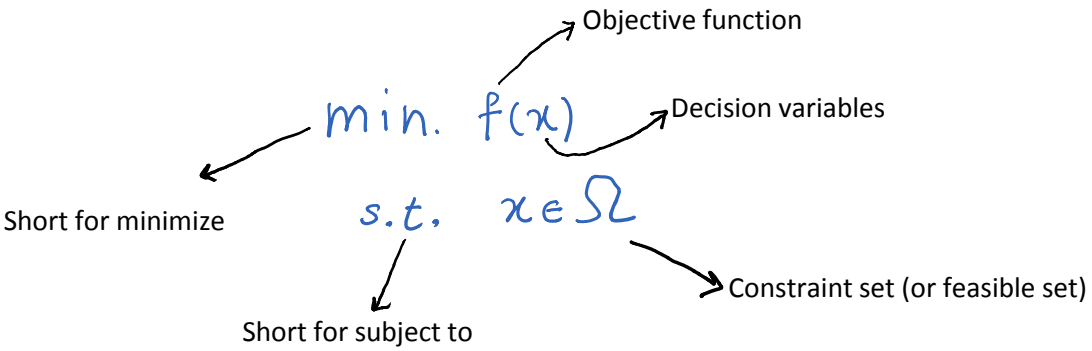
Lec3p1, ORF363/COS323

PRINCETON UNIVERSITY
ORFE
Instructor:
Amir Ali Ahmadi

This lecture:

- Optimization problems - basic notation and terminology
- Unconstrained optimization
  - The Fermat-Weber problem
  - Least squares
- First and second order necessary conditions for optimality
- Second order sufficient condition for optimality
- Solution to least squares

- An optimization problem in general (or abstract) form:



In this class (unless otherwise stated), we have:

$$x \in \mathbb{R}^n, \quad f: \mathbb{R}^n \to \mathbb{R}, \quad \Omega \subseteq \mathbb{R}^n.$$

Typically, some description of f and Ω is given as input to us.

Optimal solution x^*: "argmin $f(x)$ s.t. $x \in \Omega$"

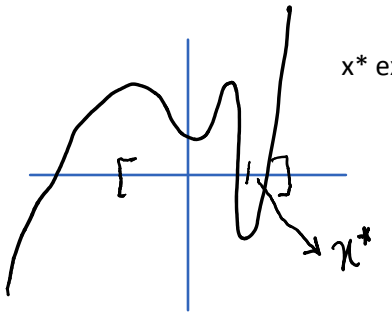
(also called the "solution" or the "global solution")

A point that minimizes f over Ω:

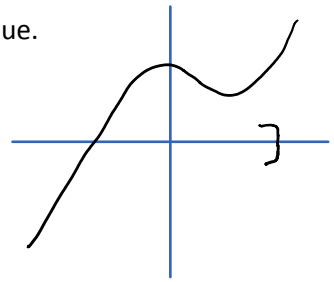
$$f(x) \geq f(x^*), \quad \forall x \in \Omega.$$

May not exist.
May not be unique.

Lec3p2, ORF363/COS323



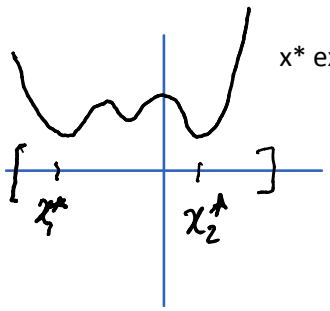
x* exists and is unique.



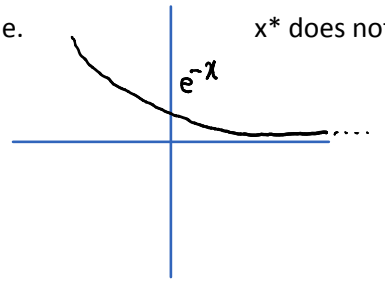
x* does not exist.

$f^* = -\infty$

Problem is "unbounded."



x* exists, but not unique.



x* does not exist.

