10701 Recitation 2: Convexity

February 2021

1 Introduction A

Optimization is at the core of many problems studied in machine learning. For example,
• maximum likelihood estimation: $\max_\theta \sum_{i=1}^n \log p_\theta(x_i)$
• linear regression: $\min_w ||Xw - y||^2$
In general:

$$\begin{aligned} &\text{minimize}_x \; f(x) \\ \text{s.t. } &g_i(x) \geq 0 \quad \forall i \in \{1, 2, \dots, k\} \\ &h_j(x) = 0 \quad \forall j \in \{1, 2, \dots, l\}. \end{aligned} \tag{1}$$

Today, we will look at a particular class of optimization problems - convex optimization.

An important terminology distinction to keep in mind throughout:
When we say **minimum**, we mean the minimum *function value.*
When we say **minimizer(s)**, we mean the input(s) that yield the minimum.

We'll use the phrase **neighborhood of a point** in certain places below. Think of this as a ball of some radius around the point. Example: in $\mathbb{R}$, $\mathcal{B}_r(x) = \{y : |y - x| < r\}$.

When we say **global minimum** of a function f, we're talking about the full domain $\mathcal{D}$. We mean a value $f(x)$ (for some $x \in \mathcal{D}$) s.t. $\forall y \in \mathcal{D}$, $f(x) \leq f(y)$.
When we say **local minimum** of a function f, we're just talking about what happens in a neighborhood of the local minimum. So we mean a value $f(x)$ (for some $x \in \mathcal{D}$) s.t. $\forall y$ in some neighborhood of x, $f(x) \leq f(y)$. That neighborhood could be arbitrarily small.

Here, we'll typically use x, y as input values, and $f(x), f(y), g(x), g(y)$ as function values. So when you see y here, think input, not output!

2 Convex Sets A

A convex set is a set $\mathcal{D}$ with the following property: $\forall x, y \in \mathcal{D}, \forall \alpha \in [0, 1]$

$$\alpha x + (1 - \alpha) y \in \mathcal{D}$$

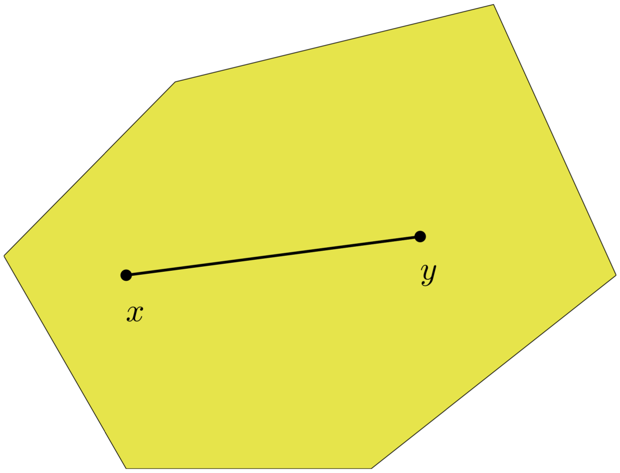


Figure 1: Source: [1]

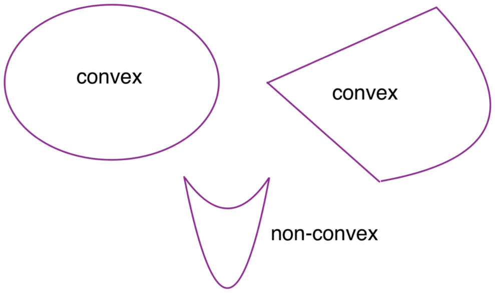


Figure 2: Source: 10-725 lecture slides by Prof. Yuanzhi Li

Example: Is a line a convex set?

Other trivial convex sets: the empty set, a single point.

Example: Let $\forall i, C_i$ is a convex set. $C = \cap_i C_i$. Is C a convex set?

