

Convexity and Second Derivatives, Gradient Descent and Acceleration

Rasmus Kyng

Lecture 3 — Monday, March 4th

In the lecture today, we didn't cover accelerated gradient descent. We'll do that next week instead.

Notation for this lecture. In this lecture, we sometimes consider a multivariate functions f whose domain is a set $S \subseteq \mathbb{R}^n$, which we will require to be open. When we additionally require that S is convex, we will specify this. Note that $S = \mathbb{R}^n$ is both open and convex and it suffices to keep this case in mind. Things sometimes get more complicated if S is not open, e.g. when the domain of f has a boundary. We will leave those complications for another time.

1 A Review of Linear Algebra from the Previous Lecture

Semi-definiteness of a matrix. The following classification of symmetric matrices will be useful.

Definition 1.1. Let \mathbf{A} be a symmetric matrix in $\mathbb{R}^{n \times n}$. We say that \mathbf{A} is:

1. *positive definite* iff $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$;
2. *positive semidefinite* iff $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ for all $x \in \mathbb{R}^n$;
3. If neither \mathbf{A} nor $-\mathbf{A}$ is positive semi-definite, we say that \mathbf{A} is *indefinite*.

Example: indefinite matrix. Consider the following matrix \mathbf{A} :

$$\mathbf{A} := \begin{bmatrix} +4 & -1 \\ -1 & -2 \end{bmatrix}$$

For $\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, we have $\mathbf{x}^\top \mathbf{A} \mathbf{x} = 4 > 0$. For $\mathbf{x} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ we have $\mathbf{x}^\top \mathbf{A} \mathbf{x} = -2 < 0$. \mathbf{A} is therefore indefinite.

In the previous lecture, we saw the Courant-Fischer theorem:

Theorem 1.2 (The Courant-Fischer Theorem). *Let \mathbf{A} be a symmetric matrix in $\mathbb{R}^{n \times n}$, with eigenvalues $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. Then*

1.

$$\lambda_i = \min_{\substack{\text{subspace } W \subseteq \mathbb{R}^n \\ \dim(W)=i}} \max_{\mathbf{x} \in W, \mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

2.

$$\lambda_i = \max_{\substack{\text{subspace } W \subseteq \mathbb{R}^n \\ \dim(W)=n+1-i}} \min_{\mathbf{x} \in W, \mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

The following theorem gives a useful characterization of (semi)definite matrices.

Theorem 1.3. *Let \mathbf{A} be a symmetric matrix in $\mathbb{R}^{n \times n}$.*

1. \mathbf{A} is positive definite iff all its eigenvalues are positive;
2. \mathbf{A} is positive semidefinite iff all its eigenvalues are non-negative;

Theorem 1.3 follows immediately from the Courant-Fischer theorem.

Example: a positive semidefinite matrix. Consider the following matrix \mathbf{A} :

$$\mathbf{A} := \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

For $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, we have $\mathbf{Ax} = \mathbf{0}$, so $\lambda = 0$ is an eigenvalue of \mathbf{A} . For $\mathbf{x} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, we have $\mathbf{Ax} = \begin{pmatrix} 2 \\ -2 \end{pmatrix} = 2\mathbf{x}$, so $\lambda = 2$ is the other eigenvalue of \mathbf{A} . As both are non-negative, by the theorem above, \mathbf{A} is positive semidefinite.

Since we are learning about symmetric matrices, there is one more fact that everyone should know about them. We'll use $\lambda_{\max}(\mathbf{A})$ denote maximum eigenvalue of a matrix \mathbf{A} , and $\lambda_{\min}(\mathbf{A})$ the minimum.

Claim 1.4. *For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\|\mathbf{A}\| = \max(|\lambda_{\max}(\mathbf{A})|, |\lambda_{\min}(\mathbf{A})|)$.*

2 Characterizations of Convexity and Optimality via Second Derivatives

We will now use the second derivatives of a function to obtain characterizations of convexity and optimality. We will begin by introducing the *Hessian*, the matrix of pairwise second derivatives of a function. We will see that it plays a role in approximating a function via a second-order Taylor expansion. We will then use *semi-definiteness* of the Hessian matrix to characterize both conditions of optimality as well as the convexity of a function.

Definition 2.1. Given a function $f : S \rightarrow \mathbb{R}$ its **Hessian** matrix at point $\mathbf{x} \in S$ denoted $\mathbf{H}_f(\mathbf{x})$ (also sometimes denoted $\nabla^2 f(\mathbf{x})$) is:

$$\mathbf{H}_f(\mathbf{x}) := \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}(1)^2} & \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}(1)\partial \mathbf{x}(2)} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}(1)\partial \mathbf{x}(n)} \\ \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}(2)\partial \mathbf{x}(1)} & \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}(2)^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}(2)\partial \mathbf{x}(n)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}(n)\partial \mathbf{x}(1)} & \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}(n)\partial \mathbf{x}(2)} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}(n)^2} \end{bmatrix}$$

Second-order Taylor expansion. When f is twice differentiable it is possible to obtain an approximation of f by quadratic functions. Our definition of $f : S \rightarrow \mathbb{R}$ being twice (Fréchet) differentiable at $\mathbf{x} \in S$ is that there exists $\nabla f(\mathbf{x}) \in \mathbb{R}^n$ and $\mathbf{H}_f(\mathbf{x}) \in \mathbb{R}^{n \times n}$ s.t.

