Advanced Course in Machine Learning Spring 2010

Online Learning with Partial Feedback

Handouts are jointly prepared by Shie Mannor and Shai Shalev-Shwartz

In previous lectures we talked about the general framework of online convex optimization and derived an algorithm for prediction with expert advice from this general framework. To apply the online algorithm, we need to know the gradient of the loss function at the end of each round. In the prediction of expert advice setting, this boils down to knowing the cost of each individual expert.

In this lecture, we show that in order to apply the online mirror descent algorithm it suffices to know an estimate of the gradient. In particular, this yields a no-regret algorithm for a famous problem called "the multi-armed bandit problem".

1 Online Mirror Descent with Estimated Gradient

Recall the online mirror descent algorithm we described in Lecture 4. Now suppose that instead of setting v_t to be a sub-gradient of $g_t(w_t)$, we shall set v_t to be a random vector with $\mathbb{E}[v_t] \in \partial g_t(w_t)$.

Algorithm 1 Online Mirror Descent with Estimated Gradient

Initialize: $w_1 \leftarrow \nabla f^\star(\mathbf{0})$
for $t = 1$ to T
  Play $w_t \in A$
  Pick v_t at random s.t. $\mathbb{E}[v_t | v_{t-1}, \ldots, v_1] \in \partial g_t(w_t)$
  Update $w_{t+1} \leftarrow \nabla f^\star \left(-\eta \sum_{s=1}^{t} v_t\right)$
end for

We now show that the analysis still holds as long as we have some bound on $\mathbb{E}[\|v_t\|_\star^2]$.

Theorem 1 *Suppose Algorithm 1 is used with a function f that is β-strongly convex w.r.t. a norm $\|\cdot\|$ on A and has $f^\star(\mathbf{0}) = 0$. Suppose the loss functions g_t are convex and that $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|v_t\|_\star^2\right] \le V^2$. Then, the algorithm run with any positive η enjoys the expected regret bound,*

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(w_t) - \min_{u \in A} \sum_{t=1}^{T} g_t(u)\right] \le \frac{\max_{u \in A} f(u)}{\eta} + \frac{\eta V^2 T}{2\beta} \,.$$

In particular, choosing $\eta = \sqrt{\frac{2\beta \max_u f(u)}{V^2 T}}$ we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(w_t) - \min_{u \in A} \sum_{t=1}^{T} g_t(u)\right] \le V\sqrt{\frac{2\max_{u\in A} f(u)\, T}{\beta}} \,.$$

Proof Apply Corollary 1 from Lecture 4 to the sequence $-\eta v_1, \ldots, -\eta v_T$ to get, for all u,

$$-\eta \sum_{t=1}^{T} \langle v_t, u\rangle - f(u) \le -\eta \sum_{t=1}^{T} \langle v_t, w_t\rangle + \frac{1}{2\beta}\sum_{t=1}^{T} \|\eta v_t\|_\star^2 \,.$$

Rearranging gives,

$$\sum_{t=1}^{T} \langle v_t, w_t - u\rangle \le \frac{f(u)}{\eta} + \frac{\eta}{2\beta}\sum_{t=1}^{T} \|v_t\|_\star^2 \,.$$

Taking expectation of both sides with respect to the randomness in choosing v_t we obtain that

$$\sum_{t=1}^{T} \mathbb{E}[\langle v_t, w_t - u\rangle] \leq \frac{f(u)}{\eta} + \frac{\eta}{2\beta} T \, \mathbb{E}\left[\tfrac{1}{T}\sum_{t=1}^{T} \|v_t\|_\star^2\right] .$$

At each round, let $\bar{v}_t = \mathbb{E}[v_t | v_{t-1}, \ldots, v_1] \in \partial g_t(w_t)$. Using the assumptions in the theorem we get that

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \bar{v}_t, w_t - u\rangle\right] \leq \frac{f(u)}{\eta} + \frac{\eta}{2\beta} T\, V^2 .$$

By convexity of g_t, $g_t(w_t) - g_t(u) \leq \langle \bar{v}_t, w_t - u\rangle$. Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(w_t) - \sum_{t=1}^{T} g_t(u)\right] \leq \frac{f(u)}{\eta} + \frac{\eta V^2 T}{2\beta} .$$

Since the above holds for all $u \in A$ the result follows. ■

2 The Multi-Armed Bandit Problem

In the multi-armed bandit problem, there are d arms, and on each online round the learner should choose one of the arms, denoted I_t, where the chosen arm can be a random variable. Then, it receives a cost of choosing this arm, $c_{t,I_t} \in [0,1]$. The vector $c_t \in [0,1]^d$ associates a cost for each of the arms, but the learner only get to see the cost of the arm it pulls. Nothing is assumed about the sequence of vectors $c_1, c_2, \ldots$. The performance of the learner is using by its regret for not always pulling the best arm,

$$\mathbb{E}\left[\sum_{t=1}^{T} c_{t,I_t}\right] - \min_i \sum_{t=1}^{T} c_{t,i} ,$$

where the expectation is over the randomness of the learner.

