2.633Convergence Guarantees for Smooth and Convex Functionstcb@cnt@theo.2.6 [**?**]@cref@cref[theo][6][2]2.6[1

JOHNS HOPKINS UNIVERSITY
Applied Mathematics and Statistics Department
**Directed Reading Program**

---

# What Makes Gradient Descent Work (Well)?

**Aaron Zoll**

## Contents

# 1 Background

A ubiquitous goal in many areas of study, application, and daily life is to find the best outcome, to optimize whatever situation is given to us. In a mathematical sense, this often resolves to minimizing an object (perhaps subject to constraints). That is, one may wish to solve the following general problem:

**Problem – General Model For Optimization**

*Given function $f\colon \mathbb{R}^n \to \mathbb{R}$ and constraint set $S \subseteq \mathbb{R}^n$, one wishes to solve the following:*

$$p_\star = \inf_{x \in S} f(x) \tag{1.1}$$

In general, the minimum may not exists (i.e., $f(x) = e^{-x}$), which is why we use the *infimum*, or the largest *lower bound*. Note that $0 = \inf_{x\in\mathbb{R}} e^{-x}$, but there is no such $x$ attaining this value. If some point does attain the minimum, we call it a *minimizer*.

**Definition 1.0.1 – Global and Local Minimizers**

*We say $x^\star$ is a global minimizer if*

$$f(x^\star) \leq f(x), \quad \forall x \in S \ .$$

*If $x^\star$ is only a minimizer **nearby**, we say $x^\star$ is a local minimizer. That is, for some $r > 0$,*

$$f(x^\star) \leq f(x), \quad \forall x \in S \cap \{x : \|x - x^\star\| \leq r\} \ .$$

However, even when a (local) minimizer exists, finding the solution $x^\star \in \operatorname{argmin}_{x\in S} f(x)$ often proves difficult. Instead, we often attempt to get *close* to optimal using an **iterative method** (an algorithm). Such methods are defined in their full generality below, but the basic idea is we utilize an *oracle* $\mathscr{O}$ to gather information about a point (i.e., evaluation of a function, its gradient, its hessian, if the point is included in a set, etc.). The information is then saved into an *informational set*. The algorithm's rules $\mathscr{M}$ dictate how to generate the next point, and a stop criterion $\mathscr{S}_\varepsilon$, based on the target accuracy, decides whether the loop repeats or returns (a possibly newly generated point) $\bar{x}$.

**Algorithm 1 – General Iterative Method (modified from [1, (1.1.3)])**

**Input**: Starting point $x^0$ and target accuracy $\varepsilon > 0$.
**Initialize**: $k = 0$, $\mathscr{I}_{-1} = \emptyset$. Here $k$ is the iteration counter and $\mathscr{I}_k$ is the *informational set*.

**Main Loop:**

1. Call oracle $\mathscr{O}$ at point $x^k$.
2. Update informational set $\mathscr{I}_k = \mathscr{I}_{k-1} \cup (x^k, \mathscr{O}(x^k))$
3. Apply rules $\mathscr{M}$ at $x^k$ to generate $x^{k+1}$
4. Check stopping criterion $\mathscr{S}_\varepsilon$. If **yes**, output $\bar{x}$. Otherwise $k \leftarrow k + 1$ and go to step 1.

While there are many ways to measure "close to optimal", a common and perhaps natural concept is to look at function value.

**Definition 1.0.2 – Nearly Optimal (suboptimality gap)**

*Suppose $\bar{x} \in S$ is the output of some* ***iterative method****. Suppose $x^\star$ solve* (1.1)*. We say $\bar{x}$ is an $\varepsilon$-optimal solution if*

$$f(\bar{x}) - f(x^\star) \leq \varepsilon \ . \tag{1.2}$$

Of course, in general, we will not know $f(x^\star)$, so this will not act as a good stopping criterion. We will find ways to mitigate this problem, either with new stopping criterion or ones we can check that imply the above.

## 1.1 Preliminary Concepts

Before we begin, it is essential to understand the basics of linear algebra and calculus to proceed with gradient descent. Many results extend beyond standard Euclidean space. We really just need the structure of a *Hilbert Space*[1], but we can consider $\mathbb{R}^n$ for simplicity.

**1.1.1 Linear Algebra Review** We begin with definition the fundamental place where all the "stuff" lives that we want to work with. In general, gradient descent will handle multiple-dimensional items: vectors. To ensure we're all on the same page, we will start from the beginning and define the necessary terms quite quickly.

In order to do linear algebra, we need numbers. Or at the very least, we need stuff to work with–something that looks familiar (to the real numbers) but is general enough to handle more *complex* settings. As we'll see, the *linearity* of *linear* algebra, needs us to be able to add, scale (multiply by a constant), and distribute the two operations. Thus, we need an ambient space to do these calculations.

**Definition 1.1.1 – Field**

*A set $\mathbb{F}$ equipped with two operations $(+, \cdot)$ is a field if*

1. *Associativity:* $\forall\ a, b, c \in \mathbb{F}$, $\begin{cases} a + (b + c) = (a + b) + c \\ a \cdot (b \cdot c) = (a \cdot b) \cdot c \end{cases}$
2. *Commutativity:* $\forall\ a, b \in \mathbb{F}$, $\begin{cases} a + b = b + a \\ a \cdot b = b \cdot a \end{cases}$
3. *Identities:* $\exists\ 0 \neq 1 \in \mathbb{F}$ *such that* $\forall x \in \mathbb{F}$, $\begin{cases} x + 0 = 0 + x = x \\ x \cdot 1 = 1 \cdot x = x \end{cases}$
4. *Inverses:* $\begin{cases} \forall\ a \in \mathbb{F}, \exists\ (-a) \in \mathbb{F}\ s.t.\ \ a + (-a) = (-a) + a = 0 \\ \forall\ 0 \neq b \in \mathbb{F}, \exists\ (b^{-1}) \in \mathbb{F}\ s.t.\ \ b \cdot (b^{-1}) = (b^{-1}) \cdot b = 1 \end{cases}$
5. *Distributivity:* $\forall\ a, b, c \in F$, $a \cdot (b + c) = a \cdot b + a \cdot c$

[1] Simply put, a Hilbert Space is one where we can do fundamental calculus and linear algebra. Limits exist and act as we would expect. Distances, angles, orthogonality, linear operators all apply here too.

The above definitions may look a bit gnarly, but simply put, a field is a set with two operations that keep elements in the set. This is called *closure*. These operations then come with the standard properties that real numbers possess (and both rational numbers and complex numbers as they are fields too). They can be undone with inverses, and there are elements that have seemingly no effect, called identities. We will find all these properties are especially useful when considered **vector spaces**. We begin with extending the concept to **lists**.

**Definition 1.1.2 – List**

*For a fixed natural number $n \in \mathbb{N}$, define an n-dimensional list over field $\mathbb{F}$ to be*

$$\mathbb{F}^n = \{(x_1, x_2, ..., x_n) : x_j \in \mathbb{F} \quad 1 \leq j \leq 1\}$$

**Remark – Coordinates**

*We call $x_j$ the $j^{th}$-coordinate*

We can then define operations over a list directly from the structure of the field.

**Definition 1.1.3 – Operations in $\mathbb{F}^n$**

- ***Addition**: $(+ : \mathbb{F}^n \times \mathbb{F}^n \to \mathbb{F}^n)$ is defined component-wise:*

$$(x_1, ..., x_n) + (y_1, ..., y_n) := (x_1 + y_1, ..., x_n + y_n)$$

- ***Scalar multiplication**: $(\cdot : \mathbb{F} \times \mathbb{F}^n \to \mathbb{F}^n)$ is component wise too:*

$$\underbrace{\lambda}_{\in \mathbb{F}} \cdot \underbrace{(x_1, ..., x_n)}_{\in \mathbb{F}^n} := (\lambda x_1, ..., \lambda x_n)$$

- ***(additive) Identity**: needs to hold that $0 + \mathbf{x} = \mathbf{x} + 0 = \mathbf{x}$, so we define $\mathbf{0} = (0, 0, ..., 0)$*
- ***(additive) Inverse**: need for $\mathbf{x} + -\mathbf{x} = \mathbf{0}$, so we define $-\mathbf{x} = (-x_1, ..., -x_n)$*

Very naturally, lists let us handle multiple elements of a field simultaneously. If we consider an objects location at a particular time, we may consider the list $(x, y, z, t)$, where $(x, y, z)$ are the three coordinate positions. If this object moves 2 units to the left and 3 units vertically over the period of 4 seconds, we can add $(-2, 0, 3, 4)$ to its previous value. Lists gives us the precisely analog to be able to handle **Vector Spaces**. For the most part, these are going to be analogous, but for mathematical maturity, we should understand that **vector spaces** are the more precise tool we want to utilize.

**Definition 1.1.4 – Vector Space (over a field)**

*A **Vector Space** over a field $\mathbb{F}$ is a set $V$ along with $\begin{cases} + : V \times V \to V \\ \cdot : \mathbb{F} \times V \to V \end{cases}$ such that the following holds*

1. *$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ for all $\mathbf{u}, \mathbf{v} \in V$*

2. $(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$ *for all* $\mathbf{u},\mathbf{v},\mathbf{w}\in V$

3. *There exists* $\mathbf{0}\in V$ *such that* $\mathbf{0}+\mathbf{v}=\mathbf{v}+\mathbf{0}=\mathbf{v}$ *for all* $\mathbf{v}\in V$

4. *For all* $\mathbf{v}\in V$ *there exists a* $(-\mathbf{v})=\mathbf{v}^{-1}\in V$ *(simply notational) such that* $\mathbf{v}+(-\mathbf{v})=(-\mathbf{v})+\mathbf{v}=0$

5. $1\mathbf{v}=\mathbf{v}$ *for all* $\mathbf{v}\in V$ *(1 being the multiplicative identity in* $\mathbb{F}$*)*

6. $\begin{cases} a(\mathbf{u}+\mathbf{v})=a\mathbf{u}+a\mathbf{v} & \forall a\in\mathbb{F},\ \mathbf{u},\mathbf{v}\in V \\ (a+b)\mathbf{v}=a\mathbf{v}+b\mathbf{v} & \forall a,b\in\mathbb{F},\ \mathbf{v}\in V \end{cases}$

**Remark – Notation**

*The condition that these functions exist can also be called "closure under addition (linearity) and closure under scalar multiplication (homogeneity)"*

$$\begin{cases} +: V\times V\to V \\ \cdot\ : \mathbb{F}\times V\to V \end{cases}$$

- *Just for ease, we typically can drop the boldness of vectors based on context. Sometimes we put an arrow over, sometimes a bar (I dislike because a bar looks like conjugation). It really just depends on what notation you choose.*
- *We also drop the* $\cdot$ *for scalar multiplication typically.*
- *Therefore* $v=\mathbf{v}=\bar{v}=\vec{v}$
- *and* $a\cdot v=av$ *going forward.*

Typically in reviewing linear algebra, one would want to study fundamental aspects like linear independence, span, bases, linear maps, eigenvalues, etc. However, we don't need any of those for standard gradient descent! We really just need to following ways to measure how close two vectors are to each other. That is, we want to measure some form of "distance" and "orthogonality" (or angle).