$$\lim_{\delta \rightarrow \mathbf{0}} \frac{\|f(\mathbf{x} + \delta) - f(\mathbf{x}) - (\nabla f(\mathbf{x})^\top \delta + \frac{1}{2} \delta^\top \mathbf{H}_f(\mathbf{x}) \delta)\|_2}{\|\delta\|_2^2} = 0.$$

This is equivalent to saying that for all δ

$$f(\mathbf{x} + \delta) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \delta + \frac{1}{2} \delta^\top \mathbf{H}_f(\mathbf{x}) \delta + o(\|\delta\|_2^2).$$

where by definition:

$$\lim_{\delta \rightarrow \mathbf{0}} \frac{o(\|\delta\|_2^2)}{\|\delta\|_2^2} = 0$$

We say that f is *continuously differentiable* on a set $S \subseteq \mathbb{R}^n$ if it is twice differentiable and in addition the gradient and Hessian are continuous on S .

As for first order expansions, we have a Taylor's Theorem, which we state in the so-called remainder form.

Theorem 2.2 (Taylor's Theorem, multivariate second-order remainder form). *If $f : S \rightarrow \mathbb{R}$ is twice continuously differentiable over $[\mathbf{x}, \mathbf{y}]$, then for some $\mathbf{z} \in [\mathbf{x}, \mathbf{y}]$,*

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top \mathbf{H}_f(\mathbf{z}) (\mathbf{y} - \mathbf{x})$$

2.1 A necessary condition for local extrema

Recall that in the previous lecture, we show the following proposition.

Proposition 2.3. *If \mathbf{x} is a local extremum of a differentiable function $f : S \rightarrow \mathbb{R}$ then $\nabla f(\mathbf{x}) = \mathbf{0}$.*

We can now give the second-order necessary conditions for local extrema via the Hessian.

Theorem 2.4. *Let $f : S \rightarrow \mathbb{R}$ be a function twice differentiable at $\mathbf{x} \in S$. If \mathbf{x} is a local minimum, then $\mathbf{H}_f(\mathbf{x})$ is positive semidefinite.*

Proof. Let us assume that \mathbf{x} is a local minimum. We know from Proposition 2.3 that $\nabla f(\mathbf{x}) = \mathbf{0}$, hence the second-order expansion at \mathbf{x} takes the form:

$$f(\mathbf{x} + \lambda \mathbf{d}) = f(\mathbf{x}) + \lambda^2 \frac{1}{2} \mathbf{d}^\top \mathbf{H}_f(\mathbf{x}) \mathbf{d} + o(\lambda^2 \|\mathbf{d}\|_2^2)$$

Because \mathbf{x} is a local minimum, we can then derive

$$0 \leq \lim_{\lambda \rightarrow 0^+} \frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda^2} = \frac{1}{2} \mathbf{d}^\top \mathbf{H}_f(\mathbf{x}) \mathbf{d}$$

This is true for any \mathbf{d} , hence $\mathbf{H}_f(\mathbf{x})$ is positive semidefinite. □

Remark 2.5. Again, for this proposition to hold, it is important that S is open.

2.2 A sufficient condition for local extrema

A local minimum thus is a stationary point and has a positive semi-definite Hessian. The converse is almost true, but we need to strengthen the Hessian condition slightly.

Theorem 2.6. *Let $f : S \rightarrow \mathbb{R}$ be a function twice differentiable at a stationary point $\mathbf{x} \in S$. If $\mathbf{H}_f(\mathbf{x})$ is positive definite then \mathbf{x} is a local minimum.*

Proof. Let us assume that $\mathbf{H}_f(\mathbf{x})$ is positive definite. We know that \mathbf{x} is a stationary point. We can write the second-order expansion at \mathbf{x} :

$$f(\mathbf{x} + \boldsymbol{\delta}) = f(\mathbf{x}) + \frac{1}{2} \boldsymbol{\delta}^\top \mathbf{H}_f(\mathbf{x}) \boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2^2)$$

Because the Hessian is positive definite, it has a strictly positive minimum eigenvalue λ_{\min} , we can conclude that $\boldsymbol{\delta}^\top \mathbf{H}_f(\mathbf{x}) \boldsymbol{\delta} \geq \lambda_{\min} \|\boldsymbol{\delta}\|_2^2$. From this, we conclude that when $\|\boldsymbol{\delta}\|_2^2$ is small enough, $f(\mathbf{x} + \boldsymbol{\delta}) - f(\mathbf{x}) \geq \frac{1}{4} \lambda_{\min} \|\boldsymbol{\delta}\|_2^2 > 0$. This proves that \mathbf{x} is a local minimum. \square

Remark 2.7. When $\mathbf{H}_f(\mathbf{x})$ is indefinite at a stationary point \mathbf{x} , we have what is known as a *saddle point*: \mathbf{x} will be a minimum along the eigenvectors of $\mathbf{H}_f(\mathbf{x})$ for which the eigenvalues are positive and a maximum along the eigenvectors of $\mathbf{H}_f(\mathbf{x})$ for which the eigenvalues are negative.

