CS292F Convex Optimization: Online Newton Method	**Spring 2020**

Lecture 16: May 20

Lecturer: Yu-Xiang Wang *Scribes: Xuandong Zhao*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

16.1 Universal Portfolio

Here we consider a model of stock market. We have a repeated investing scenario: for $t = 1, 2, \cdots$, there is a vector of prices ratio which we will call $\mathbf{r}_t \in \mathbb{R}_+^n$, where $\mathbf{r}_t(i) = \frac{\text{price of stock } i \text{ at time } t}{\text{price of stock } i \text{ at time } t-1} = \frac{\text{Price}_t(i)}{\text{Price}_{t-1}(i)}$. We define $\mathbf{x}_t \in \Delta_n$ is the asset allocation on N stocks. And we have the update $W_{t+1} = W_t \cdot \left(\mathbf{r}_t^\top \mathbf{x}_t\right)$ The total wealth after T days is:

$$W_T = W_1 \cdot \prod_{t=1}^{T} (\mathbf{r}_t^\top \mathbf{x}_t)$$

So we have:

$$\log(\frac{W_T}{W_1}) = \sum_{t=1}^{T} \log(\mathbf{r}_t^\top \mathbf{x}_t) = -\sum_{t=1}^{T} f_t(\mathbf{x}_t),$$

where $f_t^{(x)} = -\log(r_t^\top x_t)$.

The total regret is:

$$\text{Regret} = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum f_t(\mathbf{u}) \tag{16.1}$$

For example, consider a market of two stocks that fluctuate wildly. The first stock increases by 100% every even day and returns to its original price the following (odd) day. The second stock does exactly the opposite: decreases by 50% on even days and rises back on odd days. Formally, we have

$$\mathbf{r}_t(1) = \left(2, \frac{1}{2}, 2, \frac{1}{2}, \ldots\right)$$

$$\mathbf{r}_t(2) = \left(\frac{1}{2}, 2, \frac{1}{2}, 2, \ldots\right)$$

Clearly, any investment in either of the stocks will not gain in the long run. However, the portfolio $\mathbf{u} = [0.5, 0.5]$ increases the wealth by a factor of 1.25 daily because:

$$W_T = W_1 \cdot \sum f_t(\mathbf{u}) = W_1 \cdot \left(\frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 0.5\right)^\top = W_1 \cdot 1.25^\top.$$

Such a mixed distribution is called a fixed rebalanced portfolio, as it needs to rebalance the proportion of total capital invested in each stock at each iteration to maintain this fixed distribution strategy.

Thus, vanishing average regret guarantees long-run growth as the best constant rebalanced portfolio in hindsight. Such a portfolio strategy is called **universal portfolio**. We have seen that the online gradient descent algorithm gives essentially a universal algorithm with regret $O(\sqrt{T})$. Can we get better regret guarantees?

16.2 Exponential Concavity

For OGD, we have Regret $= GD\sqrt{T} = O(\sqrt{nT})$. For FTRL (entropy regularizer), we have Regret $= O(\sqrt{T \log n})$. We want to have better regret bounds. Recall that when f_t is m-strongly convex, OGD is $\frac{G^2}{m} \log T$ regret guarantee. However, $f_t = -\log(\mathbf{r}_t^\top \mathbf{x})$ is not strongly convex. So we want to introduce **exponential concavity**.

Definition 16.1. *A convex function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is defined to be α-exp-concave over $\mathcal{K} \subseteq \mathbb{R}^n$ if the function g is concave, where $g : \mathcal{K} \mapsto \mathbb{R}$ is defined as*

$$g(\mathbf{x}) = e^{-\alpha f(\mathbf{x})}$$

Lemma 16.2. *A twice-differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is α-exp-concave at $\mathbf{x}$ if and only if*

$$\nabla^2 f(\mathbf{x}) \succeq \alpha \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top$$

It is similar as strong convexity: $\nabla^2 f(\mathbf{x}) \succeq m \cdot I$.

