MS&E 213 / CS 269O
Chapter 3 - Convexity*

By Aaron Sidford (sidford@stanford.edu)

December 29, 2020

In the last chapter we saw that gradient descent can compute an ϵ-critical point at rate independent of dimension given a gradient oracle for a smooth function. We obtained this result by showing that if $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, then it is the case that

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2 \tag{0.1}$$

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

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

and therefore, repeated gradient descent steps eventually find a point with small gradient (since otherwise too much function progress is made). While this works well to compute a critical point, it doesn't yield any global optimality guarantees, i.e. compute an ϵ-optimal point, without further assumptions than smoothness.

Here we show how to add assumptions so that we can prove gradient descent doesn't just compute an ϵ-critical point, but rather it achieves an ϵ-optimal point as well. First, we motivate the assumption we make (namely *(strong) convexity*), then we prove several equivalent definitions, and then we use it to analyze gradient descent.

1 Assumptions for Proving Global Optimality

So how do we turn gradient descent from an algorithm that computes ϵ-critical points to an algorithm that computes ϵ-optimal points? It seems like we need to make another assumption. Below we discuss a few natural assumptions to make.

1.1 Assumption #1 - Lower Bound Hessian

One way that we proved the upper bound, (0.1), in the prevoius chapter was by assuming that f was twice differentiable and that $z^\top \nabla^2 f(x) z \leq L\|z\|_2^2$ for all $x, z \in \mathbb{R}^n$. We saw that twice differentiability of f implied that for all $x, y \in \mathbb{R}^n$ and $x_\alpha = x + \alpha(y - x)$ for $\alpha \in [0, 1]$ the following formula held

*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.

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

Consequently, the assumption $z^\top \nabla^2 f(x) z \leq L\|z\|_2^2$ applied to this equality yielded (0.1). In turn, this allowed us show that a gradient descent step made progress proportional to the norm of the gradient. Unfortunately, this did not let us provable compute ϵ-optimal points as we had no way to relate the progress of a gradient descent step, i.e. the norm of the gradient, to the distance of the given point from optimality.

One natural attempt to fix this, is to assume a lower bound on $z^\top \nabla^2 f(x) z$. In particular, we could assume that for some $\mu \geq 0$ that $z^\top \nabla^2 f(x) z \geq \mu\|z\|_2^2$. Integrating this would imply that

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|x-y\|_2^2$$

for all $x, y \in \mathbb{R}^n$ and thus possibly allow us to relate the norm of the gradient to our current function error.

1.2 Assumption #2 - Lower Bound on Taylor Expansion

Another natural way to achieve global guarantees on the function progress would be to leverage the fact that we have already proven that gradient descent on a smooth function allows us to compute ϵ-critical points, that is points where the norm of the gradient is sufficiently small. More precisely, we could assume that if $\nabla f(x) = 0$ for some $x \in \mathbb{R}^n$ then x is a minimizer of f. To try to make this more quantitative, we note that assuming

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|x-y\|_2^2$$

for all $x, y \in \mathbb{R}^n$ would allow us to conclude that if $\nabla f(x) = 0$ then x is a minimizer of f (and the unique one when $\mu > 0$). Thus we could just try assuming the above formula directly. (Further as we saw in the preceding chapter, in the case that f is twice differentiable this condition is equivalent to assuming that $z^\top \nabla^2 f(x) z \geq \mu\|z\|_2^2$ for all x.)

1.3 Assumption #3 - Seeing the Minima

One more way we could fix our assumptions is to take a closer look at why exactly gradient descent might get stuck, i.e. not make sufficient progress to obtain an ϵ-optimal point. The issue with gradient descent is that for any point if we move in the direction of the minimizer of f it might be the case that the function increase rather than decreases. Thus it might be the case that locally, moving in the direction of the minimum doesn't help. More broadly, the issue is that if we look at the line between two points, i.e. $x_t = x + t(y-x) = t \cdot y + (1-t) \cdot x$ for arbitrary $x, y \in \mathbb{R}^n$ and $t \in [0,1]$, it might be the case that the value of the function along the line, i.e. $f(x_t)$, lies in some places above the line between these function values, i.e. $f(x_t) \geq f(x) + t \cdot (f(y) - f(x)) = (1-t) \cdot f(x) + t \cdot f(y)$. We could fix this by assuming that

$$f(t \cdot y + (1-t) \cdot x) \leq t \cdot f(y) + (1-t) \cdot f(x)$$

for all $x, y \in \mathbb{R}^n$ and $t \in [0,1]$, meaning that the value of the function always lies underneath the line between two points. We could even make this stronger and say that for some $\mu \geq 0$ the function value along the line between two point is at most the value of the quadratic between these function values.

$$f(t \cdot y + (1-t) \cdot x) \leq t \cdot f(y) + (1-t) \cdot f(x) - \frac{\mu}{2} \cdot t \cdot (1-t) \cdot \|x-y\|_2^2\,.$$

2 Convexity

It turns out that each of the possible assumptions discussed in the previous section are equivalent to a notion known as *μ-strong convexity*, or just *convexity* when $\mu = 0$. Here we formally define μ-strong convexity and prove these equivalences.