**Definition 1.1.5 – Norms**

*A* ***norm*** *over a vector space* $V$ *is a function* $\|\cdot\|: V\to\mathbb{F}$ *satisfying*

- *(nonnegativity):* $\|x\|\geq 0$*, for all* $x\in V$*.* $\|x\|=0$ *if and only if* $x=0$*.*
- *(positive homogeneity):* $\|\lambda x\|=|\lambda|\cdot\|x\|$ *for any* $x\in V$ *and* $\lambda\in\mathbb{F}$*.*
- *(triangle inequality):* $\|x+y\|\leq\|x\|+\|y\|$ *for all* $x,y\in F$*.*

The above definition gives us a notion of length of a vector. Since vector spaces have addition operations and suitable inverses, we can "subtract" two vectors to produce a new one. Applying the norm to that vector yields a proper distance.

**Definition 1.1.6 – Distance**

*Given $x, y \in V$, the distance between $x$ and $y$, induced by norm $\|\cdot\|$, is defined as*

$$d_{\|\cdot\|}(x, y) = \|y - x\| \ .$$

**Remark – Examples of Norms**

*The following are all norms in $\mathbb{R}^n$:*

- $\|x\|_1 := \sum_{j=1}^{n} |x_j|$
- $\|x\|_2 := \sqrt{(\sum_{j=1}^{n} x_j^2)}$
- $\|x\|_\infty := \max_j |x_j|$
- $\|x\|_p := (\sum_{j=1}^{n} |x_j|^p)^{\frac{1}{p}}, \ \ p \geq 1$

The final object we will need from linear algebra is a way of relating two vectors to each other. In a sense, this next tool is all we need, as we can then suitably *induce* a norm. As we'll see, an **inner product** will be the best tool to associate two (possibly non-distinct) vectors.

**Definition 1.1.7 – Inner Product**

*An **inner product** over a vector space $V$ is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ satisfying*

1. *(positive definiteness): $\langle x, x \rangle \geq 0$ for any $x \in V$. $\langle x, x \rangle = 0$ if and only if $x = 0$.*
2. *(linearity): $\langle ax + by, z \rangle = a \langle x, z \rangle + b \langle y, z \rangle$ for $a, b \in F$ and $x, y, z \in V$.*
3. *(quasi-symmetry): $\langle x, y \rangle = \overline{\langle y, x \rangle}$ for $x, y \in V$*

The last requirement says that flipping the order in the inner product *conjugates* the value. This is really only notable when the underlying field in complex. We will only consider $\mathbb{R}^n$ going forward, in which the inner product (now just a dot product) is now linear in both components and fully symmetric!

**Remark – Inducing a norm**

*Earlier mentioned, the inner product is all we need. In a similar vain to inducing a distance from a norm, one can induce a norm from an inner product:*

$$\|x\| := \sqrt{\langle x, x \rangle}$$

Given this relationship between inner products and (induced) norms, we can derive an incredibly powerful and useful result that lets us bound the size of an inner product (an "association" of two vectors) to the product of their magnitudes.

**Proposition 1.1.8 – Cauchy Schwarz Inequality**

*The inner product and norms of two vectors are related as follows:*

$$|\langle x, y\rangle| \leq \|x\|\|y\| \tag{1.3}$$

*Proof.* We will consider the case where $\mathbb{F} = \mathbb{R}$ and $V = \mathbb{R}^n$. However, the result holds in generality. Consider $x, y \in \mathbb{R}^n$ and $t \in \mathbb{R}$. Then,

$$\begin{aligned} 0 &\leq \langle x + ty, x + ty\rangle && \text{(by positive definiteness)} \\ &= \|x\|^2 + 2t\langle x, y\rangle + t^2\|y\|^2 \ . && \text{(linearity and definition of the induced norm)} \end{aligned}$$

Note the above is a quadratic in $t$. Here, $x$, $y$, and thus their norms are just fixed scalars. Since the quadratic is facing upwards ($t^2 \geq 0$) and is nonnegative for all $t$, it must hold that the discriminant (the $b^2 - 4ac$ part in the square root of the quadratic formula) is *nonpositive*. Therefore:

$$\begin{aligned} (2\langle x, y\rangle)^2 - 4\|x\|^2\|y\|^2 &\leq 0 \\ 4(\langle x, y\rangle)^2 &\leq 4\|x\|^2\|y\|^2 \ . \end{aligned}$$

Dividing by 4 and taking the square root of both sides yields the desired result. □

We can then immediately define the notion of an *angle* between two vectors. The above inequality gives us that

$$\frac{|\langle x, y\rangle|}{\|x\|\|y\|} \leq 1, \quad \forall x, y \in V \ .$$

In other words,

$$-1 \leq \frac{\langle x, y\rangle}{\|x\|\|y\|} \leq 1 \ .$$

Therefore, we can define the angle as follows:

**Definition 1.1.9 – Angle Between Vectors**

*For any $x, y \in V$, define the angle $\theta_{x,y}$ as*

$$\theta_{x,y} = \cos^{-1}\left(\frac{\langle x, y\rangle}{\|x\|\|y\|}\right) \ .$$

We use cosine in the above definition for two reasons. The first is that this result naturally extends from law of cosines. Secondly, we get a natural association for two vectors forming a "right angle". When the inner product of two (non-zero) vectors is zero, then the angle between them is precisely $90^o$ (or $\pi/2$ in radians). Lastly, we introduce the notion of orthogonality. In a sense, this means that two vectors are "as independent" as they can get. However, for our application, orthogonality will tie to the idea of *descent*.

**Definition 1.1.10 – Orthogonality**

*Two vectors $x, y \in V$ are said to be* ***orthogonal*** *if*

$$\langle x, y\rangle = 0$$

**1.1.2 Calculus Review** With the basic notions of linear algebra out of the way, we can relax ourselves to consider a more familiar setting: $\mathbb{R}^n$. However, I would recommend keeping in mind that almost all of the following will extend to any setting that allows a suitable inner product (even infinite dimensional spaces!). With our restriction to $\mathbb{R}^n$, we will also consider the standard *Euclidean inner product*:

$$\langle x, y \rangle = x_1 y_1 + \cdots + x_n y_n \ . \tag{1.4}$$

Note that the induced norm is then

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + \cdots + x_n^2} \ . \tag{1.5}$$

We can now move onto defining the star of the show. The **gradient**. Note that the following definition is not entirely useful for us as is. It doesn't quite explain why gradient leads to descent, nor does it imply anything about "pointing in the steepest direction." We will observe the latter two claims more in the following sections, but without further ado:

**Definition 1.1.11 – Multidimension Derivatives**

*The **partial derivative** of a function $f : \mathbb{R}^n \to \mathbb{R}$, at the point $x = (x_1, \ldots, x_n)$ is defined as*

$$\frac{\partial f(x)}{\partial x_j} := \lim_{h \to 0} \frac{f(x_1, \ldots, x_j + h, \ldots, x_n) - f(x_1, \ldots, x_n)}{h} \ . \tag{1.6}$$

*If each partial derivative exists for all $j = 1, \ldots, n$, the the function is said to be **differentiable** at $x$. If all the partials exist for all $x \in \mathbb{R}^n$, the function itself is said to be differentiable.*

---

Given differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the gradient is itself a function $\nabla f(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ defined as

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

Note that the partial derivative above is defined in the following way: how much does the function value change if we move a small amount in the direction of a single coordinate, relative to how much we moved? In a sense, this is very similar to the standard "rise over run" calculation of a slope, but just pointing in a particular direction. This direction is entirely arbitrary, but useful because of how we constructed our function (if you know further linear algebra, this is precisely by focusing on a particular basis element). However, we could have perturbed in any direction, and found the relative change in function value.

**Definition 1.1.12 – Directional Derivative**

*Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and unit vector $v \in \mathbb{R}^n$ (that is, $\|v\| = 1$), the directional derivative at $x \in \mathbb{R}^n$ is defined as*

$$f'(x; v) := \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h} \tag{1.7}$$

While the definition of a gradient might not seem immediately useful, we do get for free a multidimensional Taylor's theorem. In a similar way to constructing the linear model of a one-dimensional function to pass through the function at a point and have the same slope, we can approximate a differentiable function at a point analogously.

**Definition 1.1.13 – Linear Approximation**

*The **linear approximation** of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ is*

$$\ell_{f,x}(y) := f(x) + \langle \nabla f(x), y - x \rangle \ .$$

Note that $\ell_{f,x}(x) = f(x)$ and $\nabla \ell_{f,x}(x) = \nabla f(x)$. It can further be shown that the error is of order $o(t)$. That is, for small distances away from $x$, the error is shrinking faster than linear.

**Proposition 1.1.14 – Taylor's Theorem (first order)**

*Given differentiable function $f$ and its linear approximation, the error is **first order**.*

$$f(y) = \ell_{f,x}(y) + o(\|y - x\|)$$

*where $o(h)$ implies that $\lim_{h \to 0^+} o(h)/h = 0$ and $o(0) = 0$.*

This proposition leads to a rather nice result about gradients. While the definition of a gradient may seem like a useful tool for collecting all the partial derivatives, the following corollary highlights the relationship between gradients and *directional derivatives*. That is, the gradient is the precise vector that **represents** the directional derivatives.

**Corollary 1.1.15 – Representation of Directional Derivatives**

*Given differentiable $f : \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$ and unit vector $v \in \mathbb{R}^n$ the following relationship holds:*

$$f'(x; v) = \langle \nabla f(x), v \rangle \tag{1.8}$$

*Proof.* Simply unpacking definitions,

$$\begin{aligned} f'(x; v) &= \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h} \\ &= \lim_{h \to 0} \frac{f(x) + \langle \nabla f(x), (x + hv) - x \rangle + o(\|(x + hv) - x\|) - f(x)}{h} \\ &= \lim_{h \to 0} \frac{\langle \nabla f(x), hv \rangle + o(\|hv\|)}{h} \\ &= \lim_{h \to 0} \langle \nabla f(x), v \rangle + \frac{o(h)}{h} = \langle \nabla f(x), v \rangle \end{aligned}$$

□

We conclude this section with a few more definitions that will be useful for our analysis. First, we define a neighborhood around a point. Often, as a consequence of the calculus involved and (lack of) structure of the function we aim to minimize, conditions are only considered locally.

