**EECS 598-005: Theoretical Foundations of Machine Learning** **Fall 2015**

# Lecture 22: Adversarial Multi-Armed Bandits

*Lecturer: Jacob Abernethy* *Scribes: Arvind Prasadan*

## 22.1 The EXP3 Algorithm[1]

EXP3 was invented in 2001 by Auer, Cesa-Bianchi, Freund, and Schapire [ACBFS02] to handle the non-stochastic, adversarial multi-arm bandit problem. The EXP3 algorithm has an expected regret bound of $\sqrt{2Tn\log n}$. In this lecture, we state the algorithm and derive this regret bound.

### 22.1.1 Algorithm

Let $\underline{\widetilde{L}}^t$ be the cumulative losses up to period $t$. To be precise, define $\underline{\widetilde{L}}^t = \sum_{k=1}^t \underline{\widetilde{l}}^t$, where $\underline{\widetilde{l}}^t$ is defined in the algorithm description below.

**for** t = 1, 2, $\cdots$, T-1, T **do**
  Sample $I_t \sim \underline{p}^t$
  Observe $l^t_{I_t}$
  Set $\underline{\widetilde{l}}^t = \left\langle 0, ..., 0, \dfrac{l^t_{I_t}}{p^t_{I_t}}, 0, ..., 0 \right\rangle$
  Set $\underline{\widetilde{L}}^t = \underline{\widetilde{L}}^{t-1} + \underline{\widetilde{l}}^t$
  **for** i = 1, 2, $\cdots$, n-1, n **do**
    Set $p_i^{t+1} = \dfrac{e^{-\eta \widetilde{L}_i^t}}{\sum\limits_{j=1}^n e^{-\eta \widetilde{L}_i^t}}$
  **end for**
**end for**

### 22.1.2 EXP3: Expected Regret

There are two facts that enable the following analysis. First, note that $\mathbb{E}_{i\sim p^t}\left[\underline{\widetilde{l}^t}\right] = \underline{l}^t$, so that $\mathbb{E}_{i\sim p^t}\left[\underline{\widetilde{L}^t}\right] = \underline{L}^t$. Moreover, $\underline{\widetilde{l}^t}$ and $p^t$ are uncorrelated.

We analyze the regret of EXP3 by looking at the potential function

$$\Phi_t = -\frac{1}{\eta}\log\left(\sum_{i=1}^n e^{-\eta \widetilde{L}_i^{t-1}}\right)$$

and taking the *expected* increase in potential across iterations.

The increase in potential from iteration $t$ to $t+1$ is

$$\Phi_{t+1} - \Phi_t = -\frac{1}{\eta}\log\left(\frac{\sum_{i=1}^n e^{-\eta \widetilde{L}_i^t}}{\sum_{i=1}^n e^{-\eta \widetilde{L}_i^{t-1}}}\right) = -\frac{1}{\eta}\log\left(\frac{\sum_{i=1}^n e^{-\eta \widetilde{L}_i^{t-1} - \eta \widetilde{l}_i^t}}{\sum_{i=1}^n e^{-\eta \widetilde{L}_i^{t-1}}}\right) = -\frac{1}{\eta}\log\left(\mathbb{E}_{i\sim p^t}\left[e^{-\eta \widetilde{l}_i^t}\right]\right)$$

[1]Credits: The following section is taken in part from Lecture 20 of EECS 598 in 2013 (Prediction and Learning: It's Only a Game): these notes were scribed by Zhihao Chen. The handwritten notes of Anthony Della Pella and Vikas Dhiman were instrumental in the creation of this document.

To proceed, we need the following fact:

**Lemma 22.1.** *For all $x \geq 0$,*

$$e^{-x} \leq 1 - x + \frac{1}{2}x^2$$

Using the fact, we see that

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &\geq -\frac{1}{\eta}\log\left(\mathbb{E}_{i\sim p^t}\left[1 - \eta\widetilde{l}_i^t + \frac{1}{2}\eta^2(\widetilde{l}_i^t)^2\right]\right) \\
&= -\frac{1}{\eta}\log\left(1 - \mathbb{E}_{i\sim p^t}\left[\eta\widetilde{l}_i^t + \frac{1}{2}\eta^2(\widetilde{l}_i^t)^2\right]\right) \\
&\geq \frac{1}{\eta}\mathbb{E}_{i\sim p^t}\left[\eta\widetilde{l}_i^t + \frac{1}{2}\eta^2(\widetilde{l}_i^t)^2\right] \qquad \text{(because } log(1-x) \leq -x\text{)} \\
&= \sum_{i=1}^n p_i^t\widetilde{l}_i^t - \frac{\eta}{2}\sum_{i=1}^n p_i^t(\widetilde{l}_i^t)^2
\end{aligned}$$

