Machine Learning Theory (CS 6783)

Lecture 16: Online Mirror Descent contd.

1 Recap

- Gradient descent best suited for Euclidean structure, ie. when ℓ_2 norms of $\mathbf{f} \in \mathcal{F}$ and $\nabla_t \in \mathcal{D}$ are small
- To get right rates one needs to look at the structures if $\mathcal{F}$ and $\mathcal{D}$. Eg. finite experts problem gradient descent only gets $\sqrt{\frac{|\mathcal{F}'|}{n}}$ rate
- Mirror descent update :

$$\nabla R(\hat{\mathbf{y}}'_{t+1}) = \nabla R(\hat{\mathbf{y}}_t) - \eta \nabla_t \quad \& \quad \hat{\mathbf{y}}_{t+1} = \operatorname*{argmin}_{\hat{\mathbf{y}}} \Delta_R(\hat{\mathbf{y}}, \hat{\mathbf{y}}_{t+1})$$

- If R is 1-strongly convex w.r.t. some norm $\|\cdot\|$ (and $\|\cdot\|_*$ its dual) then using MD we get

$$\mathbf{R}_n \leq O\left(\sqrt{\frac{(\sup_{\mathbf{f}\in\mathcal{F}} R(f)) \cdot \sup_{\nabla\in\mathcal{D}} \|\nabla\|_*^2}{n}}\right)$$

  Structure of $\mathcal{F}$ and $\mathcal{D}$ captured via $(\sup_{\mathbf{f}\in\mathcal{F}} R(f))$ and $\sup_{\nabla\in\mathcal{D}} \|\nabla\|_*^2$, Eg. in the experts setting using negative entropy, $R(\mathbf{f}) = \sum_{i=1}^d \mathbf{f}(i) \log \mathbf{f}(i) - 1$ MD recovers exponential weights algorithm.

2 Strongly Convex Objectives and Faster Rates

The bounds we showed for online mirror descent and online gradient descent are tight for online linear optimization and as discussed before can also be used for online convex optimization in general. However, if one knows in advance that the objectives are curved, strongly convex for instance, then the rates can be improved and in fact using the same mirror descent/gradient descent algorithm but with step sizes that vary with round t.

Example : Regularized SVM

$$\ell(\mathbf{f}, (\mathbf{x}, y)) = \max\{1 - y \cdot \mathbf{f}^\top \mathbf{x}, 0\} + \frac{\lambda}{2} \|\mathbf{f}\|_2^2$$

More generally, in this section we shall assume that for each $z \in \mathcal{Z}$, the loss $\ell(\cdot, z)$ is λ-strongly convex w.r.t. norm $\|\cdot\|_2$, that is,

$$\ell(\mathbf{f}', z) \leq \ell(\mathbf{f}, z) + \left\langle \nabla \ell(\mathbf{f}', z), \mathbf{f}' - \mathbf{f} \right\rangle - \frac{\lambda}{2} \left\|\mathbf{f} - \mathbf{f}'\right\|_2^2$$

The results just as easily extend to mirror descent for other norms.

Algorithm : GD update:

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

and we use $\hat{\mathbf{y}}_1 = 0$

Claim 1. *If we use the online gradient descent algorithm with $\eta_t = \frac{1}{\lambda t}$ then, whenever the losses are λ-strongly convex, we get*

$$\mathrm{Reg}_n \leq \frac{B^2 \log n}{2\lambda n}$$

Proof. Consider any $\mathbf{f}^* \in \mathcal{F}$, by strong convexity of loss, we have that,

$$\begin{aligned}
\ell(\hat{\mathbf{y}}_t, z_t) - \ell(\mathbf{f}^*, z_t) &\leq \langle \nabla_t, \hat{\mathbf{y}}_t - \mathbf{f}^* \rangle - \frac{\lambda}{2} \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 \\
&= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} + \hat{\mathbf{y}}'_{t+1} - \mathbf{f}^* \right\rangle - \frac{\lambda}{2} \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 \\
&= \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 - \frac{\lambda}{2} \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 \\
&= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} \right\rangle + \frac{1}{\eta_t} \left\langle \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}, \hat{\mathbf{y}}'_{t+1} - \mathbf{f}^* \right\rangle - \frac{\lambda}{2} \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 \\
&= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} \right\rangle + \frac{1}{2\eta_t} \left(\|\mathbf{f}^* - \hat{\mathbf{y}}_t\|_2^2 - \left\|\mathbf{f}^* - \hat{\mathbf{y}}'_{t+1}\right\|_2^2 - \left\|\hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\|_2^2 \right) - \frac{\lambda}{2} \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 \\
&\leq \frac{\eta_t}{2} \|\nabla_t\|_2^2 + \frac{1}{2\eta_t} \left\|\hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\|_2^2 + \frac{1}{2\eta_t} \left(\|\mathbf{f}^* - \hat{\mathbf{y}}_t\|_2^2 - \left\|\mathbf{f}^* - \hat{\mathbf{y}}'_{t+1}\right\|_2^2 - \left\|\hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\|_2^2 \right) - \frac{\lambda}{2} \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 \\
&\leq \frac{\eta_t}{2} \|\nabla_t\|_2^2 + \frac{1}{\eta_t} \left(\|\mathbf{f}^* - \hat{\mathbf{y}}_t\|_2^2 - \left\|\mathbf{f}^* - \hat{\mathbf{y}}'_{t+1}\right\|_2^2 \right) - \frac{\lambda}{2} \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 \\
&\leq \frac{\eta_t}{2} B^2 + \frac{1}{2\eta_t} \left(\|\mathbf{f}^* - \hat{\mathbf{y}}_t\|_2^2 - \left\|\mathbf{f}^* - \hat{\mathbf{y}}'_{t+1}\right\|_2^2 \right) - \frac{\lambda}{2} \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|^2
\end{aligned}$$

