Class #1

Tuesday, January 21, 2025 10:23 AM

What is Optimization?

Consider a function $f: \underbrace{\mathbb{R}^n}_{\text{dom } f} \to \mathbb{R}$

e.g., $f(x) = \log(x)$

$f: \mathbb{R} \to \mathbb{R}$

$\text{dom } f = (0, \infty)$

Optimization is concerned with finding the maximum or minimum value that f takes:

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

↳ If min (infimum) or max (supremum) exists

e.g., $f(x) = \log(x)$

$\min_x f(x) \Rightarrow$ does not exist

$f(x) = x^2 \Rightarrow \min_x f(x) = 0$

$x_n = \frac{1}{n}$; $n = 1, 2, \dots$

$\min_n x_n$ does not exist

but $\inf_n x_n = 0$

(or) Finding the x that gives the minimum or maximum value of f

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

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

e.g.: $f(x) = x^2$

$\min_x f(x) = 0$

$\arg\min_x f(x) = 0$

$f(x) = x^2 + 1$

$\min_x f(x) = 1$

$\arg\min_x f(x) = 0$

$f(x) = (x-1)^2$

$\min_x f(x) = 0$

$\arg\min_x f(x) = 1$

Practically, the function $f(x)$ represents some real-world quantity that we care about and $x \in \mathbb{R}^n$ corresponds to n parameters of the problem that alter the quantity of interest.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Example: Portfolio optimization $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ $x_1 \rightarrow$ investment in stock 1; $x_n \rightarrow$ investment in stock n

$f: \mathbb{R}^n \rightarrow \mathbb{R}$

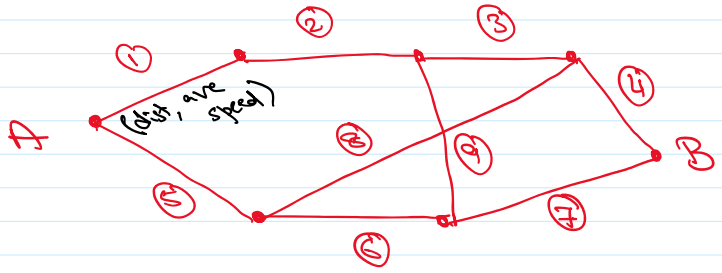
↳ Profit / loss after 6 months

$\arg\max_x f(x)$

Constraints: Total budget $\Rightarrow \sum_{i=1}^{n} x_i \leq B$

$x_i \geq 0$

Example: Route optimization w/ respect to minimization of time.



option 1: ① → ② → ③ → ⑤

option 2 ① → ② → ⑨ → ⑦

$f(x)$ = Time needed to go from A to B

↳ encoding the path

$f(x): \{0,1\}^9 \to \mathbb{R}$

$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_9 \end{bmatrix} \Rightarrow$ $x_i = 0$ if i^{th} edge is not taken

$= 1$ if i^{th} edge is taken

$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow$ option 1

$f(x) = \frac{dist(1)}{ave.speed(1)} \times x_1 + \frac{dist(2)}{av.speed(2)} \times x_2 + \cdots + \frac{dist(9)}{av.speed(9)} \times x_9$

↳ time taken to traverse edge 1

This problem involves optimizing a 9-dimensional binary vector $x \in \{0,1\}^9$

$\underset{x \in \{0,1\}^n}{\operatorname{argmin}} f(x)$ and $\min_x f(x)$

↳ This actually is not over all $\{0,1\}^9$

We have constraints that x must correspond to a valid route from A to B

$\mathcal{X} = \{ x : x \text{ corresponds to a valid route from A to B.} \}$

↳ Much smaller than 2^9

$\arg\min_{x \in \mathcal{X}} f(x)$

↳ Constraint set in optimization

↳ This is an example of combinatorial optimization.

Other Examples

① IC design layout and routing of wires

② Machine learning

③ Robot motion planning

Summary

Optimization problems are everywhere

① Modeling the problem into an optimization framework that is solvable.

→ Theory

② Recognizing the nature of the optimization framework and manipulating it so that it is efficient to solve.

→ Theory / Implementation

③ Using an appropriate numerical optimization method to help us solve the problem ⇒ Run-time

→ Theory

④ Understanding the 'goodness' of the solution provided by the algorithm.

~~~~~ x ~~~~~ x ~~~~~

Mathematical optimization: Foundational concepts

① Unless otherwise stated, $f(\cdot)$ is assumed to attain its minimum or maximum value.

② Without loss of generality, we will stick with $\min_x f(x)$ or $\max_x f(x)$, rather than $\arg\min_x f(x)$ or $\arg\max_x f(x)$

An optimization problem will be written as
~~~~~

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

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

but when the variable is obvious then we write

$\min f(x)$ or $\max f(x)$

$f \Rightarrow$ Objective function

$x \Rightarrow$ Optimization Variable

Unconstrained optimization

↳ $\min_{x \in \mathbb{R}^n} f(x)$ or $\min_{x \in \text{dom } f} f(x)$

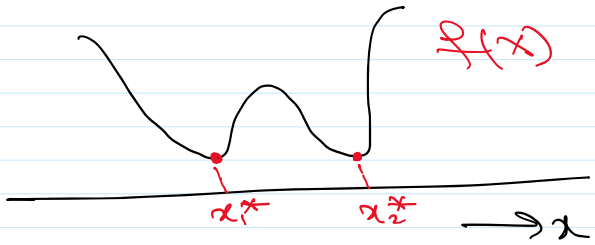
↳ "x" should be searched over the entire domain of f.

Solution $\Rightarrow$ A solution to an optimization problem $\min_{x \in \mathbb{R}^n} f(x)$ is often denoted by $x^* \in \mathbb{R}^n$ and it satisfies:

$$f(x^*) \leq f(x) \quad \forall \, x \in \mathbb{R}^n$$

x^* is referred to as a minimizer or an optimal value.

We can have multiple solutions; $x_1^*, x_2^*, \ldots$



We can have 'no' Solution.

e.g; $f(x) : (0, \infty) \to \mathbb{R}$

$$f(x) = -\frac{1}{x}$$

$\min_x f(x) \Rightarrow$ no x is going to give us the minimum value.

multiple Solutions

$\Rightarrow f(x) = \cos(x)$

$\min_x \cos(x)$

$x^* \in \{\pm\pi, \pm 3\pi, \pm 5\pi, \ldots\}$

Class #2

Friday, January 24, 2025 10:20 AM

* An optimization algorithm is a numerical method that helps us numerically find a value of x close to an optimal x^* in a reasonable amount of time as a function of 'n' and other problem parameters.

↳ Mathematical programming

* All optimization problems that are min or max can be thought of as minimization problems only.

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

$\max_x \cap$

$$\arg\max_x f(x) = \arg\min_x -f(x) = \min_x \cup$$

$\min_x \max_y f(x,y) \Rightarrow$ not the focus of this course

Constrained optimization

$$\min_{x \in \text{domf}} \boxed{f_0(x)}$$ → objective function

Subject to → constraint functions

$f_i(x) \le b_i$, $i = 1, \dots, m$

we have 'm' constraints — known constraint values

$$\iff \min_{x \in X} f_0(x) \quad : \quad X \text{ is a subset of domf}$$

$$X = \{ x : f_i(x) \le b_i, \ i = 1, \dots, m \}$$

e.g. $\min (x-4)^2$

e.g. $\min_{0 \le x \le 8} (x-4)^2$

$\mathcal{X} = \{ x : x \in [0,8] \}$

$0 \le x \le 8$

$\Rightarrow x \le 8$

$\Rightarrow -x \le 0$

$f_1(x) = x \Rightarrow b_1 = 8$

$f_2(x) = -x \Rightarrow b_2 = 0$

$\min_{x \in \text{dom}\, f_0} (x-4)^2$

s.t. $x \le 8$

$-x \le 0$

Any $x \in \mathcal{X}$ is termed a 'feasible solution'.

In unconstrained optimization, all x's are feasible.

Classes of optimization Problems

Depending on $f_0(x)$, $f_i(x)$, $i = 1, \dots, m$ (and the domain of $f_0(x)$), we characterize an optimization problem to be from a certain class of problems, all of which can be solved using a particular set of numerical tools.

① Combinatorial optimization

⇒ x takes on values from a discrete set (i.e., dom f corresponds to a discrete set).

$x \in \{0,1\}^n$

② Linear programming

The objective function $f_0(x)$ as well as the constraint functions $f_i(x)$, $i=1,\dots,m$, are linear function of x.

⇓

Integer programming

↳ Linear programs in which x takes only integer values.

③ Convex programming

↳ The function $f_0(x)$, the constraint functions $f_i(x)$ $i, i=1,\dots,m$, are all convex.

A function f is convex if and only if

$$f(\alpha x + \beta y) \leq \alpha f(x) + \beta f(y)$$

$\forall \ \alpha, \beta \geq 0$, $\alpha + \beta = 1 \Rightarrow \alpha = 1-\beta$

$\Leftrightarrow f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y)$

$\forall \ \theta \in [0,1]$

$x, y \in \text{dom} f$ and $\theta x + (1-\theta) y \in \text{dom} f$

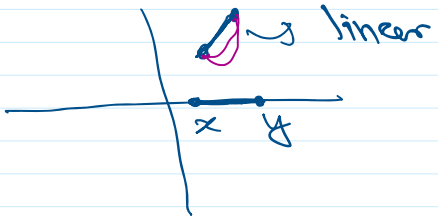
$\Leftrightarrow$ dom f has to satisfy the convexity property that $\forall\ x, y \in \text{dom} f$

$\theta x + (1-\theta) y \in \text{dom} f \quad \forall \ \theta \in [0,1]$

In other words, domf has to be a convex set.

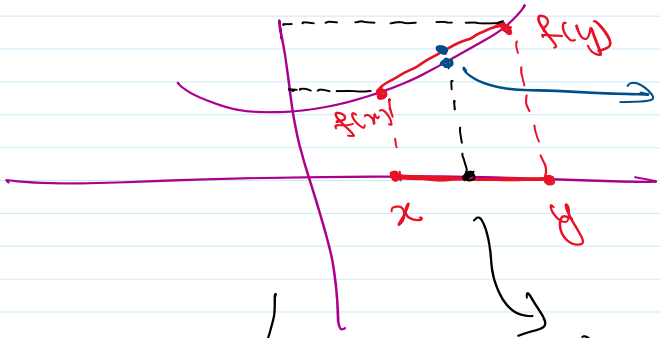
Linear function : $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$

All linear functions are convex



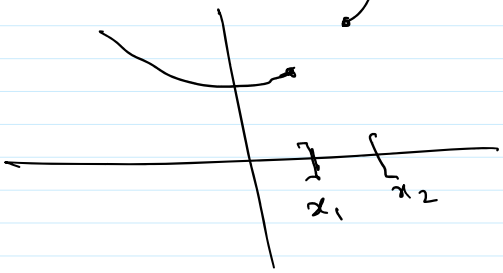
$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$

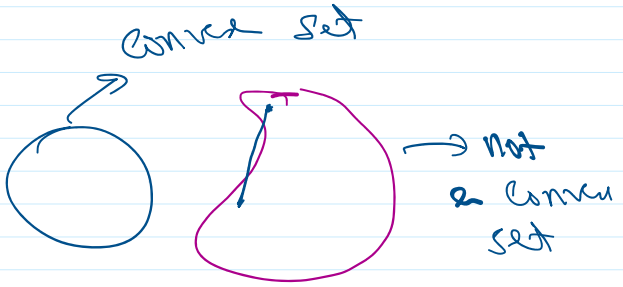
$\theta \in [0,1]$



$(0.5)x + (1-0.5)y$

E.g.;





(4) Non Convex optimization

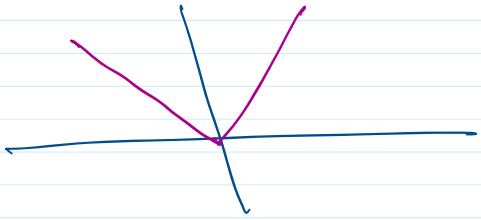
↳ Any optimization that is not convex

(5) Nonlinear programming

↳ Any optimization that is not linear.

⑥ Nonsmooth optimization

↳ when $f_0(x)$ has discontinuous derivatives



⑦ Stochastic optimization

↳ When $f_0(x)$ or $f_i(x)$ have randomness involved in them.

Global optimization method

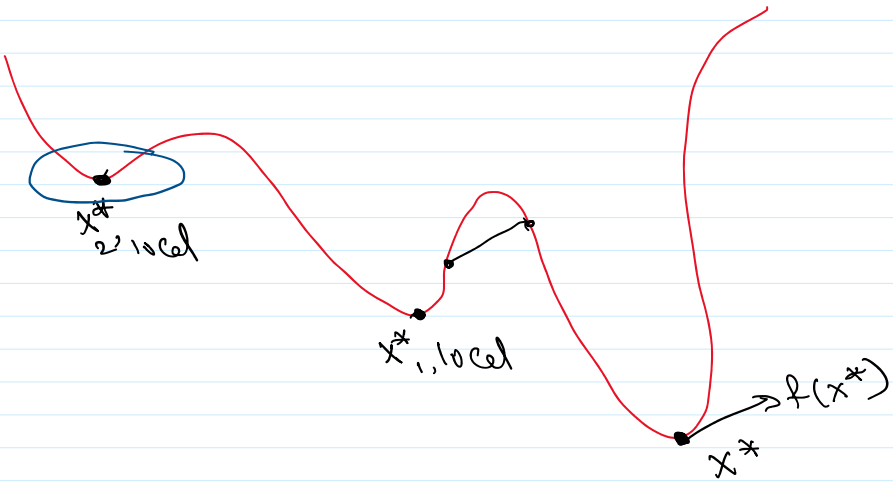
↳ A solver that provides an optimal solution x^*.

Convex optimization ⇒ we will see that global optimization is straightforward.

Nonconvex optimization ⇒ This is not computationally feasible in most cases (NP-hard).

Local optimization

↳ Find a solution x^*_{local} that is optimal in some neighborhood of x^*_{local}.



Why Convex Optimization?

1. Many real-world problems tend to be convex.
2. Convex optimization leads to global solutions.
3. The solvers are "efficient".
4. Many nonconvex problems can be 'relaxed' to convex problems and then solved globally.
5. Convex optim. can be used as an 'initialization scheme for nonconvex problems.

Class of Solvers

1. Zeroth-order methods
   - ↳ solve the problem by having access to function values only.

2. First-order methods
   - ↳ Make use of first-order partial derivatives

↳ Make use of first-order partial derivatives of $f(x)$.

$$\frac{\partial f(x)}{\partial x_i}, \quad i = 1, \dots, n$$

↳ Use gradient information of the function

(3) Second-order methods

↳ Use both gradient information and second-order partial derivatives.

$$\frac{\partial f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, \dots, n$$

↳ Use Hessian information

Examples of Convex Optimization

(1) Least Squares problem

$$f(x) = \|Ax - b\|_2^2 = \sum_{i=1}^{k} (a_i^T x - b_i)^2$$

A: $\mathbb{R}^{k \times n}$ matrix

b: $\mathbb{R}^{k \times 1}$ vector

↳ Quadratic program

$b \Rightarrow$ Observations vector (k observations)

$A \Rightarrow$ known matrix

Find closest x s.t. $\boxed{Ax \geq b}$

If $k \geq n$ and A is full rank

$x^* = (A^T A)^{-1} A^T b$

↳ Analytical solution

↳ we can use convex optimization tools to solve this in $O(n^3 k)$ time.