For example:

$$f(x) = -\log(\mathbf{r}_t^\top \mathbf{x})$$

$$\nabla \left(-\log(\mathbf{r}_t^\top \mathbf{x})\right) = -\frac{\mathbf{r}_t}{\mathbf{r}_t^\top \mathbf{x}}$$

$$\nabla^2 \left(-\log(\mathbf{r}_t^\top \mathbf{x})\right) = -\frac{0 - \mathbf{r}_t \mathbf{r}_t^\top}{(\mathbf{r}_t^\top \mathbf{x})^2} = \frac{\mathbf{r}_t \mathbf{r}_t^\top}{(\mathbf{r}_t^\top \mathbf{x})^2} = \nabla \cdot \nabla^\top$$

$$\nabla^2 \succeq 1 \cdot \nabla \cdot \nabla^\top \Rightarrow \text{1-exp-concave}$$

Another example: f is m-strongly convex in domain $\mathcal{K}$, s.t. $\|\nabla f\|_2 \le G$. We have:

$$\nabla f \nabla f^\top \preceq G^2 I \preceq \frac{G^2}{m} \nabla^2 f, \text{for all } \mathbf{x} \in \mathcal{K}$$

The first bound is for G-Lipschitz and the second bound is for m-strongly convexity. Then we get

$$\nabla^2 f \succeq \frac{m}{G^2} \nabla f \nabla f^\top \Rightarrow f \text{ is } \frac{m}{G^2}\text{-exp-concave.}$$

Lemma 16.3. *Let $f : \mathcal{K} \to \mathbb{R}$ be an α-exp-concave function, and D, G denote the diameter of $\mathcal{K}$ and a bound on the (sub)gradients of f respectively. The following holds for all $\gamma \le \frac{1}{2} \min\left\{\frac{1}{4GD}, \alpha\right\}$ and all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$*

$$f(\mathbf{x}) \ge f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{\gamma}{2} (\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$$

Proof. We do the proof based on [1]. Since $\exp(-\alpha f(\mathbf{x}))$ is concave and $2\gamma \le \alpha$ by definition, it follows from Lemma 16.2 that the function $h(\mathbf{x}) \triangleq \exp(-2\gamma f(\mathbf{x}))$ is also concave. Then by the concavity of $h(\mathbf{x})$

$$h(\mathbf{x}) \le h(\mathbf{y}) + \nabla h(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$$

Plugging in $\nabla h(\mathbf{y}) = -2\gamma \exp(-2\gamma f(\mathbf{y}))\nabla f(\mathbf{y})$ gives

$$\exp(-2\gamma f(\mathbf{x})) \leq \exp(-2\gamma f(\mathbf{y})) \left[1 - 2\gamma \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})\right]$$

Simplifying gives

$$f(\mathbf{x}) \geq f(\mathbf{y}) - \frac{1}{2\gamma} \log\left(1 - 2\gamma \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})\right)$$

Next, note that $\left|2\gamma \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})\right| \leq 2\gamma GD \leq \frac{1}{4}$ and that for $|z| \leq \frac{1}{4}$,$-\log(1-z) \geq z + \frac{1}{4}z^2$. Applying the inequality for $z = 2\gamma \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$ implies the lemma. □

16.3 Online Newton step

Algorithm 1 online Newton step

Input: convex set $\mathcal{K}, T, \mathbf{x}_1 \in \mathcal{K} \subseteq \mathbb{R}^n$, parameters $\gamma, \varepsilon > 0, A_0 = \varepsilon \mathbf{I}_n$
for $t = 1$ to T **do**
  Play $\mathbf{x}$
  Observe loss $f_t(\mathbf{x}_t)$, receive $\nabla_t = \nabla f_t(\mathbf{x}_t)$
  Rank-1 update: $A_t = A_{t-1} + \nabla_t \nabla_t^\top$
  Newton step and projection:

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \tfrac{1}{\gamma} A_t^{-1} \nabla_t$$
$$\mathbf{x}_{t+1} = \prod\nolimits_{\mathcal{K}}^{A_t} (\mathbf{y}_{t+1}) = \mathrm{argmin}_{x \in \mathcal{K}} \|\mathbf{y}_{t+1} - \mathbf{x}\|_{A_t}^2$$

return

Theorem 16.4. *Algorithm 1 with parameters* $\gamma = \min\left\{\frac{1}{4GD}, \alpha\right\}$ *and* $\epsilon = \frac{1}{\gamma^2 D^2}$, *guarantees for* $(T > 4)$

$$\mathrm{Regret}_T \leq 5\left(\frac{1}{\alpha} + GD\right) n \log T$$

To prove Theorem 16.4 we begin by proving the following:

Lemma 16.5. *The regret of online Newton step (with appropriate choice of parameters) is bounded by*

$$\mathrm{Regret}_T \leq 4\left(\frac{1}{\alpha} + GD\right)\left(\sum_{t=1}^{T} \nabla_t A_t^{-1} \nabla_t + 1\right)$$

Proof. Let $\mathbf{x}^* \in \mathcal{K}$ be the best decision in hindesight. By Lemma 16.3 we have for our choice of γ :

$$f_t(\mathbf{x}^*) \geq f_t(\mathbf{x}_t) + \nabla_t^\top (\mathbf{x}^* - \mathbf{x}_t) + \frac{\gamma}{2} (\mathbf{x}^* - \mathbf{x}_t)^\top \nabla_t \nabla_t^\top (\mathbf{x}^* - \mathbf{x}_t)$$

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*) \leq \nabla_t^\top (\mathbf{x}_t - \mathbf{x}^*) - \frac{\gamma}{2} (\mathbf{x}^* - \mathbf{x}_t)^\top \nabla_t \nabla_t^\top (\mathbf{x}^* - \mathbf{x}_t) \tag{16.2}$$

By the update rule of $\mathbf{x}_{t+1} = \prod_{\mathcal{K}}^{A_t} (\mathbf{x}_t - \frac{1}{\gamma} A_t^{-1} \nabla_t)$

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_{A_t}^2 &\leq \|\mathbf{y}_{t+1} - \mathbf{x}^*\|_{A_t}^2 \\
&= \left\|\mathbf{x}_t - \frac{1}{\gamma} A_t^{-1} \nabla_t - \mathbf{x}^*\right\|_{A_t}^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_{A_t}^2 + \frac{1}{\gamma^2} \left\|A_t^{-1} \nabla_t\right\|_{A_t}^2 - \frac{2}{\gamma} (\mathbf{x}_t - \mathbf{x}^*)^\top A_t A_t^{-1} \nabla_t
\end{aligned}$$

By move it around:

$$\frac{2}{\gamma}\left(\mathbf{x}_t-\mathbf{x}^*\right)^\top \nabla_t \le \frac{1}{\gamma^2}\left\|A_t^{-1}\nabla_t\right\|_{A_t}^2+\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_{A_t}^2-\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_{A_t}^2$$

Multiply $\frac{\gamma}{2}$ on both sides and sum over $t = 1, 2, \ldots, T$

$$\begin{aligned}\sum_{t=1}^{T}\left(\mathbf{x}_t-\mathbf{x}^*\right)^\top \nabla_t &\le \frac{1}{2\gamma}\sum_t\|\nabla_t\|_{A_t^{-1}}^2+\frac{\gamma}{2}\left\|\mathbf{x}_1-\mathbf{x}^*\right\|_{A_1}^2+\frac{\gamma}{2}\sum_{t=2}^{T}(\mathbf{x}_t-\mathbf{x}^*)^\top(A_t-A_{t-1})(\mathbf{x}_t-\mathbf{x}^*)-\frac{\gamma}{2}\left\|\mathbf{x}_{T+1}-\mathbf{x}^*\right\|_{A_T}^2\\ &\le \frac{1}{2\gamma}\sum_t\|\nabla_t\|_{A_t^{-1}}^2+\frac{\gamma}{2}\left\|\mathbf{x}_1-\mathbf{x}^*\right\|_{A_1}^2+\frac{\gamma}{2}\sum_{t=2}^{T}(\mathbf{x}_t-\mathbf{x}^*)^\top\nabla_t\nabla_t^T(\mathbf{x}_t-\mathbf{x}^*)+0\end{aligned}$$

for $A_t = A_{t-1} + \nabla_t \nabla_t^T$ and dropping the negative term.

