Unconstrained Multivariate Optimality and Convexity

Antonio Frangioni

Department of Computer Science
University of Pisa
https://www.di.unipi.it/~frangio
mailto:frangio@di.unipi.it

Computational Mathematics for Learning and Data Analysis
Master in Computer Science – University of Pisa

A.Y. 2024/25

Outline

- Back to $f : \mathbb{R}^n \to \mathbb{R}$, i.e., $f(x_1, x_2, \ldots, x_n) = f(x)$
- Of course need f L-c (exact definition later)
- Very bad news: no algorithm can work in less than $\Omega((LD/\varepsilon)^n)$ [3, p. 413]
- Curse of dimensionality: not really doable unless $n = 3/5/10$ tops
- Can make it in $O((LD/\varepsilon)^n)$, multidimensional grid with small enough step: the standard approach to hyperparameter optimization (but D, L unknown)
- If f analytic, clever (spatial) B&B can give global optimum
- If f black-box (typically $\Longrightarrow$ no derivatives), many effective heuristics can give good (not provably optimal) solutions [8]
- In both cases, complexity grows "fast" in practice as n grows
- Finding good global solutions hard in practice, proving optimality even worse

- Back to $f : \mathbb{R}^n \to \mathbb{R}$, i.e., $f(x_1, x_2, \ldots, x_n) = f(x)$
- Of course need f L-c (exact definition later)
- Very bad news: no algorithm can work in less than $\Omega((LD/\varepsilon)^n)$ [3, p. 413]
- Curse of dimensionality: not really doable unless $n = 3/5/10$ tops
- Can make it in $O((LD/\varepsilon)^n)$, multidimensional grid with small enough step: the standard approach to hyperparameter optimization (but D, L unknown)
- If f analytic, clever (spatial) B&B can give global optimum
- If f black-box (typically $\Longrightarrow$ no derivatives), many effective heuristics can give good (not provably optimal) solutions [8]
- In both cases, complexity grows "fast" in practice as n grows
- Finding good global solutions hard in practice, proving optimality even worse unless f convex $\Longrightarrow$ global $\equiv$ local

- Local optimization much better

- Results in general surprisingly analogous to (multivariate) quadratic case: most (but not all) convergence results are dimension-independent $\equiv$ complexity does not explicitly depends on n (if it does, not exponentially)

- Not completely surprising: linear / quadratic models a staple

- Does not mean all local algorithms are fast:
  - convergence speed may be rather low ("badly linear" or worse)
  - cost of f / derivatives computation necessarily increases with n: for large $n \approx 10^9$, even $O(\, n^2\,)$ is too much (will see)
  - some dependency on n may be hidden in $O(\cdot)$ constants

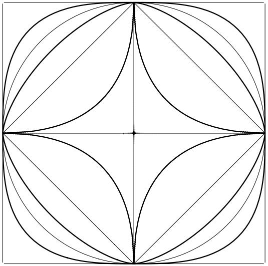
- Yet, large-scale local optimization is doable if you have derivatives

- Except, derivatives in $\mathbb{R}^n$ are significantly more complex

Outline

- Fundamental (easy) concept: $\mathcal{B}(x\,,\,r) := \{\, z \in \mathbb{R}^n \,:\, \|\, z - x \,\| \leq r \,\}$
  ball, center $x \in \mathbb{R}^n$, radius $r > 0$ = points "close" to x in the chosen norm

- Euclidean norm just one member of a large family:
  $\|\, x \,\|_p := (\, \sum_{i=1}^n |\, x_i \,|^p \,)^{1/p}$ p-norm, $p > 0$
  - Euclidean $\equiv \|\, x \,\|_2$, $\|\, x \,\|_1 := \sum_{i=1}^n |\, x_i \,|$ (Lasso)
  - $\lim_{p \to \infty} \equiv \|\, x \,\|_\infty := \max\{\, |\, x_i \,| \,:\, i = 1, \ldots, n \,\}$
  - $\lim_{p \to 0} \equiv \|\, x \,\|_0 := \#\{\, i \,:\, |\, x_i \,| > 0 \,\}$ (not norm)



- Other norms $\exists$ besides p-norm (matrix norms ...)

- Pictured $S(\,\|\cdot\|_p\,,\,1\,) \equiv \mathcal{B}_p(\,0\,,\,1\,)$, $p = 0, 1/2, 1, 3/2, 2, 3, \infty$ (grow with p)

- The norm defines the topology of $\mathbb{R}^n$, but doesn't really matter:
  all is "$\exists$ ball", "$\forall$ small ball", and all norms are equivalent [9]
  $\forall \,\|\cdot\|,\ |||\cdot|||\ \exists\, 0 < \alpha < \beta$ s.t. $\alpha\|\, x \,\| \leq |||\, z \,||| \leq \beta\|\, x \,\|\ \ \forall\, x, z \in \mathbb{R}^n$

- Limit of sequence $\{ x_i \} \subset \mathbb{R}^n$:

$$\lim_{i \to \infty} x_i = x \quad \equiv \quad \{ x_i \} \to x$$

$$\iff \quad \forall \varepsilon > 0 \ \exists h \text{ s.t. } d(x_i , x) \leq \varepsilon \ \forall i \geq h$$

$$\iff \quad \forall \varepsilon > 0 \ \exists h \text{ s.t. } x_i \in \mathcal{B}(x , \varepsilon) \ \forall i \geq h$$

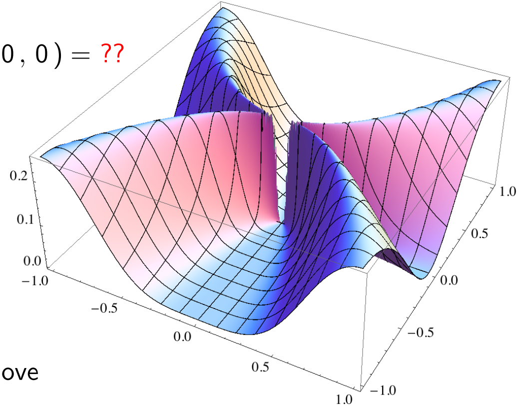
$$\iff \quad \lim_{i \to \infty} d(x_i , x) = 0$$

- Points of $\{ x_i \}$ eventually all come arbitrarily close to x

- Note that $\mathbb{R}^n$ "exponentially larger" than $\mathbb{R}$ $\Longrightarrow$
  there are many more ways for $\{ x_i \} \to x$ in $\mathbb{R}^n$ than in $\mathbb{R}$

- This may lead to more tricky situations / concepts

- Same definitions:
  - f continuous at x: $\{x_i\} \to x \implies \{f(x_i)\} \to f(x)$
  - $f \in C^0$: continuous $\forall\, x \in \mathbb{R}^n$
- There are "many" different $\{x_i\} \to x$, the limit must be $=$ for all
- Not sufficient to only consider "simple" sequences
- $f(x_1, x_2) = \left[\dfrac{x_1^2 x_2}{x_1^4 + x_2^2}\right]^2 \quad f(0, 0) = ??$
- Limit $=$ "on straight lines" $\forall\, [d_1, d_2] \in \mathbb{R}^2$ $\lim_{k\to\infty} f(d_1/k, d_2/k) = 0$
- Limit $\neq$ on "curved" line $\lim_{k\to\infty} f(1/k, 1/k^2) = 1/4$



Exercise: Prove the two limits above

- $f : \mathbb{R}^n \to \mathbb{R}$, directional derivative at $x \in \mathbb{R}^n$ along direction $d \in \mathbb{R}^n$:

  $$\frac{\partial f}{\partial d}(x) := \lim_{t \to 0} \frac{f(x + td) - f(x)}{t} = \varphi'_{x,d}(0)$$

- Scales linearly with $\| d \|$: $\frac{\partial f}{\partial \beta d}(x) = \beta \frac{\partial f}{\partial d}(x)$ (sounds familiar?) (**check**)
- One-sided directional derivative: $\lim_{t \to 0_\pm} \ldots = [\varphi_{x,d}]'_\pm(0)$
- The derivative of the (x, d)−tomography (in 0): how can it be computed?
- Special case: partial derivative of f w.r.t. x_i at $x \in \mathbb{R}^n$

  $$\frac{\partial f}{\partial x_i}(x) := \lim_{t \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + t, x_{i+1}, \ldots, x_n) - f(x)}{t} = [f_x^i]'(x_i) = \frac{\partial f}{\partial u^i}(x)$$

- The derivative of the restriction of f to x_i is easy to compute: just

  $f'(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$ treating x_j for $j \neq i$ as constants

- Gradient = (column) vector of all partial derivatives, "easy to compute" [6]

  $$\nabla f(x) := \left[\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right]^T \in \mathbb{R}^n$$

- $f(x) = \langle b, x \rangle \implies \nabla f(x) = b$

- $f : \mathbb{R}^n \to \mathbb{R}$, directional derivative at $x \in \mathbb{R}^n$ along direction $d \in \mathbb{R}^n$:

  $$\frac{\partial f}{\partial d}(x) := \lim_{t \to 0} \frac{f(x + td) - f(x)}{t} = \varphi'_{x,d}(0)$$