**Definition 1.1.16 – Ball Neighborhood $B_{\|\cdot\|}(x, r)$**

*Given a point $x \in \mathbb{R}^n$, a radius $r \geq 0$ and a norm $\|\cdot\|$, we define the ball around that point:*

$$B_{\|\cdot\|}(x, r) := \{y \in \mathbb{R}^n : \|y - x\| < r\} \ .$$

When we consider the standard Euclidean norm $\|\cdot\|_2$, we can drop the subscript. We can consider the slightly larger set $\hat{B}_{\|\cdot\|}(x, r) := \{y \in \mathbb{R}^n : \|y - x\| \leq r\}$ as well. Below are visualizations of different normed balls in $\mathbb{R}^2$. Again, we will really only consider the 2-norm, as the standard

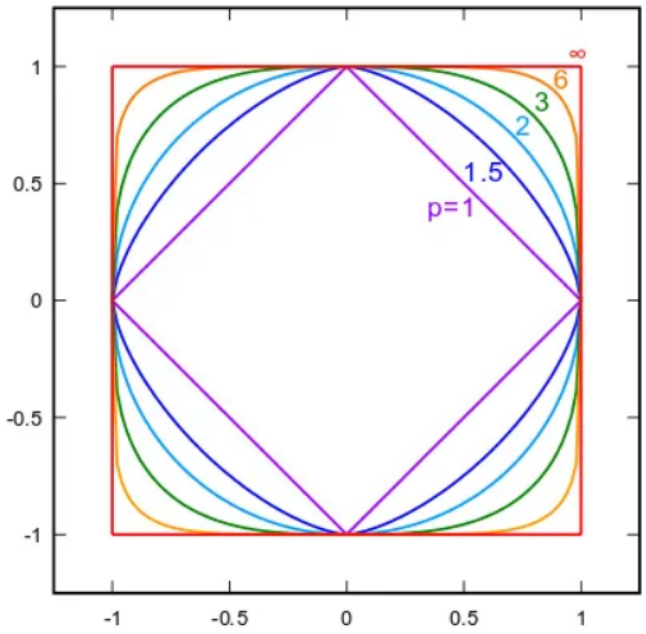


Figure 1: p-norms in $\mathbb{R}^2$

inner product induces it, and we will drop the subscript to mean $\|\cdot\| = \|\cdot\|_2$ unless otherwise noted. However, it is worth noting that other settings do utilize different norms.

As we run our iterative method, we will naturally obtain a sequence of values. If we utilize the optimality condition outlined in 1.0.2, it makes sense to look at the sequence $f(x^k) - f(x^\star)$ where $x^k$ is the $k^{th}$ iterate of some method.

**Definition 1.1.17 – Relaxation Sequence**

*If a sequence $\{a_k\}_{k=0}^{\infty}$ satisfies*

$$a_{k+1} \leq a_k \ ,$$

*we call the sequence a **relaxation sequence**.*

Note that since $x^\star$ is a minimizer, it should hold that $f(x^k) - f(x^\star) \geq 0$. That is, the sequence $\{f(x^k) - f(x^\star)\}_{k=0}^{\infty}$ is bounded below. If it is also a relaxation sequence, then the following lemma states the sequence must converge! If zero is the *best lower bound*. That is, if for any $\delta > 0$, there is some $N$ where $f(x^N) - f(x^\star) < \delta$, then $\{f(x^k) - f(x^\star)\}_{k=0}^{\infty}$ converges to 0 and the iterative method

converges to a minimizer!

**Proposition 1.1.18 – Monotone Convergence Theorem**

*Let $\{a_k\}_{k=0}^{\infty}$ be a relaxation sequence of real numbers with $\inf_k a_k = 0$ then the sequence converges to zero and we write*

$$a_k \to 0 \ .$$

To conclude this section, we must define notions of convergence. We can suppress the rigorous definitions for an real analysis course, but in optimization, it is useful to characterize how quickly a sequence converges.

**Definition 1.1.19 – Notions of Convergence**

*Suppose $a_k \to 0$. If there exists constant $0 < \tau < 1$ and $C > 0$ such that*

$$a_k \leq C\tau^k \ ,$$

*we say $\{a_k\}_{k=0}^{\infty}$ convergence **linearly**.*

---

If instead there are constants $C > 0$ and $q > 0$ such that

$$a_k \leq \frac{C}{k^q} \ ,$$

we say the sequence converges **sublinearly** with rate $q$.

Some may note that this would imply exponential decay, and it does! However, we say the convergence is linear because $a_{k+1} = \tau a_k$. For sublinear rates, we do not guarantee as good of convergence, but unfortunately, most iterative schemes we see will be sublinear.

**1.1.3 (Structured) Classes of Functions** The above tools will allow us to analyze methods like gradient descent, but so far, we have no guarantees that any iterative method will converge. Certain classes of functions behave better than others; gradient descent converges faster given structure. While optimizers aim to find the best solution to the most accurate-to-reality model, we often must sacrifice the precision of the model for the convergence of a solution. Aspects like continuity and differentiability help, but we can do better.

**Definition 1.1.20 – Convex Function**

*A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **convex** if for any $t \in [0, 1]$*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \quad \forall x, y \in \mathbb{R}^n \ .$$

The above can be written slightly differently, one can say equivalently, that

$$f(y + t(y - x)) \leq f(y) + t(f(y) - f(x)) \ .$$

That is, if one draws a line segment from $f(y)$ to $f(x)$ it will always upper bound the function evaluated at any point between $y$ and $x$. For an interactive 3-D plot of a convex function, click here.

In general, convex functions are a great place to start. While they need not be differentiable, they must be continuous. Furthermore, minima are unique, and when a local minimizer exists, it

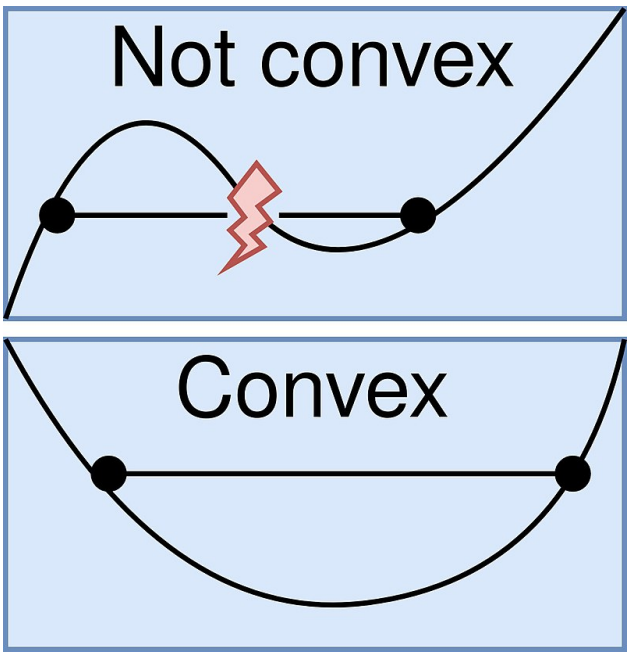


Figure 2: Convex Functions

must also be a global one. One way to think about convex functions is that they "curve upwards." However, there is a more rigorous characteristic of convex functions.

**Proposition 1.1.21 – Lower Linearization**

*Suppose $f : \mathbb{R}^n \to R$ is convex and differentiable. Then for any $x \in \mathbb{R}^n$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle , \quad \forall y \in \mathbb{R}^n.$$

*Proof.* Suppose $f$ is convex. Then, for all $t \in (0, 1)$, $tf(y) + (1-t)f(x) \geq f(ty + (1-t)x)$. In other words,

$$f(y) \geq \frac{f(x + t(y - x)) - f(x)}{t} + f(x) .$$

Letting $t \to 0$ and using Corollary 1.1.15 gives the desired characterization. □

What the above proposition states is that the local linear approximation outlined in Proposition 1.1.14 is a *lower bound*. That is, for any convex function $f$, the linear approximation $\ell_{f,x}$ gives a **global lower bound** on the function itself. That means that any minima must lower *above* that line. As we will see, many iterative methods utilize this idea of successively maximizing lower bounds and minimizing upper bounds to squeeze in on the optimal value.

While convex function give a great place to start, often they do not proving rich enough structure for gradient descent to work effectively. The following class of functions really sets the stone for gradient descent to work well.

**Definition 1.1.22 – $L$-smooth Functions**

*A function $f : \mathbb{R}^n \to \mathbb{R}$ is **$L$-smooth** if there exists an $L < \infty$ such that*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n \ .$$

There is another way to state the above definition. Utilizing the class of *Lipschitz* functions:

$$|f(x) - f(y)| \leq L\|x - y\|, \quad x, y \in \mathbb{R}^n.$$

One can say that $f$ is $L$-smooth if and only if $\nabla f$ is $L$-Lipschitz. In essence, Lipschitz functions are ones whose gradients are bound in size. That is, the function isn't doesn't change too rapidly. For $L$-smooth functions, one can say that its gradients don't change too rapidly. In short:

$f$ is $L-$smooth $\implies$ we can trust nearby gradients.

This will be the key to utilizing gradient descent! If the gradients change at a controlled pace, then when we "move in the direction of a gradient" we can ensure that this movement is in the "right direction" for a enough time to take that step. More to see later,

Notably $L$-smooth functions possess many properties. One of the more useful characterizations is that they can be bounded above *globally* by a quadratic. What globally means in this context is that the curvature, or the coefficient attached to the analogous $x^2$ term, is precisely related to $L$.

**Lemma 1.1.23 – Descent Lemma**

*Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is $L$-smooth and convex. Then,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n \ .$$

That is, once we know $f$ has $L$-Lipschitz continuous gradient, and it is convex, we can immediately know that any quadratic centers at the point $(x, f(x))$ with the exact same gradient at $x$, and curvature $L$, will upper bound the entire function. We will utilize this proposition in the next chapter to motive **gradient descent**. To play around with this quadratic upper bound, click here.

*Proof.* Consider the function $\phi_x(t) = f(x + t(y - x))$ for $t \in [0, 1]$. Note that $\phi_x(0) = f(x)$ and $\phi_x(1) = f(y)$. Furthermore, the chain rule gives us that $\phi_x'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle$. Therefore, by fundamental theorem of calculus,

$$f(y) - f(x) = \phi(1) - \phi(0) = \int_0^1 \phi'(t)dt = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle \, dt$$

Subtracting $\langle \nabla f(x), y - x \rangle$ from both sides, we get

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \, dt \\ &\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \|y - x\| dt \\ &\leq \int_0^1 L \|(x + t(y - x)) - x\| \|y - x\| dt \\ &= \int_0^1 Lt \|y - x\|^2 dt \\ &= \frac{L}{2} \|y - x\|^2 \end{aligned}$$

Where the first inequality comes from Cauchy Schwarz Eq. (1.3). The second comes from the Lipschitz continuous gradient. And the rest is simple rearrangement. □

**Remark**

*While $L$-smooth functions are always differentiable, they may not admit a second derivative. However, if a function is $L$-smooth and twice differentiable, then the following holds:*

$$\|f''(x)\| \leq L, \quad \forall x \in \mathbb{R}^n \ .$$

Note that this $L$ isn't unique! If $f$ is $L$-smooth, then it is $\hat{L}$-smooth for any $\hat{L} \geq L$.

Conversely, we define strongly convex functions as those possessing a uniform quadratic lower bound.

**Definition 1.1.24 – Strongly Convex Functions**

*A function $f : \mathbb{R}^n \to \mathbb{R}$ is $\mu$-strongly if there exists a $\mu \geq 0$ such that*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n \ .$$

There are plenty of other characterizations of strong convexity, but the following helps visualize the lower bound on its curvature, while sharing a dual perspective to $L$-smoothness.

**Remark**

*If a function is $\mu$-strongly convex and twice differentiable, then the following holds:*

$$\|f''(x)\| \geq \mu, \quad \forall x \in \mathbb{R}^n \ .$$

Note that this $\mu$ isn't unique! If $f$ is $\mu$ -strongly convex, then it is $\hat{\mu}$-smooth for any $\hat{\mu} \leq \mu$.

