CS 683 — Learning, Games, and Electronic Markets **Spring 2007**

Notes from Week 7: Learning with Many Experts

Instructor: Robert Kleinberg *5-9 Mar 2007*

1 The Hannan-Kalai-Vempala Algorithm

Sometimes one needs to solve a version of the best-expert problem in which the number of experts is exponential in the size of the problem's natural representation. For example, consider the problem of choosing a route to take from home to work every day. The road network forms a directed graph, with edge delays that vary from day to day. The number of paths joining two nodes is exponential in general. An algorithm originally discovered by Hannan, and rediscovered by Kalai and Vempala, demonstrates how to learn the best strategy efficiently, provided that:

- Strategies can be represented as vectors in a low-dimensional vector space. The costs of different strategies are defined by a linear function on this vector space.
- There is an efficient algorithm for minimizing linear functions on the set of strategies.

1.1 Notation

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

$$D = \sup_{x,y \in \mathscr{S}} \|x - y\|_1.$$

An oblivious adversary specifies a sequence of *cost vectors* $c_1, c_2, \ldots, c_T$. Let

$$A = \sup_{1 \le t \le T} \|c_t\|_1.$$

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

$$R = \sup_{1 \le t \le T,\, x,y \in \mathscr{S}} |c_t \cdot (x - y)|.$$

For convenience we will define the notation

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

If ALG is a (possibly randomized) algorithm which chooses a sequence of strategies $x_1, x_2, \ldots, x_T \in \mathscr{S}$, define

$$\mathsf{ALG}(i..j) = \sum_{t=i}^{j} c_t \cdot x_t.$$

For any vector c we will define $M(c)$ to be an arbitrary element of $\arg\min_{x \in \mathscr{S}} c \cdot x$.

We will be interested in algorithms which learn to approximately minimize the average cost of the chosen strategies. We will present algorithms which satisfy a multiplicative bound

$$\mathbf{E}\left[\sum_{t=1}^{T} c_t \cdot x_t\right] \leq (1 + O(\varepsilon A))\mathbf{E}[c_{1..T} \cdot M(c_{1..T})] + O\left(\frac{D}{\varepsilon}\log(n)\right) \tag{1}$$

or an additive bound

$$\mathbf{E}\left[\sum_{t=1}^{T} c_t \cdot x_t\right] \leq \mathbf{E}[c_{1..T} \cdot M(c_{1..T})] + O\left(\sqrt{DRAT}\right). \tag{2}$$

1.2 "Follow the leader" and "Follow the perturbed leader"

One natural algorithm for this problem is "follow the leader" (FTL), which always picks the strategy which has performed best in the past, i.e. the algorithm which selects x_t according to the rule

$$x_{t+1} = M(c_{1..t-1}).$$

This algorithm performs very well when the cost functions are i.i.d. samples from a distribution, but can perform very poorly when the cost functions are chosen adversarially.

Example 1. Suppose $n = 2$, $\mathscr{S} = \{(0,1), (1,0)\}$, $c_1 = \left(\frac{1}{3}, \frac{2}{3}\right)$, and for $t > 1$,

$$c_t = \begin{cases} (1,0) & \text{if } t \text{ is even} \\ (0,1) & \text{if } t \text{ is odd.} \end{cases}$$

For both elements of $\mathscr{S}$, the cost grows at a rate of $T/2 + O(1)$, but FTL always chooses the strategy whose cost in the current period is 1, resulting in a cumulative loss of $T - O(1)$.