- Scales linearly with $\| d \|$: $\frac{\partial f}{\partial \beta d}(x) = \beta \frac{\partial f}{\partial d}(x)$ (sounds familiar?) (**check**)
- One-sided directional derivative: $\lim_{t \to 0_\pm} \ldots = [\varphi_{x,d}]'_\pm(0)$
- The derivative of the $(x, d)-$tomography (in 0): how can it be computed?
- Special case: partial derivative of f w.r.t. x_i at $x \in \mathbb{R}^n$

  $$\frac{\partial f}{\partial x_i}(x) := \lim_{t \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + t, x_{i+1}, \ldots, x_n) - f(x)}{t} = [f^i_x]'(x_i) = \frac{\partial f}{\partial u^i}(x)$$

- The derivative of the restriction of f to x_i is easy to compute: just

  $f'(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$ treating x_j for $j \neq i$ as constants

- Gradient = (column) vector of all partial derivatives, "easy to compute" [6]

  $$\nabla f(x) := \left[\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right]^T \in \mathbb{R}^n$$

- $f(x) = \frac{1}{2} x^T Q x + qx \implies \nabla f(x) = Qx + q$

- f differentiable at x if $\exists$ linear function $\phi(h) = \langle b, h \rangle + f(x)$ s.t.

$$\lim_{\| h \| \to 0} \frac{| f(x+h) - \phi(h) |}{\| h \|} = 0 \; [\implies \phi(0) = f(x) \implies c = f(x)]$$

$\varphi \equiv$ "first order model" of f at x, the error "vanishes faster than linearly"

- f differentiable at $x \implies b = \nabla f(x)$ [5, Th. 5.3.6]
  $\implies \frac{\partial f}{\partial x_i}(x)$ exists $\forall\, i$ (but $\Longleftarrow$ not true)
  $\implies$ first-order model of f at x: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$

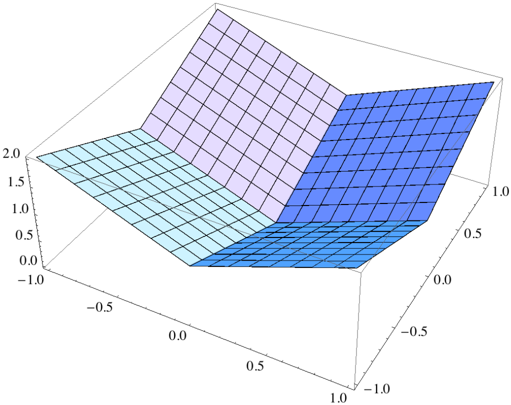
- f differentiable at $x \implies \nabla f(x)$ gives all $\frac{\partial f}{\partial d}$ [5, Ex 5.3.19]:
  $\forall d \in \mathbb{R}^n \quad \frac{\partial f}{\partial d}(x) = \langle \nabla f(x), d \rangle \quad (\Longleftarrow \exists)$

- [5, Th. 5.3.10,Th. 5.3.7] $\exists\, \delta > 0$ s.t. $\forall i\; \frac{\partial f}{\partial x_i}(z)$ continuous $\forall z \in \mathcal{B}(x, \delta)$
  $\implies f$ differentiable at $x \implies f$ continuous at x

- $\frac{\partial f}{\partial x_i} \in C^0 \implies f$ differentiable everywhere $\equiv f \in C^1$
  (but $\not\Longleftarrow$, $\exists$ weird f differentiable with discontinuous $\frac{\partial f}{\partial x_i}$ [5, Ex. 5.3.9])

- (non)differentiability in $\mathbb{R}^n$ is much weirder than in $\mathbb{R}$

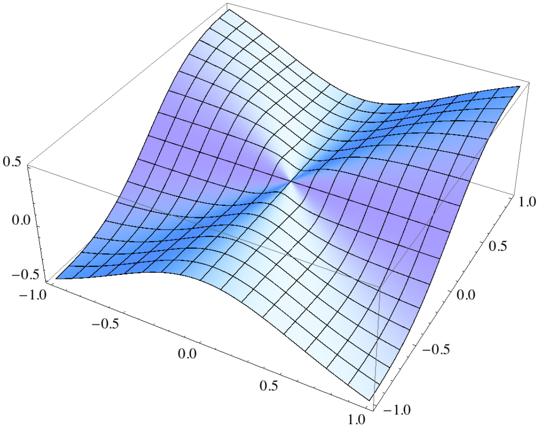
- $f(x_1, x_2) = \| [x_1, x_2] \|_1 = |x_1| + |x_2|$
- f continuous everywhere (why?)
- $\exists\, d \in \mathbb{R}^n$ s.t. $\nexists \frac{\partial f}{\partial d}(0, 0)$
- f non differentiable in $[0, 0]$



Exercise: where else f is non differentiable? Prove it is not

Non-differentiability II

- $f(x_1, x_2) = \dfrac{x_1^2 x_2}{x_1^2 + x_2^2}$
- Can take $f(0, 0) = 0$ as
  $$\lim_{[x_1, x_2] \to [0, 0]} f(x_1, x_2) = 0$$
- $\exists \frac{\partial f}{\partial d}$ $\forall d \in \mathbb{R}^n$, but
  f non differentiable in $[0, 0]$



Exercise: prove $\lim_{x \to 0} f(x) = 0$, first "along lines" then in general

Exercise: prove all this (hint: compute $[\partial f / \partial d](0, 0)$ for generic $d = [d_1, d_2]$, prove it cannot have the form $\langle v, d \rangle$ for any v)

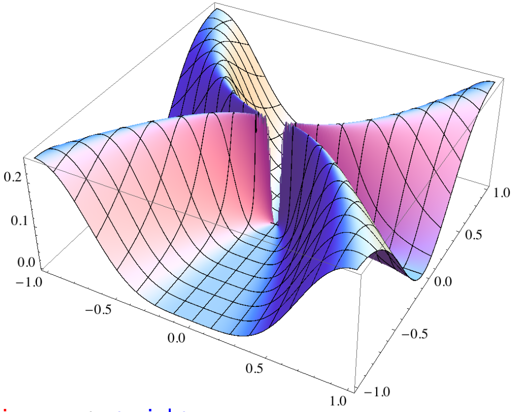
Exercise: alternatively, compute ∇f and prove it is not continuous in $[0, 0]$ (hint: look at picture of $\partial f / \partial x_2$ for directions where the limit is $\neq$)

- $f(x_1, x_2) = \left[\dfrac{x_1^2 x_2}{x_1^4 + x_2^2} \right]^2$

- f not continuous $\Longrightarrow$ not differentiable at $[0, 0]$

- $\frac{\partial f}{\partial d}(0, 0) = 0 \;\; \forall d \in \mathbb{R}^n$

- $\nexists \nabla f$, but $\exists v \, (= 0)$ s.t. $\frac{\partial f}{\partial d} = \langle v, d \rangle \; \forall d \in \mathbb{R}^n$



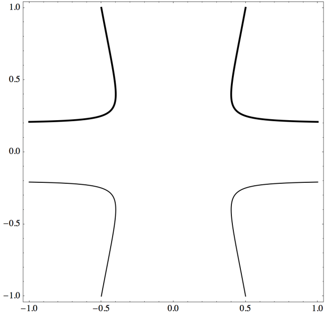
- f does nasty things on curved lines, not straight ones

Exercise: prove $\frac{\partial f}{\partial d}(0, 0) = 0$

- In $\mathbb{R}^2$, $L(L_x, f(x))$ is a line passing by x and $\nabla f(x) \perp L(L_x, f(x))$

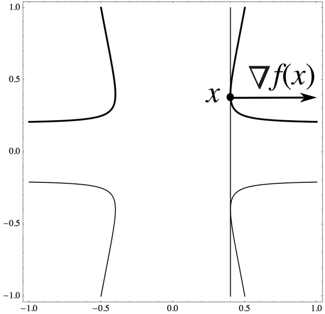
- In $\mathbb{R}^n$, $L(L_x, f(x))$ is a surface passing by x and $\nabla f(x) \perp L(L_x, f(x))$

$$f(x_1\ x_2) = \frac{x_1^2 x_2}{x_1^2 + x_2^2} \quad , \quad \nabla f(x) = \left[\frac{2x_1 x_2^3}{(x_1^2 + x_2^2)^2}, \frac{x_1^2(x_1^2 - x_2^2)}{(x_1^2 + x_2^2)^2} \right]^T$$



- In $\mathbb{R}^n$, $L(L_x, f(x))$ is a surface passing by x and $\nabla f(x) \perp L(L_x, f(x))$

$$f(x_1\ x_2) = \frac{x_1^2 x_2}{x_1^2 + x_2^2} \quad , \quad \nabla f(x) = \left[\frac{2x_1x_2^3}{(x_1^2+x_2^2)^2}, \frac{x_1^2(x_1^2 - x_2^2)}{(x_1^2+x_2^2)^2} \right]^T$$