$f^* = 0$

Optimal value: $f^* := f(x^*)$ (if x* exists)

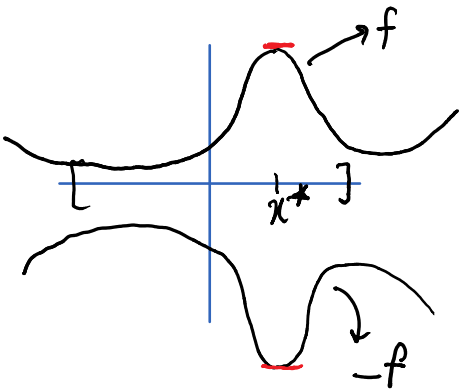
But can be well-defined even if x* doesn't exist.

- See the lower right picture above.
- In such a scenario, the term "minimum" is often replaced by "infimum".

- An important case where x* is guaranteed to exist:
  - f continuous and Ω compact (i.e., closed and bounded).
- This is known as the Weierstrass theorem.

- What if we want to maximize an objective function instead?
  - Just multiply f by a minus sign:

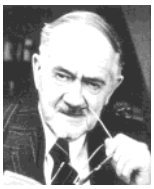
$$\max_{\text{s.t.}} f(x) = -\min_{\text{s.t.}} -f(x)$$
$$x \in \Omega \qquad x \in \Omega$$



Optimal solution doesn't change.
Optimal value only changes sign.

Unconstrained optimization: $\Omega = \mathbb{R}^n$

Decision variables are not constrained at all. The goal is only to minimize the objective function.

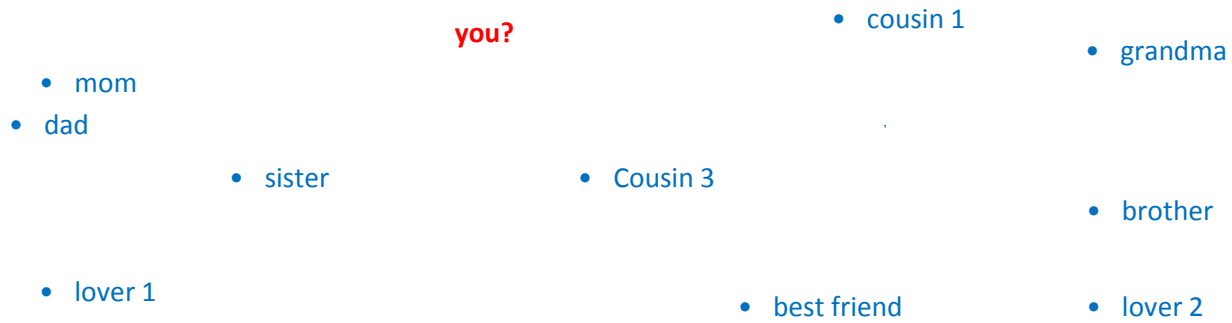


Fermat (1607-1665)

Weber (1868-1958)

Example 1: The Fermat-Weber problem.

You have a list of loved ones who live in given locations in the US. You would like to decide where to live so you are as close to them all as possible; say, you want to minimize the sum of distances to each person.



Location of person i: $x_i \in \mathbb{R}^2$ (given) $i = 1, \dots, N$

Your location: $y \in \mathbb{R}^2$ (decision variable)

$$\min_y \sum_{i=1}^N \|y - x_i\|_2$$

(Recall: $u \in \mathbb{R}^n$, $\|u\| = \sqrt{u^T u} = \sqrt{u_1^2 + \dots + u_n^2}$)

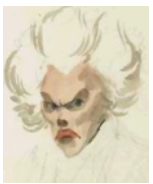
- Variant: also given weights w_i for each person
  (your mom says you should care more about her than lover 1)

$$\min_y \sum_{i=1}^N w_i \|y - x_i\|_2 \qquad (w_i \in \mathbb{R})$$

- Many other applications: e.g., Princeton is deciding on the location of a new gym and wants to minimize distance to dormitories, giving priority to undergrads,…

- As we'll see later, this optimization problem is "easy" to solve, not because it is unconstrained (as there are many terribly hard unconstrained problems!), but because it has a nice structure (called convexity).
- If at the same time you wanted to be "far" from some subset of your friends and family, this would have been a very hard problem to solve!
- Optimization theory is full of instances where a tiny variation in the problem formulation changes the problem completely from being very easy to solve to being very hard to solve. It takes a trained eye to detect this. By the end of the course, you will learn techniques that will help you make such distinctions.
- But we are getting way ahead of ourselves. For one thing, we haven't even formalized what it means for an optimization problem to be "easy" or "hard". Let's forget this for now and move on to another unconstrained optimization problem---one of the most widely-encountered in science and engineering.

