Background

Antonio Frangioni

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

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

Outline

- X any set, $f : X \to \mathbb{R}$ any function: optimization problem

$$(P) \qquad f_* = \min\{\, f(\, x\,) \;:\; x \in X \,\}$$

- X feasible region, f objective function, $\nu(P) = f_*$ optimal value

- "min" w.l.o.g.: $\min\{\, f(\, x\,) \;:\; x \in X \,\} = -\max\{\, -f(\, x\,) \;:\; x \in X \,\}$, (but $\min\{\, f(\, x\,) \,\} \neq \max\{\, f(\, x\,) \,\}$, often rather different problems)

- $x \in X$ feasible solution; often $X \subset F$, $x \in F \setminus X$ unfeasible solution

- $f_* \leq f(\, x\,)\, \forall x \in X$, $\forall v > f_* \,\exists x \in X$ s.t. $f(\, x\,) < v$

- We want any optimal solution: $x_* \in X$ such that $f(\, x_*\,) = f_*$

- Impossible (X inaccessible cardinal, f non computable function, $\ldots$)

- Even with very simple f / X, x_* may just not exist

- “Bad case” I: $X = \emptyset$ (“empty”)

  1. $\min\{\, x \,:\, x \in \mathbb{R} \wedge x \leq -1 \wedge x \geq 1 \,\}$

  There just is no solution (which may be important to know)

- “Bad case” II: $\forall M \, \exists x_M \in X$ s.t. $f(x_M) \leq M$ (“unbounded [below]”)

  2. $\min\{\, x \,:\, x \in \mathbb{R} \wedge x \leq 0 \,\}$

  There are solutions as good as you like (which may be important to know)

- Not really bad cases, just things that can happen

- Solving an optimization problem actually three different things:
  - Finding x_* and proving it is optimal (how??)
  - Proving $X = \emptyset$ (how??)
  - Constructively proving f unbounded below on X (how??)

- Things can be worse: not empty, not unbounded, but no x_* either:

  3. $\min\{\, x \;:\; x \in \mathbb{R} \wedge x > 0 \,\}$ (“bad” X)
  4. $\min\{\, 1 / x \;:\; x \in \mathbb{R} \wedge x > 0 \,\}$ (“bad” f and X)
  5. $\min \left\{ f(x) = \begin{cases} x & \text{if } x > 0 \\ 1 & \text{if } x = 0 \end{cases} \;:\; x \in [\,0\,,\,1\,] \right\}$ (“bad” f)

- Assumptions needed on f and X to ensure “things work”

- Something of an hair-splitting exercise: typically “$x \in \mathbb{R}$” actually mean “$x \in \mathbb{Q}$” with up to k digits precision

- Many (but not all) problems go away if goal is “just” to find approximately optimal $\bar{x}$ and prove it (how??)

  $f(\bar{x}) - f_* \leq \varepsilon$ (absolute) or $(\, f(\bar{x}) - f_* \,) / |\, f_* \,| \leq \varepsilon$ (relative) error

  and some ε is required anyway in most cases

- Already "$f : X \to \mathbb{R}$" a rather strong assumption:
  can "condense the value of x with a single number"

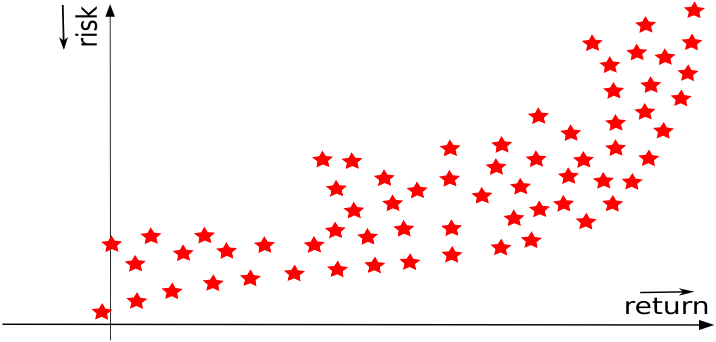
- Often you need more than one, say

$$(P) \qquad \min \{ [f_1(x) , f_2(x)] \; : \; x \in X \}$$

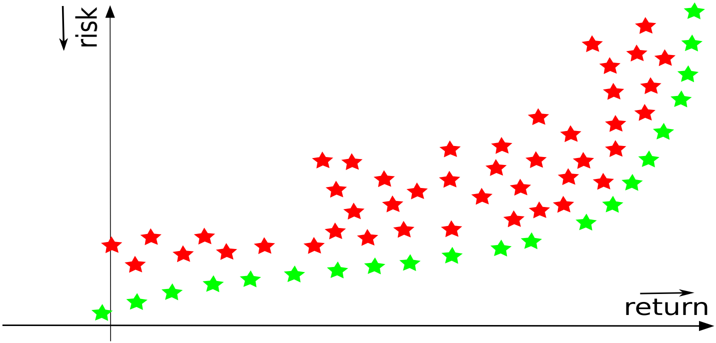
  with f_1, f_2 contrasting and/or with incomparable units (apples vs. oranges)

- Textbook example: portfolio selection problem
  - X = set of financial instruments portfolios I can buy
  - $f_1(x)$ = expected return of portfolio x (€)
  - $f_2(x)$ = risk of portfolio x not achieving the expected return (%, CVAR, . . .)

