University of California, Berkeley Lecturer: Benjamin Recht

CS227C/STAT260 Convex Optimization and Approximation: Optimization for Modern Data Analysis

February 23, 2016 Notes: Benjamin Recht

Lecture 9: Constrained Optimization and the Projected gradient methods

In constrained optimization, we aim to find a point x which achieves the smallest value of some function f subject to the requirement that x lives in some specified set Ω.

Constrained optimization lets us design considerably more rich and complex optimization problems. The constraints could simply be bounds on the values of the variables, but could model temporal dependencies, resource constraints, or statistical models. In this first introductory lecture, we will focus on case when Ω is a simple convex set. We will move to more complicated scenarios in the coming weeks.

1 The Minimum Principle

What does it mean for $x_\star$ to minimize f over a set Ω? We say $x_\star$ minimizes f over Ω if $x_\star \in \Omega$ and

$$f(x_\star) \le f(z) \quad \forall\ z \in \Omega\,.$$

That is, $x_\star$ is an element of Ω and attains the lowest function value. We say that $x_\star$ is a *local minimizer* if $x_\star \in \Omega$ and there is a convex neighborhood $\mathcal{N}$ of $x_\star$ such that

$$f(x_\star) \le f(z) \quad \forall\ z \in \mathcal{N} \cap \Omega\,.$$

These definitions seem reasonable enough, but now how do we check if we have a minimizer or local minimizer? When $\Omega = \mathbb{R}^d$, we could simply check if $\nabla f(x) = -$. But with constraints, this is not the case. For example, suppose we want to minimize x^2 subject to $x \in [1, 2]$. Then clearly the minimum is $x = 1$, but the gradient at 1 is 2. The following proposition provides a solution.

Proposition 1 (The Minimum Principle) *If f is smooth and $x_\star$ locally minimizes f on a closed convex set Ω, then $\langle \nabla f(x_\star), z - x_\star \rangle \ge 0$ for all $z \in \Omega$. If f is convex, then the converse holds.*

Proof $\langle \nabla f(x_\star), z - x_\star \rangle$ is the directional derivative of f in the direction $z - x_\star$. If this directional derivative is negative, then we can move from $x_\star$ to z along the line segment connecting them and decrease the function with a sufficiently small step. That is, there exists a $t > 0$ such that $f((1-t)x_\star + tz) < f(x_\star)$, But since Ω (and also $\mathcal{N} \cap \Omega$) is convex, $(1-t)x_\star + tz \in \Omega$ contradicting the optimality of $x_\star$.

For the converse, assume f is convex and $\langle \nabla f(x_\star), z - x_\star \rangle \ge 0$ for all $z \in \Omega$. Then for $z \in \Omega$

$$f(z) \ge f(x_\star) + \langle \nabla f(x_\star), z - x_\star \rangle \ge f(x_\star)$$

proving $x_\star$ is optimal. ■

Note that this proposition proves that 1 minimizes x^2 over $[1, 2]$ because we have that $z - 1 > 0$ for all $z \in [1, 2]$.

In the case that f is strongly convex, we can use the Minimum Principle to prove that f will have a unique minimizer over Ω.

Proposition 2 *Suppose f is strongly convex. Then f has a unique minimizer over the closed convex set Ω.*

Proof Let x_1 and x_2 minimize f over Ω. Suppose $x_1 \neq x_2$, then we have

$$f(x_2) \geq f(x_1) + \langle \nabla f(x_1), x_2 - x_\star \rangle + \tfrac{m}{2}\|x_1 - x_2\|^2 > f(x_1)$$

which contradicts the optimality of x_2. Thus, x_1 must equal x_2. ■

While the Minimum Principle looks like a difficult condition to check, we will discuss an algorithm below that will find an $x_\star$ approximately satisfying this condition.

2 Euclidean Projection

Let Ω be a closed, convex set. The *Euclidean projection* of a point x onto Ω is the closest point in Ω to x. Denote this point by $\Pi_\Omega(x)$. Note that $\Pi_\Omega(x)$ is the solution of a constrained optimization problem:

$$\Pi_\Omega(x) = \arg\min\{\|z - x\| \ : \ z \in \Omega\}$$

That is, $\Pi_\Omega(x)$ is the solution to the optimization problem

$$\begin{array}{ll} \text{minimize}_z & \frac{1}{2}\|z - x\|^2 \\ \text{subject to} & z \in \Omega \end{array} .$$

Since the cost function of this problem is strongly convex, this proves that $\Pi_\Omega(x)$ is unique for all x.

Using the minimum principle, we can compute a variety of projections onto simple sets.

Example 1: The nonnegative orthant The nonnegative orthant is the set of vectors which are nonnegative in all coordinates.

$$\Omega = \{x \ : \ x_i \geq 0 \ \forall \ i = 1, \ldots, d\}$$

Note that Ω is a closed, convex cone.