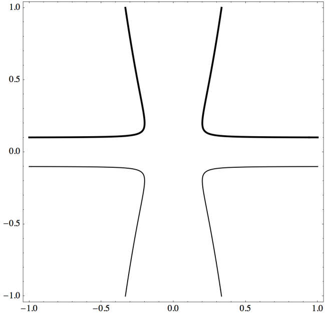


- f differentiable at $x \implies$

  $L(L_x, f(x)) \perp L(f, f(x)) \perp \nabla f(x)$

- In $\mathbb{R}^n$, $L(L_x, f(x))$ is a surface passing by x and $\nabla f(x) \perp L(L_x, f(x))$

$$f(x_1\ x_2) = \frac{x_1^2 x_2}{x_1^2 + x_2^2} \quad , \quad \nabla f(x) = \left[\frac{2x_1 x_2^3}{(x_1^2 + x_2^2)^2}, \frac{x_1^2(x_1^2 - x_2^2)}{(x_1^2 + x_2^2)^2} \right]^T$$

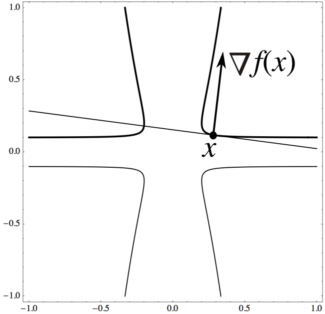


- f differentiable at $x \implies$

  $L(L_x, f(x)) \perp L(f, f(x)) \perp \nabla f(x)$

- In $\mathbb{R}^n$, $L(L_x, f(x))$ is a surface passing by x and $\nabla f(x) \perp L(L_x, f(x))$

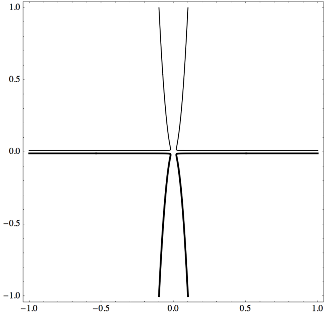
$$f(x_1\ x_2) = \frac{x_1^2 x_2}{x_1^2 + x_2^2} \quad , \quad \nabla f(x) = \left[\frac{2x_1 x_2^3}{(x_1^2 + x_2^2)^2}, \frac{x_1^2(x_1^2 - x_2^2)}{(x_1^2 + x_2^2)^2} \right]^T$$



- f differentiable at $x \implies$
  $L(L_x, f(x)) \perp L(f, f(x)) \perp \nabla f(x)$
- f differentiable at $x \implies$
  $L(f, f(x))$ "smooth"

- In $\mathbb{R}^n$, $L(L_x, f(x))$ is a surface passing by x and $\nabla f(x) \perp L(L_x, f(x))$

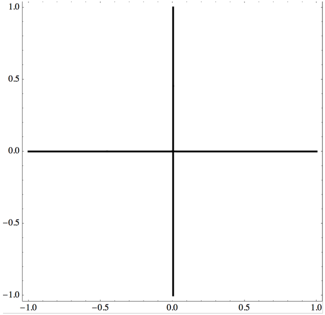
$$f(x_1\ x_2) = \frac{x_1^2 x_2}{x_1^2 + x_2^2} \quad , \quad \nabla f(x) = \left[\frac{2x_1 x_2^3}{(x_1^2 + x_2^2)^2} , \frac{x_1^2(x_1^2 - x_2^2)}{(x_1^2 + x_2^2)^2} \right]^T$$



- f differentiable at x $\Longrightarrow$
  $L(L_x, f(x)) \perp L(f, f(x)) \perp \nabla f(x)$
- f differentiable at x $\Longrightarrow$
  $L(f, f(x))$ "smooth"
- As $x \to \bar{x}$ where f non differentiable,
  $L(f, f(x))$ "less and less smooth"

- In $\mathbb{R}^n$, $L(L_x, f(x))$ is a surface passing by x and $\nabla f(x) \perp L(L_x, f(x))$

$$f(x_1\ x_2) = \frac{x_1^2 x_2}{x_1^2 + x_2^2} \quad , \quad \nabla f(x) = \left[\frac{2x_1 x_2^3}{(x_1^2+x_2^2)^2} , \frac{x_1^2(x_1^2 - x_2^2)}{(x_1^2+x_2^2)^2} \right]^T$$



- f differentiable at $x \implies$
  $L(L_x, f(x)) \perp L(f, f(x)) \perp \nabla f(x)$
- f differentiable at $x \implies$
  $L(f, f(x))$ "smooth"
- As $x \to \bar{x}$ where f non differentiable,
  $L(f, f(x))$ "less and less smooth"
- f non differentiable at $x \implies$
  $L(f, f(x))$ has "kinks"

- f differentiable $\implies$ all relevant objects in $\mathbb{R}^{n+1}$ and $\mathbb{R}^n$ are smooth
- f non differentiable $\implies$ kinks appear and things break

- Vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$, $f(x) = [\, f_1(x), f_2(x), \ldots, f_m(x) \,]$

- Partial derivative: usual stuff, except with extra index

$$\frac{\partial f_j}{\partial x_i}(x) = \lim_{t \to 0} \frac{f_j(x_1, \ldots, x_{i-1}, x_i + t, x_{i+1}, \ldots, x_n) - f_j(x)}{t}$$

- Jacobian := matrix of all $m \cdot n$ partial derivatives

$$Jf(x) := \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix} = \begin{bmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix}$$

$= m \times n$ matrix with gradients as rows

- Will come in handy later on for constrained optimization

- A special vector-valued function is particularly important already

- $\frac{\partial f}{\partial x_i} : \mathbb{R}^n \to \mathbb{R} \implies$ has partial derivatives itself
- Second order partial derivative (just do it twice)
  $$\frac{\partial^2 f}{\partial x_j \partial x_i} \qquad \frac{\partial^2 f}{\partial x_i \partial x_i} = \frac{\partial^2 f}{\partial x_i^2} = [\, f_x^i \,]''$$
- $\nabla f(\, x\,) : \mathbb{R}^n \to \mathbb{R}^n \implies$ has a Jacobian: Hessian (matrix) of f at x

$$\nabla^2 f(\, x\,) := J\nabla f(\, x\,) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\, x\,) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(\, x\,) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\, x\,) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(\, x\,) & \frac{\partial^2 f}{\partial x_2^2}(\, x\,) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_2}(\, x\,) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\, x\,) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\, x\,) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\, x\,) \end{bmatrix}$$

$O(\, n^2\,)$ to store and (at least) compute (unless sparse), bad when n large

- $\frac{\partial f}{\partial x_i} : \mathbb{R}^n \to \mathbb{R} \implies$ has partial derivatives itself
- Second order partial derivative (just do it twice) $\qquad \dfrac{\partial^2 f}{\partial x_j \partial x_i} \qquad \dfrac{\partial^2 f}{\partial x_i \partial x_i} = \dfrac{\partial^2 f}{\partial x_i^2} = [\, f_x^i \,]''$
- $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n \implies$ has a Jacobian: Hessian (matrix) of f at x

$$\nabla^2 f(x) := J\nabla f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) & \dots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}$$

$O(\, n^2 \,)$ to store and (at least) compute (unless sparse), bad when n large

- $f(x) = \langle\, b \,,\, x \,\rangle \implies \nabla^2 f(x) = 0$

- $\frac{\partial f}{\partial x_i} : \mathbb{R}^n \to \mathbb{R} \implies$ has partial derivatives itself
- Second order partial derivative (just do it twice) $\qquad \dfrac{\partial^2 f}{\partial x_j \partial x_i} \qquad \dfrac{\partial^2 f}{\partial x_i \partial x_i} = \dfrac{\partial^2 f}{\partial x_i^2} = [\, f_x^i \,]''$
- $\nabla f(\, x \,) : \mathbb{R}^n \to \mathbb{R}^n \implies$ has a Jacobian: Hessian (matrix) of f at x

$$\nabla^2 f(\, x \,) := J \nabla f(\, x \,) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\, x \,) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(\, x \,) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\, x \,) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(\, x \,) & \frac{\partial^2 f}{\partial x_2^2}(\, x \,) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_2}(\, x \,) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\, x \,) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\, x \,) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\, x \,) \end{bmatrix}$$

$O(\, n^2 \,)$ to store and (at least) compute (unless sparse), bad when n large

- $f(\, x \,) = \frac{1}{2} x^T Q x + q x \implies \nabla^2 f(\, x \,) = Q$

- $\frac{\partial f}{\partial x_i} : \mathbb{R}^n \to \mathbb{R} \implies$ has partial derivatives itself
- Second order partial derivative (just do it twice) $\dfrac{\partial^2 f}{\partial x_j \partial x_i}$ $\dfrac{\partial^2 f}{\partial x_i \partial x_i} = \dfrac{\partial^2 f}{\partial x_i^2} = [\, f_x^i \,]''$
- $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n \implies$ has a Jacobian: Hessian (matrix) of f at x

$$\nabla^2 f(x) := J\nabla f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) & \dots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}$$

$O(n^2)$ to store and (at least) compute (unless sparse), bad when n large

- $f(x) = \frac{1}{2} x^T Q x + qx \implies \nabla^2 f(x) = Q$
- Second-order model = first-order model + second-order term (= better)

$$Q_x(z) = L_x(z) + \tfrac{1}{2}(z - x)^T \nabla^2 f(x)(z - x)$$

a (non-homogeneous) quadratic function $\implies$ simple

- [5, Th. 5.3.3] $\exists\, \delta > 0$ s.t. $\forall z \in \mathcal{B}(\, x \,,\, \delta \,)$