Definition 1 ((μ-strong) convexity)**.** We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is *μ-strongly convex* for $\mu \geq 0$ if and only if for all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$ we have that

$$f(t \cdot y + (1-t) \cdot x) \leq t \cdot f(y) + (1-t) \cdot f(x) - \frac{\mu}{2} \cdot t \cdot (1-t) \cdot \|x - y\|_2^2 \,. \tag{2.1}$$

We say that f is *convex* if this holds for $\mu = 0$.

In the remainder of this section we show that this is formally equivalent to the assumptions presented in Section 1.

Lemma 2 (Convexity of Differentiable Functions)**.** *A differentiable function f is μ-strongly convex for $\mu \geq 0$ if and only if*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2 \text{ for all } x, y \in \mathbb{R}^n \,. \tag{2.2}$$

Proof. First suppose (2.2) holds. The for all all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$ it is the case that if $x_t = x + t(y - x)$ then

$$f(y) \geq f(x_t) + \nabla f(x_t)^\top (y - x_t) + \frac{\mu}{2}\|y - x_t\|_2^2 = f(x_t) + (1-t) \cdot \nabla f(x_t)^\top (y - x) + \frac{\mu}{2}(1-t)^2\|y - x\|_2^2$$

and

$$f(x) \geq f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{\mu}{2}\|x - x_t\|_2^2 = f(x_t) - t \cdot \nabla f(x_t)^\top (y - x) + \frac{\mu}{2}t^2\|y - x\|_2^2 \,.$$

Since $t(1-t)^2 + t^2(1-t) = t(1-t)$ adding a t multiple of the first equation to the $1-t$ multiple of the second equation then yields (2.1).

On the other hand, suppose that f is μ-strongly convex. Let $x, y \in \mathbb{R}^n$ be arbitrary and let $x_t = x + t(y - x)$ then, by the definition of strong convexity,

$$f(y) \geq \frac{f(x_t) - (1-t) \cdot f(x) + \frac{\mu}{2} \cdot t \cdot (1-t) \cdot \|y - x\|_2^2}{t} = f(x) + \frac{\mu}{2} \cdot (1-t) \cdot \|y - x\|_2^2 + \frac{f(x_t) - f(x)}{t}$$

However, as we have already shown that for $g(t) = f(x_t)$ it is the case that $g'(t) = \nabla f(x_t)^\top (y - x)$ we see that taking the limit of the above as $t \to 0$ yields the desired result. □

Lemma 3 (Convexity of Twice Differentiable Functions)**.** *A twice differentiable function f is μ-strongly convex for $\mu \geq 0$ if and only if*

$$z^\top \nabla^2 f(x) z \geq \mu \|z\|_2^2 \text{ for all } x, z \in \mathbb{R}^n \tag{2.3}$$

Proof. [1]First suppose 2.3 holds. Then for all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$ it is the case that if $x_t = x + t(y - x)$ then

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^\top (y - x) + \int_0^1 \int_0^t (y - x)^\top \nabla^2 f(x_\alpha)(y - x) d\alpha dt \\ &\geq f(x) + \nabla f(x)^\top (y - x) + \int_0^1 \int_0^t \mu \|y - x\|_2^2 d\alpha dt \\ &= f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2 \end{aligned}$$

[1] This is also a corollary of the general equivalence between quadratic upper and lower bounds on functions in the previous chapter and bounds on $\nabla^2 f(x)$.

and the result follows from Lemma 2.

