Lecture 7 Machine Learning & Optimization

- Most important consequence of convex optim probs:
  local minima are global minima
- Strict convexity and strong convexity:
  strict convexity $\Rightarrow$ the minimizer (if any) is unique
- First and second-order characterizations of convex functions
- Examples of convex functions using
- Operations that presere convexity of functions
- Examples of convex optimization problems

Local minima of convex optimization problems are global minima

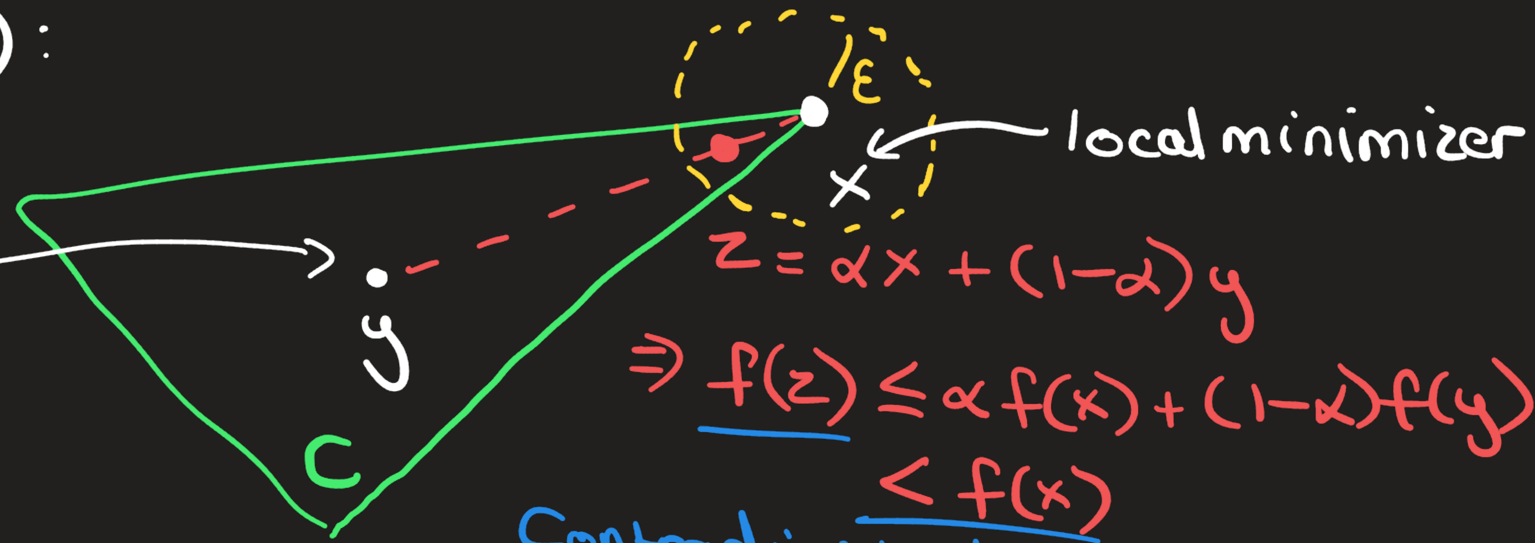
A point x is a local minimizer of a function f if there exists an $\varepsilon > 0$ such that when $\|y-x\|_2 \leq \varepsilon$ and $y \in \mathrm{dom}(f)$, then $f(y) \geq f(x)$.

A point x is a global minimizer of f if for all $y \in \mathrm{dom}(f)$, $f(y) \geq f(x)$.

Prf (sketch) of fact that local minimizers of convex optim probs are global minima

Consider a local minimizer of f on a convex set $C \subseteq \mathrm{dom}(f)$:

Assume there exists a point y such that $f(y) < f(x)$



$z = \alpha x + (1-\alpha) y$

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

$< f(x)$

Contradiction!

