Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: hardt+ee227c@berkeley.edu

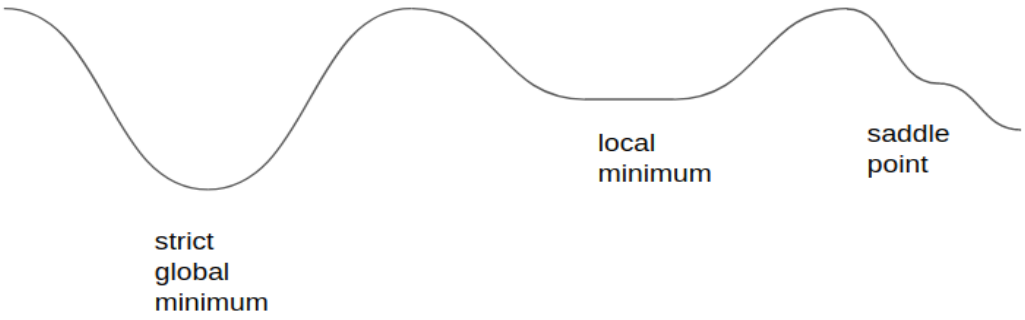
Graduate Instructor: Max Simchowitz

Email: msimchow+ee227c@berkeley.edu

October 15, 2018

17 Non-convex problems

This lecture provides the important information on how non-convex problems differ from convex problems. The major issue in non-convex problems is that it can be difficult to find the global minimum because algorithms can easily get stuck in the possibly numerous local minima and saddle points.



17.1 Local minima

Definition 17.1 (Local minimum)**.** A point x^* is an unconstrained *local minimum* if there exist $\epsilon > 0$ such that $f(x^*) \leqslant f(x)$ for all x with $\|x - x^*\| < \epsilon$.

Definition 17.2 (Global minimum)**.** A point x^* is an unconstrained *global minimum* if $f(x^*) \leqslant f(x)$ for all x.

For both definitions, we say "strict" if these inequalities are strict.

Proposition 17.3 (Necessary Conditions for local minimum)**.** *Let x^* be an unconstrained local minimum of $f\colon \mathbb{R}^n \to \mathbb{R}$ and assume f is continuously differentiable (C^1) in an open set containing x^*. Then*

1. $\nabla f(x^*) = 0$ *(First-Order Necessary Condition)*
2. *If in addition f is twice continuously differentiable in an open set around x^*, then $\nabla^2 f(x^*) \succeq 0$. (Second Order Necessary Condition)*

Proof of Proposition 17.3. Fix any direction $d \in \mathbb{R}^n$.

1. $g(\alpha) := f(x^* + \alpha d)$. Then

$$\begin{aligned} 0 &\leqslant \lim_{\alpha \to 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} \\ &= \frac{\partial g(0)}{\partial \alpha} \\ &= d^\top \nabla f(x^*) \end{aligned} \tag{1}$$

   Inequality 1 follows because x^* is a local minimum, $0 \leqslant f(x^* + \alpha d) - f(x^*)$ for sufficiently small alpha. So, we can construct a sequence with only positive α that converges to x^* such that each element $0 \leqslant \frac{f(x^* + \alpha_n d) - f(x^*)}{\alpha_n}$ which implies that statement given that f is locally differentiable.

   Since d is arbitrary, this implies that $\nabla f(x^*) = 0$.

2. First we represent $f(x^* + \alpha d) - f(x^*)$ using the 2nd order Taylor expansion.

$$\begin{aligned} f(x^* + \alpha d) - f(x^*) &= \alpha \nabla f(x^*)^\top d + \frac{\alpha^2}{2} d^\top \nabla^2 f(x^*) d + O(\alpha^2) \\ &= \frac{\alpha^2}{2} d^\top \nabla^2 f(x^*) d + O(\alpha^2) \end{aligned}$$

Now we do the following

$$\begin{aligned}
0 &\leqslant \lim_{\alpha \to 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha^2} \\
&= \lim_{\alpha \to 0} \frac{1}{2} d^\top \nabla^2 f(x^*) d + \frac{O(\alpha^2)}{\alpha^2} \\
&= \frac{1}{2} d^\top \nabla^2 f(x^*) d
\end{aligned}$$

Because d is arbitrary, this implies that $\nabla^2 f(x^*) \succeq 0$ (Positive semidefinite).

