6.883: Online Methods in Machine Learning

Alexander Rakhlin

LECTURE 3

To recap, we talked about Perceptron in the separable case, and introduced a surrogate hinge loss function for the nonseparable case. We sketched the connection to gradient descent, which we shall make precise today. One curious outcome of the Perceptron mistake bound is that the dimension of the space does not enter the bound, while we feel that the problem should be harder in high dimension. The dimension, however, enters implicitly in the margin assumption. In high dimensions, the magnitude of w^* that separates the data with margin of 1 might need to be quite large.

0.1 Review of convex optimization

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if

$$f(\alpha u + (1-\alpha)v) \le \alpha f(u) + (1-\alpha) f(v)$$

for any $\alpha \in [0,1]$ and $u, v \in \mathbb{R}^d$ (or restricted to a convex set). For a differentiable function, convexity is equivalent to monotonicity

$$\langle \nabla f(u) - \nabla f(v), u - v \rangle \ge 0. \tag{1}$$

where

$$\nabla f(u) = \left(\frac{\partial f(u)}{\partial u_1}, \dots, \frac{\partial f(u)}{\partial u_d} \right)$$

It holds that for a convex differentiable function

$$f(u) \ge f(v) + \langle \nabla f(v), u - v \rangle . \tag{2}$$

A subdifferential set is defined (for a given v) precisely as the set of all vectors ∇ such that

$$f(u) \ge f(v) + \langle \nabla, u - v \rangle . \tag{3}$$

for all u. The subdifferential set is denoted by $\partial f(v)$. A subdifferential will often substitute the gradient, even if we don't specify it.

If $f(v) = \max_i f_i(v)$ for convex differentiable f_i, then, for a given v, whenever $i \in \operatorname{argmax}_i f_i(v)$, it holds that

$$\nabla f_i(v) \in \partial f(v).$$

(Prove it!) We conclude that the subdifferential of the hinge loss $\max\{0, 1 - y_t \langle w, x_t \rangle\}$ with respect to w is

$$-y_t x_t \cdot \mathbf{1}\{y_t \langle w, x_t \rangle < 1\} . \tag{4}$$

A function f is L-Lipschitz over a set S with respect to a norm $\|\cdot\|$ if

$$\|f(u) - f(v)\| \leq L \|u - v\|$$

for all $u, v \in S$. A function f is β-smooth if its gradient maps are Lipschitz

$$\|\nabla f(v) - \nabla f(u)\| \leq \beta \|u - v\|,$$

which implies

$$f(u) \leq f(v) + \langle \nabla f(v), u - v \rangle + \frac{\beta}{2} \|u - v\|^2.$$

(Prove that the other implication also holds.) The dual notion to smoothness is that of strong convexity. A function f is σ-strongly convex if

$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v) - \frac{\sigma}{2} \alpha(1-\alpha) \|u - v\|^2,$$

which means

$$f(u) \geq f(v) + \langle u - v, \nabla f(v) \rangle + \frac{\sigma}{2} \|u - v\|^2.$$

0.2 Gradient Descent

Gradient descent in its most basic form is the following iterative procedure.

Algorithm 1 Gradient Descent

Input: $\eta > 0$
Init: $w_1 = 0$
for t=1,...,T **do**
  $w_{t+1} = w_t - \eta \nabla f(w_t)$.
end for

The gradient may be substituted with any subgradient if $\partial f(w_t)$ is not a singleton. There are various motivations for this update. One that will be used later again follows by re-writing the update as an optimization problem

$$\operatorname*{argmin}_w \eta \left[f(w_t) + \langle \nabla f(w_t), w - w_t \rangle \right] + \frac{1}{2} \|w - w_t\|^2 \tag{5}$$

which gives an interpretation of minimizing a linear approximation but also staying close to previous solution.