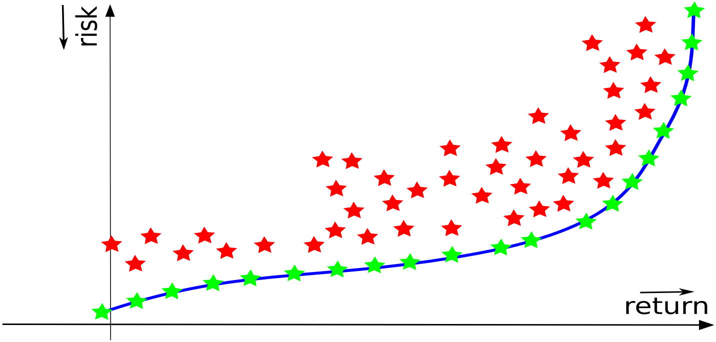
- Countless many others:
  - car cost vs. flashiness vs. km/l vs. # seats vs. trunk space . . .
  - # separated points vs. margin in SVM
  - . . .



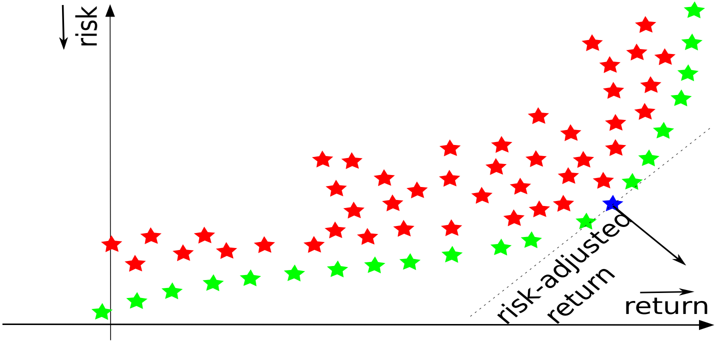
- No “best” solution, only



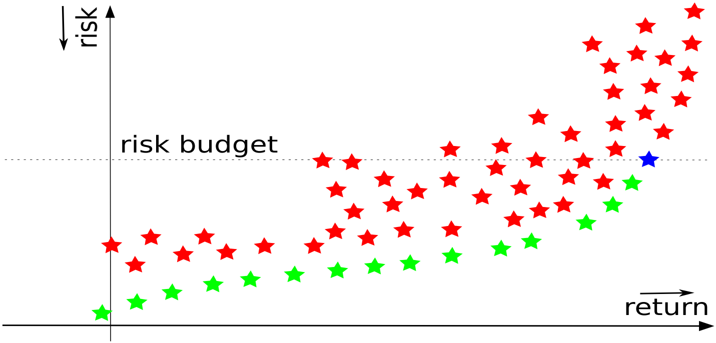
- No “best” solution, only non-dominated ones on the



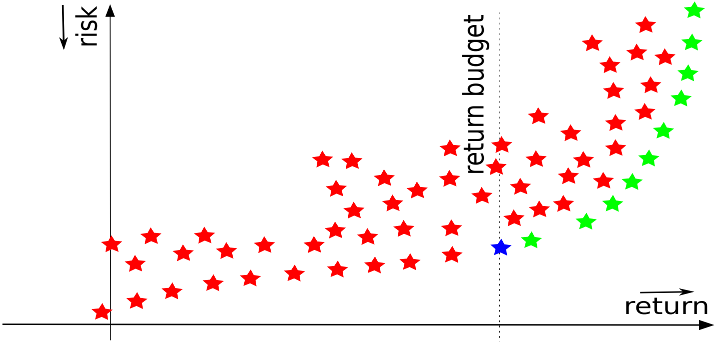
- No "best" solution, only non-dominated ones on the Pareto frontier
- Two practical solutions:



- No "best" solution, only non-dominated ones on the Pareto frontier
- Two practical solutions: maximize risk-adjusted return,
  a.k.a. scalarization $\min\left\{ f_1(x) + \alpha f_2(x) \ : \ x \in X \right\}$ (which α??)



- No "best" solution, only non-dominated ones on the Pareto frontier
- Two practical solutions: maximize return with budget on maximum risk, a.k.a. budgeting $\min \{ f_1(x) \ : \ f_2(x) \leq \beta_2 \ , \ x \in X \}$ (which β_2??)



- No "best" solution, only non-dominated ones on the Pareto frontier
- Two practical solutions: minimize risk with budget on minimum return, a.k.a. budgeting $\min\left\{ f_2(x) \;:\; f_1(x) \leq \beta_1 \,,\, x \in X \right\}$ (which β_1??)
- All a bit fuzzy, but it's the nature of the beast
- We always assume this done if necessary at modelling stage (cf. SVM)

Outline

- Since we minimize/maximize stuff, infima/suprema are important
- "$f : X \to \mathbb{R}$" precisely because $\mathbb{R}$ totally ordered:

  $$\forall x, y \in X, \text{ either } f(x) \leq f(y) \text{ or } f(y) \leq f(x)$$

  ($\mathbb{R}^k$ is not such for $k > 1$, cf. multi-objective)
- $S \subseteq \mathbb{R}$, $\underline{s} = \inf S \iff \underline{s} \leq s \;\; \forall s \in S \;\; \wedge \;\; \forall t > \underline{s} \; \exists s \in S \text{ s.t. } s \leq t$
- $S \subseteq \mathbb{R}$, $\bar{s} = \sup S \iff \bar{s} \geq s \;\; \forall s \in S \;\; \wedge \;\; \forall t < \bar{s} \; \exists s \in S \text{ s.t. } s \geq t$

- Since we minimize/maximize stuff, infima/suprema are important
- "$f : X \to \mathbb{R}$" precisely because $\mathbb{R}$ totally ordered:

  $$\forall\, x, y \in X, \text{ either } f(x) \leq f(y) \text{ or } f(y) \leq f(x)$$

