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

By Aaron Sidford (sidford@stanford.edu)

October 17, 2020

In the last chapter we saw that unconstrained function minimization is impossible without any assumptions on the objective function. Further, we saw 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 our objective function, namely smoothness, and show that although computing ϵ-optimal points may still require an exponential number of queries with just a zero-order or first-order oracle, we can create descent algorithms that achieve some local optimality 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 we consider our study of unconstrained function minimization. As usual, we let $f \in \mathbb{R}^n \to \mathbb{R}$ denote our objective function, assume that we can only access through some restrictive oracle model, and consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) \,.$$

Further, throughout this section we let $f_* \stackrel{\text{def}}{=} \inf_{x \in \mathbb{R}} f(x)$.

Whereas in the last section we considered value oracles for f and assumed that f is continuous, here we consider *gradient oracles*, that is oracle which when queried at a point output the gradient of the function at the point, and make the additional assumption that f is differentiable. First we briefly recall the definition of a gradient and a differentiable function.

Definition 1 (Differentiability and Gradients)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ that is differentiable at $x \in \mathbb{R}^n$ we let $\nabla f(x) \in \mathbb{R}^n$ where $\nabla f(x)_i = \frac{\partial}{\partial x_i} f(x)$ for all $i \in [n]$ denote the gradient of f at x. Recall that f is

*These notes are a work in progress. They are not necessarily a subset or superset of the in-class material, there may also be occasional *TODO* comments which demarcate material I am thinking of adding in the future, and citations are often omitted. These notes are intended converge to a superset of the class material that is TODO-free with a more complete set of citations and pointers to the literature. 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 Piazza or contact me directly at sidford@stanford.edu.

differentiable at $x \in \mathbb{R}^n$ if and only if the following holds for some vector $g \in \mathbb{R}^n$[1]

$$\lim_{h \to \vec{0} \in \mathbb{R}^n} \frac{\left|f(x+h) - f(x) - g^\top h\right|}{\|h\|_2} = 0\,. \tag{1.1}$$

Note that when this holds, $g = \nabla f(x)$. Further, we say that f is differentiable if and only if it is differentiable at all $x \in \mathbb{R}^n$.

Consequently, we see that locally a gradient measures how quickly a function changes when we move in a direction. Considering $y \stackrel{\text{def}}{=} x + h$ in (1.1) it says that locally $f(y) \approx f(x) + \nabla f(x)^\top (y - x)$. If the gradient could change arbitrarily quickly, it wold be difficult to use this for minimization. However, if it doesn't change too quickly, then one might hope that by moving opposite of the direction of the gradient, i.e. considering $y = x - \eta \nabla f(x)$, we might be able to sufficiently decrease the function's value (as predicted locally by differentiability, i.e. the first-order Taylor approximation $f(x) + \nabla f(x)^\top (y - x)$).

Analogous to how in the last chapter we quantified continuity with Lipschitz continuity, in this chapter we quantify differentiability through the following natural and popular definition of *smoothness*. This definition given below is formally assuming that the gradient is Lipschitz continuous in ℓ_2.

Definition 2. Differentiable[2] $f : \mathbb{R}^n \to \mathbb{R}$ is *L-smooth* if and only for all $x, y \in \mathbb{R}^n$ we have that

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

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

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 was the case of Lipschitz functions, there are functions that have value 1 everywhere except for a small region where it smoothly drops below this value (though smoothness does imply that such a ball is larger). Consequently, it still takes 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. Indeed, as the following standard lemma shows, whenever the gradient of a differentiable function is non-zero then a small enough step in the direction away from the gradient decreases the function locally.

Lemma 3. *If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable $\nabla f(x) \neq \vec{0}$ then all small enough $\eta > 0$*

$$f(x - \eta \nabla f(x)) < f(x)\,.$$

Consequently, if x_* is a minimizer of f then $\nabla f(x_*) = 0$

Proof. For all $\eta > 0$ let $h_\eta = \eta \nabla f(x)$. Differentiability implies

$$0 = \lim_{\eta \to 0^+} \frac{|f(x - h_\eta) - f(x) - \nabla f(x)^\top h_\eta|}{\|h_\eta\|_2} = \frac{1}{\|\nabla f(x)\|_2} \lim_{\eta \to 0^+} \left| \frac{1}{\eta}[f(y - h_\eta) - f(x)] - \|\nabla f(x)\|_2^2 \right|\,.$$