The example demonstrates that FTL can perform badly in cases in which it is "badly synchronized" with the input sequence $c_1, c_2, \ldots, c_T$. This suggests using randomization to avoid such synchronization problems. Let c_0 be a random vector sampled from some probability distribution p on $\mathbb{R}^n$. Think of c_0 is an extra cost vector which the algorithm "hallucinates" at time 0. The algorithm "follow the perturbed

leader" (FPL_p) is the same as "follow the leader" except that it minimizes the sum of actual cost and hallucinated cost, i.e. it samples c_0 randomly from distribution p and then chooses its strategy at time t using the rule

$$x_t = M(c_{0..t-1}).$$

Let $\varepsilon > 0$ be any positive number. We will be analyzing the performance of FPL_p when p is one of the distributions μ, α with the following two density functions:

$$\begin{aligned} d\mu(x) &= \left(\frac{\varepsilon}{2}\right)^n e^{-\varepsilon\|x\|_1} \\ d\alpha(x) &= \begin{cases} \left(\frac{\varepsilon}{2}\right)^n & \text{if } \|x\|_\infty \le \frac{1}{\varepsilon} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

A random sample from μ is generated by sampling each coordinate independently using the following procedure: draw a random positive number y from the exponential distribution $\Pr(y > r) = e^{-\varepsilon r}$, and change its sign from positive to negative with probability 1/2. A random sample from α is generated by sampling each coordinate independently from the uniform distribution on $\left[-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}\right]$.

1.3 "Be the perturbed leader"

Define an algorithm "Be the perturbed leader" (BPL) analogously to FPL, except that it has one-step lookahead. (Hence it is not an online algorithm!) In other words, BPL chooses $x_t = M(c_{0..t})$.

Let $\text{OPT}(1..T) = c_{1..T} \cdot M(c_{1..T})$. Our analysis of FPL will be based on three facts.

1. $\text{BPL}(1..T)$ is not much greater than $\text{OPT}(1..T)$. Specifically,

$$\text{BPL}(1..T) \le \text{OPT}(1..T) + \mathbf{E}\left[c_0 \cdot (M(c_{1..T}) - M(c_0))\right].$$

2. $\text{FPL}(1..T)$ is not much greater than $\text{BPL}(1..T)$. (Reason: the distribution of $c_{0..t}$ and $c_{0..t-1}$ are so similar that their minimizers are very closely related.)

3. $\mathbf{E}\left[c_0 \cdot (x - y)\right]$ is small for any $x, y \in \mathscr{S}$.

The remainder of this section is devoted to proving the first of these three facts.

Lemma 1. *For all $i \le j$,*

$$\sum_{t=i}^{j} c_t \cdot M(c_{i..t}) \le c_{i..j} \cdot M(c_{i..j}).$$

Proof. The proof is by induction on $j - i$. When $j - i = 0$ the lemma is trivial since both sides of the inequality are equal to $c_i \cdot M(c_i)$. For $j - i > 1$, the induction hypothesis (together with the definition of $M(c_{i..j-1})$) yields the inequality

$$\sum_{t=i}^{j-1} c_t \cdot M(c_{i..t}) \leq c_{i..j-1} \cdot M(c_{i..j-1}) \leq c_{i..j-1} \cdot M(c_{i..j}).$$

Adding $c_j \cdot M(c_{i..j})$ to both sides, we obtain

$$\sum_{t=i}^{j} c_t \cdot M(c_{i..t}) \leq c_{i..j} \cdot M(c_{i..j})$$

as desired. □

Corollary 2. $\mathrm{BPL}(1..T) \leq \mathrm{OPT}(1..T) + \mathbf{E}\left(c_0 \cdot [M(c_{1..T}) - M(c_0)]\right)$.

Proof. Applying Lemma 1 with $i = 0, j = T$ leads to the inequality

$$\sum_{t=0}^{T} c_t \cdot M(c_{0..t}) \leq c_{0..T} \cdot M(c_{0..T}) \leq c_{0..T} \cdot M(c_{1..T}).$$

Subtracting $c_0 \cdot M(c_0)$ from both sides, we obtain

$$\begin{aligned} \sum_{t=1}^{T} c_t \cdot M(c_{0..t}) &\leq c_0 \cdot M(c_{1..T}) + c_{1..T} \cdot M(c_{1..T}) - c_0 \cdot M(c_0) \\ &= \mathrm{OPT}(1..T) + c_0 \cdot [M(c_{1..T}) - M(c_0)]. \end{aligned}$$

The corollary follows by taking the expected value of both sides. □

1.4 Comparing BPL and FPL

As we said, comparison of BPL and FPL will be accomplished by showing that $c_{0..t-1}$ and $c_{1..t}$ have very similar distributions. For this, we need a measure of similarity of distributions. Actually, two measures of similarity will be useful because we're trying to prove a multiplicative bound (1) and an additive bound (2).

