MS&E 213 / CS 269O : Chapter 2 - Smooth Functions*

By Aaron Sidford (sidford@stanford.edu)

May 7, 2017

In the last chapter we saw that unconstrained function minimization is impossible without any assumptions on the objective function and that if all we assume is that our function is Lipschitz continuous that unconstrained function minimization is possible, but an exponential dependence on dimension is required. Here we consider a different assumption on or objective function, namely smootheness, and show that although computing ϵ-suboptimal points may still require an exponential number of queries with just a 0-order or 1-order oracle, we can create descent algorithms that achieve some local suboptimality under these assumptions.

There are several goals of this section. Primarily this section is meant to introduce *smoothness*, a natural assumption we will use on objective functions, and *gradient descent*, an extremely popular algorithm in theory and in practice. Secondarily, we will review some multivariable calculus, analysis, and possibly linear algebra that we will use repeatedly throughout the class.

1 Smoothness

In this chapter, as usual consider the unconstrained minimization problem

$$\min_x f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is our objective function that we can only access through some restrictive oracle model.

Whereas we considered value oracles in the last chapter, here we assume that we have a *gradient oracle*, that is an oracle that on some query point gives us the gradient of the function at the point. First we briefly call the definition of a gradient and different function.

Definition 1 (Gradient of Differentiable Function)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ that is differentiable at $x \in \mathbb{R}^n$ we let $\bigtriangledown f(x) \in \mathbb{R}^n$ where $\bigtriangledown f(x)_i = \frac{\partial}{\partial x_i} f(x)$ for all $i \in [n]$ denote the gradient of f at x. We recall that f is differentiable at x if and only if the following holds[1]

$$\lim_{h \to \vec{0} \in \mathbb{R}^n} \frac{\left| f(x+h) - f(x) - \bigtriangledown f(x)^\top h \right|}{\|h\|_2} = 0 \,.$$

Thus we see that locally a gradient measures how quickly a function changes when we move in a direction. If the gradient could change arbitrarily quickly this wouldn't be helpful for minimization. However, if it doesn't change too quickly then one might hope that by moving opposite of the direction of the gradient we might be able to sufficiently decrease the functions value. The way we formally quantify this in this chapter is through the following popular definition of *smoothness*.

*These notes are a work in progress. They are not necessarily a subset or superset of the in-class material and there may also be occasional *TODO* comments which demarcate material I am thinking of adding in the future. These notes will converge to a superset of the class material that is TODO-free. Your feedback is welcome and highly encouraged. If anything is unclear, you find a bug or typo, or if you would find it particularly helpful for anything to be expanded upon, please do not hesitate to post a question on the discussion board or contact me directly at sidford@stanford.edu.

[1] Note that the choice of norm here is arbitrary is f is defined over a finite dimensional vector space.

Definition 2. We say a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *L-smooth* if for all $x, y \in \mathbb{R}^n$ we have that

$$\| \bigtriangledown f(x) - \bigtriangledown f(y)\|_2 \leq L \cdot \|x - y\|_2$$

Where we recall that $\bigtriangledown f(x)$ is the gradient of f at x, i.e. $\bigtriangledown f(x) \in \mathbb{R}^n$ with $[\bigtriangledown f(x)]_i = \frac{\partial}{\partial x_i} f(x)$.

Now as usual, a natural question to ask is how many queries to a gradient oracle are needed to minimize a smooth function? Unfortunately, as with the case of Lipschitz functions, nothing stops the input being a function that is 1 everywhere except for a small ball where it smoothly drops below this value. Consequently, it still take an exponential in dimension number of queries to minimize a smooth function.

However, as we have argued it seems that smoothness should allow us to make progress locally by moving in the direction of the gradient. If the gradient is large at a point it seems like one can make much progress decreasing the value of the function by by moving against the direction of the gradient. In the next section (and the bulk of this chapter) we formalize this by giving an algorithm, *gradient descent*, that always makes progress (i.e. decreases function value) and computes an ϵ-critical point (that is a point where the norm of the gradient is small) at a rate free of dimension.