Consequently, for all small enough $\eta > 0$ the point $x_\eta = x - h_\eta$ satisfies

$$\left| \frac{1}{\eta}[f(x_\eta) - f(x)] - \|\nabla f(x)\|_2^2 \right| \leq \frac{1}{2}\|\nabla f(x)\|_2^2 \text{ and } \frac{1}{\eta}[f(x_\eta) - f(x)] \leq -\frac{1}{2}\|\nabla f(x)\|_2^2\,.$$

□

[1] Note that since f is defined over a finite dimensional vector space the choice of norm here is arbitrary as all norms in finite dimensions are equivalent, in the sense that they are within a finite multiplicative ratio of each other.

[2] Note that this assumption of differentiable is slightly redundant. If f is differentiable in every coordinate and $\|\nabla f(x) - \nabla f(y)\|_2 \leq L \cdot \|x - y\|_2$ for all $x, y \in \mathbb{R}^n$ then f is differentiable. Note however, if the condition $\|\nabla f(x) - \nabla f(y)\|_2 \leq L \cdot \|x - y\|_2$ for all $x, y \in \mathbb{R}^n$ was not assumed then f is not necessarily differentiable.

Further, if the gradient is large at a point and the gradient doesn't change too quickly, then it seems like one could quantify this decrease in the function 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. As the preceding standard lemma shows, having gradient 0 is necessary for a point to be a minimizer of the function. However, it is not sufficient. Consequently, the goal of this section can be viewed as obtaining a descent method that leveraging smoothness, approximately obtains one of the criteria for minimizing

2 How to Locally Minimize Smooth Functions?

So how do we "locally minimize a smooth function"? A natural idea here is *gradient descent*, which consists of simply starting at an initial point $x_0 \in \mathbb{R}^n$ and iterating, i.e. repeatedly applying, $x_{k+1} = x_k - \eta_k \nabla f(x_k)$ for $k \geq 0$. Smoothness implies that the gradient can't change too quickly, so if the gradient is large at iteration k, i.e. for x_k, and we pick η_k appropriately, the gradient should stay decently large and we may hope to decrease the function value by such a step.

Gradient descent is a natural well-studied widely-applicable method. There are many strategies for picking the η_k, known as step sizes), and we will see a variety of methods inspired by gradient descent later in the course. However, while different *step-size schedules*, i.e. strategies for picking η_k, can be beneficial in different contexts, in this course we will often consider simple *fixed step size* schemes, i.e. where $\eta_k = \eta$ for some η for all steps.

Let's analyze this method. Formally, let $x_{k+1} = x_k - \eta \nabla f(x)$ for some step size $\eta > 0$, some initial point $x_0 \in \mathbb{R}^n$, and iteration $k \geq 0$. To analyze this method, later in this chapter we prove the following lemma.

Lemma 4. *Let f be L-smooth and suppose that $y = x - \eta \nabla f(x)$ then*

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

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

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

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 this 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 5 (Gradient Descent Computes Critical Points)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth, let $x_0 \in \mathbb{R}^n$, and let $\epsilon > 0$. Consider the procedure* gradient descent*, which for all $k \geq 0$ outputs x_k if $\|\nabla f(x_k)\|_2 \leq \epsilon$ and otherwise lets $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$. This procedure outputs an ϵ-critical point, i.e. a point x such that $\|\nabla f(x)\|_2 \leq \epsilon$, and makes at most $\lceil \frac{2L \cdot [f(x_0) - f_*]}{\epsilon^2} \rceil$ gradient queries where $f_* = \inf_{x \in \mathbb{R}^n} f(x)$.*

Proof. For all $k \geq 0$ applying Lemma 4 with $x = x_k$ and $y = x_{k+1}$ yields that

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

By summing this inequality $k = 0, ..., T-1$ we have that

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

Consequently,

$$\frac{1}{T} \sum_{i=0}^{T-1} \|\nabla f(x_i)\|_2^2 \leq \frac{2L \cdot [f(x_0) - f_*]}{T} .$$

Note that this implies that $\|\nabla f(x_i)\|_2^2 \leq \frac{2L \cdot [f(x_0) - f_*]}{T}$ for some $i \in \{0, ...T-1\}$. Therefore the algorithm outputs a critical point in at most $T = \lceil 2L \cdot [f(x_0) - f_*] \cdot \epsilon^{-2} \rceil$ iterations. Since, in each iteration the procedure at most one gradient query is needed the result follows. □

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. It also 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 optimal? This was a long-standing open problem, however recently it was indeed shown to be optimal [1].

