Online Learning: Bandit Setting

Daniel Khashabi

Summer 2014
Last Update: October 20, 2016

1 Introduction

[TODO]

2 Bandits

2.1 Stochastic setting

Suppose there exists unknown distributions $\nu_1, \ldots, \nu_K$, such that the loss at each iteration is chosen as $l_{i,t} \sim \nu_i$. [1] Therefore the mean for each of these distributions can be represented as $\mu_k = \mathbb{E}[l_{k,t}]$. Denote the least expected loss with $\mu^* = \min_{k \in \{1,\ldots,K\}} \mu_k$. Define $\tau_i(t)$ to be the number of times arm i has been pulled. More formally:

$$\tau_i(t) = \sum_{s=1}^{t} \mathbf{1}\left\{I_s = i\right\}$$

Lemma 1. *The (pseudo) regret can be written as*

$$\bar{R}(T) = \sum_{k=1}^{K} \Delta_k \mathbb{E}\left[\tau_i(T)\right]$$

where Δ_k is the difference between the mean loss of the action chosen and the action with minimum loss: $\Delta_k = \mu_k - \mu^*$.

[1] The stationarity assumption is implicit here; the distributions are not changed across time horizon.

Proof.

$$\begin{aligned}
\bar{R}(T) = \mathbb{E}\left[\sum_{i=1}^{T} l_{I_t,t}\right] - \min_{k\in\{1,\dots,K\}} \mathbb{E}\left[\sum_{i=1}^{T} l_{k,t}\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[l_{I_t,t}|I_t\right]\right] - T\mu^* \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mu_{I_t} - \mu^*\right] \\
&= \sum_{t=1}^{T} \sum_{k\in\{1,\dots,K\}} (\mu_k - \mu^*)\,\mathbb{P}(I_t = k) \\
&= \sum_{k\in\{1,\dots,K\}} \Delta_k \sum_{t=1}^{T} \mathbb{P}(I_t = k) \\
&= \sum_{k\in\{1,\dots,K\}} \Delta_k \mathbb{E}\left[\tau_k(T)\right]
\end{aligned}$$

□

Therefore the only thing we need to worry about is $\tau_k(T)$, the number of times each armed is pulled.

Exploration first: Consider a simple strategy: sample each arm for C many times (in any order), then start decision making decisions. Denote the empirical estimate mean for each action with $\hat{\mu}_k$.

Here is the suggested algorithm:

1. For a fixed probability $\delta \in (0, 1]$, sample each arm for C times, and computer their empirical mean $\hat{\mu_k}$
2. For the rest of the $T - KC$ remaining iterations, do the action which has the minimum loss:

$$\hat{k} = \arg\min_{k\in\{1,\dots,K\}} \hat{\mu}_{k,t}$$

Using the Hoeffding bound [2] we know that:

$$|\hat{\mu}_k - \mu_k| < \sqrt{\frac{\log(2/\delta)}{2C}},\ \forall k \in \{1, ..., K\}$$

which means that, the bigger the size of samples C are, the better our estimate of the means are, as expected. Define $\Delta = \min\{\Delta_i : \Delta_i > 0\}$ which is the minimum difference between the true action means. In order to guarantee that our we always choose the correct action with mean μ^*, we need to make sure that our estimates satisfy $|\hat{\mu}_k - \mu_k| < \Delta/2$:

$$\sqrt{\frac{\log(2/\delta)}{2C}} < \Delta/2 \Rightarrow C > \frac{2\ln(2/\delta)}{2\Delta^2}$$

The regret incurred for the first KC iterations of sampling actions is

[2] See `http://web.engr.illinois.edu/~khashab2/learn/concentration.pdf`

Optimisim in the face of uncertainty (α-UCB) Define our estimate of each mean until time t to be $\hat{\mu}_{k,t} = \frac{1}{\tau_k(t)} \sum_{s=1}^{t} l_{k,s} \mathbf{1}\{I_s = k\}$. Define the upper-confidence bound on action (arm) k at time t to be $U_{k,t} = \hat{\mu}_{k,t} + \sqrt{\frac{\alpha \log(t)}{2\tau_k(t)}}$, where $\alpha > 0$ is a parameter which controls the upper-bound estimates and we will set it later. At each iteration choose the actions to be $I_t = \arg\max_{k \in \{1,\dots,K\}} U_{k,t-1}$. With this choice we know that

$$\mathbb{E}[\Delta_{I_t} | I_t] = \mu^* - \mu_{I_t} \leq^{w.h.p.} U_{k^*,t-1} - \mu_{I_t} \leq \max_{k \in \{1,\dots,K\}} U_{k,t-1} - \mu_{I_t}$$

