Lecture 13: Convex Optimization

Akshay Krishnamurthy
akshay@cs.umass.edu

October 19, 2017

1 Recap

With SVM and more generally with surrogate losses, we have been talking about doing ERM with a convex loss function. The main point of using a convex loss function is computational, but how do we actually optimize such things? Specifically, given a loss function $\ell(y', y)$ that is convex in its first argument, we often face the ERM problem

$$\text{minimize}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

We now understand many of the statistical properties of the minimizer. For example, using Rademacher complexity, we can get bounds on the estimation error. Then, for binary classification, by understanding the calibration properties we can understand the relationship to the 0/1 loss that we might actually be interested in.

Today we will look at some computational aspects of this problem. Specifically we will discuss the most basic algorithm to compute the minimizer.

2 Convex Analysis

We've been using convexity at various points throughout the course, but here are some definitions that will be useful especially today.

Definition 1 (Convex Set)**.** *A set $C \subset \mathbb{R}^d$ is* convex *if $x, y \in C \Rightarrow tx + (1-t)y \in C$ for all $0 \leq t \leq 1$.*

Definition 2 (Convex Function)**.** *A function $F : \mathbb{R}^d \to \mathbb{R}$ is* convex *if $dom(F) \subset \mathbb{R}^d$ is convex and $f(tx+(1-t)y) \leq tf(x) + (1-t)f(y)$ for all $0 \leq t \leq 1$ and all $x, y \in dom(f)$.*

A general optimization problem takes the form

$$\text{minimize}_{x \in D} f(x) \text{ s.t. } g_i(x) \leq 0, i = 1, \ldots, m, h_j(x) = 0, j = 1, \ldots, r.$$

The problem is called convex if (1) D is a convex set, (2) f, g_i are all convex functions, (3) h_j are all affine functions $h_j(x) = \langle a_j, x \rangle + b_j$. We mostly focus on convex optimization problems.

Theorem 3 (Local optima implies global optima)**.** *Let P be a convex optimization problem and let $x^\star \in D$ be a point such that $f(x^\star) \leq f(y)$ for all feasible y with $\|x^\star - y\| \leq \rho$. Then $f(x^\star) \leq f(y)$ for all feasible y.*

Proof. The proof is by contradiction. Assume that there is some feasible z such that $f(z) < f(x^\star)$. Then take $y = tx^\star + (1-t)z$ for $t \in (0, 1)$ close to 1. We claim this point is feasible. The affine constraints are satisfied due to linearity, since both $x^\star$ and z are feasible. As for the inequality constraints, by convexity we get

$$g_i(tx^\star + (1-t)z) \leq tg_i(x^\star) + (1-t)g_i(z) \leq 0$$

Hence y is feasible. However, the objective value is strictly smaller than $f(x^\star)$, since

$$f(tx^\star + (1-t)z) \leq tf(x^\star) + (1-t)f(z) < f(x^\star).$$

For t close to one, we will get $\|x^\star - y\| \leq \rho$, which is a contradiction. □

2.1 Characterizations of convex functions.

There are two useful ways to characterize convex functions, both of which follow from the definitions.

Theorem 4. *If f is differentiable, then f is convex if and only if $dom(f)$ is convex and*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

Proof. The proof here requires re-arranging the zero-th order definition

$$f(ty + (1-t)x) \leq tf(y) + (1-t)f(x) \Rightarrow f(x + t(y-x)) - f(x) \leq t(f(y) - f(x))$$
$$\Rightarrow \frac{f(x + t(y-x)) - f(x)}{t} \leq f(y) - f(x).$$

Now take limit as $t \to 0$ and you'll get a derivative, but you have to apply chain rule so you'll get $\nabla f(x)^T(y-x)$.

For the other direction, take $z = tx + (1-t)y$ and we get two inequalities

$$f(x) \geq f(z) + \nabla f(z)^T (x - z) \qquad f(y) \geq f(z) + \nabla f(z)^T (y - z)$$

Adding t times the first inequality to $(1-t)$ times the second gives us the desired inequality

$$tf(x) + (1-t)f(y) \geq f(z) + \nabla f(z)^T (tx + (1-t)y - z) = f(tx + (1-t)y)$$ □

As a consequence, it is now easy to see that if f is convex and differentiable, then $\nabla f(x) = 0$ implies that f is a global optimum.

This property also inspires the definition of *strong convexity*.

Definition 5. *A function f is λ-strongly convex if its domain is convex, and for all x, y*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\lambda}{2} \|y - x\|_2^2$$

It is much easier to optimize strongly convex functions. The canonical example is $f(x) = \frac{1}{2}(x-a)^2$ which is 1 strongly convex.

Theorem 6. *If f is twice differentiable, then f is convex if and only if $dom(f)$ is convex and $\nabla^2 f(x) \succeq 0$ for all $x \in dom(f)$.*

Exercise: Prove on your own.