Summing over we have,

$$\begin{aligned}
\sum_{t=1}^{n} \ell(\hat{\mathbf{y}}_t, z_t) - \ell(\mathbf{f}^*, z_t) &\leq \frac{B^2}{2} \sum_{t=1}^{n} \eta_t + \sum_{t=1}^{n} \frac{1}{2\eta_t} \left(\|\mathbf{f}^* - \hat{\mathbf{y}}_t\|_2^2 - \left\|\mathbf{f}^* - \hat{\mathbf{y}}'_{t+1}\right\|_2^2 \right) - \frac{\lambda}{2} \|\hat{\mathbf{y}}_t - \mathbf{f}^*\|_2^2 \\
&\leq \frac{B^2 \log n}{2\lambda} + \frac{1}{2} \|\hat{\mathbf{y}}_1 - \mathbf{f}^*\|_2^2 \left(\frac{1}{\eta_1} - \lambda \right) + \frac{1}{2} \sum_{t=2}^{n} \|\mathbf{f}^* - \hat{\mathbf{y}}_t\|_2^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \lambda \right) \\
&= \frac{B^2 \log n}{2\lambda}
\end{aligned}$$

Dividing through by n we prove the claim. □

2.1 Exp-concave losses and Online Newton Method

All losses are not made equal, some are more special! We saw how one can get faster rates for strongly convex losses. However strong convexity of the loss is a rather strong assumption. It is possible to get faster rates for losses that are not strongly convex but still have some nice properties. As an example consider linear prediction with squared loss in d dimensions. That is

$\ell(\mathbf{f}, (\mathbf{x}, y)) = (\mathbf{f}^\top \mathbf{x} - y)^2$. This loss is not strongly convex as a function of $\mathbf{f}$ w.r.t. any norm (don't confuse this with strong convexity of $(\hat{y} - y)^2$ w.r.t. $\hat{y}$). However this loss does have curvature in the direction we care about.

Throughout this subsection assume that $\mathcal{F} \subset \mathbb{R}^d$ s.t. $\|\mathbf{f}\|_2 \leq 1$.

Assumption 2. *Assume that the loss ℓ is such that, for any z and any $\mathbf{f}, \mathbf{f}' \in \mathcal{F}$,*

$$\ell(\mathbf{f}', z) \leq \ell(\mathbf{f}, z) + \left\langle \nabla \ell(\mathbf{f}', z), \mathbf{f}' - \mathbf{f} \right\rangle - \frac{\beta}{2} (\mathbf{f}' - \mathbf{f})^\top \left(\nabla \ell(\mathbf{f}', z) \right) \left(\nabla \ell(\mathbf{f}', z) \right)^\top (\mathbf{f}' - \mathbf{f})$$

A sufficient condition for the above is that loss ℓ is what is referred to as exp-concave and 1-Lipschitz (ie. $\|\nabla \ell(\mathbf{f}, z)\|_2 \leq 1$). ℓ is said to be α-exp-concave if for all z, $\exp(-\alpha \ell(\cdot, z))$ is a concave function. In this case $\lambda \leq \frac{1}{2} \min\{\frac{1}{4}, \alpha\}$
Examples : linear prediction with squared loss $\beta = 1$, Logistic loss $\beta = O(e^{-R})$, ...

Algorithm : Use arbitrary $\hat{\mathbf{y}}_1 \in \mathcal{F}$ and use $A_1 = \sigma I_d$ (I_d is identity matrix)

$$A_{t+1} = A_t + \nabla_t^\top \nabla_t \qquad \hat{\mathbf{y}}'_{t+1} = \hat{\mathbf{y}}_t - \eta A_{t+1}^{-1} \nabla_t \qquad \hat{\mathbf{y}}_{t+1} = \operatorname*{argmin}_{\hat{\mathbf{y}} \in \mathcal{F}} (\hat{\mathbf{y}} - \hat{\mathbf{y}}'_{t+1})^\top A_{t+1} (\hat{\mathbf{y}} - \hat{\mathbf{y}}'_{t+1})$$

Think of the above as MD with R varying over time. Specifically $R_t(\mathbf{f}) = \frac{1}{2} \mathbf{f}^\top A_{t-1} \mathbf{f}$.