Taking expectations on both sides of the above equation, we have:

$$\begin{aligned}
\mathbb{E}[\Phi_{t+1} - \Phi_t] &\geq \mathbb{E}\left[\sum_{i=1}^n p_i^t\widetilde{l}_i^t - \frac{\eta}{2}\sum_{i=1}^n p_i^t(\widetilde{l}_i^t)^2\right] \\
&= \sum_{i=1}^n p_i^t l_i^t - \frac{\eta}{2}\mathbb{E}\left[p_{I_t}^t\left(\frac{l_{I_t}^t}{p_{I_t}^t}\right)^2\right] \\
&= \underline{p}^t \cdot \underline{l}^t - \frac{\eta}{2}\mathbb{E}\left[\frac{(l_{I_t}^t)^2}{p_{I_t}^t}\right] \\
&= \underline{p}^t \cdot \underline{l}^t - \frac{\eta}{2}\sum_{i=1}^n (l_i^t)^2 \\
&\geq \underline{p}^t \cdot \underline{l}^t - \frac{\eta n}{2}
\end{aligned}$$

Now, we sum the differences in potential to get

$$\mathbb{E}[\Phi_{T+1} - \Phi_1] = \mathbb{E}\left[\sum_{t=1}^T(\Phi_{t+1} - \Phi_t)\right] \geq \sum_{t=1}^T \underline{p}^t \cdot \underline{l}^t - \frac{T\eta n}{2}$$

Moreover,

$$\mathbb{E}[\Phi_{T+1} - \Phi_1] \leq \mathbb{E}\left[\widetilde{L}_{i^*}^T - \left(-\frac{1}{\eta}\log n\right)\right] = L_{i^*}^T + \frac{1}{\eta}\log n$$

Combining the two inequalities, we get

$$\mathbb{E}\ \text{Regret}_T(EXP3) = \sum_{t=1}^T \underline{p}^t \cdot \underline{l}^t - L_{i^*}^T \leq \frac{1}{\eta}\log n + \frac{T\eta n}{2} \qquad (*)$$

**Theorem 22.2.**

$$\mathbb{E}\ \textit{Regret}_T(EXP3) \leq \sqrt{2Tn\log n}$$

**Proof:** Choose $\eta = \sqrt{\frac{2\log n}{Tn}}$ in $(*)$. ■

## 22.2 Progress after EXP3

### 22.2.1 Bubeck *et al*: EXP2 With John's Exploration [BCBK12]

In the title, 'John's Exploration' refers to the 'John Ellipsoid': Given a set of points, we may define their convex hull $K$. The ellipsoid of maximal volume contained inside $K$ is the John Ellipsoid. John's Theorem characterizes when this ellipsoid is the unit ball in $\mathbb{R}^n$.

Given a learning rate $\eta$, mixing coefficient $\gamma$, and action set $\mathcal{A}$ with distribution $\mu$, we may define the following algorithm.

Let $n = |\mathcal{A}|$ and $X^+$ denote the pseudoinverse of a matrix $X$.

Set $q_1 = \left(\frac{1}{n}, \cdots, \frac{1}{n}\right) \in \mathbb{R}^n$
**for** t = 1, 2, $\cdots$, T-1, T **do**
  Let $p_t = (1-\gamma)q_t + \gamma\mu$
  Choose an action $a_t \sim p_t$
  Let $P_t$ be the covariance matrix $\mathbb{E}_{a\sim p_t}\left[aa^T\right]$ and compute $P_T^+$
  Estimate the loss $\widetilde{l}_t = P_t^+(a_t a_t^T)l_t$
  Update $q_{t+1}(a) = \frac{\exp\left(-\eta\langle a, \widetilde{l}_t\rangle\right)q_t(a)}{\sum_{b\in\mathcal{A}}\exp\left(-\eta\langle b, \widetilde{l}_t\rangle\right)q_t(b)}$
**end for**

When $\mu$, $\gamma$, and $\eta$ are chosen based on the geometry of $\mathcal{A}$, a regret bound of $O(\sqrt{nT})$ is obtained.

