E1 245: Online Prediction & Learning **Fall 2014**

Lecture 21 — October 16

Lecturer: Aditya Gopalan *Scribe: Rahul R*

21.1 Recap

In the last lecture, we introduced the framework of online convex optimization in which the number of experts is exponential in the size of the problem's natural representation. An efficient algorithm for the linear costs called Follow-The-Perturbed-Leader (FPL) was introduced and a broad outline for proving its general regret bound was discussed. The proof of the general regret bound introduces a non-causal algorithm called Be-the-Perturbed-Leader (BPL) for bounding the difference between the optimal policy, OPT(T) and FPL.

In this lecture, we will prove the general regret bound for FPL. The proof technique uses three lemmas for bounding the regret for FPL algorithm.

21.2 Problem setup

The set of experts, $\mathbb{D}$, is a bounded subset of R^N. Let D be the l_1- diameter of $\mathbb{D}$.

$$D = \sup_{d,d' \in \mathbb{D}} \|d - d'\|_1$$

An oblivious adversary specifies a sequence of cost vectors $c_1, c_2, ..., c_T$, $c_t \in \mathbb{C} \subseteq \mathbb{R}^N$. Let

$$A = \sup_{d,d' \in \mathbb{D}} \|c\|_1$$

An online algorithm chooses a sequence of strategies $x_1, x_2, ..., x_T$. The cost of strategy x at time t is the dot product $c_t.x$. Let

$$R = \sup_{d \in \mathbb{D}, c \in \mathbb{C}} |\langle d, c \rangle|$$

We will define the notation

$$c_{i:j} = \sum_{t=i}^{j} c_t$$

and a function $M(.)$ which finds the minimum cost given the total cost vectors, i.e.,

$$M(c) = \arg\min_{d \in \mathbb{D}} \langle d, c \rangle$$

Let $OPT(1:T) = \langle c_{1:T}, M(c_{1:T}) \rangle$

21.3 Generic regret bound for FPL

We will make use of the following claims to prove the FPL regret bound.

1. BPL(T) is close to OPT(T). Specifically,
$$\mathbb{E}[BPL(T) - OPT(T)] \leq \mathbb{E}[\ \langle\ p_0, M(C_{1:T}) - M(p_0)\ \rangle\ ] \tag{21.1}$$
2. FPL(T) is not much greater than BPL(T), the reason being that the distribution of $c_{0:t}$ and $c_{0:t-1}$ are so similar that their minimizers are very closely related.
3. $\mathbb{E}[\langle p_0, x-y\rangle]$ is small for any $x, y \in \mathbb{D}$

21.3.1 Validating Claim 1

To prove Claim 1 we will make use of the following lemma.

Lemma 21.1.
$$\forall i \leq j,\ \ \sum_{t=i}^{j} \langle c_t, M(c_{i:t})\rangle \leq \langle c_{i:j}, M(c_{i:j})\rangle \tag{21.2}$$

Proof: Proof by induction on $j-i$.

Base case, $j-i=0$:
L.H.S= $\langle c_i, M(c_i)\rangle$= R.H.S

Induction hypothesis:
$$\sum_{t=i}^{j-1} \langle c_t, M(c_{i:t})\rangle \leq \langle c_{i:j-1}, M(c_{i:j-1})\rangle$$

Now add $\langle c_j, M(c_{i:j})\rangle$ to both sides,
$$\sum_{t=i}^{j} \langle c_t, M(c_{i:t})\rangle \leq \langle c_{i:j-1}, M(c_{i:j-1})\rangle + \langle c_j, M(c_{i:j})\rangle$$

Now, $M(c_{i:j-1}) = \arg\min_{d\in\mathbb{D}} \langle d, c_{i:j-1}\rangle$. Hence, $\langle c_{i:j-1}, M(c_{i:j-1})\rangle \leq \langle c_{i:j-1}, d\rangle, \forall d \in \mathbb{D}$. In particular, $\langle c_{i:j-1}, M(c_{i:j-1})\rangle \leq \langle c_{i:j-1}, M(c_{i:j})\rangle$. Hence,

$$\sum_{t=i}^{j}\langle c_t, M(c_{i:t})\rangle \leq \langle c_{i:j-1}, M(c_{i:j})\rangle + \langle c_j, M(c_{i:j})\rangle$$
$$\leq \langle c_{i:j}, M(c_{i:j})\rangle$$