  $\frac{\partial^2 f}{\partial x_j \partial x_i}(\, z \,)$ and $\frac{\partial^2 f}{\partial x_i \partial x_j}(\, z \,)$ exist and are continuous at x

  $\Longrightarrow\ \frac{\partial^2 f}{\partial x_j \partial x_i}(\, x \,) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\, x \,) \ \equiv\ \nabla^2 f$ symmetric

  $\Longrightarrow$ all eigenvalues of $\nabla^2 f(\, x \,)$ real

- Yet, extremely difficult to construct examples of not symmetric $\nabla^2 f$

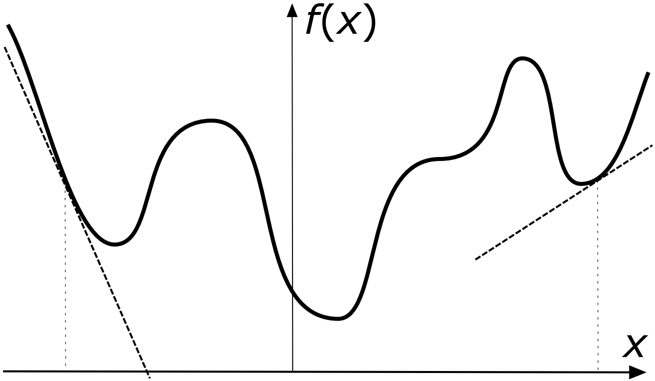
- $f \in C^2 :=$ $\nabla^2 f(\, x \,)$ continuous everywhere $\equiv\ \partial^2 f / \partial x_j \partial x_i \in C^0\ \forall\, i \,,\, j$

  $\Longrightarrow\ \nabla^2 f(\, x \,)$ symmetric everywhere and

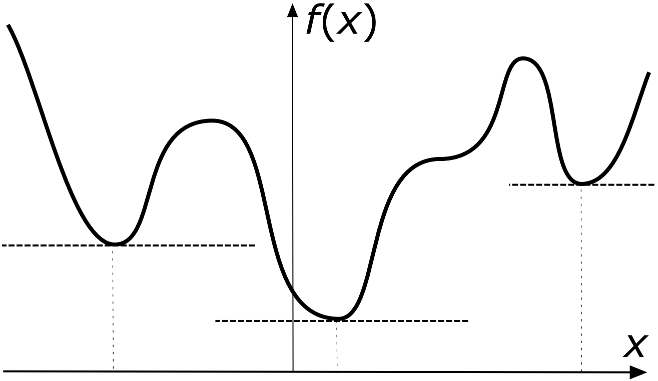
  $\nabla f(\, x \,) \in C^1 \Longrightarrow \nabla f(\, x \,) \in C^0 \Longrightarrow f(\, x \,) \in C^0$

- C^2 (strictly speaking C^3) is the best class ever for optimization, but it is sometimes necessary to make do with (much) less than that

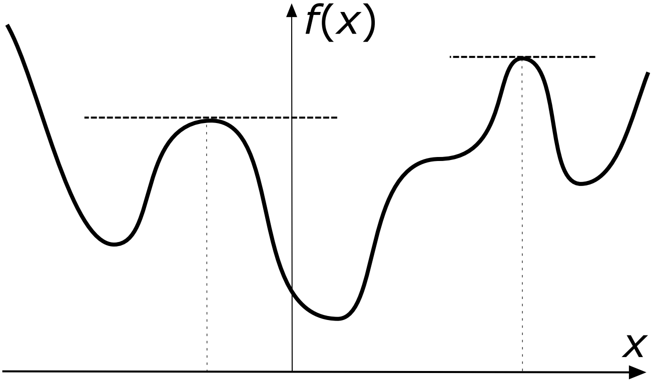
Outline



- If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum

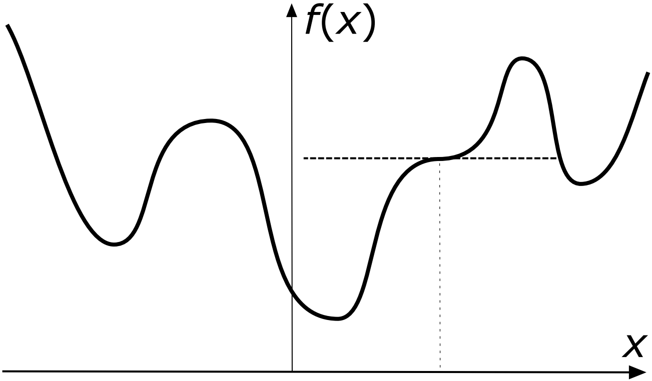


- If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)



- If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)
- However, $f'(x) = 0$ also in local (hence global) maxima

Recall: local optimality and derivatives, graphically



- If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)
- However, $f'(x) = 0$ also in local (hence global) maxima

  ... as well as in saddle points

- f differentiable at x and x local minimum $\Longrightarrow \nabla f(x) = 0$
  $\equiv$ stationary point ($\not\Leftarrow$, previous pictures for $n = 1$)
- The proof, because theorems' proofs breed algorithms
- By contradiction: x local minimum but $\nabla f(x) \neq 0$
- Prove x not local minimum not straightforward ($\nexists \equiv \forall$ /):
  $\forall \varepsilon > 0$ "small enough" $\exists z \in \mathcal{B}(x, \varepsilon)$ s.t. $f(z) < f(x)$
  $\equiv$ have to construct ∞-ly many z better then x arbitrarily close to it
- Luckily all the z can be taken along a single $d \in \mathbb{R}^n$: $z = x + \alpha d$, $\alpha > 0$
- Can choose d, use "best" one: steepest descent direction at x
  $\equiv$ d with $\| d \| = 1$ s.t. $\frac{\partial f}{\partial d}(x)$ is most negative
  $\equiv$ the (normalised) anti-gradient $-\nabla f(x)$ ($/ \| \nabla f(x) \|$)

Exercise: prove $-\nabla f(x) / \| \nabla f(x) \|$ is the steepest descent direction at x

Exercise: Why are we insisting that $\| d \| = 1$? Discuss

- Tomography $\varphi(\alpha) = \varphi_{x, -\nabla f(x)}(\alpha)$ (better not normalise d)
- Want to prove: $\exists\, \bar{\alpha} > 0$ s.t. $\varphi(\alpha) < f(x) = \varphi(0)\ \forall\, \alpha \in [0, \bar{\alpha}]$ (⚡)
- Remainder of first-order model at z: $R(z - x) = f(z) - L_x(z)$
- Definition of $f \in C^1$: $\lim_{h \to 0} R(h) / \| h \| = 0 \ \equiv\ R(h) \to 0$ "faster than $h \to 0$"
- $\varphi(\alpha) = f(x - \alpha \nabla f(x)) = f(x) + \langle -\alpha \nabla f(x), \nabla f(x) \rangle + R(-\alpha \nabla f(x))$
  $= f(x) - \alpha \| \nabla f(x) \|^2 + R(-\alpha \nabla f(x))$

  negative term linear in α + (possibly) positive "more than linear" one
- As $\alpha \to 0$ ($\Longrightarrow \| h = -\alpha \nabla f(x) \| \to 0$), it is clear who wins:

  $\lim_{\alpha \to 0} R(-\alpha \nabla f(x)) / \| \alpha \nabla f(x) \| = \lim_{h \to 0} R(h) / \| h \| = 0$

  $\equiv\ \forall \varepsilon > 0\, \exists\, \bar{\alpha} > 0$ s.t. $R(-\alpha \nabla f(x)) / \alpha \| \nabla f(x) \| \leq \varepsilon\ \ \forall\, \alpha \in [0, \bar{\alpha}]$
- Take $\varepsilon < \| \nabla f(x) \|$ to get $R(-\alpha \nabla f(x)) < \alpha \| \nabla f(x) \|^2 \ \Longrightarrow$

  $\varphi(\alpha) = f(x) - \alpha \| \nabla f(x) \|^2 + R(-\alpha \nabla f(x)) < f(x)$
- Proof shows: a small enough step along $-\nabla f(x)\ (\neq 0)$ yields a better z

- Stationary point $\not\Longrightarrow$ local minimum: how to tell them apart?
- First-order model can't, it is "flat": need to look at curvature of f
- If f were quadratic I would know: look at eigenvalues of $Q = \nabla^2 f(x)$
- Obvious idea: approximate f with a quadratic function = second-order model = $Q_x(z) = L_x(z) + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x)$
- $\nabla Q_x(x) = \nabla L_x(x) = \nabla f(x) \Longrightarrow \nabla Q_x(x) = 0$ (**check**)
- Hence, $\nabla^2 f(x) \succeq 0 \iff x$ (global) minimum of Q_x
- Can prove it (almost) holds for f, too:

  $f \in C^2$: x local minimum $\Longrightarrow \nabla^2 f(x) \succeq 0$
- Requires second-order Taylor's theorem [5, Th. 5.4.9]:

  $$f(z) = L_x(z) + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x) + R(z - x)$$

  with $\lim_{h \to 0} R(h) / \| h \|^2 = 0 \equiv R(h) \to 0$ faster than "$h^2 \to 0$"

