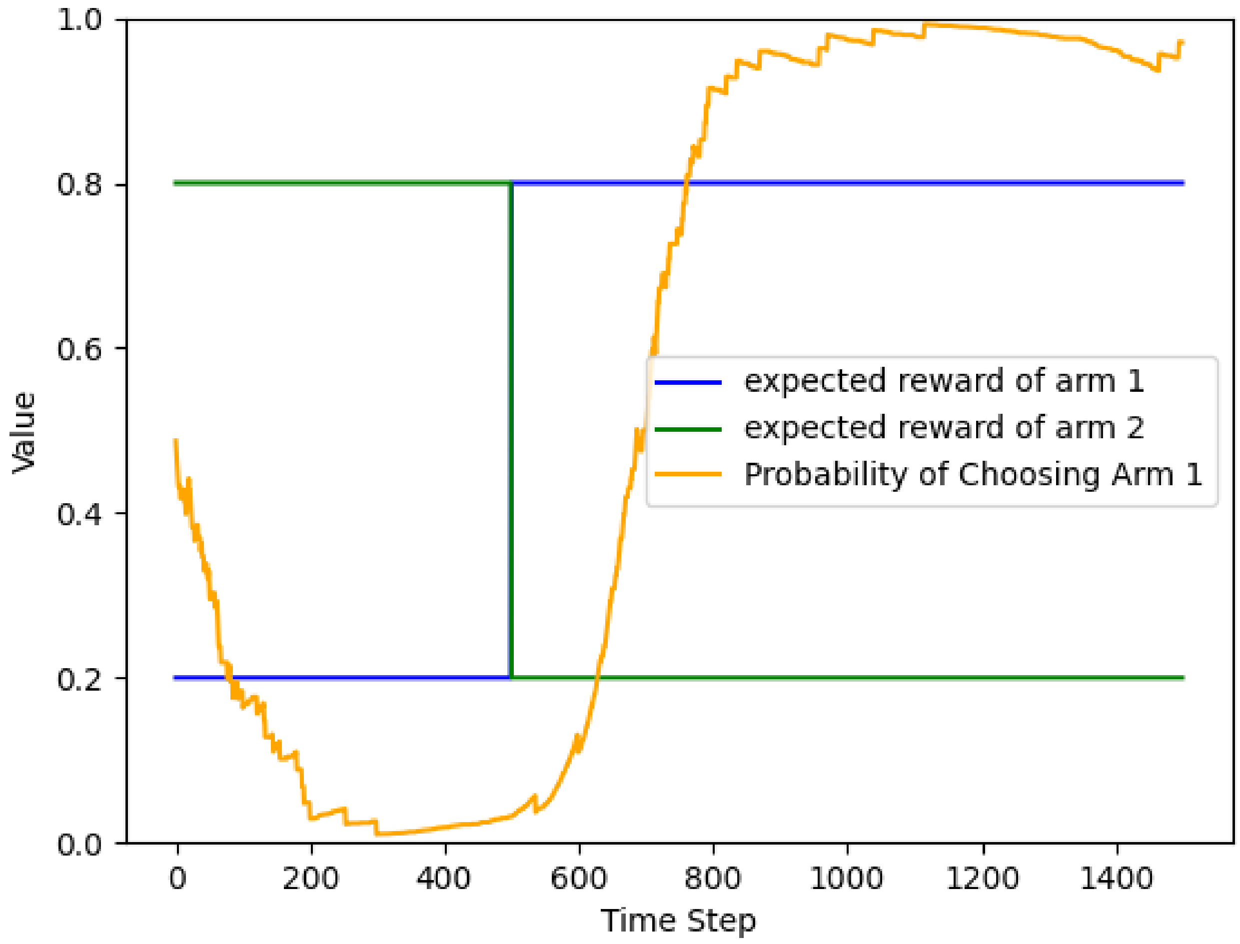
baseline

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

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

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

✗ ✓

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

Biasing

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

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

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

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

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

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

Define for all a:

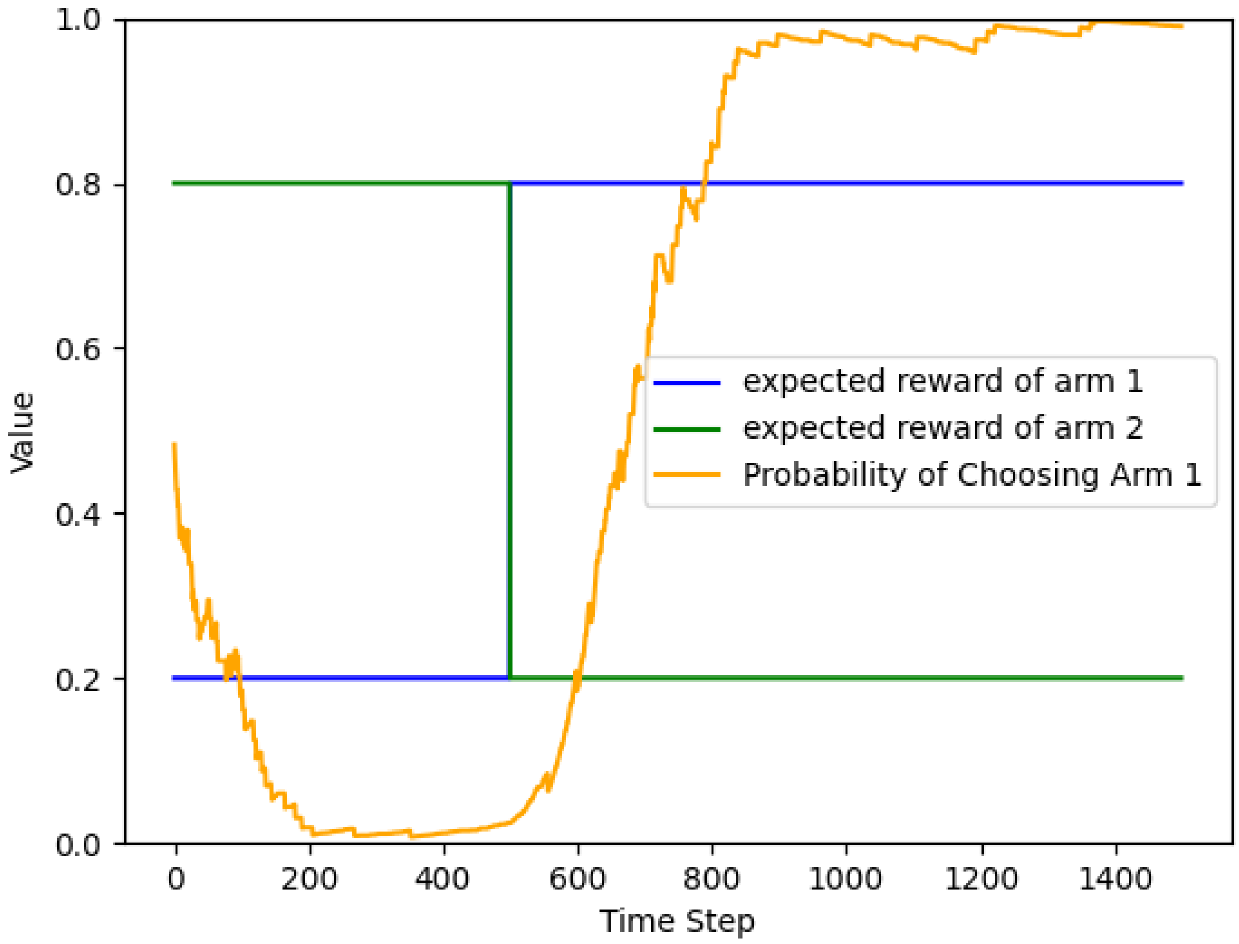
$$\hat{r}_t(a) = \frac{r_t(a) - 1}{\pi_t(a) + \eta}\mathbb{I}\{a_t = a\}$$

Update policy:

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

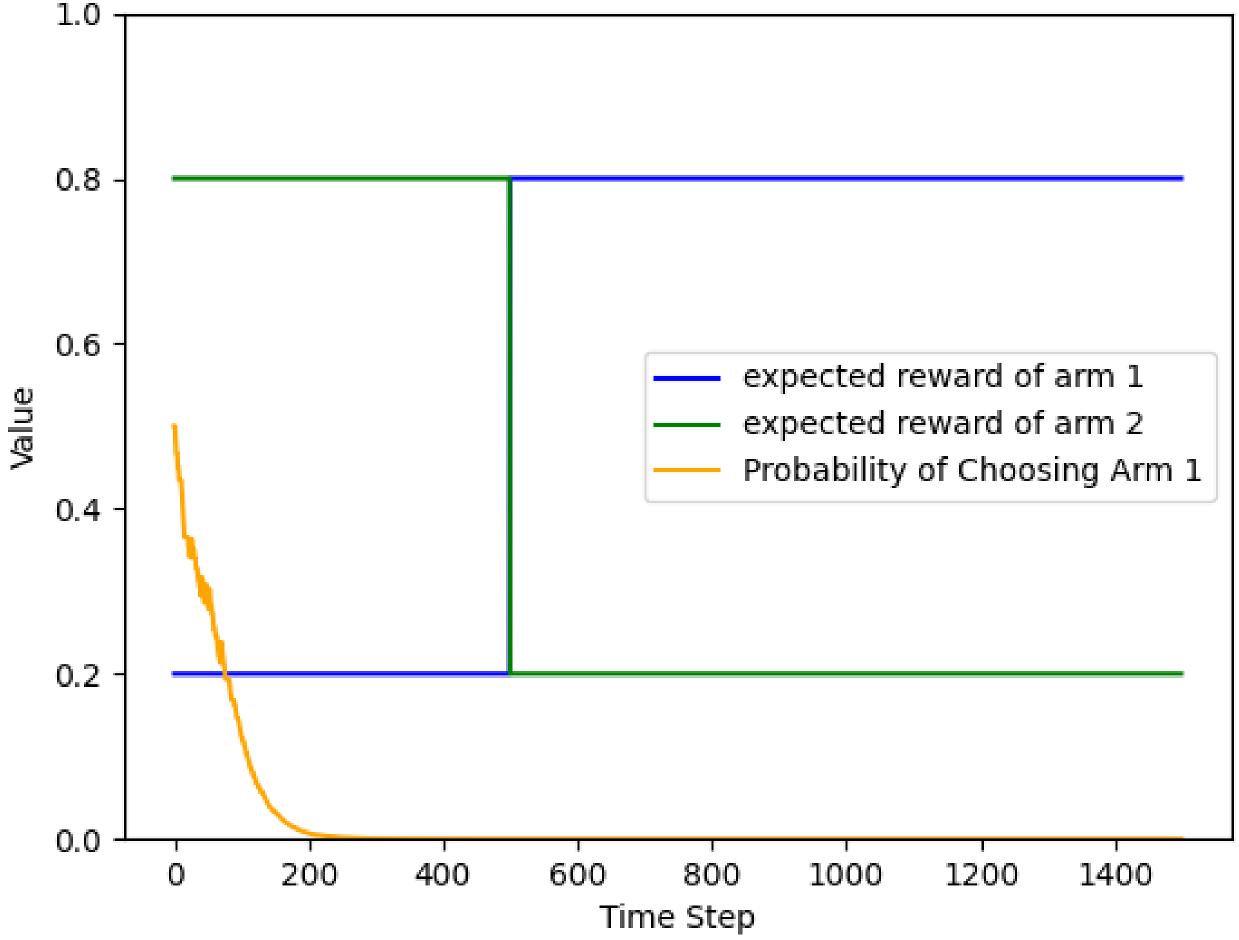
EXP3-IX

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



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

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

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

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

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

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

Summary

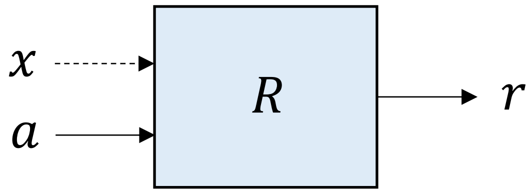
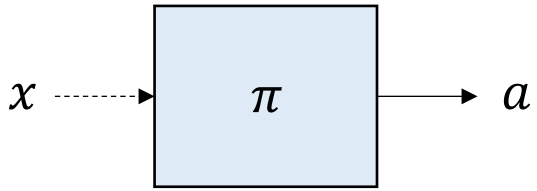
- Exponential weight update (EWU) is an effective algorithm for full-information setting. It guarantees sublinear regret even when the environment changes over time.