On the other hand suppose that f is μ-strongly convex and let $x, z \in \mathbb{R}^n$ be arbitrary. Define $x_t = x + tz$ for all $t \in \mathbb{R}$ and let $g(t) = f(x_t)$. We have $g'(t) = \nabla f(x_t)^\top z$ and $g''(t) = z^\top \nabla^2 f(x_t) z$. We have that

$$g''(0) = \lim_{t \to 0} \frac{g'(t) - g'(0)}{t} = \lim_{t \to 0} \frac{(\nabla f(x_t) - \nabla f(x))^\top z}{t} = \lim_{t \to 0} \frac{(\nabla f(x_t) - \nabla f(x))^\top (x_t - x)}{t^2}\,.$$

However, by Lemma 2 we know that

$$f(x_t) \geq f(x) + \nabla f(x)^\top (x_t - x) + \frac{\mu}{2} \|x_t - x\|_2^2$$

and

$$f(x) \geq f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{\mu}{2} \|x_t - x\|_2^2\,.$$

Adding these and using that $x_t - x = tz$ we have

$$(\nabla f(x_t) - \nabla f(x))^\top (x_t - x) \geq \mu t^2 \cdot \|z\|_2^2$$

yielding the desired result. □

Since the Hessian of a matrix is always symmetric, the condition that $z^\top \nabla^2 f(x) z \geq \mu \|z\|_2^2$ for all $x, z \in \mathbb{R}^n$ is equivalent to assuming that the smallest eigenvalue of $\nabla^2 f(x)$ is always at least μ. Recall that for symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ the condition that $z^\top \mathbf{A} z \geq 0$ for all $\mathbf{A}$, i.e. all eigenvalues of $\mathbf{A}$ being non-negative, is known as $\mathbf{A}$ be positive semidefinite (PSD). Consequently, Lemma 3 also show that twice differentiable $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its Hessian is PSD at all points.

2.1 Examples and Properties

Convex functions occur in many settings, are closed under a variety of operations, and have a number of interesting properties. Here we provide a number of examples convex functions, operations for creating convex functions, and properties of convex functions.

Lemma 4. *$f_1(x) = x$ is convex, $f_2(x) = x^2$ is 2-strongly convex, $f_p(x) = x^p$ is convex for all even p, and $f_e(x) = \exp(x)$ is convex. Further, for any norm $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ the function $f_{\|\cdot\|}(x) = \|x\|$ is convex.*

Proof. Note that f_1, f_p, and f_e are all twice differentiable with $f_1''(x) = 0$, $f_2''(x) = 2$, $f_p''(x) = p(p-1)x^{p-2}$, and $f_e''(x) = \exp(x)$. Since thee quantities are all non-negative for even p and all x they are all convex by Lemma 3 and f_2 is μ-strongly convex.

Further, for al $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$ we have by triangle inequality and absolute homogeneity of norms that

$$\begin{aligned} f_{\|\cdot\|}(t \cdot x + (1-t) \cdot y) &= \|t \cdot x + (1-t) \cdot y\| \leq \|t \cdot x\| + \|(1-t) \cdot y\| \\ &= |t| \cdot \|x\| + |1-t| \cdot \|y\| = t \cdot f_{\|\cdot\|}(x) + (1-t) \cdot f_{\|\cdot\|}(y)\,. \end{aligned}$$

□

Lemma 5. *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be functions that are μ_f-strongly convex and μ_g-strongly convex respectively for $\mu_f, \mu_g \geq 0$. Further, let $\alpha \in \mathbb{R}_{\geq 0}$, $a \in \mathbb{R}^n$, and $\mathbf{A} \in \mathbb{R}^{n \times m}$ be arbitrary. The following hold*

- *(1) $h : \mathbb{R}^n \to \mathbb{R}$ defined as $h(x) = f(x) + g(x)$ for all $x \in \mathbb{R}$ is $\mu_f + \mu_g$-strongly convex.*
- *(2) $h : \mathbb{R}^n \to \mathbb{R}$ defined as $h(x) = \max\{f(x), g(x)\}$ for all $x \in \mathbb{R}$ is $\min\{\mu_f, \mu_g\}$-strongly convex.*

- *(3) $h : \mathbb{R}^m \to \mathbb{R}$ defined as $h(x) = c \cdot f(\mathbf{A}x - a)$ for all $x \in \mathbb{R}$ is $c \cdot \mu_f \cdot \lambda$-strongly convex where λ is the smallest eigenvalue of $\mathbf{A}^\top \mathbf{A}$.*

Proof. (1) For all $x, y \in \mathbb{R}^n$ and $t \in [0,1]$ we have

$$\begin{aligned} h(t \cdot x + (1-t) \cdot y) &= f(t \cdot x + (1-t) \cdot y) + g(t \cdot x + (1-t) \cdot y) \\ &\leq t \cdot f(x) + (1-t) \cdot f(y) - \frac{\mu_f}{2} \cdot t \cdot (1-t) \|x - y\|_2^2 \\ &\quad + t \cdot g(x) + (1-t) \cdot g(y) - \frac{\mu_g}{2} \cdot t \cdot (1-t) \|x - y\|_2^2 \\ &= t \cdot h(x) + (1-t) \cdot h(y) - \frac{\mu_f + \mu_g}{2} \cdot t \cdot (1-t) \|x - y\|_2^2 \,. \end{aligned}$$

(2) Homework this year (will be added afterwards).