2.3 Characterization of convexity

Definition 2.8. For a convex set $S \subseteq \mathbb{R}^n$, we say that a function $f : S \rightarrow \mathbb{R}$ is **strictly convex on S** if for any two points $\mathbf{x}_1, \mathbf{x}_2 \in S$ and any $\theta \in (0, 1)$ we have that:

$$f(\theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2) < \theta f(\mathbf{x}_1) + (1 - \theta) f(\mathbf{x}_2).$$

Theorem 2.9. *Let $S \subseteq \mathbb{R}^n$ be open and convex, and let $f : S \rightarrow \mathbb{R}$ be twice continuously differentiable.*

1. *If $\mathbf{H}_f(\mathbf{x})$ is positive semi-definite for any $\mathbf{x} \in S$ then f is convex on S .*
2. *If $\mathbf{H}_f(\mathbf{x})$ is positive definite for any $\mathbf{x} \in S$ then f is **strictly** convex on S .*
3. *If f is convex, then $\mathbf{H}_f(\mathbf{x})$ is positive semi-definite $\forall \mathbf{x} \in S$.*

Proof.

1. By applying Theorem 2.2, we find that for some $\mathbf{z} \in [\mathbf{x}, \mathbf{y}]$:

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} \left((\mathbf{y} - \mathbf{x})^\top \mathbf{H}_f(\mathbf{z}) (\mathbf{y} - \mathbf{x}) \right)$$

If $\mathbf{H}_f(\mathbf{z})$ is positive semi-definite, this necessarily implies that:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

and from Theorem 3.5 we get that f is convex.

2. if $H_f(\mathbf{x})$ is positive definite, we have that:

$$f(\mathbf{y}) > f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$

Applying the same idea as in Theorem 3.5 we can show that in this case f is **strictly** convex.

3. Let f be a convex function. For $\mathbf{x} \in S$, and some small $\lambda > 0$, for any $\mathbf{d} \in \mathbb{R}^n$ we have that $\mathbf{x} + \lambda \mathbf{d} \in S$. From the Taylor expansion of f we get:

$$f(\mathbf{x} + \lambda \mathbf{d}) = f(\mathbf{x}) + \lambda \nabla f(\mathbf{x})^\top \mathbf{d} + \frac{\lambda^2}{2} \mathbf{d}^\top H_f(\mathbf{x}) \mathbf{d} + o(\lambda^2 \|\mathbf{d}\|_2^2).$$

From Lemma 3.5 we get that if f is convex then:

$$f(\mathbf{x} + \lambda \mathbf{d}) \geq f(\mathbf{x}) + \lambda \nabla f(\mathbf{x})^\top \mathbf{d}.$$

Therefore, we have that for any $\mathbf{d} \in \mathbb{R}^n$:

$$\frac{\lambda^2}{2} \mathbf{d}^\top H_f(\mathbf{x}) \mathbf{d} + o(\|\lambda \mathbf{d}\|^2) \geq 0$$

Dividing by λ^2 and taking $\lambda \rightarrow 0^+$ gives us that for any $\mathbf{d} \in \mathbb{R}^n$: $\mathbf{d}^\top H_f(\mathbf{x}) \mathbf{d} \geq 0$. □

Remark 2.10. It is important to note that if S is open and f is strictly convex, then $H_f(\mathbf{x})$ may still (only) be positive semi-definite $\forall \mathbf{x} \in S$. Consider $f(x) = x^4$ which is strictly convex, then the Hessian is $H_f(x) = 12x^2$ which equals 0 at $x = 0$.

3 Gradient Descent - An Approach to Optimization?

We have begun to develop an understanding of convex functions, and what we have learned already suggests a way for us to try to find an approximate minimizer of a given convex function.

Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and differentiable, and we want to solve

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

We would like to find \mathbf{x}^* , a global minimizer of f . Suppose we start with some initial guess \mathbf{x}_0 , and we want to update it to \mathbf{x}_1 with $f(\mathbf{x}_1) < f(\mathbf{x}_0)$. If we can repeatedly make updates like this, maybe we eventually find a point with nearly minimum function value, i.e. some $\tilde{\mathbf{x}}$ with $f(\tilde{\mathbf{x}}) \approx f(\mathbf{x}^*)$?

Recall that $f(\mathbf{x}_0 + \boldsymbol{\delta}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2)$. This means that if we choose $\mathbf{x}_1 = \mathbf{x}_0 - \lambda \nabla f(\mathbf{x}_0)$, we get

$$f(\mathbf{x}_0 + \boldsymbol{\delta}) = f(\mathbf{x}_0) - \lambda \|\nabla f(\mathbf{x}_0)\|_2^2 + o(\lambda \|\nabla f(\mathbf{x}_0)\|_2)$$

And because f is convex, we know that $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$ unless we are already at a global minimum. So, for some small enough $\lambda > 0$, we should get $f(\mathbf{x}_1) < f(\mathbf{x}_0)$ unless we're already at a global minimizer. This idea of taking a step in the direction of $-\nabla f(\mathbf{x}_0)$ is what is called *Gradient Descent*. But how do we choose λ each time? And does this lead to an algorithm that quickly reaches a point with close to minimal function value? To get good answers to these questions, we need to assume more about the function f that we are trying to minimize.