Unpacking the condition $\langle \Pi_\Omega(x) - x, z - \Pi_\Omega(x) \rangle \geq 0$, we must have that

$$[\Pi_\Omega(x) - x]_i \geq 0$$

for all coordinates. Note that simply setting

$$[\Pi_\Omega(x)]_i = \begin{cases} x_i & x_i \geq 0 \\ 0 & x_i < 0 \end{cases}$$

satisfies the Minimum Principle.

Example 2: Unit norm ball Let

$$\Omega = \{x \ : \ \|x\| \leq 1\}$$

To compute $\Pi_\Omega(x)$, note that we require $\langle \Pi_\Omega(x) - x, z - \Pi_\Omega(x) \rangle \geq 0.$ for all $z \in \Omega$. One can readily check that

$$\Pi_\Omega(x) = \frac{x}{\|x\|}$$

satisfies the Minimum Principle.

One of the most useful properties of the Euclidean projection is the fact that projections are *nonexpansive* in the following sense:

Proposition 3 *Let Ω be a closed convex set. Then*

$$\|\Pi_\Omega(x) - \Pi_\Omega(y)\| \le \|x - y\|$$

for all $x, y \in \mathbb{R}^d$.

Proof By the minimum principle, $\Pi_\Omega(x)$ satisfies

$$\langle \Pi_\Omega(x) - x, z - \Pi_\Omega(x) \rangle \ge 0$$

for all $z \in \Omega$. Now we can write out

$$\begin{aligned}
\|x - y\|^2 &= \|(x - \Pi_\Omega(x)) - (y - \Pi_\Omega(y)) + \Pi_\Omega(x) - \Pi_\Omega(y)\|^2 \\
&= \|(x - \Pi_\Omega(x)) - (y - \Pi_\Omega(y))\|^2 + \|\Pi_\Omega(x) - \Pi_\Omega(y)\|^2 \\
&\qquad + 2\langle \Pi_\Omega(x) - x, \Pi_\Omega(y) - \Pi_\Omega(x) \rangle + 2\langle \Pi_\Omega(y) - y, \Pi_\Omega(x) - \Pi_\Omega(y) \rangle \\
&\ge \|(x - \Pi_\Omega(x)) - (y - \Pi_\Omega(y))\|^2 + \|\Pi_\Omega(x) - \Pi_\Omega(y)\|^2
\end{aligned}$$

completing the proof. ■

3 The projected gradient algorithm

The projected gradient algorithm combines a projection step with a gradient step. This lets us solve a variety of constrained optimization problems with simple constraints.

We will aim to solve the constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \Omega \end{array} \tag{1}$$

where f is smooth and Ω is convex. Let us assume as usual that ∇f is Lipschitz so that $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$.

Let us define a projected gradient scheme to solve this problem. Let $\alpha_0, \ldots, \alpha_T, \ldots,$ be a sequence of positive step sizes. Choose $x_0 \in X$, and iterate

$$x_{k+1} = \Pi_\Omega(x_k - \alpha_k \nabla f(x_k))\,. \tag{2}$$

The algorithm simply alternates between taking gradient steps and then taking projection steps.

The key idea behind this algorithm is summed up by the following proposition

Proposition 4 *Let f be differentiable and convex and let Ω be convex. $x_\star$ is an optimal solution of (1) if and only if $x_\star = \Pi_\Omega(x_\star - \alpha \nabla f(x_\star))$ for all $\alpha > 0$.*

Proof $x_\star$ is an optimal solution if and only if $\langle \nabla f(x_\star), x - x_\star \rangle \ge 0$ for all $x \in \Omega$. This is equivalent to

$$\langle x_\star - (x_\star - \alpha \nabla f(x_\star)), x - x_\star \rangle \ge 0\,,$$

which, by the Minimum Principle is equivalent to $x_\star$ being the optimal solution of

$$\begin{array}{ll} \text{minimize}_z & \frac{1}{2}\|z - x_\star\|^2 \\ \text{subject to} & z \in \Omega \end{array}$$

in other words, $x_\star = \Pi_\Omega(x_\star - \alpha \nabla f(x_\star))$. ■

For non-convex f, we see that a fixed point of the projected gradient iteration is a stationary point of h. We first analyze the convergence of this projected gradient method for arbitrary smooth f, and then focus on strongly convex f.