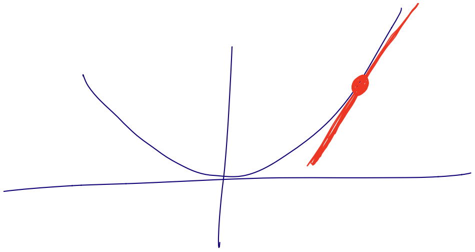
Solution: Consider $x, y \in C = \cap_i C_i \implies \forall i, x, y \in C_i$.
Since C_i is a convex set,

$$\begin{aligned}
&\implies \forall i, \forall \alpha \in [0,1], \alpha x + (1-\alpha)y \in C_i \\
&\implies \forall \alpha \in [0,1], \alpha x + (1-\alpha)y \in \cap_i C_i \\
&\implies \forall \alpha \in [0,1], \alpha x + (1-\alpha)y \in C
\end{aligned}$$

So, C is a convex set.

2.1 Why does it matter if a set is convex?

We'll show an example later!



3 Taylor Expansions/Motivation S

Analytic function. A function f is called *analytic* if, for all x_0 in f's domain, the Taylor series $g_{x_0}(x)$ centered at x_0 converges to the true function value $f(x)$ for all x in some neighborhood of x_0.

$$f(x) = g_{x_0}(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n.$$

expand in terms of derivatives

You might have seen the Taylor expansion of e^x around 0 before:

$$e^x = \underset{f(0)}{1} + \underset{f'(0)\cdot x}{x} + \underset{f''(0)/2\cdot x^2}{\frac{x^2}{2!}} + \boxed{\frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!}} + \dots = \boxed{\sum_{n=0}^{\infty} \frac{x^n}{n!}}.$$

Fun fact. A function can have infinitely-many derivatives (this is called "smooth") and actually still not be analytic! But these are pathological cases, for the most part...

That's all to say that the Taylor expansion of a function is **SUPER** useful. It appears everywhere!

Often, though, the higher-order terms decay (get very small) pretty quickly as $n \to \infty$, and so we can approximate the function pretty well in a neighborhood of x_0 just using the linear (1st-order) approximation or quadratic (2nd-order) approximation.

Therefore, in optimization, we often deal with extremely complicated functions by treating them as *linear*. Maybe this seems crazy! These functions aren't linear! But within a small-enough region of a particular point, things are "close enough" to linear. (This is how people mistakenly thought the Earth was flat for many years!)

This motivates us to frequently talk about how a function behaves compared to its linearization around a point... up next!

4 Convex Functions

4.1 Checking Convexity S

Definition (convex function). A function f over a convex set D is convex if it satisfies the following property:

$$\forall x, y \in D, \forall \alpha \in [0,1],\ f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

Concave $\geq$

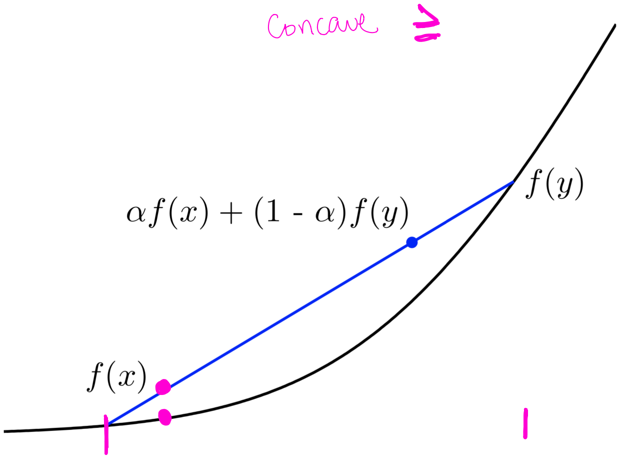


Figure 3: Function of average smaller than average of functions. Source: [1]

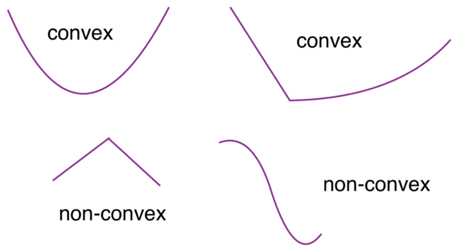
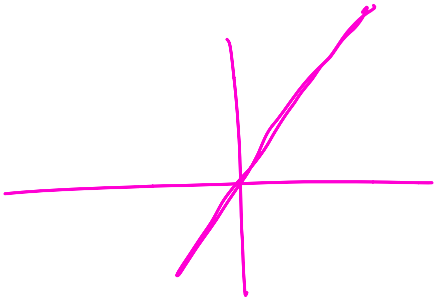


Figure 4: Convex and non-convex functions. Source: 10-725 lecture slides by Prof. Yuanzhi Li



Example: Is a line (let's say $f(x) = x$) a convex function?

