Machine Learning Theory (CS 6783)

Lecture 15: Online Mirror Descent

1 Recap

$\mathcal{F}$ is a convex subset of a vector space.

For $t = 1$ to $n$

- Learner picks $\hat{y}_t \in \mathcal{F}$
- Receives instance $\nabla_t \in \mathcal{D}$
- Suffers loss $\langle \hat{y}_t, \nabla_t \rangle$

End

The goal again is to minimize regret:

$$\text{Reg}_n := \frac{1}{n} \sum_{t=1}^{n} \langle \hat{y}_t, \nabla_t \rangle - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \langle f, \nabla_t \rangle$$

- **Online Gradient Descent Algorithm**:
  
  $$\hat{y}_{t+1} = \Pi_{\mathcal{F}} (\hat{y}_t - \eta \nabla_t)$$

- $\eta = \frac{R \sqrt{B}}{\sqrt{n}}$ and $\hat{y}_1 = 0$, then $\text{Reg}_n \leq \frac{RB}{\sqrt{n}}$ where $\sup_{f \in \mathcal{F}} \|f\|_2 \leq R$ and $\sup_{\nabla \in \mathcal{D}} \|\nabla\|_2 \leq B$

- Matches bound using sequential Rademacher complexity (both upper and lower bounds)

2 Online Mirror Descent

Is the online gradient descent algorithm always the right thing to use? Let us look at the finite experts problem. $\mathcal{F} = \Delta(\mathcal{F}^\prime)$ and $\langle f, \nabla_t \rangle = \mathbb{E}_{f^\prime \sim f} [\ell(f^\prime, (x_t, y_t))]$. Notice that in this setting, for any $f \in \Delta(\mathcal{F}^\prime)$, $\|f\|_2 \leq \|f\|_1 = 1$. However note that $\|\nabla_t\|_2 = \sqrt{\sum_{f^\prime \in \mathcal{F}^\prime} |\ell(f^\prime, (x_t, y_t))|} \leq \sqrt{|\mathcal{F}^\prime|}$. Hence GD bound can only given a rate of

$$\text{Reg}_n \leq \sqrt{\frac{|\mathcal{F}^\prime|}{n}}$$

But we know that a rate of $\sqrt{\log |\mathcal{F}^\prime|/n}$ is possible? 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 $\lambda$-strongly convex w.r.t. norm $\|\cdot\|$ if \( \forall f, f' \),

\[
R \left( \frac{f + f'}{2} \right) \leq \frac{R(f) + R(f')}{2} - \frac{\lambda}{2} \|f - f'\|^2
\]

This can equivalently be written as:

\[
R(f') \leq R(f) + \langle \nabla R(f'), f' - f \rangle - \frac{\lambda}{2} \|f - f'\|^2
\]

Bregman Divergence w.r.t. function $R$:

\[
\Delta_R(f'|f) = R(f') - R(f) - \langle \nabla R(f'), f' - f \rangle
\]

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

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

\[
\nabla R(\hat{y}_{t+1}) = \nabla R(\hat{y}_t) - \eta \nabla_t \quad \hat{y}_{t+1} = \arg \min_{\hat{y} \in F} \Delta_R(\hat{y} | \hat{y}_{t+1})
\]

Equivalently, \( \hat{y}_{t+1} = \arg \min_{\hat{y} \in F} \eta \langle \nabla_t, \hat{y} \rangle + \Delta_R(\hat{y} | \hat{y}_t) \)

and we use \( \hat{y}_1 = \arg \min_{\hat{y} \in F} R(\hat{y}) \)

**Bound:**

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

\[
\text{Reg}_n \leq \sqrt{\frac{2B^2 \sup_{f \in F} R(f)}{n}}
\]

**Proof.** Consider any $f^* \in F$, we have that,

\[
\langle \nabla_t, \hat{y}_t \rangle - \langle \nabla_t, f^* \rangle = \langle \nabla_t, \hat{y}_t - \hat{y}_{t+1} + \hat{y}_{t+1} - f^* \rangle
\]

\[
= \langle \nabla_t, \hat{y}_t - \hat{y}_{t+1} \rangle + \langle \nabla_t, \hat{y}_{t+1} - f^* \rangle
\]

By the mirror descent update, \( \nabla_t = \frac{1}{\eta} (\nabla R(\hat{y}_t) - \nabla R(\hat{y}_{t+1})) \)

\[
= \langle \nabla_t, \hat{y}_t - \hat{y}_{t+1} \rangle + \frac{1}{\eta} \langle \nabla R(\hat{y}_t) - \nabla R(\hat{y}_{t+1}), \hat{y}_{t+1} - f^* \rangle
\]

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

\[
= \langle \nabla_t, \hat{y}_t - \hat{y}_{t+1} \rangle + \frac{1}{\eta} (\Delta_R(f^*|\hat{y}_t) - \Delta_R(f^*|\hat{y}_{t+1}) - \Delta_R(\hat{y}_t|\hat{y}_{t+1}))
\]
\[ \langle a, b \rangle \leq \|a\| \|b\| \leq \frac{\eta}{2} \|b\|_2^2 + \frac{1}{2\eta} \|a\|^2 \]

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

Summing over we have,

\[
\sum_{t=1}^{n} \langle \nabla_t, \hat{y}_t \rangle - \sum_{t=1}^{n} \langle \nabla_t, f^* \rangle \leq \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla_t\|_2^2 + \frac{1}{\eta} \sum_{t=1}^{n} (\Delta_R(f^* | \hat{y}_t) - \Delta_R(f^* | \hat{y}'_{t+1}))
\]

Replacing by projection only decreases the Bregman divergence

\[
\leq \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla_t\|_2^2 + \frac{1}{\eta} \sum_{t=1}^{n} (\Delta_R(f^* | \hat{y}_1) - \Delta_R(f^* | \hat{y}_{n+1}))
\]

Dividing through by \( n \) we prove the claim.

\[ \square \]

### 2.1 Examples

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

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

Note that

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

In this case, it is not too hard to check that \( R \) is strongly convex w.r.t. \( \|\cdot\|_1 \). Also note that \( \sup_{f \in \Delta(F')} R(f) \leq \log |F'| \) (achieved at the uniform distribution).
\textbf{\(\ell_p\) and \textit{Schatten}_p\ norms}  Let us consider \(\mathcal{F}\) to be unit ball under \(\ell_p\) norm and \(\mathcal{D}\) to be unit ball under dual norm. Let \(p \in (1, 2]\), then one can use \(R(f) = \frac{1}{p-1} \|f\|_p^2\) and this function is strongly convex w.r.t. \(\ell_p\) norm. For matrices with analogous Schatten \(p\) norm, use the \(R(f) = \frac{1}{p-1} \|f\|_{Sp}^2\).

\textbf{Remark 2.1.} For \(\ell_1\) norm one can use \(R(f) = \frac{1}{p-1} \|f\|_p^2\) 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 \(\ell_\infty\) norm of \(\nabla t\)’s.