Below illustrates smooth and strong convexity. The function $f(x)$ (in orange) can be lower bounded by any linear approximation (in blue). This is a property of convexity. However, it can further be lower bounded by a quadratic approximation (in red), as well as upper bounded by a quadratic (in green). If the center point, that is where all the curves intersect, were shifted, we would still get lower/upper bounds with the same curvature! You can play around with strongly convex functions here. Note that unlike smooth functions, strongly convex ones need *not* be differentiable!

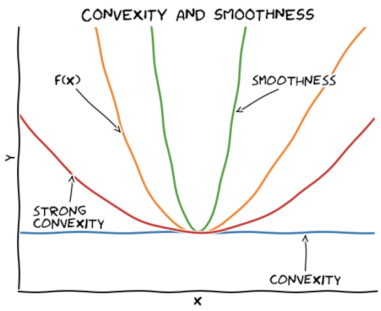


Figure 3: Smoothness and Strong Convexity

## 1.2 Why *Gradient* Descent

Let's begin with mentioning two important properties of the gradient. The first will require some new work (and a quite technical definition), but the second we have already seen before.

**Definition 1.2.1 – Sublevel Sets and their Tangents**

*We denote the* ***sublevel set*** *for given function $f : \mathbb{R}^n \to \mathbb{R}$ at $\alpha \in \mathbb{R}$ by*

$$\mathscr{L}_f(\alpha) := \{x \in \mathbb{R}^n : f(x) \leq \alpha\}$$

Consider the set of directions **tangent** to $\mathscr{L}_f(f(x))$ at $x$:

$$S_f(x) := \left\{ s \in \mathbb{R}^n : s = \lim_{k\to\infty} \frac{y_k - x}{\|y_k - x\|}, \text{ for some } y_k \to x \text{ with } f(y_k) = f(x)\ \forall k \right\}$$

The above definition is fairly involved, but the general idea is that $\mathscr{L}_f(\alpha)$ denotes all the points in $\mathbb{R}^n$ that get *mapped* to a value at most $\alpha$. Consider the pair $(x, f(x))$, we can consider all the directions that do not change the function value.

Considering a mountain range, the sublevel set are all the places you could stand that would be below a specific altitude, and the tangents are direction you could walk without changing your elevation!

**Lemma 1.2.2 – Gradients Point away from Level Sets**

*If $s \in S_f(x)$, then $\langle \nabla f(x), s\rangle = 0$. That is,*

$$\nabla f(x) \text{ is } \textbf{orthogonal} \text{ to the level sets at } x$$

*Proof.* Since $f(y_k) = f(x)$, we have

$$f(y_k) = f(x) + \langle \nabla f(x), y_k - x\rangle + o(\|y_k - x\|) = f(x)\ .$$

Therefore, $\langle \nabla f(x), y_k - x\rangle + o(\|y_k - x\|) = 0$. Dividing by $\|y_k - x\|$ and taking limits yields the result. □

The above lemma states the gradient points away from the directions that *don't* change the function value. In particular, they point *away as much as they can.*

**Remark**

*Thus, if we want to change function value (either increase or decrease), we should move in the direction of the gradient (or perhaps its negative).*

The following property will further justify this claim. Recall from Corollary 1.1.15 that the gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ "represents" all the directional derivatives. That is, given a gradient at a point, $\nabla f(x)$, one can evaluate the "rate of change" in any direction by simply evaluating an inner product:

$$f'(x; v) = \langle \nabla f(x), v \rangle \ .$$

If we apply Cauchy Schwarz inequality 1.1.8, we see that

$$-\|\nabla f(x)\|\|v\| \leq \langle \nabla f(x), v \rangle \leq \|\nabla f(x)\|\|v\| \ .$$

Since $\|v\|$ is a unit vector, we can see that the most we can decrease (or increase) is $\|\nabla f(x)\|$, and this value can be achieved by choosing

$$v = -\frac{\nabla f(x)}{\|\nabla f(x)\|} \ .$$

**Remark – Steepest Descent**

*Moving in the direction of the negative gradient, $-\nabla f(x)$, gives the fastest local decrease of the function $f(\cdot)$ at the point $x$!*

We can now conclude this section with perhaps the **most fundamental** fact in optimization theory.

**Theorem 1.1 – First-Order (local) Optimality Conditions**

*Let $x^\star$ be a local minimum to a differentiable function $f(\cdot)$. Then*

$$\nabla f(x^\star) = 0 \ .$$

*Proof.* Since $x^\star$ is a local minimizer of $f(\cdot)$, there exists an $r > 0$ where $f(y) \geq f(x^\star)$ for all $y \in B(x^\star; r)$. Since $f$ is differentiable, by Proposition 1.1.14,

$$f(y) = f(x^\star) + \langle \nabla f(x^\star), y - x^\star \rangle + o(\|y - x^\star\|) \geq f(x^\star), \quad \forall y \in B(x^\star; r).$$

Considering the inequality, subtracting $f(x^\star)$ and again dividing by $\|y - x^\star\|$ and taking limits as needed, we see that

$$0 \leq \langle \nabla f(x^\star), v \rangle = f'(x; v), \quad \forall v \in \mathbb{R}^n! \tag{1.9}$$

Considering $v = -\nabla f(x^*)/\|\nabla f(x^*)$ implies that

$$\|\nabla f(x^*)\| \leq 0 \ .$$

By the nonnegativity of the norm, it must hold that

$$\|\nabla f(x^*)\| = 0 \Longrightarrow \nabla f(x^\star) = 0 \ .$$

□

Note that (1.9) implies that all directional derivatives are nonnegative. That is, any direction that one moves from the local minimizer must increase (or remain constant) the function value.

Further note that the converse may not hold in general. The gradient could be zero at $x$, but $f(x)$ may not be a minimum. Consider the function

$$f(x,y) = x^2 - y^2, \quad \nabla f(0,0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

However, if $f$ is also assumed to be convex, then the converse holds! Convexity ensures that

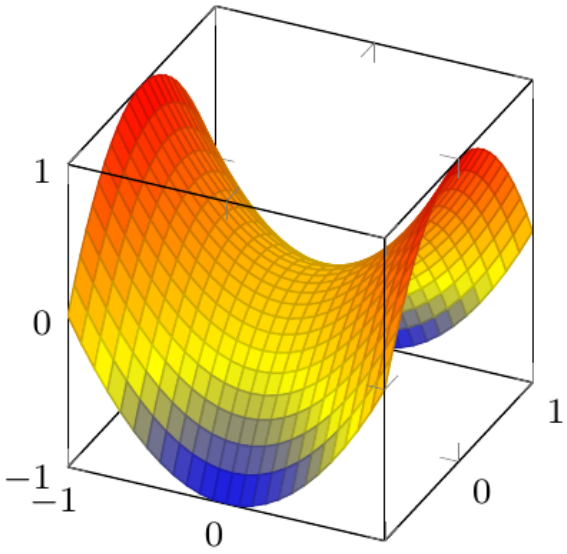


Figure 4: Graph of $f(x,y) = x^2 - y^2$

$$f(y) \geq f(x^\star) + \langle \nabla f(x^\star), y - x^\star \rangle \,, \quad \forall y \in \mathbb{R}^n \,,$$

and when $\nabla f(x^\star) = 0$, we get

$$f(y) \geq f(x^\star), \quad \forall y \in \mathbb{R}^n \,.$$

Since this holds for all $y$, not just in a small neighborhood around $x^\star$, this result implies that local minimizers for convex function are in fact global. Thus, we get our second big theorem.

**Theorem 1.2 – First-Order Optimality Conditions**

*Let $f(\cdot)$ be a* ***convex*** *differentiable function. Then the following are equivalent.*

1. *$x^\star$ is a local minimum*
2. *$x^\star$ is a global minimum*
3. *$\nabla f(x^\star) = 0$*

Therefore, we can describe the second way to check for optimality, one we can actually perform in an algorithm as it assumes no knowledge of $x^\star$.

**Definition 1.2.3 – Nearly Optimal (gradient norm)**

*Let $\bar{x}$ be the output of some iterative method. Given target accuracy $\varepsilon > 0$, we say $\bar{x}$ is $\varepsilon$-optimal if*

$$\|\nabla f(\bar{x})\| \leq \varepsilon$$

We now have all the tools and knowledge to introduce our **Gradient Descent** method. The concept is simple, but the theory goes very deep. There are many ways we can tweak this method to ensure better convergence rates. Further assumptions on $f(\cdot)$ can also speed up there algorithm. However, in its base form, the algorithm works as follows:

**Algorithm 2 – Gradient Descent**

**Input**: $x^0 \in \mathbb{R}^n$, target accuracy $\varepsilon > 0$ and stepsizes $h_k$

**For** $k = 0, 1, 2, \ldots$

1. $x^{k+1} \leftarrow x^k - h_k \nabla f(x^k)$

One may ask: when does Algorithm 2 terminate? How many iterations does it take? What structure and geometry of $f(\cdot)$ is desired? How must we choose stepsizes $h_k$? Can we guarantee anything about the suboptimality gap (how close is $f(\bar{x})$ to $f(x^\star)$)? What about size of the gradients $\|\nabla f(\bar{x})\|$? All of this will be discussed in the next section.

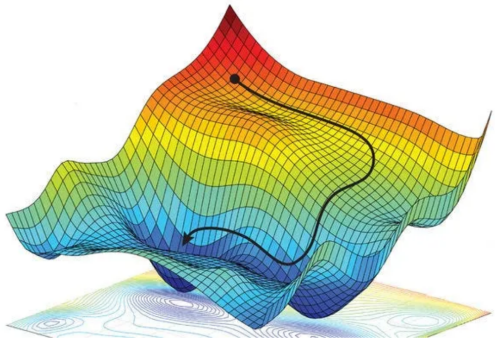

Figure 5: Medium

## 2 Convergence Guarantees for Gradient Descent

This section will primarily focus on the convergence results for applications of Algorithm 2. Note that this algorithm is essentially only one line, with barely any freedom to alter its procedure. We really only have leverage over the initial point, $x^0 \in \mathbb{R}^n$, the target accuracy $\varepsilon > 0$, and the stepsizes $h_k$. This information alone will typically not guarantee much. We won't know how many iterations to run to ensure small suboptimality gap or small gradient norms. We may not even guarantee convergence at all! However, with *known structure* of the function we are aiming to minimize, we can readily solve these issues.

## 2.1 Smooth Functions

For ease of notation, we will define the following classes of function. Later on, we will further restrict to **convex** functions with the same properties.

**Definition 2.1.1 – The Class of $L$-Smooth Functions**

*Let $C_L(\mathbb{R}^n)$ denote the class of differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ with Lipschitz continuous gradient*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n \ .$$

We first note a pleasant property of smooth functions. While there exist methods to generalize the notion of a gradient to nondifferntiable functions, $L$-smooth functions are always differentiable. Therefore, it makes sense to apply gradient descent to such functions. Unpacking the above definition, it says that for $x$ and $y$ close together, say $y = x + hv$ for unit vector $v \in \mathbb{R}^n$ and $h$ small. Then $\|\nabla f(x + hv) - \nabla f(x)\| \leq hL$. When considering $v = -\nabla f(x)$, a gradient step, we note that

$$\|\nabla f(x^{k+1}) - \nabla f(x^k)\| \leq hL \ .$$

**That is, the gradient (or direction we are moving) where we end up each iteration is not too different from where we started**. In other words, we can "trust our gradients" each step. If $L$ is large, we just need to make $h$ smaller. In fact, we can *optimize* this value of $h$ to guarantee the most descent each step!

