18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 11

Scribe: Kevin Li — Oct. 14, 2015

# 2. CONVEX OPTIMIZATION FOR MACHINE LEARNING

In this lecture, we will cover the basics of convex optimization as it applies to machine learning. There is much more to this topic than will be covered in this class so you may be interested in the following books.

*Convex Optimization* by Boyd and Vandenberghe
*Lecture notes on Convex Optimization* by Nesterov
*Convex Optimization: Algorithms and Complexity* by Bubeck
*Online Convex Optimization* by Hazan

The last two are drafts and can be obtained online.

## 2.1 Convex Problems

A convex problem is an optimization problem of the form $\min_{x \in \mathcal{C}} f(x)$ where $f$ and $\mathcal{C}$ are convex. First, we will debunk the idea that convex problems are easy by showing that virtually all optimization problems can be written as a convex problem. We can rewrite an optimization problem as follows.

$$\min_{\mathcal{X} \in \mathcal{X}} f(x) \quad \Leftrightarrow \quad \min_{t \geq f(x), x \in \mathcal{X}} t \quad \Leftrightarrow \quad \min_{(x,t) \in \text{epi}(f)} t$$

where the epigraph of a function is defined by

$$\text{epi}(f) = \{(x, t) \in \mathcal{X} \times \mathbb{R} : t \geq f(x)\}$$

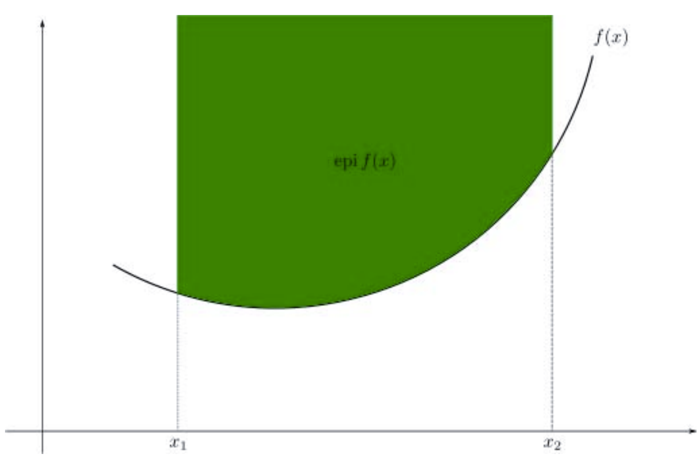


Figure 1: An example of an epigraph.
Source: https://en.wikipedia.org/wiki/Epigraph_(mathematics)

Now we observe that for linear functions,

$$\min_{x \in D} c^\top x = \min_{x \in \mathsf{conv}(D)} c^\top x$$

where the convex hull is defined

$$\mathsf{conv}(D) = \{y : \exists N \in \mathbb{Z}_+, x_1, \ldots, x_N \in D, \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1, y = \sum_{i=1}^N \alpha_i x_i\}$$

To prove this, we know that the left side is a least as big as the right side since $D \subset \mathsf{conv}(D)$. For the other direction, we have

$$\begin{aligned}
\min_{x \in \mathsf{conv}(D)} c^\top x &= \min_N \min_{x_1,\ldots,x_N \in D} \min_{\alpha_1,\ldots,\alpha_N} c^\top \sum_{i=1}^N \alpha_i x_i \\
&= \min_N \min_{x_1,\ldots,x_N \in D} \min_{\alpha_1,\ldots,\alpha_N} \sum_{i=1}^N \alpha_i c^\top x_i \geq \min_{x \in D} c^\top x \\
&\geq \min_N \min_{x_1,\ldots,x_N \in D} \min_{\alpha_1,\ldots,\alpha_N} \sum_{i=1}^N \alpha_i \min_{x \in D} c^\top x \\
&= \min_{x \in D} c^\top x
\end{aligned}$$

Therefore we have

$$\min_{x \in \mathcal{X}} f(x) \Leftrightarrow \min_{(x,t) \in \mathsf{conv}(\mathrm{epi}(f))} t$$

which is a convex problem.

Why do we want convexity? As we will show, convexity allows us to infer global information from local information. First, we must define the notion of *subgradient*.

**Definition (Subgradient):** Let $\mathcal{C} \subset \mathbb{R}^d$, $f : \mathcal{C} \to \mathbb{R}$. A vector $g \in \mathbb{R}^d$ is called a *subgradient* of $f$ at $x \in \mathcal{C}$ if

$$f(x) - f(y) \leq g^\top (x - y) \qquad \forall y \in \mathcal{C}\,.$$

The set of such vectors $g$ is denoted by $\partial f(x)$.

Subgradients essentially correspond to gradients but unlike gradients, they always exist for convex functions, even when they are not differentiable as illustrated by the next theorem.