In the following subsection, we will see some conditions that suffice. But there are also many other settings where one can show that some form of gradient descent converges.

3.1 A Quantitative Bound on Changes in the Gradient

Definition 3.1. Let $f : S \rightarrow \mathbb{R}$ be a differentiable function, where $S \subseteq \mathbb{R}^n$ is convex and open. We say that f is β -gradient Lipschitz iff for all $\mathbf{x}, \mathbf{y} \in S$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \beta \|\mathbf{x} - \mathbf{y}\|_2.$$

We also refer to this as f being β -smooth.

Proposition 3.2. Consider a twice differentiable $f : S \rightarrow \mathbb{R}$. Then f is β -gradient Lipschitz if and only if for all $\mathbf{x} \in S$ (except a measure zero set), $\lambda_{\max}(\mathbf{H}_f(\mathbf{x})) \leq \beta$.

You will prove this in Exercise 1 of this week's exercises.

Proposition 3.3. Let $f : S \rightarrow \mathbb{R}$ be a β -gradient Lipschitz function. Then for all $\mathbf{x}, \mathbf{y} \in S$,

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

To prove this proposition, we need the following result from multi-variate calculus. This is a restricted form of the fundamental theorem of calculus for line integrals.

Proposition 3.4. Let $f : S \rightarrow \mathbb{R}$ be a differentiable function, and consider \mathbf{x}, \mathbf{y} such that $[\mathbf{x}, \mathbf{y}] \in S$. Let $\mathbf{x}_\theta = \mathbf{x} + \theta(\mathbf{y} - \mathbf{x})$. Then

$$f(\mathbf{y}) = f(\mathbf{x}) + \int_{\theta=0}^1 \nabla f(\mathbf{x}_\theta)^\top (\mathbf{y} - \mathbf{x}) d\theta$$

Now, we're in a position to show Proposition 3.3

Proof of Proposition 3.3. Let $f : S \rightarrow \mathbb{R}$ be a β -gradient Lipschitz function. Consider arbitrary $\mathbf{x}, \mathbf{y} \in S$ such that $[\mathbf{x}, \mathbf{y}] \in S$

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \int_{\theta=0}^1 \nabla f(\mathbf{x}_\theta)^\top (\mathbf{y} - \mathbf{x}) d\theta \\ &= f(\mathbf{x}) + \int_{\theta=0}^1 \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) d\theta + \int_{\theta=0}^1 (\nabla f(\mathbf{x}_\theta) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) d\theta \end{aligned}$$

Next we use Cauchy-Schwarz pointwise.

We also evaluate the first integral.

$$\leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \int_{\theta=0}^1 \|\nabla f(\mathbf{x}_\theta) - \nabla f(\mathbf{x})\| \|\mathbf{y} - \mathbf{x}\| d\theta$$

Then we apply β -gradient Lipschitz and note $\mathbf{x}_\theta - \mathbf{x} = \theta(\mathbf{y} - \mathbf{x})$.

$$\leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \int_{\theta=0}^1 \beta \theta \|\mathbf{y} - \mathbf{x}\|^2 d\theta.$$

$$= f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\beta}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

□

3.2 Analyzing Gradient Descent

It turns out that just knowing a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and β -gradient Lipschitz is enough to let us figure out a reasonable step size for Gradient Descent and let us analyze its convergence.

We start at a point $\mathbf{x}_0 \in \mathbb{R}^n$, and we try to find a point $\mathbf{x}_1 = \mathbf{x}_0 + \boldsymbol{\delta}$ with lower function value. We will let our upper bound from Proposition 3.3 guide us, in fact, we could ask, what is the *best* update for minimizing this upper bound, i.e. a $\boldsymbol{\delta}$ solving

$$\min_{\boldsymbol{\delta} \in \mathbb{R}^n} f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \boldsymbol{\delta} + \frac{\beta}{2} \|\boldsymbol{\delta}\|^2$$

We can compute the best according to this upper bound by noting first that function is convex and continuously differentiable, and hence will be minimized at any point where the gradient is zero. Thus we want $\mathbf{0} = \nabla_{\boldsymbol{\delta}} \left(f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \boldsymbol{\delta} + \frac{\beta}{2} \|\boldsymbol{\delta}\|^2 \right) = \nabla f(\mathbf{x}_0) + \beta \boldsymbol{\delta}$, which occurs at $\boldsymbol{\delta} = -\frac{1}{\beta} \nabla f(\mathbf{x}_0)$.

Plugging in this value into the upper bound, we get that $f(\mathbf{x}_1) \leq f(\mathbf{x}_0) - \frac{\|\nabla f(\mathbf{x}_0)\|_2^2}{2\beta}$.

Now, as our algorithm, we will start with some guess \mathbf{x}_0 , and then at every step we will update our guess using the best step based on our Proposition 3.3 upper bound on f at \mathbf{x}_i , and so we get

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \frac{1}{\beta} \nabla f(\mathbf{x}_i) \text{ and } f(\mathbf{x}_{i+1}) \leq f(\mathbf{x}_i) - \frac{\|\nabla f(\mathbf{x}_i)\|_2^2}{2\beta}. \quad (1)$$

Let us try to prove that our algorithm converges toward an \mathbf{x} with low function value.