  ($\mathbb{R}^k$ is not such for $k > 1$, cf. multi-objective)
- $S \subseteq \mathbb{R}$, $\underline{s} = \inf S \iff \underline{s} \leq s \;\; \forall s \in S \;\; \wedge \;\; \forall t > \underline{s} \, \exists\, s \in S$ s.t. $s \leq t$
- $S \subseteq \mathbb{R}$, $\bar{s} = \sup S \iff \bar{s} \geq s \;\; \forall s \in S \;\; \wedge \;\; \forall t < \bar{s} \, \exists\, s \in S$ s.t. $s \geq t$
- Issue: $\inf S / \sup S$ may not exist in $\mathbb{R}$
- Set of extended reals: $\overline{\mathbb{R}} = \{ -\infty \} \cup \mathbb{R} \cup \{ +\infty \}$ (usually just $\mathbb{R}$)
- For all $S \subseteq \mathbb{R}$, $\sup / \inf S \in \overline{\mathbb{R}}$
- $\inf S = -\infty$ just a convenient notation for "there is no (finite) inf"
- $\inf \emptyset = \infty$, $\sup \emptyset = -\infty$
- Should write "$\inf\{ f(x) \ldots$", but we want optimal solutions (if any)

- We often do iterations, hence produce sequences $v_1, v_2, \ldots$ (think sequence of iterates $\{ x_i \} \subset X$ and $v_i = f(x_i)$)
- Typically we can't get f_* in finite time ($\exists\, i \;\; v_i = f_*$), but we can "get as close as we want": there in the limit
- $\lim_{i \to \infty} v_i = v \iff \forall \varepsilon > 0 \, \exists\, h$ s.t. $| v_i - v | \leq \varepsilon \; \forall i \geq h$
- A sequence may not have limit: are we "not converging"?
- Any monotone sequence has a limit (monotone algorithms are good)
- The obvious way to make $\{ v_i \}$ monotone: keep aside the best

  $v^*_i = \min\{ v_h \; : \; h \leq i \}$ (best value at iteration i)
- $v^*_1 \geq v^*_2 \geq v^*_3 \geq \ldots \Longrightarrow v^*_\infty = \lim_{i \to \infty} v^*_i \geq f_*$ (asymptotic estimate)
- $\lim_{i \to \infty} v^*_i = v^*_\infty = f_* \Longrightarrow \{ v_i \}$ minimizing sequence (of values)

Forcing monotonicity on sequences in $\mathbb{R}$: the hard way

- Extract monotone sequences from $\{ v_i \}$ "the hard way":
  $\underline{v}_i = \inf\{ v_h \,:\, h \geq i \} \quad , \quad \bar{v}_i = \sup\{ v_h \,:\, h \geq i \}$
- $\underline{v}_1 \leq \underline{v}_2 \leq \underline{v}_3 \leq \ldots$, $\bar{v}_1 \geq \bar{v}_2 \geq \bar{v}_3 \geq \ldots \implies$ they still have a limit
- $\liminf_{i \to \infty} v_i := \lim_{i \to \infty} \underline{v}_i = \sup_i \underline{v}_i$
- $\limsup_{i \to \infty} v_i := \lim_{i \to \infty} \bar{v}_i = \inf_i \underline{v}_i$
- $\bar{v}_i \geq \underline{v}_i \implies \limsup_{i \to \infty} v_i \geq \liminf_{i \to \infty} v_i$
- $\lim_{i \to \infty} v_i = v \iff \limsup_{i \to \infty} v_i = v = \liminf_{i \to \infty} v_i$
- $\liminf_{i \to \infty} v_i = f_* \implies \{ v_i \}$ minimizing sequence (of values)
- A stronger definition: $\liminf_{i \to \infty} v_i = f_* \implies \lim_{i \to \infty} v_i^* = f_*$

Exercise: Prove the result

Exercise: Prove that $\Longleftarrow$ does not hold. Discuss the significance if v_i is the sequence of values of iterates of a minimization algorithm

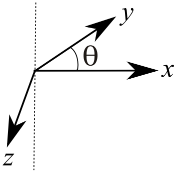
Outline

- Single numbers are not enough (except for objective function values)
- Euclidean space $\mathbb{R}^n := \{ [x_1, x_2, \ldots, x_n] : x_i \in \mathbb{R} \quad i = 1, \ldots, n \}$
- $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \ldots \mathbb{R}$, Cartesian product of $\mathbb{R}$ n times
- $x \in \mathbb{R}^n$ usually considered "column vector" $\in \mathbb{R}^{n \times 1}$ (a "^{T}" needed)
- Closed under sum and scalar multiplication

$$x + y := [x_1 + y_1, \ldots, x_n + y_n] \ , \ \alpha x := [\alpha x_1, \ldots, \alpha x_n]$$

- Finite vector space: each $x \in \mathbb{R}^n$ can be obtained from a finite basis (canonical base is u_i having 1 in position i and 0 elsewhere)
- Not all vector spaces are finite
- Not a totally ordered set
- Concept of "limit" requires topology: "what is close to what"

- scalar product of $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$:
  $\langle x , y \rangle := y^T x = \sum_{i=1}^n x_i y_i = x_1 y_1 + \cdots + x_n y_n$
  (will often succumb to temptation to write it just "yx" or "$y \cdot x$")
- Properties $\equiv$ definition of scalar product:
  1. $\langle x , y \rangle = \langle y , x \rangle \quad \forall x, y \in \mathbb{R}^n$ (symmetry)
  2. $\langle x , x \rangle \geq 0 \quad \forall x \in \mathbb{R}^n \ , \ \langle x , x \rangle = 0 \iff x = 0$
  3. $\langle \alpha x , y \rangle = \alpha \langle x , y \rangle \quad \forall x \in \mathbb{R}^n, \alpha \in \mathbb{R}$
  4. $\langle x + y , z \rangle = \langle x , z \rangle + \langle y , z \rangle \quad \forall x, y, z \in \mathbb{R}^n$