**Theorem:** If $f : \mathcal{C} \to \mathbb{R}$ is convex, then for all $x$, $\partial f(x) \neq \emptyset$. In addition, if $f$ is differentiable at $x$, then $\partial f(x) = \{\nabla f(x)\}$.

*Proof.* Omitted. Requires separating hyperplanes for convex sets. $\square$

**Theorem:** Let $f, \mathcal{C}$ be convex. If $x$ is a local minimum of $f$ on $\mathcal{C}$, then it is also global minimum. Furthermore this happens if and only if $0 \in \partial f(x)$.

*Proof.* $0 \in \partial f(x)$ if and only if $f(x) - f(y) \leq 0$ for all $y \in \mathcal{C}$. This is clearly equivalent to $x$ being a global minimizer.

Next assume $x$ is a local minimum. Then for all $y \in \mathcal{C}$ there exists $\varepsilon$ small enough such that $f(x) \leq f\Big((1-\varepsilon)x + \varepsilon y\Big) \leq (1-\varepsilon)f(x) + \varepsilon f(y) \implies f(x) \leq f(y)$ for all $y \in \mathcal{C}$. $\square$

Not only do we know that local minimums are global minimums, looking at the subgradient also tells us where the minimum can be. If $g^\top(x-y) < 0$ then $f(x) < f(y)$. This means $f(y)$ cannot possibly be a minimum so we can narrow our search to $y$s such that $g^\top(x-y)$. In one dimension, this corresponds to the half line $\{y \in \mathbb{R} : y \leq x\}$ if $g > 0$ and the half line $\{y \in \mathbb{R} : y \geq x\}$ if $g < 0$ . This concept leads to the idea of gradient descent.

### 2.2 Gradient Descent

$y \approx x$ and $f$ differentiable the first order Taylor expansion of $f$ at $x$ yields $f(y) \approx f(x) + g^\top(y-x)$. This means that

$$\min_{|\hat{\mu}|_2 = 1} f(x + \varepsilon\hat{\mu}) \approx \min f(x) + g^\top(\varepsilon\hat{\mu})$$

which is minimized at $\hat{\mu} = -\frac{g}{|g|_2}$. Therefore to minimizes the linear approximation of $f$ at $x$, one should move in direction opposite to the gradient.

Gradient descent is an algorithm that produces a sequence of points $\{x_j\}_{j\geq 1}$ such that (hopefully) $f(x_{j+1}) < f(x_j)$.

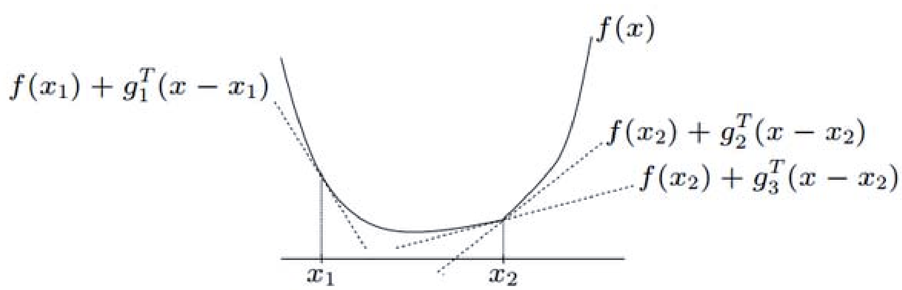


Figure 2: Example where the subgradient of $x_1$ is a singleton and and the subgradient of $x_2$ contains multiple elements.
Source: https://optimization.mccormick.northwestern.edu/index.php/Subgradient_optimization

**Algorithm 1** Gradient Descent algorithm

**Input:** $x_1 \in \mathcal{C}$, positive sequence $\{\eta_s\}_{s\geq 1}$

**for** $s = 1$ to $k-1$ **do**

$x_{s+1} = x_s - \eta_s g_s\,, \quad g_s \in \partial f(x_s)$

**end for**

**return** Either $\bar{x} = \frac{1}{k}\sum_{s=1}^{k} x_s$ or $x^\circ \in \underset{x\in\{x_1,\dots,x_k\}}{\operatorname{argmin}} f(x)$

**Theorem:** Let $f$ be a convex $L$-Lipschitz function on $\mathbb{R}^d$ such that $x^* \in \operatorname{argmin}_{\mathbb{R}^d} f(x)$ exists. Assume that $|x_1 - x^*|_2 \leq R$. Then if $\eta_s = \eta = \frac{R}{L\sqrt{k}}$ for all $s \geq 1$, then

$$f(\frac{1}{k}\sum_{s=1}^{k} x_s) - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

and

$$\min_{1\leq s\leq k} f(x_s) - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