Example 2: Least squares.



Legendre
(1752-1833)

Gauss
(1777-1855)

Given: A mxn matrix

b mx1 vector

Solve: $\min_x \; ||Ax-b||^2$

By default, ||.|| always represents the 2-norm; i.e., $||.||_2$.

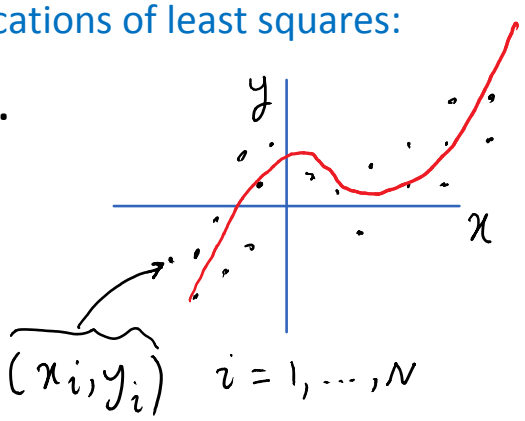
In expanded notation, we are solving:

$$\min_x \sum_{i=1}^{m} (a_i^T x - b_i)^2$$

$$A = \begin{pmatrix} — & a_1^T & — \\ & \vdots & \\ — & a_m^T & — \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

Some applications of least squares:

Data fitting.



Given: (x_i, y_i) $i = 1, \dots, N$

Fit a (say, cubic) polynomial: $p(x) = c_3 x^3 + c_2 x^2 + c_1 x + c_0$

decision variables

$$\min_{c_0,\dots,c_3} \sum_{i=1}^{N} (p(x_i) - y_i)^2$$

Quick notation exercise: convince yourself that this is a least squares problem.

Overdetermined system of linear equations.

A simple linear predictor for the stock price of a company:

$S(t)$: Stock price at day t

$$S(t) = a_1 S(t-1) + a_2 S(t-2) + a_3 S(t-3) + a_4 S(t-4)$$

We have three months of daily stock price data to train our model (lots of 5-day windows). How to find the best $a_1, \dots, a_4$ for future prediction?

$$\underbrace{\begin{pmatrix} S(4) & S(3) & S(2) & S(1) \\ S(5) & S(4) & S(3) & S(2) \\ \vdots & & & \\ S(N-1) & S(N-2) & S(N-3) & S(N-4) \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} S(5) \\ S(6) \\ \vdots \\ S(N) \end{pmatrix}}_{b}$$

$$\min_x \|Ax - b\|^2$$

A and b are given from data. (N would be 90.)

Optimality conditions

Unconstrained local and global minima.

Consider a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

A point x^* is said to be a:

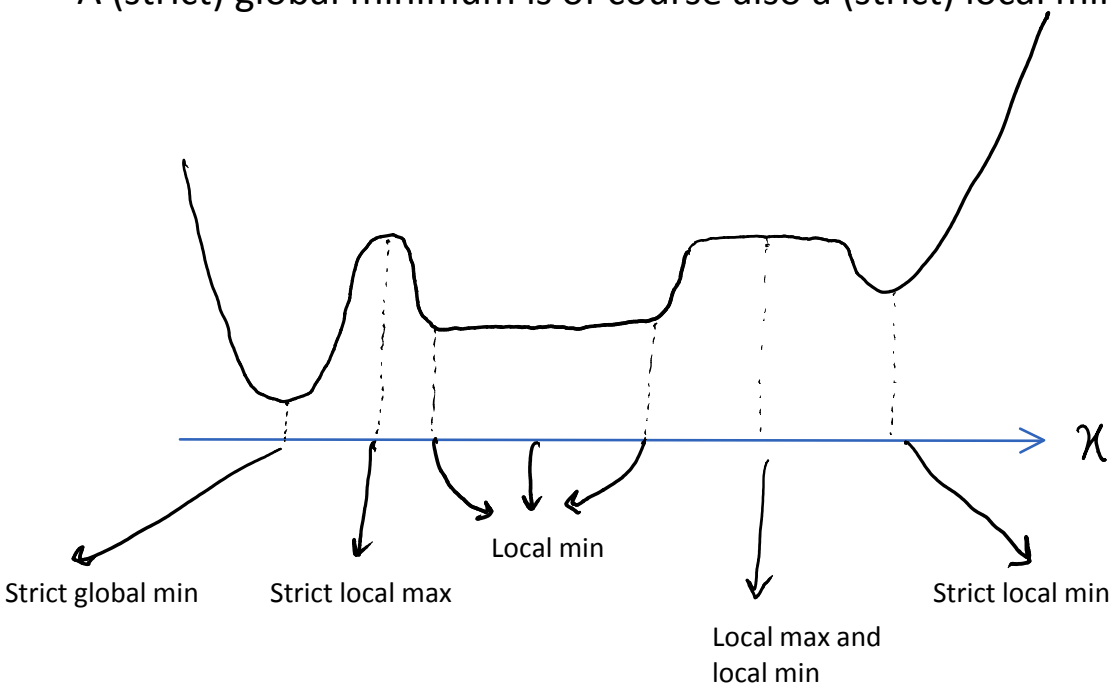
Local minimum: if $\exists \delta > 0$, s.t. $f(x^*) \leq f(x)$ $\forall x$ with $\|x - x^*\| < \delta$.

Strict local minimum: if $\exists \delta > 0$, s.t. $f(x^*) < f(x)$ $\forall x \neq x^*$ with $\|x - x^*\| < \delta$.

Global minimum: if $f(x^*) \leq f(x)$ $\forall x \in \mathbb{R}^n$.

Strict global minimum: if $f(x^*) < f(x)$ $\forall x \in \mathbb{R}^n$, $x \neq x^*$.

- Local/global maxima defined analogously.
- A (strict) global minimum is of course also a (strict) local minimum.



No global max in this case. Problem is unbounded above.

- In general, finding local minima is a less ambitious goal than finding global minima.
- Luckily, there are important problems where we can find global minima efficiently.
- On the other hand, there are problems where finding even a local minimum is intractable.
- These statements should become more concrete as the course progresses.

First and second order conditions for local optimality

Optimality conditions are results that give us some structural information about the properties of optimal solutions. To understand the proofs that follow, make sure you are comfortable with the following notions:

- The gradient vector
- The chain rule
- The Hessian matrix
- Taylor series approximation

See lecture notes of the previous lecture or Sections 5.3-5.6 of [CZ13].

Notation reminder: $f(x) := f(x_1, \dots, x_n) \quad (f: \mathbb{R}^n \to \mathbb{R})$

$$\nabla f(x) := \begin{pmatrix} \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_n \end{pmatrix}$$

The gradient vector (nx1 vector)

$$\nabla^2 f(x) := \begin{pmatrix} \partial^2 f / \partial x_1 \partial x_1 & \partial^2 f / \partial x_1 \partial x_2 & \cdots & \partial^2 f / \partial x_1 \partial x_n \\ \partial^2 f / \partial x_1 \partial x_2 & \partial^2 f / \partial x_2 \partial x_2 & \cdots & \partial^2 f / \partial x_2 \partial x_n \\ \vdots & & & \\ \partial^2 f / \partial x_n \partial x_1 & \cdots & & \partial^2 f / \partial x_n \partial x_n \end{pmatrix}$$

The Hessian matrix (nxn symmetric matrix)
Notation of [CZ13]: $D^2 f$

Theorem. **(First Order Necessary Condition for (Local) Optimality)**
If x^* is an unconstrained local minimizer of a differentiable function $f: \mathbb{R}^n \to \mathbb{R}$, then we must have:

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


Fermat
(1607-1665)

Proof. Consider some $y \in \mathbb{R}^n$. Define $g: \mathbb{R} \to \mathbb{R}$ as

$$g(\alpha) = f(x^* + \alpha y).$$

By the chain rule, we have

$$\frac{dg}{d\alpha}(\alpha) = y^T \nabla f(x^* + \alpha y).$$

So

$$\frac{dg}{d\alpha}(0) = y^T \nabla f(x^*). \quad (1)$$

But we also have, by definition,

$$\frac{dg}{d\alpha}(0) = \lim_{\alpha \downarrow 0} \frac{f(x^* + \alpha y) - f(x^*)}{\alpha} \geq 0, \quad (2)$$

↳ why?

Well, because by local optimality of x^*, $\exists \delta > 0$, s.t.

$$f(x^* + \alpha y) \geq f(x^*), \quad \forall\, 0 < \alpha \leq \delta.$$

Now, $(1) + (2) \Rightarrow y^T \nabla f(x^*) \geq 0.$ (3)

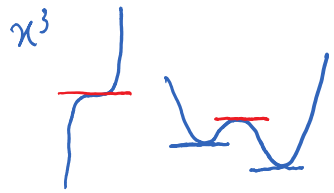
Since y was arbitrary, replace y by $-y$ and repeat the same argument to get

$$-y^T \nabla f(x^*) \geq 0. \quad (4)$$

$$(3) + (4) \Rightarrow y^T \nabla f(x^*) = 0, \quad \forall\, y \in \mathbb{R}^n.$$

$$\Rightarrow \nabla f(x^*) = 0.$$

□



Remarks:

- This condition is necessary but *not* sufficient for local optimality.
- Nevertheless, it is useful because any local minimum must satisfy this condition. So, we can look for local (or global) minima only among points that make the gradient of the objective function vanish.
- We will see later that in presence of an important concept called *convexity*, this condition is in fact sufficient for local (and global!) optimality.
- Terminology: A point x that satisfies $\nabla f(x) = 0$ is called a *stationary point* or a *critical point* of f.

Second order conditions.

The statements of our second order optimality conditions involve the notions of psd and pd matrices. Let's recap these concepts.

Linear algebra interlude.

(See the last lecture if you need more review.)

Symmetric matrix: $A = A^T$ (A^T denotes the transpose of A.)

e.g., $\begin{bmatrix} -5 & 2 \\ 2 & 3 \end{bmatrix}$ symmetric $\begin{bmatrix} -5 & 2 \\ 1 & 3 \end{bmatrix}$ not symmetric

Theorem. Eigenvalues of a real symmetric matrix are real.

Proof. See, e.g., Theorem 3.2 in Section 3.2 of [CZ13].

A square matrix *A* is said to be:

- Positive semidefinite (psd) if: $x^T A x \geq 0, \ \forall x \in \mathbb{R}^n$
- Positive definite (pd) if: $x^T A x > 0, \ \forall x \in \mathbb{R}^n, x \neq 0$

Notation: A psd: $A \succeq 0$ A pd: $A \succ 0$.

Recall that when we talk of positive semidefiniteness (or positive definiteness), we assume with no loss of generality that our matrix is symmetric: If A was not symmetric, we could take its "symmetric part".

$$x^T A x = x^T \underbrace{\left(\frac{A^T + A}{2}\right)}_{\text{symmetric part of } A} x \quad \text{(verify this equation)}$$

Theorem. A matrix is positive semidefinite if and only if all its eigenvalues are nonnegative. A matrix is positive definite if and only if all its eigenvalues are positive.

Proof. See, e.g., Theorem 3.7 in Section 3.4 of [CZ13].