Definition 1. For two distributions p, q on $\mathbb{R}^n$, their multiplicative distance (denoted by $d_\times(p, q)$) is the minimum δ such that their density functions satisfy

$$\begin{aligned} dp(x) &\leq (1 + \delta) dq(x) \\ dq(x) &\leq (1 + \delta) dp(x) \end{aligned}$$

for all x. Their additive distance (denoted by $d_+(p, q)$) is the minimum δ such that there exists a probability distribution μ on pairs $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$\mu(x \neq y) \leq \delta$$

and for all measurable $S \subseteq \mathbb{R}^n$

$$\begin{aligned} \mu(x \in S) &= p(S) \\ \mu(y \in S) &= q(S). \end{aligned}$$

(This distribution μ is called a *coupling* of p and q.)

Lemma 3. *Let p, q be two distributions on $\mathbb{R}^n$.*

- *For any $f : \mathscr{S} \to [-R, R]$,*

$$\mathbf{E}_{c \leftarrow p}[f(M(c))] \leq \mathbf{E}_{c \leftarrow q}[f(M(c))] + R d_+(p, q).$$

- *For any $f : \mathscr{S} \to \mathbb{R}_+$,*

$$\mathbf{E}_{c \leftarrow p}[f(M(c))] \leq (1 + d_\times(p, q))\mathbf{E}_{c \leftarrow q}[f(M(c))]$$

Proof. For the first part, let (c, c') be sampled from a joint distribution μ whose marginals are p, q (respectively) and such that $\mu(x \neq y) \leq \delta$. Then

$$\begin{aligned} \mathbf{E}_{c \leftarrow p}[f(M(c))] - \mathbf{E}_{c \leftarrow q}[f(M(c))] &= \mathbf{E}[f(M(c)) - f(M(c'))] \\ &= \mathbf{E}[f(M(c)) - f(M(c')) \,|\, c = c']\, \mu(c = c') \\ &\quad + \mathbf{E}[f(M(c)) - f(M(c')) \,|\, c \neq c']\, \mu(c \neq c'). \end{aligned}$$

The first term on the right side is 0, and the second term is bounded above by $2R\delta$ since $f(x) - f(y) \leq 2R$ for all $x, y \in \mathscr{S}$, and $\mu(c \neq c') \leq \delta$.

For the second part,

$$\mathbf{E}_{c \leftarrow p}[f(M(c))] = \int f(M(c)) dp(c) \leq \int f(M(c))(1+\delta) dq(c) = (1+\delta)\mathbf{E}_{c \leftarrow q}[f(M(c))].$$

□

Corollary 4. *Suppose $d_+(c_0, c + c_0) \leq \delta$ for all $c \in \{c_1, \ldots, c_T\}$. Then*

$$\mathrm{FPL}(1..T) \leq \mathrm{BPL}(1..T) + \delta RT.$$

Suppose $d_\times(c_0, c + c_0) \leq \delta$ for all $c \in \{c_1, \ldots, c_T\}$. Moreover suppose $c \cdot x \geq 0$ for all $c \in \{c_1, \ldots, c_T\}$ and $x \in \mathscr{S}$. Then

$$\mathrm{FPL}(1..T) \leq (1 + \delta) \cdot \mathrm{BPL}(1..T).$$

Proof. Recall that $\mathrm{FPL}(1..T)$ is the expected value of $\sum_{t=1}^T c_t \cdot M(c_{0..t-1})$ and $\mathrm{BPL}(1..T)$ is the expected value of $\sum_{t=1}^T c_t \cdot M(c_{0..t})$. The corollary follows by using Lemma 3 to compare the sums term-by-term. □

