Lecture 17 - 11-05-2020

1.1 Strongly convex loss functions

We will saw with OGD but we will see Support Vector Machine(SVM). Very popular learning model.
We will see SVM next to see the part of linear predictor and also speak about Kernel function used with linear predictor to obtain non -linear classifier from a linear classifier.

ℓ is σ-SC if $\quad \forall u, w$:

$$\ell(w) - \ell(u) \leq \nabla \ell(w)^T (w - u) - \frac{\sigma}{2} \| w - u\|^2$$

1.1.1 OGD for Strongly Convex losses

Init: $w_1 = (0, ..., 0)$
For $t = 1, 2...$
$w_{t+1} = w_t - \frac{1}{\sigma t} \nabla \ell_t(w_t) \qquad \eta_t = \frac{1}{\eta t}$
(no projection steps)

$$\ell_t(w_t) - \ell_t(u) \leq \nabla \ell_t(w_t)^T (w - u) - \frac{\sigma}{2} \|w_t - u\|^2 \; =$$

$$= \; -\frac{1}{\eta_t} (w_{t+1} - w_t)^T (w_t - u) - \frac{\sigma}{2} \|w_t - u\|^2 \; =$$

$$= \frac{1}{\eta_t} \left(\frac{1}{2} \|w_t - u\|^2 - \frac{1}{2} \|w_{t+1} - u\|^2 + \frac{1}{2} \|w_{t+1} - w_t\|^2 \right) - \frac{\sigma}{2} \|w_t - u\|^2$$

$$R_T(u) \; \leq \; \frac{1}{2\,\eta_1} \|w_1 - u\|^2 - \frac{1}{2\,\eta_{T+1}} \|w_{T+1} - u\|^2 - \frac{\sigma}{2} \|w_1 - u\|^2 +$$

$$+ \; \frac{1}{2} \sum_{t=1}^{T-1} \|w_{t+1} - u\|^2 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \sigma \right) + \frac{1}{2} \|w_{T+1} - u\|^2 \left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_T} \right) + \frac{G^2}{2} \sum_{t=1}^{T} \eta_t$$

where red terms cancel out, **blue** (sum) instead is 0 since $\sigma(t+1) - \sigma\, t - \sigma$

$$G = \max_t \|\nabla \ell_t(w_t)\|$$

$$R_T(U) \leq \frac{1}{2} (\sigma - \sigma) \|w_1 - u\|^2 + \frac{G^2}{2} \sum_{t=1}^{T} \frac{1}{\sigma\, t} \; =$$

$$R_T(U) \leq \frac{G^2}{2} \sum_{t=1}^{T} \frac{1}{\sigma\, t}$$

We know that $\sum_{t=1}^{T} \frac{1}{T} \leq \ln(T+1)$ so:

$$R_T(U) \;\leq\; \frac{G^2}{2\,\sigma} \ln(T+1)$$

$\frac{R_T(U)}{T}$ **vanishes at rate** $\frac{\ln T}{T} << \frac{1}{\sqrt{T}}$ **provided** $max_t \|\nabla \ell_t(w_t)\|$ **remains bounded**
We assume it in special case.

Where are these SC losses?
Minimising strongly convex version of standard convex losses helps a lot.
We will see how Regularitation imply Stability. Before studing SVM and stability we going to do something before.

1.1.2 Relate sequential risk and statistical risk

It is important: I have this algorith that control sequential risk and regret but I am also courious to use this algorithms.

We assume:
Data $(x_t, y_)$ drawm i.i.d. from fixed unknown D.
Convex loss function ℓ.
For example compare square loss and hinge loss(convex upper bound on 0-1 *loss*:

$$\ell(\hat{y}, y = (\hat{y}, y)^2 \qquad \ell(\hat{y}, y) = [1 - \hat{y}\, y]_+$$

We will focus on linear predictors $h(x) = f(w^T x)$ (easily to analise with OGD framework).
Risk $\ell_D(w) = \mathbb{E}\left[\, \ell(w^T X, Y) \,\right]$
where $\hat{y} = w^T X$
Assume we have a training set S of example $(X_1, Y_1)...(X_m, Y_m)$ (in maiusc since are random sequence of data point from a distribution)

$$Convex \quad \ell_t(w) = \ell(w^T X_t, Y_t) \qquad t = 1, ...m$$

Became a sequence of convex losses.
I run OGD on $\ell_1, \ell_2, ..., \ell_m$ and get $w_1, ..., w_m \quad \|w_t\| \leq U$
OGD projects onto:

$$\{U \in \mathbb{R}^d : \|u\| \leq U\} \qquad U^* = arg \min_{u:\|u\| \leq U} \ell_D(u)$$

where U^* is the best linear predictor in class.
So i take a bunch of predictors but i need one, so I take the average of those (since the expected value is convex):

$$\bar{w} = \frac{1}{m} \sum_{t=1}^{m} w_t$$

I want to study the variance error:

$$\ell_D(\bar{w}) - \ell_D(u^*) \ ?$$

I am using Online Learning.
Using Jensen inequality:

$$\ell_D(\bar{w}) = \mathbb{E}\left[\ell(\bar{w}^T X, Y\right] \ \leq \ \mathbb{E}\left[\frac{1}{m} \sum_{t=1}^{m} \ell(w_t^T X, Y)\right] = \frac{1}{m} \sum_{t} \mathbb{E}\left[\ell(w_t^T X, Y)\right]$$

where $\mathbb{E}\left[\ell(w_t^T X, Y)\right]$ is equals to $\ell_D(w_t)$

$$\ell_D(\bar{w}) \leq \frac{1}{m} \sum_{t=1}^{n} \ell_D(w_t) \qquad \textbf{for any given training set } (x_1, y_1)...(x_m, y_m)$$

I want to look at the difference:

$$\ell_D(w_t) - \ell(w_t^T X_t, Y_t)$$

$\ell_D = \mathbb{E}\left[\, \ell(w_t^T X, Y) \,\right]$
Now I fix $t-1$ example in the training set $\quad (X_1, Y_1)...(X_{t-1}, Y_{t-1})$
w_t is **determined** by $(X_1, Y_1), ...(X_{t-1}, Y_{t-1})$
(X_t, Y_t) is distribuited like any $(X, Y) \sim D$

$$\mathbb{E}_{t-1}\left[\,\cdot\,\right] = \mathbb{E}\left[\,\cdot\,|(X_1, Y_1)...(X_{t-1}, Y_{t-1}\right] \qquad z_t = \ell_D(w_t) - \ell(w_t^T X_t, Y_t)$$

$$\frac{1}{m} \sum_{t=1}^{m} \mathbb{E}_{t-1}\left[Z_t\right] = 0$$

I want to show the average of $\ell_D(w_t)$ is equal to average of $\ell(w_t^T X_t, Y)$
I want to prove:

$$\frac{1}{m} \sum_{t=1}^{m} \ell_D(w_t) \ \leq \ \frac{1}{m} \sum_{t=1}^{m} \ell(w_t^T X_t, Y_t) + \sqrt{\frac{1}{m} \ln \frac{1}{\delta}} \qquad \textit{with high probability w.r.t. } S$$

where (red part) is the sequential risk of OGD.

$$\frac{1}{m}\sum_{t=1}^{m} Z_t \leq \sqrt{\frac{1}{m}\ln\frac{1}{\delta}} \qquad \textbf{with prob. at least } 1-\delta$$

I know that $\mathbb{E}_{t-1}[Z_t] = 0$

$$|Z_t| \in [0, M] \quad \Rightarrow \quad \frac{1}{m}\sum_{t=1}^{m} Z_t \leq M\sqrt{\frac{2}{m}\ln\frac{1}{\delta}} \quad w.p\ 1-\delta$$

Version of Chernoff-Hoffdiwg bounds for sums of dependent random variables.

$$\frac{1}{m}\sum_{t=1}^{m} \ell_D(w_t) \ \leq \ \frac{1}{m}\sum_{t=1}^{m} \ell_t(w_t) + M\sqrt{\frac{2}{m}\ln\frac{1}{\delta}} \quad w.p\ 1-\delta$$

This tells me that$\ell_D(\bar{w})$ is controlled by the sequential risk of OGD + $O\left(\frac{1}{\sqrt{m}}\right)$
Variance Error for $(w^T x - y)^2 \qquad \|x_t\| \leq X, \quad |y_t| \leq U\,X$

$$G = \max_t \|\nabla \ell_t(w_t)\| \leq 4\,(U\,X)^2$$

$$\ell_D(\bar{w}) \leq \min_{u:\|u\|\leq U} \frac{1}{m}\sum_{t=1}^{m} \ell_D(u) + 8\,(U\,X)^2\sqrt{\frac{2}{m}} + 4\,(U\,X)^2\sqrt{\frac{2}{m}\ln\frac{1}{\delta}}$$

where red is **OGD analysis**

$$\ell_D(\bar{w}) \ \leq \ \min \frac{1}{m}\sum_{t=1}^{m} \ell_t(u) + 12\,(U\,X)^2\sqrt{\frac{2}{m}\ln\frac{1}{\delta}} \quad \textit{with prob. } 1-\delta$$

By C-H bounds:

$$\textit{where} \quad \min \frac{1}{m}\sum_{t=1}^{m} \ell_t(u) \ \leq \ \frac{1}{m}\sum_{t=1}^{m} \ell_t(u^*) \ \leq \ \ell_D(u^*) + 4\,(U\,X)^2\sqrt{\frac{1}{2\,m}\ln\frac{1}{\delta}}$$

where the sum is the test error of u^*
At the end:

$$\ell_D(\bar{w}) \leq \ell_D(u^*) + 16\,(U\,X)^2\sqrt{\frac{1}{m}\ln\frac{1}{\delta}} \qquad w.p\ 1-\delta$$

Even with m large, I can run it since i bounded in the small "ball".