  $\equiv$ the remainder vanishes "faster than quadratically"

- By contradiction: $f \in C^2$, x local minimum but $\nabla^2 f(x) \not\succeq 0 \;\equiv$
  $\exists\, d$ s.t. $d^T \nabla^2 f(x) d < 0$ (w.l.o.g. $\|\, d \,\| = 1$)
- d = direction of negative curvature, $\varphi(\alpha) = \varphi_{x,d}(\alpha)$
- Second-order Taylor + $\nabla f(x) = 0 \;\equiv\; L_x(z) = f(x) \implies$
  $$\varphi(\alpha) = f(x) + \tfrac{1}{2}\alpha^2 d^T \nabla^2 f(x) d + R(\alpha d)$$
  negative quadratic term in α + (possibly) positive "more than quadratic" one
- As α ($= \|\, h = \alpha d \,\|$ since $\|\, d \,\| = 1$) $\to 0$, it is clear who wins:
  $\lim_{\alpha \to 0} R(\alpha d) / \alpha^2 = \lim_{h \to 0} R(h) / \|\, h \,\|^2 = 0 \;\equiv$
  $\forall \varepsilon > 0 \,\exists\, \bar{\alpha} > 0$ s.t. $R(\alpha d) \leq \varepsilon \alpha^2 \;\; \forall\, \alpha \in [\, 0 \,,\, \bar{\alpha} \,]$
- Take ($0 <$) $\varepsilon < -\frac{1}{2} d^T \nabla^2 f(x) d$ to get $R(\alpha d) < -\frac{1}{2}\alpha^2 d^T \nabla^2 f(x) d$
  $\implies \varphi(\alpha) = f(x) + \frac{1}{2}\alpha^2 d^T \nabla^2 f(x) d + R(\alpha d) < f(x) \quad \forall\, \alpha \in [\, 0 \,,\, \bar{\alpha} \,]$ ↯
- In a local minimum, there cannot be directions of negative curvature:
  "when the first derivative is 0, second-order effects prevail"

- Necessary condition almost also sufficient: $f \in C^2$,

  $$\nabla f(x) = 0 \text{ and } \nabla^2 f(x) \succ 0 \implies x \text{ local minimum}$$

- Avoids "bad case" $d^T \nabla^2 f(x) d = 0 \equiv$ zero-curvature direction
  $\equiv x$ saddle point $\approx f''(x) = 0$: would need even higher-order derivatives

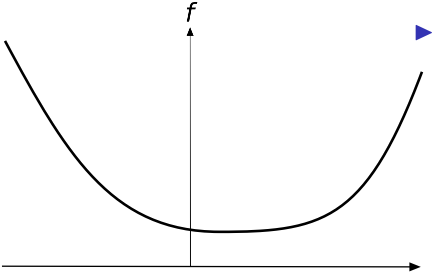
- Proof: second-order Taylor $f(x+d) = f(x) + \frac{1}{2} d^T \nabla^2 f(x) d + R(d)$ with
  $\lim_{d \to 0} R(d) / \| d \|^2 = 0 \equiv \forall \varepsilon > 0 \, \exists \, \delta > 0$ s.t. $R(d) / \| d \|^2 \geq -\varepsilon$
  $\equiv R(d) \geq -\varepsilon \| d \|^2 \;\; \forall \, d$ s.t. $\| d \| < \delta$

  $\lambda_n > 0$ min eigenvalue of $\nabla^2 f(x) \implies d^T \nabla^2 f(x) d \geq \lambda_n \| d \|^2$

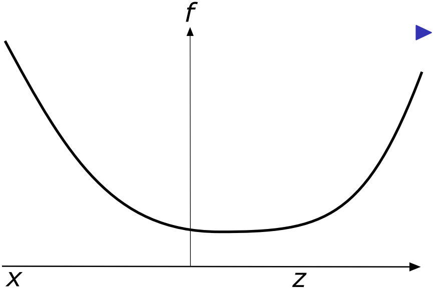
  Take $\varepsilon < \lambda_n / 2$: then, $\forall \, d$ s.t. $\| d \| < \delta$
  $f(x+d) = f(x) + \frac{1}{2} d^T \nabla^2 f(x) d + R(d) \geq f(x) + \frac{\lambda_n - \varepsilon}{2} \| d \|^2$

- It proves more than we asked: f grows "at least quadratically around x"
  $\exists \, \delta > 0$ and $\gamma > 0$ s.t. $f(z) \geq f(x) + \gamma \| z - x \|^2 \;\; \forall \, z \in \mathcal{B}(x, \delta)$
  $\equiv$ strong (local) optimality

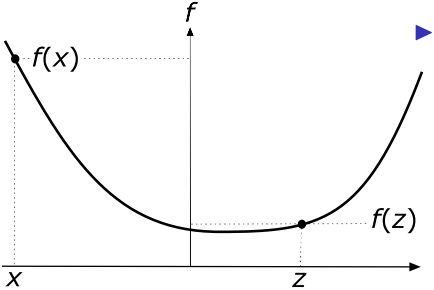
Outline



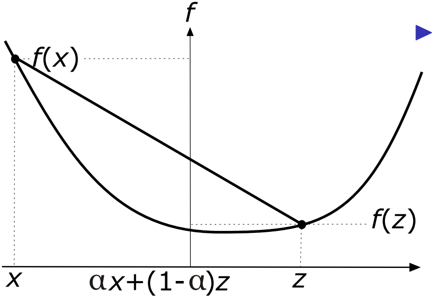
- f convex $\equiv$



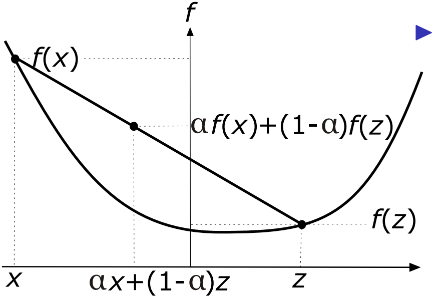
- f convex $\equiv \; \forall\, x\,,\, z \in \mathbb{R}^n$,



- f convex $\equiv \forall x, z \in \mathbb{R}^n$,

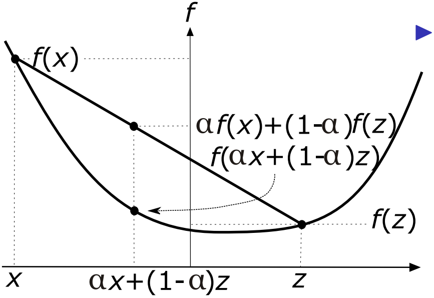


- f convex $\equiv \ \forall x, z \in \mathbb{R}^n \ , \ \alpha \in [0,1]$



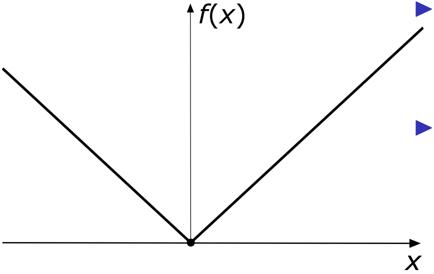
- f convex $\equiv \forall x, z \in \mathbb{R}^n$, $\alpha \in [0,1]$

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

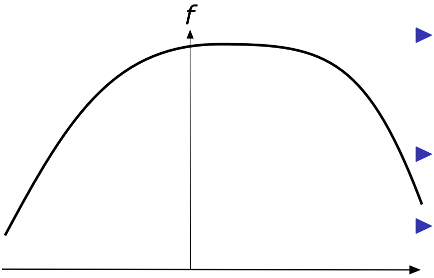


- f convex $\equiv \ \forall x, z \in \mathbb{R}^n \ , \ \alpha \in [0,1]$

  $\alpha f(x) + (1-\alpha) f(z) \geq f(\alpha x + (1-\alpha) z)$



- f convex $\equiv \forall x, z \in \mathbb{R}^n$, $\alpha \in [0,1]$
  $\alpha f(x) + (1-\alpha) f(z) \geq f(\alpha x + (1-\alpha) z)$
- Convex $\not\Longrightarrow C^1$ (ex. $\|\cdot\|_1$)



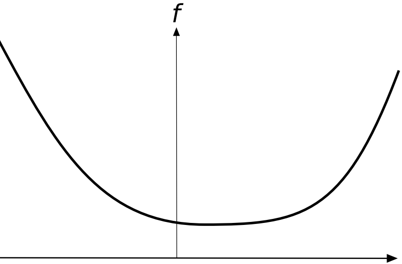
- f convex $\equiv\ \forall\, x\,,\, z \in \mathbb{R}^n\ ,\ \alpha \in [0,1]$

  $\alpha f(\,x\,) + (1-\alpha) f(\,z\,) \geq f(\,\alpha x + (1-\alpha) z\,)$
- Convex $\not\Rightarrow C^1$ (ex. $\|\cdot\|_1$)
- f concave $\equiv -f$ convex

- $\max\{\, f(\,x\,)\ :\ x \in \mathbb{R}^n\,\} = +\infty$ (unless $f(\,x\,) = c$); sounds familiar?
- In fact, f quadratic convex $\equiv\ Q \succeq 0$
- Exactly the opposite for f concave ($Q \preceq 0$): as a great man said, "(convex) optimization is a one-sided world"
- Only f both convex and concave: linear
- How do you tell if a function is convex?