Variants of LS

Weighted LS

$$f(x) = \sum_{i=1}^{k} w_i (a_i^T x - b_i)^2 \qquad w_i \geq 0$$

$$= \| W(Ax-b) \|_2^2 \text{, where}$$

$$W = \text{diag}(w_1, w_2, \ldots, w_k)$$

$$= \| \underbrace{WA}_{A} x - \underbrace{Wb}_{b} \|_2^2$$

Regularized LS

$$f(x) = \|Ax-b\|_2^2 + \rho \|x\|_2^2$$

$\|x\|_2^2$ ↓ regularizer

regularization parameter $\rho \geq 0$

$f(x)$ ↓

Ridge regression

Linear Programming

$\min_x \; c^T x$ $\quad f_0(x)$ is linear

subject to $a_i^T x \leq b_i$, $i = 1, \ldots, m$

$f_1(x) \longrightarrow f_m(x)$ linear

Typical

$f_1(x) \longrightarrow f_m(x)$ linear

Typical $O(n^2 m)$ complexity if $m \geq n$

$$\min_x \max_{i=1,\ldots,k} |a_i^T x - b_i| = \min_x \|Ax-b\|_\infty$$

Chebyshev approximation

Compare LS $\min_x \|Ax-b\|_2^2$

Review of key linear algebra concepts

Inner product

Given $x, y \in \mathbb{R}^n$, the standard inner product is

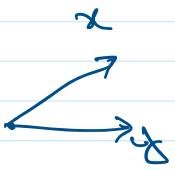
$$\langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i$$

and $x^T x = \|x\|_2^2$, where $\|\cdot\|_2$ is Euclidean norm

↳ Euclidean norm

Angle between two vectors: θ

$$x^T y = \|x\|_2 \|y\|_2 \cos(\theta)$$



If $x^T y = 0 \iff x \perp y$

If $x^T y > 0 \iff x$ and y make an acute angle

Euclidean norm: $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

Cauchy-Schwarz Inequality

$$|x^T y| \leq \|x\|_2 \|y\|_2$$

Inner Product between matrices

Given $X, Y \in \mathbb{R}^{m \times n}$, inner product on matrices

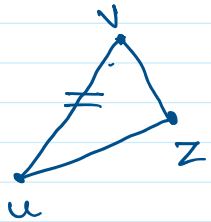
$$\langle X, Y \rangle = \text{trace}(X^T Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} Y_{ij}$$

$$\langle X, X \rangle = \|X\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2$$

What is a norm? Focus on $\mathbb{R}^n$

A function $f: \mathbb{R}^n \to \mathbb{R}$ with $\text{dom} f = \mathbb{R}^n$ is called a norm if:

- $f(x) \geq 0 \quad \forall \; x \in \mathbb{R}^n$ — non-negativity
- $f(x) = 0$ only if $x = 0$ — definiteness
- $f(tx) = |t| f(x) \quad \forall \; t \in \mathbb{R}$, — homogeneity
- $f(x+y) \leq f(x) + f(y)$, — Triangle inequality



$\|x\|_0$

$\|x\|_2 \Rightarrow$ Euclidean norm $\Rightarrow$ l_2-norm

l_p-norm, $p \geq 1$

$$\|x\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p \right)^{1/p}$$

l_1-norm $\Rightarrow \|x\|_1 = \sum_{i=1}^{n} |x_i|$

$\|x\|_1 = 2$

$\|x\|_2 = \sqrt{2}$

$\|x\|_\infty = 1$

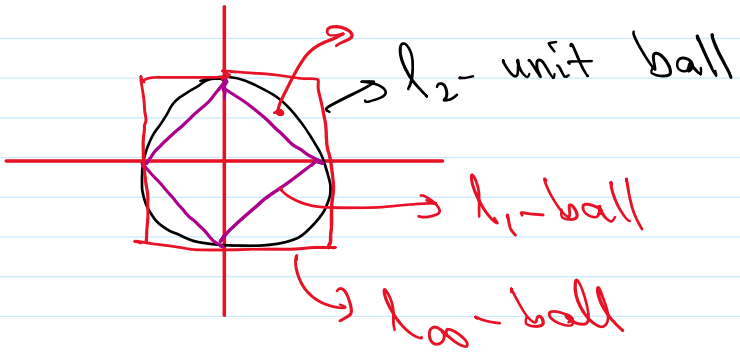
$p \to \infty$

l_∞-norm $\Rightarrow \|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$

↳ Chebyshev norm

Unit ball in $\mathbb{R}^n$

$$B = \{ x \in \mathbb{R}^n, \ \|x\|_p \leq 1 \}$$



All norms induce distances (metrics) on the vector space.

$\text{dist}(x,y) = \|x-y\|_p$

- $\text{dist}(x,y) = \text{dist}(y,x)$
- $\text{dist}(x,y) \geq 0 \quad \forall\ x,y$
- $\text{dist}(x,y) = 0$ only if $x=y$
- $\text{dist}(x,y) \leq \text{dist}(x,z) + \text{dist}(z,y)$

Space of Symmetric matrices : $S^n \Rightarrow n \times n$ Symmetric matrices

Eigenvalue Decomposition

A matrix $M \in \mathbb{R}^{n \times n}$ said to have an eigenvalue decomposition if

$$M = U \Lambda U^{-1} \Rightarrow M^T = U^{-T} \Lambda^T U^T$$
$$= U \Lambda U^T$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$

$\lambda_1, \ldots, \lambda_n \Rightarrow$ eigenvalues of M

Columns of $U \Rightarrow$ eigenvectors of $M \Rightarrow u_1, \ldots, u_n$

If u_i is an eigenvector $\Leftrightarrow$

$$Mu_i = \lambda_i u_i$$

Let $M \in S^n \Rightarrow$ Then M always has an EVD and the eigenvalues are always real.

$$M^T = M \Leftrightarrow U \Lambda U^T = M$$

$U^T U = U U^T = I$

The eigenvectors are orthogonal to each other.
They are basis of $\mathbb{R}^n$.

$S_+^n \subset S^n$: S_+^n : Positive Semidefinite (symmetric matrices)
↳ PSD
↳ $\lambda_1, \dots, \lambda_n \geq 0$

$S_{++}^n \subset S^n$: S_{++}^n : Positive definite (symmetric matrices)
$\lambda_1, \lambda_2, \dots, \lambda_n > 0$ ↳ PD

If $P \in S_+^n \Leftrightarrow P \succeq 0$

$P \in S_{++}^n \Leftrightarrow P \succ 0$

If $P \in S_+^n \Leftrightarrow x^T P x \geq 0 \quad \forall x \in \mathbb{R}^n$

If $P \in S_{++}^n \Leftrightarrow x^T P x > 0 \quad \forall x \in \mathbb{R}^n$

If $P \in S_+^n \Rightarrow$ Define $P^{1/2}$

$$P^{1/2} = U \Lambda^{1/2} U^T$$

↳ $\text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_n})$

$P^{1/2} P^{1/2} = P$

$(\dots^{1/2} \dots^T)(\dots^{1/2} \dots^T)$

$$(U\Lambda^{1/2}U^T)(U\Lambda^{1/2}U^T) = P$$

$$= I$$

$$U\Lambda^{1/2}\Lambda^{1/2}U^T = U\Lambda U^T = P$$

P-Quadratic norm

Let $P \in S^n_+$

$$\|x\|_P = \underbrace{(x^TPx)^{1/2}}_{(x^Tx)^{1/2} = \|x\|_2} = \|P^{1/2}x\|_2$$

$P = I$

$$\Leftrightarrow \|x\|_2 = \|x\|_I$$

$P \succ Q \Leftrightarrow P - Q \succ 0 \Rightarrow$ Notation

Norms on matrices $\Rightarrow X \in \mathbb{R}^{m\times n}$

$$\|X\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2}$$

$$\|X\|_{sav} = \sum_{i=1}^{m}\sum_{j=1}^{n}|X_{ij}|$$

$$\|X\|_{max} = \max\{|X_{ij}|,\ i=1,\dots,m;\ j=1,\dots,n\}$$

All norms in finite-dimensional spaces (in particular, $\mathbb{R}^n$)

All norms in finite-dimensional spaces (in particular, $\mathbb{R}^n$) are "equivalent".

There always exists scalars α and β (which may depend on n) for given norms $\|\cdot\|_a$ and $\|\cdot\|_b$ s.t.

$$\alpha \|x\|_a \leq \|x\|_b \leq \beta \|x\|_a$$

e.g. $\|x\|_1 \leq n \|x\|_\infty$

$\hookrightarrow \beta = n$

Class #4

Friday, January 31, 2025 10:51 AM

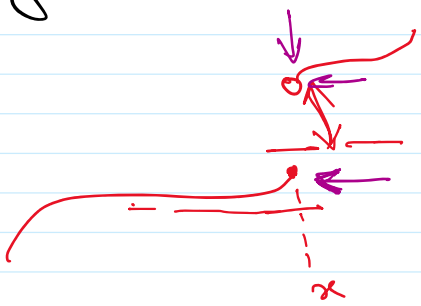
Calculus Concepts

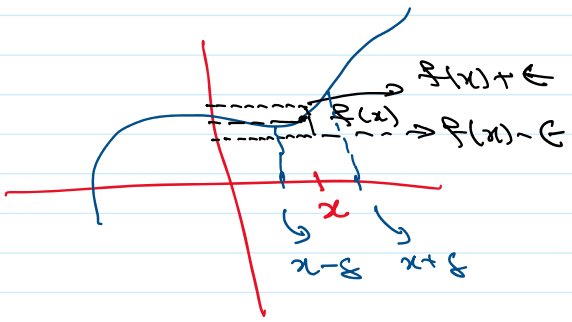
Continuous functions

A function $f: \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $x \in \mathbb{R}^n$ if for any $\epsilon > 0$, $\exists\ \delta > 0$ such that

$$\|y - x\|_2 \leq \delta \implies \|f(y) - f(x)\|_2 \leq \epsilon$$

for all $y \in \text{dom} f$.





$\iff$ Given any sequence $x_1, x_2, \ldots$

Let $\lim_{n\to\infty} x_n = x$

f is continuous at x if and only if

$$\lim_{n\to\infty} f(x_n) = f\left(\lim_{n\to\infty} x_n\right) = f(x)$$

f is a continuous function if it is continuous at all $x \in \text{dom} f$

Derivative of a function $f: \mathbb{R}^n \to \mathbb{R}^m$

Suppose $f: \mathbb{R}^n \to \mathbb{R}^m$ and $x \in \text{int}(\text{dom} f)$

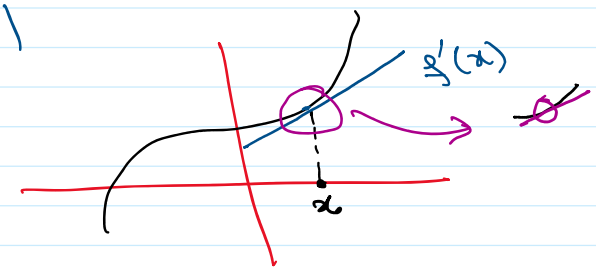
The function f is differentiable at $x \in \mathbb{R}^n$ if $\exists$ a matrix $Df(x) \in \mathbb{R}^{m \times n}$ that satisfies:

$$\lim_{\substack{z \in \text{dom} f,\ z \neq x \\ z \to x}} \frac{\| f(z) - f(x) - Df(x)(z-x) \|_2}{\| z - x \|_2} = 0$$

$Df(x)$ is called derivative of f at x
↳ when $m > 1 \Rightarrow$ Jacobian of f.

Special case: $n = 1$, $m = 1$

$$\lim_{\substack{z \in \text{dom} f,\ z \neq x \\ z \to x}} \frac{| f(z) - f(x) - f'(x)(z-x) |}{|z - x|} = 0$$



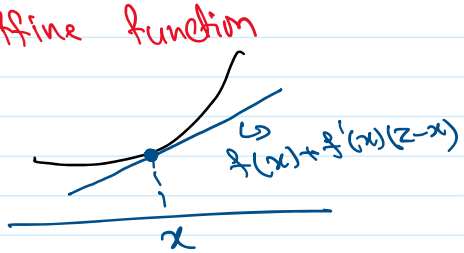
The function f is differentiable if $\text{dom} f$ is open and f is differentiable at every $x \in \text{dom} f$.

Derivatives provide first-order (or linear) approximation of f at x

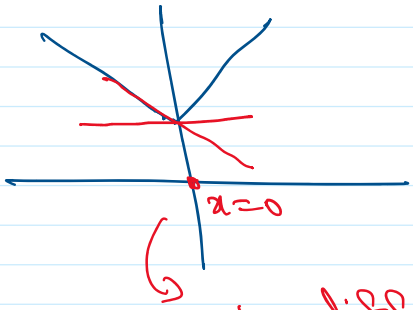
$$\hat{f}(z) = f(x) + Df(x)(z-x)$$

↳ linear function of z

→ Affine function



$Df(x)$ is unique.



First-order approximation of a function

Not differentiable at $x=0$

$f(x) = 3x + 2$

$f'(x) = 3 \Rightarrow$ $\hat{f}(z) = 3x + 2 + 3(z-x)$

$= 3z + 2 + 3x - 3x$

$= 3z + 2$

$$[Df(x)]_{ij} = \frac{\partial f_i(x)}{\partial x_j} \; ; \quad \begin{matrix} i = 1, \dots, m \\ j = 1, \dots, n \end{matrix}$$

$f(x) : \mathbb{R}^2 \to \mathbb{R} \; ; \quad f(x) = x_1^2 + x_2^2$

$Df(x) : \mathbb{R}^{1\times 2} \; ; \quad [Df(x)]_{11} = \frac{\partial f(x)}{\partial x_1} = 2x_1$

$[Df(x)]_{12} = \frac{\partial f(x)}{\partial x_2} = 2x_2$

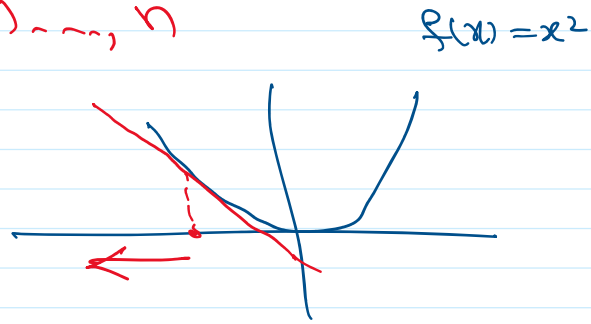
$$Df(x) = \begin{bmatrix} 2x_1 & 2x_2 \end{bmatrix}$$

when $m = 1 \Rightarrow Df(x)$ is a row vector of length n

∇

$Df(x)^T = \nabla f(x) \Rightarrow$ Gradient of the function at x

$$[\nabla f(x)]_i = \frac{\partial f(x)}{\partial x_i}, \quad i = 1, \ldots, n$$



Scalar-valued functions

$$\hat{f}(z) = f(x) + \nabla f(x)^T (z - x)$$

$\hat{f}(z) \to f(z)$ as $z \to x$

Example

① $f(x) = \frac{1}{2} x^T P x + q^T x + r \quad ; \quad f: \mathbb{R}^n \to \mathbb{R}$

$P \in S^n$ and $q \in \mathbb{R}^n$ and $r \in \mathbb{R}$

$g(x) = ax^2 + bx + c$

$g'(x) = 2ax + b$

$Df(x)$?

