CSE599s, Spring 2012, Online Learning Lecture 12 - 05/03/2012

The Multi-armed Bandit Problem

Lecturer: Ofer Dekel *Scribe: Yanping Huang*

1 Model for learning with Expert Advice

Recall the problem of learning with expert advice. Let N be the number experts. The Player (our online learning algorithm) maintains a probability distribution $\mathbf{w}_t = \{w_{t,i}\}_{i=1,\dots,N}$, $w_{t,i} \geq 0$, and $\sum_{i=1}^{N} w_{t,i} = 1$. On each round,

1. Player randomly picks expert $I_t \sim \mathbf{w}_t$.
2. Adversary picks loss $\mathbf{g}_t \in [0, G]^N$. $\mathbf{g}_t = \{g_{t,i}\}_{i=1}^{N}$ where $g_{t,i}$ is the loss associated with following the advice from expert i.
3. Player suffers loss g_{t,I_t}.

Since the player adopts a random strategy, we cares about the expected loss per time: $f_t = \mathsf{E}[g_{t,I_t}] = \sum_i w_{t,i} g_{t,i}$. The goal of the player is is to keep as small as possible the regret with respect to a single fixed distribution $\mathbf{w}^\star$:

$$\text{Regret}(T) = \sum_{t=1}^{T} \mathbf{w}_t \cdot \mathbf{g}_t - \sum_t \mathbf{w}^\star \cdot \mathbf{g}_t \tag{1}$$

Consider the normalized Exponentiated Gradient algorithm: The above algorithm is equivalent to Follow

Algorithm 1 Normalized Exponentiated Gradient(EG)

Choose $\eta > 0$
Inistialize $\mathbf{w}_1 = (1/N, \dots, 1/N)$
Update rule: $\forall i$, $w_{t+1,i} = \frac{\exp(-\eta g_{1:t,i})}{\sum_{j=1}^{N} \exp(-\eta g_{1:t,j})} = \frac{w_{t,i} \exp(-\eta g_{t,i})}{\sum_{j=1}^{N} w_{t,j} \exp(-\eta g_{t,j})}$

The Regularized Leader (FTRL) algorithm with entropic regularization $R(\mathbf{w}) = 1/\eta \sum_i w_i \log(w_i)$, and enjoys a regret bound

$$\text{Regret}(T) \leq \eta G^2 T + \frac{\log(N)}{\eta} \tag{2}$$

where $\|\mathbf{g}_t\|_\infty \leq G$. In particular, setting $\eta = \frac{\sqrt{\log(n)}}{G\sqrt{2T}}$, we have $\text{Regret}(T) \leq G\sqrt{2T \log(N)}$. The entropic regularization used in the above algorithm is $1/\eta$-strongly convex with respect to the l_1-norm.

2 The Multi-armed Bandit Problem

In the multi-armed bandit problem, there are N slots (or experts) at a rigged casino, and on each online round the player

1. random randomly chooses slot $I_t \sim \mathbf{w}_t$
2. suffers a loss g_{t,I_t}

The vector $\mathbf{g}_t = \{g_{t,i}\} \in [0, G]^N$ associates a loss for each of the slot. But the player only gets to see the cost of the slot it chooses. This problem is similar to the above learning with expert advice. The only difference is that the player does not get to see the cost of experts he didn't pick. Nothing is assumed about the sequence of vectors $\mathbf{g}_1, \ldots, \mathbf{g}_T$.

The problem nicely captures the exploration-exploitation tradeoff. On one hand, the play would like to choose the slot which he believe has the lowest cost based on previous rounds. On the other hand, it may be better to explore other arms and find the arms with smaller losses.

To approach this multi-armed bandit problem, we use the exponentiated gradient (EG) algorithm described in the previous section. However, the player doesn't have access to full information of $\mathbf{g}_t$, he only sees the value g_{t,I_t}. The trick is the estimate $\mathbf{g}_t$ using the fact that the player plays randomly. We define the random vector $\tilde{\mathbf{g}}_t$:

$$\tilde{g}_{t,i} = \begin{cases} g_{t,i}/w_{t,i}, & i = I_t \\ 0. & \text{otherwise} \end{cases} \tag{3}$$

We indeed have that $\tilde{\mathbf{g}}_t$ is an unbiased estimator of the $\mathbf{g}_t$ because

$$\mathsf{E}[\tilde{g}_{t,i}|\mathbf{w}_t] = \mathsf{E}[\tilde{g}_{t,i}|\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{t-1}] = \sum_{j=1}^{N} \tilde{g}_{t,i} w_{t,j} = g_{t,i}. \tag{4}$$

We update $\mathbf{w}_t$ using the update rule of the EG algorithm. The resulting algorithm is given below The above

Algorithm 2 Multi-Armed Bandit Algorithm (EXP3 [1])

Input parameter $\eta \in [0, 1]$,
Initialize: $\mathbf{w}_1 = (1/N, \ldots, 1/N)$
for $t = 1, 2, \ldots$ **do**
  pick $I_t \sim \mathbf{w}_t$ and suffer g_{t,I_t}
  Update $w_{t+1,i} = w_{t,i} \exp(-\eta \tilde{g}_{t,i}) / \sum_{j=1}^{N} w_{t,j} \exp(-\eta \tilde{g}_{t,j})$
end for

multi-armed bandit algorithm enjoys the same regret bound as in EG algorithm: $\text{Regret}(T) \leq \eta \tilde{G}^2 T + \frac{\log(N)}{\eta}$ where $\|\tilde{\mathbf{g}}_t\|_\infty \leq \tilde{G}$. Note that since $w_{t,i} \in [0, 1]$ could be arbitrary small, we have unbounded $\tilde{G}$, thus unbound regret.