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

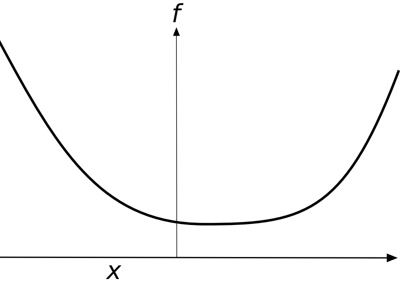
Exercise: Justify why that property is called "monotone"



- $f \in C^1$ convex $\iff$

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

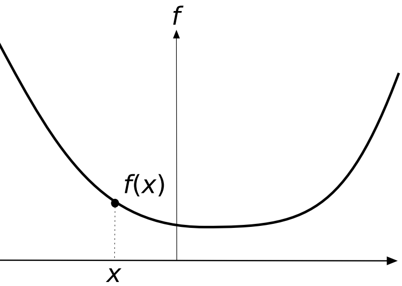
Exercise: Justify why that property is called "monotone"



- $f \in C^1$ convex $\iff$

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

Exercise: Justify why that property is called "monotone"

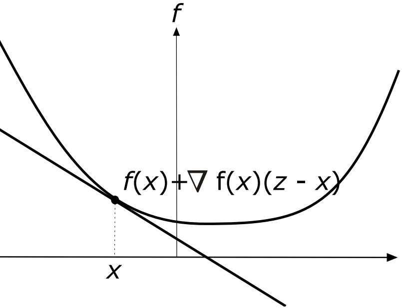


- $f \in C^1$ convex $\iff$

  $L_x(z) = f(x)$

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

Exercise: Justify why that property is called "monotone"

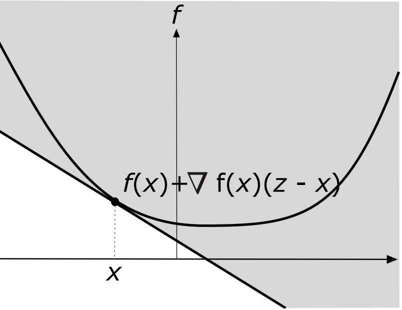


- $f \in C^1$ convex $\iff$
  $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$

Exercise: prove $\Longrightarrow$ "by prime principles"

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

Exercise: Justify why that property is called "monotone"



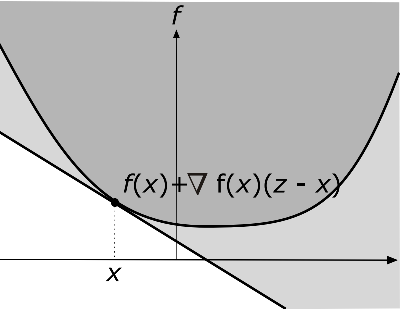
- $f \in C^1$ convex $\iff$
  $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$

Exercise: prove $\Longrightarrow$ "by prime principles"

- Geometrically: the epigraph is an half-space

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

Exercise: Justify why that property is called "monotone"



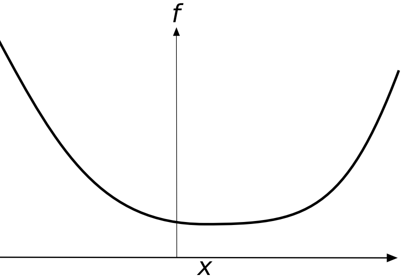
- $f \in C^1$ convex $\iff$
  $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$

Exercise: prove $\implies$ "by prime principles"

- Geometrically: the epigraph is an half-space that contains that of f ($epi(L_x) \supseteq epi(f)$)

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

Exercise: Justify why that property is called "monotone"



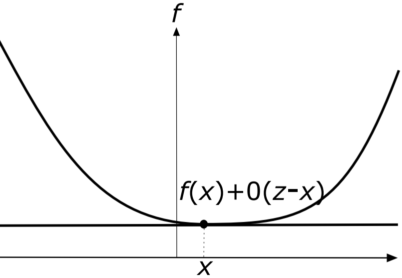
- $f \in C^1$ convex $\iff$
  $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$

Exercise: prove $\Longrightarrow$ "by prime principles"

- Geometrically: the epigraph is an half-space that contains that of f ($epi(L_x) \supseteq epi(f)$)
- $\nabla f(x) = 0$

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

Exercise: Justify why that property is called "monotone"



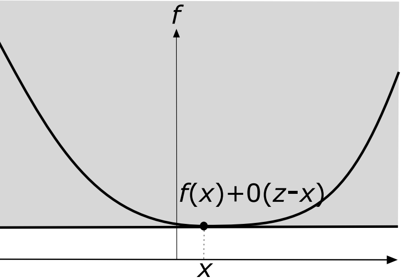
- $f \in C^1$ convex $\iff$ $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$

Exercise: prove $\implies$ "by prime principles"

- Geometrically: the epigraph is an half-space that contains that of f ($epi(L_x) \supseteq epi(f)$)
- $\nabla f(x) = 0 \implies$

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

Exercise: Justify why that property is called "monotone"



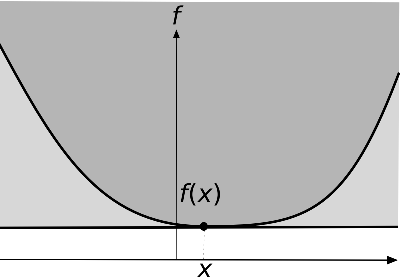
- $f \in C^1$ convex $\iff$
  $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$

Exercise: prove $\Longrightarrow$ "by prime principles"

- Geometrically: the epigraph is an half-space that contains that of f ($epi(L_x) \supseteq epi(f)$)
- $\nabla f(x) = 0 \implies$

- $f \in C^1$ convex $\iff$ ∇f monotone: $\langle \nabla f(z) - \nabla f(x), z - x \rangle \geq 0 \;\; \forall x, z$

Exercise: Justify why that property is called "monotone"



- $f \in C^1$ convex $\iff$
  $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$

Exercise: prove $\Longrightarrow$ "by prime principles"

- Geometrically: the epigraph is an half-space that contains that of f ($epi(L_x) \supseteq epi(f)$)
- $\nabla f(x) = 0 \implies f(z) \geq f(x) \;\; \forall z \in \mathbb{R}^n$
- $f \in C^1$ convex: $\nabla f(x) = 0 \iff x$ global minimum
- $f \in C^2$: f convex $\equiv$ $\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathbb{R}^n$
- $f \in C^2$ with $\nabla^2 f \succeq \tau I$ with $\tau > 0$ the best case for optimization
- Sometimes the best way to prove f convex, unless it is by construction

- Some functions are (more or less obviously) convex:
  1. $f(x) = bx + c$ (affine) $\Longleftrightarrow$ both convex and concave (**check**) [nontrivial]
  2. $f(x) = \frac{1}{2}x^T Qx + qx$ (quadratic) convex $\Longleftrightarrow$ $Q \succeq 0$
  3. $f(x) = e^{ax}$ for any $a \in \mathbb{R}$
  4. restricted to $x \geq 0$, $f(x) = -\ln(x)$
  5. restricted to $x \geq 0$, $f(x) = x^a$ for $a \geq 1$ or $a \leq 0$
  6. $f(x) = \|x\|_p$ for $p \geq 1$
  7. $f(x) = \max\{x_1, \ldots, x_n\}$
  8. $Q \in \mathbb{R}^{n \times n}$ symmetric, eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$:
     $f_m(Q) = \sum_{i=1}^{m} \lambda_i$ (sum of m largest eigenvalues)

Exercise: Prove 3., 4. and 5.; for the latter, which a make x^a convex on all $\mathbb{R}$?

Exercise: is $f(x) = \min\{x_1, \ldots, x_n\}$ convex?