Example: If f, g are convex functions, then show the convexity / non-convexity of the following:

- $f + g$

using the definition!

WTS (def. of convexity)

$\forall \alpha \in [0,1], \quad \forall x, y \in D,$

$$\underbrace{(f+g)(\alpha x + (1-\alpha)y)} \leq \underbrace{\alpha (f+g)(x) + (1-\alpha)(f+g)(y)}$$

$$\shortparallel$$

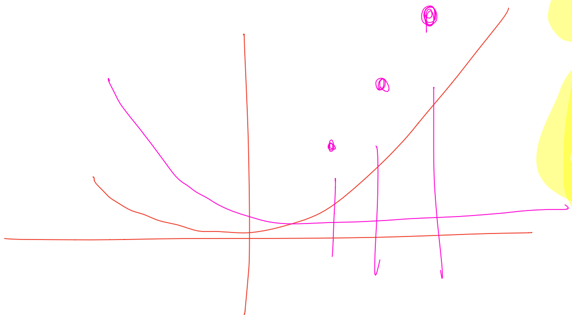
$$f(\alpha x + (1-\alpha)y) + g(\alpha x + (1-\alpha)y)$$

$$\shortparallel$$

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

$$=$$

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



Solution: $f + g$ is convex.

$$\begin{aligned}\forall\alpha \in [0,1], &\alpha \times (f(x) + g(x)) + (1-\alpha) \times (f(y) + g(y)) \\ &= (\alpha f(x) + (1-\alpha) f(y)) + (\alpha g(x) + (1-\alpha) g(y)) \\ &\geq f(\alpha x + (1-\alpha)y) + g(\alpha x + (1-\alpha)y) \text{ since } f, g \text{ are convex} \\ &\implies \alpha(f+g)(x) + (1-\alpha)(f+g)(y) \geq (f+g)(\alpha x + (1-\alpha)y)\end{aligned}$$

- $\max\{f, g\}$

Solution: $\max\{f, g\}$ is convex.

$$\forall \alpha \in [0,1], \alpha \max\{f(x), g(x)\} + (1-\alpha)\max\{f(y), g(y)\} = \begin{cases} \alpha f(x) + (1-\alpha) f(y) \\ \alpha g(x) + (1-\alpha) g(y) \\ \alpha f(x) + (1-\alpha) g(y) \\ \alpha g(x) + (1-\alpha) f(y) \end{cases}$$

Since f, g are convex, $\alpha f(x)+(1-\alpha)f(y) \geq f(\alpha x+(1-\alpha)y)$ and $\alpha g(x)+(1-\alpha)g(y) \geq g(\alpha x+(1-\alpha)y)$.

- Case I: $\alpha f(x)+(1-\alpha)f(y) \geq f(\alpha x+(1-\alpha)y)$. Also, $\alpha f(x)+(1-\alpha)f(y) \geq \alpha g(x)+(1-\alpha)g(y) \geq g(\alpha x + (1-\alpha)y)$.

$$\implies \alpha f(x) + (1-\alpha)f(y) \geq \max\{f(\alpha x + (1-\alpha)y), g(\alpha x + (1-\alpha)y)\}$$

- Case II: $\alpha g(x)+(1-\alpha)g(y) \geq g(\alpha x+(1-\alpha)y)$. Also, $\alpha g(x)+(1-\alpha)g(y) \geq \alpha f(x)+(1-\alpha)f(y) \geq f(\alpha x + (1-\alpha)y)$.

$$\implies \alpha g(x) + (1-\alpha)g(y) \geq \max\{f(\alpha x + (1-\alpha)y), g(\alpha x + (1-\alpha)y)\}$$

- Case III: $\alpha f(x)+(1-\alpha)g(y) \geq \alpha f(x)+(1-\alpha)f(y)$ since $g(y) \geq f(y)$. Also, $\alpha f(x)+(1-\alpha)g(y) \geq \alpha g(x) + (1-\alpha)g(y)$ since $f(x) \geq g(x)$. So, $\alpha f(x) + (1-\alpha)g(y) \geq \max\{f(\alpha x + (1-\alpha)y), g(\alpha x + (1-\alpha)y)\}$.
- Case IV: Similar to Case III.