- Geometric interpretation: $\langle x , y \rangle = \| x \| \cdot \| y \| \cdot \cos(\theta)$
  1. $x \perp y \iff \langle x , y \rangle = 0$
  2. $\langle x , y \rangle > 0 \equiv$ "x and y point in the same direction"
- More general: $\langle x , y \rangle_M := y^T M x$ with $M \succ 0$ ($x \longrightarrow M^{-1/2} x$)
- Other spaces (matrices, integrable functions, random variables, . . .)
- Not just theoretical stuff (recall SVM)

- Euclidean norm: $\| x \| := \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{\langle x \, , \, x \rangle}$ (induced by $\langle \cdot \, , \, \cdot \rangle$)

- Properties $\equiv$ definition of norm:

  1. $\| x \| \geq 0 \quad \forall x \in \mathbb{R}^n$, $\| x \| = 0 \iff x = 0$
  2. $\| \alpha x \| = |\alpha| \| x \| \quad \forall x \in \mathbb{R}^n, \alpha \in \mathbb{R}$
  3. $\| x + y \| \leq \| x \| + \| y \| \quad \forall x, y \in \mathbb{R}^n$ (triangle inequality)

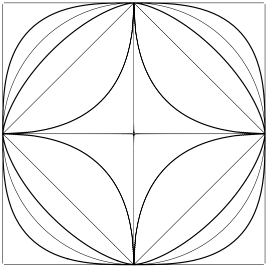
- $| \langle x \, , \, y \rangle |^2 \leq \| x \| \| y \| \quad \forall x, y \in \mathbb{R}^n$ (Cauchy-Schwarz inequality)

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

- $2 \| x \|^2 + 2 \| y \|^2 = \| x + y \|^2 + \| x - y \|^2$ (Parallelogram Law)

A useful generalization: p-norm

- Just the "most natural" among many:
  - $\| x \|_1 := \sum_{i=1}^{n} | x_i |$
  - $\| x \|_\infty := \max\{ | x_i | \,:\, i = 1, \dots, n \}$
  - $\| x \|_0 := | \{ i \,:\, | x_i | > 0 \} |$
  - Other ones (e.g. for matrices . . .)



- Many (but not all) derive from p-norm:
  $\| x \|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$
- Convex for $p \geq 1$, nonconvex for $p < 1$
- $\| \cdot \|_1$ "best convex approximation" of $\| \cdot \|_0$ (compressed sensing, . . .)
- $\langle x , y \rangle^2 \leq \| x \|_p \| y \|_q \quad 1/p + 1/q = 1$ (Hölder's inequality)
- Hereafter "$\| \cdot \| = \| \cdot \|_2$", but all norms are topologically equivalent:

$$\exists\, 0 < \alpha < \beta \text{ s.t.} \qquad \alpha \| x \|' \leq \| x \| \leq \beta \| x \|' \qquad \forall\, x$$

(because $\mathbb{R}^n$ is a finite vector space)

- Euclidean distance between x and y

$$d(\, x \,,\, y \,) := \|\, x - y \,\| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

("norm of x when y is the origin")

- Properties $\equiv$ definition of distance:
  1. $d(\, x \,,\, y \,) \geq 0 \quad \forall x, y \in \mathbb{R}^n \;,\; d(\, x \,,\, y \,) = 0 \iff x = y$
  2. $d(\, \alpha x \,,\, 0 \,) = |\alpha| d(\, x \,,\, 0 \,) \quad \forall x \in \mathbb{R}^n, \alpha \in \mathbb{R}$
  3. $d(\, x \,,\, y \,) \leq d(\, x \,,\, z \,) + d(\, z \,,\, y \,) \quad \forall x, y, z \in \mathbb{R}^n$ (triangle inequality)

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

- The distance/norm defines the topology of the vector space, but doesn't really matter: all is "$\exists$ ball", "$\forall$ small ball", and all norms are equivalent

- Finally, 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
- No obvious lim inf / lim sup ($\mathbb{R}^n$ is not totally ordered)

Exercise: Would $\liminf_{i \to \infty} d(x_i , x) = 0$ and/or $\limsup_{i \to \infty} d(x_i , x) = 0$ make sense? Which?

- We want to solve (P) $f_* = \min\{ f(x) \ : \ x \in X \}$ with $X \subseteq \mathbb{R}^n$
- Construct a minimizing sequence: $\{ x_i \}$ s.t. $\{ f(x_i) \} \to f_*$
- We know it is not enough for having "solved" (P), recall
  3. $\min\{ x \ : \ x \in \mathbb{R} \wedge x > 0 \}$ $\{ x_i = 1/i \}$ $\{ f(x_i) \} \to 0$
  4. $\min\{ 1/x \ : \ x \in \mathbb{R} \wedge x > 0 \}$ $\{ x_i = i \}$ $\{ f(x_i) \} \to 0$
- Minimizing sequences, but no optimal solution
- Want conditions that ensure

$$\{ f(x_i) \} \to f_* \quad \Longrightarrow \quad \{ x_i \} \to x_* \in X \text{ optimal solution}$$

- Two different problems
- There are more (remember the other cases)

