CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 24**

Online convex optimization: ridge regression, adaptivity

Lecturer: Sasha Rakhlin *Scribe: Alekh Agarwal, Lester Mackey*

1 Lecture Outline

In the previous lectures we saw a general scheme for deriving regret bounds for online learning algorithms that try to minimize a regularized loss at every time step. We also argued that the linear loss is the optimal loss for the adversary if it is constrained to play convex loss functions. The goal of this lecture is to obtain logarithmic regret bound when the adversary plays sub-optimally in that it plays curved loss functions. Further, we will see how a time varying learning rate can be used to smoothly interpolate between the regret regimes of $\log T$ and $\sqrt{T}$ corresponding to curved and linear losses respectively.

2 Curvature in Online learning

2.1 A rewriting of the regret bound

Last time we saw the game, where at every step we solve the problem:

$$x_{t+1} = \text{argmin}_{x \in \mathbb{R}^n} \quad \eta \sum_{s=1}^{t} \ell_s(x) + R(x) \tag{1}$$

Then we showed last time that $\forall u \in \mathbb{R}^n$:

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) = \eta^{-1} D_{\Phi_0}(u, x_1) - \eta^{-1} D_{\Phi_T}(u, x_{T+1}) + \eta^{-1} \sum_{t=1}^{T} D_{\Phi_t}(x_t, x_{t+1}) \tag{2}$$

where $\Phi_0 = R$, $\Phi_t = \Phi_{t-1} + \eta \ell_t(\cdot)$.

Also, we showed in an earlier lecture on properties of Bregman divergences that:

$$\begin{aligned} D_{\Phi_t}(x_t, x_{t+1}) &= D_{\Phi_t^*}(\nabla \Phi_t(x_{t+1}), \nabla \Phi_t(x_t)) \\ &= D_{\Phi_t^*}(0, \nabla \Phi_t(x_t)) \end{aligned} \tag{3}$$

where the second line follows from the fact that x_{t+1} is the minimizer of (1). Also, we have that:

$$\begin{aligned} \nabla \Phi_t(x_t) &= \nabla \Phi_{t-1}(x_t) + \eta \nabla \ell_t(x_t) \\ &= \eta \nabla \ell_t(x_t) \end{aligned} \tag{4}$$

where the second line is again using the fact that x_t minimizes (1) at time t. Combining (3) and (4), we get:

$$D_{\Phi_t}(x_t, x_{t+1}) = D_{\Phi_t^*}(0, \eta \nabla \ell_t(x_t)) \tag{5}$$

This might look like just some algebraic manipulation at the first sight. But note that the divergence we have rewritten is exactly the divergence that gets added up over time. In particular, we will now exploit this simple form to obtain interesting conclusions for the specific case of online ridge regression, which is an instance of a strongly convex loss.

2.2 Online Ridge Regression

Suppose the game is defined as follows:

For $t = 1 \dots T$
  We pick $w_t \in \mathbb{R}^n$
  Adversary picks $(x_t, y_t), \quad (x_t \in \mathbb{R}^n, y_t \in \mathbb{R}, \|x_t\|_2 \leq B)$
  We suffer $\frac{1}{2}(w_t^\top x_t - y_t)^2 = \ell_t(w_t)$.

This game is exactly online ridge regression within a bounded ℓ_2 ball if the regularizer is $\frac{1}{2}\|\cdot\|^2$. We will in fact assume that the y's are bounded too. Further, we assume that $\exists p^*$ such that at every time, $\ell_t(w_t) \leq p^*$. It isn't clear how the boundedness of x and y implies this directly as the weights can still be large, but the assumption is needed for our analysis and will be clarified in the later lectures. Then the regret of this game is:

$$R_T = \frac{1}{2}\sum_{t=1}^{T}(w_t^\top x_t - y_t)^2 - \frac{1}{2}\sum_{t=1}^{T}(x_t^\top w^* - y_t)^2 \tag{6}$$

where w^* is the minimizer of the cumulative loss. Our goal is to demonstrate that $\frac{R_T}{T} \leq O\left(\frac{\log T}{T}\right)$ which is significantly better than the $O\left(\frac{1}{\sqrt{T}}\right)$ rate that we obtained without any assumptions on the losses.