Mathematically speaking: Convexity-preserving operations [2, § 3.2]

1. f, g convex, δ, $\beta \in \mathbb{R}_+ \implies \delta f + \beta g$ convex (non-negative combination)

2. $\{ f_i \}_{i \in I}$ (∞-ly many) convex functions $\implies f(x) = \sup_{i \in I} \{ f_i(x) \}$ convex

3. f convex $\implies f(Ax + b)$ convex (pre-composition with linear mapping)

4. $f : \mathbb{R}^n \to \mathbb{R}$ convex, $g : \mathbb{R} \to \mathbb{R}$ convex increasing $\implies g(f(x))$ convex (post-composition with increasing convex function)

5. f_1, f_2 convex $\implies f(x) = \inf\{ f_1(x_1) + f_2(x_2) \,:\, x_1 + x_2 = x \}$ convex (infimal convolution)

6. g convex $\implies f(x) = \inf\{ g(z) \,:\, Az = x \}$ convex (value function of convex constrained problem)

7. $g(x, z) : \mathbb{R}^{n+m} \to \mathbb{R}$ convex $\implies f(x) = \inf\{ g(x, z) \,:\, z \in \mathbb{R}^m \}$ convex (partial minimization)

8. $f(x)$ convex $\implies p(x, u) = uf(x / u)$ convex on $u > 0$ (perspective or dilation function of f)

Exercise: Prove 1. "from prime principles" (at least 2., 3. analogous)

- $n = 1$: f unimodal $\iff$ quasiconvex [1, Ex. 3.57] $\equiv$
  $\alpha f(x) + (1 - \alpha) f(z) \leq \max\{ f(x), f(z) \}$ (??)
- f quasiconvex $\iff \forall$ nonempty sublevel set $S(f, l) = \{ x : f(x) \leq l \}$ is a (possibly, infinite) interval (in fact a convex set, will see) [1, Th. 3.5.2]

Exercise: Prove: f convex $\implies f$ quasiconvex, $\impliedby$ not true

- Issue: algebra of quasiconvex (not convex) functions "weaker"
- f quasiconvex, $\delta \in \mathbb{R}_+ \implies \delta f$ quasiconvex true
- But f, g quasiconvex $\implies f + g$ quasiconvex false

Exercise: Prove the two statements above

- No (or much weaker) Disciplined QuasiConvex Programming [7], f "naturally" quasiconvex unlikely
- Does not mean impossible, you may be lucky, in fact NN often $\approx$ quasiconvex

Outline

- Multivariate global optimality very hard (exponential in theory & practice)
- Multivariate local optimality "easy" with the right (first-order) information: $f \in C^1$ (but one often has to make do with less, will see)
- Local optimization ≈ nonlinear system $\nabla f(x) = 0$, surely nontrivial
- "f simple" (quadratic) $\Longrightarrow$ "$\nabla f(x) = 0$ simple" (linear system): quadratic models are going to be useful
- However, stationary points not always local minima (may be maxima)

- Multivariate global optimality very hard (exponential in theory & practice)
- Multivariate local optimality "easy" with the right (first-order) information: $f \in C^1$ (but one often has to make do with less, will see)
- Local optimization $\approx$ nonlinear system $\nabla f(x) = 0$, surely nontrivial
- "f simple" (quadratic) $\Longrightarrow$ "$\nabla f(x) = 0$ simple" (linear system): quadratic models are going to be useful
- However, stationary points not always local minima (may be maxima)
- Only theoretically safe case: f convex $\Longrightarrow$

  every stationary point is local $\equiv$ global minimum
- Always keep it convex if possible, better if C^1, better still if C^2

- Multivariate global optimality very hard (exponential in theory & practice)
- Multivariate local optimality "easy" with the right (first-order) information: $f \in C^1$ (but one often has to make do with less, will see)
- Local optimization $\approx$ nonlinear system $\nabla f(x) = 0$, surely nontrivial
- "f simple" (quadratic) $\Longrightarrow$ "$\nabla f(x) = 0$ simple" (linear system): quadratic models are going to be useful
- However, stationary points not always local minima (may be maxima)
- Only theoretically safe case: f convex $\Longrightarrow$

  every stationary point is local $\equiv$ global minimum
- Always keep it convex if possible, better if C^1, better still if C^2
- For learning, local optimality is typically enough (f "not adversarial")

- Multivariate global optimality very hard (exponential in theory & practice)
- Multivariate local optimality "easy" with the right (first-order) information: $f \in C^1$ (but one often has to make do with less, will see)
- Local optimization $\approx$ nonlinear system $\nabla f(x) = 0$, surely nontrivial
- "f simple" (quadratic) $\implies$ "$\nabla f(x) = 0$ simple" (linear system): quadratic models are going to be useful
- However, stationary points not always local minima (may be maxima)
- Only theoretically safe case: f convex $\implies$
  every stationary point is local $\equiv$ global minimum
- Always keep it convex if possible, better if C^1, better still if C^2
- For learning, local optimality is typically enough (f "not adversarial")
- Time to move to multivariate algorithms

[1] M.S. Bazaraa, H.D. Sherali, C.M. Shetty *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, 2006

[2] S. Boyd, L. Vandenberghe *Convex Optimization*, https://web.stanford.edu/~boyd/cvxbook Cambridge University Press, 2008

[3] P. Hansen, B. Jaumard "Lipschitz Optimization" in *Handbook of Global Optimization – Nonconvex optimization and its applications*, R. Horst and P.M. Pardalos (Eds.), Chapter 8, 407–494, Springer, 1995

[4] J. Nocedal, S.J. Wright, *Numerical Optimization – second edition*, Springer Series in Operations Research and Financial Engineering, 2006

[5] W.F. Trench, *Introduction to Real Analysis* https://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_REAL_ANALYSIS.PDF Free Hyperlinked Edition 2.04, December 2013

[6] AutoDiff Org: https://www.autodiff.org

[7] CVX: https://cvxr.com

[8] DFL: https://www.iasi.cnr.it/~liuzzi/DFL

[9] Wikipedia – Norm https://en.wikipedia.org/wiki/Norm_(mathematics)

Outline

- For $y = 1/k \to 0$, $f(d_1 y, d_2 y) = [d_1^2 d_2 y^3 / ((d_1 y)^4 + (d_2 y)^2)]^2 \to 0$ (the degree of the numerator is $>$ of the min degree at the denominator, i.e., the numerator goes to 0 faster than the denominator) however chosen d_1 and d_2. In the second case $f(y, y^2) = [y^4 / (y^4 + y^4)]^2 = 1/4$ **[back]**

- $\frac{\partial f}{\partial \beta d}(x) = \lim_{t \to 0}(f(x + t(\beta d)) - f(x)) / t =$
$= \lim_{t \to 0} \beta(f(x + (t\beta)d)) - f(x)) / (\beta t)$. $p = \beta t$, $t \to 0 \implies p \to 0$
$\implies \frac{\partial f}{\partial \beta d}(x) = \lim_{p \to 0} \beta(f(x + pd) - f(x)) / p = \beta \frac{\partial f}{\partial d}(x)$ **[back]**

- In all points $[0, x_2]$: for $d = [1, 0]$, $\varphi[0, x_2], d(\alpha) = |\alpha| + |x_2|$ is nondifferentiable in 0, i.e., $\partial f / \partial d \nexists$; analogous for $[x_1, 0]$ **[back]**

- Fix any $[\, d_1\,,\, d_2\,]$: $\lim_{t\to 0} f(\, td_1\,,\, td_2\,) = \lim_{t\to 0} \frac{t^3 d_1^2 d_2}{t^2(d_1^2+d_2^2)} = 0$. For the general result we use the definition of limit: for any $\varepsilon > 0$ we find $\delta > 0$ s.t. $\|\,[\, x_1\,,\, x_2\,]\,\| \leq \delta \implies |\, f(\, x_1\,,\, x_2\,)\,| \leq \varepsilon$. $\|\,[\, x_1\,,\, x_2\,]\,\| = \sqrt{x_1^2 + x_2^2} \leq \delta$ implies $|\, x_2\,| \leq \delta$. Hence,
$$|\, f(\, x_1\,,\, x_2\,) - 0\,| \leq |\, x_2\,| \left(\frac{x_1^2}{x_1^2 + x_2^2} \right) \leq |\, x_2\,| \leq \delta$$
whenever $\|\,[\, x_1\,,\, x_2\,]\,\| \leq \delta$; thus, taking $\delta = \varepsilon$ works, proving that the limit is indeed 0 however chosen the converging sequence. [**back**]

- $\dfrac{\partial f}{\partial [\, d_1\,,\, d_2\,]}(\,0\,,\,0\,) = \lim\limits_{t\to 0} \dfrac{f(\, td_1\,,\, td_2\,) - f(0\,,\,0)}{t} = \lim\limits_{t\to 0} \dfrac{t^3 d_1^2 d_2}{t^3(d_1^2 + d_2^2)} =$
$= f(\, d_1\,,\, d_2\,)$, clearly not a linear function [**back**]

- $\nabla f(\, x_1\,,\, x_2\,) = \left[\, \dfrac{\partial f}{\partial x_1}\,,\, \dfrac{\partial f}{\partial x_2}\,\right] = \left[\, \dfrac{2x_1 x_2^3}{(\, x_1^2 + x_2^2\,)^2}\,,\, \dfrac{x_1^2(\, x_1^2 - x_2^2\,)}{(\, x_1^2 + x_2^2\,)^2}\,\right]$; for
$g(\, x_1\,,\, x_2\,) = \partial f \,/\, \partial x_2$, it is easy to check that $g(\,\alpha\,,\,0\,) = 1$ while $g(\,0\,,\,\alpha\,) = 0$, i.e., the limit along the directions $[\,1\,,\,0\,]$ and $[\,0\,,\,1\,]$ is different [**back**]