There are many possible values of $h_k$ one can choose. However, let's start by considering three possible strategies.

**Remark – Possible Stepsize Rules**

1. *The sequence $h_k$ is chosen in advance. For example,*

$$h_k = h > 0, \qquad \text{(constant step)}$$
$$h_k = \frac{h}{\sqrt{k+1}}, \qquad \text{(decaying step)}$$

2. *Full relaxation (find the greastest functional decrease):*

$$h_k = \operatorname*{argmin}_{h \geq 0} f(x^k) - h\nabla f(x^k)$$

3. *The Armijo Rule: find $x^{k+1} = x^k - h\nabla f(x^k)$ such that*

$$\alpha \left\langle \nabla f(x^k), x^k - x^{k+1} \right\rangle \leq f(x^k) - f(x^{k+1}) \leq \beta \left\langle \nabla f(x^k), x^k - x^{k+1} \right\rangle \tag{2.1}$$

*where $0 < \alpha < \beta < 1$ are fixed*

Comparing these strategies, we note that the first is by far the simplest. It is used in many algorithms, especially ones where minimizer are guaranteed to be global if they exists (i.e., convex optimization). The second is completely theoretical as even in one-dimensional settings, finding the minimizer $h^\star$ may be infeasible. The third is complex, but used in the many practical algorithms. The geometric interpretation is explained below.

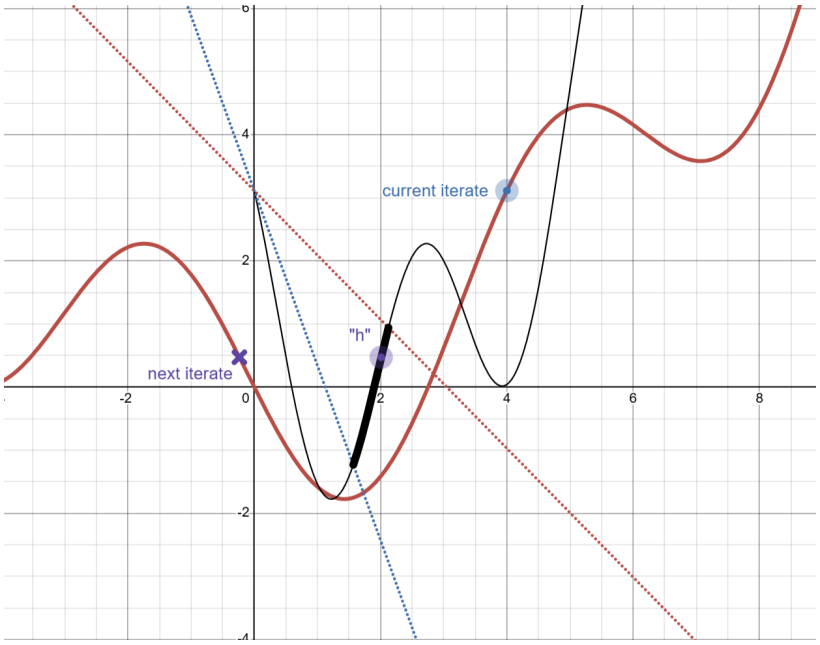


Figure 6: Geometric Interpretation of Armijo Rule

Consider the iterates $x^k$ and $x^{k+1} = x^k - h_k \nabla f(x^k)$. Now consider the function value at $f(x^{k+1})$. This is a one-dimensional function of the stepsize $h_k$. Define

$$\phi(h) := f(x^{k+1}) = f(x^k - h\nabla f(x^k)), \quad h \geq 0 \ . \tag{2.2}$$

Recall the conditions for the Armijo rule, as well as how we construct $x^{k+1}$ result in $h$ being a suitable stepsize if

$$\phi_2(h) \leq \phi(h) \leq \phi_1(h), \quad \phi_1(h) = f(x^k) - \alpha h \|\nabla f(x^k)\|^2, \quad \phi_2(h) = f(x^k) - \beta h \|\nabla f(x^k)\|^2 \ . \tag{2.3}$$

**Exercise 1**

*Check that the above condition implies the Armijo rule* (2.1). *That is given* $0 < \alpha < \beta < 1$, *for any* $h$ *satisfying*

$$\phi_2(h) \leq \phi(h) \leq \phi_1(h) \ ,$$

*then*

$$\alpha \left\langle \nabla f(x^k), x^k - x^{k+1} \right\rangle \leq f(x^k) - f(x^{k+1}) \leq \beta \left\langle \nabla f(x^k), x^k - x^{k+1} \right\rangle$$

Geometrically, the Armijo rule can be seen as finding an $h$ that lies in the black region. The thick red line is the objective function $f(x)$, and the thin black function is $\phi(x)$. The upper and bounding linear functions $\phi_1$ and $\phi_2$ are the dotted red and blue respectively. The desmos link is attached if you wish to play around.

**2.1.1 Optimizing $h_k$ for the three different methods.** Let's estimate the performance of gradient descent for $L$-smooth functions $f \in C_L(\mathbb{R}^n)$ that are bounded below. Consider a single gradient step

$$x^{k+1} = x^k - h\nabla f(x^k) \ .$$

Recalling an equivalent characterization of $L$-smoothness, the result from Lemma 1.1.23 says

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &= f(x^k) + \left\langle \nabla f(x^k), x^k - h\nabla f(x^k) - x^k \right\rangle + \frac{L}{2}\|x^k - h\nabla f(x^k) - x^k\|^2 \\ &= f(x^k) - h\|\nabla f(x^k)\|^2 + \frac{h^2}{2}L\|\nabla f(x^k)\|^2 \\ &= f(x^k) - h(1 - \frac{h}{2}L)\|\nabla f(x^k)\|^2 \ . \end{aligned}$$

Therefore, to get the most decrease each step, we simply need to solve the one-dimensional problem:

$$\min_h -h(1 - \frac{h}{2}L) \Longrightarrow h^\star = \frac{1}{L} \ .$$

**Exercise 2**

*Verify that $h^\star$ is in fact the minimizer of $\Delta(h) = -h(1 - \frac{h}{2}L)$ (hint: consider the first and second derivatives).*

We have now proved the following lemma.

**Lemma 2.1.2 – Descent Lemma II**

*For $f \in C_L(\mathbb{R}^n)$, one step of gradient descent decreases the value of the objective function*

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$

Considering our stepsize scheme, for constant stepsizes $h_k = h = \alpha/L$ for $\alpha \in (0, 2)$, then

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

Of course, the optimal choice is still $h_k = \frac{1}{L}$, or choosing $\alpha = 1$, but we can guarantee descent for any $\alpha \in (0, 2)$.

**Remark**

*Note that setting $q_{x^k,\alpha}(x) := f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{L}{2\alpha}\|x - x^k\|^2$ acts as a quadratic approximation for $f(\cdot)$ at $x^k$. For $\alpha = 1$, this is the precisely the descent lemma. Regardless, computing its minimizer*

$$\operatorname{argmin} q_{x^k,\alpha}(x) = x^k - \frac{\alpha}{L}\nabla f(x^k)$$

*demonstrates that minimizing the quadratic upper bound yields the exact method for gradient descent!*

---

For $\alpha > 1$, we may overshoot the minimizer, but this is not an issue, as long as we make descent. Another desmos graph to see this in action!

Considering the Armijo rule, we note that from Exercise 1 that

$$f(x^{k+1}) \geq f(x^k) - \beta h_k\|\nabla f(x^k)\|^2 \ .$$

However, we have previously seen that

$$f(x^{k+1}) \leq f(x^k) - h_k(1 - \frac{h_k}{2})\|\nabla f(x^k)\|^2 \ ,$$

which concludes that $h_k \geq \frac{2}{L}(1-\beta)$. Using the other condition of Exercise 1, we have

$$f(x^{k+1}) \leq f(x^k) - \alpha h_k\|\nabla f(x^k)\|^2 \leq f(x^k) - \frac{2}{L}\alpha(1-\beta)\|\nabla f(x^k)\|^2 \ .$$

Letting $\omega = 2\alpha(1-\beta)$, we again recover the notation that

$$f(x^{k+1}) \leq f(x^k) - \frac{\omega}{L}\|\nabla f(x^k)\|^2 \ , \tag{2.4}$$

achieving the optimal decrease when $\alpha = \beta = \frac{1}{2}$. We are now ready to conclude something about gradient descent.

**Theorem 2.1 – Gradient Norm Bound**

*For $f \in C_L(\mathbb{R}^n)$, running gradient descent with stepsizes $h_k$ resulting in (2.4) yields the following inequality*

$$g_N^\star \leq \sqrt{\frac{L(f(x^0) - f^\star)}{\omega(N+1)}} \ , \tag{2.5}$$

*where $g_N^\star := \min_{0\leq k\leq N}\|\nabla f(x^k)\|$*

*Proof.* Since (2.4) holds for each $k = 0, \dots, N$, we can sum each inequality

$$\begin{aligned}
f(x^1) &\leq f(x^0) - \frac{\omega}{L}\|\nabla f(x^0)\|^2 \\
+\ f(x^2) &\leq f(x^1) - \frac{\omega}{L}\|\nabla f(x^1)\|^2 \\
+\ f(x^3) &\leq f(x^2) - \frac{\omega}{L}\|\nabla f(x^2)\|^2 \\
+ &\quad\ \vdots \\
+\ f(x^{N+1}) &\leq f(x^N) - \frac{\omega}{L}\|\nabla f(x^N)\|^2 \ ,
\end{aligned}$$

noting that we can cancel most of the terms

$$\begin{aligned}
\cancel{f(x^1)} &\leq f(x^0) - \frac{\omega}{L}\|\nabla f(x^0)\|^2 \\
+\ \cancel{f(x^2)} &\leq \cancel{f(x^1)} - \frac{\omega}{L}\|\nabla f(x^1)\|^2 \\
+\ \cancel{f(x^3)} &\leq \cancel{f(x^2)} - \frac{\omega}{L}\|\nabla f(x^2)\|^2 \\
+ &\quad\ \vdots \\
+\ f(x^{N+1}) &\leq \cancel{f(x^N)} - \frac{\omega}{L}\|\nabla f(x^N)\|^2 \\
\implies &\boxed{f(x^{N+1}) \leq f(x^0) - \frac{\omega}{L}\sum_{k=0}^{N}\|\nabla f(x^k)\|^2}
\end{aligned}$$

Rearranging, we can conclude that

$$\frac{\omega}{L}\sum_{k=0}^{N}\|\nabla f(x^k)\|^2 \leq f(x^0) - f(x^{N+1}) \leq f(x^0) - f^* \ , \tag{2.6}$$

as $f^*$ is a lower bound for all of $f(\cdot)$. Therefore, the series is bounded as $N \to \infty$, and consequently,

$$\|\nabla f(x^k)\| \to 0, \quad \text{as} \quad k \to \infty \ .$$

However, we can say more than *existence* of convergence, we can quantify the rate! Considering that an average of nonnegative values upper bounds the minimum (the average of a bunch of test scores clearly is higher than the lower score), we can rearrange (2.6) to conclude

$$g_N^\star \leq \sqrt{\frac{L(f(x^0) - f^\star)}{\omega(N+1)}} \ .$$

□

As we saw, the optimal decrease achieves $\omega = \frac{1}{2}$, so our first convergence bound says that running Algorithm 2 with $h_k = h = \frac{1}{L}$ results in

$$\min_{0\leq k\leq N}\|\nabla f(x^k)\| \leq \sqrt{\frac{2L(f(x^0) - f^\star)}{N+1}} \ .$$