$\nabla f(x) = Df(x)^T$

$Df(x) = 2 \cdot \frac{1}{2} x^T P + q^T$

$\nabla f(x) = Px + q$

② $f(x) = \log(\det x) \quad ; \quad \det x > 0$

↳ $\text{dom } f = S^n_{++}$ ⇒ Positive definite matrices

$Df(x)$ or $\nabla f(x)$?

Make use of the fact that $Df(x)$ provides a linear approximation of $f(z)$, when z is 'close' to x.

Consider $Z \in S^n_{++}$ such that Z is close to X

$\Rightarrow Z = X + \Delta X$, where $\Delta X \to 0$ as $Z \to X$

Note: we cannot claim that $\Delta X \in S^n_{++}$

Goal: Show that $f(Z) = f(X) + Df(X)(Z-X)$

as $Z \to X \Rightarrow Df(X)$ would be our derivative

$$Z = X + \Delta X$$

$$= \left(X^{1/2} \left(I + X^{-1/2} \Delta X X^{-1/2} \right) X^{1/2} \right)$$

$$\underbrace{\log\det(Z)}_{f(Z)} = \log\det\left(X^{1/2} \left(I + X^{-1/2} \Delta X X^{-1/2} \right) X^{1/2} \right)$$

$$= \log\left[(\det X^{1/2}) \left(\det\left(I + X^{-1/2} \Delta X X^{-1/2} \right) \right) (\det X^{1/2}) \right]$$

$$= \log\left[\frac{\det X^{1/2} \cdot \det X^{1/2}}{\det X} \cdot \det\left(I + X^{-1/2} \Delta X X^{-1/2} \right) \right]$$

$$= \underbrace{\log\det X}_{f(X)} + \underbrace{\log\det\left(I + X^{-1/2} \Delta X X^{-1/2} \right)}_{\bigcirc} (Z-X)$$

$I + X^{-1/2} \Delta X X^{-1/2}$ $\Rightarrow$ This has eigenvalue decomposition

$\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of $X^{-1/2} \Delta X X^{-1/2}$

eigenvalues of $(I + A) = 1 +$ eigenvalues (A)

$1 + \lambda_i$, $i = 1, \ldots, n$ are the eigenvalues of

$1+\lambda_i$, $i=1,\dots,n$ are the eigenvalues of

$$\det\left(I + X^{-1/2}\Delta X X^{-1/2}\right)$$

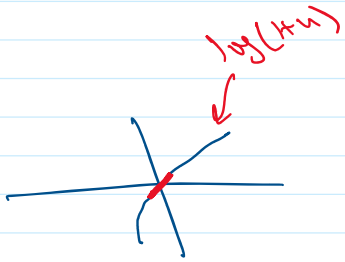
$$= \prod_{i=1}^{n}(1+\lambda_i)$$

$$f(Z) = f(X) + \log\left(\prod_{i=1}^{n}(1+\lambda_i)\right)$$

$$= f(X) + \sum_{i=1}^{n}\log(1+\lambda_i)$$

Since $\Delta X \to 0 \Rightarrow X^{-1/2}\Delta X X^{-1/2} \to 0$

$\Rightarrow \lambda_i \to 0 \ \forall i$



$\Rightarrow$ As $Z \to X$, $\lambda_i \to 0$

$\log(1+u) \geq u$ for u very small

$$\Rightarrow \sum_{i=1}^{n}\log(1+\lambda_i) \geq \sum_{i=1}^{n}\lambda_i \quad \text{when } Z \to X$$

$$f(Z) = f(X) + \sum_{i=1}^{n}\lambda_i \quad \text{as } Z \to X$$

λ_i: eigenvalues of $X^{-1/2}\Delta X X^{-1/2}$

Σ eigenvalues of $A = \mathrm{tr}(A)$

$$= f(X) + \mathrm{tr}\left(X^{-1/2}\Delta X X^{-1/2}\right)$$

$\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$

$$= f(x) + \operatorname{tr}\left(X^{-1/2} X^{-1/2} \Delta X\right)$$

$$= f(x) + \operatorname{tr}\left(X^{-1} \Delta X\right)$$

$$\hookrightarrow (z - x)$$

$$f(z) = f(x) + \langle X^{-1}, z - x \rangle$$

$$\nabla f(x) = X^{-1}$$ ▨

Chain rule

It applies in higher dimensions also.

scalar: $h(u) = g(f(u))$

$$h'(u) = g'(f(u))\, f'(u)$$

If $h(x) = g(f(x)) \quad ; \quad x \in \mathbb{R}^n$

$$Dh(x) = \underset{f(x)}{Dg}\,(f(x)) \cdot \underset{x}{D} f(x)$$

Example: Say $g(x) = f(Ax+b)$

$$Dg(x) = Df(Ax+b) \cdot D(Ax+b)$$

$$= Df(Ax+b) \cdot A$$

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

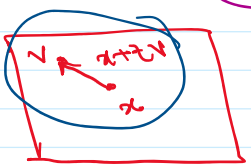
Directional derivative of a function

Let $f: \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$

Directional derivative along a vector $v \in \mathbb{R}^n$ at $x \in \mathbb{R}^n$ is defined as follows:

$$\tilde{f}(t) = f(x+tv) : \mathbb{R} \to \mathbb{R}$$

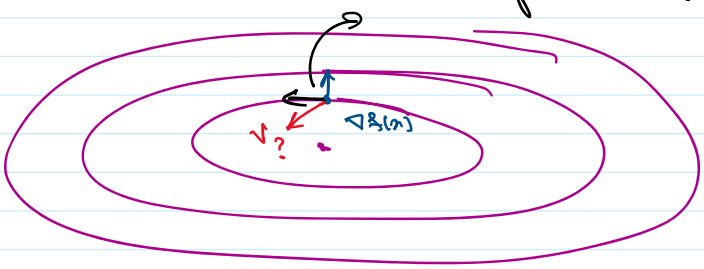
$$f\left(x + t\frac{v}{\|v\|}\right)$$



$$\tilde{f}'(t) = Df(x+tv) \cdot D(x+tv)$$
$$= \nabla f(x+tv)^T \cdot v$$

Directional derivative of $f(x)$ in the direction v is defined as $\tilde{f}'(0) = \nabla f(x)^T v$

rate of change = 0



Algorithms for unconstrained Optimization

$f: \mathbb{R}^n \to \mathbb{R}$; dom f

$$\min_{x \in \text{dom } f} f(x)$$

$$\min_{x \in \text{dom } f} f(x)$$

Assume the minimum value is attained by $f(x)$.

Define $p^* = \min_x f(x)$

An optimization algorithm is an iterative method that produces a sequence of points $x^{(0)}, x^{(1)}, \ldots, \in \text{dom } f$ such that

$$f(x^{(k)}) \longrightarrow p^* \quad \text{as} \quad k \to \infty$$

In the case when $\operatorname{Arg\,min}_x f(x)$ is unique ($= x^*$) then we also hope that $x^{(k)} \to x^*$ as $k \to \infty$

Suppose $f(\cdot)$ is continuous

$$f(x^{(k)}) \longrightarrow f(x^*) = p^*$$

Question: when to terminate the algorithm?

maybe: $f(x^{(k)}) - p^* \leq \epsilon$ for ϵ small $\Rightarrow$ terminate

↳ we do not know $p^* \Rightarrow$ not practical solution.

Search Direction-based Iterative Optimization Algorithms

Pseudo Code

Initialize: $x^{(0)} \in \text{dom } f$

$k \leftarrow 0$

while Stopping Criterion not Satisfied

$$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$$

$$k \leftarrow k+1$$

do

$t^{(k)} \in \underbrace{\mathbb{R}_{++}}_{>0}$ → Step size → (learning rate in Stochastic optimization / machine learning)

$\Delta x^{(k)}$ → Search direction at time k

$x^{(k)}$ ⇒ Iterate

Main challenge: ① How to pick $\Delta x^{(k)}$?
② How to pick $t^{(k)}$?

Class #6

Friday, February 7, 2025 10:20 AM

Stopping Criterion

- Objective function stops changing significantly

$$|f(x^{(k)}) - f(x^{(k-1)})| \leq \epsilon \quad \text{for } \epsilon \text{ small}$$

e.g., $\epsilon = 10^{-8}$

- Iterates stop changing significantly

$$\|x^{(k)} - x^{(k-1)}\| \leq \epsilon \quad \text{for } \epsilon \text{ small}$$

- Function gradient evaluated at the iterates becomes very small

$$\|\nabla f(x^{(k)})\| \leq \epsilon \quad \text{for } \epsilon \text{ small}$$

Descent Optimization Methods

An optimization method is termed a 'descent method' if

$$f(x^{(k+1)}) < f(x^{(k)}) \quad \forall \ k, \text{ except when } x^{(k)} = x^*.$$

↳ strict inequality

when we initialize at $x^{(0)}$

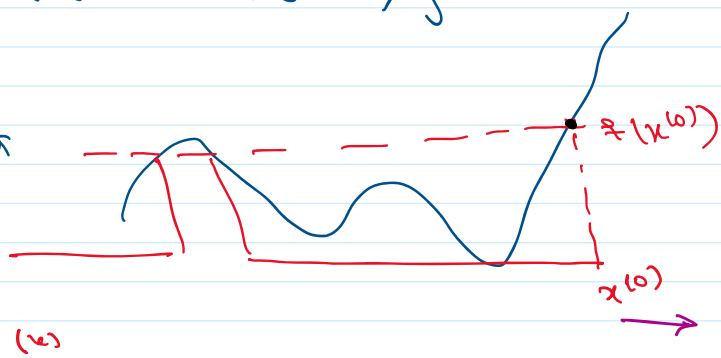
All iterates $x^{(k)}$ for a descent method stay with the set

$$\{ \quad \}$$

the set

$$S = \{x : f(x) \le f(x^{(0)})\}$$

Sublevel set of $f(x)$ at $f(x) = f(x^{(0)})$



$$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$$

In descent methods, $\Delta x^{(k)}$ is called descent direction.

What direction is a descent direction?

We will focus on all continuously differentiable functions and answer this question.

$$C^1 = \{f(x) : f(x) \text{ has continuous derivatives at all } x \in \text{dom } f\}$$

① Make use of Taylor's theorem

Since $f \in C^1$, we have that

$$f(z) = f(x) + \nabla f(x)^T (z-x) + h(x)\|z-x\|_2$$

where $h(x) \to 0$ as $z \to x$ faster than $\|z-x\|_2$

$$f(z) = f(x) + \nabla f(x)^T (z-x) + o(\|z-x\|_2)$$

e.g. $n^{-(2+\epsilon)}$

e.g. $g(n) = n^{-(2+\epsilon)}$

$g(n) = o(n^{-2})$

$$\frac{h(x)\|z-x\|_2}{\|z-x\|_2} \to 0$$

Take

$x = x^{(k)}$

$z = x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$

$$f(x^{(k+1)}) = f(x^{(k)}) + \nabla f(x^{(k)})^T (t^{(k)}\Delta x^{(k)}) + o(\|t^{(k)}\Delta x^{(k)}\|_2)$$

$$= f(x^{(k)}) + t^{(k)}\nabla f(x^{(k)})^T \Delta x^{(k)} + \underbrace{o(t^{(k)}\|\Delta x^{(k)}\|_2)}_{\geq 0}$$

Let $t^{(k)} \to 0$ (very small)

$$f(x^{(k+1)}) = f(x^{(k)}) + t^{(k)}\nabla f(x^{(k)})^T \Delta x^{(k)}$$

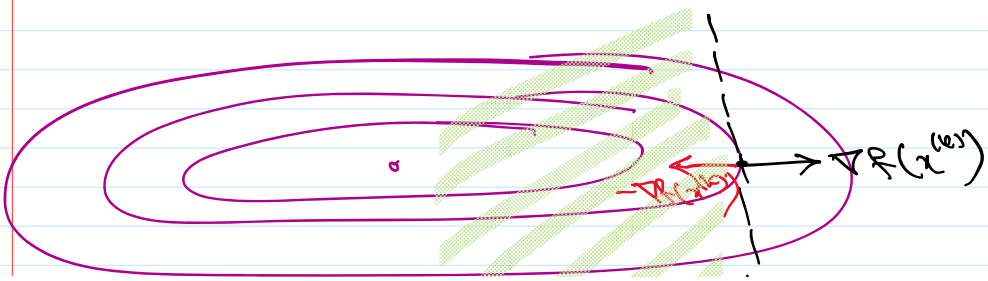
$f(x^{(k+1)}) < f(x^{(k)})$

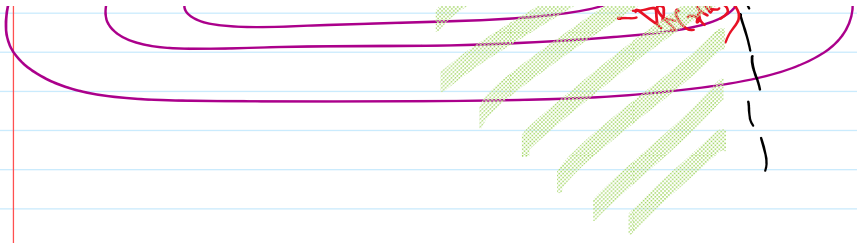
$\iff \nabla f(x^{(k)})^T \Delta x^{(k)} < 0$

$\iff \left[-\nabla f(x^{(k)})\right]^T \Delta x^{(k)} > 0$ ★

$\Delta x^{(k)}$ is a descent if and only if

It makes an acute angle with $-\nabla f(x^{(k)})$





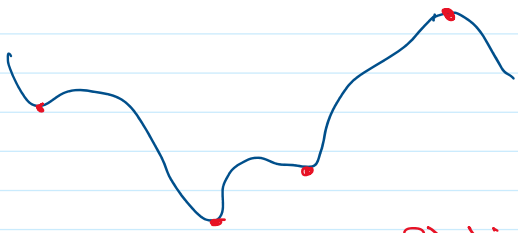
In particular, $\Delta x^{(k)} = -\nabla f(x^{(k)})$ is a descent direction

In a descent method, the optimization method "might" not be able to reduce function value further when

$$\nabla f(x^{(k)}) = 0$$

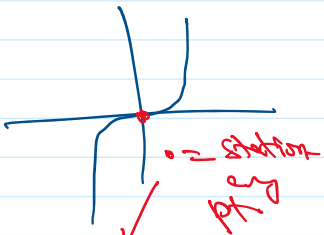
Strict use of a descent method means that any x for which $\nabla f(x^{(k)}) = 0$ is a fixed point of the method.

