LECTURE #14

LECTURER: ELAD HAZAN SCRIBE: LYDIA LIU

1 Regularization

1.1 RFTL

In the last lecture we discussed RFTL, an algorithm that arose naturally in the online learning community. There is clear intuition to motivate it too—we may obtain more stable solutions across iterations of the online learning algorithm by adding a regularization function.

To recap, the RFTL update is:

$$x_{t+1} := \arg\min_{x \in \mathcal{K}} \left\{ \eta \sum_{i=1}^{t} \nabla_i \cdot x + R(x) \right\}$$

and the regret bound is:

$$Regret(RFTL) \leq \frac{1}{\eta}[R(x_1) - R(x^*)] + 2\eta \sum_{t=1}^{T} \|\nabla_t\|^{*}_{\nabla^2 R(z_t)}{}^2 = O(\sqrt{T})$$

1.2 Mirrored Descent

Given R, $\nabla R : \mathbb{R}^d \to \mathbb{R}^d$ is a vector field, the updates for the Mirrored Descent algorithm are:

$$\nabla R(y_{t+1}) = \nabla R(x_t) - \eta \nabla_t$$
$$x_{t+1} = \Pi_{\mathcal{K}}^{B_R}(y_{t+1}) = \arg\min_{x \in \mathcal{K}} B_R(x, y_{t+1})$$

Unlike RFTL, the intuition behind Mirrored Descent (MD) seems less clear, but one can show under fairly general conditions that

$$x_t^{MD} = x_t^{RFTL} \text{ with the same regularization function } R$$

Also, note that if R is the squared euclidean norm $\|\cdot\|^2$, then MD is the gradient descent algorithm. If R is negative entropy, $\sum_i x_i \log x_i$, then MD is the multiplicative weights algorithm.

If we also optimize the parameter, η, we get the same regret bound:

$$Regret(MD) = Regret(RFTL) \leq 2\sqrt{2D_R \sum_{t=1}^{T} \|\nabla_t\|^{*}_{\nabla^2 R(z_t)}{}^2}$$

1.3 Motivating adaptive regularization

This leads us to the question: what is the best R to choose to minimize regret? Clearly, it is more important to optimize the term $\sum_{t=1}^{T} \|\nabla_t\|^{*}_{\nabla^2 R(z_t)}{}^2$ than the term D_R, since the former is a sum that increases with T.

We saw in SGD that $\overline{x}_t = \frac{1}{T}\sum_t x_t$ and that

$$\mathbb{E}[f(\overline{x}_t)] \leq \min_{x^*} f(x^*) + \frac{regret_T}{T} \text{ where } \frac{regret_T}{T} \approx \frac{1}{\sqrt{T}}$$

which is state-of-the-art.

If we apply matrix-norm regularization, i.e. $R(x) = \frac{1}{2}x^T Ax$, in RFTL, then the average regret term will be $\frac{\sqrt{D_R \sum_t \|\nabla_t\|^*_{A^{-1}}{}^2}}{T}$. Thus it is reasonable for us to try to optimize regret over the choice of matrix A.

Here we sketch the idea for a simplified example. Suppose $\nabla_t \in \begin{pmatrix} \pm 1 \\ \pm 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \subseteq \mathbb{R}^d$. We introduce an important definition for the set of matrices that we want to restrict ourselves to consider.

Definition 1.1. The spectohedron is the set of matrices

$$\mathcal{S}_n := \{X : X \succeq 0, Tr(X) \leq 1, X \in \mathbb{R}^{n \times m}\}$$

What is the best $A \in \mathcal{S}_n$ minimizing $\frac{\sqrt{\sum_t \|\nabla_t\|^*_{A^{-1}}{}^2}}{T}$ in this case? Since ∇_t is only non-zero in its first 2 coordinates, it makes sense to have non-negative weights only in the top left 2×2 submatrix of A.

When restricted only to the set $\mathcal{S}_n$, we can in fact learn the best A for regularization and get approximately the same asymptotic performance as gradient descent.

For the rest of this lecture, we will not concern ourselves with the question of whether a matrix is invertible, since we can perturb a singular matrix with $\delta\mathbf{I}$ where δ is vanishing, or just take the pseudoinverse.

2 AdaGrad

We introduce the AdaGrad algorithm:

- Initialize $S_0 = G_0 = \delta\mathbf{I}$, $x_1 \in \mathcal{K}$
- For $t = 1$ to T, do:
  1. Predict x_t, suffer loss $f_t(x_t)$
  2. Update:

$$S_t = S_{t-1} + \nabla_t \nabla_t^T,\ G_t = \sqrt{S_t}$$
$$y_{t+1} = x_t - G_t^{-1}\nabla_t$$
$$x_{t+1} = \Pi_{\mathcal{K}}^{G_t}(y_{t+1})$$

Projection step 'optional' because in reality we never step outside of $\mathcal{K}$

A note on computational efficiency: another version of AdaGrad deals with the time consuming matrix square root and inversion steps by defining $\hat{S}_t = diag(S_t), G_t = \sqrt{\hat{S}_t}$, so everything can be accomplished in linear time. The regret bound for this version is asymptotically the same as the usual AdaGrad (only slightly worse theoretical guarantees), and is popular in real world applications. (Note: $\|\cdot\|^*_A{}^2 = \|\cdot\|_{A^{-1}}{}^2$).