If we want to guarantee that for a fixed $\varepsilon > 0$, that we will find some iterate $x^k$ whose gradient is sufficiently small, we simply need to upper bound our convergence rate by $\varepsilon$ and solve. That is, we consider

$$\min_{0\leq k\leq N}\|\nabla f(x^k)\| \leq \sqrt{\frac{2L(f(x^0) - f^\star)}{N+1}} < \varepsilon \ ,$$

and conclude that for

$$N \geq \frac{2L(f(x^0) - f^\star)}{\varepsilon^2} \implies \min_{0\leq k\leq N}\|\nabla f(x^k)\| < \varepsilon \ .$$

**Exercise 3**

*Let $A, C$ be $n \times n$ matrices and $b, d \in \mathbb{R}^n$. Consider*

$$f(x) = \|Ax - b\|^2 + \|Cx - d\|^2 \ .$$

*Explain how you would find a vector $\bar{x}$ such that $\|\nabla f(\bar{x})\| < 0.001$ (hint: find a value $L$ such that*

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \quad \forall x, y \in \mathbb{R}^n$$

*and an upper bound on $f(x^0) - f^\star$, for any chosen $x^0$).*

It may be useful to know the definition of norm of a matrix:

$$\|A\| := \sup_{x\neq 0}\frac{\|Ax\|}{\|x\|} \ ,$$

and subsequently for all $x \in \mathbb{R}^n$,

$$\|Ax\| \leq \|A\|\|x\| \ .$$

Note that in general, our goal is relatively tame: we only want to approach a *local* minimum. Still, even this goal can be unreachable to gradient descent. If a function has a saddle point, that is $\nabla f(x) = 0$ yet $x$ is neither a local minimum or maximum, we very well may approach that point, and then stay there for any future iterate (why?). In order to achieve stronger goals, with perhaps faster rates of convergence, we need stronger assumptions on the function we aim to minimize.

## 2.2 Smooth and *Convex* Functions

In this section, we still consider the general optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{2.7}$$

where the objective function is $L$-smooth (and hence differentiable). Recall that previously, we had aimed to solve this problem under fairly weak assumptions on $f(\cdot)$. Adopting the same assumptions as Nesterov [1, Assumption 2.1.1-2.1.3], we aim to construct a class of differentiable functions of differentiable functions $\mathscr{F}$ we wish to work with. From the conclusion of the last section, we note that the optimality conditions was simply not strong enough to guarantee a global minimum. In fact, a zero gradient wasn't even strong enough to ensure a local one! Let's first address this issue.

**Assumption 1 – $\mathscr{F}$ has nicer optimality conditions**

*For any $f \in \mathscr{F}$, the first-order optimality condition is sufficient to ensure a global minimizer. That is,*

$$\nabla f(\bar{x}) = 0 \Longrightarrow f(\bar{x}) \leq f(y), \ \forall y \in \mathbb{R}^n \ .$$

The next property we would like to see is some level of *invariance* for elements in our function class. That is, we would wish to be able to construct new functions from others in "simple" ways. Similar to the structure of vector spaces, we introduce only simple invariant operations for our function class.

**Assumption 2 – $\mathscr{F}$ is a cone (like a vector space, but nonnegative scaling)**

*If $f_1, f_2 \in \mathscr{F}$, and $\alpha, \beta \geq 0$, then*

$$\alpha f_1 + \beta f_2 \in \mathscr{F} \ .$$

Note we only consider nonnegative scaling $\alpha$ and $\beta$, as we would like $x^2$ to be in our class and not $-x^2$. Essentially, negatives turn minimizers into maximizers, so we arbitrarily pick the former and restrict ourselves accordingly.

Finally, let's add some basic functions to our class.

**Assumption 3 – $\mathscr{F}$ contains affine (shifted linear) functions**

*For any $\alpha \in \mathbb{R}$ and $s \in \mathbb{R}^n$,*

$$f(x) := \alpha + \langle s, x \rangle \in \mathscr{F} \ .$$

Note that for linear functions $\nabla f(x) = 0$ if and only if $f$ was constant, meaning that any $x$ is a global minimizer.

These three assumptions turn out to be exactly what we need to introduce our function class. Let $f \in \mathscr{F}$ and consider the perturbed function (which is in $\mathscr{F}$ by Assumptions 2 and 3) for a fixed

$x^0 \in \mathbb{R}^n$,

$$\phi(y) = f(y) - \left\langle \nabla f(x^0), y \right\rangle \ .$$

Note that

$$\nabla \phi(y)\Big|_{y=x^0} = \nabla f(x^0) - \nabla f(x^0) = 0 \ ,$$

so by Assumption 1, $x^0$ must be a global minimizer of $\phi(\cdot)$. Therefore, for any $y$,

$$f(y) - \left\langle \nabla f(x^0), y \right\rangle = \phi(y) \geq \phi(x^0) = f(x^0) - \left\langle \nabla f(x^0), x^0 \right\rangle ,$$

recovering the condition

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \ , \quad \forall x, y \in \mathbb{R}^n \tag{2.8}$$

since $x^0$ was arbitrary. We note this result has the precise result of convex functions demonstrated by Proposition 1.1.21. In fact, these functions are precisely the convex ones described earlier.

**2.2.1 $\mathscr{F}$ is the set of convex functions.** We already showed that these assumptions imply a lower linearization, so now we demonstrate the other direction. We will need a couple auxiliary lemmas to begin, showing the above condition is sufficient to satisfy all through assumptions. Therefore, functions that possess lower linearizations coincide with ones in $\mathscr{F}$.

**Lemma 2.2.1 – Assumption 1 is satisfied**

*Suppose differentiable $f : \mathbb{R}^n \to \mathbb{R}$ satisfies (2.8) and $\nabla f(x^\star) = 0$, then $x^\star$ is a global minimizer of $f(\cdot)$ on $\mathbb{R}^n$.*

*Proof.* For any $y \in \mathbb{R}^n$,

$$f(y) \geq f(x^\star) + \langle \nabla f(x^\star), y - x^\star \rangle = f(x^\star) \ .$$

□

**Lemma 2.2.2 – Assumption 2 is satisfied**

*Suppose differentiable $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ satisfies (2.8) and $\alpha, \beta \geq 0$. Then $\alpha f_1 + \beta f_2$ also satisfies (2.8).*

*Proof.* For any $x, y \in \mathbb{R}^n$, we have

$$f_1(y) \geq f_1(x) + \langle \nabla f_1(x), y - x \rangle \ .$$
$$f_2(y) \geq f_2(x) + \langle \nabla f_2(x), y - x \rangle \ .$$

Multiplying the first line by $\alpha$ and the second by $\beta$ and adding the two inequalities shows the desired result. □

**Lemma 2.2.3 – Assumption 3 is satisfied**

*For any $f$ satisfying (2.8), linear map $A : \mathbb{R}^n \to \mathbb{R}^m$, and vector $b \in \mathbb{R}^m$, then the map*

$$\phi(x) = f(Ax + b)$$

*also satisfies (2.8).*

*Proof.* Let $x, y \in \mathbb{R}^n$, $\bar{x} = Ax + b$, and $\bar{y} = Ay + b$. Note that from the chain rule,

$$\nabla\phi(x) = A^T \nabla f(Ax + b) \ ,$$

so we have

$$\begin{aligned}
\phi(y) = f(\bar{y}) &\geq f(\bar{x}) + \langle \nabla f(\bar{x}), \bar{y} - \bar{x} \rangle \\
&= \phi(x) + \langle \nabla f(\bar{x}), Ay - Ax \rangle \\
&= \phi(x) + \left\langle A^T \nabla f(\bar{x}), y - x \right\rangle \\
&= \phi(x) + \langle \nabla\phi(x), y - x \rangle \ ,
\end{aligned}$$

where the first inequality comes from $f$ satisfying (2.8), and the result follow from the definitions of $\bar{x}$, $\bar{y}$ and $A^T$ (one definition is that $\langle x, Ay \rangle = \langle A^T x, y \rangle$ for all $x$ and $y$). □

With the first characterization that

$$\mathscr{F} = \{f : f \text{ satisfies (2.8)}\}$$

out of the way, we can prove our characterization that this class of functions we want to analyze is precisely the set of convex functions.

**Theorem 2.2 – $\mathscr{F}$ is the set of differentiable convex functions**

*A function $f$ belongs to $\mathscr{F}$ if and only if for any $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \ . \tag{2.9}$$

*Proof.* Given $x, y \in \mathbb{R}^n$ and any $t \in [0, 1]$, define $x_t = tx + (1-t)y$. Let $f \in \mathscr{F}$, then

$$\begin{aligned}
f(x_t) &\leq f(y) - \langle \nabla f(x_t), y - x_t \rangle = f(y) - t \langle \nabla f(x_t), y - x \rangle \\
f(x_t) &\leq f(x) - \langle \nabla f(x_t), x - x_t \rangle = f(x) + (1-t) \langle \nabla f(x_t), y - x \rangle \ .
\end{aligned}$$

Multiplying the first equation by $(1-t)$ and the second by $t$ and adding results, we cancel the inner product terms and get the result

$$f(x_t) = f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \ .$$

On the other hand, if (2.9) holds for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0, 1]$, then for any $\alpha \in [0, 1)$ we can rearrange the expression to yield

$$\begin{aligned}
f(y) \geq \frac{1}{1-t} \left[ f(tx + (1-t)y) - tf(x) \right] &= f(x) + \frac{1}{1-t} \left[ f(tx + (1-t)y) - f(x) \right] \\
&= f(x) + \frac{f(x + (1-t)(y-x)) - f(x)}{1-t} \ .
\end{aligned}$$

This holds for all $t$, and letting $t \to 1$, we recover the directional derivative, which from Corollary 1.1.15 shows that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \ ,$$

and consequently $f \in \mathscr{F}$. □

There are several other equivalent characterizations of convex functions, which we will now discuss.

**2.2.2 Equivalent Characterizations for Convex and Smooth Functions** In order to analyze the convergence guarantees for gradient descent in the smooth and convex setting, it will be especially useful to utilize the following equivalent characterizations. All the proofs for the rates in which gradient descent converge have the same flavor:

1. Consider the gradient step $x^{k+1} = x^k - h_k \nabla f(x^k)$.
2. Utilize one of the properties of smooth and convex functions.
3. Substitute, rearrange, bound.

Different results may utilize different characterizations, in that some will tell us how function values are behaved, others may indicate growth on the difference between gradients.. Regardless, the following characterizations will prove extremely useful in analyzing our algorithm.

We begin by considering equivalences for convex functions.

