Math 408 Nonlinear Optimization

Doug Kang

Winter 2022

My personal lecture notes for Math 408 Winter 22: Nonlinear Optimization, instructed by Dmitriy Drusvyatskiy at the University of Washington. Lecture notes are combination of transcription of the lecture given by the respective professor alongside my personal notes and additions. Textbook used for the class was *Introduction to Nonlinear Optimization* by Amir Beck.

Contents

1 January 3rd, 2022

1.1 Logistics

- Zoom lectures will be recorded but office hours will not.
- Once we go back to class (in person) nothing will be recorded.
- Piazza for discussion / questions, moderated by TA who will grade the homework.
- Three components to grading: HW (35%, Exam 1, Exam 2).
- Homeworks will be due Sunday night and posted / submitted through Gradescope.
- Late homework will be accepted through Tuesday but heavily penalized unless you have a documented emergency.
- Exam 1 will cover up through Feb 4th, and Exam 2 will be what we cover after. Exam 2 will not be cumulative (it will test what we cover after Exam 1).
- No makeup exams!

1.2 Introduction to Course

Although Math 407 (linear programming) is a prerequisite for this course, we will not use anything from linear programming in this course. In that sense, this course is not a continuation of Math 407. The main techniques for this class come from: linear algebra, multivariate calculus, and point-set topology. For the first week of this class we will review what we need from those topics in order.

1.3 Review (Chapter 1 in Textbook)

1.3.1 The Vector Space $\mathbb{R}^n$

$\mathbb{R}^n$ is the set of n-dimensional column vectors $x = (x_1, \dots, x_n)$ with real components, which is an ordered list of n numbers. There is the component-wise addition operation and the scalar-vector product.

Addition:

$$x + y = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

Scalar Multiplication:

$$\lambda x = \begin{bmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{bmatrix}$$

where $x_1, \dots, x_n, \lambda \in \mathbb{R}$.

Important Vectors:

The standard basis vectors of $\mathbb{R}^n$ and the zero vector:

$$\begin{aligned} e_1 &= (1,0,0,\dots,0) \\ e_2 &= (0,1,0,\dots,0) \\ &\vdots \\ e_n &= (0,0,0,\dots,0,1) \\ 0 &= (0,0,\dots,0). \end{aligned}$$

Important Subsets of $\mathbb{R}^n$:

The *nonnegative orthant* is the subset of $\mathbb{R}^n$ consisting of all vectors in $\mathbb{R}^n$ with nonnegative components and is denoted by $\mathbb{R}^n_+$:

$$\mathbb{R}^n_+ = \{(x_1,\dots,x_n) \mid x_1,\dots,x_n \geq 0\}.$$

The *positive orthant* consists of all the vectors in $\mathbb{R}^n$ with positive components and is denoted by $\mathbb{R}^n_{++}$:

$$\mathbb{R}^n_{++} = \{(x_1,\dots,x_n) \mid x_1,\dots,x_n > 0\}.$$

For $x, y \in \mathbb{R}^n$, the *closed line segment* between x and y is a subset of $\mathbb{R}^n$ denoted by $[x, y]$:

$$[x,y] = \{\lambda x + (1-\lambda)y \mid \lambda \in [0,1]\}.$$

The *unit-simplex* is the subset of $\mathbb{R}^n$ comprising all nonnegative vectors whose sum is 1:

$$\Delta_n = \{x \in \mathbb{R}^n_+ \mid x_1 + \cdots + x_n = 1\}.$$

One can consider the unit simplex as encoding a probability distribution on n points. Or as weights or proportions you put on n different objects. The *polyhedron* is defined as:

$$P = \{x \mid a_i^T x \leq b_i \ \forall i = 1,\dots,k\}$$

where $a_1,\dots,a-K \in \mathbb{R}^n$ and $b_1,\dots,b_k \in \mathbb{R}$.

1.3.2 The Vector Space $\mathbb{R}^{m\times n}$

The Vector Space $\mathbb{R}^{m\times n}$ is the set of all $m \times n$ (real) matrices.

Review: $A+B$, λA for $\lambda \in \mathbb{R}$, AB, Av, $\operatorname{tr}(A) = \sum_{i=1}^n A_{ii}$ if $A \in \mathbb{R}^{n\times n}$, $\det(A)$.

Important Subsets of $\mathbb{R}^{m\times n}$:

Symmetric Matrices are defined as:

$$S^n = \{A \in \mathbb{R}^{n\times n} \mid A^T = A\}.$$

Positive Semidefinite Matrices are defined as:

$$S_+^n = \{A \in S^n \mid x^T Ax \geq 0 \; \forall x \in \mathbb{R}^n\}.$$

Positive Definite Matrices are defined as:

$$S_{++}^n = \{A \in S^n \mid x^T Ax > 0 \; \forall x \in \mathbb{R}^n \setminus 0\}$$

In other words this means $A \in S_+^n \iff \inf_{x \in \mathbb{R}^n} x^T Ax = \sum_{i,j} A_{ij} x_i x_j \geq 0$. Or a quadratic function defined by the entries of A being nonnegative.

Notation: $A \in S_+^n \iff A \succeq 0$, $A \in S_{++}^n \iff A \succ 0$.

2 January 5th, 2022

2.1 The Vector Space $\mathbb{R}^{m\times n}$ (continued)

Last time we ended by talking about important matrices.

Example 2.1.1. How to tell if a matrix is positive semidefinite: $A \succeq 0$. Let

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}.$$

To check if $A \succeq 0$ we must check that $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$. We have

$$\begin{aligned} x^T Ax &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} x_1 + 2x_2 \\ 2x_1 + x_2 \end{bmatrix} \\ &= x_1(x_1 + 2x_2) + x_2(2x_1 + x_2) \\ &= x_1^2 + 2x_1x_2 + 2x_2x_1 + x_2^2. \end{aligned}$$

Notice that this is just the sum of the entries of A. In general we have

$$x^T Ax = \sum_{i,j} A_{ij}x_ix_j.$$

So in our example we have

$$x^T Ax = x_1^2 + 4x_1x_2 + x^2$$

and we want to know if this is ≥ 0 for all $x_1, x_2 \in \mathbb{R}$. It is not because let $x_1 = -1$ and $x_2 = 1$ and we have $x^T Ax = -2$. Hence, the matrix is not positive semidefinite.

What about the matrix

$$A = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}?$$

Then we have

$$\begin{aligned} x^T Ax &= 2x_1^2 + 2x_1x_2 + 2x_2x_1 + 2x_2^2 \\ &= 2x_1^2 + 4x_1x_2 + 2x_2^2 \\ &= 2(x_1 + x_2)^2 \geq 0. \end{aligned}$$

So in this case, A is positive semidefinite. But A is not positive definite since $x_1 = 1$ and $x_2 = -1$ gives us $x^T Ax = 0$.

Exercise: Check that

$$A = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

is positive definite. We have

$$x^T A x = 2(x_1 + x_2)^2 + 2x_1^2 + 2x_2^2.$$

Example 2.1.2. Is the following matrix (assume we flip across the diagonal) positive semidefinite?

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 7 \\ & 8 & 5 & 6 & 4 \\ & & 9 & 77 & \pi \\ & & & 52 & 17 \\ & & & & 81 \end{bmatrix}$$

This would get ugly because we are now working with a five dimensional matrix. We will find an easier way to check this later on.

There is one last set of matrices we'll need: *orthogonal matrices.*

$$O^n = \{A \in \mathbb{R}^{n \times n} \mid A^T A = I\}.$$

This implies that $AA^T = I$. $A^T A = I$ means

$$A_i^T A_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

2.2 Dot Product and Norms

Writing $x^T y$ is annoying. Instead define the *dot product* to be

$$\langle x, y \rangle := x^T y.$$

Basic Properties:

1. Symmetry: $\langle x, y \rangle = \langle y, x \rangle$.
2. Additivity: $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$.
3. Homogeneity: $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$ for $\lambda \in \mathbb{R}$.
4. Positive Definiteness: $\langle x, x \rangle \geq 0 \ \forall x$ and $\langle x, x \rangle = 0$ if and only if $x = 0$.

Definition 1. A *norm* $\| \cdot \|$ on $\mathbb{R}^n$ is a function on $\mathbb{R}^n$ satisfying

1. Nonnegativity: $\|x\| \geq 0 \ \forall x \in \mathbb{R}^n$ and $\|x\| = 0$ if and only if $x = 0$.
2. Positive Homogeneity: $\|\lambda x\| = |\lambda| \|x\| \ \forall x \in \mathbb{R}^n$ and $\forall \lambda \in \mathbb{R}$.
3. Triangle Inequality: $\|x + y\| \leq \|x\| + \|y\| \ \forall x, y \in \mathbb{R}^n$.

Example 2.2.1. Different norms. Let $x = (x_1, \ldots, x_n)$. Then we have the *Euclidean norm* or l_2 *norm*:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\langle x, x \rangle}.$$

We have the l_1 *norm*:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

We have the *infinity norm*:

$$\|x\| = \max_{i=1,\ldots,n} |x_i|.$$

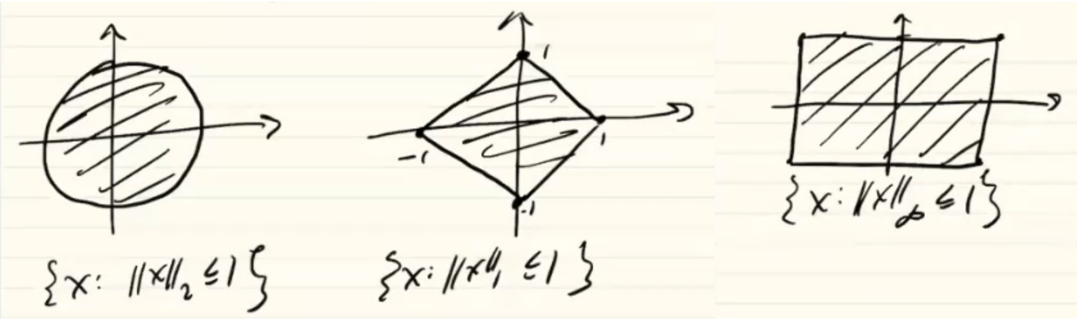
More generally, we have the l_p *norms* (for ≥ 1) defined by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

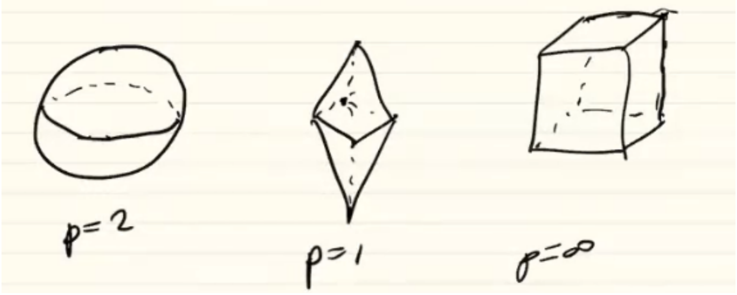
Aside: $\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$. Also, that the l_p norm satisfies the triangle inequality is called the Minkowsky inequality:

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p \right)^{1/p}.$$

One way to visualize norms is to draw the set of vectors x with $\|x\| \leq 1$.



In 3D we have:



Lemma 1. *(Cauchy-Schwartz) We have $\langle x, y \rangle \leq \|x\|_2 \cdot \|y\|_2$. Equality holds if and only if x and y are linearly independent.*

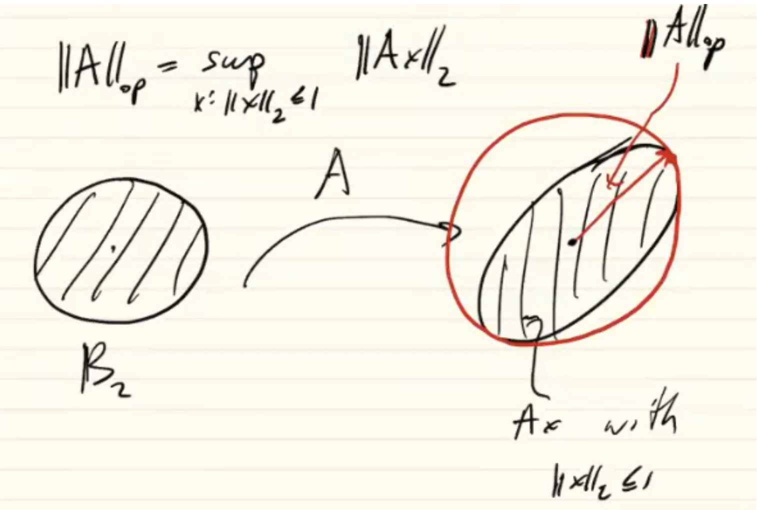
3 January 7th, 2022

3.1 Matrix Norms

Next we can think of the size of a matrix.

Definition 2. We define the *operator norm* of $A \in \mathbb{R}^{m\times n}$ as

$$\|A\|_{\text{op}} = \sup_{x:\|x\|_2\leq 1} \|Ax\|_2.$$


$\|A\|_{op} = \sup_{x:\|x\|_2 \leq 1} \|Ax\|_2$
$\|A\|_{op}$
A
B_2
Ax with $\|x\|_2 \leq 1$

So we apply A to every x in the unit ball and it becomes a deformed ellipse. Then, in that ellipse we choose the point that has the biggest norm. We are measuring the biggest amount by which we can increase the norm.

Remark 3.1.1. Note that $\|Ax\|_2 \leq \|A\|_{\text{op}}\|x\|_2$ $\forall x \in \mathbb{R}^n$. Why? Because $\|A(\frac{x}{\|x\|})\|_2 \leq \|A\|_{\text{op}}$. Then multiply by $\|x\|_2$.

The operator norm is a norm not on the matrix, but on the linear map induced. It does not depend on the basis chosen. In other words, it is not measuring the size of the matrix, but of the linear map.

Definition 3. The *Frobenius Norm* of $A \in \mathbb{R}^{m\times n}$ is defined as

$$\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2} = \|\text{vec}(A)\|_2.$$

You can check that: $\|A\|_F = \sqrt{\text{tr}(A^TA)}$.

3.2 Eigenvalue Decompositions

Definition 4. Let $A \in S^n$. Value $\lambda \in \mathbb{R}$ is an *eigenvalue* of A if $A - \lambda I$ is singular. Then any $0 \neq u \in Null(A - \lambda I)$ is called an eigenvector of A.

Note that $0 = (A - \lambda I)u \iff Au = \lambda u$. The eigenvalues of A are the roots of $p(\lambda) = \det(A - \lambda I)$ since $\det(A - \lambda I) = 0$ if and only if A is singular.