Examples: $A = \begin{bmatrix} 2 & 4 \\ 4 & 5 \end{bmatrix}$ MATLAB: eig([2 4;4 5])

$\lambda_1 \approx -0.77, \quad \lambda_2 = 7.77 \implies$ not psd

$A = \begin{bmatrix} 2 & 1 \\ 1 & 8 \end{bmatrix} \rightarrow \lambda_1 \approx 1.8, \ \lambda_2 \approx 8.2 \implies$ pd (and psd)

$A = \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \rightarrow \lambda_1 = 0, \ \lambda_2 = 5 \implies$ psd, but not pd

Recall our easy test in dimension 2:

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad \text{psd} \Leftrightarrow \begin{cases} a \geq 0, c \geq 0 \\ ac - b^2 \geq 0 \end{cases}, \quad \text{pd} \Leftrightarrow \begin{cases} a > 0 \\ ac - b^2 > 0 \end{cases}$$

This generalizes to *n* dimensions using the concepts of principal minors and leading principal minors; see Section 3.4 of [CZ13].

Theorem. **(Second Order Necessary Condition for (Local) Optimality)** If x^* is an unconstrained local minimizer of a twice continuously differentiable function $f: \mathbb{R}^n \to \mathbb{R}$, then, in addition to $\nabla f(x^*) = 0$, we must have:

$$\nabla^2 f(x^*) \succeq 0.$$

(i.e., the Hessian at x is positive semidefinite.)

Proof. Consider some $y \in \mathbb{R}^n$. For $\alpha > 0$, the second order Taylor expansion of f around x^* gives

$$f(x^* + \alpha y) = f(x^*) + \alpha y^T \nabla f(x^*) + \frac{\alpha^2}{2} y^T \nabla^2 f(x^*) y + o(\alpha^2).$$

Since $\nabla f(x^*)$ must be zero (as previously proven), we have

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

By definition of local optimality of x^*, the left hand side is ≥ 0, for α sufficiently small.

$$\Rightarrow \lim_{\alpha \downarrow 0} \frac{1}{2} y^T \nabla^2 f(x^*) y + \frac{o(\alpha^2)}{\alpha^2} \geq 0.$$

But $\lim_{\alpha \downarrow 0} \frac{o(\alpha^2)}{\alpha^2} = 0$

$$\Rightarrow y^T \nabla^2 f(x) y \geq 0$$

(and y was arbitrary). □

"Little o" notation: see [CZ13], Section 5.6 or our previous lecture.

Theorem. **(Second Order Sufficient Condition for (Local) Optimality)**

Suppose $f: \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, $\nabla f(x^*) = 0$, and

$$\nabla^2 f(x^*) \succ 0,$$

(i.e., the Hessian at x^* is positive definite), then x^* is a strict local minimum of f.

Proof. Let λ be the minimum eigenvalue of $\nabla^2 f(x^*)$.

$$\Rightarrow \quad \nabla^2 f(x^*) - \lambda I \succeq 0 \qquad (\text{why?})$$

$$\Rightarrow \quad y^T \nabla^2 f(x^*) y \geq \lambda \|y\|^2, \quad \forall y \in \mathbb{R}^n.$$

Once again, Taylor expansion yields

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

$$\geq \frac{1}{2} \lambda \|y\|^2 + o(\|y\|^2)$$

$$= \|y\|^2 \left(\frac{\lambda}{2} + \frac{o(\|y\|^2)}{\|y\|^2} \right).$$

Since $\lim_{\|y\| \to 0} \frac{o(\|y\|^2)}{\|y\|^2} = 0$,

$$\exists \delta > 0, \text{ s.t. } \frac{o(\|y\|^2)}{\|y\|^2} < \frac{\lambda}{2}, \quad \forall y \text{ with } \|y\| \leq \delta, y \neq 0.$$

Hence,

$$f(x^*+y) > f(x^*), \quad \forall y \text{ with } \|y\| \leq \delta.$$

But this by definition means that x^* is a strict local minimum. □

Remarks.

- $\nabla f(x^*)=0$, $\nabla^2 f(x^*) \not\succ 0$ is not sufficient for local optimality.

$f(x) = x^3$  $f'(0) = f''(0) = 0.$

- $\nabla^2 f(x) \not\succ 0$ is not necessary for (even strict global) optimality.

$f(x) = x^4$  $f''(0)=0$, but $x=0$ strict minimum.