Lemma 5. *Let c be any vector such that $\|c\|_1 \leq A$. If c_0 is a random sample from distribution α, then*

$$d_+(c_0, c + c_0) \leq \varepsilon A.$$

If c_0 is a random sample from distribution μ,

$$d_\times(c_0, c + c_0) \leq e^{\varepsilon A} - 1.$$

Proof. Assume without loss of generality that every component of the vector c is non-negative. Let $\mathbf{1}$ denote the vector $(1, 1, \ldots, 1)$. Let

$$c_0' = \begin{cases} c_0 & \text{if } \|c_0 - c\|_\infty \leq \frac{1}{\varepsilon} \\ \left(\frac{1}{\varepsilon}\right)\mathbf{1} + c - c_0 & \text{otherwise.} \end{cases}$$

It is an exercise to check that the random vector c_0' has the same distribution as $c + c_0$, and that $\Pr(c_0 \neq c_0') \leq \varepsilon A$. This proves the first part of the lemma.

The second part of the lemma follows from the calculation

$$\begin{aligned} d\mu(x + c) &= \left(\frac{\varepsilon}{2}\right)^n e^{-\varepsilon\|x+c\|_1} \\ &\leq \left(\frac{\varepsilon}{2}\right)^n e^{-\varepsilon\|x\|_1 + \varepsilon\|c\|_1} \\ &\leq d\mu(x) \cdot e^{\varepsilon A}. \end{aligned}$$

□

1.5 Bounding the expectation of $c_0 \cdot (x - y)$

Lemma 6. *If $\|x - y\|_1 \leq D$ and $\mathbf{E}[\|c_0\|_\infty] \leq M$ then*

$$\mathbf{E}[c_0 \cdot (x - y)] \leq DM.$$

Proof. The inequality $\|v\|_\infty \|w\|_1 \geq v \cdot w$ holds for all vectors v, w, because

$$v \cdot w = \sum v_i w_i \leq \sum |v_i||w_i| \leq \sum \|v\|_\infty |w_i| = \|v\|_\infty \|w\|_1.$$

The lemma follows by applying this identity with $v = c_0$ and $w = x - y$. □

Lemma 7. *If c_0 is sampled from distribution α,*

$$\mathbf{E}[\|c_0\|_\infty] \leq \frac{1}{\varepsilon}.$$

If c_0 is sampled from distribution μ,

$$\mathbf{E}[\|c_0\|_\infty] \leq O\left(\frac{\log n}{\varepsilon}\right)$$

provided that $n \geq 3$.

Proof. The first statement is obvious. The second statement is because the absolute values of the coordinates of c_0 are independent exponentially distributed random variables with mean $\frac{1}{\varepsilon}$. Such a random variable y satisfies

$$\mathbf{E}(y - r \,|\, y > r) = \frac{1}{\varepsilon}$$

for all $r > 0$. Letting $y_i = |(c_0)_i|$, we have

$$\mathbf{E}\left(y_i - \frac{\ln(n)}{\varepsilon} \,\middle|\, y_i > \frac{\ln(n)}{\varepsilon} \right) = \frac{1}{\varepsilon}.$$

It follows that

$$\begin{aligned}
\mathbf{E}(\|c_0\|_\infty) &= \mathbf{E}(\max_i y_i) \\
&= \frac{\ln(n)}{\varepsilon} + \mathbf{E}\left(\max_i \left(y_i - \frac{\ln(n)}{\varepsilon} \right) \right) \\
&\leq \frac{\ln(n)}{\varepsilon} + \sum_i \mathbf{E}\left(\max\left\{ y_i - \frac{\ln(n)}{\varepsilon}, 0 \right\} \right) \\
&= \frac{\ln(n)}{\varepsilon} + \sum_i \Pr\left(y_i > \frac{\ln(n)}{\varepsilon} \right) \mathbf{E}\left(y_i - \frac{\ln(n)}{\varepsilon} \,\middle|\, y_i > \frac{\ln(n)}{\varepsilon} \right) \\
&= \ln(n)/\varepsilon + n \cdot (1/n) \cdot (1/\varepsilon) \\
&\leq 2\ln(n)/\varepsilon
\end{aligned}$$

assuming $n \geq 3$. □

1.6 Putting it all together

For FPL_α, with $\varepsilon = \sqrt{D/RAT}$, we have

$$\begin{aligned}
\mathrm{FPL}_\alpha(1..T) &\leq \mathrm{BPL}_\alpha(1..T) + 2\varepsilon RAT \\
&\leq \mathrm{OPT}(1..T) + 2\varepsilon RAT + \frac{D}{\varepsilon} \\
&= \mathrm{OPT}(1..T) + O(\sqrt{DRAT}).
\end{aligned}$$