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

3 Upper and Lower Bounds from Smoothness

While we could prove Lemma 4 directly, such a proof might look a little mysterious. I think there are few principled pieces in proving this result and it is helpful to break down the proof into these more principled pieces to prove the lemma. Moreover, the techniques we use in this section, integrating to varied degrees. Taylor expansions between points and applying Cauchy Schwarz, will arise multiple times throughout the course.

We start with a lemma giving an integral formula for the difference between a function and its first order Taylor approximation. This is a natural result in multivariable calculus, however we prove it from first principles.

Lemma 6. *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*

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

and therefore

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

Proof. Let $g(t) \stackrel{\text{def}}{=} f(x_t)$ for $t \in [0, 1]$. Now, I claim that $g'(t) = \nabla f(x_t)^\top (y - x)$. Later in the course I may simply state things like this without proof; it follows near immediately by chain-rule for multivariable differentiable functions. However, as a small refresher, we'll prove facts like this from first principles in these notes.

Note that since f is differentiable, by assumption we know that for all $x \in \mathbb{R}^n$ it is the case that for any norm $\|\cdot\|$

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

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

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

Consequently

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

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

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

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

What does this lemma say? It says we can bound how good an approximation the Taylor series expansion around a point is 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 7 (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 6 we have the following

Lemma 8. *If $f:\mathbb{R}^n\to\mathbb{R}$ is L-smooth then for all $x, y\in\mathbb{R}^n$ we have that*

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

Proof. For all t let $x_t = x + t(y - x)$. By Lemma 6 and Cauchy Schwarz we have

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

Consequently, by the smoothness of f we have

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

□

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

Proof of Lemma 4. By Lemma 8 we have that for all $x, y \in \mathbb{R}^n$ we have that

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

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

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

as desired. □

4 Geometry and a General Algorithm Framework

Interestingly, the analysis in the preceding section gives us an alternative way to think about smoothness and gradient descent. In particular, Lemma 8 implies that for all x, y we have

$$\begin{aligned}f(y) \leq U_x(y) &\stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y - x\|_2^2 \text{ and} \\ f(y) \geq L_x(y) &\stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y-x) - \frac{L}{2}\|y - x\|_2^2\,.\end{aligned}$$

In other words, smoothness implies that at any point x if we add a quadratic penalty to the value of the first-order Taylor approximation evaluated at y, i.e. $U_x(y)$, then this quadratic upper bounds y. Similarly, if we subtract a quadratic penalty to the value of the first-order Taylor approximation evaluated at y, i.e. $L_x(y)$, then this lower upper bounds y. Consequently, with one value and gradient query at x for a smooth function we can compute upper and lower bounds, i.e. $U_x(y)$ and $L_x(y)$, that everywhere upper and lower bounds f.

Importantly, the upper bound $U_x(y)$ that smoothness implies has the property that $U_x(x) = f(x)$ and consequently, $\min_{y \in \mathbb{R}^n} U_x(y) \leq f(x)$. Interestingly, there is a natural general descent procedure associated with any such upper bound. Consider the following general setup, where for some function $f : \mathbb{R}^n \to \mathbb{R}$ we are given an *upper-bound oracle*, that when queried at some point x it outputs a function $U_x : \mathbb{R}^n \to \mathbb{R}$ with the properties that $U_x(y) \geq f(y)$ for all $y \in \mathbb{R}^n$, i.e. it is an upper bound, and $U_x(x) = f(x)$, i.e. it has the

same value of $f(x)$ at x.[3] Such an oracle gives an upper bound of f that cannot be too much of an upper bound as it is constrained to have the same value as the function at the point queried.

