Machine Learning Theory (CS 6783)

Lecture 12: Online Convex Optimization/Learning

1 Online Convex Optimization Setting

For the purpose of this lecture let us modify the online learning protocol a bit (this can be done w.l.o.g.). First, Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, that is the instance space pair. Let $\mathcal{F}$ be a convex subset of a vector space. $\ell : \mathcal{F} \times \mathcal{Z} \mapsto \mathbb{R}$ is the loss function. For each $z \in \mathcal{Z}$ let $\ell(\cdot, z)$ be a convex function.

```
For $t = 1$ to $n$
    Learner picks $\hat{\mathbf{y}}_t \in \mathcal{F}$
    Receives instance $z_t \in \mathcal{Z}$
    Suffers loss $\ell(\hat{\mathbf{y}}_t, z_t)$
End
```

The goal again is to minimize regret :

$$\mathrm{Reg}_n := \frac{1}{n}\sum_{t=1}^{n} \ell(\hat{\mathbf{y}}_t, z_t) - \inf_{\mathbf{f}\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(\mathbf{f}, z_t)$$

2 Examples

Online Linear SVM In the case of SVM we are interested in linear predictors with constraint on the ℓ_2 norm of the predictor. In this case, $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \{\pm 1\}$. $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\ell(\mathbf{f}, (\mathbf{x}, y)) = \max\{0, 1 - y \cdot \mathbf{f}^\top \mathbf{x}\}$, $\mathcal{F} = \{\mathbf{f} : \|\mathbf{f}\|_2 \le R\}$. Feel free to change hinge loss to any convex loss line square loss, logistic loss etc. Also feel free to replace the constraint $\|\mathbf{f}\|_2 \le R$ by some other convex constraint. Regret is given by

$$\mathrm{Reg}_n = \frac{1}{n}\sum_{t=1}^{n} \max\{0, 1 - y_t \cdot \hat{\mathbf{y}}_t^\top \mathbf{x}_t\} - \inf_{\mathbf{f}\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \max\{0, 1 - y_t \cdot \mathbf{f}_t^\top \mathbf{x}_t\}$$

Regularized Linear Prediction Another set of problems that automatically fits the online convex optimization framework are regularized loss minimization problem. Here again $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y}$ could be say $[-1, 1]$. Now consider the case when $\ell(\mathbf{f}, (x, y)) = \phi(\mathbf{f}^\top \mathbf{x}, y) + \mathbf{R}(\mathbf{f})$. Where $\phi : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is some loss convex in first argument. $\mathbf{R} : \mathcal{F} \mapsto \mathbb{R}$ is a convex function. As an example think of the regularized version of SVM or online ridge regression, or online Lasso.

Experts Problem In the experts problem, we assume we have a set of N experts. Let $\phi : [N] \times \mathcal{Z} \mapsto [-1,1]$ be any arbitrary loss function of your choice that maps each expert to its loss on given instance.. Now let us define $\mathcal{F} = \Delta_N$ as the set of distributions over N experts (which is of course a convex set). Noe that for any $f \in \mathcal{F} = \Delta_N$, its loss is given by $\ell(f,z) = \mathbb{E}_{i \sim f}[\phi(i,z)] = \sum_{i=1}^N f_i \cdot \phi(i,z)$ which is clearly linear in f. In this case clearly regret is given by

$$\begin{aligned}
\mathrm{Reg}_n &= \frac{1}{n}\sum_{t=1}^n \ell(\hat{\mathbf{y}}_t, z_t) - \inf_{\mathbf{f}\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^n \ell(\mathbf{f}, z_t) \\
&= \frac{1}{n}\sum_{t=1}^n \mathbb{E}_{g_t \sim \hat{\mathbf{y}}_t} \phi(g_t, z_t) - \inf_{\mathbf{f}\in\Delta_N} \frac{1}{n}\sum_{t=1}^n \mathbb{E}_{g\sim f}[\phi(g, z_t)] \\
&= \frac{1}{n}\sum_{t=1}^n \mathbb{E}_{g_t \sim \hat{\mathbf{y}}_t} \phi(g_t, z_t) - \inf_{\mathbf{f}\in\Delta_N} \mathbb{E}_{g\sim f}\left[\frac{1}{n}\sum_{t=1}^n \phi(g, z_t)\right] \\
&= \frac{1}{n}\sum_{t=1}^n \mathbb{E}_{g_t \sim \hat{\mathbf{y}}_t} \phi(g_t, z_t) - \min_{i\in[N]} \frac{1}{n}\sum_{t=1}^n \phi(i, z_t)
\end{aligned}$$

