Optimization Methods for Data Science

Instructor: Rayan Saab

Course webpage:

math.ucsd.edu/~rsaab

→ teaching → this course

Grading

- 2 Midterms
- HW assignments
- Final Exam

check webpage for dates

Why Opt. for Data Science?

Let's start with an informal defn of data science, as the science (or sometimes art) of extracting knowledge (whatever that is) from data.

Often, the goal is to make the best (i.e. optimal) decision

General Examples:

- optimize cost or revenue subject to constraints
- partition data (cluster it) in some optimal way

- decide how to classify (e.g. healthy vs sick) "optimally".

Specific Examples:

* Building (fantasy) sports teams given a cap.

* recommender systems, e.g. Netflix, Amazon, making "optimal" recommendations

* Image recognition

* Speech recognition

* Airline route planning

etc...

* All deep learning applications (Almost all machine learning has optimization at its core).

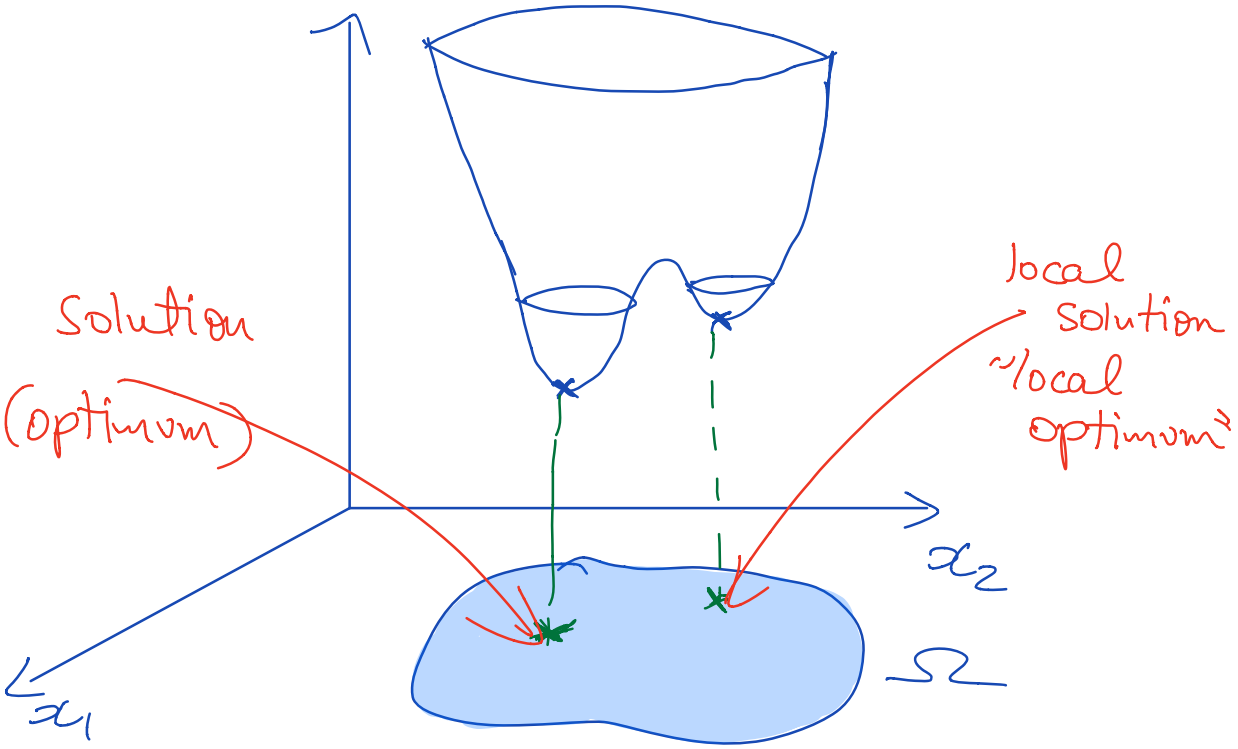
Introduction:

Our goal will often be to solve problems of the form

$$\min_{x \in \Omega} f(x) \quad (1)$$

objective function $f : \mathbb{R}^n \to \mathbb{R}$

constraint, often $\Omega \subset \mathbb{R}^n$



Constrained vs unconstrained optimization:

When Ω is the whole space i.e., $\Omega = \mathbb{R}^n$, we say the opt. problem is unconstrained

Example:

$$\min_{x_1, x_2} \; 2x_1^2 + x_2^4 - 2x_2^2 + 1$$

When Ω is a strict subset of $\mathbb{R}^n$ we say the optimization problem is constrained.