3.1 General Case

Let f_* denote the optimal value of (1). Suppose we set $\alpha_k = 1/M$ for all k with $M \geq L$. Then we have

$$\|x_{k+1} - x_k\| \leq \sqrt{\frac{2(f(x_0) - f_\star)}{M(k+1)}}\,. \tag{3}$$

This expression confirms that we will find a point x where

$$\|\Pi_\Omega(x - \alpha \nabla f(x)) - x\| \leq \epsilon\,.$$

To verify this inequality, note that for any x, y,

$$f(x) = f(x) \leq f(y) + \nabla f(y)^T(x-y) + \frac{M}{2}\|x-y\|^2 =: \ell(x;y)$$

for any $M \geq L$. This is just Taylor's series. Note that the minimizer of $\ell(x;y)$ (with respect to x) over Ω is equal to

$$\Pi_\Omega(y - 1/M \nabla f(y))\,.$$

and also note that $\ell(x;y)$ is strongly convex with parameter M.

Now we have the chain of inequalities

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq f(x_k) - \ell(x_{k+1};x_k) \\ &= \ell(x_k;x_k) - \ell(x_{k+1};x_k) \\ &\geq \frac{M}{2}\|x_{k+1} - x_k\|^2 \end{aligned}$$

Summing these inequalities up for $k = 1, \ldots, n$, we have

$$\sum_{k=0}^{n} \|x_{k+1} - x_k\|^2 \leq \frac{2}{M}(f(x_0) - f_\star)$$

and the conclusion follows.

3.2 Strongly Convex Case

Let's now assume that f is strongly convex with strong convexity parameter m:

$$f(z) \geq f(x) + \nabla f(x)^T (z-x) + \frac{m}{2}\|z-x\|^2 \,. \tag{4}$$

Let $x_\star$ denote the optimal solution of (1). $x_\star$ is unique because of strong convexity. Observe that

$$\begin{aligned} \|x_{k+1} - x_\star\| &= \|\Pi_\Omega(x_k - \alpha_k \nabla f(x_k) - \Pi_\Omega(x_\star - \alpha_k \nabla f(x_\star))\| && (5) \\ &\leq \|x_k - \alpha_k \nabla f(x_k) - x_\star + \alpha_k \nabla f(x_\star)\| && (6) \end{aligned}$$

Here, the first equality follows by the definition of x_{k+1} and because $x_\star$ is optimal (see Proposition 4). (6) follows from Proposition 3.

Since f is strongly convex and has a Lipschitz continuous gradient, it follows that for all vectors x and y and all positive scalars η

$$\|x - \eta \nabla f(x) - (y - \eta \nabla f(y))\| \leq \max\{|1-\eta L|, |1-\eta m|\}\|x-y\| \,. \tag{7}$$

To see this, note that there exists a $\hat{t} \in [0,1]$ such that

$$\|x - \eta \nabla f(x) - (y - \eta \nabla f(y))\| = \left\|\left(I - \eta \nabla^2 f(x + \hat{t}(y-x))\right)(y-x)dt\right\| \,. \tag{8}$$

From this, it follows that

$$\|x - \eta \nabla f(x) - (y - \eta \nabla f(y))\| \leq \sup_z \|I - \eta \nabla^2 f(z)\|\|y - z\| \,. \tag{9}$$

Note that the minimum eigenvalue of $\nabla^2 f(z)$ is at least m and the maximum eigenvalue is at least L. Therefore the eigenvalues of $I - \eta\nabla^2 f(z)$ are at most $\max(1-\eta L, 1-\eta m)$ and at least $\min(1-\eta L, 1-\eta m)$. Therefore, $\|I - \eta\nabla^2 f(z)\| \leq \max(|1-\eta L|, |1-\eta m|)$.

In particular, using this upper bound in (6), we have

$$\|x_{k+1} - x_\star\| \leq \max\{|1-\alpha_k L|, |1-\alpha_k m|\}\|x-y\| \,. \tag{10}$$

Note that $\alpha_k = \frac{2}{L+m}$ minimizes the right hand side for all k. Setting α_k to this value, we find that

$$\|x_{k+1} - x_\star\| \leq \left(\frac{L-m}{L+m}\right)\|x_k - x_\star\| \tag{11}$$

or, denoting $\kappa = \frac{L}{m}$ and $D_0 = \|x_0 - x_\star\|$,

$$\|x_k - x_\star\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k D_0 \tag{12}$$

That is, for strongly convex f and arbitrary Ω, the projected gradient algorithm converges at a linear rate under a constant step-size policy.

3.3 Nesterov Iteration, no proof

We can even define an *accelerated* version of the proximal gradient method. Iterations take the form:

$$\begin{aligned} \xi_{k+1} &= \Pi_\Omega \left(y_k - \alpha \nabla f(y_k) \right) \\ y_k &= \xi_k + \beta(\xi_k - \xi_{k-1}) \end{aligned} \tag{13}$$

Note that when $\Pi_\Omega = I$, we recover the standard Nesterov algorithm. When $\beta = 0$, we recover the proximal gradient method. This method will converge in

$$O\left(\sqrt{\frac{L}{m}} \log(1/\epsilon) \right)$$

iterations for strongly convex functions. We will prove this when we return to Nesterov's method after the midterm.