We will measure this by looking at

$$\text{gap}_i = f(\mathbf{x}_i) - f(\mathbf{x}^*)$$

where \mathbf{x}^* is a global minimizer of f (note that there may not be a unique minimizer of f). We will try to show that this function value gap grows small. Using $f(\mathbf{x}_{i+1}) - f(\mathbf{x}_i) = \text{gap}_{i+1} - \text{gap}_i$, we get

$$\text{gap}_{i+1} - \text{gap}_i \leq -\frac{\|\nabla f(\mathbf{x}_i)\|_2^2}{2\beta} \quad (2)$$

If the gap_i value is never too much bigger than $\frac{\|\nabla f(\mathbf{x}_i)\|_2^2}{2\beta}$, then this should help us show we are making progress. But how much can they differ? We will now try to show a limit on this.

Recall that in the previous lecture we showed the following theorem.

Theorem 3.5. *Let S be an open convex subset of \mathbb{R}^n , and let $f : S \rightarrow \mathbb{R}$ be a differentiable function. Then, f is convex if and only if for any $\mathbf{x}, \mathbf{y} \in S$ we have that $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$.*

Using the convexity of f and the lower bound on convex functions given by Theorem 3.5, we have that

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_i) + \nabla f(\mathbf{x}_i)^\top (\mathbf{x}^* - \mathbf{x}_i) \quad (3)$$

Rearranging gets us

$$\begin{aligned} \text{gap}_i &\leq \nabla f(\mathbf{x}_i)^\top (\mathbf{x}_i - \mathbf{x}^*) \\ &\leq \|\nabla f(\mathbf{x}_i)\|_2 \|\mathbf{x}_i - \mathbf{x}^*\|_2 \end{aligned} \tag{4}$$

by Cauchy-Schwarz.

At this point, we are essentially ready to connect Equation (2) with Equation (4) and analyze the convergence rate of our algorithm.

However, at the moment, we see that the change $\text{gap}_{i+1} - \text{gap}_i$ in how close we are to the optimum function value is governed by the norm of the gradient $\|\nabla f(\mathbf{x}_i)\|_2$, while the size of the gap is related to *both* this quantity and the distance $\|\mathbf{x}_i - \mathbf{x}^*\|_2$ between the current solution \mathbf{x}_i and an optimum \mathbf{x}^* . Do we need both or can we get rid of, say, the distance? Unfortunately, with this algorithm and for this class of functions, a dependence on the distance is necessary. However, we can simplify things considerably using the following observation, which you will prove in the exercises (Exercise 2):

Claim 3.6. *When running Gradient Descent as given by the step in Equation (1), for all i $\|\mathbf{x}_i - \mathbf{x}^*\|_2 \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2$.*

Combining this Claim with Equation (2) and Equation (4),

$$\text{gap}_{i+1} - \text{gap}_i \leq -\frac{1}{2\beta} \cdot \left(\frac{\text{gap}_i}{\|\mathbf{x}_0 - \mathbf{x}^*\|_2} \right)^2 \tag{5}$$

At this point, a simple induction will complete the proof of following result.

Theorem 3.7. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a β -gradient Lipschitz, convex function. Let \mathbf{x}_0 be a given starting point, and let $\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$ be a minimizer of f . The Gradient Descent algorithm given by*

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \frac{1}{\beta} \nabla f(\mathbf{x}_i)$$

ensures that the k th iterate satisfies

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{2\beta \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{k+1}.$$

Carrying out this induction is one of this week's exercises (Exercise 3).

4 Accelerated Gradient Descent

It turns out that we can get an algorithm that converges substantially faster than Gradient Descent, using an approach known as *Accelerated Gradient Descent*, which was developed by Nesterov [Nes83]. This algorithm in turn improved on some earlier results by Nemirovski and Yudin [NY83]. The phenomenon of acceleration was perhaps first understood in the context of quadratic functions, minimizing $\mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbf{x}^\top \mathbf{b}$ when \mathbf{A} is positive definite – for this case, the Conjugate Gradient algorithm was developed independently by Hestenes and Stiefel [HS⁺52] (here at ETH!), and by Lanczos [Lan52]. In the past few years, providing more intuitive explanations of

acceleration has been a popular research topic. Today’s lecture is based on an analysis of Nesterov’s algorithm developed by Diakonikolas and Orecchia [DO19].

We will adopt a slightly different approach to analyzing this algorithm than what we used in the previous section for Gradient Descent.

We will use \mathbf{x}_0 to denote the starting point of our algorithm, and we will produce a sequence of iterates $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$. At each iterate \mathbf{x}_i , we will compute the gradient $\nabla f(\mathbf{x}_i)$. However, the way we choose \mathbf{x}_{i+1} based on what we know so far will now be a little more involved than what we did for Gradient Descent.

To help us understand the algorithm, we are going to introduce two more sequences of iterates $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k \in \mathbb{R}^n$ and $\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k \in \mathbb{R}^n$.

The sequence of \mathbf{y}_i ’s will be constructed to help us get as low a function value as possible at $f(\mathbf{y}_i)$, which we will consider our current solution and the last iterate \mathbf{y}_k will be the output solution of our algorithm.

