CS 7545: Machine Learning Theory **Fall 2018**

Lecture 12: Online Convex Optimization (Cont.)

Lecturer: Jacob Abernethy *Scribes: Adrien Saremi and Bhavesh Khamesra*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

12.0.1 Review:

In the last class, we saw the general algorithm for an online convex optimization method.

Algorithm 1: Let $K \subseteq \mathbb{R}^d$ be a convex decision set.

```
for $t = 1...T$
    Algorithm selects $x_t \in K$
    Nature chooses convex function $f_t : K \to \mathbb{R}$
```

$$\mathrm{Reg_T} := \sum_{t=1}^{T} f_t(x_t) - \min_{x \in K} \sum_{t=1}^{T} f_t(x)$$

Examples:

1. Boosting
2. Statistical Problems: Batch (IID) framework

12.0.2 Batch IID Framework

Let us now look at the Batch framework in more detail. Given:

- Independent and identically distributed dataset $(x_1, y_1), ..., (x_T, y_T) \in D$ s.t. $x_t \in X$ and $y_t \in Y$.
- Hypothesis space $\mathcal{H}$ with functions which map elements of X to Y
- Loss Function: $\ell : \mathcal{H} \times X \times Y \to \mathbb{R}$.

Assumption: Let $\mathcal{H} = \{h_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ and $\ell(h_\theta, (x, y))$ be a convex function[1] in θ. For eg. consider a set of neural networks models forming the hypothesis space $\mathcal{H}$ and θ can be their parameters like learning rate, number of hidden layers, etc.

Objective: Let $\mathcal{L}(h) = \mathbb{E}_{(x,y)\sim D}\left[\ell\left(h, (x, y)\right)\right]$. We want to find some $\hat{h}_t$ [2] s.t. $\mathcal{L}(\hat{h}_t) - \min_{h \in \mathcal{H}} \mathcal{L}(h) < \epsilon$. The overall quantity on the LHS is called "excess risk“.

Classical Algorithm: Empirical Risk Management - Pick $\hat{h}_T$ by minimizing h over empirical sample i.e.

$$\hat{h}_T = \arg\min_{h \in \mathcal{H}} \frac{1}{T} \sum_{t=1}^{T} \ell\left(h, (x, y)\right)$$

[1] In realistic scenarios, this may not always be true i.e. $\ell\left(h_\theta, (x, y)\right)$ may not be convex function

[2] Quantities denoted by ˆ are functions of empirical data i.e. are an estimate

Under previous assumption, we can write an algorithm for which excess risk is less than regret.

Alternate Algorithm: Stochastic Online Convex Optimization -

Algorithm 2:

```
For t = 1...T:
    $\theta_t$ chosen using the online convex optimization algorithm
    $(x_t, y_t)$ revealed
    $f_t(\theta_t) = l\left(\theta_t, (x_t, y_t)\right)$

Output: $\hat{h} = h_{\frac{1}{T}\sum_{t=1}^{T}\theta_t}$
```

For $\hat{h}$ defined by the output above, we make the following claim:

Claim 12.1

$$\mathcal{L}(\hat{h}) - \min_{h^* \in \mathcal{H}} \mathcal{L}(h^*) \leq \mathbb{E}\left[\frac{\text{Reg}_T}{T}\right]$$

Proof: Let $\bar{\theta_T} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$. Then,

$$\begin{aligned}\mathcal{L}(\hat{h}) &= \mathcal{L}\left(h_{\frac{1}{T}\sum_{t=1}^{T}\theta_t}\right) = \mathcal{L}\left(h_{\bar{\theta_T}}\right)\\ &= \mathbb{E}_{(x,y)\in D}\left[\ell\left(h_{\hat{\theta}_T}, (x,y)\right)\right]\\ &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{(x,y)\in D}\left[\ell\left(h_{\theta}, (x,y)\right)\right]\end{aligned}$$

where we used the Jensen's inequality in the last step. To move forward, notice that x, y are IID variables and our algorithm cannot distinguish between the input data (be it training or testing data). Hence, the expectation would remain unchanged for testing and training datasets. Thus, replacing $(x, y) \to (x_t, y_t)$:

$$\mathcal{L}\left(\hat{h}\right) \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{(x_t,y_t)\in D}\left[\ell\left(h_{\theta_t}, (x_t, y_t)\right)\right]$$

Note that loss function is defined as $f_t(\theta_t) = l\left(\theta_t, (x_t, y_t)\right)$. Substituting in above equation, we get

$$\begin{aligned}\mathcal{L}\left(\hat{h}\right) &\leq \frac{1}{T}\mathbb{E}_{(x_t,y_t)\in D}\sum_{t=1}^{T}\left[f_t\left(\theta_t\right)\right]\\ &\leq \frac{1}{T}\mathbb{E}_{(x_t,y_t)\in D}\left[\min_{\theta\in\Theta}\sum_{t=1}^{T}f_t\left(\theta\right)\right] + \frac{\text{Reg}_T}{T}\\ &\leq \frac{1}{T}\min_{\theta\in\Theta}\mathbb{E}_{(x_t,y_t)\in D}\left[\sum_{t=1}^{T}\ell\left(h_\theta, x_t, y_t\right) + \text{Reg}_T\right]\end{aligned}$$

Since θ is now chosen to the argument where RHS is minimum and again using the fact (x_t, y_t) is IID,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\ell\left(h_\theta, (x_t, y_t)\right)\right] = \mathbb{E}\left[\ell\left(h_\theta, (x, y)\right)\right] \tag{12.1}$$

