CMSC 39600: Online Algorithms
Lecture 5
Course Instructor: Adam Kalai
Date: October 8, 2004

Online gradient descent

1 Background

In this lecture, we will present Zinkevich's Online Convex Optimization analysis of gradient descent. As background, let us recall the definition of the gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$. The gradient itself is a function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, which, evaluated at x is:

$$\nabla f(x) = \left(\frac{df}{dx_1}(x), \frac{df}{dx_2}(x), \dots, \frac{df}{dx_n}(x) \right).$$

In one dimension, the gradient is just the derivative. A function is *differentiable* if the gradient exists.

Also a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$f(\alpha x + (1-\alpha)x') \leq \alpha f(x) + (1-\alpha) f(x'), \forall x, y \in \mathbb{R}^n, \alpha \in [0,1].$$

Geometrically, this means that if you draw a line segment between the vectors $(x, f(x))$ and $(x', f(x'))$, the function lies below this line segment. Also, for a convex differentiable function, we have at any x,

$$f(x') - f(x) \geq \big(\nabla f(x)\big) \cdot (x' - x). \tag{1}$$

Geometrically, the above statement means that if you draw the tangent plane to the function at point x, the function lies above it. (Think about it in one dimension....)

The gradient of a function gives the direction of steepest increase. Thus a natural minimization algorithm is to go in the direction opposite the gradient a certain amount. For any $\eta \in \mathbb{R}$, the *gradient descent* algorithm (for minimizing a function f) with learning rate η, chooses a sequence of points $x^1, x^2, \dots$, such that,

$$x^{t+1} = x^t - \eta \nabla f(x^t).$$

In many cases we are restricted to a bounded convex set $S \subset \mathbb{R}^n$, which is the range of $f : S \to \mathbb{R}$. In this case, the above rule may lead us to choose a point $x \notin S$, which is a problem. A common trick is to move to the closest point to x in the set S, which we call $\Pi_S(x)$, i.e.,

$$\Pi_S(x) = \arg\min_{x' \in S} \|x - x'\|.$$

So the official gradient descent update is,

$$x^{t+1} = \Pi_S(x^t - \eta \nabla f(x^t)).$$

The parameter η is often called the learning rate and the set S is the feasible set.

2 Foreground

Zinkevich gave a nice analysis of online gradient descent. He showed the following. Assume $S \subset \mathbb{R}^n$ is a closed convex set of diameter at most D. This means that for every $x, x' \in S, \|x - x'\| \leq D$. Then,

Theorem 1 (Zinkevich 02) *Consider any sequence of differentiable functions* $f^1, f^2, \ldots, f^T : S \to \mathbb{R}$ *such that* $\|\nabla f^t(x)\| \leq G$ *for any* $1 \leq t \leq T, x \in S$, *i.e.* G *is an upper bound on the gradient magnitudes (known in advance). Then for the following sequence,* $x^1, x^2, \ldots, x^T$,

$$x^{t+1} = \Pi_S(x^t - \eta \nabla f^t(x^t)),$$

where $\eta = \frac{D}{G\sqrt{T}}$, *the total of the functions is,*

$$\sum_{t=1}^{T} f^t(x^t) \leq \min_{x \in S} \left(\sum_{t=1}^{T} f^t(x) \right) + DG\sqrt{T}$$

Proof. The convex function has a minimum (at least 1) somewhere in S. Translate space so that 0 is this minimum. This is without loss of generality because translations will not change how the algorithm works. We now need to bound the regret with respect to $\sum f^t(0)$.

We use the "potential function" method of proof. Define the potential $\Phi_t = -\frac{1}{2\eta}\|x^t\|^2$. We will show that,

$$f^t(x^t) - f^t(x^\star) + \Phi_{t+1} - \Phi_t \leq \eta G^2/2. \tag{2}$$

Intuitively, we cannot guarantee that on any particular function $f^t(x^t)$ won't be much larger than $f^t(x^\star)$. However, if it is, this means that x^{t+1} will be much closer to $x^\star$ than x^t was.

Summing from $t = 1$ to T, the sum telescopes and we have,

$$\sum_{t=1}^{T} \big(f^t(x^t) - f^t(0)\big) + \Phi_{T+1} - \Phi_1 \leq \eta G^2 T/2.$$

Since $-\frac{D^2}{2\eta} \leq \Phi_t \leq 0$, this gives,

$$\sum_{t=1}^{T} f^t(x^t) \leq \sum_{t=1}^{T} f^t(0) + \frac{D^2 T}{2\eta} + \frac{\eta G^2 T}{2}$$

Plugging in $\eta = \frac{D}{G\sqrt{T}}$ gives the theorem.

It remains to prove (2). This requires two parts. First, we argue that $\|\Pi_S(x)\| \leq \|x\|$. This will follow from the convexity of S and the fact that the origin is assumed to be in S.

Using the vector law of cosines, $\|u+v\|^2 = \|u\|^2 + \|v\|^2 + 2u \cdot v$,

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \frac{1}{2\eta}(\|x^{t+1}\|^2 - \|x^t\|^2) \\
&\leq \frac{1}{2\eta}(\|x^t - \eta\nabla f^t(x^t)\|^2 - \|x^t\|^2) \\
&= \frac{1}{2\eta}(\|x^t\|^2 + \eta^2\|\nabla f^t(x^t)\|^2 - 2\eta\nabla f^t(x^t) \cdot x^t - \|x^t\|^2) \\
&\leq \frac{1}{2}\eta G^2 - \nabla f^t(x^t) \cdot x^t
\end{aligned}$$

Using (1), we have that

$$f^t(0) - f^t(x^t) \geq \nabla f^t(x^t) \cdot (0 - x^t).$$

Combining the last two inequalities gives (2). $\square$