The sequence of \mathbf{v}_i ’s will be constructed to help us get a lower bound on $f(\mathbf{x}^*)$.

By combining the upper bound on the function value of our current solution $f(\mathbf{y}_i)$ with a lower bound on the function value at an optimal solution $f(\mathbf{x}^*)$, we get an upper bound on the gap $f(\mathbf{y}_i) - f(\mathbf{x}^*)$ between the value of our solution and the optimal one. Finally, each iterate \mathbf{x}_i , which will be where we evaluate gradient $\nabla f(\mathbf{x}_i)$, is chosen through a trade-off between wanting to reduce the upper bound and wanting to increase the lower bound.

The upper bound sequence: \mathbf{y}_i ’s. The point \mathbf{y}_i will be chosen from \mathbf{x}_i to minimize an upper bound on f based at \mathbf{x}_i . This is similar to what we did in the previous section. We let $\mathbf{y}_i = \mathbf{x}_i + \boldsymbol{\delta}_i$ and use choose $\boldsymbol{\delta}_i$ to minimize the upper bound $f(\mathbf{y}_i) \leq f(\mathbf{x}_i) + \nabla f(\mathbf{x}_i)^\top \boldsymbol{\delta}_i + \frac{\beta}{2} \|\boldsymbol{\delta}_i\|^2$, which gives us

$$\mathbf{y}_i = \mathbf{x}_i - \frac{1}{\beta} \nabla f(\mathbf{x}_i) \text{ and } f(\mathbf{y}_i) \leq f(\mathbf{x}_i) - \frac{\|\nabla f(\mathbf{x}_i)\|_2^2}{2\beta}.$$

We will introduce a notation for this upper bound

$$U_i = f(\mathbf{y}_i) \leq f(\mathbf{x}_i) - \frac{\|\nabla f(\mathbf{x}_i)\|_2^2}{2\beta}. \tag{6}$$

Philosophizing about lower bounds¹. A crucial ingredient to establishing an upper bound on gap_i was a lower bound on $f(\mathbf{x}^*)$.

In our analysis of Gradient Descent, in Equation (4), we used the lower bound $f(\mathbf{x}^*) \geq f(\mathbf{x}_i) - \|\nabla f(\mathbf{x}_i)\|_2 \|\mathbf{x}_i - \mathbf{x}^*\|_2$. We can think of the Gradient Descent analysis as being based on a tension between two statements: Firstly that “a large gradient implies we quickly approach the optimum” and secondly “the function value gap to optimum cannot exceed the magnitude of the current gradient (scaled by distance to opt)”.

¹YMMV. People have a lot of different opinions about how to understand acceleration, and you should take my thoughts with a grain of salt.

This analysis does not use that we have seen many different function values and gradients, and each of these can be used to construct a lower bound on the optimum value $f(\mathbf{x}^*)$, and, in particular, it is not clear that the last gradient provides the best bound. To do better, we will try to use lower bounds that take advantage of all the gradients we have seen.

Definition 4.1. We will adopt a new notation for inner products that sometimes is more convenient when dealing with large expressions: $\langle \mathbf{a}, \mathbf{b} \rangle \stackrel{\text{def}}{=} \mathbf{a}^\top \mathbf{b}$.

The lower bound sequence: v_i 's. We can introduce weights $a_i > 0$ for each step and combine the gradients we have observed into one lower bound based on a weighted average. Let us use $A_i = \sum_{j \leq i} a_j$ to denote the sum of the weights. Now a general lower bound on the function value at any $\mathbf{v} \in \mathbb{R}^n$ is :

$$f(\mathbf{v}) \geq \frac{1}{A_i} \sum_{j \leq i} a_j (f(\mathbf{x}_j) + \langle \nabla f(\mathbf{x}_j), \mathbf{v} - \mathbf{x}_j \rangle)$$

However, to use Cauchy-Schwarz on each individual term here to instantiate this bound at \mathbf{x}^* does not give us anything useful. Instead, we will employ a somewhat magical trick: we introduce a regularization term

$$\phi(\mathbf{v}) \stackrel{\text{def}}{=} \frac{\sigma}{2} \|\mathbf{v} - \mathbf{x}_0\|_2^2.$$

We will choose the value $\sigma > 0$ later. Now we derive our lower bound L_i

$$\begin{aligned} f(\mathbf{x}^*) &\geq \frac{1}{A_i} \left(\phi(\mathbf{x}^*) + \sum_{j \leq i} a_j f(\mathbf{x}_j) + \langle a_j \nabla f(\mathbf{x}_j), \mathbf{x}^* - \mathbf{x}_j \rangle \right) - \frac{\phi(\mathbf{x}^*)}{A_i} \\ &\geq \min_{\mathbf{v} \in \mathbb{R}^n} \left\{ \frac{1}{A_i} \left(\phi(\mathbf{v}) + \sum_{j \leq i} a_j f(\mathbf{x}_j) + \langle a_j \nabla f(\mathbf{x}_j), \mathbf{v} - \mathbf{x}_j \rangle \right) \right\} - \frac{\phi(\mathbf{x}^*)}{A_i} \\ &= L_i \end{aligned}$$

We will let \mathbf{v}_i be the \mathbf{v} obtaining the minimum in the optimization problem appearing in the definition of L_i , so that

$$L_i = \frac{1}{A_i} \left(\phi(\mathbf{v}_i) + \sum_{j \leq i} a_j f(\mathbf{x}_j) + \langle a_j \nabla f(\mathbf{x}_j), \mathbf{v}_i - \mathbf{x}_j \rangle \right) - \frac{\phi(\mathbf{x}^*)}{A_i}$$

How we will measure convergence. We have designed the upper bound U_i and the lower bound L_i such that $\text{gap}_i = f(\mathbf{y}_i) - f(\mathbf{x}^*) \leq U_i - L_i$.