- Given $S \subseteq \mathbb{R}^n$, interior/boundary points of S:
  - $x \in int(\,S\,) \equiv$ interior of S $:= \exists r > 0$ s.t. $\mathcal{B}(\,x\,,\,r\,) \subseteq S$
  - $x \in \partial(\,S\,) \equiv$ boundary of S $:= \forall r > 0\, \exists y, z \in \mathcal{B}(\,x\,,\,r\,)$ s.t. $y \in S \wedge z \notin S$

  note: $x \in int(\,S\,) \Longrightarrow x \in S$, but $x \in \partial S \not\Longrightarrow x \in S$
- S open if $S = int(\,S\,)$: "I have no points on the boundary"
- $cl(\,S\,) \equiv$ closure of $S := int(\,S\,) \cup \partial S$: "me and my boundary"
- $S \subseteq \mathbb{R}^n$ closed if $S = cl(\,S\,) \equiv \mathbb{R}^n \setminus S$ (the complement) open: "all points on my boundary are mine"
- $int(\,S\,) \neq \emptyset \Longrightarrow S$ full dimensional
- Sometimes, relative interior useful

- S closed $\iff$ $\forall\, S \supset \{\, x_i \,\} \to x \implies x \in S$

  $\equiv$ all limit points of sequences in S are in S

- Algebra of open/closed sets:
  - $\{\, S_i \,\}$ (infinitely many) open sets $\implies \bigcup_i S_i$ open
  - S_1 and S_2 are open $\implies S_1 \cap S_2$ is open
  - $\{\, S_i \,\}$ (infinitely many) closed sets $\implies \bigcap_i S_i$ closed
  - S_1 and S_2 are closed $\implies S_1 \cup S_2$ is closed

Exercise: prove $\mathbb{R}^n$ and $\emptyset$ are both closed and open (hint: what boundary?)

Exercise: exhibit a set that is neither open nor closed

Exercise: $\{\, S_i \,\}$ (infinitely many) open sets $\implies \bigcap_i S_i$ open: true?

Exercise: $\{\, S_i \,\}$ (infinitely many) closed sets $\implies \bigcup_i S_i$ closed: true?

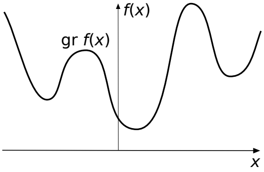
- $S \subseteq \mathbb{R}^n$ is bounded $:= \exists r > 0$ s.t. $S \subseteq \mathcal{B}(0, r)$
- Closed + bounded = compact
- Sequences $\{ x_i \} \subset \mathbb{R}^n$ and $\{ n_i \} \subseteq \mathbb{N}$, subsequence $\{ x_{n_i} \} \subseteq \{ x_i \}$
- Sequence $\{ x_i \}$: x is an accumulation point if $\exists \{ x_{n_i} \} \to x \equiv \liminf_{i \to \infty} d(x_i , x) = 0$
- Bolzano-Weierstrass Theorem: if $S \subseteq \mathbb{R}^n$ is compact, then any sequence $\{ x_i \} \subseteq S$ has an accumulation point $x \in S$
- X compact $\Longrightarrow$ any minimizing sequence has one accumulation point (candidate to be an optimal solution)

- Somewhat surprising: $|\{x_i\}| = \aleph_0 \ll \aleph_1 = |\mathbb{R}^n|$
- In hindsight, not: it is very hard to keep all points far from each other
- Unitary hypercube $[0,1]^n$, divide each side equally: $(h+1)^n$ points give a regular mesh of h^n hypercubes, each of volume $(1/h)^n$
- Again, thanks to $\mathbb{R}^n$ being a finite vector space
- "Closed" and "bounded" each solve a separate issue, recall
  3. $\min\{\, x \,:\, x \in \mathbb{R} \wedge x > 0 \,\}$ $\{\, x_i = 1/i \,\}$ $\{\, f(x_i) \,\} \to 0$
  4. $\min\{\, 1/x \,:\, x \in \mathbb{R} \wedge x > 0 \,\}$ $\{\, x_i = i \,\}$ $\{\, f(x_i) \,\} \to 0$
- But not all issues solved yet, recall
  5. $\min\{\, f(x) \,:\, x \in [\,0\,,\,1\,]\,\}$ with $f(x) = \begin{cases} x & \text{if } x > 0 \\ 1 & \text{if } x = 0 \end{cases}$

Outline

- $f : D \to \mathbb{R}$, domain $D = \mathrm{dom}(\,f\,)$ may not be all $\mathbb{R}^n$
- Equivalent for minimization: $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, $f(\,x\,) = \infty$ for $x \notin D$
  (usually OK to ignore $\mathrm{dom}(f)$ if f extended-real-valued)

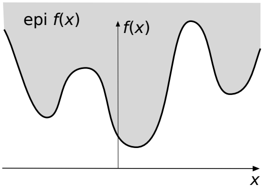
- $f : D \to \mathbb{R}$, domain $D = \mathrm{dom}(f)$ may not be all $\mathbb{R}^n$
- Equivalent for minimization: $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, $f(x) = \infty$ for $x \notin D$ (usually OK to ignore dom(f) if f extended-real-valued)



- f "lives" in $\mathbb{R}^{n+1}$
- $\mathrm{gr}(f) = \{(f(x), x) \ : \ x \in \mathrm{dom}(f)\}$ (graph)

- $f : D \to \mathbb{R}$, domain $D = \text{dom}(f)$ may not be all $\mathbb{R}^n$

- Equivalent for minimization: $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, $f(x) = \infty$ for $x \notin D$
(usually OK to ignore dom(f) if f extended-real-valued)