Theorem 1. *If $A \in S^n$ is symmetric, then the polynomial*

$$p(\lambda) = \det(A - \lambda I)$$

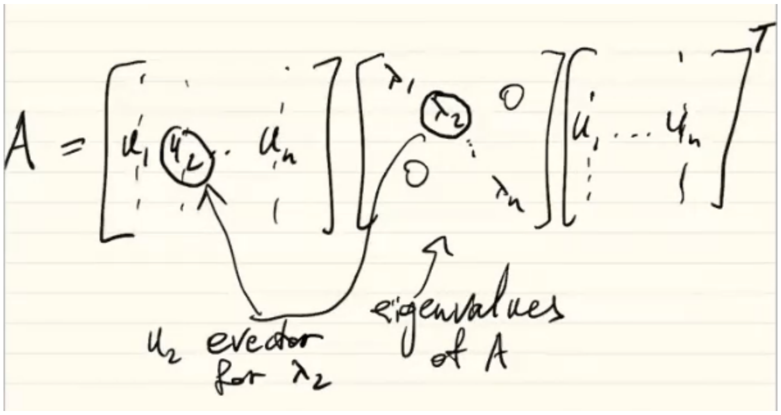
has exactly n real roots counting multiplicity and we denote them by

$$\lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_n(A).$$

So

$$p(\lambda) = (\lambda - \lambda_1(A))(\lambda - \lambda_2(A))(\ldots)(\lambda - \lambda_n(A)).$$

Theorem 2. *(Spectral Decomposition) Let $A \in S^n$. Then there exists a diagonal matrix $U \in \mathbb{R}^{n \times n}$ and a diagonal matrix $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ satisfying $A = U\Lambda U^T$.*


$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_n \end{bmatrix} \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix}^T$
u_2 evector for λ_2
eigenvalues of A

Remark 3.2.1.

$$\begin{aligned} \text{Tr}(A) &= \lambda_1(A) + \cdots + \lambda_n(A) \\ \det(A) &= \lambda_1(A) \cdot \lambda_2(A) \ldots \lambda_n(A). \end{aligned}$$

Remark 3.2.2.

$$\begin{aligned} \langle Bu, z \rangle &= (Bu)^T z = u^T B^T z \\ &= (u^T (B^T z)^T \\ &= (B^T z)^T u = \langle u, B^T z \rangle. \end{aligned}$$

Theorem 3. *(Rayleigh-Ritz) Let $A \in S^n$. Then*

$$\lambda_n(A)\|x\|_2^2 \leq \langle Ax, x\rangle \leq \lambda_1(A)\|x\|_2^2.$$

Equality holds for the corresponding eigenvectors.

Proof. Suppose x is an eigenvector corresponding to $\lambda_1(A)$. Then $Ax = \lambda_1(A) \cdot x$. So

$$\begin{aligned} \langle Ax, x\rangle &= \langle \lambda_1(A)x, x\rangle \\ &= \lambda_1(A)\langle x, x\rangle \\ &= \lambda_1(A)\|x\|_2^2. \end{aligned}$$

Same argument for equality on the left. Now let's prove the two inequalities for any $x \in \mathbb{R}^n$. Take a spectral decomposition $A = U\Lambda U^T$. Set $y = U^T x$. Then

$$\begin{aligned} \langle Ax, x\rangle &= \langle (U\Lambda U^T)x, x\rangle \\ &= \langle U(\Lambda U^T x), x\rangle \\ &= \langle \Lambda U^T x, U^T x\rangle \\ &= \langle \Lambda y, y\rangle \\ &= \sum_{i=1}^n \lambda_i y_i^2. \end{aligned}$$

Clearly

$$\lambda_n\|y\|^2 \leq \langle Ax, x\rangle \leq \lambda_1\|y\|_2^2.$$

We claim that $\|y\| = \|x\|$. We have

$$\begin{aligned} \|y\|_2^2 = \langle y, y\rangle &= \langle U^T x, U^T x\rangle \\ &= \langle x, (UU^T)x\rangle \\ &= \langle x, x\rangle = \|x\|_2^2. \end{aligned}$$

□

This completes our review of linear algebra.

4 January 10th, 2022

4.1 Topological Concepts

Definition 5. The *open ball* with center x and radius r is defined by

$$B(x, r) = \{y \in \mathbb{R}^n \mid \|y - x\|_2 < r\}.$$

The *closed ball* with center x and radius r is defined by

$$B[x, r] = \{y \in \mathbb{R}^n \mid \|y - x\|_2 \leq r\}.$$

Definition 6. A point $x \in U \subseteq \mathbb{R}^n$ is called an *interior point of* U if there exists $r > 0$ such that $B(x, r) \subseteq U$.

The set of all interior points of U is called the *interior of* U and is denoted by

$$\text{int}(U) = \{x \in U \mid x \text{ is an interior point of } U\}.$$

Example 4.1.1. Let's look at the interior of some sets.

1. $\text{int}(\mathbb{R}^n_+) = \mathbb{R}^n_{++}$
2. $\text{int}(B[x, r]) = B(x, r)$
3. $\text{int}(\Delta_n) = \varnothing$.

Definition 7. $U \subseteq \mathbb{R}^n$ is an *open set* if $U = \text{int}(U)$.

Example 4.1.2. Examples of open sets:

1. $B(x, r)$ is open.
2. $\mathbb{R}^n_{++}$ is open.

Basic properties of open sets:

1. Union of any number of open sets is open.
2. Intersection of finitely many open sets is open.

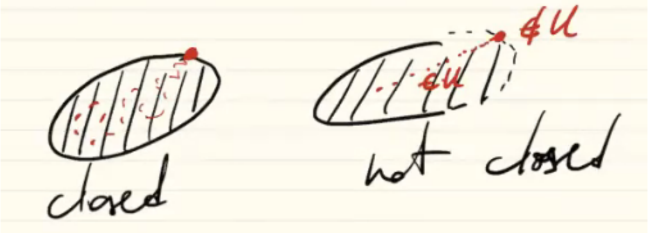
Definition 8. A set U is *closed* if $U^c := \{x \in \mathbb{R}^n \mid x \notin U\}$ is open.

Example 4.1.3. Examples of closed sets:

1. $B[x, r]$ is closed.
2. $\mathbb{R}^n_+$ is closed.
3. Δ^n is closed.

Note that we can have sets that are neither open nor closed. The next Lemma is a different way of thinking about closed sets.

Lemma 2. *U is closed if and only if for every sequence $\{x_i\}$ in U converging to some point x, it must be that $x \in U$.*



Definition 9. Given a set $U \in \mathbb{R}^n$, a *boundary point* of U is defined as

$$\mathrm{bd}(U) = \{x \in U \mid B(x, r) \cap U \neq \varnothing, B(x, r) \cap U^c \neq \varnothing \text{ for all } r > 0\}.$$

U is open if and only if $U \cap \mathrm{bd}(U) = \varnothing$. And U is closed if and only if $\mathrm{bd}(U) \subset U$.

Definition 10. The *closure* of a set $U \in \mathbb{R}^n$ is defined as

$$\mathrm{cl}(U) = U \cup \mathrm{bd}(U).$$

An equivalent definition of closure:

Lemma 3. *The closure of a set $U \in \mathbb{R}^n$ is the smallest closed set containing U:*

$$cl(U) = \bigcap \{V \mid U \subseteq V, V \text{ is closed}\}.$$

Example 4.1.4. Examples of closure:

1. $\mathrm{cl}(\mathbb{R}^n_{++}) = \mathbb{R}^n_{+}$.
2. $\mathrm{cl}(B(x, r)) = B[x, r]$.
3. $\mathrm{cl}((x, y)) = [x, y]$.

Definition 11.

- $U \subseteq \mathbb{R}^n$ is *bounded* if there exists $R > 0$ such that $U \subseteq B(0, R)$.
- $U \subseteq \mathbb{R}^n$ is *compact* if U is closed and bounded.

Definition 12. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous if $\lim_{y \to x} f(y) = f(x)$ for all x.*

I assume you know what it means for $\lambda_i \in \mathbb{R}$ to converge to some $\lambda \in \mathbb{R}$ (epsilon-delta definition from analysis). The notation $y \to x$ where $y \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$ means that the jth coordinate of y converges to the jth coordinate of x.

Proposition 1. *If f is continuous, then*

$$(\textit{level set })[f = r] = \{x \in \mathbb{R}^n \mid f(x) = r\} \tag{1}$$

$$(\textit{sublevel set })[f \leq r] = \{x \in \mathbb{R}^n \mid f(x) \leq r\} \tag{2}$$

are both closed.

Example 4.1.5. The set $\{(x, y) \mid \sin(xy)^2 + x + y = 1\}$ is closed since it is the level set of a continuous function.

Theorem 4. *(Bolzano-Weierstrass) Any continuous function $f : U \to \mathbb{R}$ defined on a compact set U attains its infimum and supremum.*

5 January 12th, 2022

5.1 Point-Set Topology (continued)

Recall that

$$\inf_x f(x) = \text{greatest valid lower bound for } f$$
$$\sup_x f(x) = \text{smallest valid upper bound for } f.$$

For example, we can look at the set of all lower bounds $L = \{r \mid r \leq f(x) \forall x\}$ which has the form $(-\infty, l)$ or $(-\infty, l]$ for some greatest lower bound l. The supremum is the same story but flipped.

Example 5.1.1. $f(x) = \frac{1}{x}$ on $(0, \infty)$. Then $L = (-\infty, 0]$ where $\inf_x f(x) = 0$.

Example 5.1.2. $f(x) = e^x$. Then $L = (-\infty, 0]$ where $\inf_x f(x) = 0$.

Example 5.1.3. $f(x) = x^2$. Then $L = (-\infty, 0]$ where $\inf_x f(x) = 0$.

We say that x is a minimizer of f if $f(\bar{x}) \leq f(x)\ \forall x$. If a minimizer $\bar{x}$ exists then $f(\bar{x}) = \inf_x f(x)$. Whereas $\inf_x f(x)$ is always defined, minimizers might not exist. For example, $f(x) = \frac{1}{x}$ has no minimizer. And $f(x) = e^x$ has no minimizer since any single value of x is not the smallest value that function takes. But $f(x) = x^2$ has minimizer 0.

If a function has no minimizer then we cannot do optimization. So the Bolzano-Weierstrass Theorem is useful because it tells us that if we restrict any continuous function to a closed and bounded interval (compact set), it has a minimizer, and hence the question is well posed.

5.2 Differentiability

The way calculus is taught is extremely old fashioned. I will try to correct some of the misunderstandings you were taught. The first misunderstanding is that in calc III, calculus is taught over two variables. That's not where calculus is used. Calculus is used in $\mathbb{R}^n$. The connection between linear algebra and calculus arises in true multivariate settings.

Let $f : U \to \mathbb{R}$ where $U \subseteq \mathbb{R}^n$ is open.

Definition 13. The *partial derivative* is defined as

$$\frac{\partial f}{\partial x_i}(x) = \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t}.$$

So equivalently, we take our function and restrict it along the ith coordinate axis. So we walk from x in the direction of the ith coordinate. And now the function is a univariate function since it's a function in one direction, and then we take the usual derivative.

Definition 14. The *gradient* is defined as

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix} \in \mathbb{R}^{n\times 1.}$$

Example 5.2.1. Let $f(x_1, x_2, x_3) = x_1^2 x_2 + x_3 x_1$. Then we have

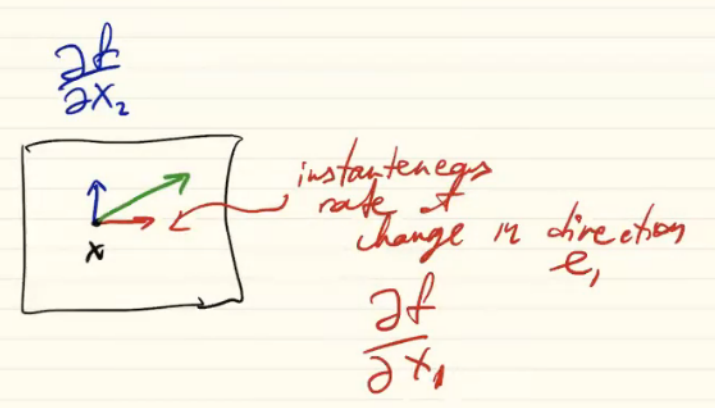
$$\nabla f(x) = \begin{bmatrix} 2x_1x_2 + x_3 \\ x_1^2 \\ x_1 \end{bmatrix} \in \mathbb{R}^{3\times 1}.$$

Definition 15. f is C^1*-smooth* if $\frac{\partial f}{\partial x_i}(x)$ exist and are continuous.

Definition 16. We define the *directional derivative* of x in direction v as

$$f'(x, v) = \lim_{t\to 0} \frac{f(x+tv) - f(x)}{t}.$$

In the below picture, the v vector representing the many other directions other than e_1, e_2 is represented by the green vector.


$\frac{\partial f}{\partial x_2}$
x
instantenegs rate of change in direction e_1
$\frac{\partial f}{\partial x_1}$

If f is C^1-smooth, then

$$f'(x, v) = \langle \nabla f(x), v\rangle = \sum_{i=1}^{n} v_i \cdot \frac{\partial f}{\partial x_i}(x).$$

Example 5.2.2. Returning to the same example, let $f(x_1, x_2, x_3) = x_1^2 x_2 + x_3 x_1$. We have

$$\nabla f(x) = \begin{bmatrix} 2x_1x_2 + x_3 \\ x_1^2 \\ x_1 \end{bmatrix} \in \mathbb{R}^{3\times 1}.$$

Then we can compute the directional derivative as

$$\begin{aligned} f'((x_1,x_2,x_3),(1,2,3)) &= \langle \nabla f(x_1,x_2,x_3),(1,2,3)\rangle \\ &= 1(2x_1x_2+x_3)+2(x_1)^2+3(x_1). \end{aligned}$$

The directional derivative is just a number representing the instantaneous rate of change of your function at a point in a certain direction.

Definition 17. Fix x, i, j and define $g(x) = \frac{\partial f}{\partial x_i}(x)$. Then we define second-order partial derivatives as

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(x) = \frac{\partial}{\partial x_i} g(x)$$

We use the Hessian matrix to encode all the second-order partial derivatives.

Definition 18. The *Hessian* of f at a point x is the $n \times n$ matrix denoted by $\nabla^2 f(x)$ with the i,jth entry defined by

$$[\nabla^2 f(x)]_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

Example 5.2.3. Returning to the same example, let $f(x_1,x_2,x_3) = x_1^2 x_2 + x_3 x_1$. We have

$$\nabla f(x) = \begin{bmatrix} 2x_1x_2 + x_3 \\ x_1^2 \\ x_1 \end{bmatrix} \in \mathbb{R}^{3\times 1}.$$

Then the Hessian is

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_3}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_3}(x) \\ \frac{\partial^2 f}{\partial x_3 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_3 \partial x_2}(x) & \frac{\partial^2 f}{\partial x_3 \partial x_3}(x) \end{bmatrix} = \begin{bmatrix} 2x_2 & 2x_1 & 1 \\ 2x_1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Notice that the matrix is symmetric, or that $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$, so the order of taking partial derivatives does not matter.

Definition 19. f is *C^2-smooth* if $\frac{\partial^2 f}{\partial x_i \partial x_j}$ exist and are continuous.

Theorem 5. *If f is C^2-smooth, then $\nabla^2 f(x)$ is symmetric. We have*

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

6 January 14th, 2022

6.1 Differentiability (continued)

Theorem 6. *If f is C^1-smooth, then*

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

The derivative can give us the best linear approximation of f at x. The best linear approximation is given by a new function:

$$g(y) = f(x) + \langle \nabla f(x), y - x \rangle.$$

Equivalently, we can let $y = x + h$ (the displacement vector) and we get

$$g(x+h) = f(x) + \langle \nabla f(x), h \rangle$$

So the key question is how good is our approximation? This is given by the difference between the function and the best linear approximation: $f(y) - g(y)$. So the Theorem is telling us that

$$\lim_{h \to 0} \frac{f(x+h) - g(x+h)}{\|h\|} = 0.$$

So the numerator tends to zero, but the question is how fast? Even if we divide this small number by another small number, the norm of displacement, the quotient goes to zero. So this means the numerator tends to zero faster than the norm of h, or any constant multiple of the norm of h. So informally, the error of approximation tends to zero faster than any linear function of the displacement vector h.