Any $x \in \text{dom } f$ for which $\nabla f(x) = 0$ is called a stationary point of f.



• = Stationary points

$f(x) = x^3$



• = Stationary pt

Saddle point

General form of descent direction

$\Delta x^{(k)}$ is a descent direction when

$$\boxed{\Delta x^{(k)} = -B^{(k)} \nabla f(x^{(k)})} \quad (*)$$

where $B^{(k)}$ is a positive definite matrix; $\in S^n_{++}$

Descent methods $\Rightarrow x^{(k+1)} = x^{(k)} - t^{(k)} B^{(k)} \nabla f(x^{(k)})$

$$\left[-\nabla f(x^{(k)})\right]^T \Delta x^{(k)}$$

$$= -\nabla f(x^{(k)})^T \left(-B^{(k)} \nabla f(x^{(k)})\right)$$

$$= \nabla f(x^{(k)})^T B^{(k)} \nabla f(x^{(k)})$$

> 0 b/c $B^{(k)}$ is PD.

Based on the choice of $B^{(k)}$, we have different names for descent methods.

① Gradient descent : $B^{(k)} = I$

$$\Leftrightarrow \Delta x^{(k)} = -\nabla f(x^{(k)})$$

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

② Newton's method

$$B^{(k)} = \text{Hessian matrix at } x^{(k)}$$

$$= \nabla^2 f(x^{(k)})$$

Hessian matrix $\Rightarrow$ matrix of second-order partial derivatives.

↳ Another type is called Quasi-Newton method, in which $B^{(k)}$ is built from $\nabla f(x^{(k)})$, but is meant to [illegible]

in which $B^{(k)}$ is built from $\nabla f(x^{(k)})$, but is meant to approximate $\nabla^2 f(x^{(k)})$

③ Steepest descent

$B^{(k)}$ is chosen based on the geometry of the function.

Issue: The previous analysis guarantees a descent direction, but only when $t^{(k)}$ is very small.

Can we use descent methods with a larger step size?
↳ we can, but we need to assume additional regularity on the function.

$C^1_L(\mathbb{R}^n) = \{ f(x) : f$ has continuous derivatives that are L-Lipschitz continuous$\}$

→ L-smooth functions

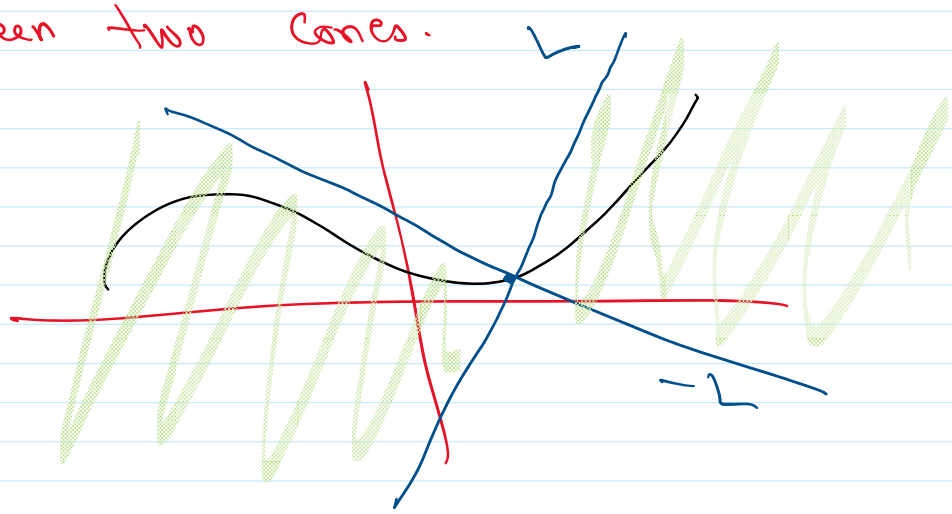
$C^1_L \subseteq C^1$

The gradients $\nabla f(x)$ of $f(x)$ are called L-Lipschitz continuous if and only if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x-y\|_2, \forall x, y \text{ domf}$$

L → Lipschitz constant

Lipschitz Continuity is a stronger form of continuity, which says that the function value always lies between two cones.



e.g.; ① $f(x) = x^2$

$f'(x) = 2x$

$|f'(x) - f'(y)| = |2x - 2y| \leq 2|x-y|$

$L = 2$

② $f(x) = x_1^2 + x_2^2 \longrightarrow x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$\nabla f(x) = 2x = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$

$\|\nabla f(x) - \nabla f(y)\|_2 = \|2x - 2y\|_2 \leq 2\|x-y\|_2$

$L = 2$

Class #7

Tuesday, February 11, 2025 10:19 AM

$f \in C^1_L(\mathbb{R}^n) \Rightarrow$ Continuously differentiable functions with Lipschitz gradients

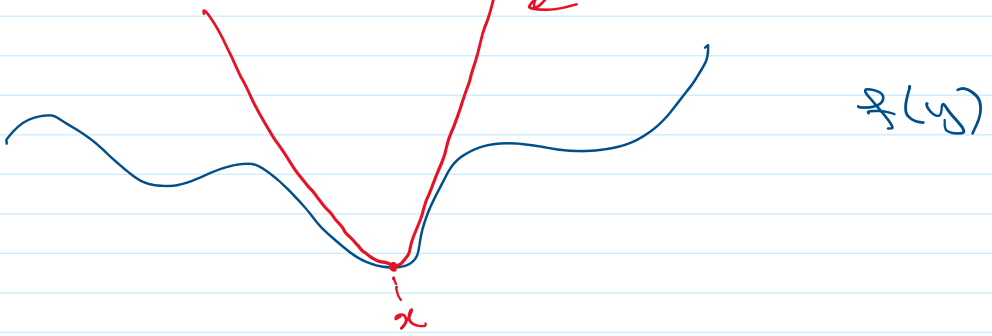
Lemma: Quadratic upper bound on $f \in C^1_L(\mathbb{R}^n)$

Let $f \in C^1_L(\mathbb{R}^n)$ and $\operatorname{dom} f$ is a convex set.

Then; $\forall\, x, y \in \operatorname{dom} f$

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

(quadratic function in y)



Proof:

Define $g(t) = f(x + t(y-x))$ for $t \in [0,1]$

Since $\operatorname{dom} f$ is convex $(x + t(y-x))$ is in the $\operatorname{dom} f$.

$$\Rightarrow x + ty - tx = (1-t)x + ty$$

Note: $g(1) = f(y)$

$g(0) = f(x)$

$$g'(t) = \nabla f(x + t(y-x))^T (y-x)$$

$$\boxed{g'(0) = \nabla f(x)^T (y-x)}$$

$$g'(0) = \nabla f(x)^T (y-x)$$

$$g'(t) - g'(0) = \nabla f(x + t(y-x))^T (y-x) - \nabla f(x)^T (y-x)$$

$$= \left[\nabla f(x + t(y-x)) - \nabla f(x)\right]^T (y-x)$$

Use Cauchy-Schwarz inequality ($a^T b \le \|a\|_2 \|b\|_2$)

$$\le \|\nabla f(x+t(y-x)) - \nabla f(x)\|_2 \, \|y-x\|_2$$

Lipschitz continuous gradients

$$\le L \|x + t(y-x) - x\|_2 \, \|y-x\|_2$$

$$g'(t) - g'(0) \le tL \|y-x\|_2^2$$

$$\int_0^1 g'(t)\,dt = g(1) - g(0)$$

$$\Rightarrow g(1) = g(0) + \int_0^1 g'(t)\,dt$$

$$\le g(0) + \int_0^1 \left(tL\|y-x\|_2^2 + g'(0)\right) dt$$

$$f(y) \le f(x) + L\|y-x\|_2^2 \left.\frac{t^2}{2}\right|_0^1 + g'(0)$$

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

Descent lemma (when $\Delta x = \nabla f(x)$)

Descent Lemma (when $\Delta x = -\nabla f(x)$)

Let $f \in C^1_L(\mathbb{R}^n)$ with $\mathrm{dom} f$ being convex. Let us consider the iterative method:

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$$

and focus on the case $\Delta x^{(k)} = -\nabla f(x^{(k)})$. Then, as long as $0 < t^{(k)} < \frac{2}{L}$, $\Delta x^{(k)}$ is a descent direction, i.e. $f(x^{(k+1)}) < f(x^{(k)})$ as long as $x^{(k)}$ is not x^*.

Remarks: A range of step sizes guarantee descent. The larger L is, the smaller is the range.

Proof: Let $y = x^{(k)} + t^{(k)} \Delta x^{(k)}$

and $x = x^{(k)}$ in the Quadratic upper bound.

$$f\left(x^{(k)} + t^{(k)} \Delta x^{(k)}\right) \le f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)^T \left(x^{(k)} + t^{(k)} \Delta x^{(k)} - x^{(k)}\right) + \frac{L}{2} \left\| x^{(k)} + t^{(k)} \Delta x^{(k)} - x^{(k)} \right\|_2^2$$

$$f\left(x^{(k+1)}\right) \le f\left(x^{(k)}\right) + t^{(k)} \nabla f\left(x^{(k)}\right)^T \Delta x^{(k)} + t^{(k)2} \frac{L}{2} \left\| \Delta x^{(k)} \right\|_2^2$$

Now put $\Delta x^{(k)} = -\nabla f(x^{(k)})$

$$f\left(x^{(k+1)}\right) \le f\left(x^{(k)}\right) - t^{(k)} \left\| \nabla f\left(x^{(k)}\right) \right\|_2^2 + t^{(k)2} \frac{L}{2} \left\| \nabla f\left(x^{(k)}\right) \right\|_2^2$$

need < 0

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

Clearly, $f(x^{(k+1)}) < f(x^{(k)})$ if and only if

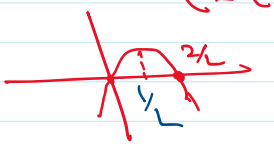
most reduction when this is the largest.

$$t^{(k)} - t^{(k)^2}\frac{L}{2} > 0$$

$$\iff t^{(k)^2}\frac{L}{2} < t^{(k)}$$

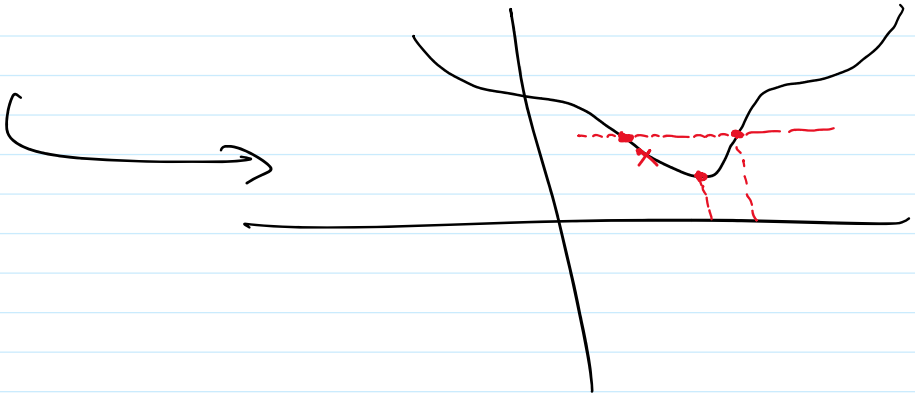
$$\iff t^{(k)} < \frac{2}{L}$$

What is the $t^{(k)}$ when 'L' is known?



The best $t^{(k)}$, when it comes to most reduction for gradient descent is $t^{(k)} = \frac{1}{L}$.

Ex: $f(x)$



General Descent Method

Initialize: $x^{(0)} \in \text{dom } f$

$k \leftarrow 0$

Repeat

1. Determine a descent direction $\Delta x^{(k)}$ (i.e., $-\nabla f(x^{(k)})^T \Delta x^{(k)} > 0$)

2. Line search: Choose a step size $t^{(k)} > 0$

3. Update the iterate: $x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$

Until stopping criterion is satisfied.

Special Cases:

(1) Gradient descent: $\Delta x^{(k)} = -\nabla f(x^{(k)})$

(2) Newton's method: $\Delta x^{(k)} = -\nabla^2 f(x^{(k)}) \nabla f(x^{(k)})$

Line Search: How to pick the step size in a descent method?

(1) Choose a fixed step size $t^{(k)} = \eta \ \forall \ k \geq 0$.

e.g., If $f \in C^1_L(\mathbb{R}^n)$ and L is known or can be computed efficiently then pick

$$\eta = \frac{1}{L} \text{ for gradient descent.}$$

In other cases, trial and error helps pick a step size.

Issues with this approach:

- Sometimes it is too costly to compute L.
- Sometimes functions are not C^1_L.
- Even when one has descent in some iterations, does not mean the step size would work in all points in domf.

Unless L is known, the longer the step size, the better 'perhaps'.

② Variable Step Size $t^{(k)}$

↳ The most common approach in the literature.

(a) Decaying step size policy

$$t^{(k)} \to 0 \quad \text{as} \quad k \to \infty$$

Typical policy

(i) $t^{(k)} \to 0$ as $k \to \infty$

(ii) $\sum_{k=0}^{\infty} t^{(k)} = \infty$

(iii) $\sum_{k=0}^{\infty} \left[t^{(k)}\right]^2 < \infty$

e.g., $t^{(k)} = \frac{\text{Const}}{k}$

↳ Often used in machine learning / Stochastic

↳ Often used in machine learning / Stochastic optimization.

(b) Search for a nice step size that reduces the objective function using a subroutine in each iteration k.

↳ often used in practical deterministic optimization problems.

↳ we will study this in detail under 'Exact line search' and 'Inexact line search'.

Exact line Search

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$$

When we have fixed $\Delta x^{(k)}$ and are looking for $t^{(k)}$, we are effectively looking at a one-dimensional function:

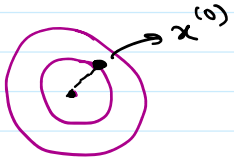
$$\tilde{f}(t) = f(x^{(k)} + t^{(k)} \Delta x^{(k)})$$

Pick $t^{(k)} = \underset{t \geq 0}{\arg\min}\ \tilde{f}(t)$

Example: Say $f(x) = x_1^2 + x_2^2$

$$\nabla f(x) = 2x$$

Let $x^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \mathbb{1}$



$$\Rightarrow x^{(1)} = x^{(0)} - 2t^{(1)} x^{(0)}$$

$= (1-2t^{(1)})\nabla f$ $= (1-\frac{1}{2}\times 2)\nabla f = 0 \cdot \nabla f = [0]$

$f^2(t) = f(x^{(1)}) = (1-2t)^2 \cdot 1 + (1-2t)^2 \cdot 1$

$= 2(1-2t)^2$

$t^{(1)} = \frac{1}{2}$

$= \underset{t \geq 0}{\operatorname{argmin}}\ f^2(t) = \underset{t \geq 0}{\operatorname{argmin}}\ 2(1-2t)^2$

$f^{2\prime}(t) = 4(1-2t)\cdot -2 = -8(1-2t)$

$f^{2\prime}(t) = 0 \Rightarrow -8(1-2t) = 0$

$t = \frac{1}{2}$

Exact line search, in which one solves a one-dim optimization problem in each iteration, works in cases where:

① The solution has an analytical form.