Showing convergence of this method for a Lipschitz function f is very easy. Suppose we run gradient descent for T steps. Define the average of the trajectory

$$\widehat{w} = \frac{1}{T} \sum_{t=1}^{T} w_t. \tag{6}$$

Next, we use convexity and linearize the function:

$$f(\widehat{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} [f(w_t) - f(w^*)] \leq \frac{1}{T} \sum_{t=1}^{T} \langle \nabla f(w_t), w_t - w^* \rangle \tag{7}$$

We unwind the recursion

$$\|w_{t+1} - w^*\|^2 = \|w_t - \eta \nabla f(w_t) - w^*\|^2 \tag{8}$$

$$= \|w_t - w^*\|^2 - 2\eta \langle \nabla f(w_t), w_t - w^* \rangle + \eta^2 \|\nabla f(w_t)\|^2 \tag{9}$$

Rearranging and summing over $t = 1, \ldots, T$,

$$\sum_{t=1}^{T} \langle \nabla f(w_t), w_t - w^* \rangle = \frac{1}{2\eta} \sum_{t=1}^{T} \left[\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2 \right] + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f(w_t)\|^2 \tag{10}$$

$$\leq \frac{1}{2\eta} \|w^*\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f(w_t)\|^2. \tag{11}$$

Lemma 1. *Suppose f is convex and L-Lipschitz, and $w^* \in \operatorname{argmin}_w f(w)$. By running gradient descent with η for T steps, we find $\widehat{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$ such that*

$$f(\widehat{w}) - f(w^*) \leq \frac{1}{2\eta T} \|w^*\|^2 + \frac{\eta L^2}{2}. \tag{12}$$

If $B \geq \|w^*\|$ is known and T is pre-specified, we may choose $\eta = \frac{B}{L\sqrt{T}}$ and then

$$f(\widehat{w}) - f(w^*) \leq \frac{BL}{\sqrt{T}}. \tag{13}$$

In the lemma we used the simple fact that f is L-Lipschitz iff the norm of any subgradient is bounded by L.

Remark 1. *We may replace $w_t = w_t - \eta \nabla f(w_t)$ with*

$$w_t = Proj(w_t - \eta \nabla f(w_t)),$$

where Proj is a Euclidean projection onto any convex set (e.g. Euclidean ball of radius B). In this case, the performance of the method is measured with respect to the best solution w^* within this set*. Convince yourself that the analysis does not change, except* (8) *is replaced with an inequality.*

Instead of fixing T and giving accuracy after T steps, we may fix target accuracy ϵ and ask for the number of steps required. Lemma 1 then says

$$\frac{B^2 L^2}{\epsilon^2}. \tag{14}$$

You may ask whether this gives us the Perceptron bound when the problem is realizable, especially since (check!) $L = \max \|x_i\|$ in that case. Unfortunately, we need a small but clever modification of the lemma to make this conclusion, and we will do so later.

0.3 Stochastic Gradient Descent (SGD)

Suppose in the gradient descent step, we only have access to an unbiased estimate ∇_t of the gradient. That is, $\mathbb{E}[\nabla_t | w_t] = \nabla f(w_t)$. Under mild conditions (almost sure boundedness of $\|\nabla_t\|$, or boundedness of $\mathbb{E}\|\nabla_t\|^2$), the GD proof goes through for SGD. Let's convince ourselves of this. Observe that (11) and the balancing of η hold true with $\nabla f(w_t)$ replaced

by ∇_t, conditionally on the random draws of the estimates $\nabla_1, \ldots, \nabla_T$. Linearization (7) also goes through since conditionally on $\nabla_{1:t-1}$,

$$f(w_t) - f(w^*) \leq \langle \mathbb{E}_t[\nabla_t], w_t - w^* \rangle .$$

Here $\mathbb{E}_t$ denotes the conditional expectation. The result follows by the linearity of the expectation and the tower property. So, SGD for convex Lipschitz functions guarantees that

$$\mathbb{E}[f(\widehat{w})] - f(w^*) \leq \frac{BG}{\sqrt{T}},$$

where $G^2 \geq \max_i \mathbb{E}\|\nabla_i\|^2$ and $B \geq \|w^*\|^2$.

On a lighter note, here is a tweet from ML_hipster (aka Mark Reid):

References