(3) For all $x, y \in \mathbb{R}^n$ and $t \in [0,1]$ we have

$$\begin{aligned} h(t \cdot x + (1-t) \cdot y) &= c \cdot f(t \cdot [\mathbf{A}x - a] + (1-t) \cdot [\mathbf{A}y - a]) \\ &\leq c \cdot \left[t \cdot f(\mathbf{A}x - a) + (1-t) \cdot f(\mathbf{A}y - a) - \frac{\mu_f}{2} \cdot t \cdot (1-t) \|\mathbf{A}(x - y)\|_2^2 \right] \\ &= t \cdot h(x) + (1-t) \cdot h(y) - \frac{c \cdot \mu_f \cdot \lambda}{2} \cdot t \cdot (1-t) \|x - y\|_2^2 \end{aligned}$$

where in the last step we used that if λ is the smallest eigenvalue of $\mathbf{A}^\top \mathbf{A}$ then

$$\|\mathbf{A}(x - y)\|_2^2 = (x - y)^\top \mathbf{A}^\top \mathbf{A}(x - y) \geq \lambda \cdot (x - y)^\top (x - y) = \lambda \|x - y\|_2^2 \,.$$

□

From these lemmas we can show that a larger family of functions are convex. For example, the lemmas immediately implies that $f(x) = a^\top x$ and $g(x) = \sum_{i \in [n]} g(a_i^\top x)$ are convex for any convex g (as these are linear transformations and sums of convex functions).

We conclude by giving two useful properties of convex functions.

Lemma 6. *If $f : \mathbb{R}^n \to \mathbb{R}$ is convex then f is continuous.*

Proof. ad □

The first is known ass

Lemma 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and $\mathcal{D}$ be a distribution over $\mathbb{R}^n$ then*

$$f\left(\mathbb{E}_{x \sim \mathcal{D}} x\right) \leq \mathbb{E}_{x \sim \mathcal{D}} f(x) \,.$$

Proof. Suppose $\mathcal{D}$ is a discrete distribution where for some $x_1, .., x_m \in \mathbb{R}^n$ we have that Pr □

2.2 Smooth (Strongly)-Convex Functions

Interestingly, we can also show that if a function is L-smooth and μ-strongly convex, this is equivalent to a particular quadratic upper and lower bound on f holding at all points.

Lemma 8. *$f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and μ-strongly convex for $\mu \geq 0$ if and only if*

$$\frac{\mu}{2}\|x-y\|_2^2 \leq f(y) - [f(x) + \nabla f(x)^\top (y-x)] \leq \frac{L}{2}\|x-y\|_2^2 \text{ for all } x, y \in \mathbb{R}^n\,. \tag{2.4}$$

Proof. From Lemma 2 and the previous chapter we have already shown that if $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and μ-strongly convex then (2.4) holds. Further, Lemma 2 shows that (2.4) implies that f is μ-strongly convex. Consequently, it just remains to show that (2.4) implies that f is L-smooth.

Suppose that (2.4) holds and let $x, y \in \mathbb{R}^n$ be arbitrary. Let $g : \mathbb{R}^n \to \mathbb{R}$ be defined for all $z \in \mathbb{R}^n$ by $g(z) = f(z) - \left[f(x) + \nabla f(x)^\top (z-x)\right]$.[2] Now, since f is convex, Lemma 4 and Lemma 5 imply that g is convex. Further, since f is differentiable so is g and since $\nabla g(x) = \vec{0}$ we have that

$$\min_{z\in\mathbb{R}^n} g(z) = g(x) = 0\,.$$

However, by assumption of (2.4), we have that for all $z, y \in \mathbb{R}^n$ it is the case that

$$\begin{aligned}
g(z) &\leq \left[f(y) + \nabla f(y)^\top (z-y) + \frac{L}{2}\|z-y\|^2\right] - \left[f(x) + \nabla f(x)^\top (z-x)\right]\\
&= f(y) - \left[f(x) + \nabla f(x)^\top (y-x)\right] + \left(\nabla f(y) - \nabla f(x)\right)^\top (z-y) + \frac{L}{2}\|z-y\|^2\\
&\leq \min_{d\in\mathbb{R}^n} f(y) - \left[f(x) + \nabla f(x)^\top (y-x)\right] + \left(\nabla f(y) - \nabla f(x)\right)^\top d + \frac{L}{2}\|d\|^2\\
&= f(y) - \left[f(x) + \nabla f(x)^\top (y-x)\right] - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2\,.
\end{aligned}$$