That is, we can think about regret as the expected loss of our algorithm compared to the loss of the single best expert in hingsight

Matrix Prediction/Collaborative Filtering Imagine we have a bunch of M users and a bunch of N products. We want to predicts ratings of users for various products in an online fashion. Eg. on round t we are given $x_t \in [M] \times [N]$ the position of the matrix we are required to predict. Learner then picks the predicted rating. Finally the true rating is revealed and learner suffers loss for predicting wrong.

$$\mathrm{Reg}_n = \frac{1}{n}\sum_{t=1}^n |\hat{\mathbf{y}}_t[x_t] - y_t| - \inf_{\mathbf{f}\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^n |\mathbf{f}[x_t] - y_t|$$

Think of $\mathcal{F}$ as a convex set where each $\mathbf{f} \in \mathcal{F}$ is an $M \times N$ matrix. Each $\hat{y}_t$ is also an $M \times N$ matrix.

2.1 Online Linear Optimization

Though we are concerned with general convex losses, it suffices (in many cases with no additional cost) to only consider online linear optimization where the loss is linear rather than general convex. The reason for this is the following. First, given any $z_1, \ldots, z_n \in \mathcal{Z}$ let $\mathbf{f}^* = \underset{\mathbf{f}\in\mathcal{F}}{\mathrm{argmin}} \sum_{t=1}^n \ell(\mathbf{f}, z_t)$. Now note that by convexity,

$$\begin{aligned}
\sum_{t=1}^n \ell(\hat{\mathbf{y}}_t, z_t) - \sum_{t=1}^n \ell(\mathbf{f}^*, z_t) &\le \sum_{t=1}^n \langle \nabla \ell(\hat{\mathbf{y}}_t, z_t), \hat{\mathbf{y}}_t - \mathbf{f}^* \rangle \\
&\le \sum_{t=1}^n \langle \nabla \ell(\hat{\mathbf{y}}_t, z_t), \hat{\mathbf{y}}_t \rangle - \inf_{\mathbf{f}\in\mathcal{F}} \sum_{t=1}^n \langle \nabla \ell(\hat{\mathbf{y}}_t, z_t), \mathbf{f} \rangle
\end{aligned}$$

Now let $\mathcal{D}$ be the subset of vectors defined as, $\mathcal{D} = \{\nabla(\mathbf{f}, z) : \mathbf{f} \in \mathcal{F}, z \in \mathcal{Z}$. Now since in the online learning protocol, learner picks $\hat{\mathbf{y}}_t \in \mathcal{F}$ and then adversary picks $z \in \mathcal{Z}$, we can simply think of adversary as directly picking any $\nabla_t \in \mathcal{D}$ directly and this only increases the bound. Thus,

$$\frac{1}{n}\sum_{t=1}^{n} \ell(\hat{\mathbf{y}}_t, z_t) - \inf_{\mathbf{f}\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(\mathbf{f}, z_t) \leq \frac{1}{n}\sum_{t=1}^{n} \langle \nabla_t, \hat{\mathbf{y}}_t \rangle - \inf_{\mathbf{f}\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \langle \nabla_t, \mathbf{f} \rangle$$

What the above means is that if we have an algorithm for online linear optimization, we can use it as an algorithm for online convex optimization assuming the instance received on round t is the gradients of the convex function at the point $\hat{\mathbf{y}}_t$.

3 Online Gradient Descent

In this example we assume $\mathcal{F} = \{\mathbf{f} : \|\mathbf{f}\|_2 \leq R\}$ and $\mathcal{D}$ is a set whose elements all have Euclidean norm bounded by B. We consider linear loss. That is at time t the loss is $\langle \nabla_t, \hat{\mathbf{y}}_t \rangle$.