- 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

		MAB	CB
Value-based	 (context, action) to reward	Mean estimation + EG, BE, IGW	Regression + EG, BE, IGW
Policy-based	 context to action distribution	KL-regularized update with reward estimators (EXP3) + baseline, bias, or uniform exploration	**Next**

Contextual Bandits

Contextual Bandits

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

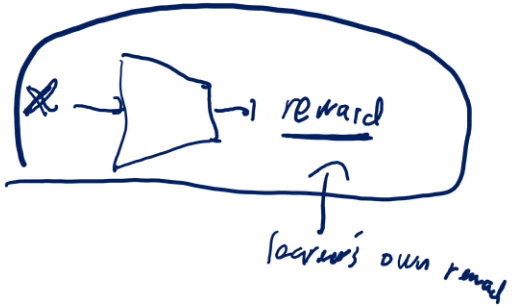
Environment generates a context $x_t \in \mathcal{X}$

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

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



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

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

MAB

$\sum_a \pi(a) \cdot b_t = b_t$

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

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

KL-Regularized Policy Updates

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

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

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

Proximal Policy Optimization (PPO) for CB

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

For $i = 1, \ldots, N$: (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 $j = 1, \ldots, M$: (10)

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

2048/64

one iteration of mirror ascent

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

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

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

Proximal Policy Optimization (PPO) for CB

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

- May replace $\mathrm{KL}\left(\pi_\theta(\cdot|x_i), \pi_{\theta_t}(\cdot|x_i)\right)$ by $\mathrm{KL}\left(\pi_{\theta_t}(\cdot|x_i), \pi_\theta(\cdot|x_i)\right)$. 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 $kl_i(\theta_t,\theta) = \boxed{\frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)} - 1 - \log\frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_t}(a_i|x_i)}}$

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, the ways to construct an unbiased estimator are not unique.

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:

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

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

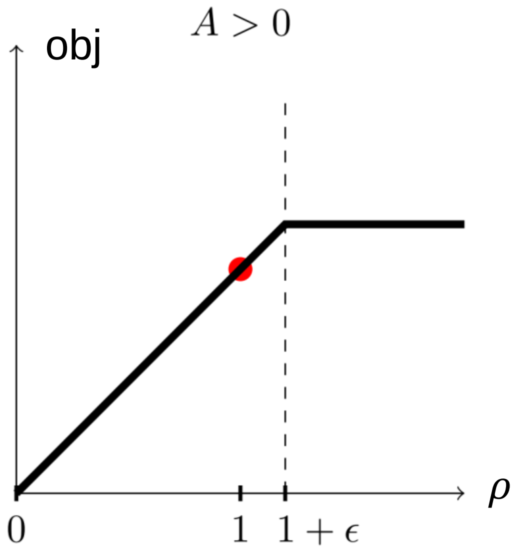
Additional Technique for PPO: Clipped Estimator

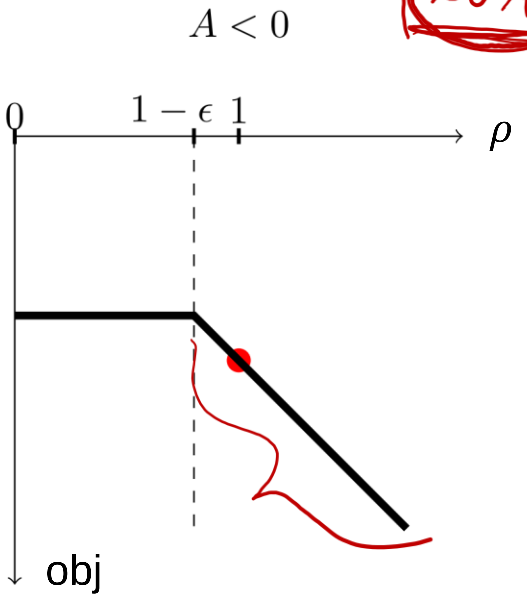
$$\rho = \frac{\pi_\theta(a|x)}{\pi_{\theta_k}(a|x)} \qquad A = r(x,a) - b(x)$$