However, by assumption

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

and therefore

$$0 \leq \min_{z\in\mathbb{R}^n} g(z) \leq \frac{L}{2}\|y-x\|^2 - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_*^2\,.$$

Rearranging and taking a square root yields the result. □

3 Bound on Distance to Optimum

Now that we have established the definition of convexity, we wish to use it to show that gradient descent converges to optimal points of a convex function. We have already shown that a gradient descent step from an arbitrary point decreases the function by an amount proportional to the squared norm of the gradient. What we want to use convexity to show is that this progress is sufficiently large relative to the current points distance to optimum.

In this section, we show how we can use the assumptions of smoothness and strong convexity to relate various measures of optimality, or difference of a point from optimum. There are three such measures of optimality that we consider. The first is one we have talked about the most, and that is *optimality* or *function error*, for a point $x \in \mathbb{R}^n$ this is just $f(x) - f_*$. The second one is simply the distance of a point to an optimum point. For a point x and minimizer x_* this is just $\|x - x_*\|_2$. We occasionally may refer to this as *residual error*. From here on we also let $X_*(f)$ denote the set of minimizers of f and may occasionally consider $\min_{x_*\in X_*}\|x - x_*\|_2$ which we will call the *distance to the minimizing set*, since this is precisely what it is.

[2] This is a general technique we will see later in course to transform a convex function to have a different minimizer. This is known as the Bregman divergence induced by f and can be used to convert an arbitrary differentiable convex function into a type of "distance measure."

(Note that when f is μ-strongly convex for $\mu > 0$ that if f has a minimizer it is unique and we will show this later.)

The last measure of optimality we will occasionally consider for a point x is simply the norm of the gradient $\|\nabla f(x)\|_2$. As we have seen, this is difficult to relate to the other two measures in general, since for non-convex functions we may have $\|\nabla f(x)\|_2 = 0$ for a point x that is not the minimizer. However, as we have already shown (and will discuss further), with convexity, this cannot happen.

We begin by proving bounds between these three measures using smoothness. We prove the lemma below simply by using that if $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth then for all $x, y \in \mathbb{R}^n$ we have

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

and then look at this inequality for different settings of x and y, i.e. either the point for which we wish to bound optimality or a minimizer.

Lemma 9. *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth then for all $x \in \mathbb{R}^n$ and $x_* \in X_*(f)$ it is the case that*

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x_*) \leq \frac{L}{2} \cdot \|x - x_*\|_2^2 \,. \tag{3.1}$$

Proof. First note that $y = x - \frac{1}{L}\nabla f(x)$ has the property that $f(y) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2$. Consequently, by definition of x_* we have that

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

This yields the first inequality in (3.1) and hows that $\|\nabla f(x_*)\|_2 = 0$ (by considering $x = x_*$)[3].

To obtain the second inequality in (3.1), note that $f(x) \leq f(x_*) + \nabla f(x_*)^\top (x - x_*) + \frac{L}{2}\|x - x_*\|_2^2$. Further, since we have just argued that $\nabla f(x_*) = \vec{0}$, the result follows. □

With lemma established, we now provide analogous bounds for μ-strongly convex functions. The proof is quite similar to the proof of the above lemma. We simply use that if $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex then for all $x, y \in \mathbb{R}^n$ we have

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

and then again, consider different settings of x and y, i.e. either the point for which we wish to bound optimality or a minimizer. This lets us prove the same sort of lemma as above, with the direction of the inequalities reversed.