Fractions can be annoying so there is a different notation where we use little o notation.

$$\text{expression} = o(t) \text{ means } \lim_{t \to 0} \frac{\text{expression}}{t} = 0.$$

Then using this notation, we have

$$\begin{aligned} f(x+h) - f(x) - \langle \nabla f(x), h \rangle &= o(\|h\|) \\ \Rightarrow f(x+h) &= f(x) + \langle \nabla f(x), h \rangle + o(\|h\|). \end{aligned}$$

What this means is that the function value at the displaced point $(x+h)$ is equal to linear the prediction plus another error term.

Theorem 7. *(Mean Value Theorem) If f is C^2-smooth, then for any x, y there exists $z \in [x, y]$ such that*

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

Notice that C^1 smoothness only guaranteed an error term $o(\|h\|)$. What the Mean Value Theorem tells us is that you can have a very precise expression for the error term, which is a quadratic function of the displacement $y-x$. Quadratic functions are always little o. We can verify this. Is there a contradiction?

$$\lim_{y\to x}\frac{\langle\nabla^2 f(z)(y-x),y-x\rangle}{\|y-x\|}=0 \text{ (why?)}$$

Theorem 8. *(Taylor's Theorem) If f is C^2-smooth, then*

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

For the best linear approximation we know $f(y)-g(y)=o(\|y-x\|)$. We want a better approximation by using a higher order polynomial, and so we use a quadratic approximation. The best quadratic approximation adds to the best linear approximation a correction term given by the hessian. We have

$$\begin{aligned}q(y)&=f(x)+\langle\nabla f(x),y-x\rangle+\frac{1}{2}\langle\nabla^2 f(x)(y-x),y-x\rangle\\&=f(x)+\sum_{i=1}^n\frac{\partial f}{\partial x_i}(x)(y_i-x_i)+\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n\frac{\partial^2 f}{\partial x_i\partial x_j}(x)(y_i-x_i)(y_j-x_j).\end{aligned}$$

For the best quadratic approximation we know

$$\begin{aligned}f(y)-q(y)&=o(\|y-x\|^2)\\ \Rightarrow\lim_{y\to x}\frac{f(y)-q(y)}{\|x-y\|^2}&=0.\end{aligned}$$

In other words, the difference between f and q tends to zero faster than a quadratic function in the displacement vector.

Example 6.1.1. Returning to the same example, let $f(x_1,x_2,x_3)=x_1^2x_2+x_3x_1$. Let's form the best linear and quadratic approximations of f at x. Remember x is fixed, so let's use the base point $x=(1,2,3)$. From before, we have

$$\nabla f(x)=\begin{bmatrix}2x_1x_2+x_3\\x_1^2\\x_1\end{bmatrix}\in\mathbb{R}^{3\times 1}$$

and the Hessian as

$$\nabla^2 f(x)=\begin{bmatrix}\frac{\partial^2 f}{\partial x_1\partial x_1}(x)&\frac{\partial^2 f}{\partial x_1\partial x_2}(x)&\frac{\partial^2 f}{\partial x_1\partial x_3}(x)\\\frac{\partial^2 f}{\partial x_2\partial x_1}(x)&\frac{\partial^2 f}{\partial x_2\partial x_2}(x)&\frac{\partial^2 f}{\partial x_2\partial x_3}(x)\\\frac{\partial^2 f}{\partial x_3\partial x_1}(x)&\frac{\partial^2 f}{\partial x_3\partial x_2}(x)&\frac{\partial^2 f}{\partial x_3\partial x_3}(x)\end{bmatrix}=\begin{bmatrix}2x_2&2x_1&1\\2x_1&0&0\\1&0&0\end{bmatrix}.$$

The best linear approximation is given by

$$\begin{aligned}
g(y) &= f(x) + \langle \nabla f(x), y - x \rangle \\
&= (1^2 \cdot 2 + 3 \cdot 1) + \left\langle \begin{bmatrix} 2 \cdot 1 \cdot 2 + 3 \\ 1^2 \\ 1 \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right\rangle \\
&= 5 + \left\langle \begin{bmatrix} 7 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} y_1 - 1 \\ y_2 - 2 \\ y_3 - 3 \end{bmatrix} \right\rangle \\
&= 5 + 7(y_1 - 1) + 1(y_2 - 2) + 1(y_3 - 3).
\end{aligned}$$

The best quadratic approximation is given by

$$\begin{aligned}
q(y) &= g(y) + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle \\
&= 5 + 7(y_1 - 1) + 1(y_2 - 2) + 1(y_3 - 3) + \frac{1}{2} \left\langle \begin{bmatrix} 2x_2 & 2x_1 & 1 \\ 2x_1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 - x_1 \\ y_2 - x_2 \\ y_3 - x_3 \end{bmatrix}, \begin{bmatrix} y_1 - x_1 \\ y_2 - x_2 \\ y_3 - x_3 \end{bmatrix} \right\rangle \\
&= 5 + 7(y_1 - 1) + 1(y_2 - 2) + 1(y_3 - 3) + \frac{1}{2} \left\langle \begin{bmatrix} 4 & 2 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 - 1 \\ y_2 - 2 \\ y_3 - 3 \end{bmatrix}, \begin{bmatrix} y_1 - 1 \\ y_2 - 2 \\ y_3 - 3 \end{bmatrix} \right\rangle \\
&= 5 + 7(y_1 - 1) + 1(y_2 - 2) + 1(y_3 - 3) + \frac{1}{2} \left\langle \begin{bmatrix} 4(y_1 - 1) + 2(y_2 - 2) + (y_3 - x_3) \\ 2(y_1 - 1) \\ 1(y_1 - x_1) \end{bmatrix}, \begin{bmatrix} y_1 - 1 \\ y_2 - 2 \\ y_3 - 3 \end{bmatrix} \right\rangle \\
&= 7y_1 + y_2 + y_3 - 7 + \\
&\frac{1}{2} \left([4(y_1 - 1 + 2(y_2 - 2) + (y_3 - x_3)](y_1 - 1) + 2(y_1 - 1)(y_2 - 2) + (y_1 - 1)(y_3 - 3) \right).
\end{aligned}$$

And that concludes the review! So where do we go from here? Our first order of business will be to write down optimality conditions for being a minimizer. Let's have a dialogue: How do we check in calc 1 if a point is a minimizer or not? We take the derivative and set it to zero. If it is not zero, then we conclude that the point is not a minimizer. If it is zero, then you take the second derivative. If the second derivative is strictly positive then you conclude it is a local minimizer. If it is negative, then it is a local maximizer. If it is zero then you are out of luck. In the multivariate setting, there is an analogue of all these statements. The derivative is know the gradient, so we must check if all partial derivatives are zero. If the answer is no, then we know the point is not a local minimizer. If the answer is yes, then we go to the second derivative, which is no longer a number, but the Hessian, a matrix of second-order partial derivatives. The right notion of positivity is now: is the matrix positive definite. If the matrix is positive definite, the point is a local minimizer. If it is not positive definite then there are possibilities. If the minimum eigenvalue is strictly negative than it is not a local minimizer. If the minimum eigenvalue is zero then you are out of luck.

However, this is not what optimization is, since nobody gives you a point and asks you if this is a minimizer. After all, why would you randomly have a point that is a supposed minimizer? A Google executive comes to you and says “find the minimizer of this function.” So checking optimality is a much easier thing to do than finding a candidate minimizer. However, it turns out that the two viewpoints are linked. If you have a certain candidate solution then two things can happen: you either certify that the point is a minimizer or you certify that it is not a minimizer. But from these optimality conditions you get more information. If the optimality conditions fail, they fail because you can learn and generate a point that is better than the point you checked. So optimality conditions are not where things stop, its where they start. If optimality conditions fail, we will use this information to develop algorithms that try to find minimizers, and this is basically what the course is.

7 January 19th, 2022

7.1 Optimality Necessary Conditions

Definition 20. Let $f : S \to \mathbb{R}$ where $S \subseteq \mathbb{R}^n$.

1. $\bar{x}$ is a *global minimizer* of f over S if $f(x) \geq f(\bar{x})$ $\forall x \in S$. Then $f(\bar{x})$ is called the *minimal value.*
2. $\bar{x}$ is a *strict global minimizer* of f over S if $f(x) > f(\bar{x})$ $\forall x \in S, x \neq \bar{x}$.

Whenever we use the word minimizer, we mean global minimizer (will not say global every time). We will primarily be happy with finding local minimizers.

Definition 21. Let $f : S \to \mathbb{R}$ where $S \subseteq \mathbb{R}^n$.

1. $\bar{x}$ is a *local minimizer* of f over S if there exists $r > 0$ such that $f(x) \geq f(\bar{x})$ $\forall x \in S \cap B(\bar{x}, r)$.
2. $\bar{x}$ is a *strict local minimizer* of f over S if there exists $r > 0$ such that $f(x) > f(\bar{x})$ $\forall x \in S \cap B(\bar{x}, r)$ and $x \neq \bar{x}$.

How can we check local optimality using derivatives?

Theorem 9. *(First-order necessary conditions) Let $\bar{x}$ be a local minimizer of $f : \mathbb{R}^n \to \mathbb{R}$. If f is differentiable at $\bar{x}$, the $\nabla f(\bar{x}) = 0$.*

Proof. Fix $i \in \{1, \ldots, n\}$. Define $\phi(t) = f(\bar{x} + te_i)$. Clearly $t = 0$ is a local minimizer of ϕ. Then since $\phi'(0) = 0$ we have

$$\phi'(0) = \lim_{t \to 0} \frac{\phi(t) - \phi(0)}{t} = \lim_{t \to 0} \frac{f(\bar{x} + te_i) - f(\bar{x})}{t} = \frac{\partial}{\partial x_i} f(\bar{x}_i) = 0.$$

□

Proof. Take $v \in \mathbb{R}^n$. Define $\phi(t) = f(\bar{x} + tv)$. Again $t = 0$ is a local minimizer of ϕ. We have

$$0 = \phi'(0) = \lim_{t \to 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t} = f'(\bar{x}, v) = \langle \nabla f(\bar{x}), v \rangle.$$

So the directional derivative in every direction is 0. Take $v = -\nabla f(\bar{x})$. Then we have

$$0 = \langle \nabla f(\bar{x}), -\nabla f(\bar{x}) \rangle = -\|\nabla f(\bar{x})\|^2$$

which implies that $\nabla f(\bar{x}) = 0$. □

Let's think a little... take any $\bar{x} \in \mathbb{R}^n$. Set $v = -\nabla f(\bar{x})$. We just found that

$$f'(\bar{x}, v) = \langle \nabla f(\bar{x}), -\nabla f(\bar{x}) \rangle = -\|\nabla f(\bar{x})\|^2.$$

In particular, if the directional derivative is zero, then we know the gradient is zero. However, what if $\bar{x}$ is not a local minimizer. Then we learn that

$$f'(\bar{x}, v) = \lim_{t \to 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t} < 0.$$

So the limit is a negative number, and so the numerator is a negative number.

Punch line: If $\bar{x}$ is a local minimizer, then $\nabla f(\bar{x}) = 0$. Otherwise, $f(\bar{x} - t\nabla f(\bar{x})) < f(\bar{x})$ for all small $t > 0$. This suggests an algorithm. If we walk a little bit in the direction given by minus the gradient, the function value goes down.

Definition 22. $\bar{x}$ is *critical* or a *stationary point* of a differentiable $f : \mathbb{R}^n \to \mathbb{R}$ if $\nabla f(\bar{x}) = 0$.

Example 7.1.1. $f(x) = x^2$ vs $f(x) = -x^2$. Both functions have $f'(0) = 0$. We don't know if it is a local minimizer.

Theorem 10. *(Second-order conditions) Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^2-smooth.*

1. *(necessary) If $\bar{x}$ is a local minimizer of f, then $\nabla f(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \succeq 0$.*
2. *(sufficient) If $\nabla f(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \succ 0$ then $\bar{x}$ is a (strict) local minimizer of f.*

Proof. Suppose $\bar{x}$ is a local minimizer. We already know $\nabla f(\bar{x}) = 0$. Fix $v \in \mathbb{R}^n$. We have the following Taylor expansion:

$$f(\bar{x} + tv) = f(\bar{x}) + \langle \nabla f(\bar{x}), tv \rangle + \frac{1}{2} \langle \nabla^2 f(\bar{x} tv, tv \rangle + o(\|tv\|^2).$$

Since we know $\nabla f(\bar{x}) = 0$ we have $\langle \nabla f(\bar{x}), tv \rangle = 0$. Then by subtracting $f(\bar{x})$ and dividing by t^2 we have

$$0 \leq \frac{f(\bar{x} + tv) - f(\bar{x})}{t^2} = \frac{1}{2} \langle \nabla^2 f(\bar{x}) v, v \rangle + \frac{o(\|tv\|^2)}{t^2}$$

where we know $0 \leq \frac{f(\bar{x}+tv) - f(\bar{x})}{t^2}$ $\forall$ small $t > 0$ since $\bar{x}$ is a local minimizer. Now let $t \to 0$. Since the term $\frac{1}{2}\langle \nabla^2 f(\bar{x}) v, v \rangle$ does not depend on zero, we just have

$$0 \leq \frac{1}{2} \langle \nabla^2 f(\bar{x}) v, v \rangle + \lim_{t \to 0} \frac{o(\|tv\|^2)}{t^2}.$$

We claim that the limit of the last term goes to zero since

$$\lim_{t \to 0} \frac{o(\|tv\|^2)}{t^2} \Rightarrow \lim_{t \to 0} \frac{o(\|tv\|^2)}{\|tv\|^2} \cdot \|v\|^2 \Rightarrow \|v\|^2 \cdot \lim_{s \to 0} \frac{o(s)}{s} = 0$$

by the definition of little o notation. Hence, we know that $0 \leq \langle \nabla^2 f(\bar{x}) v, v \rangle$ $\forall v$ which implies that $\nabla^2 f(\bar{x}) \succeq 0$ as desired. □

8 January 21st, 2022

8.1 Optimality Sufficient Conditions

We continue with the proof to the second part, or the sufficient second order optimality condition.

Proof. Let $\nabla f(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \succ 0$. We claim there exists $\varepsilon > 0$ such that $\nabla^2 f(x) \succ 0$ $\forall x \in B_\varepsilon(\bar{x})$, or that $\bar{x}$ is a strict local minimizer of f.

Suppose otherwise. Then there is a sequence $x_i \to \bar{x}$ and $\nabla^2 f(x_i)$ is not positive definite. So there is $v_i \neq 0$ such that $v_i^T \nabla^2 f(x_i) v_i \leq 0$. Let's replace v_i by $\frac{v_i}{\|v_i\|}$. But now since v_i are all on a unit sphere, they admit a limit point $\bar{v}$ of unit norm. (Since the sphere is compact and Bolzano-Weierstrass theorem tells us that every bounded sequence has a convergent subsequence). So taking a subsequence we may assume $v_i \to \bar{v}$. So $0 \geq \lim_{i \to 0} v_i^T \nabla^2 f(x_i) v_i = \bar{v}^T \nabla^2 f(\bar{x}) \bar{v} > 0$ where the first inequality is by assumption and the latter equality is by C^2-smoothness. This is a contradiction. Therefore the claim is correct.