- Strictly speaking, defining $\frac{\partial f}{\partial d}(0\,,\,0)$ requires $f(0\,,\,0)$, which is undefined. However, we can take any generic direction $d = [\,d_1\,,\,d_2\,] \neq 0$ and prove that $\lim_{\alpha \to 0} f(\alpha d) = d_1^4 d_2^4 \alpha^4 \,/\, (\,d_2^2 + d_1^4 \alpha^2\,)^2 = 0$ however chosen d. In fact, if either $d_2 = 0$ or $d_1 = 0$ the numerator is always 0 while the denominator is not (they cannot be both 0). If they are both nonzero, the numerator goes to 0 while the denominator goes to $d_2^4 > 0$. Thus, only looking along lines it would be safe to define $f(0\,,\,0) = 0$ by continuity, and therefore to have $\frac{\partial f}{\partial d}(0\,,\,0) = 0$ for all $d \neq 0$, which gives $\frac{\partial f}{\partial d}(0\,,\,0) = \langle\,[\,0\,,\,0\,]\,,\,d\,\rangle$ **[back]**

- We know that $\frac{\partial f}{\partial d}(x) = \langle\,\nabla f(x)\,,\,d\,\rangle = \|\,\nabla f(x)\,\|\|\,d\,\|\cos(\theta) =$ $= \|\,\nabla f(x)\,\|\cos(\theta)$ (as $\|\,d\,\| = 1$). Clearly, this number is minimum when $\cos(\theta)$ is, i.e., $\theta = \pi \;\equiv\; \cos(\theta) = -1$. This corresponds to d being collinear to $\nabla f(x)$ with opposite direction, i.e., $d = -\nabla f(x) \,/\, \|\,\nabla f(x)\,\|$ **[back]**

- Because $\frac{\partial f}{\partial \beta d} = \beta \frac{\partial f}{\partial d}$, hence $\|\,d\,\| \to \infty \implies \frac{\partial f}{\partial d} \to -\infty$ (with right d) **[back]**

- $Q_x(z) = f(x) + \langle \nabla f(x), z - x \rangle + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x) \implies$ $\nabla Q_x(z) = \nabla f(x) + \nabla^2 f(x)(z - x)$, thus evaluated at $z = x$ gives $\nabla f(x)$. The derivation handily reveals that $\nabla Q_x(z)$ is a linear (vector) function of z that coincides with $\nabla f(x)$ at $z = x$, i.e., it is the first-order model of ∇f at x (in fact it uses the "gradient of the gradient", that is, the Hessian) [**back**]

- In the univariate case the condition is $(f'(z) - f'(x))(z - x) \geq 0$, i.e., "$f'(z) - f'(x)$ and $z - x$ have the same sign". In other words, $z \geq x \implies f'(z) \geq f'(x)$ and $z \leq x \implies f'(z) \leq f'(x)$, i.e., f' is monotone nonincreasing [**back**]

- $\forall \alpha \in [0, 1]\ \alpha f(z) + (1 - \alpha)f(x) \geq f(\alpha z + (1 - \alpha)x) \implies$
$\alpha(f(z) - f(x)) + f(x) \geq f(\alpha(z - x) + x) \implies$
$f(z) - f(x) \geq [f(\alpha(z - x) + x) - f(x)] / \alpha$
send $\alpha \to 0$ to get $\frac{\partial f}{\partial(z-x)}(x) = \langle \nabla f(x), z - x \rangle$ [**back**]

- This is surprisingly nontrivial. We want to prove: f both concave and concave (BCC) $\iff f(x) = \langle b, x \rangle + c$ for some $b \in \mathbb{R}^n$, $c \in \mathbb{R}$.
  BCC $\equiv f((1-\alpha)x + \alpha z)$ [both $\geq$ and $\leq \implies$] $= (1-\alpha)f(x) + \alpha f(z)$
  $f(x) = \langle b, x \rangle + c \implies f((1-\alpha)x + \alpha z) = \langle b, (1-\alpha)x + \alpha z \rangle + c =$
  $(1-\alpha)\langle b, x \rangle + \alpha\langle b, z \rangle + [(1-\alpha)c + \alpha c] =$
  $(1-\alpha)(\langle b, x \rangle + c) + \alpha(\langle b, x \rangle + c) = (1-\alpha)f(x) + \alpha f(z)$; note how this crucially depends on $(1-\alpha) + \alpha = 1$, it would not be true for generic $\gamma x + \delta z$
  For $\impliedby$, define $g(x) = f(x) - f(0)$ so that $g(0) = 0$. Since f is BCC, then also g is (trivial, or see point 1. in next slide). Hence
  $0 = g(0) = g((1-(1/2))x + (1/2)(-x)) =$
  $\quad = (1-(1/2))g(x) + (1/2)g(-x) \implies g(-x) = -g(x)$ (antisymmetric)
  We now prove: i. $g(\gamma x) = \gamma g(x)$, ii. $g(x+z) = g(x) + g(z)$
  For i., $0 \leq \gamma \leq 1 \implies g(\gamma x) = g(\gamma x + (1-\gamma)0) =$
  $= \gamma g(x) + (1-\gamma)g(0) = \gamma g(x)$. If $\gamma > 1$, then $g(x) = g((1/\gamma)\gamma x) =$
  $= g((1/\gamma)\gamma x + (1 - 1/\gamma)0) = (1/\gamma)g(\gamma x) + (1 - 1/\gamma)g(0) =$
  $= (1/\gamma)g(\gamma x)$; multiply both sides by γ to get $\gamma g(x) = g(\gamma x)$. Finally, if $\gamma < 0$ then $g(\gamma x) = g((-\gamma)(-x)) = (-\gamma)g((-x))$ (using the previous results with $-\gamma > 0$) $= (-\gamma)(-g(x))$ (using $g(-x) = -g(x)$) $= \gamma g(x)$

For ii., $g(x+z) = g((1/2)2x + (1/2)2z) = (1/2)g(2x) + (1/2)(2z) =$
$= (1/2)2g(x) + (1/2)2(z) = g(x) + g(z)$ (using i. with $\gamma = 2$)
i. and ii. are the alternative definition of linear function, hence $\exists\, b \in \mathbb{R}^n$ s.t. $g(x) = \langle b, x \rangle$; thus, $f(x) = g(x) + f(0)$ is affine with $c = f(0)$, as desired [**back**]

- $[e^{a\cdot}]'(x) = ae^{ax}$, which is positive increasing if $a > 0$, negative increasing if $a < 0$. $[-\ln(\cdot)]'(x) = -1/x$, which is negative increasing. $[\cdot^a]'(x) = ax^{a-1}$; for $a < 0$ this is negative increasing, for $a \geq 1$ this is positive increasing. Only positive even integer a make x^a convex on all $\mathbb{R}$, since then ax^{a-1} is positive increasing (as the second derivative, $a(a-1)x^{a-2}$, is always positive). [**back**]

- No: consider $f(x_1, x_2) = \min\{x_1, x_2\}$ on the line $x_1 + x_2 = 0 \equiv x_2 = -x_1$, i.e., $\min\{x_1, -x_1\} = -|x_1|$ which is concave (and not linear, hence it cannot be convex) [**back**]

- $\alpha f(x) + (1-\alpha)f(z) \geq f(\alpha x + (1-\alpha)z) \implies$
  $\delta[\alpha f(x) + (1-\alpha)f(z)] \geq \delta f(\alpha x + (1-\alpha)z)$.
  $\alpha g(x) + (1-\alpha)g(z) \geq g(\alpha x + (1-\alpha)z) \implies$
  $\beta[\alpha g(x) + (1-\alpha)g(z)] \geq \beta g(\alpha x + (1-\alpha)z)$.
  Hence, $\delta[\alpha f(x) + (1-\alpha)f(z)] + \beta[\alpha g(x) + (1-\alpha)g(z)] =$
  $= \alpha(\delta f(x) + \beta g(x)) + (1-\alpha)(\delta f(z) + \beta g(z)) \geq$
  $\delta f(\alpha x + (1-\alpha)z) + \beta g(\alpha x + (1-\alpha)z)$ **[back]**

- Take x s.t. $f(x) \leq l$, z s.t. $f(z) \leq l$, and any $\alpha \in [0, 1]$: then, by convexity $f(\alpha x + (1-\alpha)z) \leq \alpha f(x) + (1-\alpha)f(z) \leq \alpha l + (1-\alpha)l = l$, i.e., $\alpha x + (1-\alpha)z \in S(f, l) \implies S(f, l)$ is a (possibly, infinite) interval (in general a convex set)
  On the other hand, consider the "downward spike function centered at c", i.e., $s_c(x) = \min\{|x-c|, 1\}$. Clearly, s_c is quasiconvex: in fact, $S(f, l) = \emptyset$ if $l < 0$, $S(f, l) = [c-l, c+l]$ if $0 \leq l < 1$, and $S(f, l) = \mathbb{R}$ if $l \geq 1$.
  However, s_0 is not convex: in fact,
  $(1/2)s_0(0) + (1/2)s_0(2) = 1/2 < 1 = s_0((1/2)0 + (1/2)2) = s_0(1)$ **[back]**

- $S(\delta f, l) = \{x : \delta f(x) \leq l\} = \{x : \delta f(x) \leq l/\delta\} = S(f, l/\delta)$: since the latter is an interval (convex set), the former also is
  To prove $\not\Leftarrow$ consider $f(x) = s_{-1}(x) + s_1(x)$ (cf. previous exercise). Clearly, $f(-1) = f(1) = 0$ but $f(x) > 0$ for all other values of x, i.e., $S(f, 0) = \{-1, 1\}$ is not an interval [**back**]