□

Setting $c_0 = p_0, i = 0$ and $j = T$ in (21.1) gives

$$\sum_{t=0}^{T}\langle c_t, M(c_{0:t})\rangle \leq \langle c_{0:t}, M(c_{0:t})\rangle$$
$$\leq \langle c_{0:t}, M(c_{1:t})\rangle$$

Substract $\langle c_0, M(c_0)\rangle$ from both sides to get

$$\begin{aligned}\text{BPL(T)} = \sum_{t=1}^{T}\langle c_t, M(c_{0:t})\rangle &\leq \langle c_{0:t}, M(c_{1:t})\rangle - \langle c_0, M(c_0)\rangle \\ &= \langle c_{1:t}, M(c_{1:t})\rangle + \langle c_0, M(c_{1:t})\rangle - \langle c_0, M(c_0)\rangle \\ &= \text{OPT(T)} + \langle c_0, M(c_{1:t}) - M(c_0)\rangle \\ &= \text{OPT(T)} + \langle p_0, M(c_{1:t}) - M(p_0)\rangle\end{aligned}$$

21.3.2 Validating Claim 2

We need to show that when $C_0 = P_0$, which is the distribution used by FPL and BPL, the distribution of $C_{0:t-1} \approx C_{0:t}$. To do this we need a metric to measuer "distance" between distributions.

1. Multiplicative distance $d_*(.,.)$
   For distributions p and q over $\mathbb{R}^N$, $d_*(p,q) \doteq \min_{\delta \geq 0} \delta$, s.t. their pdfs satisfy

$$dp(x) \leq (1+\delta)dq(x)$$
$$dq(x) \leq (1+\delta)dp(x)$$

2. Additive distance $d_+(.,.)$
   For distributions p and q over $\mathbb{R}^N$, $d_+(p,q) \doteq \min_{\delta \geq 0} \delta$, s.t. there exist a joint distribution μ (coupling distribution) on $\mathbb{R}^N x \mathbb{R}^N$ satisfying $\forall S \subseteq \mathbb{R}^N$, $\mu(S,\mathbb{R}^N) = p(S)$, $\mu(\mathbb{R}^N, S) = q(S)$ and

$$\mu\{(x,y) : x \neq y\} \leq \delta$$

Lemma 21.2. *Let p,q be two probability distributions on $\mathbb{R}^N$. Then,*

1. *For any $f:\mathbb{D}\to[-R,R]$*

$$\mathbb{E}_{c\sim p}[f(M(C))] \le \mathbb{E}_{c\sim q}[f(M(C))] + 2Rd_+(p,q) \tag{21.3}$$

2. *For any $f:\mathbb{D}\to\mathbb{R}_+$*

$$\mathbb{E}_{c\sim p}[f(M(C))] \le [1+d_*(p,q)]\mathbb{E}_{c\sim q}[f(M(C))] \tag{21.4}$$

Proof: For (21.3), let $(C,C') \sim \mu_{\mathbb{R}^N x \mathbb{R}^N}$ s.t. $\mu(C\neq C')\le\delta$, then

$$\begin{aligned}\mathbb{E}_{C\sim p}[f(M(C))] - \mathbb{E}_{C\sim q}[f(M(C))] &= \mathbb{E}_\mu[f(M(C)) - f(M(C'))]\\ &= \mathbb{E}_\mu[f(M(C)) - f(M(C'))|C\neq C']\mu(C\neq C')\\ &\le 2R\delta\end{aligned}$$