Lemma 2. *If $I_t = i \neq i^*$ (incorrect action) then $U_{i,t} \geq U_{i^*,t}$. The mistake might be at least due to one of the following events.*

1. $A_1(t) = \{U_{i^*,t} \leq \mu^*\}$*: the upper-bound estimate on the true action is too small.*
2. $A_2(t) = \{U_{i,t} \geq \mu^* (= \mu_i + \Delta_i)\}$*: the upper-bound estimate on action i is too big.*
3. $A_3(t) = \{\tau_i(t) \leq \frac{2\alpha \ln T}{\Delta_i^2}\}$*: number of samples from action i is too small.*

Proof. To prove it, we can use proof by contradiction. Suppose all of the above are false; we will show that essentially $I_t = i^*$:

$$\begin{aligned}
U_{i^*,t} &> \mu^* && A_1 \text{ is false} \\
&= \mu_i + \Delta_i \\
&> \mu_i + \sqrt{\frac{2\alpha \ln T}{\tau_i(t)}} && A_3 \text{ is false} \\
&\geq \mu_i + \sqrt{\frac{2\alpha \ln t}{\tau_i(t)}} \\
&\geq \hat{\mu}_{i,t} + \sqrt{\frac{\alpha \ln t}{2\tau_i(t)}} && A_2 \text{ is false}
\end{aligned}$$

□

Lemma 3. $\mathbb{P}(A_1(t)) \leq t^{1-\alpha}$ *and* $\mathbb{P}(A_2(t)) \leq t^{1-\alpha}$

Proof. Again we use the Hoeffding bound:

$$\mathbb{P}\left(\hat{\mu}_{k^*,t} - \mu^* \leq \sqrt{\frac{\alpha \ln t}{2s}}\right) \leq t^{-\alpha}$$

where s is the number of times the action i is sampled. Now we can show that

$$\begin{aligned}
\mathbb{P}(A_1(t)) = \mathbb{P}\left(\hat{\mu}_{k^*,t} - \mu^* \leq \sqrt{\frac{\alpha \ln t}{2\tau_k(t)}}\right) &\leq \mathbb{P}\left(\bigcup_{s=1}^{t} \left\{\hat{\mu}_{k^*,t} - \mu^* \leq \sqrt{\frac{\alpha \ln t}{2s}}\right\}\right) \\
&\leq \sum_{s=1}^{t} \mathbb{P}\left(\hat{\mu}_{k^*,t} - \mu^* \leq \sqrt{\frac{\alpha \ln t}{2s}}\right) \\
&\leq \sum_{s=1}^{t} t^{-\alpha} = t^{1-\alpha}
\end{aligned}$$

The claim $\mathbb{P}(A_2(t)) \leq t^{1-\alpha}$ can be proved in a similar way. □

Lemma 4.

$$\mathbb{E}[\tau_k(T)] \leq \frac{2\alpha \log T}{\Delta_i^2} + \frac{2}{\alpha - 2}, \quad \forall k \neq k^*$$

Proof.

$$\begin{aligned}
\mathbb{E}[\tau_k(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{I_t = k\}\right] = \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{I_t = k, \tau_k(t-1) \leq t_0\} + \sum_{t=1}^{T} \mathbf{1}\{I_t = k, \tau_k(t-1) > t_0\}\right] = \\
&\leq t_0 + \sum_{t=t_0+1}^{T} \mathbb{P}\{I_t = k, \tau_k(t-1) > t_0\} =
\end{aligned}$$

Now define $t_0 = \frac{2\alpha \ln T}{\Delta_i^2}$. Therefore the event A_3 of lemma 2 has been satisfied. We can further simplify the previous equation:

$$\begin{aligned}
\mathbb{E}[\tau_k(T)] &\leq t_0 + \sum_{t=t_0+1}^{T} \mathbb{P}\{I_t = k, \tau_k(t-1) > t_0\} \\
&= t_0 + \sum_{t=t_0+1}^{T} \mathbb{P}\{A_1(t) \vee A_2(t)\} \\
&\leq t_0 + \sum_{t=t_0+1}^{T} [\mathbb{P}\{A_1(t)\} + \mathbb{P}\{A_2(t)\}] \\
&\leq t_0 + 2\sum_{t=t_0+1}^{T} t^{1-\alpha} \leq t_0 + 2\sum_{t=1}^{\infty} t^{1-\alpha} \leq t_0 + \frac{2}{\alpha - 2}
\end{aligned}$$