This problem nicely captures the exploration-exploitation tradeoff. On one hand, we would like to pull the arm which, based on previous rounds, we believe has the lowest cost. On the other hand, maybe it better to explore the arms and find another arm with a smaller cost.

To approach the multi-armed bandit problem we use the general result derived in the previous section. Let the loss function be $g_t(w) = \langle w, c_t\rangle$ and note that if w_t is a probability vector and $I_t \sim w_t$, then $g_t(w_t) = \mathbb{E}[c_{t,I_t}]$. The gradient of the loss is c_t, but we don't know the value of all elements of c_t. To estimate the gradient we shall define a vector v_t s.t.

$$v_{t,j} = \begin{cases} c_{t,j}/w_{t,j} & \text{if } j = I_t \\ 0 & \text{else} \end{cases} .$$

Clearly, $\mathbb{E}[v_t] = c_t$. Additionally,

$$\mathbb{E}[\|v_t\|_\infty^2] \leq \sum_i w_{t,i} (c_{t,i})^2 / w_{t,i}^2 \leq \sum_i 1/w_{t,i} .$$

To ensure that this quantity is not excessively large we will define the set of allowed distributions to be $A = \{w : w_i \in [\gamma, 1], \sum_i w_i = 1\}$, where γ is a parameter to be defined later. Thus, $\mathbb{E}[\|v^t\|_\infty^2] \leq 1/\gamma$. Applying Theorem 1 we obtain that for all $u \in A$

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(w_t)\right] \leq \sum_{t=1}^{T} g_t(u) + \sqrt{\frac{2\log(d)\, T}{\gamma}} .$$

Finally, Let $C_i = \sum_{t=1}^T c_{t,i}$ and note that for each i if we set u to be s.t. $u_i = 1 - (d-1)\gamma$ and $u_j = \gamma$ then

$$\sum_{t=1}^{T} g_t(u) = C_i + \gamma \sum_{j \neq i} (C_j - C_i) \leq C_i + \gamma d T \ .$$

So, overall,

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(w_t)\right] \leq C_i + \gamma d T + \sqrt{\frac{2\log(d)\,T}{\gamma}} \ .$$

Setting $\gamma = (2\log(d)T/(d^2T^2))^{1/4} = (2\log(d)/(d^2T))^{1/4}$ we obtain the regret bound

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(w_t)\right] \leq C_i + O\left((\log(d) d^2 T^3)^{1/4}\right) = \tilde{O}(d^{1/2}\, T^{3/4}) \ .$$

3 An improved Multi-Armed Bandit Predictor

We now derive another algorithm, called EXP3 (which stands for "exponential-weight algorithm for exploration and exploitation), that enjoys a regret bound of $O(\sqrt{T})$. The algorithm is due to Auer, Cesa-Bianchi, Freund, and Schapire.

Remark: Throughout this section, we think about c_t as *gain* that we'd like to *maximize* rather than a cost. One can derive a result for minimizing a cost by defining $c_{t,i} \leftarrow 1 - c_{t,i}$ for all t and i.

Algorithm 2 EXP3

Parameter: $\gamma \in (0,1]$
Initialize: $w_1 = (1,\ldots,1)$
for $t = 1$ to T
  Set $Z_t = \sum_{j=1}^d w_{t,j}$
  Set $p_{t,i} = (1-\gamma) w_{t,i}/Z_t + \gamma/d$
  Pull I_t randomly according to p_t
  Receive cost $c_{t,I_t} \in [0,1]$
  Let v_t be the vector with $v_{t,j} = \frac{c_{t,j}}{p_{t,j}}\, \mathbb{1}_{[I_t = j]}$
  Update: $w_{t+1,j} = w_{t,j} e^{\gamma v_{t,j}/d}$
end for

Theorem 2 *For any* $\gamma \in (0,1)$ *and* $j \in [d]$ *we have*

$$\sum_t c_{t,j} - \mathbb{E}[C_{\text{exp3}}] \leq (e-1)\gamma \sum_t c_{t,j} + \tfrac{1}{\gamma} d \ln(d)$$

Proof We have

$$\begin{aligned}
\frac{Z_{t+1}}{Z_t} &= \sum_{i=1}^{d} \frac{w_{t+1,i}}{Z_t} \\
&= \sum_{i=1}^{d} \frac{w_{t,i}}{Z_t} e^{\gamma v_{t,j}/d} \\
&\leq \sum_{i=1}^{d} \frac{w_{t,i}}{Z_t} \left(1 + \gamma v_{t,j}/d + (e-2)(\gamma v_{t,i}/d)^2\right) \ ,
\end{aligned}$$

where in the last inequality we used the inequality $e^x \leq 1 + x + (e-2)x^2$ which holds for $x \leq 1$.