Now take $\varepsilon > 0$ such that $\nabla^2 f(x) \succ 0$ $\forall x \in B_\varepsilon(\bar{x})$. Then for any $x \in B_\varepsilon(\bar{x})$ by the Mean Value Theorem, there exists $z \in [x, y]$ such that $f(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{1}{2} \langle \nabla^2 f(z)(x - \bar{x}), x - \bar{x} \rangle$. So the claim gives us this connection telling us how the function value deviates form its linearization by a second order term given by the Hessian at a midpoint. But since we assumed that $\nabla f(\bar{x}) = 0$ we know that the second inner product term $\langle \nabla f(\bar{x}), x - \bar{x} \rangle = 0$ and since we assumed that $\nabla^2 f(x) \succ 0$ we know that the last inner product term is greater than zero. So $f(x)\ f(\bar{x})$ $\forall x \in B_\varepsilon(\bar{x})$ and therefore $\bar{x}$ is a strict local minimizer of f. □

So if someone hands you a point $\bar{x}$ and asks you if it is a local minimizer, you check that the gradient is equal to zero and the hessian is positive semidefinite. The latter part is the hard part, so next we will develop more tools to recognize psd matices.

8.2 Recognizing Positive Semidefinite (PSD) Matrices

There are two main tools. The first is based on eigenvalues ($\lambda_n(A)$) is the minimal eigenvalue of A.

Theorem 11.

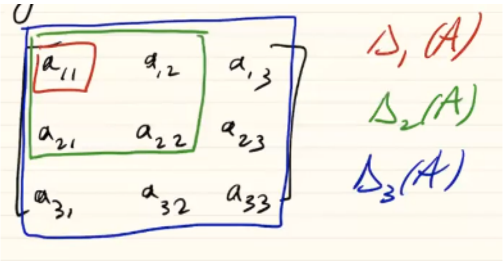
$$A \succeq 0 \iff \lambda_n(A) \geq 0$$
$$A \succ 0 \iff \lambda_n(A) > 0.$$

Corollary 1. *A is PSD and not PD* $\iff \lambda_n(A) = 0 \iff A \succeq 0$ *and A is singular.*

In other words, the distinction between psd and pd is that you might have a positive semidefinite matrix with a nullspace. If that happens the matrix is positive semidefinite, but not positive definite.

The second approach is based on principal minors.

Definition 23. If $A \in \mathbb{R}^{n\times n}$, then the determinant of the upper $k \times k$ submatrix of A is the *kth principal minor*, and is denoted by $\Delta_k(A)$.



Theorem 12. *$A \in S^n$ is positive definite if and only if* $\Delta_1(A) > 0, \Delta_2(A) > 0, \dots, \Delta_n(A) > 0$.

Note: Analagous statement fails for checking PSD. As an example:

$$A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}.$$

Then $\Delta_1(A) = \Delta_2(A) = 0$ but A is not PSD since

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}^T A \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} 0 \\ -1 \end{bmatrix} = -1.$$

Example 8.2.1. Let

$$A = \begin{bmatrix} 4 & 2 & 3 \\ 2 & 3 & 2 \\ 3 & 2 & 4 \end{bmatrix}.$$

Is A positive definite? We have

$$\begin{aligned} \Delta_1 &= \det[4] = 4 > 0 \\ \Delta_2 &= \det \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix} = 8 > 0 \\ \Delta_3 &= \det(A) = 13 > 0. \end{aligned}$$

So we conclude that $A \succ 0$. Note that this strategy only works for positive definiteness. If you needed to check positive semi-definiteness, you would need to compute this eigenvalues.

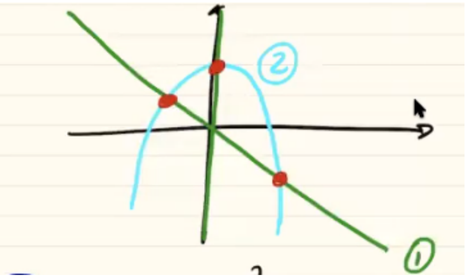
Example 8.2.2. Consider $f(x_1, x_2) = 2x_1^3 + 3x_2^2 + 3x_1^2 x_2 - 24x_2$. Let's classify the critical points of f and classify them as local minimizers or not. We have

$$\nabla f(x_1, x_2) = \begin{bmatrix} 6x_1^2 + 6x_1 x_2 \\ 6x_2 + 3x_1^2 - 24. \end{bmatrix}$$

Set $\nabla f(x_1, x_2) = 0$ and solve for x_1, x_2. In general this is not always an easy problem as you are solving a nonlinear set of equations, however in this constructed example it is easy. This is the system:

$$\begin{cases} 6x_1^2 + 6x_1x_2 = 0 \\ 6x_2 + 3x_1^2 - 24 = 0 \end{cases} \Rightarrow \begin{cases} x_1^2 + x_1x_2 = 0 \\ 2x_2 + x_1^2 - 8 = 0 \end{cases}$$

The first equation can be factored as $x_1(x_1+x_2) = 0$ which implies that $x_1 = 0$ or $x_1 = -x_2$. The second equation gives us $x_2 = -\frac{1}{2}x_1^2 + 4$. Here is a graph where the red points are the critical points.



Now we just need to compute the coordinates of the red points.

Points of Intersection: When $x_1 = 0$ we have $x_2 = 4$ so one of the points is (0,4). When $x_1 = -x_2$, by plugging that into the second equation we have $x_2 = -\frac{1}{2}x_2^2 + 4$ which implies that $x_2^2 + 2x_2 - 8 = 0 \Rightarrow (x_2 + 4)(x_2 - 2) = 0 \Rightarrow x_2 = 2$ or $x_2 = -4$. Hence the other two points are $(-2, 2), (4, -4)$.

Next we need to compute the Hessian. We have

$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} 12x_1 + 6x_2 & 6x_1 \\ 6x_1 & 6 \end{bmatrix} = 6 \begin{bmatrix} 2x_1 + x_2 & x_1 \\ x_1 & 1 \end{bmatrix}.$$

The relevant matrices we must check for positive semidefiniteness are attaiend by plugging in the three points of intersection to the Hessain. For the points $(0, 4), (-2, 2), (4, -4)$ respectively we have

$$\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -2 & -2 \\ -2 & 1 \end{bmatrix}, \begin{bmatrix} 4 & 4 \\ 4 & 1 \end{bmatrix}.$$

For the first matrix we have

$$\det(4) = 4 > 0$$
$$\det(\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}) = 4 > 0$$

so the matrix is positive definite. For the second matrix we have $\det(-2) = -2 < 0$ so the matrix is not positive definite. For the last matrix we have

$$\det(4) = 4 > 0$$
$$\det(\begin{bmatrix} 4 & 4 \\ 4 & 1 \end{bmatrix}) = -12 < 0$$

so the matrix is not positive definite. So we know that the point $(0, 4)$ is a local minimizer. For the second matrix although it is not positive definite, it could be positive semidefinite, and you would have to compute the eigenvalues. If you do this the eigenvalues show that the matrix is not positive semidefinite. Similarly, for the third matrix the eigenvalues show that the matrix is not positive semidefinite.

Is the point $(0, 4)$ a global minimizer of $f(x_1, x_2)$? In general, this is a hard question and there is no easy recipe to check. It must be done ad hoc. However there will be certain sufficient conditions. But in this example, from ad hoc analysis we can plug in $x_2 = 0$ and we get $f(x_1, 0) = 2x_1^3$. So along the x_1 axis the function is a simple cubic, and we know cubics are not bounded from below. So $f(-k, 0) = -2k^3 \to -\infty$ as $k \to \infty$. So if we walk along in the negative direction the function is not bounded from below, so it has no global minimizer.

9 January 24th, 2022

9.1 Optimality Conditions (continued)

Definition 24. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^2-smooth function. We say that f is *convex* if $\nabla^2 f(x) \succeq 0 \ \forall x \in \mathbb{R}^n$.

Convexity is the multivariate analogue of concave up in lower dimensions. As a review, there are several equivalent conditions for concave up.

- $f'(x)$ is increasing.
- $f''(x) \geq 0 \ \forall x$.
- Tangent lines are below the graph.
- $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$ secant lines are above the graph.

We will return to the multivariate analogues of these definitions, but for now we will use the fact that a function that is C^2-smooth is convex if the Hessian is positive semidefinite everywhere.

Theorem 13. *Suppose f is C^2-smooth and $\nabla^2 f(x) \succeq 0 \ \forall x$. Then the following are equivalent.*

1. *$\bar{x}$ is a global minimizer of f.*
2. *$\bar{x}$ is a local minimizer of f.*
3. *$\bar{x}$ is a critical point of f.*

Proof. Clearly, $1 \implies 2 \implies 3$ done. Let us verify $3 \implies 1$. Suppose $\bar{x}$ is a critical point of f. Then $\nabla f(\bar{x}) = 0$ by definition. Take any $x \in \mathbb{R}^n$. Then by the Mean Value Theorem there is $z \in [\bar{x}, x]$ such that $f(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{1}{2} \langle \nabla^2 f(z)(x - \bar{x}), x - \bar{x} \rangle$. Then since $\nabla f(\bar{x}) = 0$ the middle term is equal to zero. And since $\nabla^2 f(x) \succeq 0 \ \forall x$ (f is convex), we know that the last term is ≥ 0. Therefore, we know that $f(\bar{x}) \leq f(x) \ \forall x$, so $\bar{x}$ is a global minimzier of f. □

Example 9.1.1. (Univariate Convex Functions)

- $f(x) = \frac{1}{2}x^2$.
- $f(x) = e^x$.
- $f(x) = e^x + \frac{1}{2}x^2$.
- $f(x) = -\sqrt{x}$ on $(0, \infty)$.
- $f(x) = \ln(1 + e^{-x})$. Check this!

Sometimes weird functions are convex. Ex 2.39 in the book. $f(x) = \|x\|^2 + \|x\|^4 + x_1x_2 + x_1x_3 + x_2x_3$ is convex.

10 January 26th, 2022

10.1 Coercive Functions

Recall: Bolzano-Weierstrass guarantees that continuous $f : S \to \mathbb{R}$ on a compact set S has a global minimizer and global maximizer. What if the underlying set is not compact and f is unbounded?

Definition 25. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function defined over $\mathbb{R}^n$. The function f is called *coercive* if for any x_i with $\|x_i\| \to +\infty$ it holds that $\lim\limits_{i\to\infty} f(x_i) = +\infty$.

Basically the function is a giant infinite bowl or valley.

Example 10.1.1. Some examples of coercive / not coercive functions.

- $f(x) = 2$ is not coercive.
- $f(x) = x^2$ is coercive.
- $f(x) = x^3$ is not coercive.
- $f(x) = e^x$ is not coercive.

Theorem 14. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous and coercive. Then f attains its infimum.*

Proof. Take any $x_0 \in \mathbb{R}^n$. Define $L = \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$.

Claim: $\inf_x f(x) = \inf_L f(x)$. Clearly $\leq$ is true. Conversely let $u = \inf_x f(x)$. By the definition of inf, $\exists x_i \in \mathbb{R}^n$ such that $f(x_i) \to u$. Note $u \leq f(x_0)$ so $x_i \in L$ also.

Claim: L is bounded. Suppose not. So L is unbounded. So there exists $x_i \in L$ such that $\|x_i\| \to +\infty$. This is a contradiction since $\lim\limits_{i\to\infty} f(x_i) = +\infty \leq f(x_0)$ where the first equality is by the definition of coercivity. □

Theorem 15. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous and coercive and let $S \subseteq \mathbb{R}^n$ be a closed set. Then f attains its infimum over S.*

Proof. The proof is identical if you take $x_0 \in S$. □

Continuous, coercive functions attain their minimizers over any closed set including $\mathbb{R}^n$.

10.2 Quadratic Functions

Definition 26. A *quadratic function* over $\mathbb{R}^n$ is of the form:

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c$$

where $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n, c \in \mathbb{R}$. In coordinates:

$$f(x_1, \dots, x_n) = \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}x_ix_j + \sum_{i=1}^{n} b_ix_i + c.$$

Example 10.2.1. We have the following quadratic function

$$\begin{aligned} f(x_1, x_2) &= 17x_1x_2 + 14x_1^2 + 4x_2 + 7 \\ &= \frac{1}{2}(34x_1x_2 + 28x_1^2) + 4x_2 + 7 \\ &= \frac{1}{2}x^TAx + b^Tx + c \end{aligned}$$

where

$$A = \begin{bmatrix} 28 & 17 \\ 17 & 0 \end{bmatrix}, b = \begin{bmatrix} 0 \\ 4 \end{bmatrix}, c = 7.$$

You can divide the coefficient of x_ix_j however you want but putting it in half is ideal so that we get a symmetric matrix.

Lemma 4. *$x^TAx = x^T(\frac{A+A^T}{2})x \ \forall x$.*

Proof. We have

$$\begin{aligned} x^TAx &= \frac{1}{2}x^TAx + \frac{1}{2}x^TAx \\ &= \frac{1}{2}x^TAx + \frac{1}{2}(x^TAx)^T \\ &= \frac{1}{2}x^TAx + \frac{1}{2}x^TA^Tx \\ &= x^T\left(\frac{A+A^T}{2}\right)x. \end{aligned}$$

□

Example 10.2.2. So in our last example we could have instead picked A to be

$$A = \begin{bmatrix} 28 & 33 \\ 1 & 0 \end{bmatrix}.$$

But then we have

$$\frac{A+A^T}{2} = \frac{\begin{bmatrix} 28 & 33 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 28 & 1 \\ 33 & 0 \end{bmatrix}}{2} = \begin{bmatrix} 28 & 17 \\ 17 & 0 \end{bmatrix}.$$

So $\frac{1}{2}x^TAx + b^Tx + c = \frac{1}{2}x^T(\frac{A+A^T}{2})x + b^Tx + c$ $\forall x$. Note that $\frac{A+A^T}{2}$ is symmetric since $(\frac{A+A^T}{2})^T = \frac{A^T+A}{2} = \frac{A+A^T}{2}$.

Punchline: By replacing A with $\frac{A+A^T}{2}$ we can always assume A is symmetric when writing $f(x) = \frac{1}{2}x^TAx + b^Tx + c$.