Example:

$$\min \; 2x_1^2 + x_2^4 - 2x_2^2 + 1$$
$$x_1 \in [1,2]$$
$$x_2 \in [3,4]$$

Instead of $\min_{x \in \Omega} f(x)$ we may write $\min_x f(x)$ subject to $x \in \Omega$

Defn: Optimal solution

* We say that x^* is a solution of (1) if
  - $x^* \in \Omega$ (feasible)
  - $\forall x \in \Omega : f(x^*) \leq f(x)$
* We say that x^* is a local solution of (1) if
  - $x^* \in \Omega$
  - There is a neighborhood N around x^* such that

$$\forall x \in N \cap \Omega : f(x^*) \leq f(x)$$

(Strict local min $\Rightarrow$ replace $\leq$ with $<$)

So now, one question is:

If someone hands you a function $f(x)$, and a point $\hat{x}$, when can you even check if $\hat{x}$ is optimal? (or even locally optimal)

To answer this and other questions (like how do we find optimal solutions), we need some definitions.

Convexity:

First, there are convex functions

then there are convex sets.

Convex functions:

We say $f: \Omega \to \mathbb{R}$ is convex if

$\forall x, y \in \Omega$:

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

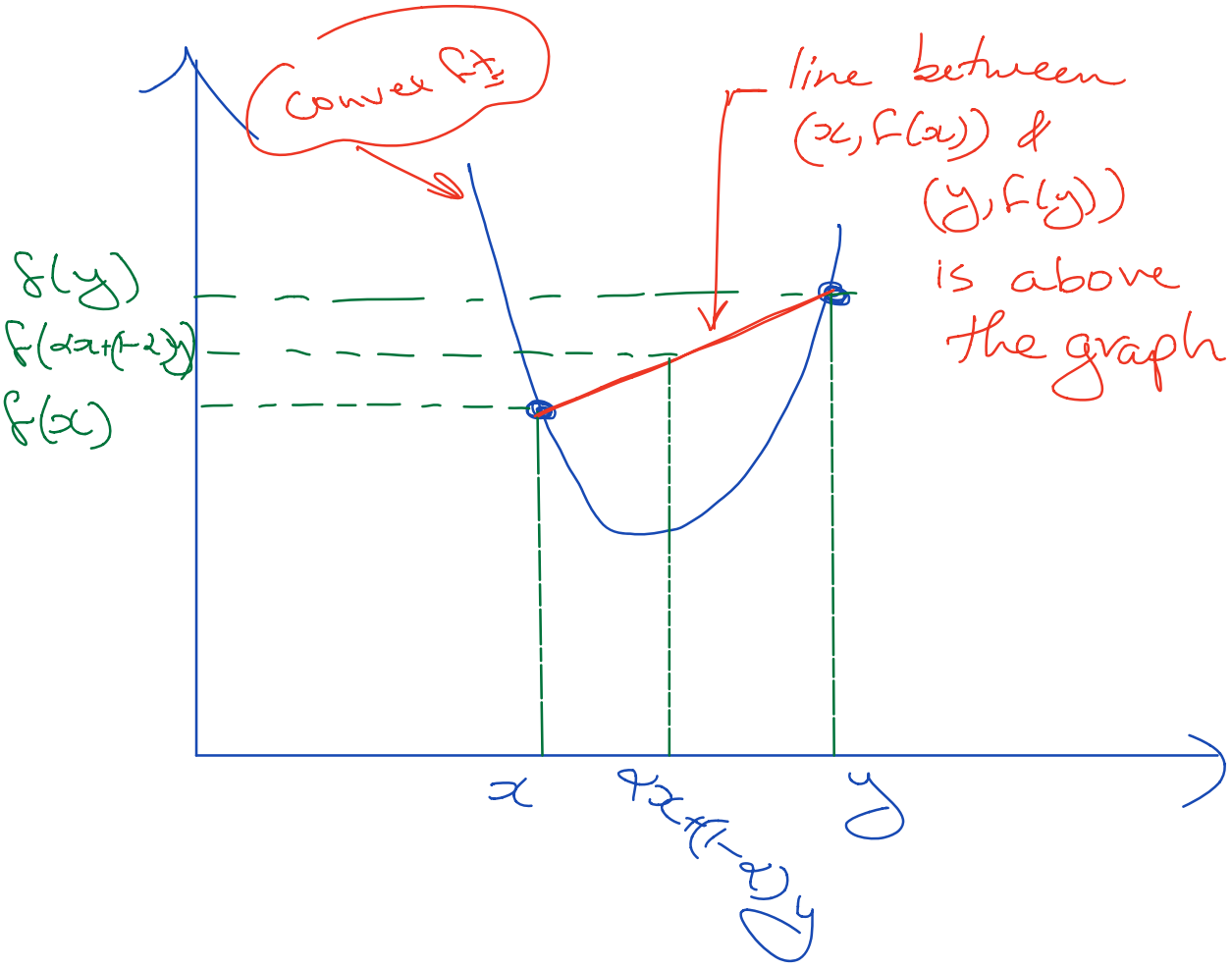
$$\alpha \in [0,1]$$

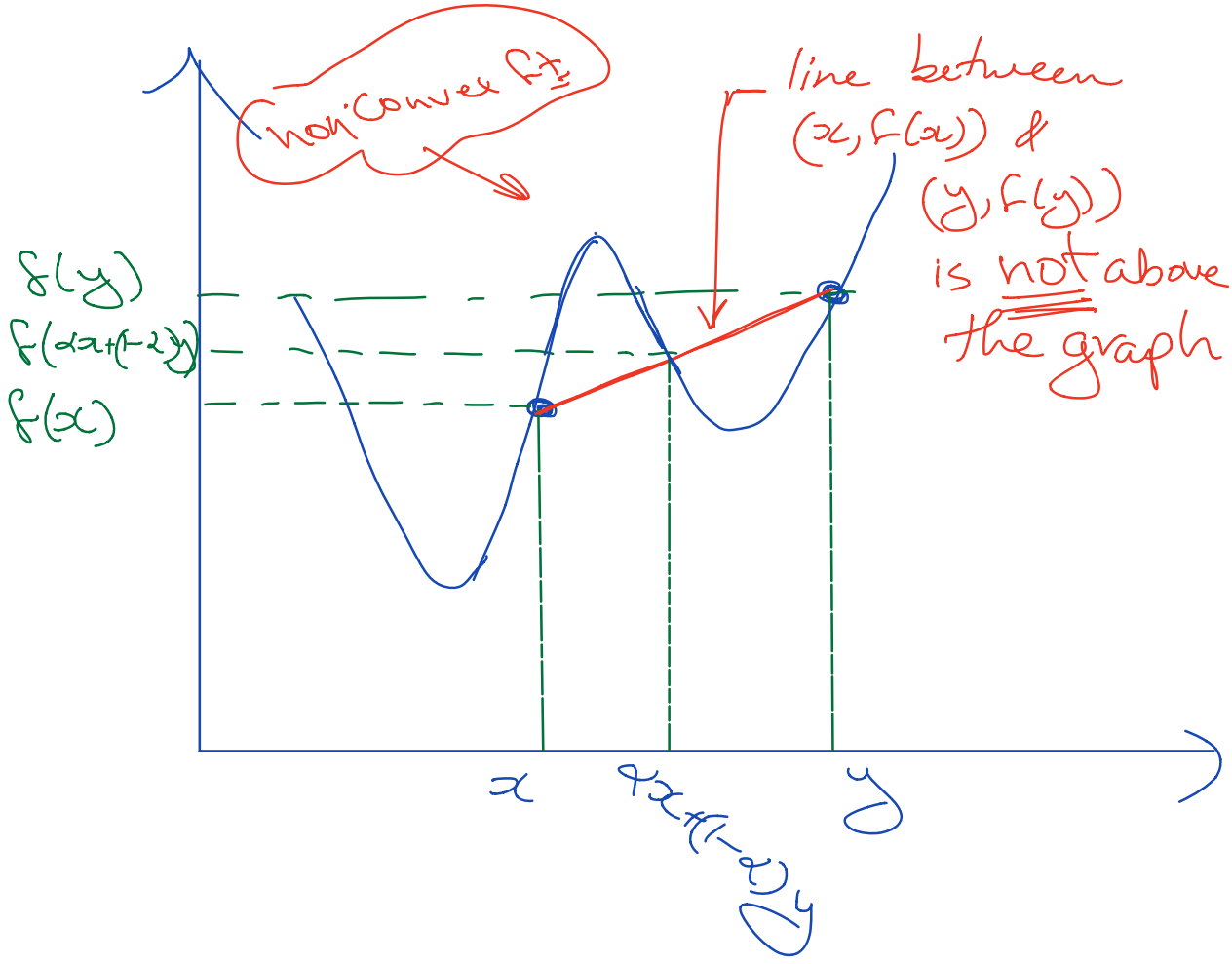
We say $f: \Omega \to \mathbb{R}$ is strictly convex