Lemma 10. *If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and μ-strongly convex for $\mu > 0$ then for $x_* \in X_*(f)$ we have*

$$\frac{1}{2\mu}\|\nabla f(x)\|_2^2 \geq f(x) - f(x_*) \geq \frac{\mu}{2} \cdot \|x - x_*\|_2^2 \,. \tag{3.2}$$

Proof. First we note that since f is differentiable we have $\nabla f(x_*) = 0$ as we showed in the previous chapter. Therefore

$$f(x) \geq f(x_*) + \nabla f(x_*)^\top (x - x_*) + \frac{\mu}{2}\|x - x_*\|_2^2$$

gives the second inequality in (3.2). Next we note that

$$f(x_*) \geq \min_y f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2 = f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

where this equality from the analogous fact in the gradient descent proof that the minimizer of the quadratic is $y = x - \frac{1}{\mu}\nabla f(x)$. □

[3] In the last chapter we proved a slightly stronger variant of this, i.e. that $\|\nabla f(x_*)\|_2 = 0$ for any differentiable f and $x_* \in X_*(f)$.

4 Gradient Descent Strongly Convex Case

We now have everything to analyze gradient descent for strongly convex functions. Note that Lemma 10 lets us lower bound the norm of the gradient at a point by the function error at that point. Consequently, combining this fact with what we showed regarding gradient descent for smooth functions in the previous chapter yields the following theorem.

Theorem 11. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function for $\mu > 0$. Then for $x_0 \in \mathbb{R}^n$ let $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ for all $k \geq 0$. Then we have*

$$f(x_k) - f_* \leq \left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f_*]$$

and consequently we can compute an ϵ-optimal point with $\lceil \frac{L}{\mu} \log(\frac{f(x_0)-f_*}{\epsilon})\rceil$ calls to a gradient oracle.

Proof. As we have seen $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$. However, from Lemma 10 we have that $\|\nabla f(x_k)\|_2^2 \geq 2\mu\,[f(x_k) - f_*]$. Consequently

$$f(x_{k+1}) - f_* \leq f(x_k) - f_* - \frac{\mu}{L}[f(x_k) - f_*] = \left(1 - \frac{\mu}{L}\right)[f(x_k) - f_*]$$

applying this repeatedly uses the bound on $f(x_k) - f_*$ and using that $1 + x \leq e^x$ for all $x \in \mathbb{R}$ and picking $k = \lceil \frac{L}{\mu} \ln(\frac{f(x_0)-f_*}{\epsilon})\rceil$ then yields the bound on the number of gradient oracle calls. □

5 Gradient Descent Non-strongly Convex Case

Here we show how to analyze gradient descent in the case when f is not strongly convex and just convex. To do this we need a new bound on the norm of the gradient. For this we prove the following.

Lemma 12. *If $f \in \mathbb{R}^n$ is differentiable and convex then for all $x \in \mathbb{R}^n$ and $x_* \in X_*(f)$ we have that*

$$f(x) - f_* \leq \|\nabla f(x)\|_2 \cdot \|x - x_*\|_2 .$$

and consequently

$$f(x) - f_* \leq \|\nabla f(x)\|_2 \cdot \inf_{x_* \in X_*} \|x - x_*\|_2 .$$

Proof. By convexity we have that

$$f_* \geq f(x) + \nabla f(x)^\top (x_* - x) \geq f(x) - \|\nabla f(x)\|_2 \cdot \min_{x_* \in X_*} \|x - x_*\|_2$$

where in the last step we used that $\nabla f(x)^\top (y - x) \geq -\left|\nabla f(x)^\top (y-x)\right|$ and $\left|\nabla f(x)^\top (x_* - x)\right| \leq \|\nabla f(x)\|_2 \cdot \|x_* - x\|_2$ by Cauchy Schwarz. □

While this lemma doesn't let us lower bound the norm of the gradient by just the function error and a constant, it does provide a lower bound on the norm of the gradient provided that the distance to the minimizer is not too large. Consequently, if we simply provide an upper bound on $\|x_k - x_*\|_2$ for all k

Using this, we have everything we need to analyze gradient descent.

Theorem 13 (Gradient Descent)**.** *Let f be a L-smooth convex function and starting from some x_0 let $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$. Then if $D_k = \max_{i \in \{0,\ldots,k\}} \min_{x_* \in X_*} \|y - x_i\|_2$ we have that*

$$f(x_{k+1}) - \min_x f(x) \leq \frac{2 \cdot L \cdot D^2}{k+4} .$$

Consequently, for $D_\infty = \sup_{k \geq 0} D_k \leq \sup_{y \,:\, f(y) \leq f(x_0)} \min_{x_* \in X_*} \|y - x_*\|_2$ we see that we can compute an ϵ-optimal point with $\lceil 2 \cdot L \cdot D_\infty^2/\epsilon\rceil$ calls to a gradient oracle.