2.2 Optimality Conditions

It is often useful to understand what properties the global optima of an optimization problem satisfy. We saw that for f convex $\nabla f(x) = 0$ implies that f is globally optimal. There is almost a converse to this statement.

Proposition 7. *If f is differentiable, then a feasible point x is optimal if and only if $\nabla f(x)^T (y - x) \geq 0$ for all feasible y.*

The proof here follows from the first order definition of convexity and the fact that the feasible set is convex. The point is that the gradient need not be zero, but the function must increase when you move from x to y.

There are many other forms of optimality conditions, and in general much more to say about convex sets, functions, and optimization. Much of it is broadly useful for machine learning, but we do not have time to cover everything. Instead let us turn to some algorithms.

3 Gradient Descent

The most basic algorithm for convex optimization is gradient descent. Let us consider the unconstrained problem

$$\text{minimize}_{x\in\mathbb{R}^d} f(x)$$

The algorithm is iterative, starting at some initial point x_0 we repeatedly update

$$x_t \leftarrow x_{t-1} - \eta_t \nabla f(x_{t-1}).$$

For some number of iterations T. Here η_t is some learning rate parameter that we will set in the proof.

Technically, the way to think about the algorithms is as repeatedly minimizing something that looks like the second order taylor approximation to f at x_{t-1}.

$$f(y) \approx f(x) + \nabla f(x)^T(y-x) + (y-x)^T \frac{\nabla^2 f(x)}{2}(y-x) \approx f(x) + \nabla f(x)^T(y-x) + \frac{1}{2t}\|y-x\|_2^2$$

Gradient descent chooses x_t to minimize this approximation evaluated at x_{t-1}.

$$x_t = \operatorname*{argmin}_y f(x_{t-1}) + \nabla f(x_{t-1})^T(y-x_{t-1}) + \frac{1}{2\eta_t}\|y-x_{t-1}\|_2^2.$$

We require one more definition.

Definition 8. *We say that a function f is μ-smooth if for all $x, y \in dom(f)$,*

$$f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2$$

Alternatively

$$\|\nabla f(x) - \nabla f(y)\| \le \mu\|x-y\|_2$$

Theorem 9. *Let f be a differentiable convex function that is μ smooth. Then gradient descent with step size $\eta_t = \eta \le 1/\mu$ satisfies*

$$f(x^{(T)}) - \min_x f(x) \le \frac{\|x^{(0)} - x^\star\|_2^2}{2\eta T}$$

Proof. Consider an iteration t and let $x^+ = x^{(t+1)}$ and $x = x^{(t)}$. Using the smoothness property, we have

$$f(x^+) \le f(x) + \nabla f(x)^T(x^+ - x) + \frac{\mu}{2}\|x^+ - x\|_2^2 = f(x) - \eta_t\|\nabla f(x)\|_2^2 + \frac{\mu\eta_t^2}{2}\|\nabla f(x)\|_2^2 \le f(x) - \frac{\eta_t}{2}\|\nabla f(x)\|_2^2$$

Here we used that $\eta_t \le 1/\mu$. This shows that if the gradient is big then we make substantial progress. We now have to relate the gradient to the distance to the optimum. By the first-order characterization of convex functions,

$$\begin{aligned}
f(x^+) &\le f(x) - \frac{\eta_t}{2}\|\nabla f(x)\|_2^2 \\
&\le f(x^\star) + \nabla f(x)^T(x - x^\star) - \frac{\eta_t}{2}\|\nabla f(x)\|_2^2 \\
&= f(x^\star) + \frac{1}{2\eta_t}\left(\|x-x^\star\|_2^2 - \|x-x^\star\|_2^2 + 2\eta_t\nabla f(x)^T(x-x^\star) - \eta_t^2\nabla f(x)^T\nabla f(x)\right) \\
&= f(x^\star) + \frac{1}{2\eta_t}\left(\|x-x^\star\|_2^2 - \|x^+-x^\star\|_2^2\right)
\end{aligned}$$

Thus since η_t is constant, we get a telescoping sum

$$\sum_{t=1}^T f(x^{(t)}) - f(x^\star) \le \frac{1}{2\eta}\sum_{t=1}^T\left(\|x^{(t-1)}-x^\star\|_2^2 - \|x^{(t)}-x^\star\|_2^2\right) \le \frac{1}{2\eta}\|x^{(0)}-x^\star\|_2^2$$

Finally, since we proved that we always make progress we know that $f(x^{(T)})$ is the smallest, and the min is smaller than average, which proves the result. □

3.1 Gradient Descent Variants

There are many gradient descent variants and analyses, that you might want to become familiar with

1. *Subgradient descent* can be used when f is not differentiable. Instead of the actual gradient (which is not defined), you just have to use any vector that provides a lower bound on the function at this point, e.g., using the first order definition of convexity. These are called subgradients.