For FPL_μ, with any $\varepsilon \leq 1/A$, we have

$$\begin{aligned}
\mathrm{FPL}_\mu(1..T) &\leq e^{\varepsilon A}\mathrm{BPL}_\mu(1..T) \\
&\leq (1 + O(\varepsilon A))\mathrm{OPT}(1..T) + \frac{2\ln(n)D}{\varepsilon}.
\end{aligned}$$

1.7 Applying this to the best expert problem

We have seen that the best expert problem is a special case of online linear optimization. Here the cost vectors satisfy $\|c\|_\infty \leq 1$ hence $\|c\|_1 \leq n$, so applying the analysis above directly would lead to the bound

$$\mathrm{FPL}_\mu(1..T) \leq (1 + O(\varepsilon n))\mathrm{OPT}(1..T) + O(\frac{\ln(n)}{\varepsilon}).$$

By comparison, we know that Hedge satisfies a similar bound but with a factor of $1 + O(\varepsilon)$ instead of $1 + O(\varepsilon n)$. It turns out that we can prove that FPL_μ satisfies the same bound (up to constant factors) by a clever trick. Replace the sequence $c_1, c_2, \ldots, c_T$ with a new sequence $c'_1, c'_2, \ldots, c'_{nT}$ by expanding each vector c_i into n consecutive vectors $(c_i)_1\vec{e}_1, (c_i)_2\vec{e}_2, \ldots, (c_i)_n\vec{e}_n$. Note that the cost vectors in the new sequence have ℓ_1-norms bounded by 1, so

$$\mathrm{FPL}_\mu(1..nT) \leq (1 + O(\varepsilon))\mathrm{OPT}(1..nT) + O(\frac{\ln(n)}{\varepsilon}).$$

$\mathrm{OPT}(1..nT)$ on this new sequence is equal to $\mathrm{OPT}(1..T)$ on the old sequence. And

$$\mathbf{E}(\mathrm{FPL}_\mu(1..T)) \leq \mathbf{E}(\mathrm{FPL}_\mu(1..nT))$$

because the probability of incurring cost $(c_i)_k$ at time $(i-1)n + k$, when the cost vector is $(c_i)_k\vec{e}_k$, is at least as great as the probability that the algorithm incurred cost $(c_i)_k$ at time i in the original sequence. (The total past cost of choice k, relative to the alternatives, only looks better in the new sequence, because some of the alternatives have already been hit with additional costs.)

2 Online convex optimization

Now suppose $\mathscr{S} \subset \mathbb{R}^n$ is a bounded convex set, and suppose that we have a sequence of convex cost functions $f_1, f_2, \ldots, f_T$. In comparison with the online linear optimization problem studied in the preceding section, we are now making a more specific assumption about the set $\mathscr{S}$ (it is convex and bounded, not just bounded) and a less specific assumption about the functions f_t (they are convex, not necessarily linear).

We will make the following additional assumptions about $\mathscr{S}$ and the functions f_t.

1. $\|x - y\|_2 \leq D$ for all $x, y \in \mathscr{S}$.
2. $\|\nabla f\|_2 \leq A$ for all $f \in \{f_1, f_2, \ldots, f_T\}$.

For any point $z \in \mathbb{R}^n$, let $P(z)$ denote the closest point of $\mathscr{S}$, i.e.

$$P(z) = \arg\min_{x \in \mathscr{S}} \|x - z\|_2.$$

We have seen in an earlier lecture that the convexity of $\mathscr{S}$ implies that the set $\arg\min_{x\in\mathscr{S}} \|x-z\|_2$ consists of a single point, so the definition of $P(z)$ is unambiguous.