As you will show in Exercise 3, we can prove the convergence of Gradient Descent directly by an induction that establishes $1/\text{gap}_i \leq C \cdot i$ for some constant C depending on the Lipschitz gradient parameter β and the distance $\|\mathbf{x}_0 - \mathbf{x}^*\|_2$.

To analyze Accelerated Gradient Descent, we will adopt a similar, but slightly different strategy, namely trying to show that $(U_i - L_i)r(i)$ is non-increasing for some positive “rate function” $r(i)$. Ideally $r(i)$ should grow quickly, which would imply that gap_i quickly gets small. We will also need

to show that $(U_0 - L_0)r(0) \leq C$ for some constant C again depending on β and $\|\mathbf{x}_0 - \mathbf{x}^*\|_2$. Then, we'll be able to conclude that

$$\text{gap}_i \cdot r(i) \leq (U_i - L_i)r(i) \leq (U_{i-1} - L_{i-1})r(i-1) \leq \dots \leq (U_0 - L_0)r(0) \leq C,$$

and hence $\text{gap}_i \leq C/r(i)$.

This framework is fairly general. We could have also used it to analyze Gradient Descent, and it works for many other optimization algorithms too.

We are going to choose our rate function $r(i)$ to be exactly A_i , which of course is no accident! As we will see, this interacts nicely with our lower bound L_i . Hence, our goals are to

1. provide an upper bound on $A_0(U_0 - L_0)$,
2. and show that $A_{i+1}(U_{i+1} - L_{i+1}) \leq A_i(U_i - L_i)$,

Establishing the convergence rate. Let's start by looking at the change in the upper bound scaled by our rate function:

$$\begin{aligned} A_{i+1}U_{i+1} - A_iU_i &= A_{i+1}(f(\mathbf{y}_{i+1}) - f(\mathbf{x}_{i+1})) - A_i(f(\mathbf{y}_i) - f(\mathbf{x}_{i+1})) + (A_{i+1} - A_i)f(\mathbf{x}_{i+1}) \quad (7) \\ &\leq A_{i+1} \left(-\frac{\|\nabla f(\mathbf{x}_{i+1})\|_2^2}{2\beta} \right) \quad \text{First term controlled by Equation (6).} \\ &\quad + A_i \langle \nabla f(\mathbf{x}_{i+1}), \mathbf{y}_i - \mathbf{x}_{i+1} \rangle \quad \text{Second term bounded by Theorem 3.5.} \\ &\quad + a_{i+1}f(\mathbf{x}_{i+1}) \quad \text{Third term uses } a_{i+1} = A_{i+1} - A_i. \end{aligned}$$

The solution \mathbf{v}_i to the minimization in the lower bound L_i turns out to be relatively simple to characterize. By using derivatives to find the optimum, we first analyze the initial value of the lower bound L_0 .

Claim 4.2.

1. $\mathbf{v}_0 = \mathbf{x}_0 - \frac{a_0}{\sigma} \nabla f(\mathbf{x}_0)$
2. $L_0 = f(\mathbf{x}_0) - \frac{a_0}{2\sigma} \|\nabla f(\mathbf{x}_0)\|_2^2 - \frac{\sigma}{2a_0} \|\mathbf{x}^* - \mathbf{x}_0\|_2^2$.

You will prove Claim 4.2 in next week's exercises (postponed from this week, because we didn't end up covering accelerated gradient descent). Noting $A_0 = a_0$, we see from Equation (6) and Part 2 of Claim 4.2, that

$$A_0(U_0 - L_0) \leq \left(\frac{a_0^2}{2\sigma} - \frac{a_0}{2\beta} \right) \|\nabla f(\mathbf{x}_0)\|_2^2 + \frac{\sigma}{2} \|\mathbf{x}^* - \mathbf{x}_0\|_2^2 \quad (8)$$

It will be convenient to introduce notation for the rescaled lower bound A_iL_i *without* optimizing over \mathbf{v} .

$$m_i(\mathbf{v}) = \phi(\mathbf{v}) - \phi(\mathbf{x}^*) + \sum_{j \leq i} a_j f(\mathbf{x}_j) + \langle a_j \nabla f(\mathbf{x}_j), \mathbf{v} - \mathbf{x}_j \rangle$$

Thus $A_iL_i - A_{i+1}L_{i+1} = m_i(\mathbf{v}_i) - m_{i+1}(\mathbf{v})$. Now, it is not too hard to show the following relationships.

Claim 4.3.