Hence, $\forall \alpha \in [0,1], \alpha \max\{f(x), g(x)\}+(1-\alpha)\max\{f(y), g(y)\} \geq \max\{f(\alpha x+(1-\alpha)y), g(\alpha x+(1-\alpha)y)\}$

- $\min\{f, g\}$

Solution: $\min\{f, g\}$ is nonconvex. Counterexample: Draw the functions: $f(x) = x^2$ and $g(x) = (x-2)^2$.

How to check convexity for differentiable functions? A

- *First order convexity condition:* Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then f is convex if and only if $\forall x, y \in \mathbb{R}^n$,

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

.

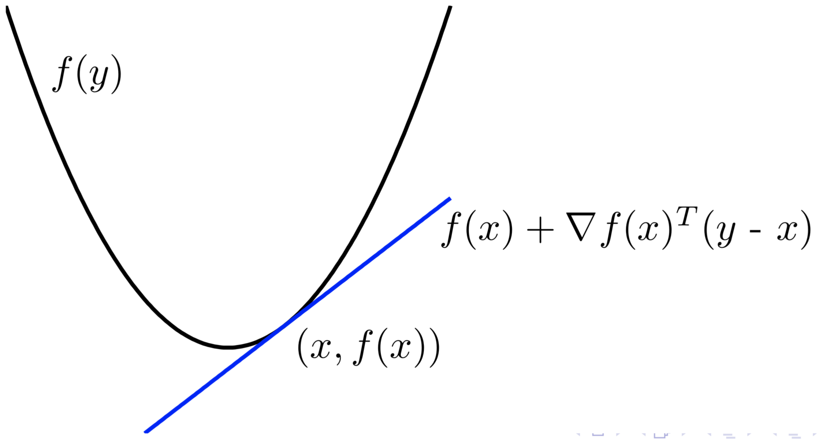


Figure 5: Source: 10-725 lecture slides by Prof. Yuanzhi Li

- *Second order convexity condition:* Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable. Then f is convex if and only if $\forall x \in \mathbb{R}^n$,

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

.

Note that the Hessian $\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots \\ \vdots & & \\ \cdots & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$ is positive semi-definite i.e. $\nabla^2 f(x) \succeq 0$ iff the eigen values of $\nabla^2 f(x)$ are all non-negative.

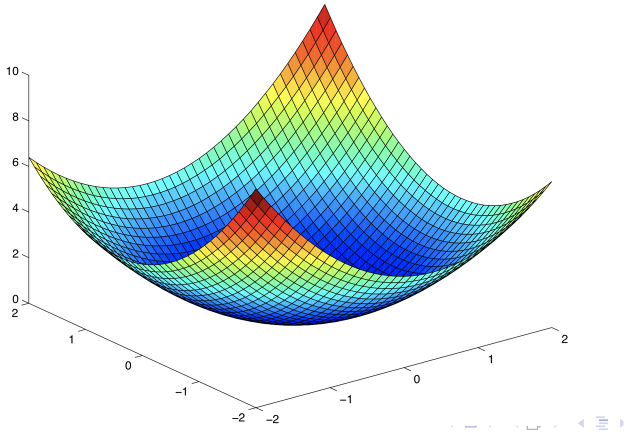


Figure 6: Source: [1]

Example: ($l2$ norm) $f(w) = \frac{1}{2}||Xw - y||^2$. Is f convex?

Solution:
$f(w) = \frac{1}{2}||Xw - y||^2 = \frac{1}{2}(Xw - y)^T(Xw - y) = \frac{1}{2}[w^T X^T X w - w^T X^T y - y^T X w + y^T y]$.
$f(w)$ is twice differentiable. $\nabla^2 f(w) = X^T X = ||X||_2^2 \geq 0$. So, f is convex.

① $f(\cdot) = \|\cdot\|_2$

0th-order condition

② $f(\alpha x + (1-\alpha) y)$

③ $\|\alpha x + (1-\alpha) y\|$

④ $\leq \|\alpha x\| + \|(1-\alpha) y\|$

⑤ $= \alpha \|x\| + (1-\alpha) \|y\|$

⑥ $= \alpha f(x) + (1-\alpha) f(y)$

Example: (MLLE for Bernoulli random variables) $f(\theta) = \log \prod_i \theta^{x_i}(1-\theta)^{(1-x_i)}$. Is f convex?