Zinkevich's algorithm produces a sequence of strategies $x_1, x_2, \ldots, x_T$ using a "gradient descent" algorithm. There is a predefined sequence of decreasing step sizes $\eta_1, \eta_2, \ldots$ and the algorithm starts with an arbitrary strategy x_1 and updates its strategy according to the rule:

$$x_{t+1} = P(x_t - \eta_t \nabla f_t(x_t)).$$

Note that this is a deterministic algorithm, in contrast to the primarily randomized learning algorithms we've seen up to this point. (The fact that we can get away with using a deterministic algorithm here is related to the fact that the strategy set $\mathscr{S}$ is convex, so instead of randomizing over a set of strategies we could simply play the strategy which is their weighted average.) Also note that we always move in the direction opposite the gradient of the *most recent* cost function. So the past history has no influence on the algorithm *except* that it influences the position of the current strategy vector x_t. It is sort of surprising that the algorithm still accomplishes a non-trivial learning task, given that it is keeping track of the past history in such an indirect way.

Theorem 8. *For every $x \in \mathscr{S}$, the sequence of strategies selected by Zinkevich's algorithm satisfies*

$$\sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{T} f_t(x) + \frac{D^2}{\eta_T} + \frac{1}{2}A^2 \sum_{t=1}^{T} \eta_t.$$

Proof. The analysis of the algorithm is inspired by the following informal train of thought. Time steps when $f_t(x_t) \leq f_t(x)$ are no problem. In a step when $f_t(x_t) > f_t(x)$, the gradient descent rule ensures that the algorithm will take a step which brings it closer to x. This suggests the following line of attack: define a potential function based on the distance from x. Show that every increase in the regret is offset by a corresponding decrease in the potential.

Consider the potential function $\Phi(y) = \|y - x\|_2^2$. For any z we have $\Phi(P(z)) \leq \Phi(z)$. This is clear when $z \in \mathscr{S}$ since $P(z) = z$. Otherwise, it follows from the fact that the triangle with vertices $z, P(z), x$ has an obtuse angle at $P(z)$; this is a fact that we worked out when studying the existence of Nash equilibria and again during the proof of Blackwell's approachability theorem.

Armed with this useful fact, let's evaluate the change in potential from time t to $t+1$.

$$\begin{aligned}
\Phi(x_{t+1}) - \Phi(x_t) &\leq \Phi(x_t - \eta_t \nabla f_t(x_t)) - \Phi(x_t) \\
&= \|(x_t - x) - \eta_t \nabla f_t(x_t)\|_2^2 - \|x_t - x\|^2 \\
&= -2\eta_t \nabla f_t(x_t) \cdot (x_t - x) + \eta_t^2 \|\nabla f_t(x_t)\|_2^2
\end{aligned}$$

By convexity of f_t, we have

$$\nabla f_t(x_t) \cdot (x_t - x) \geq f_t(x_t) - f_t(x).$$

Hence

$$\Phi(x_{t+1}) - \Phi(x_t) \leq -2\eta_t \left[f_t(x_t) - f_t(x)\right] + A^2\eta_t^2.$$

Rearranging terms,

$$f_t(x_t) - f_t(x) \leq \frac{\Phi(x_t) - \Phi(x_{t+1})}{2\eta_t} + \frac{1}{2}A^2\eta_t. \tag{3}$$

We now want to manipulate things so that when we sum over t, the $\Phi(\cdot)$ terms form a telescoping sum. The easiest way to accomplish this is to replace the first term on the right side of (3) with its upper bound $(\Phi(x_t) - \Phi(x_{t+1}))/(2\eta_T)$. (Here we are using the fact that the sequence $\eta_1, \eta_2, \ldots$ is decreasing.) Thus

$$\begin{aligned}
\sum_{t=1}^{T}[f_t(x_t) - f_t(x)] &\leq \frac{1}{2\eta_T}\sum_{t=1}^{T}[\Phi(x_t) - \Phi(x_{t+1})] + \frac{1}{2}A^2\sum_{t=1}^{T}\eta_t \\
&= \frac{1}{2\eta_T}[\Phi(x_1) - \Phi(x_{T+1})] + \frac{1}{2}A^2\sum_{t=1}^{T}\eta_t \\
&\leq \frac{D^2}{\eta_T} + \frac{1}{2}A^2\sum_{t=1}^{T}\eta_t.
\end{aligned}$$