Proof. Let $\epsilon_k \overset{\text{def}}{=} f(x_k) - f_*$. We have already argued by smoothness that $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$ and therefore $\epsilon_{k+1} \leq \epsilon_k - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$. Further, by Lemma 12 and the definition of D_k we have that $\min_{x_* \in X_*} \|x_k - x_*\|_2 \leq D$ and thus $\epsilon_k \leq \|\nabla f(x_k)\|_2 \cdot D_k$.

Combining yields that

$$\epsilon_{k+1} \leq \epsilon_k - \frac{\epsilon_k^2}{2 \cdot L \cdot D_k^2}$$

and therefore

$$\frac{1}{\epsilon_{k+1}} - \frac{1}{\epsilon_k} \geq \frac{\epsilon_k - \epsilon_{k+1}}{\epsilon_k \epsilon_{k+1}} \geq \frac{\epsilon_k}{2 \cdot L \cdot D_k^2 \cdot \epsilon_{k+1}} \geq \frac{1}{2 \cdot L \cdot D_k^2}\,.$$

Now, by Lemma 9 we know that $f(x_0) - f_* \leq \frac{L}{2}\|x_0 - x_*\|_2^2$ for all $x_* \in X_*(f)$. Consequently, $\epsilon_0 \leq \frac{L}{2}D_0^2$ by Lemma 9 and since D_k increases monotonically we have that

$$\frac{1}{\epsilon_k} \geq \frac{1}{\epsilon_0} + \frac{k}{2 \cdot L \cdot D_k^2} \geq \frac{k+4}{2 \cdot L \cdot D_k^2}\,.$$

Since $f(x_{k+1}) \leq f(x_0)$ for all k and D_k increases monotonically we have the desired result. □

6 Other Progress Measures

Lemma 14. *If f is a L-smooth μ-strongly convex function and $x_* \in X_*(f)$ then if $y = x - \eta\nabla f(x)$ for $\eta \in [0, \frac{1}{L}]$ the following holds*

$$\|y - x_*\|_2^2 \leq \|x - x_*\|_2^2\,.$$

Proof. By definitions we have that

$$\begin{aligned}\|y - x_*\|_2^2 &= \|x - x_* - \eta\nabla f(x_k)\|_2^2 \\ &= \|x - x_*\|_2^2 + 2\eta\nabla f(x)^\top(x_* - x) + \eta^2\|\nabla f(x)\|_2^2\,.\end{aligned}$$

Now $f(x_*) \geq f(x) + \nabla f(x)^\top(x_* - x)$ by convexity and $\|\nabla f(x)\|_2^2 \leq 2L \cdot [f(x) - f(x_*)]$ by smoothness. Consequently

$$\begin{aligned}\|y - x_*\|_2^2 &\leq \|x - x_*\|_2^2 - 2\eta\,[f(x) - f_*] + 2\eta^2 L\,[f(x) - f_*] \\ &= \|x - x_*\|_2^2 - 2\eta(1 - \eta L) \cdot [f(x) - f_*]\,.\end{aligned}$$

Since $f(x) - f_* \geq 0$ by definition of f_* and $2\eta(1 - \eta L) \geq 0$ by assumption on η we have that

$$-2\eta(1 - \eta L) \cdot [f(x) - f_*] \leq 0$$

□

7 Summary and Extensions

In these notes we bounded the performance for gradient descent on smooth convex functions. In all of the settings we considered the algorithm, gradient descent, remained unchanged. It was only the analysis that changed. Thus we can combine the bounds to get a clean statement for the performance of gradient descent. To simplify our statements we use the following helper lemma.

Lemma 15. *If f is a L-smooth μ-strongly convex function for any $\mu \geq 0$ and $x_* \in X_*(f)$ then if $y = x - \eta\nabla f(x)$ for $\eta \in [0, \frac{1}{L}]$ the following holds*

$$\|y - x_*\|_2^2 \leq (1 - \eta\mu)\,\|x - x_*\|_2^2\,.$$

Proof. By definitions we have that

$$\begin{aligned}\|y - x_*\|_2^2 &= \|x - x_* - \eta\nabla f(x_k)\|_2^2 \\ &= \|x - x_*\|_2^2 + 2\eta\nabla f(x)^\top(x_* - x) + \eta^2\|\nabla f(x)\|_2^2 \,.\end{aligned}$$