How would we use such an upper-bound oracle to minimize f? Given an initial point x_0, there is a natural procedure one can consider of repeatedly applying $x_{k+1} = \text{argmin}_{x_k \in \mathbb{R}^n} U_{x_k}(x)$ (provided $U_y(x)$ has a minimizer). Note that by the assumptions of an upper-bound oracle we have

$$f(x_{k+1}) \leq U_{x_k}(x_{k+1}) = U_{x_k}(x_k) + \left[\min_{x_k} U_{x_k}(x) - U_{x_k}(x_k)\right] = f(x_k) - \Delta_k \tag{4.1}$$

where $\Delta_k \stackrel{\text{def}}{=} -[\min_{x_k} U_{x_k}(x) - U_{x_k}(x_k)]$. Now, since $U_{x_k}(x_k) = f(x_k)$ we see that $\Delta_k \geq 0$. Consequently, this method is a descent method, i.e. $f(x_k)$ decreases monotonically. Further, just saw in the proof of Lemma 5 we can sum 4.1 for $k = 0$ to $T-1$ to obtain that

$$f_* - f(x_0) \leq f(x_T) - f(x_0) \leq -\sum_{k=0}^{T-1} \Delta_k \text{ where } f_* \stackrel{\text{def}}{=} \text{argmin}_x f(x)$$

and conclude that

$$\frac{1}{T}\sum_{k=0}^{T-1} \Delta_k \leq \frac{f(x_0) - f_*}{T} \text{ and } \exists i_* \in \{0, 1, ..., T-1\} \text{ with } \Delta_{i_*} \leq \frac{f(x_0) - f_*}{T}\,. \tag{4.2}$$

Interestingly, gradient descent is in fact an instance of this general framework. Consider again

$$U_x(y) \stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2\,.$$

Note that this function is differentiable and has the property that as $\|y\|_2 \to \infty$ goes to infinity then $U(y) \to \infty$. Consequently, the minimizer of $U(y)$ occurs when $\nabla U(y) = \vec{0}$. However, $\nabla U(y) = \nabla f(x) + L(y - x)$ and consequently $\nabla U(y) = \vec{0}$ if and only if $y = x - \frac{1}{L}\nabla f(x)$. Consequently, we see that $y = x - \frac{1}{L}\nabla f(x)$ is the minimizer of $U_x(y)$.[4]This implies that gradient descent with step size $\frac{1}{L}$, i.e. $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$, is the same as the descent algorithm we provided for upper bound oracles, i.e. $x_{k+1} = \text{argmin}_x U_{x_k}(x)$! In other words, gradient descent simply considers the quadratic upper bound implied by smoothness, minimizes this upper bound, and repeats! Further, the analysis we saw for critical point point computation is a special case of (4.2) where we note that $\Delta_k = \frac{1}{2L}\|\nabla f(x_k)\|_2^2$ in this special case.

5 Second Order Explanation of Gradient Descent

Another useful way to think about smoothness and the convergence of gradient descent is through second-order Taylor approximations to f (in the case that f is twice differentiable). Recall the following definition of the Hessian and twice differentiability.

Definition 9 (Hessian)**.** For $f \in \mathbb{R}^n \to \mathbb{R}$ that is twice differentiable at $x \in \mathbb{R}^n$ we let $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ where $[\nabla^2 f(x)]_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$ for all $i, j \in [n]$ denote the *Hessian* of f at x. Recall that f is twice differentiable at x if and only if the following holds for some matrix $\mathbf{H}$

$$\lim_{h \to 0} \frac{\|\nabla f(x + h) - \nabla f(x) - \mathbf{H}h\|}{\|h\|} = 0\,.$$

[3]More generally, $\min_{y \in \mathbb{R}^n} U_x(y) \leq f(x)$, suffices however we state the definition here for consistency with what is yielded by smoothness and the geometry it implies.

[4]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. It also follows from the fact that $U(y) = f(x) - \frac{1}{2L}\|\nabla f(y)\|_2^2 + \frac{L}{2}\|y - (x - \frac{1}{L}\nabla f(x))\|_2^2$. We will discuss this in greater details in later chapters when we formally define convexity and acceleration.

Note that when this holds $\mathbf{H} = \nabla^2 f(x)$. Further, we say f is twice differentiable if and only if it is twice differentiable at all $x \in \mathbb{R}^n$.

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

Lemma 10. *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*

$$\nabla f(x_t) - \nabla f(x) = \int_0^t \nabla^2 f(x_\alpha)(y - x) \cdot d\alpha$$

Proof. Let $g_i(\alpha) \stackrel{\text{def}}{=} [\nabla f(x_\alpha)]_i$ for $\alpha \in [0, t]$. Now, we show that $g_i'(\alpha) = \vec{1}_i^\top \nabla^2 f(x_\alpha)(y - x)$. (This could also be proven more concisely by chain-rule, but we include a proof from the definition for now.)