**Theorem 2.3 – Equivalent Characterizations of Convex Functions**

*For differentiable $f : \mathbb{R}^n \to \mathbb{R}$, any $x, y \in \mathbb{R}^n$ and parameter $t \in [0,1]$, the following are equivalent:*

*i) $f$ is convex, denoted $f \in \mathscr{F}(\mathbb{R}^n)$*

*ii) (Jensen's inequality) $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$*

*iii) (First-order lower bound) $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$*

*iv) (Monotonicity of the gradient) $\langle f(x) - f(y), x - y \rangle \geq 0$*

*Proof.* We have already demonstrated $i) \iff ii) \iff iii)$. It suffices to conclude $iii) \iff iv)$. From the first-order lower bound, we have for any $x, y \in \mathbb{R}^n$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad f(x) \geq f(y) + \langle \nabla f(y), y - x \rangle \ .$$

Adding the two inequalities together yields $iv)$.

Now suppose $iv)$ holds. Define $x_t = x + t(y - x)$ for $t \in [0,1]$.

$$\begin{aligned}
f(y) &= f(x) + \int_0^1 \left\langle \nabla f(\underbrace{x + t(y-x)}_{x_t}), y - x \right\rangle dt \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x_t) - \nabla f(x), y - x \rangle \, dt \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \frac{1}{t} \langle \nabla f(x_t) - \nabla f(x), x + t(y-x) - x \rangle \, dt \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \underbrace{\int_0^1 \frac{1}{t} \langle \nabla f(x_t) - \nabla f(x), x_t - x \rangle \, dt}_{\geq 0} \\
&\geq f(x) + \langle f(x), y - x \rangle \ ,
\end{aligned}$$

where the first equation holds by the fundamental theorem of calculus (why?), the second equality adds and subtracts the linear term $\langle \nabla f(x), y = x \rangle$, and finally we simplify and utilize the monotonicty of the gradient. □

We discussed extensively properties *ii)* and *iii)*, but *iv)* also shares some insight on convex functions. This characterization roughly states that the gradients are increasing. That is, if $y > x$ then $\nabla f(y) > \nabla f(x)$. This can work in higher-dimensions too by considering only increasing one component and keeping the rest static.

Now we move onto smooth functions. There are several more properties of smooth functions, but below are the ones we will find particularly useful.

**Theorem 2.4 – Equivalent Characterizations of Smooth and Convex Functions**

*For differentiable $f : \mathbb{R}^n \to \mathbb{R}$ and any $x, y \in \mathbb{R}^n$, the following are equivalent:*

*i) $f$ is $L$-smooth and convex over $\mathbb{R}^n$, denoted $f \in \mathscr{F}_L(\mathbb{R}^n)$*

*ii) (Lipschitz continuous gradient)* $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

*iii) (Quadratic upper bound)* $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$

*iv) (Upper bounded monotonicity)* $\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\|x - y\|^2$

*v) (Cocoercivity)* $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$

*vi) (Lower bounded monotonicity)* $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$

*Proof.* As *i)* $\iff$ *ii)* by definition, we can prove the result by showing

$$ii) \implies iii) \implies iv) \implies v) \implies vi) \implies ii) \ .$$

- *ii)* $\implies$ *iii)*: Fundamental theorem of calculus.
- *iii)* $\implies$ *iv)*: Swap the roles of $x$ and $y$ and add the inequalities.
- *iv)* $\implies$ *v)*: Fix $x \in \mathbb{R}^n$ and define $\phi_x(y) := f(y) - \langle \nabla f(x), y\rangle$ (this should look familiar). We have already seen this function is convex with a minimizer at $y^\star = x$. Furthermore,

$$\begin{aligned}\langle \nabla \phi_x(y_1) - \nabla \phi_x(y_2), y_1 - y_2\rangle &= \langle (\nabla f(y_1) - \nabla f(x)) - (\nabla f(y_2) - \nabla f(x)), y_1 - y_2\rangle \\ &= \langle \nabla f(y_1) - \nabla f(y_2), y_1 - y_2\rangle \\ &\leq L\|y_1 - y_2\| \ .\end{aligned}$$

  Therefore, $\phi_x$ satisfies *iv)*. Again using fundamental theorem of calculus (see exercise below to fix this missing details), we can obtain *iii)* from *iv)*. Consequently, it also holds that

$$\phi_x(y_2) \leq \phi_x(y_1) + \langle \nabla \phi_x(y_1), y_2 - y_1\rangle + \frac{L}{2}\|y_2 - y_1\|^2 \ .$$

  Simply renaming, we have that for any $y \in \mathbb{R}^n$, consider

$$\phi_x(z) \leq \phi_x(y) + \langle \nabla \phi_x(y), z - y\rangle + \frac{L}{2}\|z - y\|^2$$

  and minimizing over $z$. We know that left hand side is minimized at $z^\star = x$, so we just need to minimize the right hand side. That is

$$f(x) - \langle \nabla f(x), x\rangle = \phi_x(x) \leq \phi_x(z) \leq \phi_x(y) + \langle \nabla \phi_x(y), z - y\rangle + \frac{L}{2}\|z - y\|^2 \ . \tag{2.10}$$

Recall, to minimize a (convex) function, it suffices to sets its gradient equal to zero, so consider the equation on the right hand side

$$Q_{x,y}(z) = \phi_x(y) + \langle \nabla \phi_x(y), z - y \rangle + \frac{L}{2}\|z - y\|^2$$

where we again recall $x$ and $y$ are fixed. Thus,

$$\nabla Q_{x,y}(z^\star) = \nabla \phi_x(y) + L(z^\star - y) = 0 \ .$$

To find $z^\star$, we substitute $\nabla \phi_x(y) = \nabla f(y) - \nabla f(x)$ and solve

$$\begin{aligned} &\nabla f(y) - \nabla f(x) + L(z^\star - y) = 0 \\ \implies &z^\star = y - \frac{\nabla f(y) - \nabla f(x)}{L} \end{aligned}$$

Substituting this value in (2.10) yields the desired bound

$$\begin{aligned} \phi_x(x) \leq Q_{x,y}(z^\star) &= \phi_x(y) + \langle \nabla \phi_x(y), z^\star - y \rangle + \frac{L}{2}\|z^\star - y\|^2 \\ &= \phi_x(y) + \left\langle \nabla \phi_x(y), \left( y - \frac{\nabla f(y) - \nabla f(x)}{L} \right) - y \right\rangle + \frac{L}{2} \left\| \left( y - \frac{\nabla f(y) - \nabla f(x)}{L} \right) - y \right\|^2 \\ &= \phi_x(y) + \left\langle \nabla \phi_x(y), \frac{\nabla f(y) - \nabla f(x)}{L} \right\rangle + \frac{L}{2} \left\| \frac{\nabla f(y) - \nabla f(x)}{L} \right\|^2 \end{aligned}$$

Recall $\phi_x(y) = f(y) - \langle \nabla f(x), y \rangle$ and $\nabla \phi_x(y) = \nabla f(y) - \nabla f(x)$, so

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle + \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \ .$$

Rearranging yields

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \ .$$

- $v) \implies vi)$: Swap the roles of $x$ and $y$ and add the inequalities.
- $vi) \implies ii)$: Considering the bound in $iv)$ and Cauchy Schwarz inequality,

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\| \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \|x - y\| \|\nabla f(x) - \nabla f(y)\| \ .$$

Multiplying both sides by $L$ and dividing by $\|\nabla f(x) - \nabla f(y)\|$ yields

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

□

The missing part of the proof above uses the fact that $iv) \implies iii)$ as well. The following exercise completes the proof.

**Exercise 4**

*Suppose for* $f : \mathbb{R}^n \to \mathbb{R}$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n \ .$$

*Show that*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \quad \forall x, y \in \mathbb{R}^n$$

*(hint: use fundamental theorem of calculus similar to the proof of Theorem 2.3, except this time upper bound the inner product).*

To summarize, the first two characterizations are how we defined $L$-smooth. The third and fourth handle upper bounding the function, either with a quadratic upper bound on the function itself or bounding the amount the gradients can change direction. The fifth and sixth bound handle the converse, bounding the function below (not by a quadratic in $x$, but by relating to the change in gradients) and the change in direction of the gradient. Combining the results, we get two two-way bounds (not symmetric, yet) stating

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

as well as

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

We can then see, both of these results conclude

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq L^2\|y - x\|^2 \implies \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \ ,$$

which is precisely the definition of $L$-smoothness. Note that $L$-smooth functions are rather restrictive, but this will be useful for ensuring that gradient descent works well.

**2.2.3 Equivalent Characterizations of Strongly Convex Functions** We now explore a dual characteristic of $L$-smoothness. Here, we use the word dual for multiple reasons, some more complex and rich in theory than others. At the very least, we will see that the characterizations "mimic" those of $L$-smooth functions quite visually. That is, when a function is smooth and strongly convex, we will now see two-sided *symmetric bounds*.

The proof of the following theorem is very similar to the previous, so we leave it without proof. However, note that some techniques need to be reversed. That is, we need to be careful (as always) with which way bounds go. That is, with Cauchy Schwarz for example, the inner products is bounded *above* by the product of the norms.

**Theorem 2.5 – Equivalent Characterizations of Strongly Convex Functions**

*For differentiable $f : \mathbb{R}^n \to \mathbb{R}$ and any $x, y \in \mathbb{R}^n$, the following are equivalent*

*i)* $f$ *is* $\mu$*-strongly convex*

*ii)* $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$

*iii)* $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$

---

Furthermore, if $f$ is $\mu$-strongly convex, then the following hold

a) $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|^2$

b) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|^2$

c) $\|\nabla f(x) - \nabla f(y)\| \geq \mu \|y - x\|$

We note that for a $\mu$-strongly convex function, similar two sides bound hold, but now the roles are reverse. Still, these bound aren't quite symmetric. That is, considering the monotonicity of the gradient

$$\mu\|x - y\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|^2 \ ,$$

where on the left hand side, we consider the squared distance of the points $x$ and $y$, while the right hand side considers the squared distance of their gradients. Regardless, strongly convex functions are very useful. In particular, the always admit a unique global minimizer!

**Exercise 5**

*Suppose differentiable $f : \mathbb{R}^n \to \mathbb{R}$ is $\mu$-strongly convex. Suppose a minimizer exists, show that it is unique.*

It will turn out that the class of functions we really would benefit from are those possessing both qualities. In this case, we get symmetric bounds

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \leq f(y) \leq f(x) + \langle f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \ . \tag{2.11}$$

We will see soon how to leverage this fact.

**2.2.4 Properties of Smooth and Strongly Convex Functions** The below exercises demonstrate the many properties of smooth and strong convex functions. When $f$ is convex and both $L$-smooth and $\mu$-strongly convex, we denoted $f \in \mathscr{F}_{L,\mu}(\mathbb{R}^n)$.