Lemma 5. *Define* $f(x) = \frac{1}{2}x^TAx + b^Tx + c$ *where* $A \in S^n$. *Then*

$$\begin{aligned}\nabla f(x) &= Ax + b\\ \nabla^2 f(x) &= A.\end{aligned}$$

Proof. We have

$$f(x_1, \ldots, x_n) = \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}x_ix_j + \sum_{i=1}^{n} b_ix_i + c.$$

Then we have

$$\frac{\partial f}{\partial x_k}(x) = (k\text{th entry of } Ax) + b_k.$$

□

Proof. (Alternate proof). Recall that the directional derivative is

$$\lim_{t\to 0}\frac{f(x+tv) - f(x)}{t} = f'(x,v) = \langle \nabla f(x), v\rangle.$$

Then we have

$$\begin{aligned}f(x+tv) &= \frac{1}{2}(x+tv)^TA(x+tv) + b^T(x+tv) + c\\ &= \frac{1}{2}(x^TAx + tx^TAv + tv^TAx + t^2v^TAv) + b^Tx + tb^Tv + c\\ &= \frac{1}{2}x^TAx + b^Tx + c + \frac{1}{2}tx^TAv + tv^TAx + \frac{1}{2}t^2v^TAv + t^2b^Tv\\ &= f(x) + t(x^TAv + b^Tv) + \frac{t}{2}^2 v^TAv.\end{aligned}$$

Then we have

$$\begin{aligned}\lim_{t\to 0}\frac{f(x+tv) - f(x)}{t} &= \lim_{t\to 0}(x^TAv + b^Tv) + \frac{t}{2}v^TAv\\ &= x^TAv + b^Tv\\ &= (Ax+b)^Tv \ \forall v\end{aligned}$$

since $\lim_{t\to 0}\frac{t}{2}v^TAv \to 0$. So $\nabla f(x)^Tv = (Ax+b)^Tv$ $\forall v$ implies that $0 = (\nabla f(x) - (Ax+b))^Tv$ $\forall v$, which gives us $\nabla f(x) - (Ax+b) = 0$, and so $\nabla f(x) = Ax + b$ as desired. □

11 January 28th, 2022

11.1 Midterm Exam Logistics

- The midterm will be very similar to the homeworks (first two).
- There will be short proofs (possibly), but there will be lots of computations like in the homework.
- There also might be questions on definitions.

11.2 Quadratic Functions (continued)

Theorem 16. $A \succeq 0 \iff$ *there exists a lower triangular matrix* L *such that* $A = LL^T$.

In general, any matrix $A = LL^T$, lower triangular or not, must be positive semidefinite. Why? We have

$$x^T(LL^T)x = (L^Tx)^T(L^Tx) = \|L^Tx\|^2 \geq 0.$$

But we can take L to be a lower triangular matrix, and the process of forming L is called a Cholesky Factorization. While you try to form L, there is a check that characterizes if A is not positive semidefinite. Now we will return to the Lemma we left off with last time and continue the proof.

Lemma 5. *Define* $f(x) = \frac{1}{2}x^TAx + b^Tx + c$ *where* $A \in S^n$. *Then*

$$\begin{aligned}\nabla f(x) &= Ax + b\\ \nabla^2 f(x) &= A.\end{aligned}$$

Proof. (First, another way of showing $\nabla f(x) = Ax + b$). Recall that the directional derivative is

$$\lim_{t\to 0} \frac{f(x+tv) - f(x)}{t} = f'(x, v) = \langle \nabla f(x), v\rangle.$$

Then we have

$$\begin{aligned}f(x+tv) &= \frac{1}{2}(x+tv)^TA(x+tv) + b^T(x+tv) + c\\ &= \frac{1}{2}(x^TAx + tx^TAv + tv^TAx + t^2v^TAv) + b^Tx + tb^Tv + c\\ &= \frac{1}{2}x^TAx + b^Tx + c + \frac{1}{2}tx^TAv + tv^TAx + \frac{1}{2}t^2v^TAv + t^2b^Tv\\ &= f(x) + t(x^TAv + b^Tv) + \frac{t^2}{2}v^TAv.\end{aligned}$$

Then we have

$$\begin{aligned}\lim_{t\to 0}\frac{f(x+tv)-f(x)}{t} &= \lim_{t\to 0}(x^TAv+b^Tv)+\frac{t}{2}v^TAv\\ &= x^TAv+b^Tv\\ &= (Ax+b)^Tv\ \forall v\end{aligned}$$

since $\lim_{t\to 0}\frac{t}{2}v^TAv \to 0$. So $\nabla f(x)^Tv = (Ax+b)^Tv\ \forall v$ implies that $0 = (\nabla f(x)-(Ax+b))^Tv$ $\forall v$, which gives us $\nabla f(x)-(Ax+b)=0$, and so $\nabla f(x) = Ax+b$ as desired.

Let's check $\nabla^2 f(x)$. Since $\nabla f(x) = Ax+b$ this implies that

$$\frac{\partial f}{\partial x_i}(x) = \sum_k A_{ik}x_k \Rightarrow \frac{\partial f}{\partial x_j \partial x_i}(x) = A_{ij} \Rightarrow \nabla^2 f(x) = A.$$

□

Theorem 17. *Let $f(x) = \frac{1}{2}x^TAx+b^Tx+c$ where $A\in S^n$, $b\in\mathbb{R}^n$, $c\in\mathbb{R}$. Then*

1. *x is a critical point $\iff$ $Ax+b=0$.*
2. *f has a global minimzer $\iff$ $A\succeq 0$ and $b\in Range(A)$*

in which case any x satisfying $Ax+b=0$ is a global minimizer.

Proof. 1. is clear from $\nabla f(x) = Ax+b$. For (2) suppose f has a global minimizer $\bar{x}$. Since $\nabla f(\bar{x}) = A\bar{x}+b=0$ then $-b = A\bar{x}$ and so $b = A(-\bar{x})$ and hence $b\in \text{Range}(A)$. Since $\bar{x}$ is a global minimizer, it is also a local minimizer, so we know $A = \nabla^2 f(\bar{x}) \succeq 0$.

Conversely, suppose that $b\in\text{Range}(A)$. Then there is $\bar{x}$ such that $A\bar{x}+b=0$ and $A\succeq 0$. First, $\bar{x}$ is a critical point but also $\nabla^2 f(x) = A\succeq 0\ \forall x$, so f is convex. Then $\bar{x}$ is a global minimizer. □

So if given a quadratic function and asked to find a minimizer, the recipe is: See if $Ax+b=0$ is solvable. If yes, then check if A is positive semidefinite. If yes, then x is a global minimizer. If $Ax+b=0$ is not solvable or A is not positive semidefinite, then f has no global minimizer.

Theorem 18. *Let $f(x) = \frac{1}{2}x^TAx+b^Tx+c$ where $A\in S^n$, $b\in\mathbb{R}^n$, and $c\in\mathbb{R}$. Then f is coercive if and only if $A\succ 0$.*

Proof. See textbook for proof. □

As a summary, quadratic functions are special because you can completely characterize their global minimizers purely using linear algebra. Computing minimizers of a quadratic functions reduces to linear algebra! This is extremely special. Next time we will look at a particular class of quadratics and applications in data fitting (regression) and in signal processing (denoising).

12 January 31st, 2022

12.1 Least Squares (Chapter 3)

Consider solving

$$Ax = b.$$

Equivalently this means find x such that $A_{i\cdot}x = b_i \ \forall i = 1, \ldots, m$ (ith row of A times x gives us the corresponding entry of b). If the system is inconsistent (A is a tall and skinny matrix with more equations than unknowns $n \ll m$) then small perturbations of the RHS immediately destroy intersections, so we can instead solve the following optimization problem with $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^n$.

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 = \underbrace{\frac{1}{2}\sum_{i=1}^{n}(A_{i\cdot}x - b_i)^2}_{\text{smooth}}$$

or

$$\min_x \frac{1}{2}\|Ax - b\|_1 = \underbrace{\frac{1}{2}\sum_{i=1}^{n}|A_{i\cdot}x - b_i|}_{\text{nonsmooth}}.$$

Goal: Analyze least squares (the first optimization problem) and illustrate some applications. We need a lemma:

Lemma 6. *We have*

1. *Null*$(A^TA) =$ *Null*(A).
2. *Range*$(A^TA) =$ *Range*(A^T).

Recall: For any $B \in \mathbb{R}^{m\times n}$ we have

$$\text{Range}(B) = (\text{Null}(B^T))^{\perp}.$$

Proof. $\supseteq$ in (1) is clear. Let $v \in \text{Null}(A^TA)$. Then $0 = A^TAv \implies 0 = v^TA^TAv = (Av)^T(Av) = \|Av\|_2^2 \implies Av = 0 \implies v \in \text{Null}(A)$.

(2) We have $\text{Range}(A^TA) = (\text{Null}(A^TA)^T))^{\perp} = \underbrace{\text{Null}(A^TA)^{\perp} = \text{Null}(A)^{\perp}}_{\text{using (1)}} = \text{Range}(A^T)$. □

Recall that a *quadratic function* over $\mathbb{R}^n$ is of the form:

$$f(x) = \frac{1}{2}x^TAx + b^Tx + c$$

Claim 1. $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ is a quadratic function. We have

$$\begin{aligned} f(x) &= \frac{1}{2}(Ax-b)^T(Ax-b) \\ &= \frac{1}{2}(x^TA^T - b^T)(Ax-b) \\ &= \frac{1}{2}x^TA^TAx - \frac{1}{2}x^TA^Tb - \frac{1}{2}b^TAx + \frac{1}{2}b^Tb \\ &= \frac{1}{2}x^T(\underbrace{A^TA}_{A})x + (\underbrace{-A^Tb}_{b})^Tx + \underbrace{\frac{1}{2}\|b\|_2^2}_{c}. \end{aligned}$$

Theorem 19. *The Least Squares problem always admits a minimizer which is any solution of the normal equations:*

$$A^TAx = A^Tb.$$

Proof. First we must show that $A^TA \succeq 0$. We did this for HW 2.

Second we must show that $(A^TA)x - A^Tb = 0$ or $(A^TA)x = A^Tb$ is solvable. The system is solvable iff $A^Tb \in \text{Range}(A^TA)$. From the lemma above we already know that $\text{Range}(A^TA) = \text{Range}(A^T)$ and clearly $A^Tb \in \text{Range}(A^T)$. □

12.2 Applications of Least Squares

Linear Fitting: Suppose we have data points $(s_i, t_i) \in \mathbb{R}^n \times \mathbb{R}$ for $i = 1, \dots, m$. We believe there exists $x \in \mathbb{R}^n$ such that $t_i \approx s_i^T x \ \forall i = 1, \dots, m$. In other words we believe that t_i is approximately a linear function of s_i where we don't know the coefficients (slope of the line). Then we may solve

$$\min_x \frac{1}{2}\sum_{i=1}^m (s_i^Tx - t_i)^2$$

or equivalently

$$\min_x \frac{1}{2}\|Sx - t\|_2^2$$

where

$$S = \begin{bmatrix} \leftarrow s_1^T \rightarrow \\ \vdots \\ \leftarrow s_t^T \rightarrow \end{bmatrix}, \ t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix}.$$

Then as we learned, solving for optimal x amounts to solving

$$S^TSx = S^Tt.$$

13 Feburary 2nd, 2022

13.1 Nonlinear Fitting

Suppose we observe $(s_i, t_i) \in \mathbb{R} \times \mathbb{R}$ for $i = 1, \dots, m$ and we believe there is a degree d polynomial $p(\cdot)$ such that $t_i \approx p(s_i)$. How do we find a good polynomial that explains the data? So set $p(s) = a_0 + a_1 s + a_2 s^2 + \cdots + a_d s^d$. We believe $t_i \approx p(s_i) = a_0 + a_1 s_i + a_2 s_i^2 + \cdots + a_d s_i^d \ \forall i = 1, \dots, m$. So solve the problem

$$\begin{aligned}\min_a \frac{1}{2}\sum_{i=1}^m (t_i - p(s_i))^2 &= \min_a \frac{1}{2}\sum_{i=1}^m (a_0 + a_1 s_i + a_2 s_i^2 + \cdots + a_d s_i^d - t_i)^2 \\ &= \min_{a=[a_0,\dots,a_d]^T} \frac{1}{2}\|Sa - t\|_2^2\end{aligned}$$

where

$$S = \begin{bmatrix} 1 & s_1 & s_1^2 & \dots & s_1^d \\ & & \vdots & & \\ 1 & s_i & s_i^2 & \dots & s_i^d \\ & & \vdots & & \\ 1 & s_m & s_m^2 & \dots & s_m^d \end{bmatrix}, \ t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix}.$$

Important pitfall: When d is very big, you can fit anything. To prevent overfitting, we can regularize by penalizing $\sum_{j=0}^d a_j^2$. Then *regularized least squares* becomes

$$\min_x \frac{1}{2}\underbrace{\|Ax - b\|_2^2}_{\text{fidelity}} + \lambda \underbrace{R(x)}_{\text{prior}}$$

for $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, \lambda \geq 0$.

Concretely, for polynomial fitting, we have

$$\min_{a=[a_0,\dots,a_d]^T} \frac{1}{2}\|Sa - t\|_2^2 + \lambda \underbrace{\sum_{j=0}^m a_j^2}_{\|a\|_2^2.}$$

So $R(a) = \|a\|_2^2$. Other typical choices for R are

- $R(x) = \frac{1}{2}\|Dx\|_2^2 = \frac{1}{2}\sum_{i=1}^p [(Dx)_i]^2$ for some $D \in \mathbb{R}^{p \times n}$.
  Forces $(Dx)_i$ to be all small on average.
- $R(x) = \|Dx\|_1 = \sum_{i=1}^p |(Dx)_i|$ for some $D \in \mathbb{R}^{p \times n}$.
  Forces many of $(Dx)_i$ to be zero (sparsity). L1 regularization is also called compressed sensing, LASSO, sparse recovery.

We claim that (RLS) with $R(x) = \frac{1}{2}\|Dx\|_2^2$ is a quadratic optimization problem. We have

$$\begin{aligned}\min_x \frac{1}{2}\|Ax-b\|_2^2 + \frac{\lambda}{2}\|Dx\|_2^2 &= \min_x \frac{1}{2}x^T A^T A x - (A^T b)^T x + \frac{1}{2}\|b\|_2^2 + \frac{\lambda}{2}x^T D^T D x \\ &= \min_x \frac{1}{2}x^T \underbrace{(A^T A + \lambda D^T D)}_{\text{If } D = I\text{, this matrix is invertible.}} x - (A^T b)^T x + \frac{1}{2}\|b\|_2^2.\end{aligned}$$

14 Feburary 7th, 2022

14.1 Midterm Review (Solutions)

Problem 2 was an exercise on the Rayleigh-Ritz Theorem. $A \in \mathbb{R}^{n \times n}$ symmetric.

(a) Show that $A - \lambda_n I \succeq 0$.

Proof. Take $x \in \mathbb{R}^n$. We have $x^T(A - \lambda_n I)x = x^T Ax - \lambda_n \|x\|_2^2 \geq 0$. □

(b) Show that $A - \lambda_n I$ is not positive definite.

Proof. We need $x \neq 0$ such that $0 = x^T(A - \lambda_n I)x$. Take x to be an eigenvector corresponding to λ_n. □

(c) Show that $A - \lambda I_n \succ 0$ $\forall \lambda < \lambda_n$.

Proof. We have

$$\begin{aligned} x^T(A - \lambda I)x &= x^T Ax - \lambda \|x\|_2^2 \\ &\geq \lambda_n \|x\|_2^2 - \lambda \|x\|_2^2 \\ &= \underbrace{(\lambda_n - \lambda)}_{>0} \|x\|_2^2. \end{aligned}$$

□

Problem 4 was a direct application of the MVT, and mirrored proofs done in class.

(a) Show that $f(y) > f(x) + \langle \nabla f(x), y - x \rangle$ $\forall x \neq y \in \mathbb{R}^n$.

Proof. MVT gives $z \in [x, y]$ such that

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \underbrace{\langle \nabla^2 f(z)(y - x), y - x \rangle}_{>0 \text{ if } y \neq x}.$$

□

(b) Any critical point $\bar{x}$ of f must be a strict global minimizer.