The last inequality comes from the fact that

$$\int_{t=1}^{\infty} t^{1-\alpha} dt = \frac{1}{\alpha - 2}$$

□

Using the results of Lemma 4 into Lemma 1 we would get:

$$\bar{R}(T) = \sum_{k=1}^{K} \Delta_k \left(\frac{2\alpha \log T}{\Delta_k^2} + \frac{2}{\alpha - 2}\right) = \sum_{k=1}^{K} \left(\frac{2\alpha}{\Delta_k}\right) \log T + \frac{2}{\alpha - 2} \sum_{k=1}^{K} \Delta_k$$

2.2 Adversarial setting

Input: decay parameters $\{\eta_t\}_{t=1}^T$.
Initialize: Uniform distribution $\mathbf{p}_1 = [p_{1,1}, \ldots, p_{1,K}]$ over the set $\{1, \ldots, K\}$.
For $t = 1, \ldots, T$:

- Draw an arm I_t based on probability distribution $\mathbf{p}_t$.
- Create the loss value $\hat{l}_{i,t}$, based on $l_{i,t}$ and $\mathbf{p}_t$.
- Update the commulative loss $\hat{L}_{i,t} = \sum_{s=1}^t \hat{l}_{i,s}$.
- Update the probability distribution over actions:

$$p_{i,t+1} = \frac{\exp\left(-\eta_t \hat{L}_{i,t}\right)}{\sum_{k=1}^K \exp\left(-\eta_t \hat{L}_{k,t}\right)}, \quad \text{for each } 1 \le i \le K$$

Lemma 5. *For any sequence of actions in Algorithm* **??**, *with non-increasing positive sequence* $\eta_1, \eta_2, \ldots, \eta_T$ *we have:*

$$\sum_{t=1}^T \sum_{k=1}^K p_{k,t}\hat{l}_{k,t} - \min_{h\in\{1,\ldots,K\}} \left(\sum_{t=1}^T p_{h,t}\hat{l}_{h,t}\right) \le \sum_{t=1}^T \frac{\eta_t}{2} \sum_{k=1}^K p_{k,t}\left(\hat{l}_{k,t}\right)^2 + \frac{\ln K}{\eta_T}$$

Proof. We prove the inequality for any decisions h and ignore the "minimum". Therefore the left side is: $\sum_{t=1}^T \sum_{k=1}^K p_{k,t}\hat{l}_{k,t} - \left(\sum_{t=1}^T p_{h,t}\hat{l}_{h,t}\right)$. The log-moment of the

$$\sum_{k=1}^K p_{k,t}\hat{l}_{k,t} = \mathbb{E}_{k\sim\mathbf{p}_t}\hat{l}_{k,t} = \mathbb{E}_{k\sim\mathbf{p}_t}\hat{l}_{k,t} + \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right) - \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right) \quad (1)$$

redundant

$$\begin{aligned}
\sum_{k=1}^K p_{k,t}\hat{l}_{k,t} &= \mathbb{E}_{k\sim\mathbf{p}_t}\hat{l}_{k,t} \\
&= \mathbb{E}_{k\sim\mathbf{p}_t}\hat{l}_{k,t} + \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right) - \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right) \\
&= \frac{1}{\eta_t}\ln\exp \mathbb{E}_{k\sim\mathbf{p}_t}\eta_t\hat{l}_{k,t} + \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right) - \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right) \\
&= \frac{1}{\eta_t}\ln\left(\exp \mathbb{E}_{k\sim\mathbf{p}_t}\eta_t\hat{l}_{k,t} \times \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right)\right) - \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right) \\
&= \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\left(\exp \mathbb{E}_{k\sim\mathbf{p}_t}\eta_t\hat{l}_{k,t} \times \exp\left(-\eta_t\hat{l}_{h,t}\right)\right) - \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right) \\
&= \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\left(\hat{l}_{h,t} - \mathbb{E}_{k\sim\mathbf{p}_t}\hat{l}_{k,t}\right)\right) - \frac{1}{\eta_t}\ln \mathbb{E}_{h\sim\mathbf{p}_t}\exp\left(-\eta_t\hat{l}_{h,t}\right)
\end{aligned}$$