**Exercise 6**

*Suppose $f_1$ is $L_1$-smooth and $f_2$ is $L_2$-smooth. Is $f_1 + f_2$ smooth? If so, what is the value of the Lipschitz constant?*

**Exercise 7**

*Suppose $f_1$ is $\mu$-strongly convex and $f_2$ is $L_2$-strongly convex. Is $f_1 + f_2$ strongly convex? If so, what is the value of the modulus (i.e. $f_1$ has modulus $\mu_1$)?*

**Exercise 8**

*Suppose $f$ is $L$-smooth and $\mu$-strongly convex. Show that $\mu \leq L$.*

**Exercise 9**

*What class of functions attain equality. That is, what functions are $L$-smooth and $\mu$-strongly convex with $L = \mu$?*

**Exercise 10**

*Suppose $f$ is $L$-smooth and $\mu$-strongly convex. Show that $\phi(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex and $(L-\mu)$-smooth.*

**Exercise 11**

*Suppose $f$ is $L$-smooth and $\mu$-strongly convex. Show that the bound for the monotonicity for the gradient can be tightened*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2 \ .$$

*It will be useful to consider the function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|^2$ and the previous exercise.*

## 2.3 Proofs for Gradient Descent Convergence Guarantees

We finally now have the tools to present the performance of gradient descent applied to the problem

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

with $f \in \mathscr{F}_L(\mathbb{R}^n)$, the set of convex and $L$-smooth functions. Recall the method below.

**Algorithm – Gradient Descent**

**Input**: $x^0 \in \mathbb{R}^n$, target accuracy $\varepsilon > 0$ and stepsizes $h_k$

**For** $k = 0, 1, 2, \ldots$

1. $x^{k+1} \leftarrow x^k - h_k \nabla f(x^k)$

In this section, we only consider the constant step size $h_k = h > 0$, but it is possible to show that different schemes exhibit similar rates of convergence. Denote $x^\star$ as an optimal point, with $f^\star = f(x^\star)$.

**Theorem 2.6 – Convergence Guarantees for Smooth and Convex Functions**

*Let $f \in \mathscr{F}_L(\mathbb{R}^n)$ and $0 < h < \frac{2}{L}$. Then the iterations of gradient descent $\{x^k\}$, with corresponding function values satisfy*

$$f(x^k) - f^\star \leq \frac{2(f(x^0) - f^\star)\|x^0 - x^\star\|^2}{2L\|x^0 - x^\star\|^2 + kh(2 - Lh)(f(x^0) - f(x^\star))}, \quad \forall k \geq 0.$$

The proof of this will be the most technical one we've seen so far, combining many characteristics of convex and smooth functions. The following is aimed to be as digestible as possible, but it still may take a few read-throughs to fully understand. Regardless, the above theorem states that the suboptimality gap decreases at a *sublinear* rate of order $\mathcal{O}(1/k)$. That is, in the long run, the distance to optimal looks at most, inversely proportional to the number of steps taken. This turns out to be far from optimal, in which a modified version of gradient descent can achieve a rate of $\mathcal{O}(1/k^2)$, implying that the distance to optimal is proportional to the *squared* number of iterates.

*Proof.* Let $r_k = \|x^k - x^\star\|$. Then

$$\begin{aligned}
r_{k+1}^2 &= \|x^{k+1} - x^\star\|^2 \\
&= \|x^k - h\nabla f(x^k) - x^\star\|^2 \\
&= \|x^k - x^\star\|^2 - 2h\left\langle \nabla f(x^k), x^k - x^\star \right\rangle + h^2\|\nabla f(x^k)\|^2 \\
&= \|x^k - x^\star\|^2 - 2h\left\langle \nabla f(x^k) - \nabla f(x^\star), x^k - x^\star \right\rangle + h^2\|\nabla f(x^k)\|^2 \\
&\leq \|x^k - x^\star\|^2 - \frac{2h}{L}\|\nabla f(x^k) - \nabla f(x^\star)\|^2 + h^2\|\nabla f(x^k)\|^2 \\
&= r_k^2 - h\left(\frac{2}{L} - h\right)\|\nabla f(x^k)\|^2
\end{aligned}$$

where we use the gradient step in the first equality, we "FOIL" in the next equality, we add $\nabla f(x^\star) = 0$ in the next equality, we apply the "lower bounded monotonicity" from *vi*) in Theorem 2.4, and we conclude by simplfying the algebra and again noting that $\nabla f(x^\star) = 0$. Notably $r_0 \geq r_1 \geq ... \geq r_k$ for all $k \geq 0$.

Considering Lemma 1.1.23, we recall the "upper bounding quadratic" nature of smooth functions to note that

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \omega\|\nabla f(x^k)\|^2 \ ,
\end{aligned}$$

where $\omega = h\left(1 - \frac{L}{2}h\right)$. Then by convexity and Cauchy-Schwarz,

$$\underbrace{f(x^k) - f^\star}_{\Delta_k} \leq \left\langle \nabla f(x^k), x^k - x^\star \right\rangle \leq \|\nabla f(x^k)\|\|x^k - x^\star\| \leq r_0\|\nabla f(x^k)\| \ ,$$

where we recall $r_0 \geq r_k = \|x^k - x^\star\|$. Considering the above inequalities (subtracting $f^\star$ on both sides, we conclude

$$\underbrace{f(x^{k+1}) - f^\star}_{\Delta_{k+1}} \leq \underbrace{f(x^k) - f^\star}_{\Delta_k} - \omega\|\nabla f(x^k)\|^2 \leq \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2 \ .$$

Rearranging (and dividing by $\frac{1}{\Delta_k \Delta_{k+1}}$) yields the recurrence

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} - \frac{\omega}{r_0^2} \cdot \frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2} \ ,$$

where the last inequality uses the fact that $\Delta_k$ is non-increasing (from the descent lemma). Summing up all the inequalities, we again get a string of telescoping cancellations, and conclude that

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_0} + \frac{\omega}{r_0^2}(k+1) \ .$$

Considering $\Delta_{k+1} = f(x^{k+1}) - f^\star$, $\Delta_0 = f(x^0) - f^\star$, $r_0 = \|x^0 - x^\star\|$, and $\omega = h\left(1 - \frac{L}{2}h\right)$, all we need to do is rearrange the expression to conclude

$$f(x^k) - f^\star \leq \frac{2(f(x^0) - f^\star)\|x^0 - x^\star\|^2}{2L\|x^0 - x^\star\|^2 + kh(2 - Lh)(f(x^0) - f(x^\star))} \ ,$$

where we are looking at iterate $k$ for simplicity. □

Recall that we can maximize the amount of decrease each gradient step by setting

$$h = \frac{1}{L},$$

and we can further note that

$$f(x^0) - f^* \leq \frac{L}{2}\|x^0 - x^\star\|^2$$

by the descent lemma (again) and noting that $\nabla f(x^\star) = 0$ (again). Therefore, we can greatly simplify the expression above. Simply plugging in $h = \frac{1}{L}$ and bounding $\|x^0 - x^\star\|^2 \geq \frac{2}{L}(f(x^0) - f(x^\star))$ in the convergence guarantee, we prove the following corollary.

**Corollary 2.3.1**

*If $h_k = h = \frac{1}{L}$ and $f \in \mathscr{F}_L(\mathbb{R}^n)$, then*

$$f(x^k) - f^\star \leq \frac{2L\|x^0 - x^\star\|^2}{k+4} \leq \frac{2L\|x^0 - x^\star\|^2}{k}$$

We can now turn our attention to that of smooth *and strongly convex* functions. Recall that a consequence of $f : \mathbb{R}^n \to \mathbb{R}$ being $L$-smooth and $\mu$-strongly convex is the two sided bound

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

We can then leverage these bounds against each other. Notably, it can be shown that every $\frac{2L}{\mu}$ iterations of gradient descent with *halve* the function gap.

**Exercise 12**

*Suppose $f \in \mathscr{F}_{L,\mu}(\mathbb{R}^n)$ (i.e., $f$ is $L$-smooth and $\mu$-strongly convex). Suppose $f(x^0) - f^\star \leq R$. Show that after $k = \frac{8L}{\mu}$ iterations $f(x^k) - f^\star \leq \frac{R}{2}$. Conclude that*

$$f(x^N) - f^\star \leq \varepsilon$$

*for any $N \geq \frac{8L}{\mu} \log_2\left(\frac{R}{\varepsilon}\right)$ (hint. use Corollary 2.3.1 and bound $\|x^0 - x^\star\|$ with respect to $R$).*

For a more explicit (and classical) convergence guarantee that doesn't require an inductive arguement we move to the following theorem.

**Theorem 2.7**

*Let $f \in \mathscr{F}_{L,\mu}(\mathbb{R}^n)$ and $0 < h \leq \frac{2}{\mu+L}$. The the iterations of gradient descent $\{x^k\}$, with corresponding function values satisfy*

$$\|x^k - x^\star\|^2 \leq \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|x^0 - x^\star\|^2 \ .$$

*If $h = \frac{2}{\mu+L}$ then the following hold for all $k \geq 0$*

$$\begin{aligned}
\|x^k - x^\star\|^2 &\leq \left(\frac{L-\mu}{L+\mu}\right)^{2k} \|x^0 - x^\star\|^2 \\
f(x^k) - f^\star &\leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k} \|x^0 - x^\star\|^2 \\
f(x^k) - f^\star &\leq \frac{L}{\mu}\left(\frac{L-\mu}{L+\mu}\right)^{2k} (f(x^0) - f^\star)
\end{aligned}$$

*Proof.* Let $r_k = \|x^k - x^\star\|$. Then

$$\begin{aligned}
r_{k+1}^2 &= \|x^k - h\nabla f(x^k) - x^\star\|^2 \\
&= r_k^2 - 2h\left\langle \nabla f(x^k), x^k - x^\star \right\rangle + h^2\|\nabla f(x^k)\|^2 \\
&= r_k^2 - 2h\left\langle \nabla f(x^k) - \nabla f(x^\star), x^k - x^\star \right\rangle + h^2\|\nabla f(x^k)\|^2 \\
&\leq r_k^2 - 2h\left[\frac{\mu L}{\mu+L}\|x^k - x^\star\|^2 + \frac{1}{\mu+L}\|\nabla f(x^k) - \nabla f(x^\star)\|^2\right] + h^2\|\nabla f(x^k)\|^2 \\
&= \left(1 - \frac{2h\mu L}{\mu+L}\right) r_k^2 - h\left(\frac{2}{\mu+L} - h\right)\|\nabla f(x^k)\|^2 \ ,
\end{aligned}$$

where we utilize similar technique to the prove of the smooth convergence guarantee, while also utilizing the strengthened result from Exercise 11. We note that $h \leq \frac{2}{\mu+L}$ to achieve our first result, and the rest follow immediately from $\nabla f(x^\star) = 0$ and the upper and lower quadratic bounds. □

**Remark**

*Often the ratio between the smoothness and strong convexity parameters is useful to consider. As shown early, if a function is $L$-smooth and $\mu$-strongly convex, then $L \geq \mu$, so $Q_f := L/\mu \geq 1$. We can reformulate our convergence guarantees with this value instead*

$$\begin{aligned}
\|x^k - x^\star\|^2 &\leq \left(\frac{Q_f - 1}{Q_f + 1}\right)^{2k} \|x^0 - x^\star\|^2 \\
f(x^k) - f^\star &\leq \frac{L}{2}\left(\frac{Q_f - 1}{Q_f + 1}\right)^{2k} \|x^0 - x^\star\|^2 \\
f(x^k) - f^\star &\leq Q_f\left(\frac{Q_f - 1}{Q_f + 1}\right)^{2k} (f(x^0) - f^\star)
\end{aligned}$$

# References

[1] Yurii Nesterov. *Lectures on Convex Optimization*. Springer Cham, Switzerland, 2 edition, 2018.