COS-511: Learning Theory **Spring 2017**

Lecture 19: Follow The Regulerized Leader

Lecturer: Roi Livni

19.1 Regularization

One aspect of Statistical Learning Theory is that learnability is captured, at least for classification, through ERM. Namely, the ERM algorithm can be considered as a universal algorithm which either solves the learning problem or the problem is not learnable.

ERM, is also known as a "Follow The Leader" approach where at each iteration we choose $\mathbf{x}_t$ so that

$$\mathbf{x}_t = \arg\min_{\mathbf{x}^* \in \mathcal{K}} \sum_{i=1}^{t-1} f_i(\mathbf{x}^*)$$

While this approach works for stochasticaly chosen f_i (with bounded Rademacher Complexity). It is easy to show that in the online setting, this approach will fail.

Indeed, while the regret bounds achievable thorugh online learning are similar in spirit to those attained in the statisitcal setting, we've seen that we need to "tailor" a different algorithms for different problems – Linear classification is solved via Online Gradient Descent, while the expert problem has been solved using an MW algorithm.

In this section we will try to connect these two algorithm by a generic meta-algorithm, where both algorithms can be considered a special case. The algorithm we will consider is called "Follow the Regulerized Leader" or "FTRL".

Algorithm 1 Follow The Regulerized Leader

Inititalization regularization function $R(\mathbf{x})$, $\eta > 0$
Let $\mathbf{x}_1 = \arg\min_{\mathbf{x} \in \mathcal{K}} \{R(\mathbf{x})\}$.
for $t = 1, 2 \dots T$ **do**
  Play $\mathbf{x}_t$ and observe cost $f_t(\mathbf{x}_t)$
  Set $\nabla_t = \nabla f_t(\mathbf{x}_t)$.
  Update

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{K}} \left\{ \eta \sum_{s=1}^{t} \nabla_s^\top \mathbf{x} + R(\mathbf{x}) \right\}$$

end for
return

We will consider in this lecture regularization function $R(\mathbf{x})$ which are twice differentiable and are strongly convex, this means that for all $\mathbf{x} \in \text{int}(\mathcal{K})$, the Hessian $\nabla^2 R(\mathbf{x})$ is strictly larger than zero. We denote the

diameter of the set $\mathcal{K}$ relative to a function R)

$$D_R = \max_{\mathbf{x},\mathbf{y}\in\mathcal{K}} \sqrt{R(\mathbf{x}) - R(\mathbf{y})}$$

We wll make use of general norms and their dual. The dual of a norm $\|\cdot\|$ is given by the following definition

$$\|\mathbf{y}\|^* = \max_{\|\mathbf{x}\|\leq 1} \mathbf{x}\cdot\mathbf{y}.$$

Every positive definite matrix A give rise to a norm $\|\mathbf{x}\|_A = \mathbf{x}^\top A\mathbf{x}$ and its dual is given by $\|\mathbf{x}\|_A{}^* = \|\mathbf{x}\|_{A^{-1}}$. The generalized Cauchy Schwartz asserts that $\mathbf{x}\cdot\mathbf{y} \leq \|\mathbf{x}\|\|\mathbf{y}\|^*$.

In our derivations, we will usually consider matrix norms with respect to $\nabla^2 R(\mathbf{x})$ (which we assume to be p.s.d). For brevity, we will use the notation

$$\|\mathbf{x}\|_{\mathbf{y}} = \|\mathbf{x}\|_{\nabla^2 R(\mathbf{y})}$$

, and similarly

$$\|\mathbf{x}\|_{\mathbf{y}}^* = \|\mathbf{x}\|_{\nabla^{-2} R(\mathbf{y})}$$

Definition 19.1. *Denote by $B_R(\mathbf{x}\|\mathbf{y})$ the Bregman divergence w.r.t a function R:*

$$B_R(\mathbf{x}\|\mathbf{y}) = R(\mathbf{x}) - R(\mathbf{y}) - \nabla R(\mathbf{y})^\top(\mathbf{x}-\mathbf{y}).$$

For twice differentiable functions, the mean value theorem and Taylor expansion asserts that the Bregman divergence equal to the second derivative at an intermediate point:

$$B_R(\mathbf{x}\|\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_{\mathbf{z}}^2$$

For some point $\mathbf{z}\in[\mathbf{x},\mathbf{y}]$. (i.e. $\mathbf{z} = \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$) for some $0\leq\alpha\leq 1$). Thus we will denote

$$B_R(\mathbf{x}\|\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_{\mathbf{x},\mathbf{y}}^2$$

Finally, we will consider projections that use the Bregman divergence as a distance instead of a norm. Formally, the projection of a point $\mathbf{y}$ to a set $\mathcal{K}$ is given by

$$\arg\min_{\mathbf{x}\in\mathcal{K}} B_R(\mathbf{x}\|\mathbf{y}).$$

19.1.1 Follow The Regulerized Leader

The algorithm we will analyze in this section is FTRL depicted in Alg. 1. It is a meta-algorithm that depends on choice of $R(\mathbf{x})$ as we will later see we can derive from FTRL OGD as well as Hedge, by proper choice of R.