■

Note that $\nabla f(x^*) = 0$ alone does not imply x^* is a local minimum. Even the necessary conditions $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$ does not imply x^* is a local minimum. This is because it could be that $\nabla^2 f(x^*) = 0$, but the 3rd order is not 0. For example in the 1d case, $x^* = 0$ for $f(x) = x^3$ satisfies these conditions, but is not a local minimum. Now, we will look at the actual sufficient conditions for a local minimum, but these conditions can only detect strict local minima.

Proposition 17.4 (Sufficient conditions for strict local minimum)**.** *Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable (C^2) over an open set S. Suppose $x \in S$ such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x) \succ 0$ (positive definite). Then, x^* is a strict unconstrained local minimum.*

Proof of Proposition 17.4. Fix $d \in \mathbb{R}^n$. Note that $d^\top \nabla^2 f(x^*) d \geqslant \lambda_{\min} \|d\|^2$, where $\lambda_{\min}$ is the smallest eigenvalue of $\nabla^2 f(x^*)$.

$$\begin{aligned}
f(x^* + d) - f(x^*) &= \nabla f(x^*)^\top d + \frac{1}{2} d^\top \nabla^2 f(x^*) d + O(\|d\|^2) && (2) \\
&\geqslant \frac{\lambda_{\min}}{2} \|d\|^2 + O(\|d\|^2) \\
&= \left(\frac{\lambda_{\min}}{2} + \frac{O(\|d\|^2)}{\|d\|^2}\right) \|d\|^2 \\
&> 0 && (3)
\end{aligned}$$

Equality 2 follows from using the 2nd Order Taylor expansion.
Inequality 3 follows for sufficiently small $\|d\|$.

Therefore, x^* must be a strict local minimum. ■

17.2 Stationary points

For non-convex problems we must accept that gradient descent cannot always find the global minimum, but can it at least find a nearby local minimum? It turns out that for many problems we care about it can usually, but not always!

Definition 17.5 (Stationary point)**.** We say a point $x \in \mathbb{R}^n$ is a stationary point of $f\colon \mathbb{R}^n \to \mathbb{R}$ if $\nabla f(x) = 0$.

Proposition 17.6. *Gradient Descent converges to a stationary point.*

Proof of Proposition 17.6. Suppose $x' = x - \eta \nabla f(x)$. From Taylor expansion we get the following

$$\begin{aligned} f(x') &= f(x) + \nabla f(x)^\top (x' - x) + O(\|x' - x\|) \\ &= f(x) - \eta \|\nabla f(x)\|^2 + O(\eta \|\nabla f(x)\|) \\ &= f(x) - \eta \|\nabla f(x)\|^2 + O(\eta) \end{aligned} \tag{4}$$

Equality 4 is justified because we control η, and $\|\nabla f(x)\|$ is a constant with respect to η.

Now we need to worry about selecting step sizes.

17.2.1 Minimization Rule / Line Search

Given a descent direction d (example $d = -\nabla f(x)$), let our step rate η be as follows

$$\eta \in \operatorname*{argmin}_{\eta \geqslant 0} f(x + \eta d)$$

Using this procedure is called **Line Search** because we search for the best step size along the direction d. However, exact line search can be expensive due to the argmin.

Instead, we can approximate this minimization by using the **Armijo Rule**. Fix

$$\gamma, s, \sigma < 1$$

Put $\eta = \gamma^m s$ where m is the smallest non-negative integer such that

$$f(x) - f(x + \gamma^m s d) \geqslant -\sigma \gamma^m s \nabla f(x)^\top d$$

Think of s as an initial learning rate. If s causes sufficient decrease then stop, otherwise keep multiplying by γ until you do. Typical choices for parameters are

$$\gamma = \frac{1}{2}, \sigma = \frac{1}{100}, s = 1$$

Notice that as long as d satisfies $-\nabla f(x)^T d > 0$ that the inequality ensures that our function sequence will decrease.