Claim 3. *Using $\eta = \frac{1}{\beta}$ and $\sigma = \frac{1}{\beta^2}$ if we run the online Newton method, we get*

$$\mathbf{R}_n \leq O\left(\frac{d \log(n+1)}{2n\beta} \right)$$

Proof sketch. Define $R_t(\mathbf{f}) = \frac{1}{2} \mathbf{f}^\top A_{t+1} \mathbf{f}$ and view the algorithm as

$$\nabla R_t(\hat{\mathbf{y}}'_{t+1}) = \nabla R_t(\hat{\mathbf{y}}_t) - \eta \nabla_t \qquad \hat{\mathbf{y}}_{t+1} = \operatorname*{argmin}_{\hat{\mathbf{y}} \in \mathcal{F}} \Delta_{R_t}(\hat{\mathbf{y}} | \hat{\mathbf{y}}'_{t+1})$$

Now note that for any $\mathbf{f}^* \in \mathcal{F}$,

$$\ell(\hat{\mathbf{y}}_t, z_t) - \ell(\mathbf{f}^*, z_t) \leq \langle \nabla_t, \hat{\mathbf{y}}_t - \mathbf{f}^* \rangle - \frac{1}{\eta} \left(\Delta_{R_t}(\mathbf{f}^* | \hat{\mathbf{y}}_t) - \Delta_{R_{t-1}}(\mathbf{f}^* | \hat{\mathbf{y}}_t) \right)$$

Following the bound from MD proof,

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

Combining we get,

$$\begin{aligned}
\ell(\hat{\mathbf{y}}_t, z_t) - \ell(\mathbf{f}^*, z_t) &\leq \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} \right\rangle + \frac{1}{\eta}\left(\Delta_{R_t}(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_{R_t}(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1}) - \Delta_{R_t}(\hat{\mathbf{y}}_t|\hat{\mathbf{y}}'_{t+1})\right) \\
&\quad - \frac{1}{\eta}\left(\Delta_{R_t}(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_{R_{t-1}}(\mathbf{f}^*|\hat{\mathbf{y}}_t)\right) \\
&= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} \right\rangle + \frac{1}{\eta}\left(\Delta_{R_{t-1}}(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_{R_t}(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1}) - \Delta_{R_t}(\hat{\mathbf{y}}_t|\hat{\mathbf{y}}'_{t+1})\right) \\
&\leq \frac{\eta}{2}\left\|\nabla_t\right\|^2_{A_t^{-1}} + \frac{1}{2\eta}\left\|\hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\|^2_{A_t} + \frac{1}{\eta}\left(\Delta_{R_{t-1}}(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_{R_t}(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1}) - \Delta_{R_t}(\hat{\mathbf{y}}_t|\hat{\mathbf{y}}'_{t+1})\right) \\
&= \frac{\eta}{2}\left\|\nabla_t\right\|^2_{A_t^{-1}} + \frac{1}{\eta}\left(\Delta_{R_{t-1}}(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_{R_t}(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1})\right) \\
&\leq \frac{\eta}{2}\left\|\nabla_t\right\|^2_{A_t^{-1}} + \frac{1}{\eta}\left(\Delta_{R_{t-1}}(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_{R_t}(\mathbf{f}^*|\hat{\mathbf{y}}_{t+1})\right)
\end{aligned}$$

Summing up and noticing the telescoping sum we get,

$$\begin{aligned}
n\mathrm{Reg}_n &\leq \frac{\eta}{2}\sum_{t=1}^{n} \nabla_t^\top \left(A_{t-1} + \nabla_t\nabla_t^\top\right)^{-1}\nabla_t + \frac{1}{\eta}\Delta_{R_1}(\mathbf{f}^*|\hat{\mathbf{y}}_1) \\
&= \frac{\eta}{2}\sum_{t=1}^{n} \nabla_t^\top \left(A_{t-1} + \nabla_t\nabla_t^\top\right)^{-1}\nabla_t + \frac{\sigma}{\eta}\left\|\mathbf{f}^* - \hat{\mathbf{y}}_1\right\|_2^2 \\
&\leq \frac{1}{2\beta}\sum_{t=1}^{n} \nabla_t^\top \left(A_{t-1} + \nabla_t\nabla_t^\top\right)^{-1}\nabla_t + \frac{4}{\beta}
\end{aligned}$$

To conclude the proof note that by matrix-determinant identity

$$\nabla_t^\top \left(A_{t-1} + \nabla_t\nabla_t^\top\right)^{-1}\nabla_t = 1 - \frac{\det(A_{t-1})}{\det(A_{t-1} + \nabla_t\nabla_t^\top)} = 1 - \frac{\det(A_{t-1})}{\det(A_t)} \leq \log\left(\frac{\det(A_t)}{\det(A_{t-1})}\right)$$

Hence,

$$n\mathrm{Reg}_n \leq \frac{1}{2\beta}\log\left(\frac{\det(A_n)}{\det(A_0)}\right) + \frac{4}{\beta} = \frac{1}{2\beta}\left(\sum_{j=1}^{d}\log(1+\lambda_j) + 4\right) \leq \frac{1}{2\beta}\left(d\log(n) + 4\right)$$

□