② It might be computationally feasible to numerically solve the 1-D problem.

But if the cost of exact line search is too much, we resort to 'inexact line search'.

↳ Backtracking (Armijo-Goldstein line search)

Class #8

Friday, February 14, 2025 10:17 AM

$f \in C^1_L(\mathbb{R}^n) \Rightarrow f$ is L-Smooth

Function is Smooth (or non-smooth)

Quadratic Function $f(x) = \frac{1}{2} x^T Q x + p^T x + r$

Generally, we focus on $f(x)$ being convex

$\Rightarrow Q$ is Positive Semi definite or Positive definite

$f(x) = ax^2$

↳ Convex when $a > 0$

Special case: $f(x) = \frac{1}{2} x^T Q x$; Q is S.P.D.

$Q \in S^n_{++}$

$f(x)$ in this case is L-Smooth

Linear algebra review (Appendix A.1.5)

Operator norm of a matrix

↳ Singular Value Decomposition of a matrix

Every matrix $A \in \mathbb{R}^{m \times n}$ has a Singular Value decomposition

$A = U \Sigma V^T$ → V is $n \times n$ matrix with orthonormal columns

↳ right singular vectors.

U: $m \times m$

Σ: diagonal matrix, $m \times n$

$m \times m$ with orthonormal columns → left singular vectors of A

diagonal matrix $m \times n$ → singular values ≥ 0

right singular vectors.

$A v_i = \sigma_i u_i$

2×3 $\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{bmatrix}$

$\Sigma^T \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

$AA^T = (U \Sigma V^T)(V \Sigma^T U^T)$

$= U \Sigma \Sigma^T U^T$ ($V^T V = I$)

$= U \Lambda U^T$

$(m \times n)(n \times m)$

$(m \times m)$

entries are σ_i^2

Symmetric

Eigenvalue decomposition of AA^T

① Singular values of A are the square root eigenvalues of AA^T / A^TA

② Left singular vectors of A are the eigenvectors of AA^T

$A^TA = V \Lambda V^T$

entries are σ_i^2

$\Sigma^T \Sigma$

$n \times n$ $n \times n$ $n \times n$

③ Right singular vectors of A are the eigenvectors of A^TA

The operator norm of a matrix A with respect to the Euclidean norm is defined as:

$$\|A\|_2 = \sup\left\{\|Ax\|_2 : \|x\|_2 \le 1\right\}$$

$$\|A\|_2 = \sigma_{max}(A) \Rightarrow \text{maximum Singular value of } A$$

$$= \sqrt{\lambda_{max}(A^TA)} = \sqrt{\lambda_{max}(AA^T)}$$

$\lambda_{max}(A^TA)$: maximum eigen value of A^TA

Submultiplicative property

$$\|Az\|_2 \le \|A\|_2 \|z\|_2$$

General Definition of Operator Norm

$$\|A\|_{a,b} = \sup\left\{\|Ax\|_a : \|x\|_b \le 1\right\}$$

Special Cases: $a = b$

1. $a = 2$
2. $a = 1$
3. $a = \infty$

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |A_{ij}|$$

$$\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |A_{ij}|$$

Problem: Let $f(x) = \frac{1}{2}x^TQx$ with $Q \in S^n$.

Might not be convex

Prove that $f(x)$ is L-Smooth and derive the

Prove that $f(x)$ is L-smooth and derive the value of L.

$f \in C^1_L$

Solution:

$$\nabla f(x) = Qx$$

Let x and $y \in \mathbb{R}^n$

$$\|\nabla f(x) - \nabla f(y)\|_2 = \|Qx - Qy\|_2$$

$$= \|Q(x-y)\|_2$$

$$\leq \underbrace{\|Q\|_2}_{L} \|x-y\|_2$$

$$L = \|Q\|_2 = \sqrt{\lambda_{max}(Q^T Q)}$$

$$= \sqrt{\lambda_{max}(Q^2)} = \sqrt{\lambda_{max}(Q)^2}$$

$$= \lambda_{max}(Q)$$

A with EVD $A = U \Lambda U^T$

$A^n = U \Lambda^n U^T$

Does Gradient Descent Converge?
If it does, at what rate?

Assume $f \in C^1_L(\mathbb{R}^n)$ and $p^* = \operatorname{argmin} f(x)$ exists.

$$\Downarrow$$

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

Put $y = x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f(x^{(k)})$

$x = x^{(k)}$

$t^{(k)} = \frac{1}{L}$

$$f(x^{(k+1)}) \le f(x^{(k)}) - \frac{1}{L}\|\nabla f(x^{(k)})\|_2^2 + \frac{1}{2L}\|\nabla f(x^{(k)})\|_2^2$$

$$f(x^{(k+1)}) \le f(x^{(k)}) - \frac{1}{2L}\|\nabla f(x^{(k)})\|_2^2$$

↳ Descent lemma with $t^{(k)} = \frac{1}{L}$ and gradient descent

---

$$\star \quad f(x^{(k+1)}) - f(x^{(k)}) \le -\frac{1}{2L}\|\nabla f(x^{(k)})\|_2^2$$

Add the above expression for $k = 0, 1, \ldots, n$

$$\sum_{k=0}^{n}\left[f(x^{(k+1)}) - f(x^{(k)})\right] \le -\frac{1}{2L}\sum_{k=0}^{n}\|\nabla f(x^{(k)})\|_2^2$$

telescoping sum

↳ $\lim_{k\to\infty} f(x^{(k+1)}) = f(x^{(n)})$

$f(x^{(n)}) - f(x^{(0)})$

$\sum_{n=0}^{\infty}(a_{n+1} - a_n) \Rightarrow$ telescoping sum

$= \lim_{N\to\infty} \sum_{n=0}^{N}(a_{n+1} - a_n)$

$$\left(a_{N+1} - a_N\right) + \left(a_N - a_{N-1}\right) + \left(a_{N-1} - a_{N-2}\right) + \cdots + \left(a_1 - a_0\right)$$

$$= \lim_{N\to\infty} \left(a_{N+1} - a_0\right) = a_\infty - a_0$$

if $\lim_{N\to\infty} a_N = a_\infty$

$$\Rightarrow f(x^{(\infty)}) - f(x^{(0)}) \leq -\frac{1}{2L} \sum_{k=0}^{\infty} \left\| \nabla f(x^{(k)}) \right\|_2^2$$

$$\sum_{k=0}^{\infty} \left\| \nabla f(x^{(k)}) \right\|_2^2 \leq 2L \cdot \left[f(x^{(0)}) - \underbrace{f(x^{(\infty)})}_{\geq f^*} \right]$$

$$\leq 2L\left(f(x^{(0)}) - f^*\right)$$

B/C $\sum_{k=0}^{\infty} \left\| \nabla f(x^{(k)}) \right\|_2^2 < \infty$

$$\Rightarrow \lim_{k\to\infty} \left\| \nabla f(x^{(k)}) \right\|_2^2 = 0 \iff \lim_{k\to\infty} \nabla f(x^{(k)}) = 0$$

↳ we have converged to a local minimum, or a saddle point.

↳ Basically a first-order stationary point.

Class #9

Tuesday, February 18, 2025 10:20 AM

When $f \in C^1_L(\mathbb{R}^n)$ and $t^{(k)} = \frac{1}{L}$

$$\|\nabla f(x^{(k)})\|_2^2 \longrightarrow 0 \quad \text{as} \quad k \to \infty$$

$$\nabla f(x^{(k)}) \longrightarrow 0$$

what about the case of variable step sizes?

① Decaying step size policy

② Step size is bounded below:

Let $\epsilon > 0$ be a fixed constant

$$\frac{\epsilon}{L} \leq t^{(k)} \leq \frac{2-\epsilon}{L}$$

↳ Same proof works.

What about the rate of convergence?

$f \in C^1_L(\mathbb{R}^n)$; $t^{(k)} = \frac{1}{L}$

↳ from the previous lecture:

$$(\star) \qquad \|\nabla f(x^{(k)})\|_2^2 \leq 2L \left[f(x^{(k)}) - f(x^{(k+1)}) \right]$$

Sum $(\star)$ from $k=1$ to $k=K$

$$\sum_{k=1}^{K} \|\nabla f(x^{(k)})\|_2^2 \leq 2L \sum_{k=1}^{K} \left[f(x^{(k)}) - f(x^{(k+1)}) \right]$$

$$\leq 2L \sum_{k=0}^{K} \underbrace{\left[f(x^{(k)}) - f(x^{(k+1)}) \right]}_{>0}$$

↳ Telescoping sum

$$\leq 2L \left(f(x^{(0)}) - \underbrace{f(x^{(K+1)})}_{\geq p^*} \right)$$

$$(\star\star) \quad \sum_{k=1}^{K} \left\| \nabla f(x^{(k)}) \right\|_2^2 \leq 2L \left(f(x^{(0)}) - p^* \right)$$

$$\sum_{k=1}^{K} \left\| \nabla f(x^{(k)}) \right\|_2^2 \geq K \min_{k \in \{1,2,\dots,K\}} \left\| \nabla f(x^{(k)}) \right\|_2^2$$

$$K \min_{k \in \{1,\dots,K\}} \left\| \nabla f(x^{(k)}) \right\|_2^2 \leq \underbrace{2L \left(f(x^{(0)}) - p^* \right)}_{\gamma > 0}$$

$$\min_{k \in \{1,\dots,K\}} \left\| \nabla f(x^{(k)}) \right\|_2^2 \leq \frac{\gamma}{K}$$

Within K iterations, we will have at least one $x^{(k)}$ such that $\left\| \nabla f(x^{(k)}) \right\|_2^2 \leq \boxed{\frac{\gamma}{K}} = O\left(\frac{1}{K}\right)$

Suppose we want $\left\| \nabla f(x^{(k)}) \right\|_2^2 \leq \epsilon$ for ϵ very small

$$\Rightarrow \quad \frac{\gamma}{K} \leq \epsilon \quad \Rightarrow \quad K \geq \frac{\gamma}{\epsilon}$$

$$\Rightarrow K = \Omega\left(\epsilon^{-1}\right)$$

Say $\epsilon = 10^{-8}$

$\Rightarrow K = O(10^8)$ iterations

Similar results hold for step size choices for general descent methods, where the step size depends on the descent direction and is strictly lower bounded by $\epsilon > 0$.

Another interpretation of Gradient Descent for $f \in C^1_L(\mathbb{R}^n)$

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

Let us derive an iteration in which x is current iterate and $y = x^+$ is the next iterate

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

we need x^+ such that $f(x^+)$ is as small as possible.

Mirror Descent proximity term (when this is replaced by another $\|\cdot\|$)

we approach this by minimizing the upper bound

$$\tilde{f}(x^+) = f(x) + \nabla f(x)^T (x^+ - x) + \frac{L}{2}\|x^+ - x\|_2^2 \quad \text{w.r.t. } x^+$$

$\|x^+ - x\|_2^2 = (x^+ - x)^T (x^+ - x)$

$= x^{+T}x^+ - x^{+T}x - x^{+T}x + x^T x$

$= x^{+T}x^+ - 2x^{+T}x + x^T x$

$$\nabla_{x^+} \tilde{f}(x^+) = 0 + \nabla f(x) + \frac{L}{2}(2x^+ - 2x + 0)$$

$$= \nabla f(x) + L(x^+ - x) = 0$$

$$L(x^+ - x) = -\nabla f(x)$$

$$\boxed{x^+ = x - \frac{1}{L}\nabla f(x)}$$

Stepsize selection when $f \in C^1_L(\mathbb{R}^n)$ but L is not know or computing it is too expensive.

⇓

Inexact line search.

Backtracking | Armijo rule | Armijo–Goldstein step

Another approach based on Wolfe conditions, but they are harder to compute and we won't study them.

$$\tilde{f}(t) = f(x^{(k)} + t\,\Delta x^{(k)}) \;;\; t \geq 0$$

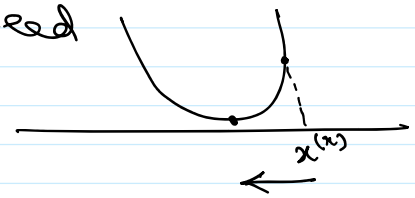
$t = 0 \Rightarrow f(x^{(k)})$

↳ Inexact search line search requires finding a value of $t^{(k)}$ such that

$$\tilde{f}(t^{(k)}) = f(x^{(k)} + t^{(k)}\Delta x^{(k)}) \text{ is}$$

Sufficiently smaller than $f(x^{(k)})$

but there has to be a guaranteed decrease.



Algorithm (Backtracking)

Input: Current iterate x
Search direction Δx

Parameters $\alpha \in (0, 0.5)$ → Sufficient decrease parameter

$\beta \in (0, 1)$

Initialize: $t \leftarrow 1$

While $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$

$t \leftarrow \beta t$

Sufficient decrease condition

The algorithm ends when $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$

< 0

↳ It depends on α.

Back tracking

$t = 1$
$t = \beta$
$t = \beta^2$

$\beta \Rightarrow$ The gridding of $(0, 1]$

↳ larger β can slow down the line search
Smaller β can end up giving you a very

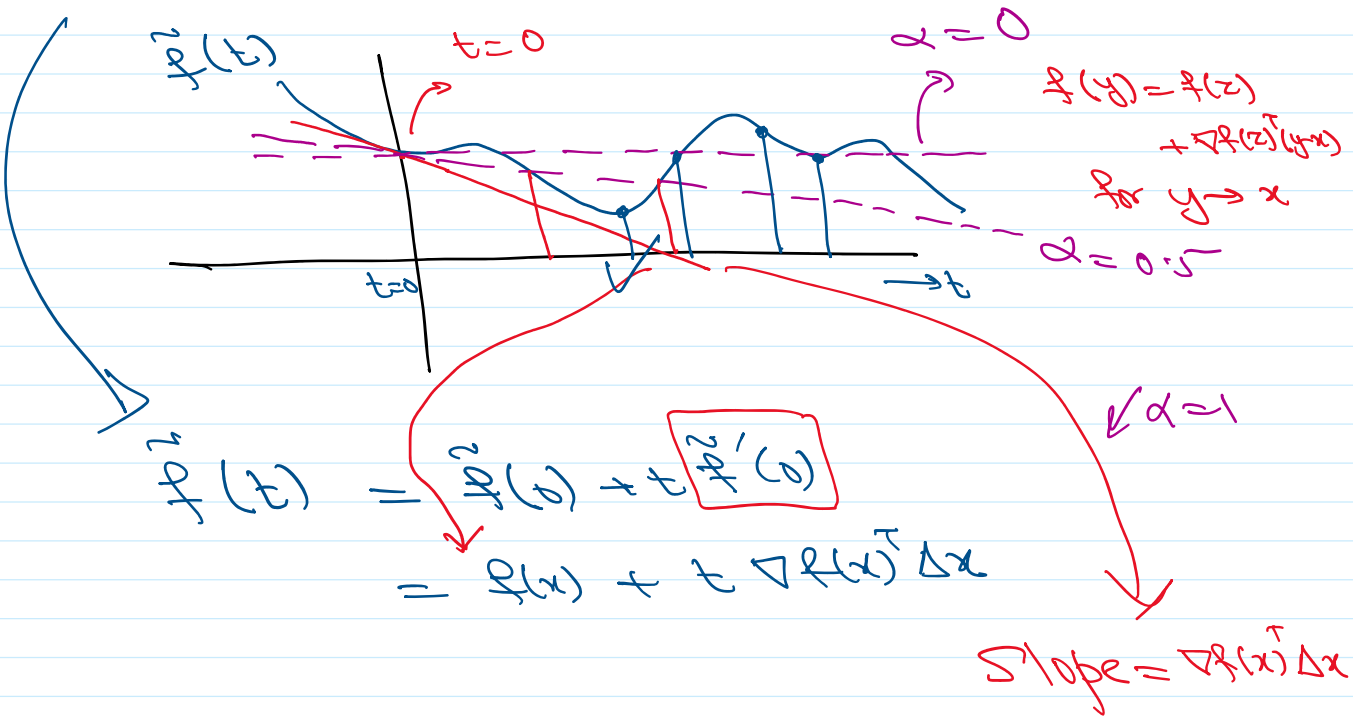
Small step size.