Algorithm :

$$\hat{\mathbf{y}}_{t+1} = \Pi_{\mathcal{F}}\left(\hat{\mathbf{y}}_t - \eta \nabla_t\right)$$

where $\Pi_{\mathcal{F}}$ is the Euclidean projection on to set $\mathcal{F}$ and $\eta > 0$ is referred to as step-size.

$$\Pi_F(\mathbf{f}) = \begin{cases} \mathbf{f} & \text{if } \|\mathbf{f}\|_2 \leq R \\ R\frac{\mathbf{f}}{\|\mathbf{f}\|_2} & \text{otherwise} \end{cases}$$

Claim 1. *If we use the online gradient descent algorithm with* $\eta = \frac{R}{B\sqrt{n}}$ *and* $\hat{\mathbf{y}}_1 = \mathbf{0}$*, then*

$$\frac{1}{n}\sum_{t=1}^{n} \langle \nabla_t, \hat{\mathbf{y}}_t \rangle - \inf_{\mathbf{f}\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \langle \nabla_t, \mathbf{f} \rangle \leq \frac{RB}{\sqrt{n}}$$

Proof. Fix any $\mathbf{f}^* \in \mathcal{F}$. Note that,

$$\|\hat{\mathbf{y}}_{t+1} - \mathbf{f}^*\|_2^2 = \|\Pi_{\mathcal{F}}\left(\hat{\mathbf{y}}_t - \eta\nabla_t\right) - \mathbf{f}^*\|_2^2 \leq \|\hat{\mathbf{y}}_t - \eta\nabla_t - \mathbf{f}^*\|_2^2 = \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 + \eta^2 \|\nabla_t\|_2^2 - 2\eta\langle \nabla_t, \hat{\mathbf{y}}_t - \mathbf{f}^* \rangle$$

Thus we can conclude that

$$\langle \nabla_t, \hat{\mathbf{y}}_t - \mathbf{f}^* \rangle \leq \frac{1}{2\eta}\left(\|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 - \|\hat{\mathbf{y}}_{t+1} - \mathbf{f}^*\|_2^2\right) + \frac{\eta}{2}\|\nabla_t\|_2^2$$

Summing we get,

$$\begin{aligned} \sum_{t=1}^{n} \langle \nabla_t, \hat{\mathbf{y}}_t - \mathbf{f}^* \rangle &\leq \frac{1}{2\eta}\sum_{t=1}^{n}\left(\|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 - \|\hat{\mathbf{y}}_{t+1} - \mathbf{f}^*\|_2^2\right) + \frac{\eta}{2}\sum_{t=1}^{n}\|\nabla_t\|_2^2 \\ &= \frac{1}{2\eta}\left(\|\hat{\mathbf{y}}_1 - \mathbf{f}^*\|_2^2 - \|\hat{\mathbf{y}}_{n+1} - \mathbf{f}^*\|_2^2\right) + \frac{\eta}{2}nB^2 \\ &\leq \frac{1}{2\eta}R^2 + \frac{\eta}{2}nB^2 \end{aligned}$$

Using the η from the claim and dividing throughout by n gives the result. □

What is the lower bound for this problem? In fact it is not hard to see that the lower bound for this problem is also $\frac{RB}{\sqrt{n}}$ at least when dimensionality is huge. To see this assume the adversary simply plays on each round vector orthogonal to current $\hat{\mathbf{y}}_t$ and also orthogonal to previous $\nabla_1, \ldots, \nabla_{t-1}$.

This algorithm is worst case optimal (in terms of computational efficiency) for SVM (even for statistical learning). Why ? Think about sample complexity and amount of time needed to read the data.

4 Online Mirror Descent

Is the online gradient descent algorithm always the right thing to use? Let us look at the finite experts problem. $\mathcal{F} = \Delta_N$ and $\langle \mathbf{f}, \nabla_t \rangle = \mathbb{E}_{i \sim \mathbf{f}}[\phi(i, z_t)]$. Notice that in this setting, for any $\mathbf{f} \in \Delta_N$, $\|\mathbf{f}\|_2 \le \|\mathbf{f}\|_1 = 1$. However note that $\|\nabla_t\|_2 = \sqrt{\sum_{i=1}^N |\phi(i, z_t)|} \le \sqrt{N}$ (assuming losses are bounded by 1). Hence GD bound can only given a rate of

$$\mathrm{Reg}_n \le \sqrt{\frac{N}{n}}$$

But is this the best rate possible? In statistical learning setting we know that $\log N$ was achievable, can we obtain that here? What is the right algorithm in general. In fact in general vector spaces, GD does not even type check!
Strongly convex function: Function R is said to be λ-strongly convex w.r.t. norm $\|\cdot\|$ if $\forall \mathbf{f}, \mathbf{f}'$,

$$R\left(\frac{\mathbf{f} + \mathbf{f}'}{2}\right) \le \frac{R(\mathbf{f}) + R(\mathbf{f}')}{2} - \frac{\lambda}{2}\|\mathbf{f} - \mathbf{f}'\|^2$$