Proof. Let $\bar{x}$ be a critical point. Then by (a), we have

$$f(y) > f(\bar{x}) + \underbrace{\langle \nabla f(\bar{x}), y - \bar{x} \rangle}_{=0} > f(\bar{x})$$

$\forall y \neq \bar{x}$. □

14.2 Nonlinear Fitting (continued)

Theorem 20. *Solutions of regularized least squares (RLS) are precisely the solutions of the linear system*

$$(A^T A + \lambda D^T D)x = A^T b.$$

Example: Denoising/Filtering

Suppose we get to observe $b = x_{\#} + w$ where $x_{\#} \in \mathbb{R}$ is the "truth" and w is a small noise.

Goal: Infer $x_{\#}$. Let's try LS:

$$\min_x \frac{1}{2} \sum_{i=1}^{n} (b_i - x_i)^2$$

yields $x = b$. What if we belive that $(x_1, \dots, x_n)$ is a discretization of a smooth function. If x_i comes from a smooth curve, we would expect $x_i - x_{i+1}$ to be small. So we can perform RLS with

$$\min_x \frac{1}{2} \sum_{i=1}^{n} (b_i - x_i)^2 + \lambda \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 = \min_x \frac{1}{2} \|b - x\|_2^2 + \frac{\lambda}{2} \|Lx\|_2^2.$$

So $R(x) = \frac{1}{2} \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 = \frac{1}{2} \|Lx\|_2^2$ where

$$L = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & -1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & & & & & \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}$$

also known as a finite difference matrix. So recalling Theorem 20, what is A in this scenario? Here, $A = I$. So the optimality condition is

$$(I + \lambda L^T L)x = b$$

so $x_{RLS} = (I + \lambda L^T L)^{-1} b$ and $I + \lambda L^T L \succ 0$ (why? possible exam question) as a consequence of Rayleigh-Ritz.

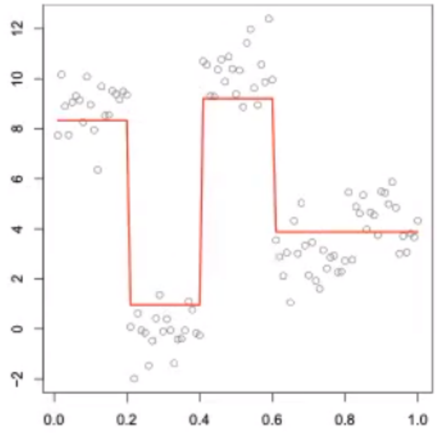
15 Feburary 9th, 2022

15.1 More Examples

What if $x_{\#}$ is piecewise constant? Then we have

$$\min_x \frac{1}{2}\|b - x\|_2^2 + \lambda \sum_{i=1}^{n} |x_i - x_{i+1}|.$$

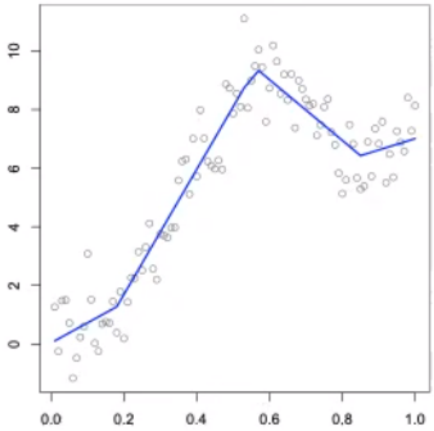
This will output the following piecewise constant structure, also called trend filtering.



What if $x_{\#}$ is piecewise linear? Then it's derivative is piecewise constant. We have

$$\min_x \frac{1}{2}\|b - x\|_2^2 + \lambda \sum_{i=1}^{n} |x_i - 2x_{i+1} + x_{i+2}|$$

which penalizes the second-order discrete derivative.



15.2 Line Search Methods (chapter 4)

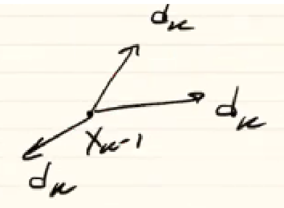
Problem:

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

where f is C^1-smooth or maybe C^2-smooth. The information we are allowed to use is $f(x)$, $\nabla f(x)$ and maybe $\nabla^2 f(x)$ at any point x. Line-search methods take the form

$$x_{k+1} = x_k + t_k d_k \text{ for } k = 1, 2, \dots$$

where $d_k \in \mathbb{R}^n$ is the *search direction* and $t_k > 0$ is the *stepsize*.

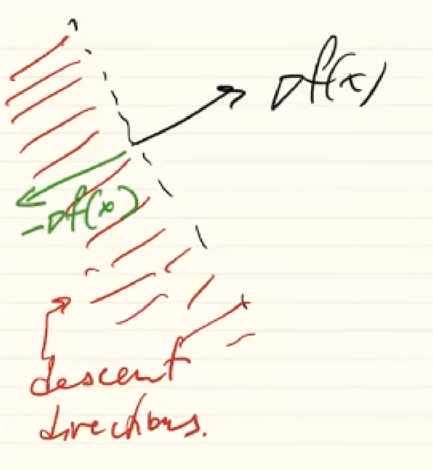


Questions:

1. How to choose d_k?
2. How to choose t_k?
3. When to stop?

We want to choose the search direction such that the function value decreases.

Definition 27. Let f be C^1-smooth. A vector $0 \neq d \in \mathbb{R}^n$ is a *descent direction* of f at x if $f'(x, d) = \langle \nabla f(x), d \rangle < 0$.



16 Feburary 11th, 2022

16.1 Line Search Methods (continued)

Lemma 7. *Let f be C^1 and let $d \in \mathbb{R}^n$ be a descent direction of f at x. Then $\forall_{\alpha\in(0,1)}$ there exists $\varepsilon > 0$ such that*

$$f(x+td) \leq f(x) + \alpha t \underbrace{\langle \nabla f(x), d\rangle}_{f'(x,d)<0} \quad \forall t \in (0,\varepsilon).$$

So $f(x+td) < f(x)$ $\forall t \in (0,\varepsilon)$.

Proof. We know

$$\begin{aligned} f(x+td) &= f(x) + t\langle \nabla f(x), d\rangle + o(t)^{1} \\ &= f(x) + \alpha t\langle \nabla f(x), d\rangle + \underbrace{(1-\alpha)t\langle \nabla f(x), d\rangle + o(t)}_{\leq 0 \text{ for all small } t>0}. \end{aligned}$$

The last two terms are ≤ 0 because

$$\frac{(1-\alpha)t\langle \nabla f(x), d\rangle + o(t)}{t} = \underbrace{(1-\alpha)f'(x,d)}_{<0} + \underbrace{\frac{o(t)}{t}}_{\to 0}.$$

□

So what this means is that if you take a sufficiently small step size in a descent direction the function value is guaranteed to go down, and we can estimate by how much. If you take a big step size, it may not go down.

The two steps are to 1. Pick the descent (search) direction 2. Pick the stepsize.

Popular Search Directions

1. Gradient Descent: $d_k = -\nabla f(x_k)$.
   Descent direction because $\langle \nabla f(x), d\rangle = -\|\nabla f(x_k)\|^2 < 0$ as long as $\nabla f(x_k) \neq 0$.
2. Newton: $d_k = -[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$.
   Descent direction if $\nabla^2 f(x_k) \succ 0$ because $\langle \nabla f(x), d\rangle = -\langle \nabla f(x), \nabla^2 f(x_k)^{-1} f(x_k)\rangle < 0$[2] as long as $\nabla f(x_k) \neq 0$.
3. Quasi-Newton: $d_k = H_k \nabla f(x)$ where H_k is an approximation of $-[\nabla^2 f(x_k)]^{-1}$.
   We won't have time to talk about these.

More on each of these later! But they give us a choice for the descent direction. The next choice we have to make is for the step size t_k. Note that choice of descent direction fundamentally changes the way the algorithm works, while the step size does not.

[2] The inverse of a psd matrix is psd since the eigenvalues of the inverse just become the reciprocals. So if the original eigenvalues are all positive, then so are their reciprocals.

Popular ways to set stepsize t_k

1. Constant: $t_k = \bar{t} \ \forall k$.

   What should $\bar{t}$ be? We'll see later that it will depend on "how smooth f is."
2. Exact line search: $t_k \in \operatorname*{arg\,min}_{t \geq 0} f(x_k + td_k)$.

   Often can't be computed.
3. Backtracking: Fix parameters $s > 0, \alpha \in (0,1), \beta \in (0,1)$. Then use the following algorithm to find t_k:

```
$t \leftarrow s$
while $f(x_k + td_k) > f(x_k) + \alpha t \langle \nabla f(x_k), d_k \rangle$ do
    $t \leftarrow \beta t$
end while
$t_k = t$.
```

17 Feburary 14th, 2022

17.1 Line Search Methods (continued)

Recall last time we ended with the backtracking algorithm to select the stepsize t_k. Let's rewrite it a little. Define $\varphi(t) = f(x_k + td_k)$.

Basic fact:

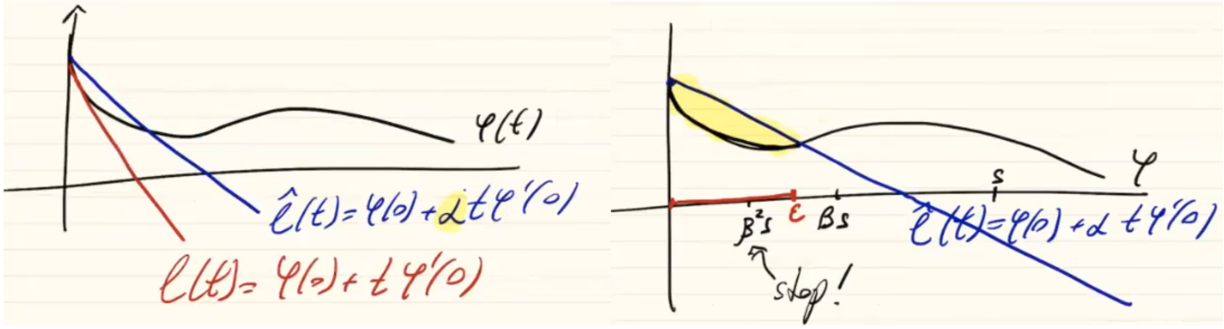
$$\varphi(t) = \varphi(0) + \varphi'(0)t + o(t).$$

This is true since recall $f(x_k + td_k) = f(x_k) + tf'(x_k, d_k) + o(t)$.

What the lemma showed:

$$\varphi(t) \leq \varphi(0) + \alpha t\varphi'(0) \ \forall \text{ small } t.$$

This is true since recall from lemma $f(x_k+td_k) \leq f(x_k)+\alpha tf'(x_k, d_k)\ \forall t \in (0, \varepsilon)$. So in each step of the backtracking algorithm we are checking the while loop condition if $\varphi(t) > \hat{\varphi}(t)$. Later on we will get a Theorem verifying that choosing a search direction and stepsize via backtracking will succeed. But next we want to spend a little more time talking about the various search directions.



17.2 Gradient Descent

Recall $d = -\nabla f(x)$ is a descent direction at x as long as $\nabla f(x) \neq 0$ because $\langle \nabla f(x), d\rangle = -\|\nabla f(x)\|^2 < 0$. Why this direction? Because it is the most obtuse direction.

Lemma 8. *Suppose f is C^1 and $\nabla f(x) \neq 0$. Then*

$$-\frac{\nabla f(x)}{\|\nabla f(x)\|_2} \in \underset{\|d\|_2=1}{\arg\min} f'(x, d).^{3}$$

[3] Note that we are just rescaling the negative gradient so that it has unit length

Proof. By Cauchy Schwarz we have

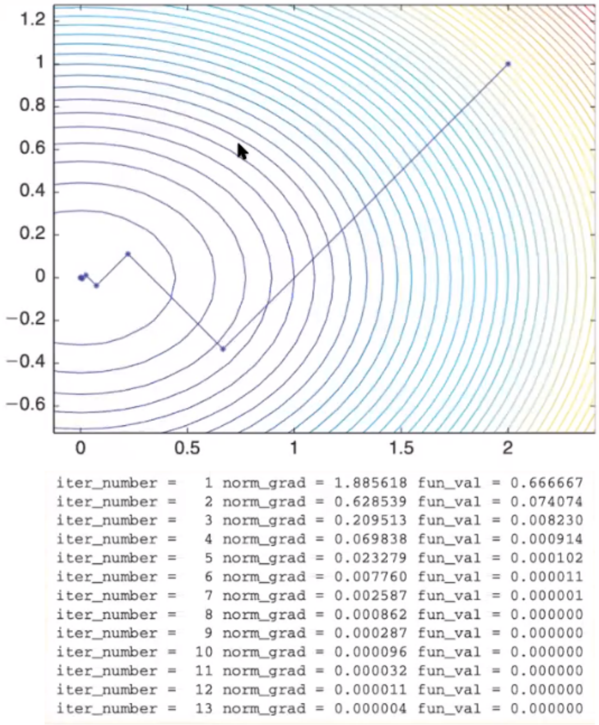
$$\begin{aligned}\min_{\|d\|=1} f'(x,d) &= \min_{\|d\|=1} \langle \nabla f(x), d\rangle \\ &\geq \min_{\|d\|_2=1} -\|\nabla f(x)\|_2 \cdot \|d\|_2 \\ &= -\|\nabla f(x)\|_2.\end{aligned}$$

Set $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$. Then

$$\begin{aligned}f'(c,d) &= \left\langle \nabla f(x), -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}\right\rangle \\ &= -\|\nabla f(x)\|.\end{aligned}$$

□

Here is a picture of grad descent with exact line search. Notice that each step is orthogonal to the contour line (the gradient is always orthogonal to the contour line, and we are walking in the direction of minus the gradient). Also notice that gradient descent exhibits the phenomenon of zig-zagging, where each direction is perpendicular to last. With respect to the training we have linear convergence. This means that at every iteration we decrease the error by a multiplicative of some constant that we call the rate of convergence, which depends on the scaling of the problem (in this example it is $\approx \frac{1}{3}$).



```
iter_number =   1 norm_grad = 1.885618 fun_val = 0.666667
iter_number =   2 norm_grad = 0.628539 fun_val = 0.074074
iter_number =   3 norm_grad = 0.209513 fun_val = 0.008230
iter_number =   4 norm_grad = 0.069838 fun_val = 0.000914
iter_number =   5 norm_grad = 0.023279 fun_val = 0.000102
iter_number =   6 norm_grad = 0.007760 fun_val = 0.000011
iter_number =   7 norm_grad = 0.002587 fun_val = 0.000001
iter_number =   8 norm_grad = 0.000862 fun_val = 0.000000
iter_number =   9 norm_grad = 0.000287 fun_val = 0.000000
iter_number =  10 norm_grad = 0.000096 fun_val = 0.000000
iter_number =  11 norm_grad = 0.000032 fun_val = 0.000000
iter_number =  12 norm_grad = 0.000011 fun_val = 0.000000
iter_number =  13 norm_grad = 0.000004 fun_val = 0.000000
```

Lemma 9. *(Zig-zag) Let $\{x_k\}$ be generated by gradient descent with exact line search.*

$$\begin{cases} t_k = \underset{t\in\mathbb{R}}{\arg\min} f(x_k - t\nabla f(x_k)) \\ x_{k+1} = x_k - t_k \end{cases}$$

Then $\langle x_{k+2} - x_{k+1}, x_{k+1} - x_k\rangle = 0 \ \forall k$.