Now for the ridge regularizer $\frac{1}{2}\|\cdot\|^2$, we have:

$$w_{t+1} = \text{argmin}_w \frac{\eta}{2}\sum_{s=1}^{t}(w^\top x_s - y_s)^2 + \frac{1}{2}\|w\|^2 \tag{7}$$

Then we have $\Phi_0(u) = \frac{1}{2}\|u\|^2$ and

$$\begin{aligned}
\Phi_t(u) &= \frac{1}{2}\|u\|^2 + \frac{\eta}{2}\sum_{s=1}^{t}(w^\top x_s - y_s)^2 \\
&= \frac{1}{2}u^\top I u + \frac{\eta}{2}u^\top(\sum_{s=1}^{t} x_s x_s^\top)u + \frac{\eta}{2}\sum_{s=1}^{t} y_s^2 - \eta\sum_{s=1}^{t} u^\top(y_s x_s) \\
&= \frac{1}{2}u^\top(I + \eta\sum_{s=1}^{t} x_s x_s^\top)u - u^\top\eta\sum_{s=1}^{t} y_s x_s + c_t
\end{aligned} \tag{8}$$

where $c_t = \sum_{s=1}^{t} y_s^2$.

We will now state a lemma that allows us to manipulate the above quantity.

Lemma 2.1. If $\Phi(u) = \frac{1}{2}u^\top M u + u^\top v + c$ for M positive definite (elliptic potential), then:

(a) $\nabla\Phi(u) = Mu + v$

(b) $\Phi^*(u) = \frac{1}{2}u^\top M^{-1}u - u^\top M^{-1}v + \frac{1}{2}v^\top M^{-1}v$

(c) $\nabla\Phi^*(u) = M^{-1}u - M^{-1}v$

(d) $D_\Phi(u, w) = (u-w)^\top M(u-w)$

(e) $D_{\Phi^*}(u, w) = (u-w)^\top M^{-1}(u-w)$

All of these follow from simple matrix algebra and will not be proved.

Let $A_t = I + \eta\sum_{s=1}^t x_s x_s^\top$. Then it is easily seen to be positive definite as the identity matrix is positive definite, and the rank-one matrices $x_s x_s^\top$ are positive semidefininte. So we have an elliptic potential in the ridge regression updates and can apply the above lemma to our case. Substituting A_t and $v_t = \eta\sum_{s=1}^t x_s y_s$ in (8) we get:

$$\begin{aligned}
\Phi_t(u) &= \frac{1}{2}u^\top A_t u - u^\top v_t + c_t \\
D_{\Phi_t^*}(0, \eta\nabla\ell_t(w_t)) &= \eta^2\nabla\ell_t(w_t)^\top A_t^{-1}\nabla\ell_t(w_t) \\
\nabla\ell_t(w_t) &= (w_t^\top x_t - y_t)x_t \\
D_{\Phi_t^*}(0, \eta\nabla\ell_t(w_t)) &= \eta^2(w_t^\top x_t - y_t)^2 x_t^\top A_t^{-1} x_t
\end{aligned}$$

By the earlier assumption of bounded losses, $(w_t^\top x_t - y_t)^2 \le 2p^*$. Further set $\eta = 1$ and $w_1 = \vec{0}$ to be the zero vector. Then using equations (2), (5) and above results for our specific potential we get:

$$\sum_{t=1}^T(\ell_t(w_t) - \ell_t(u)) \le \frac{1}{2}\|u\|^2 + \sum_{t=1}^T 2p^* x_t^\top A_t^{-1} x_t \tag{9}$$

Now intuitively we expect A_t to grow linearly with t as it keeps on accumulating identical terms over time, so that A_t^{-1} goes as $1/t$ giving us $\log T$ regret. We will now make this intuition concrete using the following lemma about quadratic forms:

Lemma 2.2. Let B be an arbitrary $n \times n$ full rank matrix and x be any vector. Let $A = B + xx^\top$. Then $x^\top A^{-1}x = x^\top(B + xx^\top)^{-1}x = 1 - \frac{\det(B)}{\det(A)}$

where $\det(A)$ is the determinant of the matrix A. In our setup, we intend to apply this lemma with $B = A_{t-1}$, $x = x_t$ and $A = A_t$ to the quadratic form of (9) which gives us:

$$\begin{aligned}
R_T &\le \frac{1}{2}\|u\|^2 + 2p^*\sum_{t=1}^T\left(1 - \frac{\det(A_{t-1})}{\det(A_t)}\right) \\
&\le \frac{1}{2}\|u\|^2 - 2p^*\sum_{t=1}^T\log\frac{\det(A_{t-1})}{\det(A_t)} \quad (\text{using } \; 1 - x \le -\log x \;\; \forall x > 0) \\
&= \frac{1}{2}\|u\|^2 + 2p^*\log\frac{\det(A_T)}{A_0} \\
&= \frac{1}{2}\|u\|^2 + 2p^*\log\det(A_T)
\end{aligned} \tag{10}$$

where the first equality follows from telescoping of terms in the summation, and the second one follows from the fact that $\det(A_0) = \det(I) = 1$.

Next we note that

$$\log(\det(A_T)) = \log\left(\det\left(I + \sum_{t=1}^T x_t x_t^\top\right)\right) = \log\left(\prod_{i=1}^n(1+\lambda_i)\right) = \sum_{i=1}^n\log(1+\lambda_i)$$

where $\lambda_1, ..., \lambda_n$ are the eigenvalues of the matrix $\sum_{t=1}^T x_t x_t^\top$. The eigenvalues of $\sum_{t=1}^T x_t x_t^\top$ are known to be equal to the eigenvalues of the Gram matrix G, where $G_{ij} = x_i^\top x_j$. This implies

$$\sum_{i=1}^n \lambda_i = \sum_{t=1}^T x_t^\top x_t \le TB^2.$$

Our expression for $\log(\det(A_T))$ is therefore maximized when $\lambda_1 = \lambda_2 = ... = \lambda_n = \frac{TB^2}{n}$. We conclude that

$$R_T \le \frac{1}{2}\|u\|^2 + 2p^* n \log\left(1 + \frac{TB^2}{n}\right) \tag{11}$$

2.3 Mirror Descent with Euclidean Norm Squared

In this section, we propose an algorithm that can exploit the suboptimal play of an adversary who plays only curved (strictly convex) loss functions. Consider again the mirror descent algorithm with $R(x) = \frac{1}{2}\|x\|^2$:

For $t = 1 \dots T$
  Play x_t
  Observe $\ell_t(\cdot)$
  Update $x_{t+1} = x_t - \eta_{t+1}\nabla\ell_t(x_t)$.

After observing $\ell_t(\cdot)$, we can calculate

$$H_t \ge 0, \quad s.t. \quad \nabla^2\ell_t \succeq H_t I \tag{12}$$
$$G_t \ge 0, \quad s.t. \quad G_t = \|\nabla\ell_t(x_t)\| \tag{13}$$

where ∇^2 is the Hessian, and $A \succeq B$ means $A - B$ is positive semidefinite.

The following lemma bounds our regret in terms of H_t and G_t.

Lemma 2.3. Define $H_{1:t} = \sum_{s=1}^t H_s$. If we set $\eta_{t+1} = \frac{1}{H_{1:t}}$ in $x_{t+1} = x_t - \eta_{t+1}\nabla\ell_t(x_t)$, then $R_T \le \frac{1}{2}\sum_{t=1}^T \frac{G_t^2}{H_{1:t}}$.

Proof. Let $x^* = \operatorname{argmin}_x \sum_{t=1}^T \ell_t(x)$ and $\nabla_t = \nabla\ell_t(x_t)$. By (12),

$$\ell_t(x_t) - \ell_t(x^*) \le \nabla_t^\top(x_t - x^*) - H_t\frac{1}{2}\|x_t - x^*\|^2.$$

Further,

$$\frac{1}{2}\|x_{t+1} - x^*\|^2 \le \frac{1}{2}\|x_t - x^*\|^2 - \eta_{t+1}\nabla_t^\top(x_t - x^*) + \eta_{t+1}^2\frac{1}{2}\|\nabla_t\|^2,$$

which implies

$$\nabla_t^\top(x_t - x^*) \le \frac{\frac{1}{2}\|x_t - x^*\|^2 - \frac{1}{2}\|x_{t+1} - x^*\|^2}{\eta_{t+1}} + \frac{\eta_{t+1}}{2}G_t^2.$$