Note that since f is twice-differentiable, by assumption we know that for all $x \in \mathbb{R}^n$ it is the case that

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

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

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

Consequently

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

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

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

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

Combining this with Lemma 6 we immediately obtain the following characterization of the difference between the function and the Taylor series expansion

Lemma 11. *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.*

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

Proof. Recall from Lemma 6

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

Consequently, since $x_1 = y$ and $x_0 = x$ we see that

$$f(y) - [\nabla f(x)^\top (y - x)] = f(x_1) - [f(x_0) + \nabla f(x_0)^\top [y - x] = \int_0^1 (\nabla f(x_t) - \nabla f(x_0))(y - x) dt \, .$$

Further, by Lemma 10

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

Combining then yields the result. □

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

Lemma 12. *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 $z^\top \nabla^2 f(x) z \leq L \cdot \|z\|_2^2$ then for all $x, y \in \mathbb{R}^n$ we have*

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

Proof. By Lemma 11 we have that if $x_t = x + t(y-x)$ for all $t \in [0,1]$ then

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

However, by assumption we have that

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

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 that for all $x, z \in \mathbb{R}^n$ it is the case that $z^\top \nabla^2 f(x) z \leq L \cdot \|z\|_2^2$. Interestingly, for twice differentiable functions we can more broadly relate the value of $z^\top \nabla^2 f(x) z / \|z\|^2$ to upper and lower bounds on how well $f(y)$ is approximated by the first order Taylor approximation about a point x evaluated at y, i.e. $f(y) + \nabla f(y)^\top (x-y)$, and upper and lower bounds on $(\nabla f(x) - \nabla f(y))^\top (x-y)$. We prove this in the following more general as we will use it and the proof that underlies it several times in the course.

Lemma 13. *For $\alpha, \beta \in \mathbb{R} \cup \{\pm\infty\}$ and twice differentiable $f : \mathbb{R}^n \to \mathbb{R}$ and any norm $\|\cdot\|$ the following three conditions are equivalent:*

$$\frac{\alpha}{2}\|x-y\|^2 \leq f(x) - f(y) - \nabla f(y)^\top (x-y) \leq \frac{\beta}{2}\|x-y\|^2 \text{ for all } x, y \in \mathbb{R}^n \tag{5.1}$$

$$\alpha\|y-x\|^2 \leq (\nabla f(x) - \nabla f(y))^\top (x-y) \leq \beta\|x-y\|^2 \text{ for all } x, y \in \mathbb{R}^n \tag{5.2}$$

$$\alpha\|z\|^2 \leq z^\top \nabla^2 f(x) z \leq \beta\|z\|^2 \text{ for all } x, z \in \mathbb{R}^n\,. \tag{5.3}$$

Proof. First suppose that (5.1) holds. For arbitrary x, y adding (5.1) with the same inequalities with x and y swapped directly yields 5.2.

Next, suppose (5.1) holds. Let $x, z \in \mathbb{R}^n$ be arbitrary and let $x_t \stackrel{\text{def}}{=} x + tz$ for all $t \in \mathbb{R}$. Then, by the definition of twice differentiability we have

$$0 = \lim_{t \to 0} \frac{\|\nabla f(x+tz) - \nabla f(x) - \nabla^2 f(x)(tz)\|}{\|tz\|}$$

Consequently, taking the limit of positive t converging to 0 and applying that $z = \frac{x_t - x}{t}$ we have

$$0 = \lim_{t\to 0^+} \frac{z^\top(\nabla f(x_t) - \nabla f(x)) - z^\top \nabla^2 f(x) tz}{t} = \lim_{t\to 0^+} \left[\frac{1}{t^2}(\nabla f(x_t) - \nabla f(x))^\top (x_t - x) - z^\top \nabla^2 f(x) z\right].$$

However, by assumption of (5.1) and that $\|x_t - x\| = t^2\|z\|$ we have

$$\alpha\|z\|^2 \le \frac{1}{t^2}(\nabla f(x_t) - \nabla f(x))^\top (x_t - x) \le \beta\|z\|^2$$

and combining yields (5.3).