For (21.4),

$$\begin{aligned}\mathbb{E}_{C\sim p}[f(M(C))] &= \int_{\mathbb{R}^N} f(M(C))dp(C)\\ &\le \int_{\mathbb{R}^N} f(M(C))(1+\delta)dq(C)\\ &\le (1+\delta)\mathbb{E}_{C\sim q}[f(M(C))]\end{aligned}$$

□

Define

$$\begin{aligned}D &:= \max_{x,y\in\mathbb{D}} \|x-y\|_1\\ A &:= \max_{c\in\mathbb{C}} \|C\|_1\\ R &:= \max_{c\in\mathbb{C},d\in\mathbb{D}} |\langle c,d\rangle|\end{aligned}$$

Corollary 21.3. 1. *Suppose $d_+(C_0, C+C_0)\le\delta$, $\forall C\in\{C_1,C_2,\ldots,C_T\}$, then*

$$\mathbb{E}[FPL_{unif}(T)] \le \mathbb{E}[BPL_{unif}(T)] + 2\delta RT$$

2. *Suppose $d_*(C_0,C+C_0)\le\delta$, $\forall C\in\{C_1,C_2,\ldots,C_T\}$. Also, suppose $\langle c,x\rangle\ge 0$, $\forall C\in\{C_1,C_2,\ldots,C_T\}$, $x\in\mathbb{D}$. Then,*

$$\mathbb{E}[FPL^*(T)] \le (1+\delta)\mathbb{E}[BPL^*(T)]$$

Proof:

$$\mathbb{E}[FPL(T)] = \mathbb{E}[\sum_{t=1}^{T}\langle C_t, M(C_{0:t-1})\rangle] = \mathbb{E}[\sum_{t=1}^{T}\langle C_t, f_t(C_0)\rangle]$$
$$\mathbb{E}[BPL(T)] = \mathbb{E}[\sum_{t=1}^{T}\langle C_t, M(C_{0:t})\rangle] = \mathbb{E}[\sum_{t=1}^{T}\langle C_t, f_t(C_0+C_t)\rangle]$$

Proof of the corollary then follows by comparing term-by-term, applying Lemma(21.2) and using translational invariance of d_+ and d_*. □

Denote by FPL(ε), the version of FPL with perturbation distribution

$$d\alpha(x) = \begin{cases} \varepsilon^N & \text{if } \|x_i\|_\infty \leq \frac{1}{2\varepsilon},\ \forall i \in \{1,2,3,\ldots,N\} \\ 0 & \text{otherwise} \end{cases}$$

Denote by $FPL^*(\varepsilon)$, the version of FPL with perturbation distribution

$$d\mu(x) = \left(\frac{\varepsilon}{2}\right)^N e^{-\varepsilon\|x\|_1}$$

Lemma 21.4. *Controlling d_+, d_* for $d_\alpha(.,.)$ and $d_\mu(.,.)$*

1. *Let $C \in \mathbb{R}^N$. If $C_0 \sim \alpha$ on $\mathbb{R}^N$, then*
$$d_+(C_0, C+C_0) \leq \varepsilon\|C\|_1$$
2. *If $C_0 \sim \mu$ on $\mathbb{R}^N$, then,*
$$d_*(C_0, C+C_0) \leq e^{-\varepsilon\|C\|_1} - 1$$

Proof: 1. Define

$$C_0 = \begin{cases} C_0 & \text{if } \|C-C_0\|_\infty \leq \frac{1}{2\varepsilon} \\ C-C_0 & \text{otherwise} \end{cases}$$

Observe that, C' has same distribution as $C+C_0$. Hence,

$$\begin{aligned} p[C_0 \neq C'] &\leq p\left[\|C-C_0\|_\infty > \frac{1}{2\varepsilon}\right] \\ &= p\left[\cup_{i=1}^{N}\{|C_0(i)-C(i)| > \frac{1}{2\varepsilon}\}\right] \\ &\leq \sum_{i=1}^{N} p\left[|C_0(i)-C(i)| > \frac{1}{2\varepsilon}\right] \\ &= \sum_{i=1}^{N}\varepsilon|C(i)| \\ &= \varepsilon\|C\|_1 \end{aligned}$$