Now we simplify left two terms in Equation 1. In the following, we use the two inequalities $\ln x \leq x - 1$ and $\exp(-x) - 1 + x \leq x^2/2$ each once:

$$\begin{aligned}
\frac{1}{\eta_t} \ln \mathbb{E}_{h\sim\mathbf{p}_t} \exp\left(-\eta_t \hat{l}_{h,t}\right) + \mathbb{E}_{k\sim\mathbf{p}_t} \hat{l}_{k,t} &\leq \frac{1}{\eta_t}\left(\mathbb{E}_{h\sim\mathbf{p}_t} \exp\left(-\eta_t \hat{l}_{h,t}\right) - 1\right) + \mathbb{E}_{k\sim\mathbf{p}_t} \hat{l}_{k,t} \\
&= \frac{1}{\eta_t} \mathbb{E}_{h\sim\mathbf{p}_t}\left(\exp\left(-\eta_t \hat{l}_{h,t}\right) - 1 + \eta_t \hat{l}_{h,t}\right) \\
&\leq \frac{1}{\eta_t} \frac{\eta_t^2}{2} \mathbb{E}_{h\sim\mathbf{p}_t} \hat{l}_{h,t}^2 = \frac{\eta_t}{2} \mathbb{E}_{h\sim\mathbf{p}_t} \hat{l}_{h,t}^2
\end{aligned}$$

Define the shorthand notation $\Phi_t(\eta) = \frac{1}{\eta} \ln \frac{1}{K} \sum_{h=1}^{K} \exp\left(-\eta \hat{L}_{h,t}\right)$.

$$\begin{aligned}
\frac{1}{\eta_t} \ln \mathbb{E}_{h\sim\mathbf{p}_t} \exp\left(-\eta_t \hat{l}_{h,t}\right) &= \frac{1}{\eta_t} \ln \sum_{h=1}^{K} p_{h,t} \exp\left(-\eta_t \hat{l}_{h,t}\right) \\
&= \frac{1}{\eta_t} \ln \sum_{h=1}^{K} \frac{\exp\left(-\eta_t \hat{L}_{h,t-1}\right)}{\sum_{k=1}^{K} \exp\left(-\eta_t \hat{L}_{k,t-1}\right)} \exp\left(-\eta_t\left(\hat{L}_{h,t} - \hat{L}_{h,t-1}\right)\right) \\
&= \frac{1}{\eta_t} \ln \frac{\sum_{h=1}^{K} \exp\left(-\eta_t \hat{L}_{h,t}\right)}{\sum_{k=1}^{K} \exp\left(-\eta_t \hat{L}_{k,t-1}\right)} = \Phi_t(\eta_t) - \Phi_{t-1}(\eta_t)
\end{aligned}$$

Summing the time index we have:

$$\sum_{t=1}^{T} \sum_{k=1}^{K} p_{k,t} \hat{l}_{k,t} - \sum_{t=1}^{T} p_{I_t,t} \hat{l}_{I_t,t} \leq \sum_{t=1}^{T} \frac{\eta_t}{2} \mathbb{E}_{h\sim\mathbf{p}_t} \hat{l}_{h,t}^2 + \sum_{t=1}^{T} \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t) - \sum_{t=1}^{T} \mathbb{E}_{I_t\sim\mathbf{p}_t} \hat{l}_{k,t}$$

$$\sum_{t=1}^{T} \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t) = \sum_{t=1}^{T-1} \left(\Phi_t\left(\eta_{t+1}\right) - \Phi_t\left(\eta_t\right)\right) - \Phi_T(\eta_T)$$

Note that $\Phi_0(\eta_1) = 0$. Also,

$$\begin{aligned}
-\Phi_T(\eta_T) &= -\frac{1}{\eta_T} \ln \frac{1}{K} \sum_{h=1}^{K} \exp\left(-\eta_T \hat{L}_{h,T}\right) \\
&= \frac{\ln K}{\eta_T} - \frac{1}{\eta_T} \ln \sum_{h=1}^{K} \exp\left(-\eta_T \hat{L}_{h,T}\right) \\
&\leq \frac{\ln K}{\eta_T} - \frac{1}{\eta_T} \ln \exp\left(-\eta_T \hat{L}_{h,T}\right) \\
&= \frac{\ln K}{\eta_T} - \sum_{t=1}^{T} \hat{l}_{h,t}
\end{aligned}$$