This can equivalently be written as:

$$R(\mathbf{f}') \le R(\mathbf{f}) + \left\langle \nabla R(\mathbf{f}'), \mathbf{f}' - \mathbf{f} \right\rangle - \frac{\lambda}{2}\|\mathbf{f} - \mathbf{f}'\|^2$$

Bregman Divergence w.r.t. function R:

$$\Delta_R(\mathbf{f}'|\mathbf{f}) = R(\mathbf{f}') - R(\mathbf{f}) - \left\langle \nabla R(\mathbf{f}), \mathbf{f}' - \mathbf{f} \right\rangle$$

Clearly if a function R is λ strongly convex, then by definition, $\Delta_R(\mathbf{f}'|\mathbf{f}) \ge \frac{\lambda}{2}\|\mathbf{f}' - \mathbf{f}\|^2$

Algorithm : Let R be any strongly convex function. We define the mirror descent update as follows :

$$\nabla R(\hat{\mathbf{y}}'_{t+1}) = \nabla R(\hat{\mathbf{y}}_t) - \eta \nabla_t \quad , \quad \hat{\mathbf{y}}_{t+1} = \underset{\hat{\mathbf{y}} \in \mathcal{F}}{\mathrm{argmin}}\, \Delta_R(\hat{\mathbf{y}}|\hat{\mathbf{y}}'_{t+1})$$

$$\text{Equivalently,} \qquad \hat{\mathbf{y}}_{t+1} = \underset{\hat{\mathbf{y}} \in \mathcal{F}}{\mathrm{argmin}}\, \eta\langle \nabla_t, \hat{\mathbf{y}} \rangle + \Delta_R(\hat{\mathbf{y}}|\hat{\mathbf{y}}_t)$$

and we use $\hat{\mathbf{y}}_1 = \underset{\hat{\mathbf{y}} \in \mathcal{F}}{\mathrm{argmin}}\, R(\hat{\mathbf{y}})$

Bound :

Claim 2. *Let R be any 1-strongly convex function. If we use the Mirror descent algorithm with $\eta = \sqrt{\frac{2\sup_{\mathbf{f}\in\mathcal{F}} R(\mathbf{f})}{nB^2}}$ then,*

$$\mathrm{Reg}_n \leq \sqrt{\frac{2B^2 \sup_{\mathbf{f}\in\mathcal{F}} R(\mathbf{f})}{n}}$$

Proof. Consider any $\mathbf{f}^* \in \mathcal{F}$, we have that,

$$\begin{aligned}\langle \nabla_t, \hat{\mathbf{y}}_t \rangle - \langle \nabla_t, \mathbf{f}^* \rangle &= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} + \hat{\mathbf{y}}'_{t+1} - \mathbf{f}^* \right\rangle \\ &= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} \right\rangle + \left\langle \nabla_t, \hat{\mathbf{y}}'_{t+1} - \mathbf{f}^* \right\rangle\end{aligned}$$

By the mirror descent update, $\nabla_t = \frac{1}{\eta}\left(\nabla R(\hat{\mathbf{y}}_t) - \nabla R(\hat{\mathbf{y}}'_{t+1})\right)$

$$= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} \right\rangle + \frac{1}{\eta}\left\langle \nabla R(\hat{\mathbf{y}}_t) - \nabla R(\hat{\mathbf{y}}'_{t+1}), \hat{\mathbf{y}}'_{t+1} - \mathbf{f}^* \right\rangle$$

For any vectors a, b, c, $\langle \nabla R(a) - \nabla R(b), b - c \rangle = \Delta_R(c|a) - \Delta_R(c|b) - \Delta_R(b|a)$

$$= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} \right\rangle + \frac{1}{\eta}\left(\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1}) - \Delta_R(\hat{\mathbf{y}}_t|\hat{\mathbf{y}}'_{t+1})\right)$$