□

Now we have to choose η_t to trade off the conflicting goals of making η_T large but making $\sum_{t=1}^{T}\eta_t$ small. The optimum trade-off (up to constant factors) is achieved by setting $\eta_t = \frac{D}{A}\sqrt{1/t}$ which leads to

$$\sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{T} f_t(x) + O\left(DA\sqrt{T}\right).$$

2.1 Dealing with concept drift

One of the great features of Zinkevich's algorithm is that it can deal with a limited amount of "concept drift," i.e. it does well not only against the benchmark of the best fixed strategy, but against a benchmark with a limited amount of power to change its strategy over time.

For a sequence $z_1, z_2, \ldots, z_T \in \mathscr{S}$ let $L(z_1, \ldots, z_T) = \sum_{t=1}^{T-1}\|z_{t+1} - z_t\|_2$. We have the following theorem.

Theorem 9. *Zinkevich's algorithm satisfies the following bound for any sequence* $z_1, z_2, \ldots, z_T$.

$$\sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{T} f_t(z_t) + \frac{D^2}{\eta_T} + \frac{2DL(z_1, \ldots, z_T)}{\eta_T} + \frac{A^2}{2} \sum_{t=1}^{T} \eta_t.$$

Proof. Define $\Phi(x, z) = \|x - z\|_2^2$. The argument is parallel to the argument given above, but with a couple of extra steps.

$$\begin{aligned} \Phi(x_{t+1}, z_{t+1}) - \Phi(x_t, z_t) &= [\Phi(x_{t+1}, z_{t+1}) - \Phi(x_{t+1}, z_t)] + [\Phi(x_{t+1}, z_t) - \Phi(x_t, z_t)] \\ &\leq [\Phi(x_{t+1}, z_{t+1}) - \Phi(x_{t+1}, z_t)] - 2\eta_t[f_t(x_t) - f_t(z_t)] + A^2\eta_t^2. \end{aligned}$$

The last line is derived from the line above it by using the technique from the proof of Theorem 8, with z_t in place of the variable labeled x in that proof.

To finish up, we must bound $\Phi(x_{t+1}, z_{t+1}) - \Phi(x_{t+1}, z_t)$. The easiest way to do this is to let $v = x_{t+1} - z_{t+1}$ and $w = x_{t+1} - z_t$ and observe that $\|v\|_2, \|w\|_2 \leq D$ hence:

$$\begin{aligned} \|v\|_2^2 - \|w\|_2^2 &= (v + w) \cdot (v - w) \\ &\leq \|v + w\|_2 \|v - w\|_2 \\ &\leq 2D\|z_{t+1} - z_t\|_2 \end{aligned}$$

Hence

$$\begin{aligned} \Phi(x_{t+1}, z_{t+1}) - \Phi(x_t, z_t) &\leq 2D\|z_{t+1} - z_t\|_2 - 2\eta_t[f_t(x_t) - f_t(z_t)] + A^2\eta_t^2 \\ f_t(x_t) - f_t(z_t) &\leq \frac{2D}{\eta_t}\|z_{t+1} - z_t\|_2 + \frac{\Phi(x_{t+1}, z_{t+1}) - \Phi(x_t, z_t)}{2\eta_t} + \frac{A^2}{2}\eta_t \\ &\leq \frac{2D}{\eta_T}\|z_{t+1} - z_t\|_2 + \frac{\Phi(x_{t+1}, z_{t+1}) - \Phi(x_t, z_t)}{2\eta_T} + \frac{A^2}{2}\eta_t \\ \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(z_t) &\leq \frac{2D}{\eta_T} L(z_1, \ldots, z_T) + \frac{D^2}{\eta_T} + \frac{A^2}{2} \sum_{t=1}^{T} \eta_t. \end{aligned}$$

□