2. We have,

$$\begin{aligned} d\mu(x+C) &= \left(\frac{\varepsilon}{N}\right)^N e^{-\varepsilon\|x+C\|_1} \\ &\leq \left(\frac{\varepsilon}{N}\right)^N e^{-\varepsilon\|x\|_1+\varepsilon\|C\|_1} \\ &= d\mu(x)e^{-\varepsilon\|C\|_1} \\ \text{i.e., } d_*(C+C_0) &\leq e^{-\varepsilon\|C\|_1}-1 \end{aligned}$$

□

21.3.3 Validating Claim 3

Lemma 21.5.

$$\begin{aligned} \mathbb{E}[\langle P_0, x-y\rangle] &\leq \|x-y\|_1 \mathbb{E}[\|P_0\|_\infty] \;\; \{ \textit{ Holder's Inequality } \} \\ &\leq D.\mathbb{E}[\|P_0\|_\infty] \end{aligned}$$

Lemma 21.6. *1. If $P_0 \sim \alpha$, then $\mathbb{E}[\|P_0\|_\infty] \leq \frac{1}{2\varepsilon}$*

2. If $P_0 \sim \mu$, then $\mathbb{E}[\|P_0\|_\infty] = O\left[\frac{\log N}{\varepsilon}\right]$

Proof: If $P_0 \sim \alpha$, the proof is immediate.
When $P_0 \sim \mu$, each coordinate of $|P_0| \sim \exp\left(\frac{1}{\varepsilon}\right)$.

$$\begin{aligned} \mathbb{E}[\|P_0\|_\infty] &= \mathbb{E}[\max_{i=1,2,\ldots,N} |P_0(i)|] \\ &\leq \frac{\log N}{\varepsilon} + \mathbb{E}\left[\max_{i=1,2,\ldots,N}\left(Y_i - \frac{\log N}{\varepsilon}\right)\right] \\ &\leq \frac{\log N}{\varepsilon} + \mathbb{E}\left[\max_{i=1,2,\ldots,N}\left(Y_i - \frac{\log N}{\varepsilon}\right)^+\right] \\ &\leq \frac{\log N}{\varepsilon} + \sum_{i=1}^N \mathbb{E}[\max\left(Y_i - \frac{\log N}{\varepsilon}, 0\right)] \\ &= \frac{\log N}{\varepsilon} + \sum_{i=1}^N P\left[Y_i \geq \frac{\log N}{\varepsilon}\right] \mathbb{E}\left[\max\left(Y_i - \frac{\log N}{\varepsilon}\right) | Y_i \geq \frac{\log N}{\varepsilon}\right] \\ &= \frac{\log N}{\varepsilon} + \sum_{i=1}^N \left(e^{-\varepsilon \frac{\log N}{\varepsilon}}\right) . \frac{1}{\varepsilon} \\ &= \frac{\log N}{\varepsilon} + + N.\frac{1}{N}.\frac{1}{\varepsilon} \\ &\leq \frac{2N\log N}{\varepsilon} \text{ when } N \geq 3 \end{aligned}$$

$\square$

Putting everything together:

$$
\begin{aligned}
\mathbb{E}[FPL_\alpha(T)] &\leq \mathbb{E}[BPL(T)] + 2\delta RT \\
&\leq \mathbb{E}[BPL(T)] + 2\varepsilon ART \\
&\leq OPT(T) + \frac{D}{2\varepsilon} + 2\varepsilon ART \\
\mathbb{E}[FPL^*(T)] &\leq (1+\delta)\mathbb{E}[BPL^*(T)] \\
&\leq e^{\varepsilon A}\mathbb{E}[BPL^*(T)] \\
&\leq e^{\varepsilon A}\left[OPT(T) + \frac{2D\log N}{\varepsilon}\right]
\end{aligned}
$$

References:

1. **Kleinberg lecture notes on FPL**

2. Kalai, Adam, and Santosh Vempala. **Efficient algorithms for online decision problems**. Journal of Computer and System Sciences 71.3 (2005): 291-307.