### 22.2.2 Abernethy *et al*: GBPA [ALT15]

Consider the following framework: The Gradient-Based Prediction Algorithm (GBPA) for Multi-Armed Bandits:

Given a differentiable convex function $\Phi$ such that $\nabla\Phi \in \Delta^N$ with $\nabla_i\Phi > 0$ for all $i$,

Initialize $\hat{G}_0 = 0$
**for** t = 1, 2, $\cdots$, T-1, T **do**
  Nature (The Adversary) chooses a loss vector $g_t \in [-1, 0]^N$
  The Learner chooses $i_t$ according to the distribution $p(\hat{G}_{t-1} = \nabla\Phi_t(\hat{G}_{t-1})$
  The Learner incurs loss $g_{t,i_t}$
  The Learner predicts $\hat{g}_t = \frac{g_{t,i_t}}{p_{i_t}(\hat{G}_{t-1})}\mathbf{e}_{i_t}$
  $\hat{G}_t = \hat{G}_{t-1} + \hat{g}_t$
**end for**

Note that GBPA includes FTRL and FTPL as special cases.

Recall that the negative Shannon Entropy is defined as $H(p) = \sum_i p_i \log p_i$, and has Fenchel Conjugate $H^*(G) = \frac{1}{\eta}\log(\sum_i e^{\eta G_i})$. With these definitions, EXP3 is merely GBPA with $\Phi$ chosen as the Fenchel Conjugate of the Shannon Entropy with update rule $p_t = \nabla H^*(G)$.

Now, define the Tsallis entropy:

$$S_\alpha(p) = \frac{1}{1-\alpha}\left(1 - \sum_{i=1}^N p_i^\alpha\right) \qquad \forall\alpha \in (0,1)$$

Note that the Shannon Entropy is recovered as the limit of the Tsallis entropy as $\alpha \to 1$. If we replace the Shannon Entropy with the Tsallis in GBPA, we have a regret bound

$$\mathbb{E}\ \text{Regret} \leq \eta\frac{N^{1-\alpha}-1}{1-\alpha} + \frac{N^\alpha T}{2\eta\alpha}$$

Choosing $\alpha = \frac{1}{2}$ yields a bound of $O(\sqrt{NT})$.

### 22.2.3 Shamir: Information-Theoretic Lower Bounds [Sha14]

Shamir analyzed the limitations of online algorithms for statistical learning and estimation. In particular, he analyzed things like memory-sample complexity trade-offs, communication-sample complexity trade-offs, and various information-theoretic characterizations of online learning. In particular, he gives a lower bound on the regret of a partial information set-up in an online learning algorithm. In particular, for $n$-dimensional loss vectors $\ell_t \in [0,1]^n$ at every iteration, assume that only $b < n$ bits are available. Then, there exists some constant $c$ such that the regret has lower bound

$$\min_{i^*} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(i_t) - \sum_{t=1}^{T} \ell_t(i_t^*)\right] \geq c \min\left\{T, \sqrt{\frac{n}{b}T}\right\}$$

### 22.2.4 Neu: High Probability Regret Bounds [Neu15]

Neu gives regret bounds for general bandit problems that hold with high probability, i.e., with probability $1-\delta$ for some small $\delta$. In particular, one application given is a modification of EXP3. Define some parameter $\gamma$ and modify EXP3 as follows: set

$$\tilde{\underline{l}}^t = \left\langle 0, ..., 0, \frac{l^t_{I_t}}{\gamma + p^t_{I_t}}, 0, ..., 0 \right\rangle$$

This modification leads to a regret bound of $O(\sqrt{NT \log \frac{N}{\delta}})$ with probability $1-\delta$.

## References

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[ALT15] Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems*, pages 2188–2196, 2015.

[BCBK12] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. *arXiv preprint arXiv:1202.3079*, 2012.

[Neu15] Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 3150–3158, 2015.

[Sha14] Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems*, pages 163–171, 2014.