2. *Projected gradient descent* can be used for constrained optimization problems, the idea is after doing the gradient step you project back onto the feasible set.

3. *Proximal gradient descent* is related to the projected case, but it works well for non-smooth regularizers or optimizations of the form $f(x) + h(x)$. The idea is to do gradient descent on the smooth part f and think of the regularizer h as a soft constraint, and "project" back onto the regularizer using

$$\text{prox}_h(x) = \operatorname*{argmin}_u h(y) + \frac{1}{2}\|u - x\|_2^2$$

   Indeed if $h(u)$ is the $0/\infty$ indicator of a constraint set, then this is just the projection operator.

4. *Gradient descent with strong convexity.* With strong convexity a much better convergence rate is possible. You will analyze this case on the homework.

5. There are also notions of *acceleration* and *momentum* that can speed up the convergence of gradient methods. The idea is to maintain the last two iterates, $x^{(t-1)}$ and $x^{(t-2)}$

$$v \leftarrow x^{(t-1)} + \frac{t-2}{t+1}(x^{(t-1)} - x^{(t-2)}), x^{(t)} \leftarrow v - \eta_t \nabla f(x^{(t-1)})$$

   This method has an $O(1/T^2)$ convergence rate, which is optimal for first-order methods.

I should also mention that there are many "second-order" methods that use Hessian information. One issue with them is that for high-dimensional problems the Hessian is quite big, so while the iteration complexity can be much better than first-order methods, the computational overhead mitigate this.

4 Stochastic Gradient Methods

Stochastic gradient descent and its variants are probably the most popular optimization algorithms in machine learning. They work best in the ERM setting, where we are trying to optimize a function with "finite-sum" structure:

$$\operatorname*{argmin}_x F(x) = \operatorname*{argmin}_x \frac{1}{T}\sum_{t=1}^{T} f_t(x)$$

Let us just think of this as just one fuction F but we get noisy gradients v_t that satisfy $\mathbb{E}[v_t|x^{(t)}] = \nabla f(x^{(t)})$. This can be achieved by sampling $t \in [T]$ and computing the gradient on f_t. The algorithm starts with $x^{(1)} = 0$, repeat for T rounds

1. Choose v_t such that $\mathbb{E}[v_t|x^{(t)}] \in \partial F(x^{(t)})$
2. Update $x^{(t+1)} \leftarrow x^{(t)} - \eta v_t$

At the end we output $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x^{(t)}$.

Theorem 10. *Let F be a convex function and let $x^\star = \operatorname{argmin}_{x:\|x\|\le B} F(x)$ and assume that $\|v_t\| \le \mu$ with probability 1. Then if SGD is run for T iterations with $\eta = \sqrt{\frac{B^2}{\mu^2 T}}$,*

$$\mathbb{E}[F(\bar{x})] - F(x^\star) \le \frac{B\mu}{\sqrt{T}}$$

Proof. By convexity

$$\mathbb{E}_{v_1^T} F(\bar{x}) - F(x^\star) \leq \mathbb{E}_{v_1^T}\left[\frac{1}{T}\sum_{t=1}^{T} F(x^{(t)}) - F(x^\star)\right]$$

By linearity of expectation

$$\begin{aligned}
\mathbb{E}_{v_1^T}\left[\frac{1}{T}\sum_{t=1}^{T} F(x^{(t)}) - F(x^\star)\right] &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{v_1^T} F(x^{(t)}) - F(x^\star) \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{v_1^t} F(x^{(t)}) - F(x^\star) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{v_1^{t-1}} \langle \mathbb{E}_{v_t}[v_t | v_1^{t-1}], x^t - x^\star \rangle \\
&= \mathbb{E}_{v_1^T}\left[\frac{1}{T}\sum_{t=1}^{T} \langle v_t, x^t - x^\star \rangle\right]
\end{aligned}$$

Next, for any sequence $v_1, \ldots, v_T$, using a similar argument to the gradient descent proof

$$\begin{aligned}
\langle x^{(t)} - x^\star, v_t \rangle &= \frac{1}{2\eta}\langle x^{(t)} - x^\star, 2\eta v_t \rangle \\
&= \frac{1}{2\eta}\left(\|x^{(t)} - x^\star\|_2^2 - \|x^{(t+1)} - x^\star\|_2^2 + \eta^2 \|v_t\|_2^2\right) \\
&= \frac{1}{2\eta}\left(\|x^{(t)} - x^\star\|_2^2 - \|x^{(t+1)} - x^\star\|_2^2\right) + \frac{\eta}{2}\|v_t\|_2^2
\end{aligned}$$

and the first term telescopes while for the second term we use $\|v_t\|^2 \leq \mu^2$. This gives,

$$\mathbb{E}_{v_1^T} f(\bar{x}) - f(x^\star) \leq \frac{1}{2\eta}\|w^\star\| + \frac{\eta}{2} T\mu^2,$$

which is optimized by our choice of η. □