Theorem 19.2. *The RFTL algorithm depicted in Alg. 1 attains for every $\mathbf{u}\in\mathcal{K}$ the followin bound on the regret :*

$$\text{Regret}_T \leq 2\eta\sum_{t=1}^{T}\|\nabla_t\|_{\mathbf{x}_t,\mathbf{x}_{t+1}}^* + \frac{D_R}{\eta}$$

In an upper bound on the local norms is known i.e. $\|\nabla_t\|^*_{\mathbf{x}_t,\mathbf{x}_{t+1}} \leq G_R$, by proper choice of η we can attain the regret bound:

$$\text{Regret}_T \leq 2D_R G_R \sqrt{2T}$$

To prove Thm. 19.2 we begin by analyzing the "stability" of the algorithm

Lemma 19.3. *For every $\mathbf{u} \in \mathcal{K}$ Alg. 1 guarantees the following regret bound*

$$\text{Regret}_T \leq \sum_{t=1}^{T} \nabla_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{1}{\eta} D_R^2$$

Proof. Denote

$$g_0(\mathbf{x}) = \frac{1}{\eta} R(\mathbf{x}) \quad g_t(\mathbf{x}) = \nabla_t \cdot \mathbf{x}.$$

By convexity we have that:

$$\sum f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*) \leq \nabla_t \cdot (\mathbf{x}_t - \mathbf{x}^*)$$

Hence we only need to bound $\sum_{t=1}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{u})$. As a first step we prove the following inequality

$$\sum_{t=0}^{T} g_t(\mathbf{u}) \geq \sum_{t=0}^{T} g_t(\mathbf{x}_{t+1}) \tag{19.1}$$

The proof is by induction, since $\mathbf{x}_1 = \arg\min_{\mathbf{x}\in\mathcal{K}} R(\mathbf{x})$, the base case follows. For the induction step, assume that statement is true for T', since $\mathbf{x}_{T'+2} = \arg\min \sum_{t=0}^{T'+1} g_t(\mathbf{x})$, we have:

$$\begin{aligned}
\sum_{t=0}^{T'+1} g_t(\mathbf{u}) &\geq \sum_{t=0}^{T'+1} g_t(\mathbf{x}_{T'+2}) \\
&= \sum_{t=0}^{T'} g_t(\mathbf{x}_{T'+2}) + g_{T'+1}(\mathbf{x}_{T'+2}) \\
&\geq \sum_{t=0}^{T'} g_t(\mathbf{x}_{t+1}) + g_{T'+1}(\mathbf{x}_{T'+2}) \\
&= \sum_{t=0}^{T'+1} g_t(\mathbf{x}_{t+1})
\end{aligned}$$

We conclude that

$$\begin{aligned}
\sum_{t=1}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{u}) &\leq \sum_{t=1}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1}) + g_0(\mathbf{u}) - g_0(\mathbf{x}_1) \\
&= \sum_{t=1}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1}) + \frac{1}{\eta}\left(R(\mathbf{u}) - R(\mathbf{x}_1)\right) \\
&\leq \sum_{t=1}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1}) + \frac{1}{\eta} D_R^2
\end{aligned}$$

□

Proof of Thm. 19.2. Denote

$$\Phi_t(\mathbf{x}) = \eta \sum_{s=1}^{t} \nabla_s \cdot \mathbf{x} + R(\mathbf{x})$$

Note that for the Bregman divergence if f and g differ by a linear term, then $B_f = B_g$, as a result we have that $B_{\Phi_t} = B_R$:

$$\begin{aligned}\Phi_t(\mathbf{x}_t) &= \Phi_t(\mathbf{x}_{t+1}) + (\mathbf{x}_t - \mathbf{x}_{t+1})\nabla\Phi_t(\mathbf{x}_{t+1}) + B_R(\mathbf{x}_t\|\mathbf{x}_{t+1}) \\ &\geq \Phi_t(\mathbf{x}_{t+1}) + B_R(\mathbf{x}_t\|\mathbf{x}_{t+1}) \\ &= \Phi_t(\mathbf{x}_{t+1}) + B_R(\mathbf{x}_t\|\mathbf{x}_{t+1})\end{aligned}$$

Where the last inequlaity is true since $\mathbf{x}_{t+1}$ is the minimizer of Φ_t, over $\mathcal{K}$. We thus have

$$\begin{aligned}B_R(\mathbf{x}_t\|\mathbf{x}_{t+1}) \leq \Phi_t(\mathbf{x}_t) - \Phi_t(\mathbf{x}_{t+1}) & \\ &\leq \Phi_{t-1}(\mathbf{x}_t) - \Phi_{t-1}(\mathbf{x}_{t+1}) + \nabla_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \\ &\leq \nabla_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1})\end{aligned}$$