1. $m_i(\mathbf{v}) = m_i(\mathbf{v}_i) + \frac{\sigma}{2} \|\mathbf{v} - \mathbf{v}_i\|_2^2$
2. $m_{i+1}(\mathbf{v}) = m_i(\mathbf{v}) + a_{i+1}f(\mathbf{x}_{i+1}) + \langle a_{i+1}\nabla f(\mathbf{x}_{i+1}), \mathbf{v} - \mathbf{x}_{i+1} \rangle$
3. $\mathbf{v}_{i+1} = \mathbf{v}_i - \frac{a_{i+1}}{\sigma}\nabla f(\mathbf{x}_{i+1})$

And again, you will prove Claim 4.3 in next week's exercises (postponed from this week, because we didn't end up covering accelerated gradient descent). *Hint for Part 1: note that $m_i(v)$ is a quadratic function, minimized at v_i and its Hessian equals σI at all v .*

Given Claim 4.3, we see that

$$\begin{aligned} A_i L_i - A_{i+1} L_{i+1} &= m_i(\mathbf{v}_i) - m_{i+1}(\mathbf{v}_{i+1}) = -a_{i+1}f(\mathbf{x}_{i+1}) - \langle a_{i+1}\nabla f(\mathbf{x}_{i+1}), \mathbf{v}_{i+1} - \mathbf{x}_{i+1} \rangle - \frac{\sigma}{2} \|\mathbf{v}_{i+1} - \mathbf{v}_i\|_2^2 \\ &= -a_{i+1}f(\mathbf{x}_{i+1}) - \langle a_{i+1}\nabla f(\mathbf{x}_{i+1}), \mathbf{v}_i - \mathbf{x}_{i+1} \rangle + \frac{a_{i+1}^2}{2\sigma} \|\nabla f(\mathbf{x}_{i+1})\|_2^2 \end{aligned} \quad (9)$$

This means that by combining Equation (8) and (9) we get

$$A_{i+1}(U_{i+1} - L_{i+1}) - A_i(U_i - L_i) \leq \left(\frac{-A_{i+1}}{2\beta} + \frac{a_{i+1}^2}{2\sigma} \right) \|\nabla f(\mathbf{x}_{i+1})\|_2^2 + \langle \nabla f(\mathbf{x}_{i+1}), A_{i+1}\mathbf{x}_{i+1} - a_{i+1}\mathbf{v}_i - A_i\mathbf{y}_i \rangle.$$

Now, this means that $A_{i+1}(U_{i+1} - L_{i+1}) - A_i(U_i - L_i) \leq 0$ if

$$A_{i+1}\mathbf{x}_{i+1} - a_{i+1}\mathbf{v}_i - A_i\mathbf{y}_i = \mathbf{0} \text{ and } A_{i+1}/\beta \geq a_{i+1}^2/\sigma$$

We can get this by letting $\mathbf{x}_{i+1} = \frac{A_i\mathbf{y}_i + a_{i+1}\mathbf{v}_i}{A_{i+1}}$, and $\sigma = \beta$ and $a_i = \frac{i+1}{2}$, which implies that $A_i = \frac{(i+1)(i+2)}{4} < a_i^2$.

By Equation (8), these parameter choices also imply that

$$A_0(U_0 - L_0) \leq \frac{\beta}{2} \|\mathbf{x}^* - \mathbf{x}_0\|_2^2.$$

Finally, by induction, we get $A_i(U_i - L_i) \leq \frac{\beta}{2} \|\mathbf{x}^* - \mathbf{x}_0\|_2^2$. Dividing through by A_i and using $\text{gap}_i \leq U_i - L_i$ results in the following theorem.

Theorem 4.4. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a β -gradient Lipschitz, convex function. Let \mathbf{x}_0 be a given starting point, and let $\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$ be a minimizer of f .*

The Accelerated Gradient Descent algorithm given by

$$\begin{aligned} a_i &= \frac{i+1}{2}, A_i = \frac{(i+1)(i+2)}{4} \\ \mathbf{v}_0 &= \mathbf{x}_0 - \frac{1}{2\beta}\nabla f(\mathbf{x}_0) \\ \mathbf{y}_i &= \mathbf{x}_i - \frac{1}{\beta}\nabla f(\mathbf{x}_i) \end{aligned}$$

$$\begin{aligned}\mathbf{x}_{i+1} &= \frac{A_i \mathbf{y}_i + a_{i+1} \mathbf{v}_i}{A_{i+1}} \\ \mathbf{v}_{i+1} &= \mathbf{v}_i - \frac{a_{i+1}}{\beta} \nabla f(\mathbf{x}_{i+1})\end{aligned}$$

ensures that the k th iterate satisfies

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{4\beta \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{(k+1)(k+2)}.$$

References

- [DO19] Jelena Diakonikolas and Lorenzo Orecchia. Conjugate gradients and accelerated methods unified: The approximate duality gap view. *arXiv preprint arXiv:1907.00289*, 2019.
- [HS⁺52] Magnus R Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(6):409–436, 1952.
- [Lan52] Cornelius Lanczos. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards*, 49(1):33–53, 1952.
- [Nes83] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.
- [NY83] A Nemirovski and D Yudin. Information-based complexity of mathematical programming. *Izvestia AN SSSR, Ser. Tekhnicheskaya Kibernetika (the journal is translated to English as Engineering Cybernetics. Soviet J. Computer & Systems Sci.)*, 1, 1983.