We state and prove the regret bound for the usual AdaGrad as follows:

Theorem 2.1. $Regret(AG)_T = O\left(\sqrt{\min_{A \in \mathcal{S}_n} \sum_{t=1}^T \|\nabla_t\|^*_A{}^2}\right)$

Proof. We use the following fact: Let $B \succeq 0$. Then

$$\arg\min_{A \in \mathcal{S}^n} A^{-1} \circ B = \frac{B^{1/2}}{Tr(B^{1/2})}$$

where for symmetric matrices M, N, $M \circ N := Tr(MN)$.

This fact leads to the following observation. Notice that there is a closed form expression for the 'best' A^{-1} norm:

$$\begin{aligned}\arg\min_{A \in \mathcal{S}^n} \sum_{t=1}^{T} {\|\nabla_t\|_A^*}^2 &= \arg\min_{A \in \mathcal{S}^n} A^{-1} \circ S_T \\ &= \frac{S_T^{1/2}}{Tr(S_T^{1/2})} \\ &= \frac{G_T}{Tr(G_T)}\end{aligned}$$

Thus, $\min \sqrt{\sum_{t=1}^{T} {\|\nabla_t\|_A^*}^2} = \sqrt{\left(\frac{G_T}{Tr(G_T)}\right)^{-1} \circ {G_T}^2} = Tr(G_T)$

It now suffices to prove:

$$Regret(AG)_T = O(Tr(G_T))$$

Define $D = \max_{u \in \mathcal{K}} \|u - x_1\|_2$.

$$\begin{aligned}\|x_{t+1} - x^*\|_{G_t}^2 &\leq \|y_{t+1} - x^*\|_{G_t}^2 \text{ (because of projection)} \\ &= \|x_t - G_t^{-1}\nabla_t - x^*\|_{G_t}^2 \\ &= \|x_t - x^*\|_{G_t}^2 - 2\nabla_t^T(x_t - x^*) + \nabla_t^T G_t^{-1} \nabla_t \\ f_t(x_t) - f_t(x^*) &\leq \nabla_t(x_t - x^*) \\ &\leq \|x_t - x^*\|_{G_t}^2 - \|x_{t+1} - x^*\|_{G_t}^2 + \nabla_t^T G_t^{-1} \nabla_t \\ \text{Summing over } t = 1 \text{ to } T,\ 2 \times Regret(AG) &\leq \sum_{t=1}^{T} \|x_t - x^*\|_{G_t}^2 - \|x_{t+1} - x^*\|_{G_t}^2 + \nabla_t^T G_t^{-1} \nabla_t \\ &= \sum_{t=1}^{T} (x_t - x^*)^T (G_t - G_{t-1})(x_t - x^*) + \sum_{t=1}^{T} \|\nabla_t\|_{G_t^{-1}}^2 + O(1)\end{aligned}$$

Looking at the first term,

$$\begin{aligned}\sum_{t=1}^{T} (x_t - x^*)^T (G_t - G_{t-1})(x_t - x^*) &= \sum_{t=1}^{T} Tr\left((G_t - G_{t-1})(x_t - x^*)(x_t - x^*)^T\right) \\ &\leq \sum_{t=1}^{T} Tr\left(G_t - G_{t-1}\right) \|(x_t - x^*)(x_t - x^*)^T\|_2 \text{ by Hölder's inequality} \\ &\leq D^2 \sum_{t=1}^{T} Tr(G_t) - Tr(G_{t-1}) \\ &\leq D^2 Tr(G_T)\end{aligned}$$

Looking at the second term, we can prove by induction that

$$\sum_{t=1}^{T} \|\nabla_t\|_{G_t^{-1}}^2 \leq 2 \sum_{t=1}^{T} \|\nabla_t\|_{G_T^{-1}}^2 = 2Tr(G_T)$$

Check that the result holds for $t = 1$. Now, assume it holds for $t = T$, check the induction hypothesis for $t = T + 1$:

$$\begin{aligned}\sum_{t=1}^{T+1} \nabla_t^T G_t^{-1} \nabla_t &\leq 2Tr(G_T) + \nabla_{T+1}^T G_{T+1}^{-1} \nabla_{T+1} \\ &\leq 2Tr((G_{T+1}^2 - \nabla_{T+1}\nabla_{T+1}^T)^{1/2}) + Tr(G_{T+1}^{-1}\nabla_{T+1}\nabla_{T+1}) \\ &\leq 2Tr(G_T)\end{aligned}$$

where the last inequality is due to the following matrix inequality:

$$2Tr((A - B)^{1/2}) + Tr(A^{-1/2}B) \leq 2Tr(A^{1/2})$$

Together, this proves that $Regret(AG)_T \leq O(1) \cdot Tr(G_T)$

□