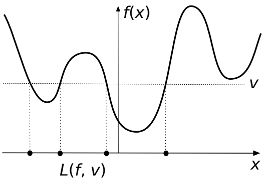


- f "lives" in $\mathbb{R}^{n+1}$

- $\text{gr}(f) = \{ (f(x), x) : x \in \text{dom}(f) \}$ (graph)
- $\text{epi}(f) = \{ (v, x) : x \in \text{dom}(f) \wedge v \geq f(x) \}$
(epigraph, we are minimizing)

- Looking at f in $\mathbb{R}^n$ requires projection:

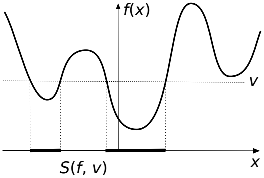
- $f : D \to \mathbb{R}$, domain $D = \mathrm{dom}(f)$ may not be all $\mathbb{R}^n$
- Equivalent for minimization: $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, $f(x) = \infty$ for $x \notin D$
  (usually OK to ignore dom(f) if f extended-real-valued)



- f "lives" in $\mathbb{R}^{n+1}$
- $\mathrm{gr}(f) = \{ (f(x), x) : x \in \mathrm{dom}(f) \}$ (graph)
- $\mathrm{epi}(f) = \{ (v, x) : x \in \mathrm{dom}(f) \wedge v \geq f(x) \}$
  (epigraph, we are minimizing)

- Looking at f in $\mathbb{R}^n$ requires projection:
  - $L(f, v) = \{ x \in \mathrm{dom}(f) : f(x) = v \}$ (level set)

- $f : D \to \mathbb{R}$, domain $D = \mathrm{dom}(\, f\,)$ may not be all $\mathbb{R}^n$
- Equivalent for minimization: $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, $f(\, x\,) = \infty$ for $x \notin D$ (usually OK to ignore $\mathrm{dom}(f)$ if f extended-real-valued)



- f "lives" in $\mathbb{R}^{n+1}$
- $\mathrm{gr}(f) = \{\, (\, f(\, x\,)\,,\, x\,) \;:\; x \in \mathrm{dom}(\, f\,)\,\}$ (graph)
- $\mathrm{epi}(f) = \{\, (\, v\,,\, x\,) \;:\; x \in \mathrm{dom}(\, f\,) \wedge v \geq f(\, x\,)\,\}$ (epigraph, we are minimizing)

- Looking at f in $\mathbb{R}^n$ requires projection:
  - $L(\, f\,,\, v\,) = \{\, x \in \mathrm{dom}(\, f\,) \;:\; f(\, x\,) = v\,\}$ (level set)
  - $S(\, f\,,\, v\,) = \{\, x \in \mathrm{dom}(\, f\,) \;:\; f(\, x\,) \leq v\,\}$ (sublevel set, we are minimizing)
- When maximizing, $f(\, x\,) = -\infty$ for $x \notin D$, superlevel set and ipograph

- The other problem: f "jumps wildly"

  5. $\min\{ f(x) \;:\; x \in [0\,,\,1] \}$ with $f(x) = \begin{cases} x & \text{if } x > 0 \\ 1 & \text{if } x = 0 \end{cases}$

  values "near" x do not reliably "predict" what happens there

- $f : \mathbb{R}^n \to \mathbb{R}$ continuous at x:
  - $\{ x_i \} \to x \Longrightarrow \{ f(x_i) \} \to f(x)$
  - $\forall \varepsilon > 0 \; \exists \delta > 0$ s.t. $| f(y) - f(x) | < \varepsilon \;\; \forall y \in \mathcal{B}(x\,,\,\delta)$

  continuous on $S \equiv \forall x \in S$, just "continuous" $\equiv S = \mathbb{R}^n$

- Intermediate value theorem: $f : \mathbb{R} \to \mathbb{R}$ continuous on $[a\,,\,b]$, $\forall\, v$ s.t.

  $\min\{ f(a)\,,\, f(b) \} \leq v \leq \max\{ f(a)\,,\, f(b) \} \; \exists\, c \in [a\,,\,b]$ s.t. $f(c) = v$

- Continuity easily preserved: f, g continuous at $x \Longrightarrow$
  - $f + g$, $f \cdot g$ continuous at x
  - $\max\{ f\,,\, g \}$, $\min\{ f\,,\, g \}$ continuous at x
  - $f \circ g \equiv f(g(\cdot))$ continuous at x

- Yet, plenty of non-continuous functions ($sign(x)$, $1 / (x - 1)$, ...)

- Weierstrass extreme value theorem (in our parlance):

  $X \subseteq \mathbb{R}^n$ compact and f continuous on $X \Longrightarrow$

  (P) has an optimal solution

- Works for both min and max

- In other words:

  $X \subseteq \mathbb{R}^n$ compact and f continuous on $X \Longrightarrow$

  all accumulation points of any minimizing sequence are optima

  and there is at least one

- Thus, "X compact and f continuous (on X)" natural assumptions

- "f continuous" basically necessary (recall 5.)

- But non-bounded X also common, and $f_* = -\infty$ happens

- f Lipschitz continuous (L.c.) on S if $\exists L > 0$ such that

$$| f(x) - f(y) | \leq L \| x - y \| \qquad \forall x, y \in S$$

globally L.c.: $S = \mathbb{R}^n$, locally L.c. at x: $\exists \varepsilon > 0$ s.t. $S \supseteq \mathcal{B}(x, \varepsilon)$

- Note: L depends on S (locally L.c. $\not\Longleftarrow$ globally L.c.)

- Lipschitz continuity $\equiv$ f cannot change too fast

  strong relationships with derivatives (see next)

- Much stronger property: Lipschitz continuity $\Longrightarrow$ continuity

Exercise: Prove it

Exercise: Exhibit simple functions that are continuous but not globally L.c.