Solution:

$$\frac{df}{d\theta} = \sum_i [\frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}]$$

$$\frac{d^2 f}{d\theta^2} = \sum_i [-\frac{x_i}{\theta^2} - (1-x_i)(\frac{-1}{(1-\theta)^2})(-1)] \tag{2}$$

$$= -[\sum_i \frac{x_i}{\theta^2} + \frac{1-x_i}{(1-\theta)^2}] \leq 0 \tag{3}$$

So, log likelihood function is concave, whereas the negative log likelihood is convex.

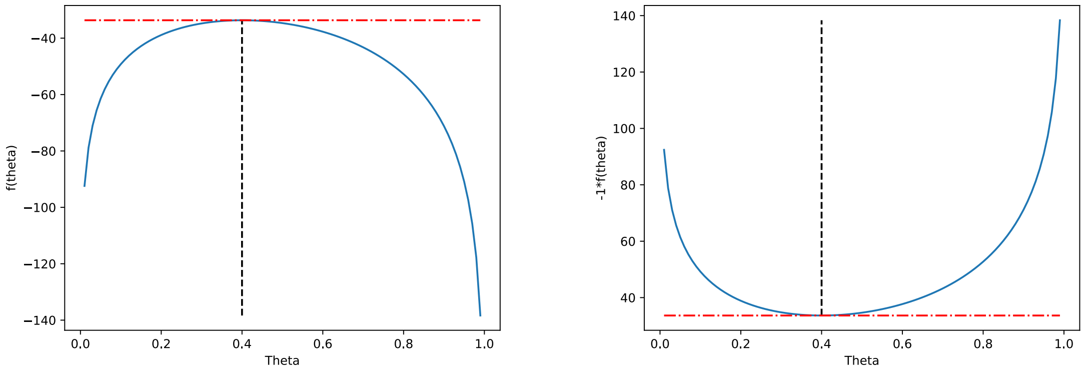


Figure 7: We have the log-likelihood function on the left and the negative log-likelihood on the right. $\theta_{MLE} = 0.4$ over the sampled x's. The red dashed line shows that the gradient goes to zero at $\theta = \theta_{MLE}$.

$f''(x) > 0$

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

4.2 Strict Convexity S

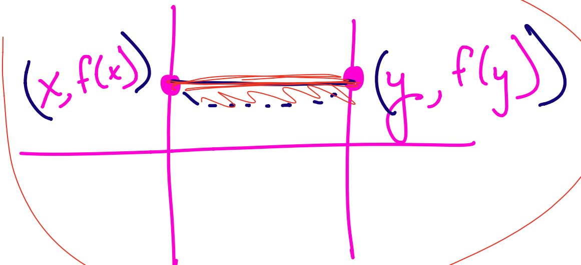
Definition (strictly convex function). A function f over a convex set D is strictly convex if it satisfies the following property:

$$\forall x, y \in D, \forall \alpha \in [0,1],\ f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y)$$

Claim: Let f be a strictly convex function. If f has a global minimizer, then that global minimizer is *unique*.

Proof: By contradiction.

Let x and y be 2 distinct $(x \neq y)$ minimizers of f. Let $f^\star := f(x) = f(y)$. By strict convexity of f:



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

Since the domain is convex, $\alpha x + (1-\alpha)y$ is in the domain. Thus, there is a point $z = \alpha x + (1-\alpha)y$ in the domain of f which yields a lower function value than $f^\star$. Therefore, x and y are not minimizers of f over its domain. Contradiction! Therefore, if x is a global minimizer of f, it must be the unique global minimizer.

But what if $\alpha x + (1-\alpha)y$ were not in the domain? CONVEX SETS

Example: $f(x) = x^2$ on $(-\infty, -a] \cup [a, \infty)$.