Substituting this in RHS of previous equation, we get:

$$\mathcal{L}(\hat{h}) \leq \min_{\theta\in\Theta} \mathbb{E}_{(x,y)\in D}\left[\ell\left(h_\theta,(x,y)\right)\right] + \mathbb{E}_{(x_t,y_t)\in D}\left[\frac{\text{Reg}_T}{T}\right] \tag{12.2}$$

$$\mathcal{L}(\hat{h}) \leq \min_{\theta\in\Theta} \mathcal{L}\left(h_\theta\right) + \mathbb{E}_{(x_t,y_t)\in D}\left[\frac{\text{Reg}_T}{T}\right] \tag{12.3}$$

■

Follow the Leader (FTL):

Recap: $x_t = \arg\min_{x\in K}\sum_{s=1}^{t-1} f_s(x)$

Claim 12.2

$$\text{Reg}_\text{T}(\text{FTL}) \leq \sum_{t=1}^{T}\left(f_t(x_t) - f_t(x_{t+1})\right)$$

Proof: To prove this, we use the method of induction. Consider the first time iteration:

$$\begin{aligned}\text{Reg}_\text{T}(\text{FTL}) &= f_1(x_1) - f_1(x^*) = f_1(x_1) - \min_{x\in K}\sum_{s=1}^{1} f_x(x)\\ &= f_1(x_1) - f_1(x_2)\end{aligned}$$

So, the base case holds. Now consider the case of $T > 1$ iterations, assuming the claim is true for $T-1$. Then, we have:

$$\begin{aligned}\text{Reg}_\text{T}(\text{FTL}) &= \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x_{T+1})\\ &= \sum_{t=1}^{T-1}\Big(f_t(x_t) - f_t(x_{T+1})\Big) + f_T(x_T) - f_T(x_{T+1})\end{aligned}$$

Since $x_t = \arg\min_{x\in K}\sum_{s=1}^{t-1} f_s(x)$, this implies $\sum_{t=1}^{T-1} f_t(x_T) \leq \sum_{t=1}^{T-1} f_t(u)\ \ \forall u \in K$. Setting $u = x_{T+1}$, we get$\sum_{t=1}^{T-1} f_t(x_T) \leq \sum_{t=1}^{T-1} f_t(x_{T+1})$. Using this relation in above equation, we get

$$\begin{aligned}\text{Reg}_\text{T}(\text{FTL}) &\leq \sum_{t=1}^{T-1}\left(f_t(x_t) - f_t(x_T)\right) + f_T(x_T) - f_T(x_{T+1})\\ &\leq \text{Reg}_{\text{T}-1}(\text{FTL}) + f_t(x_T) - f_t(x_{T+1})\end{aligned}$$

Using the induction, we can recursively expand the above inequality to get

$$\begin{aligned}\text{Reg}_\text{T}(\text{FTL}) &\leq \sum_{t=1}^{T-1}\left(f_t(x_t) - f_t(x_{t+1})\right) + f_T(x_T) - f_T(x_{T+1})\\ &= \sum_{t=1}^{T}\left(f_t(x_t) - f_t(x_{t+1})\right)\end{aligned}$$

■

Gaussian Density Estimation:

Let us look at an example of FTL. Consider the following algorithm:

Algorithm 3:

- Data z_t are revealed one by one
- Algorithm predicts μ_t, the mean of the data
- Loss function:$f_t(\mu_t) = \|\mu_t - z_t\|_2^2$

To choose the mean, we use FTL and define mean as $\mu_t = \arg\min_\mu \sum_{s=1}^{t-1} \frac{1}{2} \|\mu - z_s\|_2^2$. Using the first derivative test, we can compute the minimum of function on RHS and after little algebraic manipulation, we get $\mu_t = \frac{1}{t-1} \sum_{s=1}^{t-1} z_s = \bar{z}_{t-1}$

Let us analyze the performance of this algorithm. We start with the definition of regret function:

$$\begin{aligned}\mathrm{Reg_T} &\leq \sum_{t=1}^{T} \left(\|\mu_t - z_t\|_2^2 - \|\mu_{t+1} - z_t\|_2^2 \right) \\ &\leq \sum_{t=1}^{T} \left(\|\bar{z}_{t-1} - z_t\|_2^2 - \|\bar{z}_t - z_t\|_2^2 \right)\end{aligned}$$

Let us focus on second term and try to write in is terms of first term.

$$\begin{aligned}\bar{z}_t &= \frac{1}{t}\sum_{s=1}^{t} z_s = \frac{t-1}{t}\bar{z}_{t-1} + \frac{1}{t}z_t \\ \implies \bar{z}_t - z_t &= \left(1 - \frac{1}{t}\right)(\bar{z}_{t-1} - z_t)\end{aligned}$$

Substituting this in expression of regret, we get

$$\mathrm{Reg_T} \leq \sum_{t=1}^{T} \|\bar{z}_{t-1} - z_t\|_2^2 - \left(1 - \frac{1}{t}\right)^2 \|\bar{z}_{t-1} - z_t\|_2^2$$

Dropping the $1/t^2$ term:

$$\begin{aligned}&\leq \sum_{t=1}^{T} \frac{2}{t} \|\bar{z}_{t-1} - z_t\|_2^2 \\ &\leq 8D \left(\sum_{t=1}^{T} \frac{1}{t} \right) \\ &\leq 8D \log T\end{aligned}$$

In second last step, we use $\|z_t\|_2^2 < D$. This result is much better then Online gradient descent whose performance is $O(\sqrt{T})$. This is because the loss function in this case is strongly convex.

Fact: If f_t is α_t-strongly convex, then

$$\mathrm{Reg}_{\mathrm{T}}(\mathrm{FTL}) \leq O\left(\sum_{t=1}^{T} \frac{1}{A_{t-1}}\right)$$

where $A_t = \sum_{s=1}^{t} \alpha_s$.