Proposition 17.7. *Assume that f if continuous and differentiable (C^1), and let $\{x_t\}$ be a sequence generated by $x_{t+1} = x_t - \eta_t \nabla f(x_t)$ where η_t is selected by the Armijo rule. Then, every limit point of $\{x_t\}$ is a stationary point.*

Proof of Proposition 17.7. Let $\bar{x}$ be a limit point. By continuity $\{f(x_t)\}$ converges to $f(\bar{x})$ and therefore:

$$f(x_t) - f(x_{t+1}) \to 0$$

By definition of Armijo rule:

$$f(x_t) - f(x_{t+1}) \geqslant -\sigma\eta_t \|\nabla f(x_t)\|^2 \tag{5}$$

Suppose for the sake of contradiction that $\bar{x}$ is not a stationary point of f. Then,

$$\limsup_{t\to\infty} -\|\nabla f(x_t)\|^2 < 0$$

By inequality 5, this must mean that $\eta_t \to 0$. This implies $\exists t_0$ such that $\forall t \geqslant t_0$

$$f(x_t) - f(x_t - \frac{\eta_t}{\gamma}\nabla f(x_t)) < \frac{\sigma\eta_t}{\gamma}\|\nabla f(x_t)\|^2$$

Because $\eta_t \to 0$, we know that after some t_0 all step sizes are chosen with a $m \geqslant 1$. Therefore, going back one iteration of Armijo rule was not good enough to satisfy the inequality or else some previous step size would have been chosen.

Now let $\tilde{\eta}_t = \frac{\eta_t}{\gamma}$ and we can continue as follows

$$\begin{aligned}
\frac{f(x_t) - f(x_t - \tilde{\eta}_t\nabla f(x_t))}{\tilde{\eta}_t} < \sigma\|\nabla f(x)\|^2 \quad &\Rightarrow \\
\nabla f(x_t - \tilde{\eta}_t\nabla f(x_t))^T\nabla f(x_t) < \sigma\|\nabla f(x)\|^2 \quad &\Rightarrow \qquad (6)\\
\|\nabla f(x_t)\|^2 \leqslant \sigma\|\nabla f(x_t)\|^2 \quad &\qquad (7)
\end{aligned}$$

Inequality 6 follows from using Mean Value Theorem (MVT)
Inequality 7 follows by taking the limit as $\eta_t \to 0 \Rightarrow \tilde{\eta}_t \to 0$

This is a contradiction because $0 < \sigma < 1$. Therefore, the limit point $\bar{x}$ is a stationary point of f. ■

Therefore, if we can use the Armijo rule to determine step sizes that guarantee that gradient descent will converge to a stationary point. ■

17.3 Saddle points

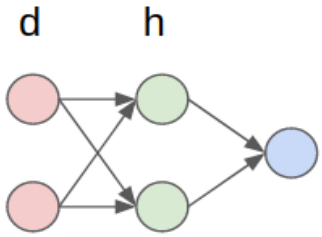
Now that we have found that gradient descent will converge to a stationary point, how concerned should we be that the stationary point is not a local minimum?

Definition 17.8 (Saddle points)**.** Saddle points are stationary points that are not local optima.

This means that if f is twice differentiable then $\nabla^2 f(x)$ has both positive and negative eigenvalues.

17.3.1 How do saddle points arise?

In most non-convex problems there exists several local minima. This is clear to see in problems that have natural symmetry such as in a two layer fully connected neural networks.



Notice that any permutation of the units of the hidden layer would preserve the same function, so we have at least $h!$ local minima. Typically a convex combination of two distinct local minima in a non-convex problem is not a local minimum. In the case where $\nabla f(x)$ is differentiable, then by Mean Value Theorem we know that there must exist another stationary point between any two local minima. So, there often exists at least one saddle point between any two distinct local minima. Hence, many local minima tends to lead to many saddle points.

However, recent work has demonstrated that saddle points are usually not a problem.

1. Gradient descent does not converge to strict saddle points from a random initialization. [GHJY15]
2. Saddle points can be avoided with noise addition. [LPP$^+$17]

References

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. *CoRR*, abs/1503.02101, 2015.

[LPP$^+$17] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *CoRR*, abs/1710.07406, 2017.