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

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

$100\,A$ $\qquad (1+\varepsilon)A \qquad \rho = 100$





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

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, reversed KL divergence, and KL estimators for computational efficiency

NPG and PG

Natural Policy Gradient

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

$$\Big\downarrow\ \eta \text{ close to zero}$$

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

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

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

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

η close to zero

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

$\boxed{\theta \in \mathbb{R}^d}$

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

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

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

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

PPO

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

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

$$F_{\theta_t} = \left[\nabla_\theta^2 \mathrm{KL}(\pi_\theta, \pi_{\theta_t})\right]_{\theta=\theta_t}$$ **(exercise)**

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

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

NPG vs. PG

expected reward of π_θ $= \boxed{\sum_a \pi_\theta(a) r_t(a)}$

$g_t = \boxed{\nabla_\theta \left(\text{expcd reward of } \pi_\theta\right)}$

NPG

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

(Vanilla) PG

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

NPG vs. PG

NPG

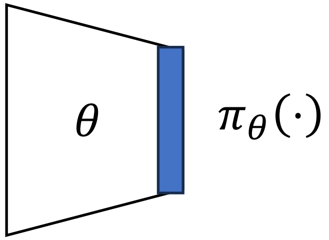
$$\theta_{t+1} = \theta_t + \eta F_t^{-1} \sum_a \nabla_\theta \pi_\theta(a) r_t(a) \Big|_{\theta=\theta_t}$$

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

PG

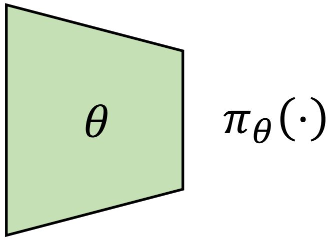
$$\theta_{t+1} = \theta_t + \eta \sum_a \nabla_\theta \pi_\theta(a)\, r_t(a) \Big|_{\theta=\theta_t}$$

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



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

$\|\theta_1 - \theta_2\|$ big



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 $\theta(1), \theta(2), \dots, \theta(A)$