Fact

If f is a convex function, its set of minimizers is a convex set

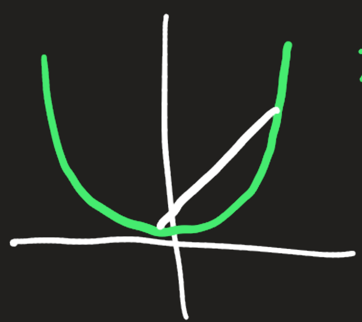
- It may be empty, e.g. $\emptyset = \underset{x>0}{\operatorname{argmin}} \frac{1}{x}$
- It may be a singleton, e.g. $0 = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\ x^2$
- It may be infinite, e.g. $[1,\infty) = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\ (1-x)_+$ 

Def Strict convexity

A function f is strictly convex if $\forall\, x, y \in \operatorname{dom}(f)$

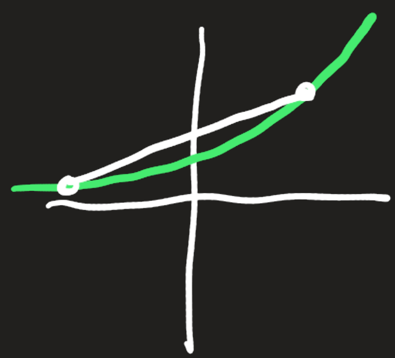
$$f(\alpha x + (1-\alpha) y) < \alpha f(x) + (1-\alpha) f(y)$$

Ex



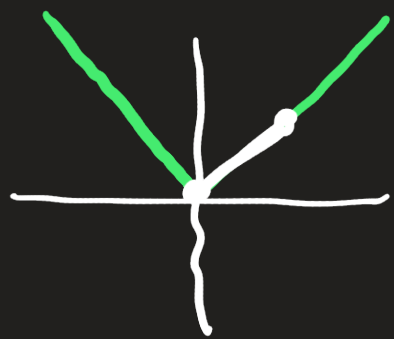
x^2

strictly cvx



e^x

strictly cvx



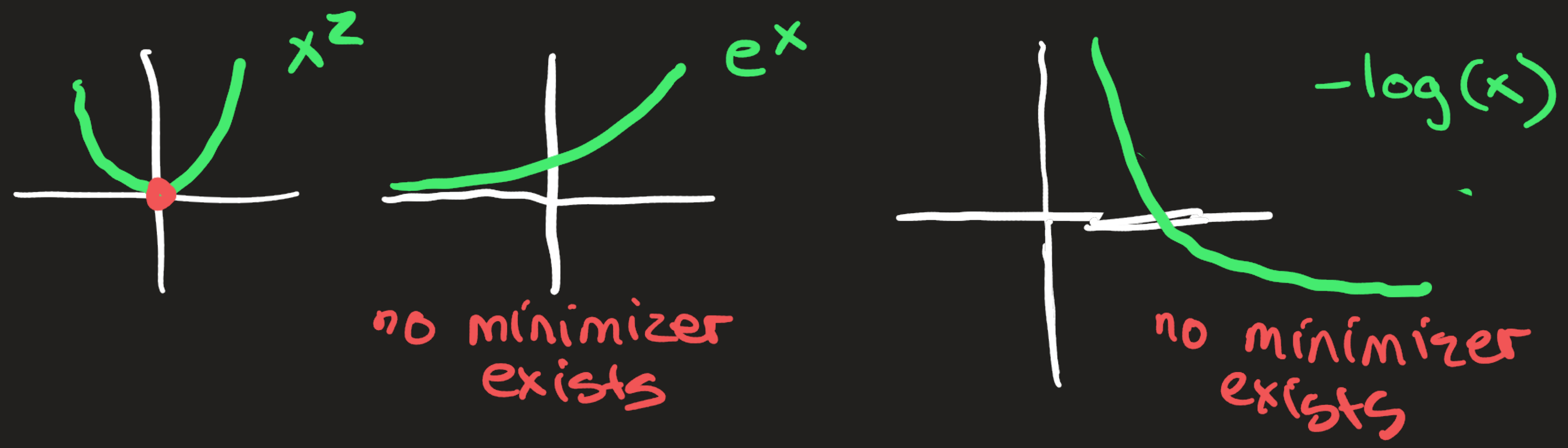
$|x|$

convex but not strictly convex

Fact

The minimizer of a strictly convex function over a convex set, if it exists, is unique.

Ex



Fact

If f and g are convex, and any of f and g are strictly convex, then $f+g$ is strictly convex

In practice, we often more commonly deal with strongly convex functions.