Finally, suppose (5.3) holds and $x, y \in \mathbb{R}^n$ are arbitrary. Let $x_t = x + t(y - x)$ for all $t \in [0, 1]$ then

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

Applying (5.3) yields that

$$\int_0^1 \int_0^t (y - x)^\top \nabla^2 f(x_\alpha)(y - x) d\alpha dt \le \int_0^1 \int_0^t \beta\|y - x\|^2 d\alpha dt = \beta\|y - x\|^2 \int_0^1 \alpha dt = \frac{\beta}{2}\|y - x\|^2\,.$$

Similarly

$$\int_0^1 \int_0^t (y - x)^\top \nabla^2 f(x_\alpha)(y - x) d\alpha dt \ge \int_0^1 \int_0^t \alpha\|y - x\|^2 d\alpha dt = \alpha\|y - x\|^2 \int_0^1 \alpha dt = \frac{\alpha}{2}\|y - x\|^2$$

and therefore 5.1 holds. □

Specialized to $\|\cdot\| = \|\cdot\|_2$, $\alpha = -\infty$, and $\beta = L$ this lemma implies that the 3 conditions that, (1) $f(x) \le f(y) + \nabla f(y)^\top (x - y) + \frac{\beta}{2}\|x - y\|_2^2$ for al $x, y \in \mathbb{R}^n$, (2) $(\nabla f(x) - \nabla f(y))^\top (x - y) \le \beta\|x - y\|_2^2$ for all $x, y \in \mathbb{R}^n$, and (3) $z^\top \nabla^2 f(x) z \le L\|z\|_2^2$ are equivalent. Further, in the case of $\|\cdot\|_2$ it is known that $\alpha\|z\|_2^2 \le z^\top \nabla^2 f(x) z \le \beta\|z\|_2^2$ is equivalent to assuming that all the eigenvalues of $\nabla^2 f(x)$ are between α and β. Moreover, the analysis of this chapter shows that these conditions are sufficient for gradient descent to provably compute a critical point in $O(L[f(x_0) - f_*]/\epsilon^2)$ gradient evaluations.

Interestingly, these conditions can be less restrictive than assuming f is L-smooth as there are functions which are not L-smooth and nevertheless satisfy these conditions. To see this, consider the following lemma which provides a Hessian based characterization of twice-differentiable L-smooth functions.

Lemma 14. *Twice differentiable $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth if and only if for all $x, z \in \mathbb{R}^n$ we have $|z^\top \nabla^2 f(x) z| \le L\|z\|_2^2$.*

Proof. If f is L-smooth then Lemma 8 implies that

$$\frac{-L}{2}\|x - y\|^2 \le f(x) - f(y) - \nabla f(y)^\top (x - y) \le \frac{L}{2}\|x - y\|^2 \text{ for all } x, y \in \mathbb{R}^n\,.$$

Consequently, Lemma 13 with $\|\cdot\| = \|\cdot\|_2$, $\alpha = -L$, and $\beta = L$ implies that

$$-L\|z\|^2 \le z^\top \nabla^2 f(x) z \le L\|z\|^2 \text{ for all } x, z \in \mathbb{R}^n$$

and $|z^\top \nabla^2 f(x) z| \le L\|z\|_2^2$ for all $x, z \in \mathbb{R}^n$ as desired.

On the other hand, suppose that $|z^\top \nabla^2 f(x) z| \le L\|z\|_2^2$ for all $x, z \in \mathbb{R}^n$. Let $x, y \in \mathbb{R}^n$ be arbitrary and let $x_t \overset{\text{def}}{=} x + t(y - x)$ for all $t \in [0, 1]$. Then,

$$\|\nabla f(y) - \nabla f(x)\|_2 = \left\|\int_0^1 \nabla^2 f(x_t)(y - x) dt\right\|_2 \le \int_0^1 \|\nabla^2 f(x_t)(y - x)\| dt$$

However,

$$\|\nabla^2 f(x_t)(y-x)\|_2 = \sqrt{(y-x)\left[\nabla^2 f(x_t)\right]^2 (y-x)}$$

and since $|z^\top \nabla^2 f(x_t) z| \leq L\|z\|_2^2$ for all t we know that all eigenvalues of $\nabla^2 f(x_t)$ have magnitude at most L and therefore all eigenvalues of $[\nabla^2 f(x_t)]^2$ lie between 0 and L^2. Consequently, $(y-x)\left[\nabla^2 f(x_t)\right]^2 (y-x) \leq L^2\|y-x\|_2^2$ and the result follows. □

These second-order characterizations of smoothness and quadratic upper bound can often be useful when trying to prove that a function has these properties. As we proceed through the course we will continue to provide such characterizations of function properties. It will be a recurring theme that the methods we provide will often continue to perform as desired, even if the assumptions we make to originally motivate them are relaxed.

References

[1] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, Jun 2019.