if $\forall x, y \in \Omega$:

$$f(\alpha x + (1-\alpha) y) < \alpha f(x) + (1-\alpha) f(y)$$

$$\alpha \in [0,1]$$


convex fct
line between (x, f(x)) & (y, f(y)) is above the graph
f(y)
f(αx+(1-α)y)
f(x)
x
αx+(1-α)y
y



Example: $f(x) = x^2$ is convex

because:

- $f(\alpha x + (1-\alpha) y) = (\alpha x + (1-\alpha) y)^2$

$$= \alpha^2 x^2 + (1-\alpha)^2 y^2 + 2\alpha(1-\alpha) xy$$

- $\alpha f(x) + (1-\alpha) f(y) = \alpha x^2 + (1-\alpha) y^2$

& so

$$\alpha f(x) + (1-\alpha) f(y) - f(\alpha x + (1-\alpha) y)$$

$$= (\alpha - \alpha^2) x^2 + [(1-\alpha) - (1-\alpha)^2] y^2 - 2\alpha(1-\alpha) xy$$

$$= \alpha(1-\alpha) x^2 + (1-\alpha)(1-(1-\alpha)) y^2 - 2\alpha(1-\alpha) xy$$

$$= \alpha(1-\alpha) [x^2 + y^2 - 2xy]$$

$$= \alpha(1-\alpha) (x-y)^2$$

$$\geq 0 \quad \forall \alpha \in [0,1], \; x, y \in \mathbb{R}$$

On the other hand :

Example : $f(x) = x^3$ is not convex on $\mathbb{R}$

because

$f(0.5) = 1/8$, $f(-1) = -1$

$f(\alpha\, 0.5 + (1-\alpha)(-1)) = f((1.5)\alpha - 1)$

$= ((1.5)\alpha - 1)^3$

while $\alpha f(0.5) = \alpha/8$

$(1-\alpha) f(-1) = \alpha - 1$

so $\alpha f(0.5) + (1-\alpha) f(-1) = \frac{9\alpha}{8} - 1$

pick $\alpha = 2/3$

$\Rightarrow f(\alpha\,(0.5) + (1-\alpha)(-1)) = 0$

$\geq \frac{9}{8} \times \frac{2}{3} - 1$

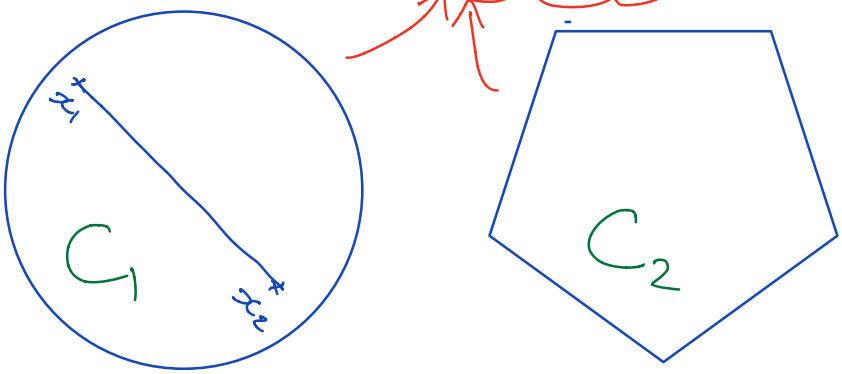
Convex Sets

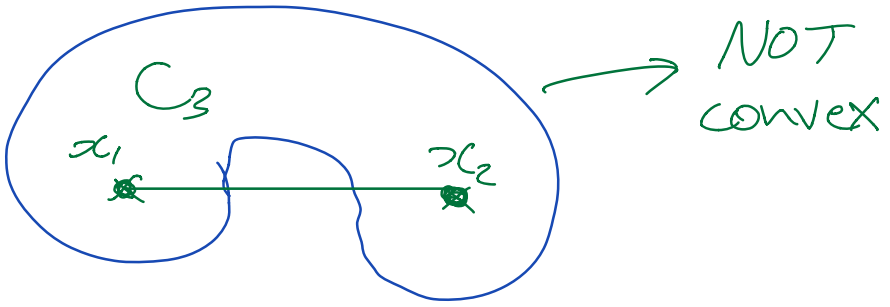
A set C is convex if
$\forall x_1, x_2 \in C$ and any $\alpha \in [0,1]$