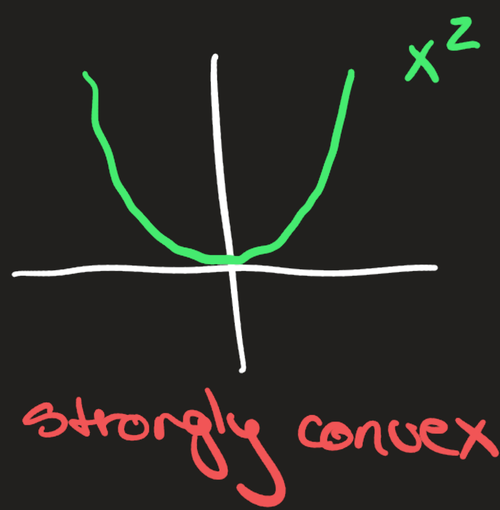
<u>Def</u> Strong convexity (Intuition: at every point the function has at least a curvature of μ)

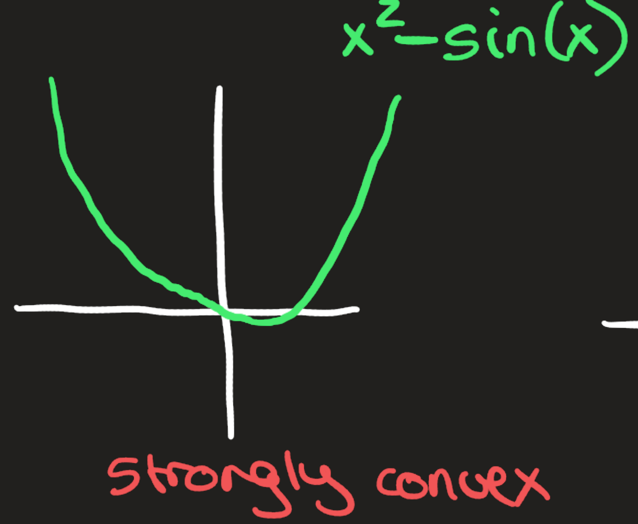
A function f is strongly convex if for all $x, y \in \mathrm{dom}(f)$,

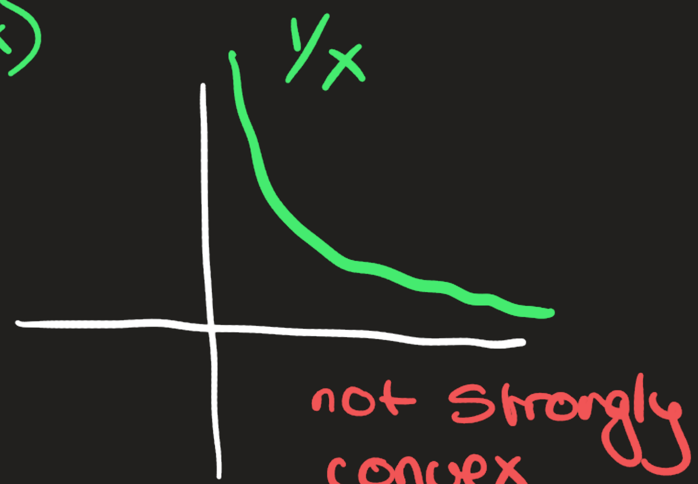
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\alpha(1-\alpha)\mu}{2} \|x-y\|_2^2$$

whenever $\alpha \in [0,1]$. Here, $\mu \in \mathbb{R}_+$ is the strong convexity constant of f.

<u>Ex</u>

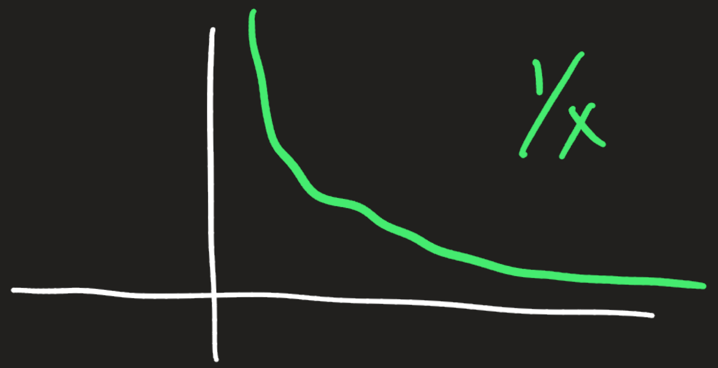






__Fact__

Strongly convex functions are strictly convex.
Consequently the minimizer of a strongly convex function over a convex set, if it exists, is unique.