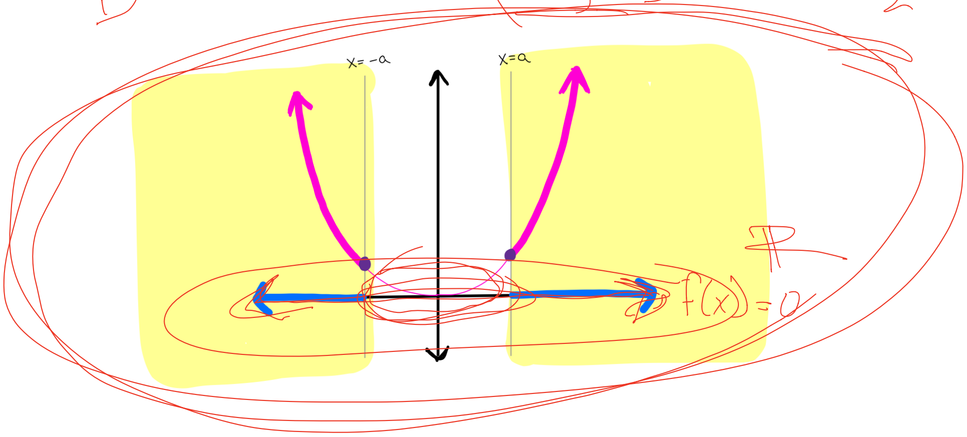
D

$(-a, a)$

$\frac{1}{2}(-a) + \frac{1}{2}(a) = 0$

$2x$

2

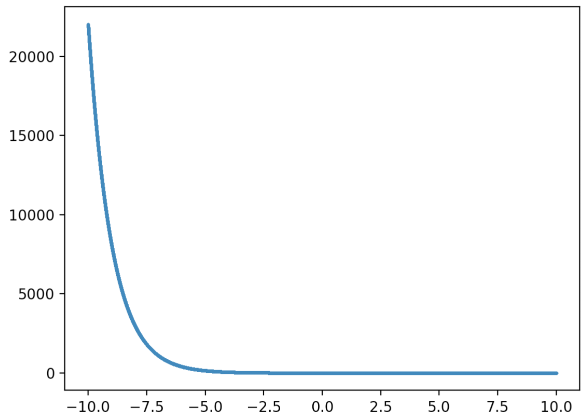


This is one example of why we focus on convex sets as our domains. More generally, certain statements we want to make about convex functions might not hold or allow for weird pathologies when the set is not convex. To be safe, we want to ***always*** be able to assume that, for any points x, y in the domain, $\alpha x + (1-\alpha)y$ is in the domain too!

Why did we say IF? ("if" f has a global minimizer)

$-e^{-x}$

$e^{-x} > 0$



Above is a plot of $y = e^{-x}$. What's the minimum?

Trick question! There's no minimum; to be a minimum, it must be attained by the function. And there's thus no minimizer. (There's an infimum, but this is a bit different.)

So it's not true that every strictly convex function on a convex domain has a minimizer.

4.3 Optimization S

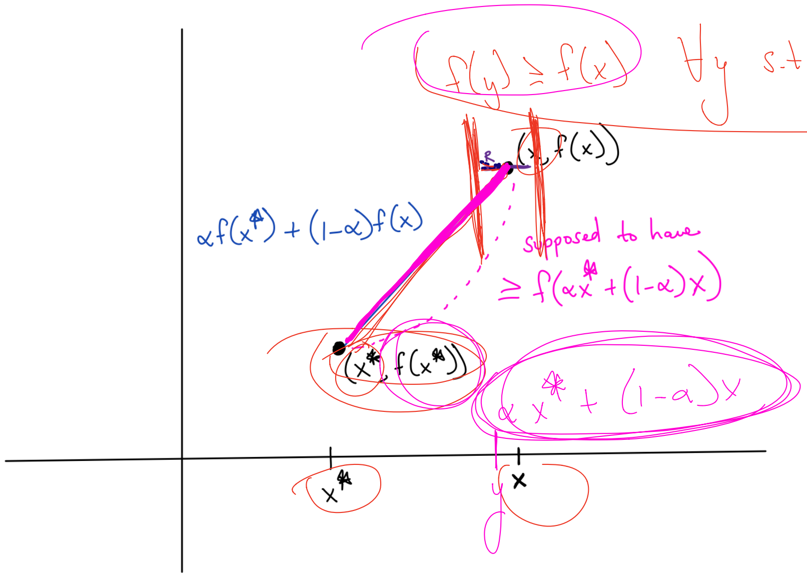
Claim: For a convex function over a convex domain, local minimizer $\implies$ global minimizer.