Combining, we get

$$\ell_t(x_t) - \ell_t(x^*) \le \frac{\frac{1}{2}\|x_t - x^*\|^2 - \frac{1}{2}\|x_{t+1} - x^*\|^2}{\eta_{t+1}} + \frac{\eta_{t+1}}{2}G_t^2 - H_t\frac{1}{2}\|x_t - x^*\|^2,$$

which summed over $t = 1, ..., T$ gives,

$$\sum_{t=1}^T(\ell_t(x_t) - \ell_t(x^*)) \le \sum_{t=1}^T\left(\frac{1}{2}\|x_t - x^*\|^2\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - H_t\right)\right) + \frac{1}{2}\sum_{t=1}^T G_t^2\eta_{t+1}.$$

Since $\forall t$, $\frac{1}{2}\|x_t - x^*\|^2 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - H_t\right) \leq 0$, we have established

$$R_T \leq \frac{1}{2}\sum_{t=1}^{T} \frac{G_t^2}{H_{1:t}}.$$

□

Note that our procedure does not require that we know the curvature of an adversary's functions in advance. With the sole requirement that our adversary play only curved functions, our algorithm provides a logarithmic regret bound. To see this, note that when $H_1 = ... = H_T = \sigma$,

$$R_T \leq \frac{1}{2}\sum_{t=1}^{T} \frac{G_t^2}{t\sigma} = O(\log(T)).$$

However, if the adversary begins to play linear functions at some time step, we do not achieve the optimal $O(\sqrt{T})$ regret guarantee for arbitrary convex functions: our regret bound instead grows like $O(T)$. Is it possible to guarantee $O(\sqrt{T})$ growth when arbitrary convex functions are played while still exploiting the suboptimality of curved functions? We explore this issue in the next section.

2.4 Adaptive Mirror Descent

Consider the following modified version of our mirror descent algorithm:

For $t = 1 \dots T$
  Play x_t
  Observe $\ell_t(\cdot)$
  Pretend you observed $\tilde{\ell_t}(\cdot) = \ell_t(\cdot) + \lambda_t \frac{1}{2}\|\cdot\|^2$
  Update $x_{t+1} = x_t - \eta_{t+1}\nabla\tilde{\ell_t}(x_t)$.

If we choose H_t and G_t according to (12) and (13) and set $\eta_{t+1} = \frac{1}{H_{1:t}+\lambda_{1:t}}$, we have

$$\sum_{t=1}^{T}(\ell_t(x_t) + \lambda_t\frac{1}{2}\|x_t\|^2) \leq \sum_{t=1}^{T}(\ell_t(x^*) + \lambda_t\frac{1}{2}\|x^*\|^2) + \frac{1}{2}\sum_{t=1}^{T}\frac{(G_t + \lambda_t D)^2}{H_{1:t}+\lambda_{1:t}},$$

where D is the diameter of the set from which vectors x_t are selected.

The purpose of the transformation $\tilde{\ell_t}$ is to inject curvature into a possibly linear function ℓ_t. [1] We want to select λ_t values which increase the curvature of linear functions without overly penalizing those functions which are already curved. That is, we want to choose λ_t values which balance the two terms in the following upper bound:

$$R_T = \sum_{t=1}^{T}\ell_t(x_t) - \sum_{t=1}^{T}\ell_t(x^*) \leq \frac{1}{2}D^2\lambda_{1:T} + \frac{1}{2}\sum_{t=1}^{T}\frac{(G_t + \lambda_t D)^2}{H_{1:t}+\lambda_{1:t}},$$

As it turns out, the right choice is given by

$$\lambda_t = \frac{1}{2}\left(\sqrt{(H_{1:t}+\lambda_{1:t})^2 + \frac{8G_t^2}{3D^2}} - (H_{1:t}+\lambda_{1:t})\right).$$

[1] Alternatively, we may view the transformation as L-2 regularization, which favors shrinking the player's actions x_t toward the origin.

This small modification to our original mirror descent algorithm allows us to adapt to the curvature of an adversary's functions and to achieve regret guarantees between $O(\log(T))$ and $O(\sqrt{T})$ depending on that curvature.