$\langle a, b \rangle \leq \|a\| \, \|b\|_* \leq \frac{\eta}{2}\|b\|_*^2 + \frac{1}{2\eta}\|a\|^2$

$$\leq \frac{\eta}{2}\|\nabla_t\|_*^2 + \frac{1}{2\eta}\left\|\hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\|^2 + \frac{1}{\eta}\left(\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1}) - \Delta_R(\hat{\mathbf{y}}'_{t+1}|\hat{\mathbf{y}}_t)\right)$$

By strangle convexity of R, $\Delta_R(\hat{\mathbf{y}}_t|\hat{\mathbf{y}}'_{t+1}) \geq \frac{1}{2}\left\|\hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\|^2$

$$\leq \frac{\eta}{2}\|\nabla_t\|_*^2 + \frac{1}{\eta}\left(\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\hat{\mathbf{y}}'_{t+1}|\hat{\mathbf{y}}_t)\right)$$

Summing over we have,

$$\sum_{t=1}^n \langle \nabla_t, \hat{\mathbf{y}}_t \rangle - \sum_{t=1}^n \langle \nabla_t, \mathbf{f}^* \rangle \leq \frac{\eta}{2}\sum_{t=1}^n \|\nabla_t\|_*^2 + \frac{1}{\eta}\sum_{t=1}^n \left(\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1})\right)$$

Replacing by projection only decreases the Bregman divergence

$$\begin{aligned}&\leq \frac{\eta}{2}\sum_{t=1}^n \|\nabla_t\|_*^2 + \frac{1}{\eta}\sum_{t=1}^n (\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_{t+1})) \\ &\leq \frac{\eta}{2}\sum_{t=1}^n \|\nabla_t\|_*^2 + \frac{1}{\eta}(\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_1) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_{n+1})) \\ &\leq \frac{\eta}{2}\sum_{t=1}^n \|\nabla_t\|_*^2 + \frac{1}{\eta}R(\mathbf{f}^*) \\ &\leq \frac{\eta}{2}nB^2 + \frac{1}{\eta}\sup_{f\in\mathcal{F}} R(\mathbf{f}) \\ &= \sqrt{2B^2 \sup_{\mathbf{f}\in\mathcal{F}} R(\mathbf{f}) n}\end{aligned}$$

Dividing through by n we prove the claim. □

4.1 Examples

Gradient Descent $R(\hat{\mathbf{y}}) = \frac{1}{2}\left\|\hat{\mathbf{y}}\right\|_2^2$. In this case mirror descent update coincides with that of Gradient descent and we recover the bound. Strong convexity is just Pythagorus theorem

Exponential Weights Let is consider the example of finite experts setting. In this setting we can consider R to be the negative entropy function,

$$R(\hat{\mathbf{y}}) = \sum_{i=1}^{d} \hat{\mathbf{y}}[i] \log(\hat{\mathbf{y}}[i]) - 1$$

Note that

$$D_R(\hat{\mathbf{y}}|\hat{\mathbf{y}}') = \mathrm{KL}(\hat{\mathbf{y}}\|\hat{\mathbf{y}}') = \sum_{i=1}^{d} \hat{\mathbf{y}}[i] \log\left(\frac{\hat{\mathbf{y}}[i]}{\hat{\mathbf{y}}'[i]}\right)$$

In this case, it is not too hard to check that R is strongly convex w.r.t. $\left\|\cdot\right\|_1$. Also note that $\sup_{\mathbf{f}\in\Delta_N} R(\mathbf{f}) \le \log N$ (achieved at the uniform distribution).

ℓ_p and Schatten$_p$ norms Let us consider $\mathcal{F}$ to be unit ball under ℓ_p norm and $\mathcal{D}$ to be unit ball under dual norm. Let $p \in (1, 2]$, then one can use $R(\mathbf{f}) = \frac{1}{p-1}\left\|\mathbf{f}\right\|_p^2$ and this function is strongly convex w.r.t. ℓ_p norm. For matrices with analogous Schatten p norm, use the $R(\mathbf{f}) = \frac{1}{p-1}\left\|\mathbf{f}\right\|_{S_p}^2$.

Remark 4.1. *For ℓ_1 norm one can use $R(\mathbf{f}) = \frac{1}{p-1}\left\|\mathbf{f}\right\|_p^2$ with $p \approx \frac{\log d}{\log d - 1}$ and hence recover a bound of form $O\left(\sqrt{\frac{B^2 \log d}{n}}\right)$ where B is the bound on ℓ_∞ norm of ∇_t's.*