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

Examples

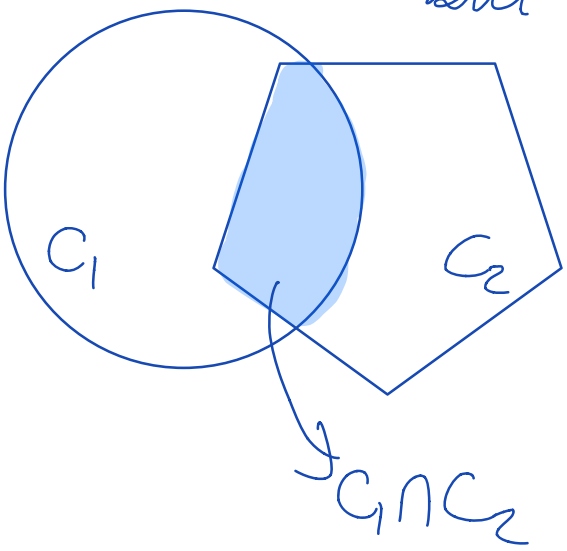
convex





Notice that $C_1 \cup C_2$ is not convex.

but $C_1 \cap C_2$ is convex.



Fact: If S, T are convex sets then $S \cap T$ is convex

Proof: HW

Fact: The intersection of any collection of (finite, countable, infinite) convex sets is convex

For convex functions defined on convex sets, local minima/maxima are global minima/maxima

Theorem: Consider the opt. problem

$$\underset{x \in \Omega}{\text{minimize}} \quad f(x)$$

where $f: \Omega \to \mathbb{R}$ is a convex fct & Ω is a convex set

then if x^* is a local min. it is also a global min.

Proof: Suppose x^* is a local min that is not global

$\Rightarrow \exists\, w \in \Omega$: $f(w) < f(x^*)$

Since f is convex: $\forall \alpha \in [0,1]$ we have

$$f(\alpha w + (1-\alpha)x^*) \leq \alpha f(w) + (1-\alpha) f(x^*)$$
$$< f(x^*) \quad \text{(why?)}$$

but this means every pt. z on the line segment from w to x^* has $f(z) < f(x^*)$

So x^* is not a local minimum, and we have a contradiction



Gradients, Hessians, Taylor's Theorem.

Recall that in single variable calculus, to find local minima/maxima of smooth functions, we needed to compute the derivative $f'(x)$ & solve $f'(x) = 0$.

With more variables (i.e. $x \in \mathbb{R}^n$) the situation is more complicated but it will still be necessary to consider gradients.

Defn:

For $f : \mathbb{R}^n \to \mathbb{R}$ the gradient

$$(\nabla f)(x) = \nabla f|_x = \nabla f(x)$$

is defined as

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_i} \right)_{i=1,\dots,n}$$

and the Hessian

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{\substack{i=1,\dots,n \\ j=1,\dots,n}}$$

Example:

for $f: \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x_1, x_2) = 2x_1^2 + x_2^4 - 2x_2^2 + 1$$

$$(\nabla f)(x_1, x_2) = \left(4x_1,\ 4x_2^3 - 4x_2\right)$$

&

$$\nabla^2 f = \begin{pmatrix} 4 & 0 \\ 0 & 12x_2^2 - 4 \end{pmatrix}$$

Recall: Directional derivatives

Remember that if, for example

$f: \mathbb{R}^3 \to \mathbb{R}$, then

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2, x_3) - f(x_1, x_2, x_3)}{h}$$

$$= \lim_{h \to 0} \frac{f(\vec{x} + h\vec{i}) - f(\vec{x})}{h}$$

So $\frac{\partial f}{\partial x_1}$ is the derivative in the direction of $\vec{i}$.