Geometric View of Backtracking

for t small enough

$$\hat{f}(t) = f(x + t\Delta x) \approx f(x) + t\nabla f(x)^T \Delta x$$

$$\hat{f}'(t) = \nabla f(x)^T \Delta x$$



what is the approximation when slope is $\alpha \nabla f(x)^T \Delta x$

$$\hat{f}(t) = f(x) + \alpha t \nabla f(x)^T \Delta x$$

Switching to Newton's Method

Second Derivative of a function $f: \mathbb{R}^n \to \mathbb{R}$.

The second derivative of f, called the Hessian of f, at $x \in \text{int dom} f$, is denoted by $\nabla^2 f(x)$

f, at $x \in \text{int dom} f$, is denoted by $\nabla^2 f(x)$ ($n \times n$ matrix) and is defined as:

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i=1,\dots,n \quad j=1,\dots,n$$

provided f is twice differentiable at x.

Gradient $z \to x \Rightarrow f(z) \gtrsim f(x) + \nabla f(x)^T (z-x)$ as $z \to x$

Hessian, by definition, is a quadratic approximation of f at x

$$f(z) \gtrsim f(x) + \nabla f(x)^T (z-x) + \underbrace{\frac{1}{2}(z-x)^T \nabla^2 f(x)(z-x)}_{\hat{f}(z)}$$

$z \to x$

$$\lim_{\substack{z \in \text{dom} f \\ z \neq x \\ z \to x}} \frac{|f(z) - \hat{f}(z)|}{\|z-x\|_2^2} = 0$$

Note: $D \nabla f(x) = \nabla^2 f(x)$

$\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n$

↳ Hessian is the derivative of the gradient

This term is the derivative of the gradient

Chain rule for Hessians

(1) Suppose $f: \mathbb{R}^n \to \mathbb{R}$, $g: \mathbb{R} \to \mathbb{R}$

$$h(x) = g(f(x))$$

$$\nabla^2 h(x) = g'(f(x))\,\nabla^2 f(x) + g''(f(x))\,\nabla f(x)\nabla f(x)^T$$

(2) Suppose $f: \mathbb{R}^n \to \mathbb{R}$

$g: \mathbb{R}^m \to \mathbb{R}$

$A \in \mathbb{R}^{n \times m}$

$b \in \mathbb{R}^n$

$$g(x) = f(Ax+b)$$

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax+b)\,A$$

(3) Define $\tilde{f}(t) = f(x+tv)$

$$\nabla^2 \tilde{f}(t) = \tilde{f}''(t) = v^T \nabla^2 f(x+tv)\,v$$

For $t=0$: $\nabla^2 \tilde{f}(0) = v^T \nabla^2 f(x)\,v$

Example: $f(x) = \frac{1}{2}x^T P x + q^T x + r$

$P \in S^n$, $q \in \mathbb{R}^n$, $r \in \mathbb{R}$

$$\nabla f(x) = Px + q$$

$$\nabla^2 f(x) = D(\nabla f(x)) = P$$

assume: $f(x) = \frac{1}{2}ax^2 + bx + c$

$f''(x) = a$

Newton's Method

It is "supposed" to be a descent method with iterations given by:

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x_{nt}$$

where: $\Delta x_{nt} = -\left[\nabla^2 f(x)\right]^{-1} \nabla f(x)$ Newton direction

⇓

Based on this, it requires:

① Function f has to be twice differentiable

↳ Typically we assume $f \in C^2$

↳ Twice continuously differentiable

② $\nabla^2 f(x)$ must be invertible $\Rightarrow \text{rank}(\nabla^2 f(x)) = n$

↳ Typical requirement is that it is invertible over every $x \in \mathbb{R}^n$

③ In order for Newton's method to be a descent method, we require that

$$\nabla^2 f(x) > 0 \quad \text{(Positive definite)}$$

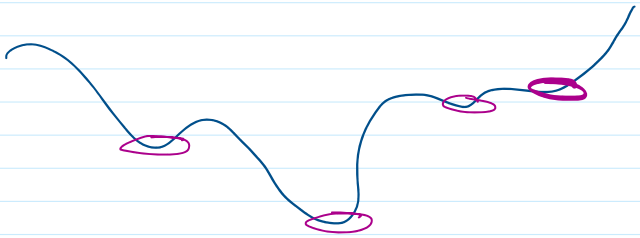
(Remember. $A > 0 \iff A^{-1} > 0$)

There are two ways to handle ③

① Assume $\nabla^2 f(x) > 0 \quad \forall x \in \mathbb{R}^n$

↳ strongly convex functions

② what about non convex functions?



In that case, we first run gradient descent for a number of iterations till $\|\nabla f(x)\|_2$ is small and then we switch to Newton iterations.

Even when $\nabla^2 f(x) \succ 0 \quad \forall x \in \mathbb{R}^n$, Newton's has some drawbacks:

① Compute and store $\nabla^2 f(x)$

② Compute inverse of $[\nabla^2 f(x)]^{-1}$

So why use it? It is extremely fast in the right regions (to be shown later).

Ways to deal with these issues

① Quasi-Newton method

② Approximate the Hessian by looking/exploiting the structure of the problem (in a fast way).

E.g., $\nabla f(x) = \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \\ \vdots \\ f_n(x_n) \end{bmatrix}$

$\Rightarrow \nabla^2 f(x) = \begin{bmatrix} f_1'(x) & & & \\ & f_2'(x_2) & & \\ & & \ddots & \\ & & & f_n'(x_n) \end{bmatrix}$

Interpretations of Newton's Method

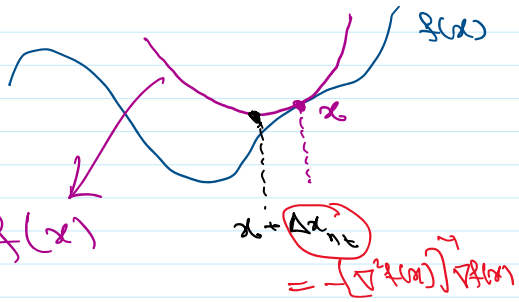
① Minimizer of the second-order approximation of f at x

① Minimizer of the second-order approximation of f at x

$$\hat{f}(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x)$$

write $y = x + v$

$$\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$



argmin

$$= -\left[\nabla^2 f(x)\right]^{-1} \nabla f(x)$$

$= -[\nabla^2 f(x)]^{-1} \nabla f(x)$

Compute ∇_v and set it equal to 0.

$$\nabla_v \left(f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v \right)$$

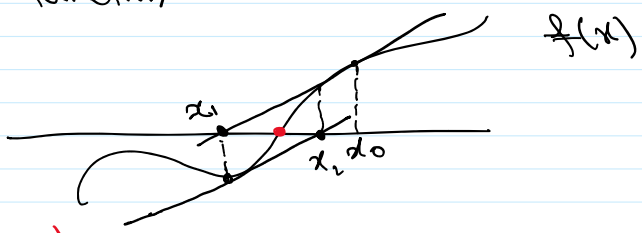
$$= 0 + \nabla f(x) + \nabla^2 f(x) v = 0$$

$$\nabla^2 f(x) v = -\nabla f(x)$$

$$v^* = -\left[\nabla^2 f(x)\right]^{-1} \nabla f(x)$$

② Newton's method is also tied to the idea of approximating the gradient $\nabla f(x)$ by a linear function and then finding the root of that linear function.

Stationary point of a function is when $\nabla f(x) = 0$



Given $f(x) : \mathbb{R} \to \mathbb{R}$

$$\underbrace{f(y)}_{0} \approx f(x) + f'(x)(y-x)$$

$$\nabla f(y) \approx \nabla f(x) + \nabla^2 f(x)(y-x)$$

$$\nabla f(y) \approx \nabla f(x) + \nabla^2 f(x)(y-x)$$

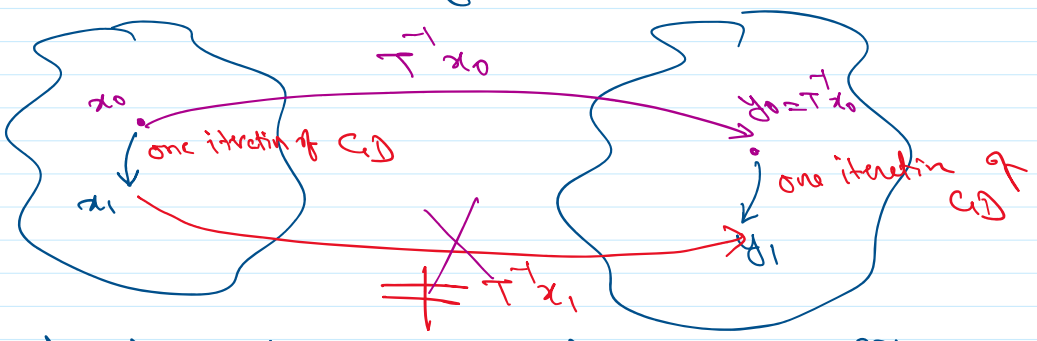
Put $y = x+v$

$$\underbrace{\nabla f(x+v)}_{=0} = \nabla f(x) + \nabla^2 f(x)\, v$$

$$\nabla^2 f(x) v = -\nabla f(x)$$

$$v = -\left[\nabla^2 f(x)\right]^{-1} \nabla f(x)$$

$x \in \mathbb{R}^n \quad \underset{T y}{\overset{T^{-1} x}{\rightleftarrows}} \quad y \in \mathbb{R}^n \qquad (T \in \mathbb{R}^{n \times n}$ is non-singular)



Gradient descent, in general, is not affine invariant. Coordinate system in gradient descent affects the algorithmic performance.

Affine invariance of Newton's Step

Suppose $T \in \mathbb{R}^{n \times n}$ is non-singular and let $y = T^{-1} x$; $\boxed{x = Ty}$

$$f(x) : \mathbb{R}^n \to \mathbb{R}$$

$$\bar{f}(y) = f(Ty) : \mathbb{R}^n \to \mathbb{R}$$

$$\bar{f}(y) = f(Ty) : \mathbb{R}^n \to \mathbb{R}$$

Then: $x + \Delta x_{nt} = T(y + \Delta y_{nt})$

Basic Assumption

Either we are close to a local optimum or we are working with the case $\nabla^2 f(x) \succ 0 \quad \forall x \in \mathbb{R}^n$.

Newton Decrement

$$\lambda(x) = \left[\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right]^{1/2}$$

is called Newton decrement.

1. Used in analysis
2. Used in stopping criterion of Newton's method

- $\lambda(x)$ is a scalar
- $\lambda(x) \geq 0$ since $\nabla^2 f(x)^{-1} \succ 0$

$\lambda(x)$ allows us to approximate how close we are to a local minimum (p^*).

Recall: $\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$

$$f(x) - \hat{f}(x+v) = -\nabla f(x)^T v - \frac{1}{2} v^T \nabla^2 f(x) v$$

$f(x)$: function at x

$\hat{f}(x+v)$: Quadratic approximation of x

$v = -\nabla^2 f(x)^{-1} \nabla f(x)$

$$f(x) - \min_v \hat{f}(x+v) \geq -\nabla f(x)^T \left(-\nabla^2 f(x)^{-1} \nabla f(x)\right)$$

$f(x) - \min_v f(x+v)$ $\underbrace{\qquad}_{p^*}$ $\approx -\nabla f(x)\left(-\nabla^2 f(x)\nabla f(x)\right)$

$-\frac{1}{2}\left(-(\nabla^2 f(x)^{-1}\nabla f(x))^T\right)$

$\nabla^2 f(x)\left(-\nabla^2 f(x)^{-1}\nabla f(x)\right)$

$$f(x) - p^* \approx \frac{\lambda(x)^2}{2}$$

When our current iteration x is very close to x^*, $\frac{\lambda(x)^2}{2}$ gives us an estimate of how far we are from the best minimum.

Class #11

Friday, February 28, 2025 9:34 AM

Newton Decrement: $\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \right)^{1/2}$

When we are close to a minimizer p^*

$\frac{\lambda(x)^2}{2} \Rightarrow$ Estimate of how close we are to p^*

$f(x) - p^* \geq \frac{\lambda(x)^2}{2}$

Some facts: ① $\nabla f(x)^T \Delta x_{nt} = -\lambda(x)^2$

$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$

$\|\nabla f(x)\|_2 \|\Delta x_{nt}\|_2 \cos(\theta)$

② $\lambda(x) = \left(\Delta x_{nt} \nabla^2 f(x) \Delta x_{nt} \right)^{1/2}$

P - Quadratic norm of a vector when $P \in S^n_{++}$

$\|x\|_P = (x^T P x)^{1/2}$

$\lambda(x) = \|\Delta x_{nt}\|_{\nabla^2 f(x)}$

③ $\lambda(x)$ is invariant under affine transformations of the function $f(x)$

Newton's Step:

$$x^{(k+1)} = x^{(k)} + t^{(k)} \underbrace{\Delta x_{nt}}_{-\underbrace{\nabla^2 f(x)^{-1}}_{\nabla^2 f(x) \succ 0} \nabla f(x)}$$

when $t^{(k)} = 1 \Rightarrow$ Full Newton Step.

↳ Full Newton step should be taken only when x is very close to the local (or global) minimum. Otherwise, the method can diverge.