Proof. We have

$$\begin{aligned} \langle x_{k+2} - x_{k+1}, x_{k+1} - x_k\rangle &= \langle -t_{k+1}, \nabla f(x_{k+1}), -t_k\nabla f(x_k)\rangle \\ &= t_k t_{k+1}\langle \nabla f(x_{k+1}), \nabla f(x_k)\rangle \end{aligned}$$

Define $\varphi() = f(x_k - t\nabla f'(x_k))$. Remember $t_k = \underset{t\in\mathbb{R}}{\arg\min}\varphi(t)$. Then by the first order necessary condition for optimality, $0 = \varphi'(t) = \langle \nabla f(\underbrace{x_k - t\nabla f(x_k)}_{x_{k+1}}), \nabla f(x_k)\rangle$. □

18 Feburary 16th, 2022

18.1 Convergence Analysis of Gradient Descent

The zig-zag behavior suggests gradient descent can be slow if the problem is "poorly scaled." For example observe the following problem:

$$\min_{x,y} x^2 + \frac{1}{100}y^2 = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{100} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Notice that the $\frac{\lambda_{max}}{\lambda_{min}} = 100$ which is big.

Definition 28. The *condition number* of $A \in S^n$ is defined by

$$K(A) = \frac{\overbrace{\lambda_1(A)}^{\text{max e-val}}}{\underbrace{\lambda_n(A)}_{\text{min e-val}}}.$$

We will see later that the convergence speed of gradient descent depends on $K(\nabla^2 f(\bar{x}))$ where $\bar{x}$ is a minimizer.

Convergence Analysis of GD problem:

$$\min_x f(x)$$

where f is C^1. We say $f \in C_L^{1,1}$ if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \; \forall x, y.$$

Theorem 21. *Suppose f is C^2 smooth. Then the following are equivalent:* [4]

1. $f \in C_L^{1,1}$
2. $\underbrace{max_{i=1,\ldots,n}|\lambda_i(\nabla^2 f(x))|}_{\|\nabla^2 f(x)\|_{op}} \leq L \; \forall x.$

To prove this we will need the following lemma

Lemma 10. *Suppose f is C^1. Then*

1. $f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y-x)), y - x \rangle dt$

Suppose f is C^2. Then

2. $\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(y - x))(y - x)dt.$

[4] We won't be tested on this proof.

Proof. For (1), set $\varphi(t) = f(x + t(y - x))$. The Fundamental Theorem of Calculus tells us

$$\underbrace{\varphi(1)}_{=f(y)} - \underbrace{\varphi(0)}_{=f(x)} = \underbrace{\int_0^1 \varphi'(t)dt}_{=\nabla f(x+t(y-x)),y-x\rangle}$$

For (2), fix i and set $\varphi(t) = \frac{\partial}{\partial x_i} f(x + t(y - x))$. The Fundamental Theorem of Calculus tells us

$$\underbrace{\varphi(1)}_{=\frac{\partial}{\partial x_i}f(y)} - \underbrace{\varphi(0)}_{=\frac{\partial}{\partial x_i}f(x)} = \int_0^1 \varphi'(t)dt$$

and $\phi'(t) = \frac{d}{dt}\frac{\partial}{\partial x_i} f(x+t(y-x)) = \sum_j \frac{\partial^2 f}{\partial x_i \partial x_j}(x+t(y-x))(x_j-y_j) = \langle$ ith row of $\nabla^2 f(x)), x-y\rangle$ by the chain rule. □

Now we prove the Theorem.

Proof. $1 \implies 2$. Suppose $f \in C_L^{1,1}$. Then using (2) from the above lemma we have

$$\nabla f(\underbrace{x + td}_{y}) - \nabla f(x) = \int_0^t \nabla^2 f(x + sd)d\,ds.$$

We know

$$\begin{aligned}\|\nabla f(x + td) - \nabla f(x)\|_2 \le Lt\|d\| &\implies \left\|\frac{1}{t}\int_0^t \nabla^2 f(x + sd)d\,ds\right\|_2 \le L\|d\|_2 \\ &\implies \|\nabla^2 f(x)d\|_2 \le L\|d\|_2\,\forall d.\end{aligned}$$

Then plug in d to be the unit eigenvector corresponding to $\lambda_i \implies |\lambda_i| \le L$ as desired.

$2 \implies 1$. Suppose $\|\nabla^2 f(x)\|_{op} \le L$. Then for any vector v we have by (2) of lemma

$$\begin{aligned}\langle \nabla f(y) - \nabla f(x), v\rangle &= \langle \int_0^1 \nabla^2 f(x + t(y - x))(y - x)dt, v\rangle \\ &= \int_0^1 \langle \nabla^2 f(x + t(y - x))(y - x), v\rangle dt \\ &\le \int_0^1 \|\nabla^2 f(x + t(y - x))(y - x)\|_2\|v\|dt \\ &\le \int_0^1 \|\nabla^2 f(x + t(y - x))\|_{op}\|y - x\|_2\|v\|dt \\ &\le L\|y - x\|_2\|v\|_2.\end{aligned}$$

Then set $v = \frac{\nabla f(y) - \nabla f(x)}{\|\nabla f(y) - \nabla f(x)\|} \implies \|\nabla f(y) - \nabla f(x)\| \le L\|y - x\|$. □

19 Feburary 18th, 2022

19.1 Logistic Regression

Note: We will not be tested on this material.

Suppose you have data $(a_i, b_i) \in \mathbb{R}^n \times \{-1, 1\}$ for $i = 1, \dots, m$ where a_i are the features and b_i are the labels. n is the number of features for each observation, and m is the number of observations. The goal is that given a new vector a, we predict their label b.

Modeling Assumptions: We believe (a_i, b_i) are drawn from some distribution and the goal is to predict $\Pr(b = 1 \mid a)$. So we introduce a parametric description of this probability. Assume

$$Pr(b = 1 \mid a) = \frac{1}{1 + e^{-f(a)}} \in (0, 1)$$

for some function f. Then we have

$$\begin{aligned} Pr(b = -1 \mid a) &= 1 - Pr(b = 1 \mid a) \\ &= 1 - \frac{1}{1 + e^{-f(a)}} \\ &= \frac{e^{-f(a)}}{1 + e^{-f(a)}} \\ &= \frac{1}{e^{f(a)} + 1} \\ &= \frac{1}{1 + e^{f(a)}}. \end{aligned}$$

Then to summarize, we say

$$Pr(b \mid a) = \frac{1}{1 + e^{-bf(a)}}$$

where $b \in \{-1, 1\}$. Next assume that $f(a) = x^T a$ for some x. We don't know x! So now,

$$Pr(b \mid a) = \frac{1}{1 + e^{-bx^T a}}.$$

So now our goal is to infer x from the observed data (a_i, b_i). We can use maximum likelihood estimation. The likelihood we observe (a_i, b_i) is

$$Pr(b_1, b_2, \dots, b_m \mid a_1, \dots, a_m) = \prod_{i=1}^{m} Pr(b_i \mid a_i) = \prod_{i=1}^{m} \frac{1}{1 + e^{-b_i(x^T a_i)}}.$$

Now we can maximize this expression over x, which is equivalent to minimizing the negative log likelihood since it is a monotone change of variables. We have

$$\max_x \prod_{i=1}^{m} \frac{1}{1 + e^{-b_i(x^T a_i)}} \iff \min_x \sum_{i=1}^{m} \log(1 + e^{-b_i(x^T a_i)}).$$

This function (call it f) in x is C^2. Actually, it is $C_L^{1,1}$ for some L. In fact, we will see that

$$0 \leq \lambda_{\min}(\nabla^2 f(x)) \leq \lambda_{\max}(\nabla^2 f(x)) \leq L.$$

So what does this tell us? So by the right side tells us that f is $C_L^{1,1}$, and the left side tells us that f is convex. So any critical point is a global minimizer. Hence, if we apply one our descent methods to find a solution we can solve the problem globally.

20 Feburary 23rd, 2022

20.1 Convergence of Gradient Descent

$C^{1,1}$ functions can be bounded above (and below) by a quadratic function over the space.

Lemma 11. *(Descent Lemma). Let $f \in C_L^{1,1}$. Then $\forall x, y \in \mathbb{R}^n$ we have*

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

Proof. Recall by the fundamental theorem of calculus that

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle \, dt.$$

Then we have

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &= \left| \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \, dt \right| \\ &\leq \int_0^1 |\langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle| \, dt \\ (Cauchy - Schwarz) &\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \cdot \|y - x\| \, dt \\ (f \in C_L^{1,1}) &\leq \int_0^1 L\|t(y - x)\| \cdot \|y - x\| \, dt \\ &= L\|y - x\|^2 \int_0^1 t \, dt \\ &= \frac{1}{2}L\|y - x\|. \end{aligned}$$

□

Lemma 12. *(Sufficient Decrease). Suppose $f \in C_L^{1,1}$. Then $\forall x \in \mathbb{R}^n$ and $t > 0$ we have*

$$f(x) - f(x - t\nabla f(x)) \geq t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x)\|^2.$$

Proof. By the descent lemma we have

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

Result follows after rearranging terms. □

In particular, $f(x - t\nabla f(x)) < f(x)$ as long as $t < \frac{2}{L}$. A good stepsize may be chosen by

$$\max_t t - \frac{Lt^2}{2}.$$

Then $1 - Lt = 0 \implies t = \frac{1}{L}$ which is an "optimal stepsize." Then

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

Lemma 13. *(Sufficient decrease of the gradient method). Let $f \in C_L^{1,1}$. Let $\{x_k\}$ be the sequence generated by gradient descent with one of the following stepsize strategies:*

1. *constant stepsize $t_k = \bar{t} \in (0, \frac{2}{L})$.*
2. *exact line search*
3. *backtracking with $s > 0$, $\alpha \in (0,1)$, $\beta \in (0,1)$.*

Then

$$f(x_k) - f(x_{k+1}) \geq M\|\nabla f(x_k)\|^2$$

where

$$M = \begin{cases} \bar{t}(1 - \frac{\bar{t}L}{2}) & \text{if } (1) \\ \frac{1}{2L} & \text{if } (2) \\ \alpha \min\{s, \frac{2(1-\alpha)\beta}{L}\} & \text{if } (3). \end{cases}$$

Proof. Recall that $f(x_{k+1}) = f(x_k + t_k d_k) = f(x_k - t\nabla f(x_k))$ with $d_k = -\|\nabla f(x_k)\|$.

(1) Substituting $x = x_k$, $t = \bar{t}$ in the sufficient decrease lemma yields

$$f(x_k) - f(x_{k+1}) \geq \bar{t}\left(1 - \frac{L\bar{t}}{2}\right)\|\nabla f(x)\|^2.$$

(2) Recall in exact line search the update requires $t_k \in \text{argmin}_{t\geq 0} f(x_k - t\nabla f(x_k))$. Then

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

where the first inequality is because $t = \frac{1}{L}$ is the max and the last inequality is from the sufficient decrease lemma with $t = \frac{1}{L}$. Rearranging yields the result.

(3) Two cases:
case 1 (1 step): Set $t_k = s$. Then by backtracking's sufficient decrease condition (Lemma 4.3 in textbook)

$$f(x_k) - f(x_k - s\nabla f(x_k)) \geq \alpha s\|\nabla f(x)\|^2.$$

case 2 (multiple steps): $t_k < s$. Then $\frac{t_k}{\beta}$ was not acceptable and so

$$f(x_k) - f\left(x_k - \frac{t_k}{\beta}\nabla f(x_k)\right) < \alpha\frac{t_k}{\beta}\|\nabla f(x_k)\|^2.$$

Then by the sufficient decrease lemma (for gradient method) with $x = x_k$, $t = \frac{t_k}{\beta}$ we have

$$f(x_k) - f(x_k - \frac{t_k}{\beta}\nabla f(x_k)) \geq \frac{t_k}{\beta}\left(1 - \frac{Lt_k}{2\beta}\right)\|\nabla f(x_k)\|^2,$$

which combined with above gives us

$$\alpha\frac{t_k}{\beta}\|\nabla f(x_k)\|^2 > f(x_k) - f\left(x_k - \frac{t_k}{\beta}\nabla f(x_k)\right) \geq \frac{t_k}{\beta}\left(1 - \frac{Lt_k}{2\beta}\right)\|\nabla f(x_k)\|^2 \tag{3}$$

$$\implies \frac{t_k}{\beta}\left(1 - \frac{Lt_k}{2\beta}\right)\|\nabla f(x_k)\|^2 < \alpha\frac{t_k}{\beta}\|\nabla f(x_k)\|^2 \tag{4}$$

$$\implies \frac{t_k}{\beta}\left(1 - \frac{Lt_k}{2\beta}\right) < \alpha\frac{t_k}{\beta} \tag{5}$$

$$\implies t_k > \frac{2(1-\alpha)\beta}{L}. \tag{6}$$

Then plug (6) back into backtracking's sufficient decrease condition and we have our result. □

21 Feburary 25th, 2022

21.1 Convergence of Gradient Descent (continued)

Theorem 22. *(convergence of the gradient method / rate of convergence of gradient norms). Let $f \in C_L^{1,1}$. Let $\{x_k\}$ be the sequence generated by gradient descent with one of the following stepsize strategies: constant stepsize $t_k = \bar{t} \in (0, \frac{2}{L})$, exact line search, or backtracking. Assume that f is bounded below over $\mathbb{R}^n$ ($\exists m \in \mathbb{R}$ such that $f(x) > m$ $\forall x \in \mathbb{R}^n$ or* $\inf f > -\infty$*). Then $\nabla f(x) \to 0$ as $k \to \infty$ and*

$$\min_{k=0,1,\dots,k} \|\nabla f(x_k)\|^2 \leq \frac{f(x_0) - \inf f}{m(k+1)}.$$

Proof. From the last Theorem, we get

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - M\|\nabla f(x_k)\|^2 \\ &\leq (f(x_{k-1}) - M\|\nabla f(x_{k-1})\|^2) - M\|\nabla f(x_k)\|^2 \\ &\leq f(x_0) - M\sum_{i=0}^{k} \|\nabla f(x_i)\|^2. \end{aligned}$$

Rearranging implies

$$\begin{aligned} M\sum_{i=0}^{k} \|\nabla f(x_i)\|^2 &\leq f(x_0) - f(x_{k+1}) \\ &\leq f(x_0) - \inf f. \end{aligned}$$

So $M\sum_{i=0}^{k} \|\nabla f(x_i)\|^2 < \infty$ and hence $\nabla f(x_i) \to 0$ as $i \to \infty$. Therefore we have

$$\min_{k=0,1,\dots,k} \|\nabla f(x_k)\|^2 \leq \frac{1}{k+1}\sum_{i=0}^{k} \|\nabla f(x_i)\|^2 \leq \frac{f(x_0) - \inf f}{m(k+1)}.$$

□

So we learned that to find x_i with $\|\nabla f(x_i)\| \leq \varepsilon$ it suffices to perform $\frac{f(x_0)-\inf f}{M\varepsilon^2}$ iterations where we multiply by $2L$ if we know the Lipschitz constant.