2 How to Locally Minimize Smooth Functions?

So how do we "locally minimize a smooth function"? A natural idea here is gradient descent. Smoothness tells us that the gradient can't change too quickly, so if the gradient is large somewhere and we move in that direction or its opposite the gradient should stay decently large and we should increase or decrease the function.

Let's analyze this method. Let et $x_{k+1} = x_k - \eta \bigtriangledown f(x)$ for some step size η. We call η here the *step size* of the method. There are many procedures for choosing the step size, however we will consider a fixed step size for now. To analyze this method we ultimately prove the following.

Lemma 3. *Let f be L-smooth and suppose that $y = x - \eta \bigtriangledown f(x)$ then*

$$\left|f(y) - \left[f(x) - \eta\| \bigtriangledown f(x)\|_2^2\right]\right| \leq \frac{\eta^2 L}{2} \cdot \| \bigtriangledown f(x)\|_2^2$$

Consequently if $\eta = \frac{1}{L}$ we have that

$$f(y) \leq f(x) - \frac{1}{2L}\| \bigtriangledown f(x)\|_2^2 .$$

This lemma says when we move η in the direction of the gradient we decrease the function value at a rate that depends linearly on η with an additive penalty that depends quadratically η. Consequently, there is always a step that makes progress, where the best progress guaranteed from thsi worst case bound coming from when we pick $\eta = \frac{1}{L}$. We defer the proof of this lemma to the next section where we give a better understanding of smooth functions and the derivation of this algorithm. Instead, here we analyze the performance of the algorithm that repeatedly applies this result, i.e. *gradient descent* with a fixed step size.

Lemma 4 (Gradient Descent Computes Critical Points)**.** *Let f be a L-smooth function for any $x_0 \in \mathbb{R}^n$ consider the algorithm that computes $x_{k+1} = x_k - \frac{1}{L} \bigtriangledown f(x_k)$ for all $k \geq 0$. For all $\epsilon > 0$ with $\frac{2L \cdot [f(x_0) - f_{\min}]}{\epsilon^2}$ call to a gradient oracle this procedure computes an ϵ-critical point, that is a point x such that $\| \bigtriangledown f(x)\|_2 \leq \epsilon$.*

Proof. For all $k \geq 0$ let $x_{k+1} = x_k - \frac{1}{L} \bigtriangledown f(x_k)$. Then we have by the previous lemma that for all k it is the case that

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\| \bigtriangledown f(x_k)\|_2^2$$

Consequently, with k gradients computation we can compute x_k such that

$$f_* - f(x_0) \leq f(x_k) - f(x_0) \leq -\frac{1}{2L}\sum_{i=0}^{k-1} \| \bigtriangledown f(x_k)\|_2^2$$

Consequently

$$\frac{1}{k}\sum_{i=0}^{k-1} \| \bigtriangledown f(x_i)\|_2^2 \leq \frac{2L \cdot (f(x_0) - f_*)}{k}$$

and therefore it must be the case that $\| \bigtriangledown f(x_i)\|_2^2 \leq \frac{2L \cdot [f(x_0) - f_{\min}]}{k}$ for some $i \in \{0, ...k-1\}$ and picking $k = \frac{2L \cdot [f(x_0) - f_{\min}]}{\epsilon^2}$ yields the result. □

Note from the proof of the above actually gave something stronger then claimed as it gives bounds on the average norm of the gradient over the life of gradient descent.

The above lemma shows that gradient descent is a descent algorithm i.e. it always makes progress) that converges to a point where the norm of the gradient is small. This is a reasonable proxy for a local minimum in many cases and thus can be useful in many settings.

A natural question is, is this dependence on ϵ tight? We actually don't know in full but this has been a recently active area of research (TODO GIVE CITATIONS).

In the rest of this chapter we take a closer look at this analysis so we can better leverage it later in the course.

3 Upper Bounds

While we could prove Lemma 3 directly, I believe such a proof would be a bit mysterious. However, I think there few principled pieces going on and here we will break down the proof into these more principled pieces to prove the lemma.