Denote $\bar{w}_{t,i} = w_{t,i}/Z_t$ and using the definition of v_t, the above implies:

$$\frac{Z_{t+1}}{Z_t} \ \leq \ 1 + \tfrac{\gamma}{d}\bar{w}_{t,I_t} v_{t,I_t} + (e-2)\left(\tfrac{\gamma}{d}\right)^2 \ w_{t,I_t}\, v_{t,I_t}^2 \ .$$

Since $\bar{w}_{t,I_t} \leq p_{t,I_t}/(1-\gamma)$, and using the definition of v_{t,I_t} we get

$$\frac{Z_{t+1}}{Z_t} \ \leq \ 1 + \tfrac{\gamma}{d(1-\gamma)} c_{t,I_t} + (e-2)\left(\tfrac{\gamma}{d}\right)^2 \ \tfrac{1}{1-\gamma}\ \tfrac{c_{t,I_t}}{p_{t,I_t}} \ .$$

Taking logarithms of both sides and using $\ln(1+x) \leq x$ we get

$$\ln\, \frac{Z_{t+1}}{Z_t} \ \leq \ \tfrac{\gamma}{d(1-\gamma)} c_{t,I_t} + (e-2)\left(\tfrac{\gamma}{d}\right)^2 \ \tfrac{1}{1-\gamma}\ \tfrac{c_{t,I_t}}{p_{t,I_t}} \ .$$

Summing over t we obtain

$$\ln\, \frac{Z_{t+1}}{Z_t} \ \leq \ \tfrac{\gamma}{d(1-\gamma)} C_{\text{exp3}} + (e-2)\left(\tfrac{\gamma}{d}\right)^2 \ \tfrac{1}{1-\gamma}\ \sum_{t=1}^{T} \tfrac{c_{t,I_t}}{p_{t,I_t}} \ .$$

On the other hand, for any action j we have

$$\ln\, \frac{Z_{t+1}}{Z_t} \ \geq \ \ln \frac{w_{T+1,j}}{Z_1} \ \geq \ \frac{\gamma}{d}\sum_{t=1}^{T} v_{t,j} - \ln d \ .$$

Combining the upper and lower bound we obtain

$$\frac{\gamma}{d}\sum_{t=1}^{T} v_{t,j} - \ln d \ \leq \ \tfrac{\gamma}{d(1-\gamma)} C_{\text{exp3}} + (e-2)\left(\tfrac{\gamma}{d}\right)^2 \ \tfrac{1}{1-\gamma}\ \sum_{t=1}^{T} \tfrac{c_{t,I_t}}{p_{t,I_t}} \ .$$

Now, take expectation of both sides (w.r.t. to the random choice of I_t). Note that $\mathbb{E}[v_t | I_{t-1}, \ldots, I_1] = c_t$ and that $\mathbb{E}[c_{t,I_t}/p_{t,I_t} | I_{t-1}, \ldots, I_1] = \sum_i c_{t,i} \leq d c_{t,j}$. Therefore,

$$\mathbb{E}\left[\frac{\gamma}{d}\sum_{t=1}^{T} c_{t,j} - \ln d\right] \leq \mathbb{E}\left[\tfrac{\gamma}{d(1-\gamma)} C_{\text{exp3}} + (e-2)\left(\tfrac{\gamma}{d}\right)^2 \ \tfrac{1}{1-\gamma}\ d\sum_{t=1}^{T} c_{t,j}\right] \ .$$

Rearranging the above gives

$$\sum_t c_{t,j} - \mathbb{E}[C_{\text{exp3}}] \leq (e-1)\gamma \sum_t c_{t,j} + \tfrac{1-\gamma}{\gamma} d \ln(d) \ ,$$

which concludes our proof. ■

Corollary 1 *Choose $\gamma = \min\{1, \sqrt{d\ln(d)/((e-1)g}\}$, then for any j s.t. $\sum_t c_{t,j} \geq g$ we have*

$$\sum_t c_{t,j} - \mathbb{E}[C_{\text{exp3}}] \leq 2\sqrt{e-1}\sqrt{gd\ln(d)} = O(\sqrt{Td\ln(d)}) \ .$$

3.1 Lower bound

Theorem 3 *For any $d \geq 2$ and $T \geq 1$ there exists a distribution over assignments of rewards such that the expected regret of any algorithm (where expectation is both with respect to the randomization of the algorithm and the assignments of rewards) is at least $\Omega(\min\{\sqrt{dT}, T\})$.*

A proof can be find in Auer et. al. paper. The idea is to define a distribution over rewards of arms as follows. Before the play begins, one action I is chosen uniformly at random to be the "good" action. The rewards of the good action are chosen i.i.d. to be 1 with probability $1/2 + \epsilon$ and 0 otherwise for some ϵ to be defined later. The rewards of the rest of the arms are chosen to be either 0 or 1 with probability $1/2$. Now, the idea is to show that any function defined on rewards in previous rounds cannot distinguish to well between rewards that come according to the distribution mentioned above and rewards that come from a uniform distribution.