Questions:

1. Does there exist a $C_L^{1,1}$ function for which gradient descent is this slow? Yes
2. Is there an algorithm with a better guaranteed efficiency? No :(. For any algorithm, there exists a "bad" function $C_L^{1,1}$ for which the algorithm needs to compute $\frac{L(f(x_0)-\inf L)}{\varepsilon^2}$ gradients before finding x_i with $\|\nabla f(x_i)\| \leq \varepsilon$. (2017 Carmon, Duchi, Hinder, Sidford).

Theorem 23. *(Linear rate). Suppose $f \in C_L^{1,1}$ and C^2 and $\lambda_n(\nabla^2 f(x)) \geq u > 0 \ \forall x \in \mathbb{R}^n$. Let $\bar{x}$ be a minimizer. Then gradient descent with $t_k = \frac{1}{L}$ satisfies*

$$\|x_{k+1} - \bar{x}\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|x_k - \bar{x}\|^2.$$[5]

Proof. We have

$$\begin{aligned} \|x_{k+1} - \bar{x}\|^2 &= \|x_k - \frac{1}{L}\nabla f(x_k) - \bar{x}\|^2 && (7) \\ &= \|(x_k - \bar{x}) - \frac{1}{L}\nabla f(x_k)\|^2 && (8) \\ &= \|x_k - \bar{x}\|^2 - \frac{2}{L}\langle \nabla f(x_k), x_k - \bar{x}\rangle + \frac{1}{L^2}\|\nabla f(x_k)\|^2. && (9) \end{aligned}$$

We now need to prove a claim:

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

Proof. By the mean value theorem there exists $z \in [x, y]$ such that

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

By Rayleigh-Ritz we know that $\langle \nabla^2 f(z)(y - z), y - x\rangle \geq \mu\|y - x\|^2$. The result follows. □

So by this claim ($x = x, y = \bar{x}$?),

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

Then returning to our original proof, we have (9) $\leq$

$$\begin{aligned} &\leq \|x_k - \bar{x}\|^2 - \frac{2}{L}\left(f(x_k) - f(\bar{x}) + \frac{\mu}{2}\|\bar{x} - x\|^2\right) + \frac{1}{L^2}\|\nabla f(x_k)\|^2 \\ &\leq \left(1 - \frac{\mu}{L}\right)\|x_k - \bar{x}\|^2 - \frac{2}{L}\underbrace{\left(f(x_k) - f(\bar{x}) - \frac{1}{2L}\|\nabla f(x_k)\|^2\right)}_{\text{If we can show this term } > 0 \text{ we are done.}}. \end{aligned}$$

Since $\bar{x}$ is a minimizer and by the sufficient decrease lemma with $t = \frac{1}{L}$, we know that

$$\begin{aligned} f(\bar{x}) \leq f\left(x_k - \frac{1}{L}\nabla f(x_k)\right) &\leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\ \implies f(x_k) - f(\bar{x}) - \frac{1}{2L}\|\nabla f(x_k)\|^2 &\geq 0. \end{aligned}$$

□

[5] Notice that the term μ/L is related to the condition number $\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$, or namely 1 over the condition number since μ is a lowerbound of the min eigenvalue and L is an upperbound on the max eigenvalue.

22 Feburary 28th, 2022

22.1 Convergence of Gradient Descent (for Convex functions)

Let $f \in C_L^{1,1}$ with $\lambda_{min}(\nabla^2 f(x)) \geq \mu > 0 \; \forall x$. Let $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$. We showed that

$$\begin{aligned}\|x_{k+1} - \bar{x}\|^2 &\leq \left(1 - \frac{\mu}{L}\right)\|x_k - \bar{x}\|^2 \\ &\leq \left(1 - \frac{\mu}{L}\right)^2 \|x_{k-1} - \bar{x}\|^2 \\ &\vdots \\ &\leq \left(1 - \frac{\mu}{L}\right)^{k+1} \|x_0 - \bar{x}\|^2.\end{aligned}$$

We can just replace $k+1$ with k to simplify notation. How big of a k do we need to get the RHS $< \varepsilon$? Recall the following algebraic inequality: $(1-q)^k < e^{-qk}$ for $q \in (0,1)$. Then we have

$$\begin{aligned}\|x_{k+1} - \bar{x}\|^2 &\leq \left(1 - \frac{\mu}{L}\right)^k \|x_0 - \bar{x}\|^2 \\ &\leq e^{-k\frac{\mu}{L}}\|x_0 - \bar{x}\|^2.\end{aligned}$$

Set $e^{-k\frac{\mu}{L}}\|x_0 - \bar{x}\|^2 = \varepsilon$ and solve for k. Dividing both sides by $\|x_0 - \bar{x}\|^2$ we get

$$\begin{aligned}e^{-k\frac{\mu}{L}} &= \frac{\varepsilon}{\|x_0 - \bar{x}\|^2} \\ \implies e^{k\frac{\mu}{L}} &= \frac{\|x_0 - \bar{x}\|^2}{\varepsilon} \\ \implies k\frac{\mu}{L} &= \ln\left(\frac{\|x_0 - \bar{x}\|^2}{\varepsilon}\right) \\ \implies k &= \frac{L}{\mu}\ln\left(\frac{\|x_0 - \bar{x}\|^2}{\varepsilon}\right).\end{aligned}$$

So to recap, when applying gradient descent to a function $f \in C_L^{1,1}$ with minimum eigenvalue of its Hessian lower bounded by μ, the number of iterations needed to generate a point such that the squared distance from the optimum is smaller than epsilon is k. Notice that the initialization and the target accuracy (ε) appear logarithmically (not polynomially!). This is a consequence of $\lambda_{min}(\nabla^2 f(x)) \geq \mu > 0 \; \forall x$ (convexity). Also recall that $\frac{L}{\mu}$ is the condition number, so the larger the condition number, the greater number of steps needed (slower the algorithm) as expected.

Questions:

1. Does there exist a $C_L^{1,1}$ function for which gradient descent is this slow? Yes
2. Is there an algorithm with a better guaranteed efficiency? Yes!!!

Some history: Lower and upper complexity bounds originates in 1982 in book of Yudin-Nemirovsky. They proved a lower complexity bound for $C_L^{1,1}$ functions with $\lambda_{min}(\nabla^2 f(x)) \geq$

$\mu > 0 \; \forall x$. They proved the following lower bound (there is a really bad function where this is the minimum number of iterations):

$$k = \sqrt{\frac{L}{\mu}} \cdot \log\left(\frac{\|x_0 - \bar{x}\|^2}{\varepsilon}\right).$$

Example 22.1.1. So if $\frac{L}{\mu} = 10,000$ then $\sqrt{\frac{L}{\mu}} = 100$. Big difference.

No algorithm is guaranteed to converge faster than this lower bound. But does there exist an algorithm whose convergence guarantees match this lower convexity bound? They posed this question but did not answer it fully. Then is Nesterov '83 gave an algorithm with efficiency matching the lower bound. Next we will present this algorithm.

22.2 Accelerated Gradient Method

Not tested on this! Set $k = 0$ and $a_0 = a_{-1} = 1$, $x_{-1} = x_0$.

```
for $t = 0, \dots, T$ do
    $u_t = x_t + a_t(a_{t-1}^{-1} - 1)(x_t - x_{t-1})$
    $x_{t+1} + u_t - \frac{1}{L}\nabla f(u_t)$
    $a_{t+1} \leftarrow \frac{\sqrt{a_t^4+4a_t^2}-u_t^2}{2}$
end for
```

So x_{t+1} is the usual gradient step but at u_t not x_t. The next point where you take the gradient step, u_t, is updated according to the rule where you average x_t's together.

Theorem 24. *This algorithm (plus tiny modification) will find x such that $\|x - \bar{x}\|^2 \leq \varepsilon$ using $k = \sqrt{\frac{L}{\mu}} \cdot \log\left(\frac{\|x_0 - \bar{x}\|^2}{\varepsilon}\right)$ iterations.*

We will not prove this, a little complicated, would take too long. But this is used all the time in industry and is way faster.

Student Question: So we spent this last bit of time talking about guaranteed convergences. What about average / amortized cases? Like one where gradient descent is guaranteed to be this fast, but in general it is?

Tanget: This a long standing issue of understanding average-case algorithms, since everything we're discussing is worst case. What about average case? The main issue is that you need to describe what average means. Average depends on what kind of distribution you choose. Very recently there has been breakthrough work (Paquette et al). Given

$$\min_x \frac{1}{2}\|Ax - b\|^2$$

where $A \in \mathbb{R}^{m \times n}$, what if A is random? They showed:

Theorem 25. *Suppose $\frac{m}{n} \to r$ as $m, n \to \infty$. And suppose A is random means bounded moments. Convergence guarantee is universal, does not depend on distribution of A, and is much better than worst case.*

22.3 Newton's Method (Ch. 5)

Basic Guarantee: If you start sufficiently close to $\bar{x}$, then the # of iterations to get $\|x_n - \bar{x}\| \leq \varepsilon$ is $\log(\log(\frac{c}{\varepsilon}))$ where c is a constant.

Disadvantages:

1. Local rates
2. Need to compute Hessian $\nabla^2 f(x)$
3. In each iteration, you need to solve a linear system of equations.

Newton's Method: Algorithm for solving a system of equations:

$$\begin{aligned} g_1(x) &= 0 \\ g_2(x) &= 0 \\ &\vdots \\ g_m(x) &= 0. \end{aligned}$$

Notation: Define a map $G : \mathbb{R}^n \to \mathbb{R}^m$ by

$$G(x) = (g_1(x), g_2(x), \ldots, g_m(x)).$$

So we want to solve $G(x) = 0 \in \mathbb{R}^m$. At each iteration we approximate the point by a best linear approximation, set it to zero, which gives us the next point. How do we form a linear approximation of G? Recall our first order approximation:

$$g_i(y) = g_i(x) + \langle \nabla g_i(x), y - x \rangle + o(\|y - x\|).$$

Then if we do this for $i = 1, \ldots, m$ we get

$$G(y) = G(x) + A(y - x) + o(\|y - x\|)$$

where

$$A = \begin{bmatrix} \nabla g_1(x)^T \\ \nabla g_2(x)^T \\ \vdots \\ \nabla g_m(x)^T \end{bmatrix} =: \nabla G(x)$$

which we call the *Jacobian.* Back to equation solving: $G(x) = 0$.

Newton's method: Let x_{k+1} be the solution of the linear system

$$G(x_k) + \nabla G(x_k)(x - x_k) = 0.$$

In particular, if $m = n$ and $\nabla G(x_k)$ is invertible, then

$$\begin{aligned} &G(x_k) + \nabla G(x_k)(x - x_k) = 0 \\ \implies &\nabla G(x_k)(x - x_k) = -G(x_k) \\ \implies &(x - x_k) = -[\nabla G(x_k)]^{-1} G(x_k). \end{aligned}$$

So

$$x_{k+1} = x_k - [\nabla G(x_k)]^{-1} G(x_k)$$

where we solve for the inverse by letting it be d_k and solving the system $\nabla G(x_k) d_k = G(x_k)$.

23 March 2nd, 2022

23.1 Newton's Method (continued)

We are doing

$$\min_x f(x)$$

where f is C^2. The goal is to solve $\nabla f(x) = 0$. Let $G(x) = \nabla f(x)$. We can verify what $\nabla G(x)$ is. By definition, the (i, j)th entry of $\nabla G(x)$ is $\frac{\partial}{\partial x_j} g_i(x) = \frac{\partial}{\partial x_j \partial x_i} f(x)$. So the Newton update is simply

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k).$$

There is an alternative way to arrive at Newton's method which is more aligned with optimization. Recall that

$$Q(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle$$

is the best quadratic approximation of f at x_k. An interesting algorithm is

$$x_{k+1} = \arg\min_x f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle$$

which is well defined if $\nabla^2 f(x_k) \succ 0$.

First order optimality conditions (taking derivative of above at x_{k+1} and setting to zero):

$$\begin{aligned} &\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0 \\ \implies\ &\nabla^2 f(x_k)(x_{k+1} - x_k) = -\nabla f(x_k) \\ \implies\ &x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k). \end{aligned}$$

Example 23.1.1. We have

$$\min_{x,y} xy + x^3 + y^3.$$

Take $(x, y) = (1, 1)$. Let's compute

$$\begin{aligned} \nabla f(x, y) &= \begin{bmatrix} y + 3x^2 \\ x + 3y^2 \end{bmatrix} \implies \nabla f(1, 1) = \begin{bmatrix} 4 \\ 4 \end{bmatrix} \\ \nabla^2 f(x, y) &= \begin{bmatrix} 6x & 1 \\ 1 & 6y \end{bmatrix} \implies \nabla^2 f(1, 1) = \begin{bmatrix} 6 & 1 \\ 1 & 6 \end{bmatrix}. \end{aligned}$$

Our update rule is

$$x_{k+1} = x_k - \underbrace{[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)}_{=d}.$$

So if $(x_k) = (1, 1)$, to find x_{k+1} you do the following: Solve the equation

$$\begin{aligned} &\nabla^2 f(x_k) d = \nabla f(x_k) \\ \implies &\begin{bmatrix} 6 & 1 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}. \end{aligned}$$

Then

$$x_{k+1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}.$$

Theorem 26. *Let G be C^1 and let $\bar{x}$ satisfy $G(\bar{x}) = 0$. Suppose there eixst $\mu, \varepsilon, L > 0$ such that*

1. $\|\nabla G(x)^{-1}\|_{op} \leq \frac{1}{\mu}\ \forall x \in B_\varepsilon(\bar{x})$.
2. $\|\nabla G(x) - \nabla G(y)\|_{op} \leq L\|x - y\|\ \forall x, y \in B_\varepsilon(\bar{x})$.[6]

Let x_k be generated by Newton's method with $x_0 \in B_\varepsilon(\bar{x})$. Then

$$\|x_{k+1} - \bar{x}\| \leq \frac{L}{2\mu} \underbrace{\|x_k - \bar{x}\|^2}_{\leq \varepsilon}.$$

If also $\varepsilon < \frac{\mu}{L}$, then

$$\|x_k - \bar{x}\| \leq \frac{2\mu}{L} \quad \underbrace{\left(\frac{1}{2}\right)^{2^k}}_{\textit{quadratic convergence}} \quad \forall k.$$

Remark 23.1.1. Not tested on this. The problems

$$\min_x f(x) \quad \min_x f(Ax)$$

where A is invertible are identical since all we have done is reparameterized the domain of the function. But gradient descent behaves completely on these two problems. For example, grad descent on $f(x, y) = x^2 + y^2$ is great but grad descent on $f(Ax) = \frac{1}{2}x^2 + \frac{1}{2^{84}}y^2$ is awful. Here $A = \begin{bmatrix} 1 & 0 \\ 0 & 2^{-42} \end{bmatrix}$. Newton's method does not suffer from the same problem. Newton's method iterates on one problem are images by A of Newton iterates applied to the other problem. In other words, Newton's method is affine invariant. This is the reason why the constant stepsize of 1 is used in Newton's method. Newton's method is naturally adaptive to the scaling of the problem. The matrix itself (Hessian inverse) is already a kind of stepsize and we don't need to choose another scaler.

[6] This is an analogue to Lipschitz continuity for the gradient but now applied to the Jacobian of the gradient which is the Hessian.