Now $f(x_*) \geq f(x) + \nabla f(x)^\top(x_* - x) + \frac{\mu}{2}\|x - x_*\|_2^2$ by μ-strong convexity and $\|\nabla f(x)\|_2^2 \leq 2L \cdot [f(x) - f(x_*)]$ by smoothness. Consequently

$$\begin{aligned}\|y - x_*\|_2^2 &\leq \|x - x_*\|_2^2 - 2\eta\left[f(x) - f_* + \frac{\mu}{2}\|x - x_*\|_2^2\right] + 2\eta^2 L\,[f(x) - f_*] \\ &= \|x - x_*\|_2^2 - 2\eta(1 - \eta L)\cdot[f(x) - f_*]\ .\end{aligned}$$

Since $f(x) - f_* \geq 0$ by definition of f_* and $2\eta(1 - \eta L) \geq 0$ by assumption on η we have that

$$-2\eta(1 - \eta L)\cdot[f(x) - f_*] \leq 0$$

□

Using this, we can summarize our results as follows.

Theorem 16. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function for $\mu \geq 0$. Let $x_0 \in \mathbb{R}^n$ and $x_* \in X_*(f)$ be arbitrary and let $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ for all $k \geq 0$. Then*

$$f(x_k) - f_* \leq \min\left\{\left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f_*]\ ,\ \frac{2L\cdot\|x_0 - x_*\|_2^2}{k+4}\right\}\ .$$

Consequently we can compute an ϵ-optimal point with $O(\lceil\min\{\frac{L}{\mu}\log(\frac{f(x_0)-f_*}{\epsilon}), \frac{L\|x_0-x_*\|_2^2}{\epsilon}\}\rceil)$ oracle calls.

Proof. Combine the theorems in this chapter regarding gradient descent and note that the $f(x_k) - f_* \leq \left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f_*]$ still holds when $\mu = 0$ as each step of gradient descent reduces the function value. □

That the distance to a minimizer doesn't increase in a step of gradient descent is a special property of working in ℓ_2. Later in the course we will consider generalizations to arbitrary norms where this does not necessarily hold.

Interestingly though, in this special case of ℓ_2 we can in fact show that every step of gradient descent decreases decreases every progress measure by a multiplicative $1 - \frac{\mu}{L}$ per step.

Lemma 17. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable, L-smooth, μ-strongly convex function for $\mu \geq 0$. For arbitrary $x \in \mathbb{R}^n$, minimizer $x_* \in X_*(f)$, and $y = x - \frac{1}{L}\nabla f(x)$ we have*

$$\begin{aligned}f(y) - f_* &\leq \left(1 - \frac{\mu}{L}\right)[f(x) - f_*]\ , \\ \|y - x_*\|_2^2 &\leq \left(1 - \frac{\mu}{L}\right)\|x - x_*\|_2^2\ ,\ \textit{and} \\ \|\nabla f(y)\|_2^2 &\leq \left(1 - \frac{\mu}{L}\right)\|\nabla f(x)\|_2\,.\end{aligned}$$

Proof. The first two inequalities follow from the calculation in Theorem 11 and Lemma 15 For the third, for all $t \in [0,1]$ let $x_t = t\cdot x + (1-t)\cdot y$. We have that

$$\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x_t)(y - x)dt = \frac{1}{L}\int_0^1 \nabla^2 f(x_t)\nabla f(x)dt\,.$$

Since

$$\nabla f(x) = \int_0^1 \mathbf{I} \nabla f(x) dt$$

this implies that

$$\|\nabla f(y)\|_2 = \left\| \int_0^1 \left[\frac{1}{L} \nabla^2 f(x_t) - \mathbf{I} \right] \nabla f(x) dt \right\|_2 \leq \int_0^1 \| \frac{1}{L} \nabla^2 f(x_t) - \mathbf{I} \|_2 \cdot \|\nabla f(x)\|_2 dt$$

where for a matrix $\mathbf{A}$ we let $\|\mathbf{A}\|_2 \overset{\text{def}}{=} \max_{x \neq 0} \|\mathbf{A}x\|_2 / \|x\|_2$ denote its operator norm. However, since f is L-smooth and μ-strongly convex we know that every eigenvalue of $\nabla^2 f(x_t)$ is between μ and L. Consequently, every eigenvalue of the symmetric matrix $\frac{1}{L} \nabla^2 f(x_t) - \mathbf{I}$ is between $-(1 - \frac{\mu}{L})$ and 0. Since $\|\mathbf{A}\|_2$ for symmetric $\mathbf{A}$ is the largest absolute value of any eigenvalue this implies that $\| \frac{1}{L} \nabla^2 f(x_t) - \mathbf{I} \|_2 \leq 1 - \frac{\mu}{L}$ and the result follows. □