Newton's Method (Assuming we are close to a local minimum or are working with a convex problem)

Input: $x^{(0)} \in \text{dom } f$; tolerance $\epsilon > 0$

Initialize: $k \leftarrow 0$

Repeat:

① $\Delta x_{nt}^{(k)} \leftarrow -\underbrace{\nabla^2 f(x^{(k)})^{-1}}_{\text{↳ make sure } \nabla^2 f(x^{(k)}) \succ 0} \nabla f(x^{(k)})$ // Compute Newton's direction

② Compute $\lambda^2(x^{(k)}) \leftarrow \nabla f(x^{(k)})^T \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$

// Compute Newton decrement Square

③ If $\frac{\lambda^2(x^{(k)})}{2} \leq \epsilon$

break

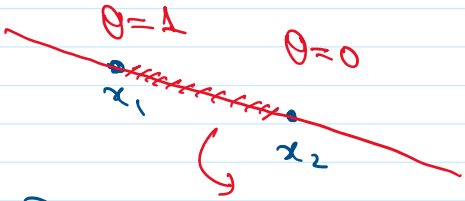
④ Line search: choose $t^{(k)}$ using back tracking line search

⑤ Update: $x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x_{nt}^{(k)}$

⑥ $k \leftarrow k+1$

Chapter 2: Convex Sets

① Line: Given $x_1 \neq x_2 \in \mathbb{R}^n$



$y = \theta x_1 + (1-\theta) x_2$; $\theta \in \mathbb{R}$

$\downarrow$

$y(\theta)$

↳ one-dimensional linear function

$y(\theta)$ is a line passing through x_1 and x_2

② Line segment: when $\theta \in [0,1]$,

$y(\theta) = \theta x_1 + (1-\theta) x_2$; is the line segment joining x_1 and x_2

$$= x_2 + \theta(x_1 - x_2)$$

$\theta \Rightarrow$ Tells us how far we are from x_2 in the direction $(x_1 - x_2)$

(3) **Affine Set**: A set $C \subseteq \mathbb{R}^n$ is affine if the line through any distinct points in C lies in the set C.

$$\forall\, x_1, x_2 \in C \text{ and } \theta \in \mathbb{R}$$

$$\theta x_1 + (1-\theta) x_2 \in C$$

↳ Linear combinations of points in C lie in C provided the coefficients of the linear combination sum to 1.

↓ simple induction argument

If C is an affine set, $x_1, \ldots, x_k \in C$ $\theta_1, \ldots, \theta_k \in \mathbb{R}$ s.t. $\sum_{j=1}^{k} \theta_j = 1$ then

$$\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k \in C$$

Affine Combination of Points

Given $x_1, \ldots, x_k \in \mathbb{R}^n$ and $\theta_1, \ldots, \theta_k \in \mathbb{R}$ s.t. $\sum_{j=1}^{k} \theta_j = 1$

$\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$ is called an affine combination.

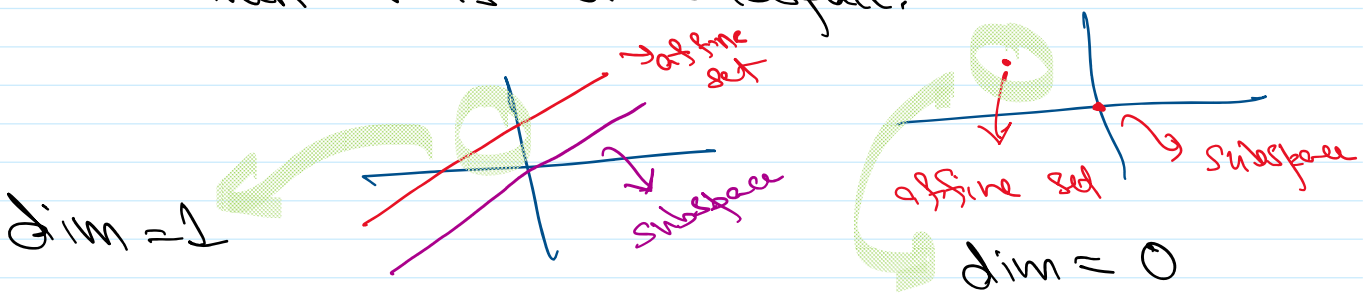
Connection to a Subspace

An affine set is a subspace shifted from the origin.

If C is an affine set and $x_0 \in C$ for any x_0 then

$$V = C - x_0 = \{x - x_0 : x \in C\}$$

↳ Then V is a subspace.



Dimension of an Affine Set

It is the dimension of the subspace associated with C.

Ex: $C = \{x : Ax = b ; b \in \mathbb{R}^n\}$

Is C affine?

Approach 1: Use the definition of an affine set.

Let x_1 and $x_2 \in C$

Show that $\theta x_1 + (1-\theta) x_2 \in C$

$x_1 \in C \Rightarrow Ax_1 = b$

$x_2 \in C \Rightarrow Ax_2 = b$

$\theta x_1 + (1-\theta) x_2 \Rightarrow \in C$?

Show $\Rightarrow A(\theta x_1 + (1-\theta) x_2) = b$

$= \theta \underbrace{Ax_1}_{b} + (1-\theta) \underbrace{Ax_2}_{b}$

$= \theta b + (1-\theta) b = b$ ✓

Approach 2:- $\underbrace{\{x : Ax = 0\}}_{\text{null space of } A}$ is a subspace

$\{x : Ax = b\}$ is simply a shifted version of $\{x : Ax = 0\} \Rightarrow C$ is an affine set.

Dimensionality of C = dimensionality of the null space of A.

Affine Hull of points in a set in $\mathbb{R}^n$

Given a set C, which may not be affine, the affine hull of C is defined as:

Given a set C, which may not be affine, the affine hull of C is defined as:

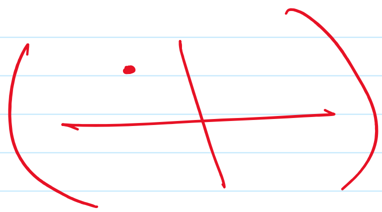
$$\text{aff } C = \{ \theta_1 x_1 + \dots + \theta_k x_k : x_1, \dots, x_k \in C, \; \theta_1 + \theta_2 + \dots \theta_k = 1 \}$$

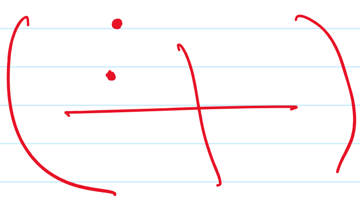
↪ Affine hull of a set C is the smallest affine set that contains C. That is, let S be any affine set such that:

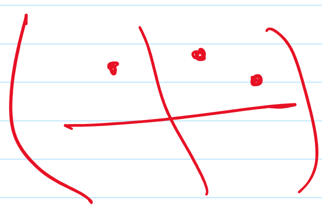
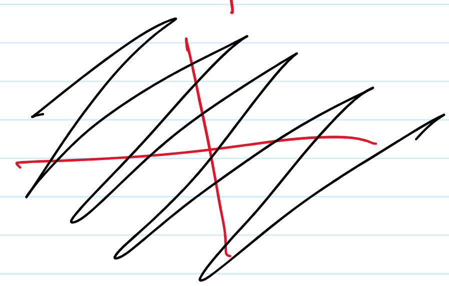
$C \subseteq S$. Then,

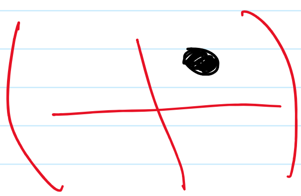
$\text{aff } C \subseteq S$

Examples:

aff  = 

aff  = 

aff  = 

aff  = $\mathbb{R}^2$

Reading: 2.1.1 and 2.1.2 (BV)

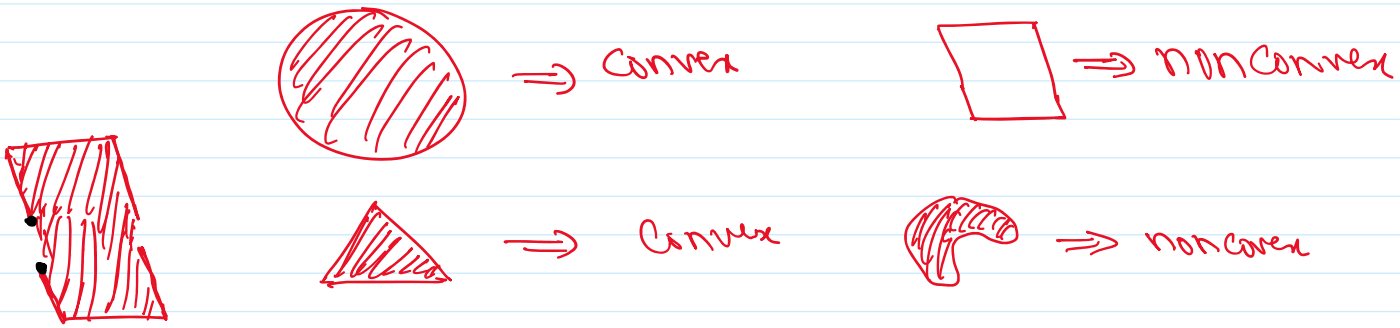
Convex Set

A set C is convex if the line segment between any two points in C, lies in C.

$$\forall\ x_1, x_2 \in C \text{ and } \theta \in [0,1]$$

$$\theta x_1 + (1-\theta) x_2 \in C$$

- Every affine set is a convex set



$\sum_{i=1}^{k} \theta_i = 1$

$\sum_{i=1}^{k} \theta_i = 1$, $\theta_i \geq 0$

linear combination

Subspaces $\subset$ affine sets $\subset$ Convex sets

Convex Combination of Points

Given $x_1, x_2, \ldots, x_k \in C$ and $\theta_1, \ldots, \theta_k \geq 0$

s.t. $\sum_{i=1}^{k} \theta_i = 1$

$\theta_1 x_1 + \cdots + \theta_k x_k$ is called a convex combination of the points in C.

Class #12

Friday, February 28, 2025 9:35 AM

A convex set implies that any convex combination of points in the set lie in the same set.

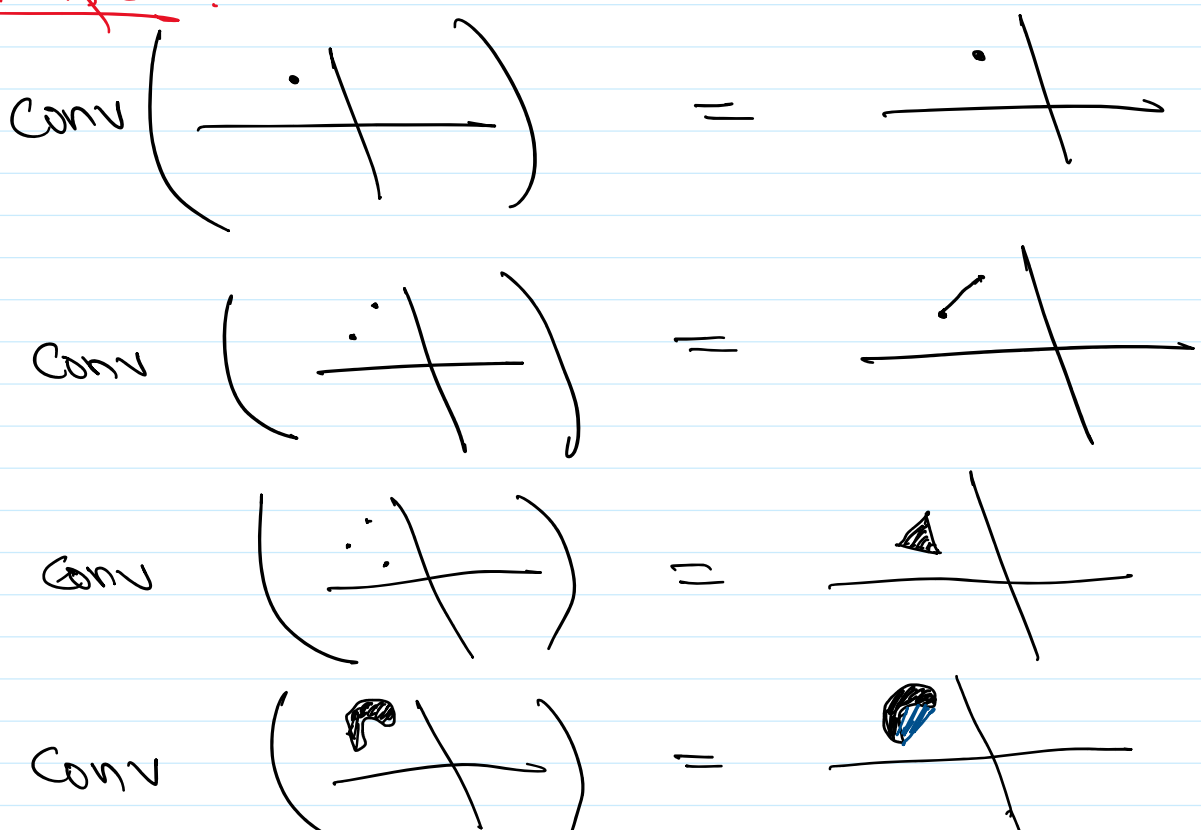
Convex hull

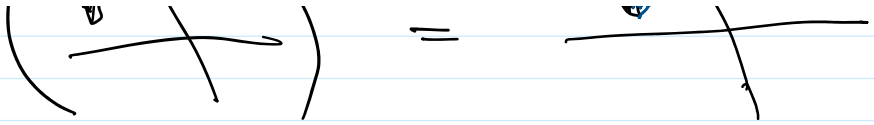
The convex hull of a set C is the set of all convex combinations of points in C

$$\text{conv } C = \left\{ \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k : x_i \in C;\ \theta_i \geq 0;\ \sum_{i=1}^{k} \theta_i = 1 \right\}$$

Convex hull of C is convex $\Leftrightarrow$ It is the smallest convex set that contains C.

Examples:



Conv  =

Convex combination $\Rightarrow$

$\theta_i \geq 0, \; i = 1, \ldots, k$
$\sum_{i=1}^{k} \theta_i = 1$
} Probability mass function

$p_x(x) \geq 0$
$\int p_x(x)\, dx = 1$
} Probability density functions

$$E[X] = \sum_{i=1}^{k} \underbrace{p_x(x_i)}_{\geq 0} x_i$$

Reading: 2.1.4 (BV)

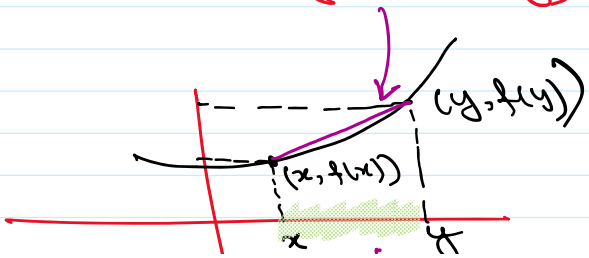
Convex functions

Let $f: \mathbb{R}^n \to \mathbb{R}$ with $\text{dom} f$ being convex. Then f is termed a convex function if $\forall\, x, y \in \text{dom} f$ and $\theta \in [0,1]$

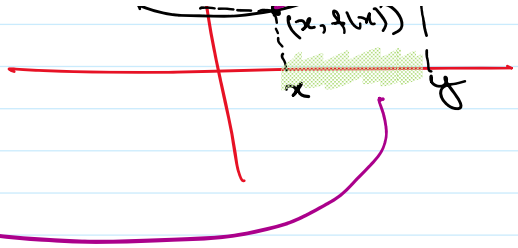
$$f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y)$$

Jensen's Inequality

$\theta x + (1-\theta) y \in \text{dom} f$



inequality

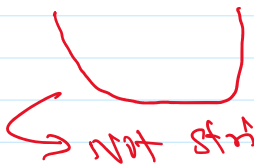


The chord (line segment) connecting $(x, f(x))$ and $(y, f(y))$ should lie above the function between x and y.