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

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

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

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

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

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

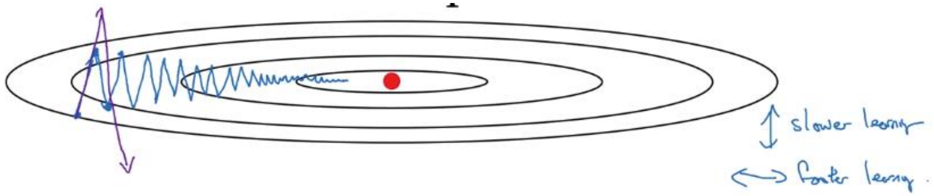
$\boxed{\theta_t(1)}$

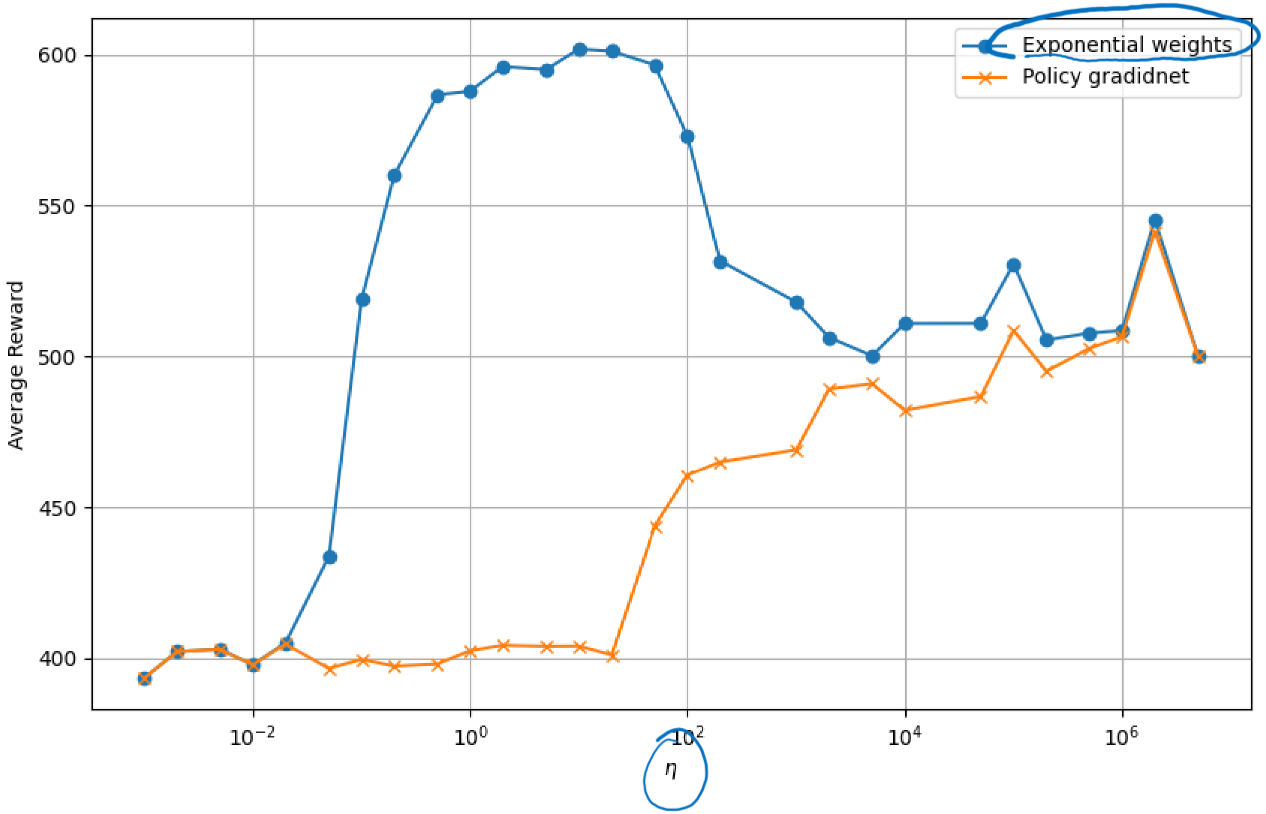
$\pi_{\theta_t}(2)\, \tilde{r}_t(a)$

Reward = $[\mathrm{Ber}(0.6), \mathrm{Ber}(0.4)]$

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

NPG and PG with bandit feedback

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

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

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

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

Or simply written as

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

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

Natural Policy Gradient

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

Receive context x_t

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

Update

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

A naïve calculation of $F_{\theta_t}^{-1}$ 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 $h_t \coloneqq F_{\theta_t}^{-1} g_t$ as a solution of a linear regression problem

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

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

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

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

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

Summary: Policy Learning in Bandits

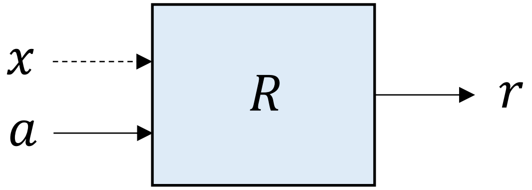
PG	PPO / NPG
$\theta_{t+1} = \operatorname*{argmax}_{\theta} \langle \pi_\theta - \pi_{\theta_t}, \hat{r}_t \rangle - \frac{1}{2\eta} \|\theta - \theta_t\|^2$	$\theta_{t+1} = \operatorname*{argmax}_{\theta} \langle \pi_\theta - \pi_{\theta_t}, \hat{r}_t \rangle - \frac{1}{\eta} \mathrm{KL}(\pi_\theta,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