Then the summation becomes:

$$\sum_{t=1}^{T} \sum_{k=1}^{K} p_{k,t} \hat{l}_{k,t} - \sum_{t=1}^{T} p_{I_t,t} \hat{l}_{I_t,t} \leq \sum_{t=1}^{T} \frac{\eta_t}{2} \mathbb{E}_{h\sim\mathbf{p}_t} \hat{l}_{h,t}^2 + \frac{\ln K}{\eta_T} + \sum_{t=1}^{T-1} \left(\Phi_t\left(\eta_{t+1}\right) - \Phi_t\left(\eta_t\right)\right)$$

We can show that $\Phi'_t(\eta) \geq 0$. Since we assumed that for any t $\eta_{t+1} \leq \eta_t$. Therefore $\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t) \leq 0$.

$$\sum_{t=1}^{T}\sum_{k=1}^{K} p_{k,t}\hat{l}_{k,t} - \sum_{t=1}^{T} p_{I_t,t}\hat{l}_{I_t,t} \leq \sum_{t=1}^{T} \frac{\eta_t}{2}\mathbb{E}_{h\sim\mathbf{p}_t}\hat{l}^2_{h,t} + \frac{\ln K}{\eta_T}$$

which is the desired result. □

Corollary 1. *For any sequence of actions in Algorithm* **??**, *with non-increasing positive sequence* $\eta_1, \eta_2, \ldots, \eta_T$ *we have:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} p_{k,t}\hat{l}_{k,t}\right] - \min_{h\in\{1,\ldots,K\}}\left(\mathbb{E}\left[\sum_{t=1}^{T} p_{h,t}\hat{l}_{h,t}\right]\right) \leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta_t}{2}\sum_{k=1}^{K} p_{k,t}\left(\hat{l}_{k,t}\right)^2\right] + \frac{\ln K}{\eta_T}$$

Proof. Take expectation from both sides and use the fact that $\mathbb{E}\left[\min\left[.\right]\right] \leq \min\left[\mathbb{E}\left[.\right]\right]$. □

2.2.1 The EXP3 algorithm

If the second step in the Algorithm **??** is $\hat{l}_{i,t} = l_{i,t}\mathbf{1}\left\{I_t = i\right\}/p_{t,i}$.

Lemma 6. *For the EXP3 algorithm the expected regret is bounded by*

$$\frac{K}{2}\sum_{t=1}^{T}\eta_t + \frac{\ln K}{\eta_T}$$

For proper choice of $\{\eta_t\}_{t=1}^{T}$ *the overall bound is*

$$\sqrt{2nK\ln K}$$

Proof.

$$\begin{aligned}
\sum_{t=1}^{T}\frac{\eta_t}{2}\sum_{k=1}^{K} p_{k,t}\left(\hat{l}_{k,t}\right)^2 &= \sum_{t=1}^{T}\frac{\eta_t}{2}\sum_{k=1}^{K} p_{k,t}\left(l_{k,t}/p_{k,t}\right)^2\mathbf{1}\left\{I_t = k\right\} \\
&= \sum_{t=1}^{T}\frac{\eta_t}{2} p_{I_t,t}\left(l_{I_t,t}/p_{I_t,t}\right)^2 \\
&= \sum_{t=1}^{T}\frac{\eta_t}{2} l^2_{I_t,t}/p_{I_t,t}
\end{aligned}$$

Since the decisions I_t are made in stochastic fashion we need to find the expectation with respect to I_t.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta_t}{2}\sum_{k=1}^{K}p_{k,t}\left(\hat{l}_{k,t}\right)^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta_t}{2}l_{I_t,t}^2/p_{I_t,t}\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\frac{\eta_t}{2}l_{I_t,t}^2/p_{I_t,t}\right] \\
&\leq \sum_{t=1}^{T}\mathbb{E}\left[\frac{\eta_t}{2}.1/p_{I_t,t}\right] \\
&= \frac{K}{2}\sum_{t=1}^{T}\eta_t
\end{aligned}$$

Which would give the general form of the bound for EXP3. If we set $\eta_t = \sqrt{\frac{2\ln K}{TK}}$ we would get the result.