We start with a lemma giving an integral formula for the difference between a function and its first order Taylor approximation.

Lemma 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, then for all $x, y \in \mathbb{R}^n$ and $x_t = x + t(y - x)$ for $t \in [0, 1]$ we have*

$$f(y) - f(x) = \int_0^1 \bigtriangledown f(x_\alpha)^\top (y - x) \cdot d\alpha$$

and therefore

$$f(y) - f(x) - \bigtriangledown f(x)^\top (y - x) = \int_0^1 \left(\bigtriangledown f(x_\alpha) - \bigtriangledown f(x)\right)^\top (y - x) \cdot d\alpha\,.$$

Proof. Let $g(t) \stackrel{\text{def}}{=} f(x_t)$ for $t \in [0, 1]$. Now, I claim that $g'(t) = \bigtriangledown f(x_t)^\top (y - x)$. Later in the course I will simply state things like this without proof. However, to get a refresh in how these sorts of proofs work, we'll work it out from scratch just for these notes.

Note that since f is differentiable, by assumption (and the fact that all norm in finite dimensions are the same up to a constant) we know that for all $x \in \mathbb{R}^n$ it is the case that for any norm $\| \cdot \|$

$$\lim_{h \to 0} \frac{\left| f(x_t + h) - f(x_t) - \bigtriangledown f(x_t)^\top h \right|}{\|h\|} = 0$$

Consequently, considering h of the form $\alpha(y-x)$ we have that

$$\begin{aligned}
0 &= \lim_{\alpha\to 0} \frac{\left|f(x_t+\alpha(y-x)) - f(x_\alpha) - \bigtriangledown f(x_t)^\top \cdot (\alpha(y-x))\right|}{\|t\cdot(y-x)\|} \\
&= \frac{1}{\|y-x\|}\cdot \lim_{\alpha\to 0} \frac{\left|f(x_t+\alpha(y-x)) - f(x_t) - \alpha\cdot\bigtriangledown f(x_t)^\top (y-x)\right|}{|\alpha|} \\
&= \frac{1}{\|y-x\|}\cdot \lim_{\alpha\to 0} \left|\frac{g(t+\alpha)-g(t)-\alpha\cdot\bigtriangledown f(x_t)^\top(y-x)}{\alpha}\right| .
\end{aligned}$$

Consequently

$$g'(t) = \lim_{\alpha\to 0}\frac{g(t+\alpha)-g(t)}{\alpha} = \bigtriangledown f(x_t)^\top (y-x)\,.$$

The claim then follows from the fundamental theorem of calculus, we have

$$f(y)-f(x) = g(1)-g(0) = \int_0^1 \bigtriangledown f(x_\alpha)^\top (y-x)\cdot d_\alpha$$

subtracting$\bigtriangledown f(x)^\top(y-x) = \int_0^1 \bigtriangledown f(x)^\top(y-x)d\alpha$ to each side of the equation then yields the result. □

What does this lemma say? It says we can explain how good a Taylor series expansion is around a point in terms of how much the gradient changes relative to the direction we move in. Thus, naturally this says that if we assume smoothness we can upper bound the Taylor series expansion in terms of a quadratic when our function is smooth. We show this formally using Cauchy Schwarz.

Lemma 6 (Cauchy Schwarz)**.** *For $x, y \in \mathbb{R}^n$ we have $\left|x^\top y\right| \le \|x\|_2 \cdot \|y\|_2$.*

Proof. Note that the claim is equivalent to $(x^\top y)^2 \le \|x\|_2^2\cdot\|y\|_2^2$. We prove this through the following

$$\begin{aligned}
\|x\|_2^2\cdot\|y\|_2^2 - (x^\top y)^2 &= \left(\sum_{i\in[n]} x_i^2\right)\cdot\left(\sum_{i\in[n]} y_i^2\right) - \left(\sum_{i\in[n]} x_i y_i\right)^2 \\
&= \sum_{i,j\in[n]} x_i^2\cdot y_j^2 - \sum_{i,j\in[n]} x_i y_i x_j y_j\,.
\end{aligned}$$