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

Review: Bandit Techniques

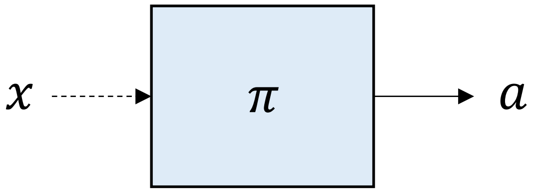
x: context, a: action, r: reward

Value-based



(context, action) to reward

Policy-based



context to action distribution

	MAB	CB
Value-based	Mean estimation + EG, BE, IGW	Regression + EG, BE, IGW
Policy-based	KL-regularized update with reward estimators (EXP3) + baseline, bias, or uniform exploration	PPO/NPG PG + baseline, bias, uniform exploration, clipping

Are we done with bandits?

- Almost, but we have a last important topic: how to deal with continuous action sets? (#actions could be infinite)
- We will go over the 4 regimes once again to deal with continuous actions

	MAB	CB
VB		
PB		

Dealing with Continuous Action Set

Continuous Action Set

Full-information feedback

Given: Action set $\Omega \subseteq \mathbb{R}^d$

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

Learner chooses a point $a_t \in \Omega$

Environment reveals a reward function $r_t\colon\ \Omega \to \mathbb{R}$

Bandit feedback

Given: Action set $\Omega \subseteq \mathbb{R}^d$

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

Learner chooses a point $a_t \in \Omega$

Environment reveals a reward value $r_t(a_t)$

Continuous Multi-Armed Bandits

With a mean estimator

	MAB	CB
VB	•	
PB		

Value-Based Approach (mean estimation)

- Use supervised learning to learn a reward function $R_{\phi}(a)$

- How to perform the exploration strategies (like ϵ-Greedy)?
  - How to find $\mathrm{argmax}_a R_{\phi}(a)$?
  - Usually, there needs to be another **policy learning procedure** that helps to find $\mathrm{argmax}_a R_{\phi}(a)$
  - Then we can explore as $a_t = \mathrm{argmax}_a R_{\phi}(a) + \sigma\, \mathcal{N}(0, I)$

Full-Information Policy learning Procedure

Gradient Ascent

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

Choose action a_t

Receive reward function $r_t\colon\ \Omega \to \mathbb{R}$

Update action $a_{t+1} \leftarrow \mathcal{P}_\Omega(a_t + \eta \nabla r_t(a_t))$

When $\pi_\theta = \mathcal{N}(\mu_\theta, \sigma^2)$, the KL-regularized policy update

$$\theta_{t+1} = \operatorname*{argmax}_{\theta} \left\{ \int \left(\pi_\theta(a) - \pi_{\theta_t}(a)\right) r_t(a)\, \mathrm{d}a - \frac{1}{\eta}\, \mathrm{KL}\left(\pi_\theta, \pi_{\theta_t}\right) \right\}$$

is close to $\mu_{\theta_{t+1}} \leftarrow \mu_{\theta_t} + \eta \nabla r_t(\mu_{\theta_t})$

Regret Bound of Gradient Ascent

Theorem. If Ω is convex and all reward functions r_t are concave, then Gradient Ascent ensures

$$\text{Regret} = \max_{a^\star \in \Omega} \sum_{t=1}^{T} r_t(a^\star) - r_t(a_t) \leq \frac{\max_{a \in \Omega} \|a\|_2^2}{\eta} + \eta \sum_{t=1}^{T} \|\nabla r_t\|_2^2$$

This can also be applied to the finite-action setting, but only ensures a $\sqrt{AT}$ regret bound (using exponential weights we get $\sqrt{(\log A)T}$)

Combining with Mean Estimator

The mean estimator R_ϕ essentially gives us a full-information reward function

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

Take action $\tilde{a}_t = \mathcal{P}_\Omega(a_t + \sigma\, \mathcal{N}(0, I))$

Receive $r_t(\tilde{a}_t)$

Update the mean estimator:

$$\phi \leftarrow \phi - \lambda \nabla_\phi \left[\left(R_\phi(\tilde{a}_t) - r_t(\tilde{a}_t) \right)^2 \right]$$

Update policy:

$$a_{t+1} = \mathcal{P}_\Omega\left(a_t + \eta \nabla_a R_\phi(a_t) \right)$$

Think of this as a continuous-action counterpart of ϵ-Greedy

Continuous Multi-Armed Bandits

Pure policy-based algorithms

	MAB	CB
VB		
PB	•	

Pure Policy-Based Approach

Gradient Ascent

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

Choose action a_t

Receive reward function r_t: $\Omega \rightarrow \mathbb{R}$

Update action $a_{t+1} \leftarrow \mathcal{P}_\Pi(a_t + \eta \boxed{\nabla r_t(a_t)})$

We face a similar problem as in EXP3: if we only observe $r_t(a_t)$, how can we estimate the **gradient**?