Exercise: Exhibit a continuous functions that is not L.c. on a compact set

- Weaker condition: f is lower[upper] semi-continuous (l.[u.]s.c.) at x if

$$\{ x_i \} \to x \implies f(x) \le \liminf_{i \to \infty} f(x_i) \;\; [\, f(x) \ge \limsup \ldots \,]$$

- Also written $\liminf_{y \to x} f(y) \ge f(x)$, $\limsup_{y \to x} f(y) \le f(x)$
- Can jump down [up] wildly, but can never jump up [down]
- Particularly useful example: indicator function of $S \subset \mathbb{R}^n$

  $\imath_S(x) = 0$ if $x \in S$, $\imath_S(x) = +\infty$ if $x \notin S$

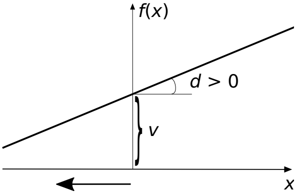
  (think $S = \mathrm{dom}(f)$) clearly not continuous, but l.s.c.
- f continuous at $x \iff f$ is both l.s.c. and u.s.c. at x
- Any l.s.c. f attains minimum (but not maximum) on any compact set X
- X compact and f l.s.c. $\implies$ all accumulation points of any minimizing sequence are optimal solutions, and there is at least one
- As a great man said: "(convex) optimization is a one-sided world"

Outline

- (the vector) "Space ($\mathbb{R}^n$) is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is."

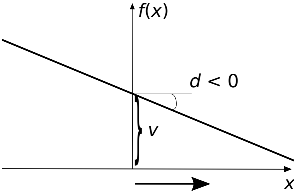
- (the vector) "Space ($\mathbb{R}^n$) is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is." Which way is x_*?

- (the vector) "Space ($\mathbb{R}^n$) is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is." Which way is x_*?



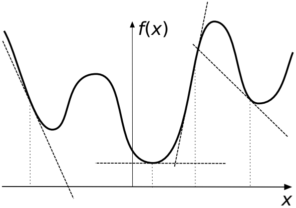
- $f(x) := dx + v : \mathbb{R} \to \mathbb{R}$ (linear) is easy: always left if $d > 0$,

- (the vector) "Space ($\mathbb{R}^n$) is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is." Which way is x_*?



- $f(x) := dx + v : \mathbb{R} \to \mathbb{R}$ (linear) is easy: always left if $d > 0$, right if $d < 0$

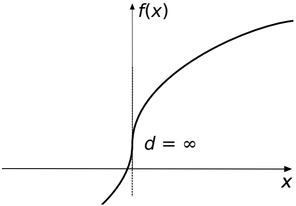
- (the vector) "Space ($\mathbb{R}^n$) is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is." Which way is x_*?



- $f(x) := dx + v : \mathbb{R} \to \mathbb{R}$ (linear) is easy: always left if $d > 0$, right if $d < 0$
- Obvious idea: use the linear function that best locally approximates f
- Trusty old derivative:
  $d = f'(x) = \lim_{t \to 0}[f(x+t) - f(x)]/t$

- Easy closed-forms for most reasonable functions

- (the vector) "Space ($\mathbb{R}^n$) is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is." Which way is x_*?

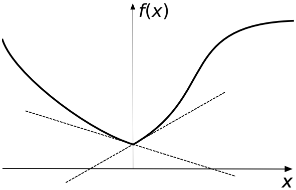


- $f(x) := dx + v : \mathbb{R} \to \mathbb{R}$ (linear) is easy: always left if $d > 0$, right if $d < 0$
- Obvious idea: use the linear function that best locally approximates f
- Trusty old derivative:
  $d = f'(x) = \lim_{t\to 0}[f(x+t) - f(x)]/t$

- Easy closed-forms for most reasonable functions

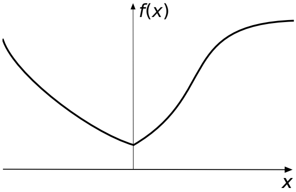
- Provided the limit is finite

- (the vector) "Space ($\mathbb{R}^n$) is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is." Which way is x_*?

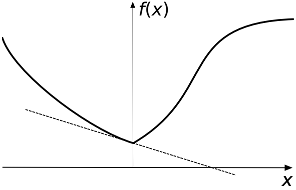


- $f(x) := dx + v : \mathbb{R} \to \mathbb{R}$ (linear) is easy: always left if $d > 0$, right if $d < 0$
- Obvious idea: use the linear function that best locally approximates f
- Trusty old derivative: $d = f'(x) = \lim_{t \to 0}[f(x+t) - f(x)] / t$

- Easy closed-forms for most reasonable functions
- Provided the limit is finite . . . and it exists at all
- f differentiable on S if $f'(x)$ exists finite $\forall x \in S$

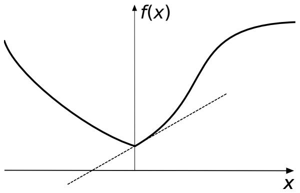


- Left and right derivatives:



- Left and right derivatives:

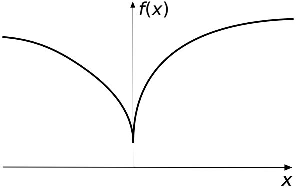
  $f'_{-}(x) = \lim_{t \to 0_-} [f(x+t) - f(x)]/t$