What if we want the deriv. in the direction of some unit vector $\vec{v}$?

$$\lim_{h \to 0} \frac{f(\vec{x} + h\vec{v}) - f(\vec{x})}{h}$$

↳ Directional derivative in the direction of the unit vector $\vec{v}$.

Directional derivatives & Gradients

Theorem: The directional deriv. of a differentiable function $f: \mathbb{R}^n \to \mathbb{R}$, in the direction of $\vec{v}$, $\|\vec{v}\| = 1$ is

$$\langle \nabla \vec{f}, \vec{v} \rangle = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} v_i$$

Example: $f(x_1, x_2, x_3) = x_1^2 e^{-x_2 x_3}$

Find the rate of change of f in the direction of $\vec{v} = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$ at $(1, 0, 0)$.

Soln: The required rate of change is $\langle \nabla f, v \rangle \big|_{(1,0,0)}$, so we need

$$\nabla f = \left(2x_1 e^{-x_2 x_3}, -x_1^2 x_3 e^{-x_2 x_3}, -x_1^2 x_2 e^{-x_2 x_3}\right)$$

$$\Rightarrow \langle \nabla f, v \rangle \big|_{(1,0,0)} = \left\langle (2, 0, 0), \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right) \right\rangle$$

$$= 2/\sqrt{3}.$$

Remark: if $\|\vec{v}\| \neq 1$, then you should normalize, i.e., use $\dfrac{\vec{v}}{\|\vec{v}\|}$.

The gradient is the direction of greatest increase

↳ If $\vec{n}$ is such that $\|n\| = 1$, then the rate of change in the direction of $\vec{n}$ is

angle between ∇f & n

$$\langle \nabla f, n \rangle = \|\nabla f\| \underbrace{\|n\|}_{1} \cos\theta$$

So if $\nabla f \neq 0$, $\langle \nabla f, n \rangle$ is maximal when $\cos\theta = 1$, i.e., $\theta = 0$

$\Leftrightarrow$ when n & ∇f are in the same direction.

Example: The direction of greatest increase of $f(x_1, x_2) = x_1^2 + x_2^3$ at $p = (2,3)$ is

$$\nabla f|_p = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right)\Big|_p = (2x_1, 3x_2^2)\Big|_{(2,3)}$$

$$= (4, 27)$$

Recall: Taylor's Theorem

We will also need Taylor's Theorem (multivariable version)

IF $f: \mathbb{R}^n \to \mathbb{R}$, f continuously differentiable then:

$$f(\underbrace{y}_{\in \mathbb{R}^n}) = f(\underbrace{x}_{\in \mathbb{R}^n}) + \nabla f(tx + (1-t)y)^T (y-x)$$

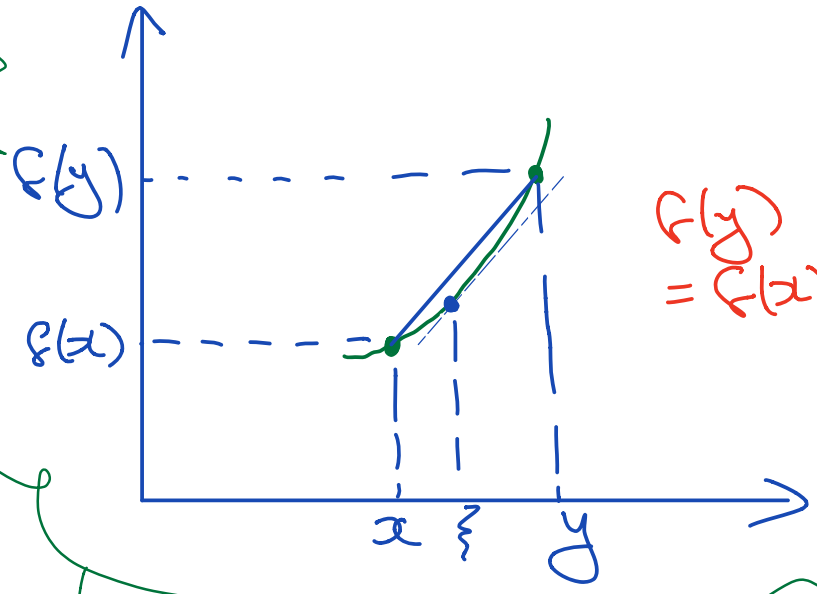
for some $t \in (0,1)$

↳ 0th order Taylor

Recall that in one variable we had

$$f(y) = f(x) + f'(\xi)(y-x)$$

$\hookrightarrow \xi \in (x, y)$



$$f(y) = f(x) + f'(\xi)(y-x)$$

In one variable

Taylor's Theorem (continued)

If $f: \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable

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

↳ 1st order Taylor approx

The one variable version:

$$f(y) = f(x) + f'(x)(y-x) + \frac{f''(\xi)(y-x)^2}{2}$$

$\xi \in (x, y)$

Equivalent conditions for convexity

Recall: We showed that convex

Functions over convex domains have the nice property that local minima are ~~global~~ minima.