(Nearly) Unbiased Gradient Estimator

Goal: construct $g_t \in \mathbb{R}^d$ such that $\mathbb{E}[g_t] \approx \nabla r_t(a_t)$ with only $r_t(a_t)$ feedback

(Nearly) Unbiased Gradient Estimator (1/3)

Uniformly randomly choose a direction $i_t \in \{1, 2, \dots, d\}$

Uniformly randomly choose $\beta_t \in \{1, -1\}$

Sample $\tilde{a}_t = a_t + \delta \beta_t \mathrm{e}_{i_t}$

Observe $r_t(\tilde{a}_t)$

Define $g_t = \frac{d r_t(\tilde{a}_t)}{\delta} \beta_t \mathrm{e}_{i_t}$

(Nearly) Unbiased Gradient Estimator (2/3)

Uniformly randomly choose s_t from the unit sphere $\mathbb{S}_d = \{s \in \mathbb{R}^d: \ \|s\|_2 = 1\}$

Sample $\tilde{a}_t = a_t + \delta s_t$

Observe $r_t(\tilde{a}_t)$

Define $g_t = \frac{d r_t(\tilde{a}_t)}{\delta} s_t$

(Nearly) Unbiased Gradient Estimator (3/3)

Choose $s_t \sim \mathcal{D}$ with $\mathbb{E}_{s\sim\mathcal{D}}[s] = 0$

Sample $\tilde{a}_t = a_t + s_t$

Observe $r_t(\tilde{a}_t)$

Define $g_t = r_t(\tilde{a}_t) H_t^{-1} s_t$ where $H_t \coloneqq \mathbb{E}_{s\sim\mathcal{D}}[ss^\top]$

Gradient Ascent with Gradient Estimator

Assume the feasible set Ω contains a ball of radius δ

Define $\Omega' = \{a \in \Omega\text{:}\ \ \mathcal{B}(a, \delta) \subset \Omega\}$

Arbitrarily pick $a_1 \in \Omega'$

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

Let $\tilde{a}_t = a_t + s_t$ where $s_t \sim \mathcal{D}$ (assume that $\|s_t\| \leq \delta$ always holds)

Receive $r_t(\tilde{a}_t)$

Define

$$g_t = (r_t(\tilde{a}_t) - b_t)H_t^{-1}s_t \qquad \text{where } H_t \coloneqq \mathbb{E}_{s\sim\mathcal{D}}[ss^\top]$$

Update policy:

$$a_{t+1} = \Pi_{\Omega'}\left(a_t + \eta g_t\right)$$

Continuous Contextual Bandits

With a regression oracle

	MAB	CB
VB		•
PB		

Combining with Regression Oracle (a bandit version of DDPG)

Assume policy parametrization $\pi_\theta(\cdot\,|x) = \mathcal{N}(\mu_\theta(x), \sigma^2)$

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

Receive context x_t

Take action $a_t = \mathcal{P}_\Omega(\mu_\theta(x_t) + \sigma\,\mathcal{N}(0, I))$

Receive $r_t(x_t, a_t)$

Update the mean estimator:

$$\phi \leftarrow \phi - \lambda \nabla_\phi \left[\left(R_\phi(x_t, a_t) - r_t(x_t, a_t)\right)^2\right]$$

Update policy:

$$\theta \leftarrow \theta + \eta \nabla_\theta R_\phi(\mu_\theta(x_t))$$

Continuous Contextual Bandits

Pure policy-based algorithms

	MAB	CB
VB		
PB		•