$$\begin{aligned}
\frac{K}{2}\sum_{t=1}^{T}\eta_t + \frac{\ln K}{\eta_T} &= \frac{K}{2}\times T\sqrt{\frac{2\ln K}{TK}} + \ln K \times \sqrt{\frac{TK}{2\ln K}} \\
&= \sqrt{\frac{TK\ln K}{2}} + \sqrt{\frac{TK\ln K}{2}} \\
&= \sqrt{2TK\ln K}
\end{aligned}$$

□

2.3 Lower bounds

2.3.1 Preliminaries

The KL divergence has ??? property.

$$\begin{aligned}\mathrm{KL}\left(p(x,y)||q(x,y)\right) &= \mathrm{KL}\left(p(x)||q(x)\right) + \mathrm{KL}\left(p(y|x)||q(y|x)\right) \\ &= \mathrm{KL}\left(p(x)||q(x)\right) + \sum_x p(x)\mathrm{KL}\left(p(y|x)||q(y|x)\right)\end{aligned}$$

The Pinsker's inequality creates connection between the KL divergence and the total variations divergence:

$$\sup_x |p(x) - q(x)| \leq \sqrt{\frac{1}{2}\mathrm{KL}\left(p||q\right)}$$

2.3.2 Lower bounding ...

Theorem 1. *Suppose $Y_{i,1}, Y_{i,1}, \ldots$ the i.i.d. sequence of costs. We want to find a lower bound on the regret. The lower bound needs to hold for any distribution of rewards (specifically the worst case of the distributions, thus inf with respect to the reward distributions). It also needs to hold to the best forecaster one can design (thus sup with respect to forecasters).*

$$\inf \sup \left(\mathbb{E} \sum_{t=1}^{T} Y_{i,t} - \min_{i \in \{0,\ldots,K\}} \mathbb{E} \sum_{t=1}^{T} Y_{i,t} \right) \geq \frac{1}{20}\sqrt{nK}$$

Proof. The idea of the proof is to analyze the behavior of any forecaster against two distributions that differ slightly: (1) in one all of the distributions are $1/2$. (2) in the other all of the arm distributions are $1/2$ except one which is $1/2 + \epsilon$.

□

Lemma 7.

$$\inf \sup \left(\mathbb{E} \sum_{t=1}^{T} Y_{i,t} - \min_{i \in \{0,\ldots,K\}} \mathbb{E} \sum_{t=1}^{T} Y_{i,t} \right) \geq T\epsilon \left(1 - \frac{1}{K} - \sqrt{\epsilon \ln \frac{1+\epsilon}{1-\epsilon}} \sqrt{\frac{T}{2K}} \right)$$

Proof. Define the loss $l_{h,t}$ representing the loss value at time t for action h. We choose action $h \in \{1, \ldots, K\}$. Define $K+1$ different games. In each of the games the distribution of losses are different. For the i-th game $1 \geq i \geq K$, all of the loss values are iid random variables distributed with Bernoulli of bias $\frac{1-\epsilon}{2}$, except the h-th arm, which is distributed with Bernoulli distribution of bias $\frac{1-\epsilon}{2}$. Also define an additional game in which all of the losses have Bernoulli distribution with bias $\frac{1-\epsilon}{2}$.

Suppose I_t is the arm played by the algorithm at time t. Denote the empirical distribution over actions up to time t with $q_t = (q_{1,t}, \ldots, q_{K,t})$:

$$q_{k,t} = \frac{1}{t} \sum_{t'=1}^{t} \mathbf{1}\left\{I_{t'} = k\right\}$$

Let J be a random variable distributed according to q_t. Define $\mathbb{P}_h$ to be the law of J, when the forecaster plays the h-th game, and we know:

$$\mathbb{P}_h\left(J=h\right)=\mathbb{E}_h\left[\frac{1}{t}\sum_{t'=1}^{t}\mathbf{1}\left\{I_{t'}=k\right\}\right]$$

where $\mathbb{E}_h[.]$ means expectation with respect to the distribution of the h-th game. The regret for the h-th game is:

$$R(T)=\mathbb{E}_h\left[\sum_{t=1}^{T}\left(l_{h,t}-l_{I_t,t}\right)\right]$$

The regret can be simplified in the following form:

$$\mathbb{E}_h\left[\sum_{t=1}^{T}\left(l_{h,t}-l_{I_t,t}\right)\right]=\epsilon T\sum_{h'\neq h}\mathbb{P}_h(J=h')$$