Now note that when $i=j$ we have $x_i^2 y_i^2 = x_i y_i x_j y_j$ and that for every setting of $i<j$ there is a term with the values of i and j reversed (which does not affect the $x_i y_i x_j y_j$ value) and therefore

$$\|x\|_2^2\cdot\|y\|_2^2 - (x^\top y)^2 = \sum_{i<j\in[n]} \left(x_i^2\cdot y_j^2 + x_j^2\cdot y_i^2 - 2\cdot x_i y_i x_j y_j\right) = \sum_{i<j\in[n]} (x_i\cdot y_j - x_j\cdot y_i)^2 \ge 0\,.$$

□

From Cauchy Schwarz and Lemma 5 we have the following

Lemma 7. *If $f:\mathbb{R}^n\to\mathbb{R}$ is L-smooth then for all $x, y\in\mathbb{R}^n$ we have that*

$$\left|f(y) - \left[f(x) + \bigtriangledown f(x)^\top (y-x)\right]\right| \le \frac{L}{2}\|y-x\|_2^2$$

Proof. For all t let $x_t = x + t(y-x)$. By Lemma 5 and Cauchy Schwarz we have

$$\begin{aligned}\left|f(y) - f(x) - \bigtriangledown f(x)^\top (y-x)\right| &= \left|\int_0^1 \left(\bigtriangledown f(x_\alpha) - \bigtriangledown f(x)\right)^\top (y-x)\cdot d_\alpha\right| \\ &\leq \int_0^1 \left|\left(\bigtriangledown f(x_\alpha) - \bigtriangledown f(x)\right)^\top (y-x)\right| \cdot d_\alpha \\ &\leq \int_0^1 \| \bigtriangledown f(x_\alpha) - \bigtriangledown f(x)\|_2 \cdot \|y-x\|_2 \cdot d_\alpha \,.\end{aligned}$$

Consequently, by the smoothness of f we have

$$\| \bigtriangledown f(x_\alpha) - \bigtriangledown f(x)\|_2 \leq L \cdot \|x_\alpha - x\|_2 = L\alpha \cdot \|y-x\|_2$$

□

Note that the proof of our early inequality follows from this.

Proof of Lemma 3. By Lemma 7 we have that for all $x, y \in \mathbb{R}^n$ we have that

$$\left|f(y) - \left[f(x) + \bigtriangledown f(x)^\top (y-x)\right]\right| \leq \frac{L}{2}\|y-x\|_2^2$$

and consequently, when $y = x - \eta \bigtriangledown f(x)$ we have

$$\left|f(y) - \left[f(x) - \eta\| \bigtriangledown f(x)\|_2^2\right]\right| \leq \frac{\eta^2 L}{2}\| \bigtriangledown f(x)\|_2^2$$

as desired. □

Note, this this gives us another way to obtain the gradient descent step. Note that

$$\text{argmin}_y U(y) \stackrel{\text{def}}{=} f(x) + \bigtriangledown f(x)^\top (y-x) + \frac{L}{2}\|y-x\|_2^2$$

has the property that for as $\|y\|_2 \to \infty$ goes to infinity then $U(y) \to \infty$. Consequently, the minimizer of $U(y)$ occurs when $\bigtriangledown U(y) = \vec{0}$. However, $\bigtriangledown U(y) = \bigtriangledown f(x) + L(y-x)$ and consequently $\bigtriangledown U(y) = \vec{0}$ if and only if $y = x - \frac{1}{L} \bigtriangledown f(x)$.[2]

4 Second Order Explanation of Non-convex Gradient Descent

Another nice way to think about smoothness and the convergence of gradient descent is through second order Taylor approximations to f. Recall the following defintion of the Hessian.