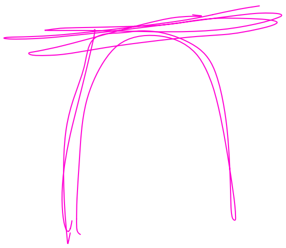
Proof: Let x be a local minimizer of f. This means that, for all y in a neighborhood of x, $f(y) \geq f(x)$.
(Proof by contradiction) Now suppose x is *not* a global minimizer. This means there exists some $x^\star$ s.t. $f(x^\star) < f(x)$.
By convexity, $\forall \alpha \in [0, 1]$, $f(\alpha x^\star + (1-\alpha)x) \leq \alpha f(x^\star) + (1-\alpha) f(x)$.
Choose α s.t. $y = \alpha x^\star + (1-\alpha)x$ is in this neighborhood of x.
Then $f(y) \leq \alpha f(x^\star) + (1-\alpha) f(x)$, by convexity of f.
But $\alpha f(x^\star) + (1-\alpha) f(x) < f(x) \leq f(y)$, so we have $f(y) < f(y)$, a contradiction.

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

$= f(x)$

$f(y) \geq f(x) \quad \forall y \text{ s.t. } |y - x| < R$


$(x, f(x))$
$\alpha f(x^*) + (1-\alpha) f(x)$
supposed to have
$\geq f(\alpha x^* + (1-\alpha)x)$
$(x^*, f(x^*))$
$\alpha x^* + (1-\alpha)x$
x^*
y
x



If f is convex, then $\nabla f(x) = 0$ if and only if x is a global minimizer of f.

A: Visualizations, `https://colab.research.google.com/drive/1EEkABrqHR33L6HnbTOH6UK8auZdez2SY?usp=sharing`

4.4 Jensen's Inequality (20 min), A

- Inequality of non-linear mappings
if f is convex, then for any distribution $\mathbf{P}$ over m,

$$\mathbb{E}_{m\sim\mathbf{P}}[f(X(m))] \geq f(\mathbb{E}_{m\sim\mathbf{P}}[X(m)])$$

- For concave functions, the inequality is reversed.

Proof: Let us consider a discrete probability distribution P. Expanding the LHS,

$$\begin{aligned} \mathbb{E}_{m\sim\mathbf{P}}[f(X(m))] &= \sum_i p_i f(X(m_i)) && (4)\\ &\geq f(\sum_i p_i X(m_i)) \text{ by the convexity of } f && (5)\\ &= f(\mathbb{E}_{m\sim\mathbf{P}}[X(m)]) && (6) \end{aligned}$$

Example: Prove that Arithmetic Mean (AM) $\geq$ Geometric Mean (GM), where AM $= \sum_{i=1}^{n} w_i x_i$ and GM $= \prod_{i=1}^{n} {x_i}^{w_i}$ such that $\forall i, w_i \geq 0, \quad \sum_i w_i = 1$.

Solution: We will use the fact that log is a concave function and apply Jensen's inequality. Let $f = \sum_{i=1}^{n} w_i x_i$, $g = \prod_{i=1}^{n} {x_i}^{w_i}$.

$$\begin{aligned} \log g &= \sum_{i=1}^{n} w_i \log x_i \\ &\leq \log(\sum_{i=1}^{n} w_i x_i) \text{ using Jensen's inequality for } \textit{log} \qquad (7) \\ &= \log f \qquad (8) \end{aligned}$$

$$\implies \log g \leq \log f \implies g \leq f \implies \text{GM} \leq \text{AM}$$

5 References:

- https://web.stanford.edu/~boyd/cvxbook/bv_cvxslides.pdf
- https://ttic.uchicago.edu/~dmcallester/ttic101-07/lectures/jensen/jensen.pdf
- 10-725 Lecture Slides by Prof. Yuanzhi Li

$$w^T X^T X w \overset{?}{\geq} 0$$

$$(Xw)^T Xw = \|Xw\|_2^2 \geq 0$$

$n \times d \; d \times 1$

$n \times 1$

$1 \times n \quad n \times 1$