Note that $\mathbb{P}(l_{h,t}=1 \text{ and } l_{I_t,t}=0)(1-0)+\mathbb{P}(l_{h,t}=0 \text{ and } l_{I_t,t}=1)(0-1)=\frac{1+\epsilon}{2}-\frac{1-\epsilon}{2}=\epsilon$. This part needs modification and more clear explanation. which can be written as

$$\epsilon T\left(1-\mathbb{P}_h(J=h)\right)$$

Averaging over all of the games we have:

$$\frac{1}{K}\sum_h\mathbb{E}_h\left[\sum_{t=1}^{T}\left(l_{h,t}-l_{I_t,t}\right)\right]=\epsilon T\left(1-\frac{1}{K}\sum_h\mathbb{P}_h(J=h)\right)$$

Note that we want a lower bound on max (not average). But since average is less than max, a good lower bound on average would also work for us. By the Pinsker's inequality we have:

$$\mathbb{P}_h(J=h)\leq\mathbb{P}_\emptyset+\sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_\emptyset||\mathbb{P}_h)}$$

and hence

$$\frac{1}{K}\sum_h\mathbb{P}_h(J=h)\leq\frac{1}{K}+\sqrt{\frac{1}{2N}\sum_h\mathrm{KL}(\mathbb{P}_\emptyset||\mathbb{P}_h)}$$

Note that in the last step we used the fact that the squared root function is concave and $\sum_h\mathbb{P}_\emptyset(J=h)=1$.

The next step is to establish the distance measure between the probability distributions of losses for different games.

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_\emptyset^T||\mathbb{P}_h^T)&=\mathrm{KL}(\mathbb{P}_\emptyset^0||\mathbb{P}_h^0)+\sum_{t=2}^{T}\sum_{y^{t-1}}\mathbb{P}_\emptyset^{t-1}(y^{t-1})\mathrm{KL}\left(\mathbb{P}_\emptyset^t(.|y^{t-1})||\mathbb{P}_h^t(.|y^{t-1})\right)\\
&=\mathrm{KL}(\mathbb{P}_\emptyset^0||\mathbb{P}_h^0)+\sum_{t=2}^{T}\left(\sum_{y^{t-1};I_t=i}\mathbb{P}_\emptyset^{t-1}(y^{t-1})\mathrm{KL}\left(\frac{1-\epsilon}{2}||\frac{1+\epsilon}{2}\right)+\sum_{y^{t-1};I_t\neq i}\mathbb{P}_\emptyset^{t-1}(y^{t-1})\mathrm{KL}\left(\frac{1+\epsilon}{2}||\frac{1+\epsilon}{2}\right)\right)\\
&=\mathrm{KL}\left(\frac{1-\epsilon}{2}||\frac{1+\epsilon}{2}\right)\mathbb{E}_\emptyset\left[\sum_{t=1}^{T}\mathbf{1}\left\{I_t=h\right\}\right]
\end{aligned}$$

□

$$\frac{1}{K}\sum_h \sqrt{\mathrm{KL}(\mathbb{P}_\emptyset||\mathbb{P}_h)} \le \sqrt{\frac{1}{K}\sum_h \mathrm{KL}(\mathbb{P}_\emptyset||\mathbb{P}_h)}$$
$$\le \sqrt{\frac{1}{K}\sum_h \mathrm{KL}(\frac{1-\epsilon}{2}||\frac{1+\epsilon}{2})\mathbb{E}_\emptyset\left[\sum_{t=1}^{T}\mathbf{1}\left\{I_t=h\right\}\right]}$$
$$\le \sqrt{\frac{T}{K}\mathrm{KL}(\frac{1-\epsilon}{2}||\frac{1+\epsilon}{2})}$$

This last step I am confused why?