not strongly convex,
but strictly convex

__Fact__

If f and g are convex, and either of f and g are μ-strongly convex, then $f+g$ is μ-strongly convex

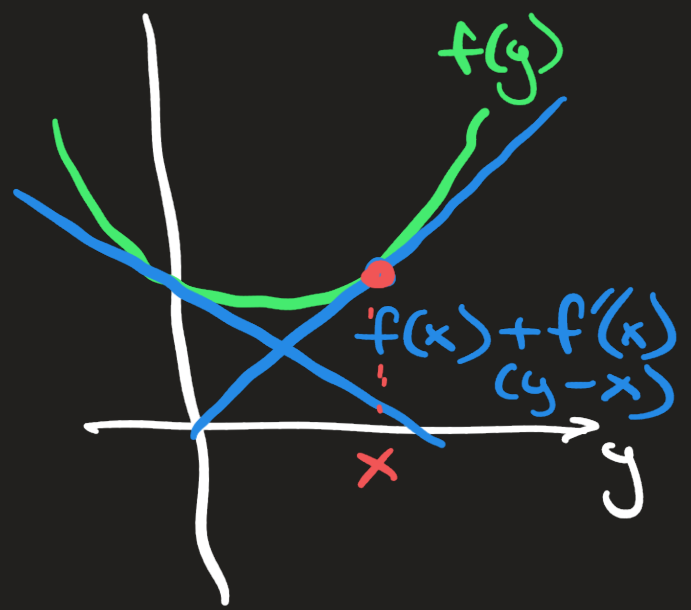
First-order characterizations of convexity

Assume f is a differentiable function

A function $f: \mathbb{R} \to \mathbb{R}$ satisfies

$$f(y) \geq f(x) + f'(x)(y-x)$$

for all $x, y \in \mathbb{R}$ iff f is convex on $\mathbb{R}$

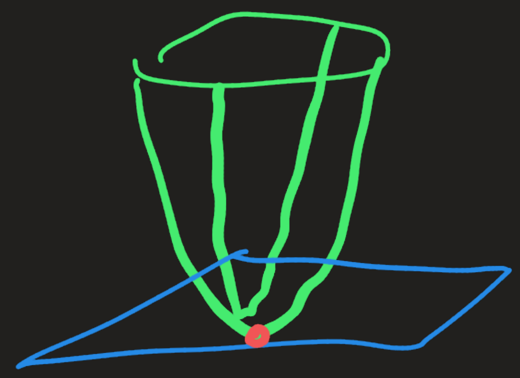


A convex function always lies on or above its first-order TS

A function $f: C \to \mathbb{R}$ where $C \subseteq \mathbb{R}^d$ is an open convex set satisfies

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

for all $x, y \in C$ iff f is convex on C



Ex

Establishing that x^2 is convex on $\mathbb{R}$:

x^2 is convex on $\mathbb{R}$ $\iff$ $\forall x,y:\ y^2 \geq x^2 + 2x(y-x)$

$\iff \forall x,y:\ y^2 \geq -x^2 + 2xy$

$\iff \forall x,y:\ y^2 + x^2 - 2xy \geq 0$

$\iff \forall x,y:\ (y-x)^2 \geq 0$

Establishing that $\frac{1}{x}$ is convex on $\mathbb{R}_{++} = \{x \mid x > 0\}$:

$\frac{1}{x}$ is convex on $\mathbb{R}_{++}$ $\iff$ $\forall x,y \in \mathbb{R}_{++}:\ \frac{1}{y} \geq \frac{1}{x} - \frac{1}{x^2}(y-x)$

$\iff \forall x,y \in \mathbb{R}_{++}:\ \frac{x^2}{y} \geq x - (y-x)$

$\iff \forall x,y \in \mathbb{R}_{++}:\ x^2 \geq yx - y(y-x)$

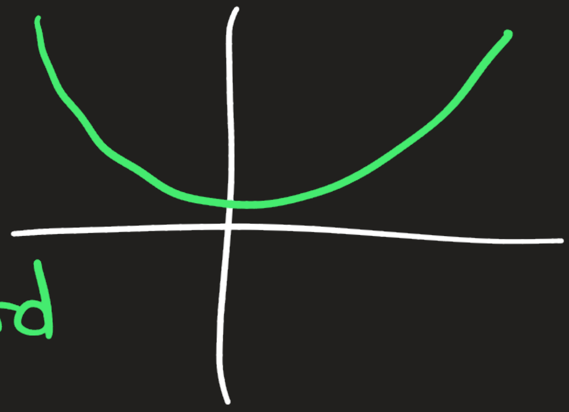
$\iff \forall x,y \in \mathbb{R}_{++}:\ x^2 \geq 2xy - y^2$