How can we easily check if a f is convex?

Theorem: If f is continuously differentiable, then f is convex if & only if $\forall x, y \in \mathbb{R}^n$ we have

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

1st order Taylor approx.

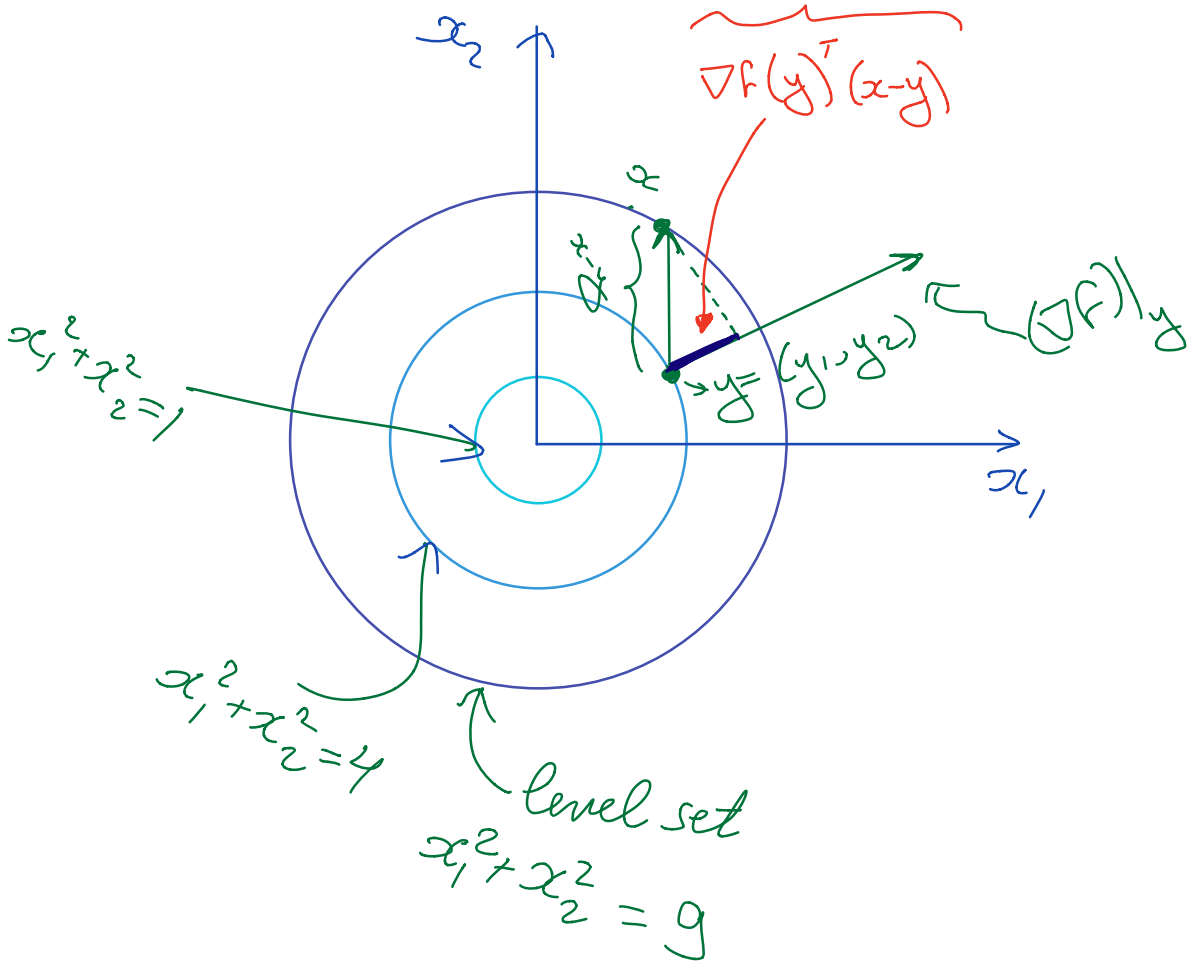
Important

Example/Illustration:

$$f(x_1, x_2) = x_1^2 + x_2^2$$

$$(\nabla f)\big|_{(x_1, x_2)} = (2x_1, 2x_2)$$

Adding $\nabla f(y)^T (x-y)$ to $f(y)$ does NOT get you to the level set of $f(x)$.