Definition 8 (Hessian)**.** For twice differentiable function $f \in \mathbb{R}^n \to \mathbb{R}$ the *Hessian* of f at x is, $\bigtriangledown^2 f(x) \in \mathbb{R}^{n\times n}$ is defined so that for all $i, j \in [n]$ we have $[\bigtriangledown^2 f(x)]_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$ and satisfies for all x, α

$$\lim_{h\to 0} \frac{\| \bigtriangledown f(x+h) - \bigtriangledown f(x) - \bigtriangledown^2 f(x)h\|}{\|h\|} = 0\,.$$

Using this we can characterize how the gradient changes between different points.

[2] This follows more generally from the fact that $U(y)$ is convex in y and therefore its minimum value is achieved at any point whose gradient is 0. We discuss this in greater details in later notes when we formally define convexity.

Lemma 9. *Let $f \in \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. Then for all $x, y \in \mathbb{R}^n$ and $x_t = x + t(y-x)$ for $t \in [0,1]$ we have*

$$\bigtriangledown f(y) - \bigtriangledown f(x) = \int_0^1 \bigtriangledown^2 f(x_\alpha)(x-y) \cdot d\alpha$$

Proof. Let $g_i(t) \stackrel{\text{def}}{=} [\bigtriangledown f(x_t)]_i$ for $t \in [0,1]$. Now, I claim that $g_i'(t) = \vec{1}_i^\top \bigtriangledown^2 f(x_t)(y-x)$.
Note that since f is differentiable, by assumption (and the fact that all norm in finite dimensions are the same up to a constant) we know that for all $x \in \mathbb{R}^n$ it is the case that for any norm $\|\cdot\|$

$$\lim_{h \to \vec{0} \in \mathbb{R}^n} \frac{\| \bigtriangledown f(x_t + h) - \bigtriangledown f(x_t) - \bigtriangledown^2 f(x_t)^\top h\|_2}{\|h\|_2} = 0$$

Consequently, considering h of the form $\alpha(y-x)$ we have that

$$\begin{aligned} 0 &= \lim_{\alpha \to 0} \frac{\| \bigtriangledown f(x_t + \alpha(y-x)) - \bigtriangledown f(x_t) - \bigtriangledown^2 f(x_t) \cdot (\alpha(y-x))\|_2}{\|\alpha \cdot (y-x)\|_2} \\ &= \frac{1}{\|y-x\|_2} \cdot \lim_{\alpha \to 0} \frac{\| \bigtriangledown f(x_t + \alpha(y-x)) - \bigtriangledown f(x_t) - \bigtriangledown^2 f(x_t) \cdot (\alpha(y-x))\|_2}{|\alpha|} \\ &= \frac{1}{\|y-x\|} \cdot \lim_{\alpha \to 0} \left| \frac{\sqrt{\sum_{i \in [n]} \left(g_i(t+\alpha) - g_i(t) - \alpha \vec{1}_i^\top \bigtriangledown^2 f(x_t)(y-x) \right)^2}}{\alpha} \right| . \end{aligned}$$

Consequently

$$g_i'(t) = \lim_{\alpha \to 0} \frac{g_i(t+\alpha) - g_i(t)}{\alpha} = 1_i^\top \bigtriangledown^2 f(x_t)(y-x) \, .$$

The claim then follows from the fundamental theorem of calculus, we have

$$[\bigtriangledown f(y) - \bigtriangledown f(x)]_i = g_i(1) - g_i(0) = \int_0^1 \vec{1}_i^\top \bigtriangledown^2 f(x_\alpha)(y-x) \cdot d_\alpha \, .$$

Furthermore, since this holds for all $i \in [n]$ the result follows. □

Combining this with Lemma 5 we immediately get the following characterization of the difference between the function and the Taylor series expansion

Lemma 10. *If f is twice differentiable then for all $x, y \in \mathbb{R}^n$ and $x_t \stackrel{\text{def}}{=} x + t(y-x)$ we have that.*

$$f(y) = f(x) + \bigtriangledown f(x)^\top (y-x) + \int_0^1 \int_0^t (y-x)^\top \bigtriangledown^2 f(x_\alpha)(y-x) d\alpha dt \, .$$