*Proof.* Using the fact that $g_s = \frac{1}{\eta}(x_{s+1} - x_s)$ and the equality $2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$,

$$\begin{aligned} f(x_s) - f(x^*) \leq g_s^\top(x_s - x^*) &= \frac{1}{\eta}(x_s - x_{s+1})^\top(x_s - x^*) \\ &= \frac{1}{2\eta}\Big[\|x_s - x_{s+1}\|^2 + \|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\Big] \\ &= \frac{\eta}{2}\|g_s\|^2 + \frac{1}{2\eta}(\delta_s^2 - \delta_{s+1}^2) \end{aligned}$$

where we have defined $\delta_s = \|x_s - x^*\|$. Using the Lipschitz condition

$$f(x_s) - f(x^*) \leq \frac{\eta}{2}L^2 + \frac{1}{2\eta}(\delta_s^2 - \delta_{s+1}^2)$$

Taking the average from 1, to $k$ we get

$$\frac{1}{k}\sum_{s=1}^{k} f(x_s) - f(x^*) \leq \frac{\eta}{2}L^2 + \frac{1}{2k\eta}(\delta_1^2 - \delta_{s+1}^2) \leq \frac{\eta}{2}L^2 + \frac{1}{2k\eta}\delta_1^2 \leq \frac{\eta}{2}L^2 + \frac{R^2}{2k\eta}$$

Taking $\eta = \frac{R}{L\sqrt{k}}$ to minimize the expression, we obtain

$$\frac{1}{k}\sum_{s=1}^{k} f(x_s) - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

Noticing that the left-hand side of the inequality is larger than both $f(\sum_{s=1}^{k} x_s) - f(x^*)$ by Jensen's inequality and $\min_{1\leq s\leq k} f(x_s) - f(x^*)$ respectively, completes the proof. □

One flaw with this theorem is that the step size depends on $k$. We would rather have step sizes $\eta_s$ that does not depend on $k$ so the inequalities hold for all $k$. With the new step sizes,

$$\sum_{s=1}^{k} \eta_s[f(x_s) - f(x^*)] \leq \sum_{s=1}^{k} \frac{\eta_s^2}{2} L^2 + \frac{1}{2}\sum_{s=1}^{k}(\delta_s^2 - \delta_{s+1}^2) \leq \left(\sum_{s=1}^{k} \eta_s^2\right)\frac{L}{2} + \frac{R^2}{2}$$

After dividing by $\sum_{s=1}^{k} \eta_s$, we would like the right-hand side to approach 0. For this to happen we need $\frac{\sum \eta_s^2}{\sum \eta_s} \to 0$ and $\sum \eta_s \to \infty$. One candidate for the step size is $\eta_s = \frac{G}{\sqrt{s}}$ since then $\sum\limits_{s=1}^{k} \eta_s^2 \leq c_1 G^2 \log(k)$ and $\sum\limits_{s=1}^{k} \eta_s \geq c_2 G\sqrt{k}$. So we get

$$\left(\sum_{s=1}^{k} \eta_s\right)^{-1} \sum_{s=1}^{k} \eta_s[f(x_s) - f(x^*)] \leq \frac{c_1 GL \log k}{2c_2\sqrt{k}} + \frac{R^2}{2c_2 G\sqrt{k}}$$

Choosing $G$ appropriately, the right-hand side approaches 0 at the rate of $LR\sqrt{\frac{\log k}{k}}$. Notice that we get an extra factor of $\sqrt{\log k}$. However, if we look at the sum from $k/2$ to $k$ instead of 1 to $k$, $\sum\limits_{s=\frac{k}{2}}^{k} \eta_s^2 \leq c_1' G^2$ and $\sum\limits_{s=1}^{k} \eta_s \geq c_2' G\sqrt{k}$. Now we have

$$\min_{1\leq s\leq k} f(x_s) - f(x^*) \leq \min_{\frac{k}{2}\leq s\leq k} f(x_s) - f(x^*) \leq \left(\sum_{s=\frac{k}{2}}^{k} \eta_s\right)^{-1} \sum_{s=\frac{k}{2}}^{k} \eta_s[f(x_s) - f(x^*)] \leq \frac{cLR}{\sqrt{k}}$$

which is the same rate as in the theorem and the step sizes are independent of $k$.

Important Remark: Note this rate only holds if we can ensure that $|x_{k/2} - x^*|_2 \leq R$ since we have replaced $x_1$ by $x_{k/2}$ in the telescoping sum. In general, this is not true for gradient descent, but it will be true for *projected gradient descent* in the next lecture.

One final remark is that the dimension $d$ does not appear anywhere in the proof. However, the dimension does have an effect because for larger dimensions, the conditions $f$ is $L$-Lipschitz and $|x_1 - x^*|_2 \leq R$ are stronger conditions in higher dimensions.