$\iff \forall x,y \in \mathbb{R}_{++}:\ (x-y)^2 \geq 0$

Second-order characterization of convexity

Assume f is second-differentiable

> A function $f: \mathbb{R} \to \mathbb{R}$ is convex iff
> $f''(x) \geq 0$ for all $x \in \mathbb{R}$



A convex function is always curved upward

> A function $f: C \to \mathbb{R}$, where $C \subseteq \mathbb{R}^d$ is an open convex set is convex on C iff
> $$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in C$$

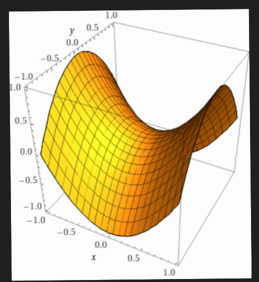
<u>Ex:</u> let $f(x) = \frac{1}{2} x^T A x$

$$= \frac{1}{2} \sum_{ij} a_{ij} x_i x_j$$

then

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

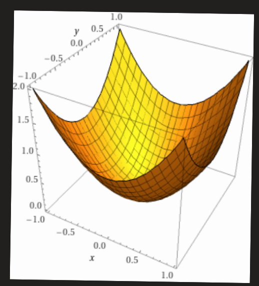
$A \succeq 0$ means f has positive curvature in all directions



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

nonconvex

$\frac{1}{2}(x^2 - y^2)$



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

convex

$\frac{1}{2}(x^2 + y^2)$

__Ex__ $f(x) = -\log(x)$ is convex on $\mathbb{R}_{++}$

because $f''(x) = \frac{1}{x^2} > 0$ when $x \in \mathbb{R}_{++}$

$f(x) = x^2 - \sin(x)$ is convex on $\mathbb{R}$

because

$$f''(x) = 2 + \sin(x) > 0 \quad \text{on } \mathbb{R}$$

__Ex__

$\mathrm{logsumexp}(\cdot)$ is convex on $\mathbb{R}^d$, $\mathrm{logsumexp}(x) = \log\left(\sum_{i=1}^{d} e^{x_i}\right)$

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

$$\left[\nabla f(x)\right]_j = \frac{1}{\sum_{i=1}^{d} e^{x_i}} e^{x_j} = \left[\mathrm{softmax}(x)\right]_j$$

We've shown a fact that is interesting in itself:

$$\nabla \operatorname{logsumexp}(x) = \operatorname{softmax}(x)$$

Using this,

$$[\nabla^2 f(x)]_{ij} = \frac{\partial}{\partial x_i}[\operatorname{softmax}(x)]_j = \frac{\partial}{\partial x_i}\left[\frac{e^{x_j}}{1^T e^x}\right]$$

$$= \begin{cases} \text{when } i=j, & \dfrac{e^{x_i}(1^T e^x) - e^{2x_i}}{(1^T e^x)^2} = \dfrac{e^{x_i}}{1^T e^x} - \dfrac{e^{2x_i}}{(1^T e^x)^2} \\ \text{when } i \neq j, & \dfrac{-e^{x_i + x_j}}{(1^T e^x)^2} \end{cases}$$

$$= \begin{cases} \operatorname{softmax}(x)_i - [\operatorname{softmax}(x)_i]^2, & i=j \\ -\operatorname{softmax}(x)_i \cdot \operatorname{softmax}(x)_j, & i \neq j \end{cases}$$

so letting $\theta = \mathrm{softmax}(x)$, we see that

$$\nabla^2 f(x) = \mathrm{Diag}(\theta) - \theta\theta^T$$

Recall that our goal is to establish that $\nabla^2 f(x) \succeq 0$ for every $x \in \mathbb{R}^d$, to conclude that $f = \mathrm{logsumexp}$ is convex on $\mathbb{R}^d$.

