Machine Learning Theory (CS 6783)

Lecture 15: Online Mirror Descent contd.

1 Recap

- 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, for linear game (convex Lipschitz) we get

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

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

- If Losses are λ-strongly convex w.r.t. ℓ_2 norm them using GD with step size $\eta_t = 1/\lambda t$ we get

$$\mathrm{Reg}_n \leq O\left(\frac{\sup_{\nabla\in\mathcal{D}} \|\nabla\|_*^2 \ \log n}{\lambda n}\right)$$

2 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 $\ell(f,(\mathbf{x},y)) = (f^\top \mathbf{x} - y)^2$. This loss is not strongly convex as a function of 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. $\|f\|_2 \leq 1$.

Assumption 1. *Assume that the loss ℓ is such that, for any z and any $f, f' \in \mathcal{F}$,*

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

A sufficient condition for the above is that loss ℓ is what is referred to as exp-concave and 1-Lipschitz (ie. $\|\nabla \ell(f,z)\|_2 \le 1$). ℓ is said to be α-exp-concave if for all z, $\exp(-\alpha \ell(\cdot, z))$ is a concave function. In this case $\lambda \le \frac{1}{2}\min\{\frac{1}{4}, \alpha\}$
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 $A_1 = I_d$ (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 $R_t(f) = \frac{1}{2} f^\top A_{t+1} f$. In this case $\Delta_{R_t}(a|b) = \frac{1}{2}(a-b)^\top A_{t+1}(a-b)$.

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

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

Proof sketch. Define $R_t(f) = \frac{1}{2} f^\top A_{t+1} f$ 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 $f^* \in \mathcal{F}$,

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

Following the bound from MD proof,

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

Combining we get,

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

Summing up and noticing the telescoping sum we get,

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

To conclude the proof note that by matrix-determinant identity we have that for any vector x and any invertible matrix B, $\det(B - xx^\top) = \det(B)(1 - x^\top B^{-1}x)$ and so using $B = A_t + \nabla_t\nabla_t^\top$ and $x = \nabla_t$ we have:

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

Hence,

$$n\mathrm{Reg}_n \le \frac{1}{2\beta}\log\left(\frac{\det(A_{n+1})}{\det(A_1)}\right) + \frac{4}{\beta} = \frac{1}{2\beta}\left(\sum_{j=1}^{d}\log(1+\lambda_j(A_{n+1})) + 4\right) \le \frac{1}{2\beta}\left(d\log(n) + 4\right)$$

□

3 Mirror Descent and Local Norms

Lemma 3. *For any twice differentiable convex R, if we run mirror descent using step size η, then*

$$n\mathrm{Reg}_n(\nabla_1, \dots \nabla_n) \le \frac{\eta}{2}\sum_{t=1}^{n}\|\nabla_t\|^2_{\nabla^2 R(z_t)^{-1}} + \frac{1}{\eta}\sup_{f\in\mathcal{F}}\Delta_R(f|\hat{y}_1)$$

where z_t is some convex combination of $\hat{\mathbf{y}}_t$ and $\hat{\mathbf{y}}'_{t+1}$ (here matrix M, $\|x\|_M^2 = x^\top Mx$)

Proof. We will recall the upper bound from the mirror descent proof of the form:

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

Now the key trick is that we start with the definition of Bregman divergence and use Taylor's theorem. Note that:

$$\Delta_R(\hat{\mathbf{y}}'_{t+1}|\hat{\mathbf{y}}_t) = R(\hat{\mathbf{y}}'_{t+1}) - R(\hat{\mathbf{y}}_t) - \left\langle R(\hat{\mathbf{y}}_t), \hat{\mathbf{y}}'_{t+1} - \hat{\mathbf{y}}_t\right\rangle$$

Now using Taylor's theorem (+ intermediate value theorem) there exists a point z_t that is some convex combination of $\hat{\mathbf{y}}'_{t+1}$ and $\hat{\mathbf{y}}_t$ such that

$$R(\hat{\mathbf{y}}'_{t+1}) - R(\hat{\mathbf{y}}_t) - \left\langle R(\hat{\mathbf{y}}_t), \hat{\mathbf{y}}'_{t+1} - \hat{\mathbf{y}}_t\right\rangle = \frac{1}{2}(\hat{\mathbf{y}}'_{t+1} - \hat{\mathbf{y}}_t)^\top\nabla^2 R(z_t)(\hat{\mathbf{y}}'_{t+1} - \hat{\mathbf{y}}_t) = \frac{1}{2}\|\hat{\mathbf{y}}'_{t+1} - \hat{\mathbf{y}}_t\|^2_{\nabla^2 R(z_t)}$$

Hence using this we can conclude that

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

Now note that for any invertible matrix M, $\| \cdot \|_{M^{-1}}$ is the dual norm to the norm $\| \cdot \|_M$ and hence using the fact (as we did in the earlier mirror descent proof) that

$$\left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} \right\rangle \leq \frac{\eta}{2} \| \nabla_t \|^2_{\nabla^2 R(z_t)^{-1}} + \frac{1}{2\eta} \| \hat{\mathbf{y}}'_{t+1} - \hat{\mathbf{y}}_t \|^2_{\nabla^2 R(z_t)}$$

we conclude that

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

Summing over t and simplifying the telescoping sum over the Bregman divergences we we obtain that

$$\begin{aligned} n\mathrm{Reg}_n(\nabla_1, \dots \nabla_n) &\leq \frac{\eta}{2} \sum_{t=1}^{n} \| \nabla_t \|^2_{\nabla^2 R(z_t)^{-1}} + \frac{1}{\eta} \left(\Delta_R(f^* | \hat{\mathbf{y}}_1) - \Delta_R(f^* | \hat{\mathbf{y}}_{n+1}) \right) \\ &\leq \frac{\eta}{2} \sum_{t=1}^{n} \| \nabla_t \|^2_{\nabla^2 R(z_t)^{-1}} + \frac{1}{\eta} \sup_{f \in \mathcal{F}} \Delta_R(f | \hat{\mathbf{y}}_1) \end{aligned}$$

□

When is this result useful? Well, if hessian is a Lipschitz continuous function then one can further bound $\| \nabla_t \|_{\nabla^2 R(z_t)^{-1}}$ approximately by say $\| \nabla_t \|_{\nabla^2 R(\hat{y}_{t+1})^{-1}}$ since z_t is between $\hat{y}_t$ and $\hat{y}_{t+1}$ which themselves are close if η is small. We will see a concrete example of this when we come to Bandit algorithms but for now, its a good tool to keep in mind.