Proof. This Recall from the definition of the gradient that

$$f(x_1) - f(x_0) = \int_0^1 \bigtriangledown f(x_t)^\top (y-x) dt \, .$$

Furthermore, we have that

$$\bigtriangledown f(x_t) - \bigtriangledown f(x_0) = \int_0^t \bigtriangledown^2 f(x_\alpha)(y-x) d\alpha \, .$$

Combining and using that $x_1 = y$ and $x_0 = x$ then yields the result. □

Thus we see that having $z^\top \bigtriangledown^2 f(x)z \leq L\|z\|_2$ also suffices for having the gradient descent guarantee.

Lemma 11. *If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable and has the property that for all $x, z \in \mathbb{R}^n$ it is the case that $z^\top \bigtriangledown^2 f(x)z \leq L \cdot \|x\|_2^2$ then for all $x, y \in \mathbb{R}^n$ we have*

$$f(y) \leq f(x) + \bigtriangledown f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2 .$$

Proof. By Lemma 10 we have that if $x_t = x + t(y - x)$ for all $t \in [0, 1]$ then

$$f(y) = f(x) + \bigtriangledown f(x)^\top (y - x) + \int_0^1 \int_0^t (y - x)^\top \bigtriangledown^2 f(x_\alpha)(y - x) d\alpha dt .$$

However, by assumption we have that

$$\int_0^1 \int_0^t (y - x)^\top \bigtriangledown^2 f(x_\alpha)(y - x) d\alpha dt \leq \int_0^1 \int_0^t L \cdot \|y - x\|_2^2 d\alpha dt$$

Since

$$\int_0^1 \int_0^t 1 \cdot d\alpha dt = \int_0^1 t \cdot dt = \frac{1}{2}$$

the result follows. □

Consequently, we have that we can compute an ϵ-critical point using gradient descent at the same rate we got for a L-smooth function provided r all $x, z \in \mathbb{R}^n$ it is the case that $z^\top \bigtriangledown^2 f(x)z \leq L \cdot \|z\|_2^2$. We conclude by showing that this assumption is more general then simply assuming smoothness.

Lemma 12. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and and suppose that f is twice differentiable at some $x \in \mathbb{R}^n$ and we have that*

$$z^\top \bigtriangledown^2 f(x)z > L \cdot \|z\|_2^2$$

for some $z \in \mathbb{R}^n$. Then, f is not L-smooth.

Proof. If f is not differentiable the claim follows trivially, so suppose without loss of generality that f is differentiable. Now, for all t let $x_t = x + t \cdot z$ and let $g(t) = f(x_t)$. We have that $g'(t) = \bigtriangledown f(x_t)^\top z$. Furthermore since f is twice differentiable at x we have that g is twice differentiable at 0 and therefore that for all $\epsilon > 0$ there exist some $\delta > 0$ such that

$$\left| \frac{g'(\delta) - g'(0)}{\delta} - g''(0) \right| \leq \epsilon$$

However, $g''(t) = z^\top \bigtriangledown^2 f(x_t)^\top z$ and therefore we have that

$$\frac{1}{\delta} \left[(\bigtriangledown f(x_\delta) - \bigtriangledown f(x_0))^\top z \right] \geq z^\top \bigtriangledown^2 f(x_t)^\top z - \epsilon$$

and by Cauchy Schwarz and our assumption

$$\| \bigtriangledown f(x_\delta) - \bigtriangledown f(x_0)\|_2 \cdot \|z\|_2 \geq \delta \cdot \left[L \cdot \|z\|_2^2 - \epsilon \right] .$$

Since $\|x_\delta - x_0\|_2 = \delta\|z\|_2$by picking even smaller ϵ we have that

$$\| \bigtriangledown f(x_\delta) - \bigtriangledown f(x_0)\|_2 \geq L \cdot \|x_\delta - x_0\| - \epsilon$$

for all ϵthereby giving the desired contradiction. □

References