Plug into equation (16.2)

$$\begin{aligned}\sum_{t=1}^{t}\left(f_t\left(\mathbf{x}_t\right)-f_t\left(\mathbf{x}^*\right)\right) &\le \frac{1}{2\gamma}\sum_t\|\nabla_t\|_{A_t^{-1}}^2+\frac{\gamma}{2}\left\|\mathbf{x}_1-\mathbf{x}^*\right\|_{A_1}^2+\frac{\gamma}{2}\sum_{t=2}^{T}(\mathbf{x}_t-\mathbf{x}^*)^\top\nabla_t\nabla_t^T(\mathbf{x}_t-\mathbf{x}^*)-\frac{\gamma}{2}\sum_{t=1}^{T}\left(\mathbf{x}_t-\mathbf{x}^*\right)^\top\nabla_t\nabla_t^\top\left(\mathbf{x}_t-\mathbf{x}^*\right)\\ &= \frac{1}{2\gamma}\sum_t\|\nabla_t\|_{A_t^{-1}}^2+\frac{\gamma}{2}(\mathbf{x}_1-\mathbf{x}^*)^\top A_1(\mathbf{x}_1-\mathbf{x}^*)-\frac{\gamma}{2}(\mathbf{x}_1-\mathbf{x}^*)^\top\nabla_t\nabla_t^T(\mathbf{x}_1-\mathbf{x}^*)\\ &= \frac{1}{2\gamma}\sum_t\|\nabla_t\|_{A_t^{-1}}^2+\frac{\gamma}{2}(\mathbf{x}_1-\mathbf{x}^*)^\top(A_1-\nabla_t\nabla_t^T)(\mathbf{x}_1-\mathbf{x}^*)\\ (A_1=\varepsilon I+\nabla_1\nabla_1^T) &= \frac{1}{2\gamma}\sum_t\|\nabla_t\|_{A_t^{-1}}^2+\frac{\gamma}{2}\varepsilon\left\|\mathbf{x}_1-\mathbf{x}^*\right\|_2^2\end{aligned}$$

For $\varepsilon = \frac{1}{\gamma^2 D^2}$, $\frac{\gamma}{2}\varepsilon\left\|\mathbf{x}_1-\mathbf{x}^*\right\|_2^2 \le \frac{\gamma}{2}\varepsilon D^2$. Then by $\gamma = \frac{1}{2}\min(\alpha, \frac{1}{4GD})$,

$$\begin{aligned}\sum_{t=1}^{t}\left(f_t\left(\mathbf{x}_t\right)-f_t\left(\mathbf{x}^*\right)\right) &\le \frac{1}{2\gamma}\left(\sum_t\|\nabla_t\|_{A_t^{-1}}^2+1\right)\\ &\le 4\left(\frac{1}{\alpha}+GD\right)\left(\sum_{t=1}^{T}\nabla_t A_t^{-1}\nabla_t+1\right)\end{aligned}$$

$\square$

Proof. of the main Theorem (Theorem 16.4): First we show that the term $\sum_{t=1}^{T}\nabla_t^\top A_t^{-1}\nabla_t$ tis upper bounded by a telescoping sum. Notice that