Strictly Convex function

If $f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$ $\forall\, x, y \in \text{dom } f$ then f is called a strictly convex function.

Ex: A linear function is convex but it is not strictly convex.



↳ not strictly convex but it is convex.

Concave functions: If $-f$ is convex then f is called concave (similarly strictly concave).

What if $\text{dom } f \neq \mathbb{R}^n$?

Extensions of convex functions on all $\mathbb{R}^n$

Let $f: \mathbb{R}^n \to \mathbb{R}$ be convex on dom f.

The convex extension of f on $\mathbb{R}^n$ is defined

The convex extension of f on $\mathbb{R}^n$ is defined as:

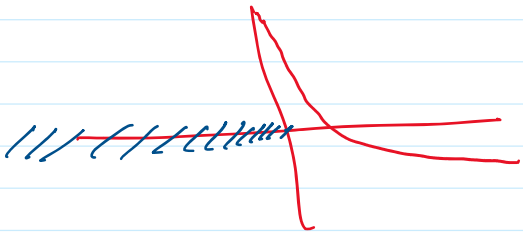
$$\tilde{f}(x) = \begin{cases} f(x), & x \in \text{dom} f \\ \infty, & x \notin \text{dom} f \end{cases}$$

↳ using this trick, we can ignore that $\text{dom} f \neq \mathbb{R}^n$

$$\text{dom} f = \{x : \tilde{f}(x) < \infty\}$$

Ex: $f(x) = -\log x$

$\text{dom} f = (0, \infty)$



Constrained Optimization trick

Let C be a convex set

Suppose we need to solve

$$\min_{x \in C} f(x) \; ; \; f : \mathbb{R}^n \to \mathbb{R}$$

$$\tilde{I}_C(x) = \begin{cases} 0, & x \in C \\ \infty, & x \notin C \end{cases} \Rightarrow \text{Convex function}$$

$$\min_{x \in \mathbb{R}^n} \left[f(x) + \tilde{I}_C(x) \right]$$

Reading: 3.1.1 and 3.1.2 (BV)

Convex function, convex combination, and Probability

Convex function, convex combination, and Probability

If $f: \mathbb{R}^n \to \mathbb{R}$ is convex

$$\Rightarrow f(\theta_1 x_1 + \cdots + \theta_k x_k) \leq \theta_1 f(x_1) + \cdots + \theta_k f(x_k)$$

$\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$

$f(\theta_1 x_1 + \cdots + \theta_k x_k) \hookrightarrow E[X]$

$\theta_1 f(x_1) + \cdots + \theta_k f(x_k) \hookrightarrow E[f(X)]$

Let x_i's be the values that at random variable X takes and θ_i are the probabilities

$$P(X = x_i) = \theta_i$$

If f is convex $\Rightarrow f(E[X]) \leq E[f(X)]$

↳ Jensen's Inequality

↳ This holds even when X is continuous random variable.

Reading: 3.1.8 (BV)

Equivalent Characterizations of Convex Functions

① A function $f: \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its domain is convex and its restriction to any line in $\mathbb{R}^n$ is convex:

Define: $g(t): \mathbb{R} \to \mathbb{R}$

$$g(t) = f(x + tv) \quad \forall\, x + tv \in \text{dom } f$$

Then f is convex $\iff$ $g(t)$ is convex for every x and v.

② First-order condition of convexity

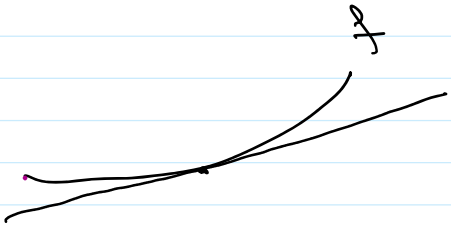
Let $f: \mathbb{R}^n \to \mathbb{R}$ be differentiable on $\operatorname{dom} f$ and $\operatorname{dom} f$ is open.

$f: \mathbb{R}^n \to \mathbb{R}$ is convex if and only if $\operatorname{dom} f$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) \quad \text{—— (*)}$$

$\forall\, x, y \in \operatorname{dom} f$

first-order approximation (linear approximation) of f around



↳ The first-order approximation must be a uniform underestimator of f.

Global optimality condition for convex functions

Let $x_0 \in \operatorname{dom} f$ be such that $\nabla f(x_0) = 0$

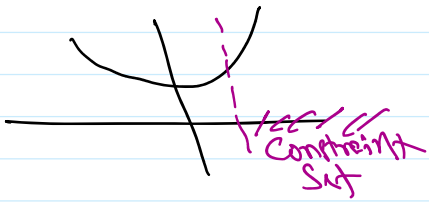
Then x_0 is a global minimizer of f.

i.e. $f(x_0) \leq f(y) \quad \forall\, y \in \operatorname{dom} f$

Proof: Take $x = x_0$ in (*) (first-order convexity condition)

$f(y) \geq f(x_0) + 0$

$\Rightarrow f(x_0) \leq f(y) \quad \forall\, y \in \text{dom} f$ ▨



Unconstrained optimization

x_0 is a global minimizer of $f \in C^1$ if and only if

$$\nabla f(x_0) = 0$$

Strictly Convex functions

$$f(y) > f(x) + \nabla f(x)^T (y - x)$$

→ A strictly convex function can only have a unique minimizer.

Indeed: Let x_1 and x_2 be two global minimizers with $\nabla f(x_1) = 0 = \nabla f(x_2)$

$f(x_1) < f(y) \quad \forall\, y \in \text{dom} f$

$f(x_1) < f(x_2)$

$\Rightarrow$ contradiction ▨

③ Monotonicity of gradients

Suppose $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable with dom$f$ being convex. Then f is convex if and only if

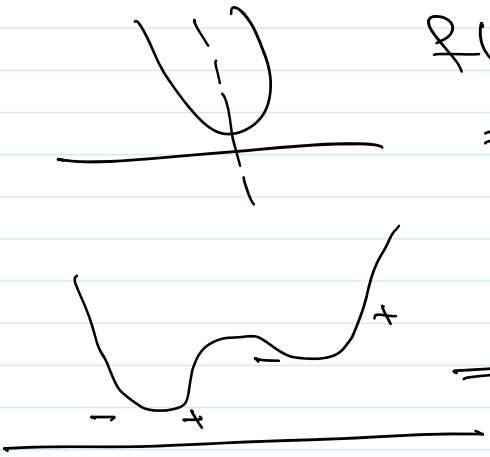
convex. Then f is convex if and only if

$$(\nabla f(x) - \nabla f(y))^T (y - x) \geq 0 \quad \forall \; x, y \in \text{dom} f$$

$> 0 \Rightarrow$ Strict Convexity

This generalizes the concept of monotonicity of functions to $\mathbb{R}^n$

E.g.:



$f(x) = x^2$

$\Rightarrow f'(x) = 2x$

$\Rightarrow$ Increasing $\Rightarrow$ Convex

$\Rightarrow$ not Convex

(4) Second-order Condition of Convexity

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable with dom$f$ being open.

f is convex if and only if dom$f$ is convex and

$$\underbrace{\nabla^2 f(x) \succeq 0}_{\text{Positive semi definite.}} \quad \forall \; x \in \text{dom} f$$

Basically, the function at every point x has non-negative curvature.

Concave : $\nabla^2 f(x) \preceq 0$

Strictly Convex functions $\not\Rightarrow \nabla^2 f(x) \succ 0$

Ex. $f(x) = x^4 \Rightarrow$ Strictly Convex.
but $f''(x) = 0$ at $x = 0$

Class #13

Tuesday, March 4, 2025 10:22 AM

Quadratic functions

$f: \mathbb{R}^n \to \mathbb{R}$ with $\text{dom } f = \mathbb{R}^n$

$$f(x) = \frac{1}{2} x^T P x + q^T x + r$$

$P \in \mathbb{R}^{n \times n}$

↳ Quadratic functions

$$\nabla^2 f(x) = P$$

A quadratic function is convex if and only if

$P \succeq 0 \quad (P \in S_+^n)$

$f(x)$ is concave when $P \preceq 0$ (negative semi definite)

when P is neither $\succeq 0$ or $\preceq 0$ $\Rightarrow$ It is neither concave nor convex

1D: $f(x) = ax^2$; $f(x) = x^2$ ↳ convex

$f(x) = -x^2 \Rightarrow$ concave

2D: $f(x) = x_1^2 - x_2^2$

Unconstrained Minimization of Convex functions

$$\min_x f(x) \quad \text{s.t. } f: \mathbb{R}^n \to \mathbb{R}$$

If f is convex $\Leftrightarrow$ $x^* : \nabla f(x^*) = 0$

Necessary and sufficient condition for solution of a convex optimization problem corresponds to solving the equation

$$\nabla f(x) = 0$$

e.g., $f(x) = (x-4)^2 \Rightarrow f'(x) = 2(x-4) = 0$

$$\Rightarrow x^* = 4$$

Why then we need optimization algorithms?

① Finding a solution to $\nabla f(x) = 0$ is not always possible analytically.

e.g., $\min_x \left[f(x) = \log\left(\sum_{i=1}^{m} \exp(a_i^T x + b_i) \right) \right]$

↳ Geometric program

$$\nabla f(x) = \frac{1}{\sum_{i=1}^{m} \exp(a_i^T x + b_i)} \sum_{i=1}^{m} \exp(a_i^T x + b_i)\, a_i$$

② Even when an analytical solution exists, computing it numerically in a direct fashion might be too costly.

e.g., $\min_x \left[f(x) = \|Ax - b\|_2^2 \right]$

$$\nabla f(x) = 2A^T(Ax - b)$$

$\Rightarrow \nabla f(x) = 0 \iff \underline{A^T A x = A^T b}$

Let's even assume $(A^T A)$ is full rank.

$$x^* = (A^T A)^{-1} A^T b$$

Convex Optimization Problems with and without Solution

① $f(x) = a^T x$

$\min_x f(x) \Rightarrow$ no solution $\Rightarrow$ function is unbounded below

② $f(x) = e^{-x}$

$\min_x f(x) \Rightarrow$ Does not exist

↳ function is bounded below but min exist

$\inf_x \{ f(x) = e^{-x} \} = 0$

Supremum (Sup) and infimum (inf) ⇒ Generalizations of max and min of sets

Let $C \subseteq \mathbb{R}$.

We say that a scalar value 'u' is an upperbound on the set C if:

$$\forall x \in C : x \leq u$$

① If the set of upper bounds is empty ⇒ C is unbounded from above

$C = \{\mathbb{N}\} \Rightarrow C$ is unbounded from above.

② If $C = \phi \Rightarrow$ set of upper bounds is $\mathbb{R}$

The set of upper bounds is always of the form $[b, \infty)$

e.g; $C = \{1, 2, 3\} \Rightarrow$ Set of upperbounds $= [3, \infty)$

Sup(C) is defined as the least upper bound in the set of upperbounds.

① $\text{Sup}\{1,2,3\} = 3 = \max$

Sup = max when Sup belongs to the set C

② $\text{Sup}\left\{1 - \frac{1}{n}\right\}_{n=1}^{\infty} = 1$

max $\Rightarrow$ does not exist

★ $\text{Sup}(\phi) = -\infty$

what if C is unbounded?

★ $\text{Sup}(C) = \infty$

Infimum of a set $\Rightarrow$ Deals with the case of lower bounds on sets

The set of lower bounds takes the form

$(-\infty, l]$

Inf(C) $\Rightarrow$ The greatest lower bound on the set

C.

e.g.; $C = \left\{1 + \frac{1}{n}\right\}_{n=1}^{\infty}$

$\inf(C) = 1 \neq \min(C)$

↳ does not exist

$\inf\{1,2,3\} = 1 = \min(C)$

* $\inf(\phi) = \infty$
* $\inf(C) = -\infty \Rightarrow$ when C is unbounded from below

$\inf(C) = -\sup(-C)$

Optimization problems should technically be initially written as inf problems, till we are convinced that inf = min

$\inf_x a^T x = -\infty \rightarrow a \neq 0$

$p^* = \inf_x e^{-x} = 0 \neq \min_x e^{-x}$

↳ This is not solvable

↳ $\arg\inf_x e^{-x} = \infty$ or no solution

↳ There is no x^* such that $f(x^*) = p^*$

Example of a Quadratic

$$f(x) = \frac{1}{2} x^T P x + q^T x + r \quad ; \quad P \succeq 0$$

$\Rightarrow f(x)$ is Convex

$$\nabla f(x) = Px + q$$

If $\min_x f(x)$ has a Solution

$\iff \nabla f(x) = 0$

$Px + q = 0 \iff Px = -q$

The Solution Set is $\{x : Px = -q\}$

What if this is not the Case?

$-q$ is in the column space of P

$\Rightarrow f(x)$ is unbounded from below and there is no Solution.

If P has full rank $\Rightarrow$ The Solution Set is non-empty; in fact there is a unique Solution.

$$x^* = -P^{-1} q$$

$P \succ 0 \Rightarrow f(x)$ is Strictly Convex

Strongly Convex Functions

SC functions $\subset$ Strictly Convex $\subset$ Convex functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ with dom$f$ being Convex.

f is Strongly Convex with parameter $m > 0$ if:

$\Leftrightarrow$ ① $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) - \frac{m}{2}\theta(1-\theta)\|x-y\|_2^2$

A function is strongly convex if and only if grows at least as fast as a quadratic function with growth parameter 'm'.

From ①, we see that

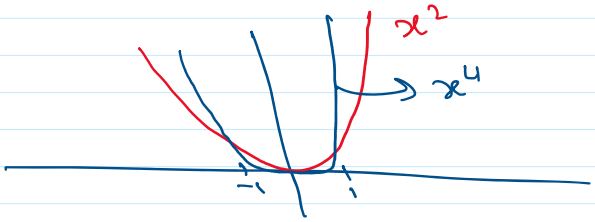
$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$

$\Rightarrow$ SC functions are Strictly Convex also

$\Rightarrow$ SC functions have a unique minimizer.

Eg: $f(x) = x^2 \rightarrow$ Strongly Convex

$f(x) = x^4 \rightarrow$ Strictly Convex $\Rightarrow$ But not Strongly Convex



$\Leftrightarrow$ ② Second-order definition of Strong Convexity

Suppose $f \in C^2$ (twice continuously differentiable) and $\text{dom} f$ is open

$f(x)$ is SC $\Leftrightarrow$ $\nabla^2 f(x) \succeq mI \quad \forall x \in \text{dom} f$

$\nabla^2 f(x) \succeq mI \Leftrightarrow \nabla^2 f(x) - mI \succeq 0$

$\Leftrightarrow \lambda_{min}(\nabla^2 f(x)) \geq m > 0$

e.g.; $f(x) = x^4 \Rightarrow f'(x) = 4x^3$

$f''(x) = 12x^2$

$f''(x) > 0 \quad \forall x \neq 0$ but

$f''(x) = 0$ for $x = 0$.

↳ Even for $x \neq 0$ $f''(x) \not\geq m$ for some $m > 0$.