Questions to keep in the back of your mind:

- How would we use all these optimality conditions to find local solutions and certify their optimality?
- Is it easy to find points satisfying these conditions? e.g., is it easy to solve $\nabla f(x)=0$?
- Suppose you certified that a given point is locally optimal, how would you go about checking if it is also globally optimal?

Exercise. State (and prove) the analogues of our three theorems for local maxima.

Now that we have a better understanding of the structure of optimal solutions for unconstrained optimization problems, let's revisit our least squares problem…

Least squares, revisited.

Given: A $m \times n$ matrix (Assume columns of A are linearly independent)

b $m \times 1$ vector

Solve: $\min_x \|Ax-b\|^2$

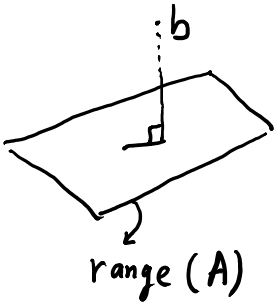
Let $f(x) = \|Ax-b\|^2 = (Ax-b)^T(Ax-b)$

$= x^TA^TAx - 2x^TA^Tb + b^Tb.$

$\nabla f(x) = 2A^TAx - 2A^Tb$

$\nabla f(x) = 0 \Rightarrow A^TAx = A^Tb$ ← Called "Normal Equations"

$\Rightarrow \boxed{x = (A^TA)^{-1}A^Tb}$



A^TA is invertible b/c its null space is just the origin:

$A^TAx = 0 \Rightarrow x^TA^TAx = 0 \Rightarrow (Ax)^T(Ax) = 0 \Rightarrow \|Ax\|^2 = 0 \Rightarrow Ax = 0$

$\Rightarrow x = 0.$

↑ Columns of A linearly independent.

$\nabla^2 f(x) = 2A^TA \succ 0$ (b/c $x^TA^TAx = \|Ax\|^2 \geq 0$ and $= 0 \Leftrightarrow x = 0$)

$\Rightarrow x = (A^TA)^{-1}A^Tb$ is a strict local minimum. □

(We will see later that it's in fact a strict global minimum b/c $\|Ax-b\|^2$ is a "convex" function.)

Exercise with optimality conditions

Find all the local minima and maxima of the following function:

$$f(x) = \frac{1}{2} x_1^2 + x_1 x_2 + 2x_2^2 - 4x_1 - 4x_2 - x_2^3.$$

(Example taken from notes by Rob Freund of MIT.)

$$\nabla f(x) = \begin{pmatrix} x_1 + x_2 - 4 \\ x_1 + 4x_2 - 4 - 3x_2^2 \end{pmatrix}$$

$\nabla f(x) = 0 \implies \bar{x} = \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \quad \tilde{x} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ ← Candidate critical points

$$\nabla^2 f(x) = \begin{pmatrix} 1 & 1 \\ 1 & 4 - 6x_2 \end{pmatrix}.$$

$$\nabla^2 f(\tilde{x}) = \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix} \rightarrow \text{indefinite}$$

b/c $\begin{pmatrix} 1 \\ 0 \end{pmatrix}^T \nabla^2 f(\tilde{x}) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 > 0$

$\begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \nabla^2 f(\tilde{x}) \begin{pmatrix} 0 \\ 1 \end{pmatrix} = -2 < 0.$

$$\nabla^2 f(\bar{x}) = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \succ 0$$ Sylvester's criterion: $1 > 0, \quad 4 - 1^2 = 3 > 0.$

$\implies \bar{x} = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$ is a strict local min & it's the only local min.

Notes:

- Optimality conditions are covered in Chapter 6 of [CZ13] in a more general setting where one also has a general constraint $x \in \Omega$. The unconstrained optimality conditions that we presented here are stated in Chapter 6 as corollaries (called the "interior case"). You are only responsible for what was covered in class.
- Least squares is covered in Section 12.1 of [CZ13]. But again, this is for further reading and my notes should have everything that I expect you to know.

References:

- [CZ13] E.K.P. Chong and S.H. Zak. An Introduction to Optimization. Fourth edition. Wiley, 2013.

- [Bert04] D.P. Bertsekas. Nonlinear Programming. Second edition. Athena Scientific, 2004.