$$\begin{aligned}
\|\nabla_t\|^2_{A_t^{-1}} &= \nabla_t^T A_t^{-1} \nabla_t \\
&= \operatorname{tr}(\nabla_t^T A_t^{-1} \nabla_t) \\
&= \operatorname{tr}(A_t^{-1} \nabla_t \nabla_t^T) \\
&= \operatorname{tr}\left(A_t^{-1}(A_t - A_{t-1})\right) \\
&= \operatorname{tr}(I - A_t^{-1} A_{t-1}) \\
&= \sum_i \left(1 - \lambda_i(A_t^{-1} A_{t-1})\right) \\
&\leq \sum_{i=1}^n \log(\lambda_i^{-1}(A_t^{-1} A_{t-1})) \\
&= \log\left(\prod_{i=1}^n \lambda_i^{-1} A_t^{-1} A_{t-1}\right) \\
&= \log\left|(A_t^{-1} A_{t-1})^{-1}\right| \\
&= \log \frac{|A_t|}{|A_{t-1}|} \\
&= \log|A_t| - \log|A_{t-1}|
\end{aligned}$$

where $\lambda_i(\cdot)$ denotes the ith eigenvalues of a matrix and the inequality is by $\varepsilon \leq \log\left(\frac{1}{1-\varepsilon}\right)$ and take $\varepsilon = 1 - \lambda_i(A_t^{-1} A_{t-1})$. Recall that by the Jordan canonical form of a matrix X, $\operatorname{tr} X = \sum_i \lambda_i(X)$ also the determinant $|X| = \prod_i \lambda_i(X)$. Notice that λ_i will be complex if A is not symmetric, but the traces and determinants will be real numbers.

Sum up $t = 1, 2, \dots, T$, we have:

$$\sum_{t=1}^T \|\nabla_t\|^2_{A_t^{-1}} = \log|A_T| - \log|A_0|$$

Since $A_T = \sum_{t=1}^T \nabla_t \nabla_t^\top + \varepsilon I_n$ and $\|\nabla_t\| \leq G$, the largest eigenvalue of A_T is at most $TG^2 + \varepsilon$. Hence the determinant of A_T can be bounded by $|A_T| \leq \left(TG^2 + \varepsilon\right)^n$. Hence recalling that $\varepsilon = \frac{1}{\gamma^2 D^2}$ and $\gamma = \frac{1}{2}\min\left\{\frac{1}{4GD}, \alpha\right\}$ for $T > 4$,

$$\sum_{t=1}^T \|\nabla_t\|^2_{A_t^{-1}} = \sum_{t=1}^T \nabla_t^\top A_t^{-1} \nabla_t \leq \log\left(\frac{TG^2 + \varepsilon}{\varepsilon}\right)^n \leq n \log\left(TG^2 \gamma^2 D^2 + 1\right) \leq n \log T$$

Plugging into Lemma 16.5 we obtain the main Theorem 16.4.

$$\text{Regret}_T \leq 5\left(\frac{1}{\alpha} + GD\right)(n \log T + 1)$$

□

16.4 Running Time

Every iteration of the online Newton step requires the computation of the matrix A_t^{-1}, which usually in time $O(n^3)$. However, by the Sherwin-Morrison Woodbury (matrix inversion lemma)[2],

$$\left(A + \mathbf{x}\mathbf{x}^\top\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{x}\mathbf{x}^\top A^{-1}}{1 + \mathbf{x}^\top A^{-1} \mathbf{x}}$$

given A_{t-1}^{-1} and ∇_t one can compute A_t^{-1} in time $O\left(n^2\right)$ using only matrix-vector and vector-vector products. The online Newton step algorithm requires $O(n^2)$ space to store the matrix A_t^{-1}. If making projections onto $\mathcal{K}$ can be solved in $O(n^2)$ time (this may or may not be the case...), the algorithm can be implemented in time and space $O(n^2)$.

Beware of the numerical instability of SMW identity though, if you are to implement this trick in practice.

16.5 Additional thoughts?

You get either $\sqrt{T \log n}$ (from the multiplicative weights updates, or FTRL with entropy regularization) or $n \log T$ (from online Newton step.)

Is it possible to get $\operatorname{poly}\log(n, T)$ regret for this problem?

References

[1] Hazan, Elad "Introduction to online convex optimization." *arXiv preprint arXiv:1909.05207 (2019).*

[2] Sherman, Jack, and Winifred J. Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix." The Annals of Mathematical Statistics 21.1 (1950): 124-127.