To obtain a better regret bound for the normalized EG algorithm, we first defines

$$\frac{1}{\eta} \log \sum_{i=1}^{N} w_{t,i} \exp(-\eta g_{t,i}) = \star_t \tag{5}$$

Note that, we can telescope $\star$:

$$\star_t = \frac{1}{\eta} \log \sum_{i=1}^{N} \frac{\exp(\eta g_{1:t-1,i})}{\sum_j \exp(-\eta g_{1:t-1,j})} \exp(-\eta g_{t,i}) \tag{6}$$

$$= \frac{1}{\eta} \log \sum_{i=1}^{N} \exp(-\eta g_{1:t,i}) - \frac{1}{\eta} \log \sum_{i=1}^{N} \exp(-\eta g_{1:t-1,i}). \tag{7}$$

Using inequalities $\log x \leq x - 1$ and $\exp(-x) \leq 1 - x + x^2/2$ for $x > 0$, we also have

$$\star_t = \frac{1}{\eta} \log \sum_{i=1}^{N} w_{t,i} \exp(-\eta g_{t,i}) + \mathbf{w}_t \cdot \mathbf{g}_t - \mathbf{w}_t \cdot \mathbf{g}_t \tag{8}$$

$$\leq \frac{1}{\eta} [\sum_{i=1}^{N} w_{t,i} \exp(-\eta g_{t,i}) - 1] + \mathbf{w}_t \cdot \mathbf{g}_t - \mathbf{w}_t \cdot \mathbf{g}_t \quad \text{Using } \log x \leq x - 1 \tag{9}$$

$$= \frac{1}{\eta} \sum_{i=1}^{N} w_{t,i} [\exp(-\eta g_{t,i} - 1 + \eta g_{t,i})] - \mathbf{w}_t \cdot \mathbf{g}_t \quad \text{Using } \textstyle\sum_i w_{t,i} = 1 \tag{10}$$

$$= \frac{1}{\eta} \sum_{i=1}^{N} w_{t,i} \eta^2 g_{t,i}^2 / 2 - \mathbf{w}_t \cdot \mathbf{g}_t \quad \text{Using } \exp(-x) \leq 1 - x + x^2/2 \tag{11}$$

$$\tag{12}$$

Summing $\star_t$ over $t = 1, \ldots, T$, we have the regret bound

$$\sum_{t=1}^{T} \star_t = \frac{1}{\eta} \log \sum_{i=1}^{N} \exp(-\eta g_{1:T,i}) - \frac{1}{\eta} \log(N) \geq -\max_i(g_{1:T,i}) \tag{13}$$

$$\sum_{t=1}^{T} \star_t \leq \frac{\eta}{2} \sum_{t=1}^{T} w_{t,i} g_{t,i}^2 - \sum_{t=1}^{T} \mathbf{w}_t \cdot \mathbf{g}_t \tag{14}$$

$$\text{Regret}(T) = \sum_{t=1}^{T} \mathbf{w}_t \cdot \mathbf{g}_t - \max_i(g_{1:T,i}) \leq \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{N} w_{t,i} g_{t,i}^2 + \log(N)/\eta \tag{15}$$

Note that we substitute $\tilde{g}_{t,i}$ with $g_{t,i} \in [0, G]$ in multi-armed bandit algorithm. The corresponding regret bound is then

$$\text{Regret}(T) \leq \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{N} w_{t,i} \tilde{g}_{t,i}^2 + \log(N)/\eta \tag{16}$$

To take the expectation of regret, we compute

$$\mathsf{E}[\tilde{g}_{t,i} | \mathbf{w}_t] = g_{t,i} \tag{17}$$

$$\mathsf{E}[\tilde{g}_{t,i}^2 | \mathbf{w}_t] = \sum_{j=1}^{N} w_{t,j} \frac{g_{t,i}^2}{w_{t,i}^2} \delta_{i,j} \leq \frac{G^2}{w_{t,i}} \tag{18}$$

$$\mathsf{E}[\mathsf{E}[\tilde{g}_{t,i}^2 | \mathbf{w}_t]] \leq G^2 \tag{19}$$

$$\mathsf{E}[\text{Regret}(T)] = \mathsf{E}[\sum_{t=1}^{T} \mathbf{w}_t \cdot \tilde{\mathbf{g}}_t] - \max_i[g_{1:T,i}] \tag{20}$$

$$\leq \mathsf{E}[\mathsf{E}[\frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{N} w_{t,i} \tilde{g}_{t,I_t}^2 | \mathbf{w}_t]] + \log(N)/\eta \tag{21}$$

$$= \frac{\eta}{2} TNG^2 + \log(N)/\eta \tag{22}$$

Setting $\eta = frac\sqrt{2 \log N} G \sqrt{NT}$, we have $\text{Regret}(T) \leq \sqrt{2TN \log N} G$

References

[1] Peter Auer, Nicolo Cesa-Bianchi, Yaov Freund, and Robert E. Schapire, "The non-stochastic multi-armed bandit problem.", in *SIAM journal on computing*, 32:48-77 2002.

[2] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan, "Online convex optimization in the bandit setting: gradient descent without a gradient", in *SODA 05: Proceedings of the sixteenth annual ACM–SIAM symposium on Discrete algorithms,*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.