Where the last inequality is true since $\mathbf{x}_t$ is the minimizer for Φ_{t-1}. Next, recall that the Bregman divergence induces a norm which we will denote as:

$$B_R(\mathbf{x}_t\|\mathbf{x}_{t+1}) = \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_{\mathbf{x}_t,\mathbf{x}_{t+1}} := \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_t.$$

By the generelized Cauchy Schwartz we have that:

$$\begin{aligned}\nabla_t \cdot (\mathbf{x}_t - \mathbf{x}_{t+1}) &\leq \|\nabla_t\|_t^*\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_t \\ &= \|\nabla_t\|_t^* \cdot \sqrt{2B_R(\mathbf{x}_t\|\mathbf{x}_{t+1})} \\ &\leq \|\nabla_t\|_t^* \cdot \sqrt{2\eta\nabla_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1})}\end{aligned}$$

After rearranging we get:

$$\nabla_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \leq 2\eta\|\nabla_t\|_t^{*2}$$

The result now follows from Lem. 19.3. □

19.1.2 Application

19.1.2.1 Deriving OGD

To derive OGD from FTRL let us consider the regularization function $R(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$. The FTRL update rule is then:

$$\begin{aligned}\text{minimize } & \frac{1}{2}\|\mathbf{x}\|^2 + \eta \sum \nabla_s \cdot \mathbf{x} \\ & \|\mathbf{x}\| \leq 1.\end{aligned}$$

Then the solution satisfies $\mathbf{x}_{t+1} = -\eta\sum_{i=1}^{s}\nabla_s = \mathbf{x}_t\eta\nabla_t$. To derive the regert bound from the analysis of FTRL we bound D_R and $\|\nabla_t\|_t^*$.

$$D_R = \frac{1}{\sqrt{2}}\sqrt{\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2} \leq \sqrt{2}\max\|\mathbf{x}\| \leq 2$$

The norm $\|\cdot\|^*$ is given by the dual norm for $\nabla^2 R(\mathbf{z}) = \mathbf{Id}$ for some point in $[\mathbf{x}_t, \mathbf{x}_{t+1}]$: hence $\|\nabla_t\|_T^* = \|\nabla_t\| = G$.

19.1.2.2 Deriving Hedge

We next show how to derive Hedge from FTRL. Let $R(\mathbf{x}) = \mathbf{x}\log\mathbf{x} = \sum \mathbf{x}_i \log \mathbf{x}_i$. $R(\mathbf{x})$ is the negative entropy function. Regulerizing distribution using $R(\mathbf{x})$ leads to an algorithm that tries to minimize the regret while prefering distributions with high entropy: such as unifrom distribution. The Hedge algorithm indeed starts with the uniform distribution which has the highest entopy amongst all distributions over a finite set. We have that $\nabla R(\mathbf{x}) = 1 + \log \mathbf{x}$. Considering the problem

$$\begin{aligned} &\text{minimize } R(\mathbf{x}) + \sum \nabla_s \cdot \mathbf{x} \\ &\text{s.t. } \sum \mathbf{x}_i = 1 \end{aligned}$$

Taking the Lagrangian, a solution will be optimal if it will satisfy

$$\eta \sum \nabla_t + 1 + \log \mathbf{x} - \lambda \cdot 1 = 0$$

in turn we have

$$\mathbf{x}_{t+1} = \Pi_{\Delta_n}\left(e^{-\eta \sum \nabla_t + \lambda}\right)$$

Thus taking the distribution proportional to $e^{-\eta \sum \nabla_t}$ will indeed lead to an optimal solution[1] Next, the regret term depends on $\max_{\mathbf{x},\mathbf{y}} \sqrt{R(\mathbf{x}) - R(\mathbf{y})} \leq 2\max_{\mathbf{x}} \sqrt{R(\mathbf{x})} = 2\sqrt{\log n}$ Finally, recall that the norm $\|\cdot\|_t$ is given by the second derivative of $R(\mathbf{z})$ at a point $\mathbf{z} \in \Delta_n$ and we have that $\nabla^2 R(\mathbf{z}) = \text{diag}(\frac{1}{\mathbf{z}})$. Hence

$$\|\nabla_t\|_t^{*2} = \nabla_t \nabla^{-2} R(\mathbf{z}) \nabla_t = \sum \mathbf{z}_i \nabla_t^2(i) \leq \max_i \nabla^2(i) = \max_i g_t(i)^2 \leq 1$$

Plugging these magnitudes into Thm. 19.2 we obtain the known regret of

$$\text{Regret}_T = O(\sqrt{T \log n}).$$

[1]Note that we neglected the constraint $\mathbf{x} \geq 0$, but since the solution satisfies this condition, we get that in particular it is optimal together with the given constraint