We know

$$\mathrm{KL}(\frac{1-\epsilon}{2}||\frac{1+\epsilon}{2}) = \epsilon \ln\frac{1+\epsilon}{1-\epsilon}$$

then

$$\frac{1}{K}\sum_h \sqrt{\mathrm{KL}(\mathbb{P}_\emptyset||\mathbb{P}_h)} \le \sqrt{\frac{T}{K}\epsilon\ln\frac{1+\epsilon}{1-\epsilon}}$$

So far the lower bound is the following: $\sup R(T) \ge \epsilon T\left(1-\frac{1}{K}-\sqrt{\frac{T}{K}\epsilon\ln\frac{1+\epsilon}{1-\epsilon}}\right)$ The final step is the ϵ-tuning of the bound. Since the lower bound holds for any $\epsilon \le 1/2$ we choose it in a way that it attains its biggest value. If we set $\epsilon = \alpha\sqrt{\frac{N}{MT}}$, where α is a real number to be tuned, this would give us the desired result.

http://cseweb.ucsd.edu/~kamalika/teaching/CSE291W11/lecture5.pdf http://courses.cs.washington.edu/courses/cse599s/12sp/scribes.html Lower bounds: http://www.stat.berkeley.edu/~bartlett/courses/2014fall-cs294stat260/lectures/bandit-lower-bound-notes.pdf http://www.ece.iisc.ernet.in/~aditya/E1245_Online_Prediction_Learning_F2014/Lecture_14_Scribe_Notes.pdf http://www.ece.iisc.ernet.in/~aditya/E1245_F14.html —¿ So good! This contains a very nice comparison between UCB and EXP3: http://www.levreyzin.com/presentations/CMU_bandits.pdf

3 Contextual Bandits

In contextual bandit unlike the standard bandits, the importance of actions are dependent on the "context" on which they are being done. In other words whether a single action is optimal or not depends on its context. A simple example is to consider two contexts "weekday" and "weekend". An action which might be optimal during "weekend" is not necessarily the best action for the "weekday".

Just like the standard bandits, in the contextual bandit problem, on each of T rounds a learner is presented with the choice of taking one of K actions. Before making the choice of action, the learner observes a feature vector (context) associated with each of its possible choices. In this setting the learner has access to a hypothesis class, in which the hypotheses receives action features (context) and predict which action will give the best reward. If the learner can guarantee to do nearly as well as the prediction of the best hypothesis in hindsight (to have low regret), the learner is said to successfully compete with that class.

Algorithm	Regret	High probability bound	Contextual	Efficient
Exp3.P	$O(T^{1/2})$	Y	N	Y
UCB	$O(T^{1/2})$	Y	N	Y
Exp4	$O(\sqrt{TK \ln N})$	N	Y	N
Epoch-greedy	$O(T^{2/3})$	Y	Y	Y
LinUCB [4]	$O()$			
Exp4.P	$O(T^{1/2})$	Y	Y	N

Table 1: Properties of popular bandit algorithms; N experts, T number of rounds, K number of possible actions.

If we ignore the contextual information we can just use the existing vanilla bandit algorithms. Therefore having the contextual information one should be able to get better guarantees. One way of looking at the contextual bandits is to think of it as means to connect to the supervised learned, which requires input features supplied by users for making predictions. An important point here is that the bandit problems are not supervised learning problems; for example in a click-or-not on one ad does not generally tell you if a different ad would have been clicked on. Instead this problem is inherently exploration-exploitation problem. That said, the solution to the contextual bandit should be intuitive and reasonable from supervised learning setting; in fact some of the well-established supervised learning techniques will come handy in analysis of contextual bandits.

Here is an approach which tries to adapt the existing bandit algorithms as blackbox. Suppose the size of the context space is bounded and is small. Run a different k-Armed Bandit for every value of context vector. The regret and amount of information required to do well scales linearly in the number of contexts. This approach is a little counter-intuitive; good supervised learning algorithms often require information which is (essentially) independent of the number of contexts (instead they depend on the complexity of the concept class define on top of the features/contexts).

One can get inspiration from supervised learning. Define a policy space $\mathcal{H}$ from which policies are chosen, and treat every policy $h(x) \in \mathcal{H}$ as a different arm. This removes an explicit dependence on the number of contexts, but it creates a linear dependence on the number of policies. Via Occams razor/VC dimension/Margin bounds, we already know that supervised learning requires experience much smaller than the number of policies.

The name *contextual bandit* is borrowed from Langford and Zhang [3] but it has been known under other names as well; e.g. *bandit problems with expert advice* [1], *associative reinforcement learning* [2].

Bibliographical notes

References

[1] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Research*, 3:397–422, 2003.

[2] Andrew G Barto and P Anandan. Pattern-recognizing stochastic learning automata. *Systems, Man and Cybernetics, IEEE Transactions on*, (3):360–375, 1985.

[3] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Adv. Neural Info. Proc. Sys. 21* (NIPS), pages 817–824, 2008.

[4] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. pages 661–670. ACM, 2010.