We will use the characterization that

$$\nabla^2 f(x) \succeq 0 \quad \text{iff} \quad z^T \nabla^2 f(x) z \geq 0 \text{ for all } z \in \mathbb{R}^d$$

To show this, notice that θ is a probability vector, so

$$\begin{aligned} z^T \nabla^2 f(x) z &= z^T \left[\mathrm{Diag}(\theta) - \theta\theta^T \right] z \\ &= \underbrace{\sum_{i=1}^{d} \theta_i z_i^2}_{\text{a second moment!}} - \underbrace{\left(\sum_{i=1}^{d} \theta_i z_i \right)^2}_{\text{a squared expectation!}} \end{aligned}$$

This looks like a variance, and variances are always nonnegative.

To make this concrete, define a r.v. φ taking values in $\{z_1, \dots, z_d\}$ with $\mathbb{P}(\varphi = z_i) = \theta_i$.

Then

$$\mathrm{Var}(\varphi) = \mathbb{E}(\varphi - \mathbb{E}\varphi)^2 = \mathbb{E}\varphi^2 - (\mathbb{E}\varphi)^2$$

$$= \sum_{i=1}^{d} \theta_i z_i^2 - \left(\sum_{i=1}^{d} \theta_i z_i\right)^2 \geq 0$$

so for any $z \in \mathbb{R}^d$,

$$z^T \nabla^2 f(x) z = \mathrm{Var}(\varphi) \geq 0$$

so $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^d$. Consequently $f(x) = \mathrm{logsumexp}(x)$ is convex on $\mathbb{R}^d$.

Whenever possible, we prefer to avoid doing the calculus to check that functions are convex. For this reason, it is helpful to know some operations that preserve convexity of functions.

Operations that preserve convexity of functions

- $\|x\|_2$ is convex on $\mathbb{R}^d$

$$\|\alpha x + (1-\alpha) y\|_2 \leq \|\alpha x\|_2 + \|(1-\alpha) y\|_2 \quad \text{by } \Delta\text{-inequality}$$
$$= \alpha \|x\|_2 + (1-\alpha) \|y\|_2$$

In fact, <u>any norm</u> is convex on $\mathbb{R}^d$

- $f(x) = a^T x + b$ (affine functions) are convex

(proof: exercise)

- $f(x) = \frac{1}{2} x^T A x + b^T x + c$ (quadratic function)

  is convex iff $\nabla^2 f(x) = \frac{A + A^T}{2} \succcurlyeq 0$

  so if A is symmetric, $f(x)$ is convex iff $A \succcurlyeq 0$

- If $f_1, \ldots, f_n$ are convex on a set C and $\alpha_1, \ldots, \alpha_n$ are nonnegative real numbers, then

  $$\sum_{i=1}^{n} \alpha_i f_i \text{ is convex on } C$$

  (proof: exercise)

- If f is convex on $\mathrm{dom}(f)$, then

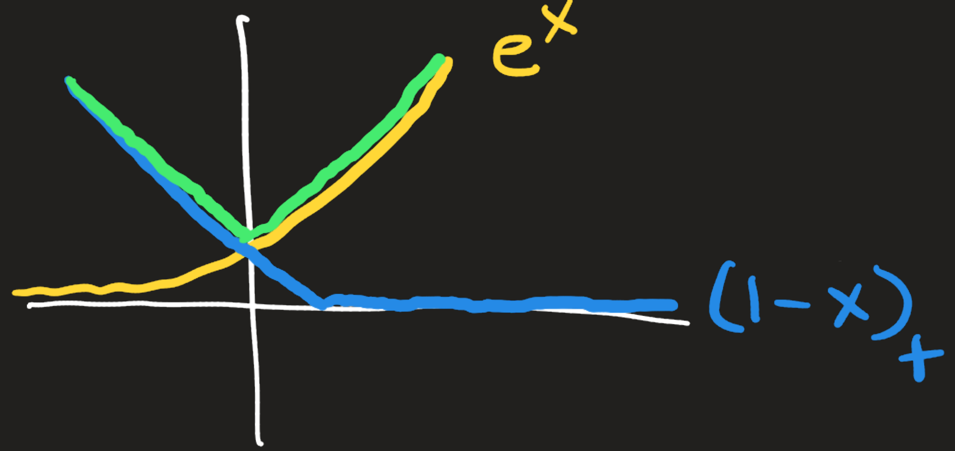
  $f(Ax+b)$ is convex on the set of x such that