Theorem

If f is twice cont. diff. then f is convex if & only if $\forall x$:

$$\nabla^2 f(x) \text{ is pos. semidef.}$$

Important

Recall: $A \in \mathbb{R}^{n \times n}$:

- A is PSD $\iff$ $x^T A x \geq 0$ $\forall x \in \mathbb{R}^n$
- A is PSD $\iff$ eigenvalues of A are ≥ 0

Example:

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

Consider $x = (x_1, x_2, x_3)$

then $x^T A x =$

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= \begin{bmatrix} 2x_1 - x_2 & -x_1 + 2x_2 - x_3 & -x_2 + 2x_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= 2x_1^2 - x_1x_2 - x_1x_2 + 2x_2^2 - x_2x_3 - x_2x_3 + 2x_3^2$$

$$= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2$$

≥ 0 for any x_1, x_2, x_3

$\Rightarrow A$ is PSD

Exercise: Check the eigenvalues of A!

Example: Show that the function $f: \mathbb{R}^3 \to \mathbb{R}^3$ given by $f(a,b,c) = a^2 + b^2 + c^2$ is convex.

Solution: Let's calculate $\nabla^2 f$:

$$\nabla f(a,b,c) = (2a, 2b, 2c)$$

$$\nabla^2 f(a,b,c) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

which is a PSD matrix for all (a,b,c)

$\Rightarrow$ f is convex.

More on convexity:

(Convexity & derivatives)

Def'n:

A mapping $g: \mathbb{R}^n \to \mathbb{R}^n$ is called monotone if $\forall x, y \in \text{domain}(g)$

$$\langle g(x) - g(y), x - y \rangle \geq 0$$

Example: Let $g(x) = 2x$

Then $\forall x, y \in \mathbb{R}^n$

recall def'n of norm!

$$\langle g(x) - g(y), x - y \rangle$$

$$= \langle 2x - 2y, x - y \rangle = 2\|x - y\|^2 \geq 0$$

$\Rightarrow g$ is monotone!

Fact: If a function is differentiable then convexity & monotonicity of gradients are equivalent!

In theorem form:

Theorem: A differentiable fn $f: \mathbb{R}^n \to \mathbb{R}$ is convex if and only if ∇f is monotone.

Proof: (I) convex $\Rightarrow$ gradient monotone:

By previous theorem, f is convex

$$\Rightarrow \begin{cases} f(y) \geq f(x) + \nabla f(x)^T (y - x) \\ \& \\ f(x) \geq f(y) + \nabla f(y)^T (x - y) \end{cases}$$

$$\Rightarrow \begin{cases} f(y) - f(x) \geq \nabla f(x)^T (y-x) \\ f(x) - f(y) \geq \nabla f(y)^T (x-y) \end{cases}$$

adding

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

$$\Rightarrow \langle \nabla f(y) - \nabla f(x), y-x \rangle \geq 0$$

(II) gradient monotone $\Rightarrow$ convex:

The idea is that we want to show that monotonicity of ∇f implies

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

(which implies convexity)

To do this, we'll parametrize the segment between x & y and define $g(t) = f(x + t(y-x))$

chain rule

$$\Rightarrow g'(t) \overset{\downarrow}{=} \nabla f(x + t(y-x))^T (y-x)$$

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

But monotonicity of ∇f implies

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

$$= t\,[g'(t) - g'(0)]$$

$$\Rightarrow \forall\, t \in [0,1] : g'(t) \geq g'(0)$$

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

$$\Rightarrow \underbrace{g(0) + \int_0^1 g'(t)\,dt}_{g(1)} \geq g(0) + g'(0)$$

$$\Rightarrow \underset{\| \; f(y)}{g(1)} \geq \underset{\|}{g(0)} + \underset{\|}{g'(0)}$$

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

which is what we wanted.

∎