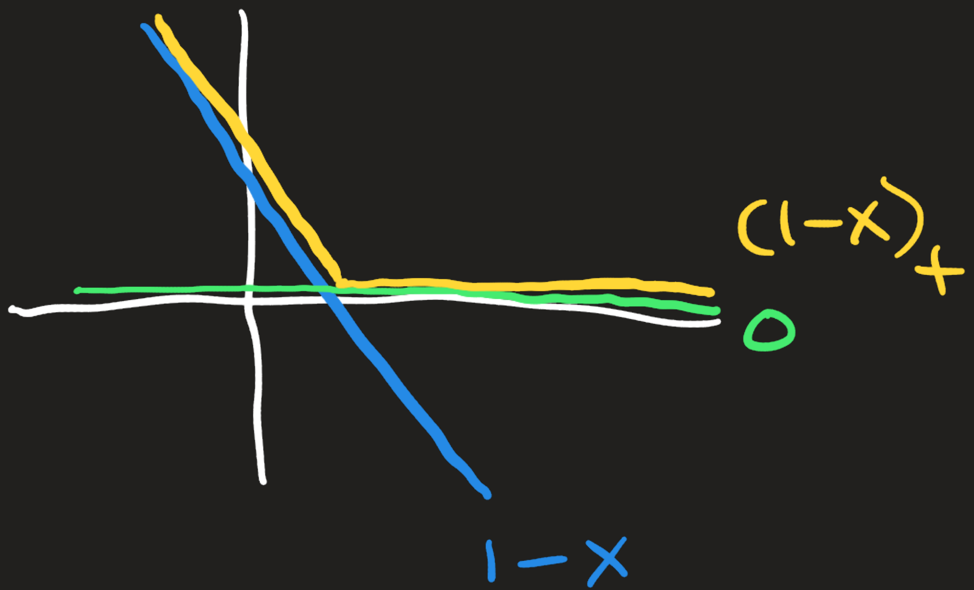
  $$Ax + b \in \mathrm{dom}(f)$$

  (proof: exercise)

- The pointwise maximum of convex functions is convex on the intersection of their domains is convex

$f(x) = \max\{e^x, (1-x)_+\}$





<u>Prf:</u> The epigraph of $f(x) = \max_{i \in I} f_i(x)$ is the set of points that lie above <u>all</u> the graphs of the f_i. That is,

$$\text{epigraph}(f) = \bigcap_{i \in I} \text{epigraph}(f_i).$$

The f_i are convex, so their epigraphs are convex. Their intersection is therefore convex. Since epigraph(f) is convex, f is convex.

- If f is a convex and non-decreasing function (i.e. $x \geq y$ implies $f(x) > f(y)$) and g is a convex function, then $f(g(x))$ is a convex function

Ex $f(x) = x^2$ is convex and non-decreasing on $\mathbb{R}_+$, so if g is convex and nonnegative, then g^2 is convex

Ex

Sometimes practitioners replace the hinge loss $\ell(\hat{y}, y) = (1 - \hat{y}y)_+$ used in SVMs with the squared hinge loss $\ell(\hat{y}, y) = (1 - \hat{y}y)_+^2$. This penalizes incorrect classifications more harshly

Cor Any squared norm is convex. In particular, the ℓ_2 regularizer term, $\lambda \|x\|_2^2$, is convex.

Examples of convex optim problems

1) Linear programming

$$\min_{\substack{Ax \le b \\ x \geq 0}} \; c^T x$$

linear functions are convex

polyhedra are convex

2) Least squares

$$\min_{x \in C} \; \|Ax - b\|_2^2$$

convex when C is convex

norm squared is convex, composition of affine function with convex is convex

e.g. when $C = \{x \geq 0\}$, we get non-negative least squares

$$\min_{x \geq 0} \; \|Ax - b\|_2^2$$

3) Regularized least squares

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda \|x\|_1$$

or

$$\min_{x \in \mathbb{R}^d} \underbrace{\|Ax - b\|_2^2 + \lambda \|x\|_2^2}_{\substack{\text{positively weighted (we take } \lambda \geq 0) \\ \text{sums of convex functions are convex}}}$$

These problems, and many more, can all be solved using the algorithms for convex optimization that we will develop over the next several weeks.