- Left and right derivatives:

  $$f'_{-}(x) = \lim_{t\to 0_-}[\,f(x+t) - f(x)\,]\,/\,t$$

  $$f'_{+}(x) = \lim_{t\to 0_+}[\,f(x+t) - f(x)\,]\,/\,t$$

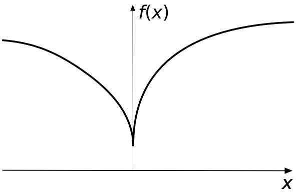


- Left and right derivatives:

  $$f'_{-}(x) = \lim_{t \to 0_{-}} [\, f(x+t) - f(x)\,] \,/\, t$$

  $$f'_{+}(x) = \lim_{t \to 0_{+}} [\, f(x+t) - f(x)\,] \,/\, t$$

- Can be as different as $-\infty$ and $+\infty$



- Left and right derivatives:

  $f'_-(x) = \lim_{t \to 0_-} [\, f(x+t) - f(x) \,] / t$

  $f'_+(x) = \lim_{t \to 0_+} [\, f(x+t) - f(x) \,] / t$

- Can be as different as $-\infty$ and $+\infty$

- f is differentiable at $x \in \operatorname{int} \operatorname{dom}(f) \iff f'_-(x) = f'_+(x)$ ($\Longleftarrow$ they $\exists$)

- f differentiable at $x \Longrightarrow f$ continuous at x

Exercise: Prove it

- f continuously differentiable: f' is also continuous

- Nondifferentiable functions happen (e.g., $|x|$, $\sqrt{x}$)

- $f : \mathbb{R}^n \to \mathbb{R}$, partial derivative of f w.r.t. x_i at $x \in \mathbb{R}^n$:

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

  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 $=$ vector of all partial derivatives

$$\nabla f(\, x \,) := \left[\, \frac{\partial f}{\partial x_1}(x) \,, \, \ldots \,, \, \frac{\partial f}{\partial x_n}(x) \,\right]$$

- Directional derivative at x along direction $d \in \mathbb{R}^n$:

$$\frac{\partial f}{\partial d}(x) := \lim_{t \to 0} \frac{f(x + td) - f(\, x \,)}{t}$$

- Of course, $\frac{\partial f}{\partial x_i} = \frac{\partial f}{\partial d}$ with $d = u_i$

- One-sided directional derivative: $\lim_{t \to 0_+} \ldots$ (generalizes f'_+ and f'_-)

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

$$\lim_{\|\, h\,\|\to 0} \frac{|\, f(x+h) - \phi(\, h\,)\,|}{\|\, h\,\|} = 0$$

- ϕ "first order approximation" of f at x
- The error in the approximation vanishes faster than linearly
- f differentiable at $x \Longrightarrow c = \nabla f(\, x\,) \;\equiv\; \phi(h) = \langle\, \nabla f(\, x\,)\,,\, h\,\rangle + f(\, x\,)$
  $\Longrightarrow$ first-order model of f at x: $L_x(\, y\,) = \nabla f(\, x\,)(\, y - x\,) + f(\, x\,)$
- Hence, f differentiable $\Longrightarrow \frac{\partial f}{\partial x_i}(x)$ exists $\forall\, i$
- The converse is not true
- More in general, f differentiable at $x \Longrightarrow \exists\, \frac{\partial f}{\partial d}(x)\;\; \forall d \in \mathbb{R}^n$, and

$$\frac{\partial f}{\partial d}(x) = \langle\, \nabla f(\, x\,)\,,\, d\,\rangle$$

Exercise: prove $-\nabla f(\, x\,)$ is the steepest descent direction at $x \equiv$ direction with most negative directional derivative

- $f : \mathbb{R}^n \to \mathbb{R}$ differentiable at $x \Longrightarrow f$ locally Lipschitz continuous at x; hence, f differentiable $\Longrightarrow f$ continuous (but $\Longleftarrow$ not true)

Exercise: Prove f differentiable $\Longrightarrow f$ continuous

- $\exists \delta > 0$ s.t. $\forall i \; \frac{\partial f}{\partial x_i}(y)$ continuous $\forall y \in \mathcal{B}(\, x \,,\, \delta \,) \Longrightarrow f$ differentiable at x

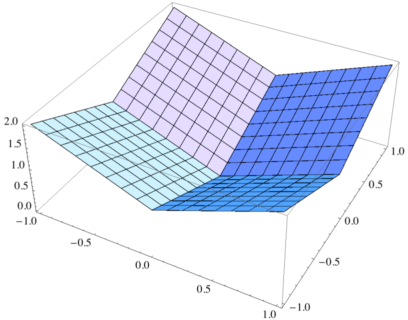
- The converse is not true ($\exists f$ differentiable with discontinuous $\frac{\partial f}{\partial x_i}$, weird)

- The good class for optimization: $C^1 := \nabla f(\, x \,)$ continuous everywhere

- $f \in C^1 \Longrightarrow f$ differentiable everywhere $\Longrightarrow f$ continuous everywhere

- Yet, nondifferentiable functions happen

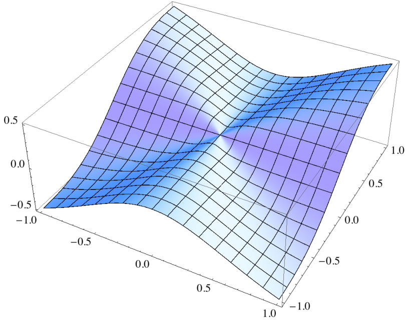
- $f(x_1, x_2) = \|\,[\,x_1\,,\,x_2\,]\,\|_1 = |\,x_1\,| + |\,x_2\,|$
- f continuous everywhere (why?)
- f non differentiable in $[\,0\,,\,0\,]$



Exercise: prove it (hint: compute 4 easy directional derivatives, prove they cannot ever have the form $\langle\, v\,,\,d\,\rangle$ for any v)

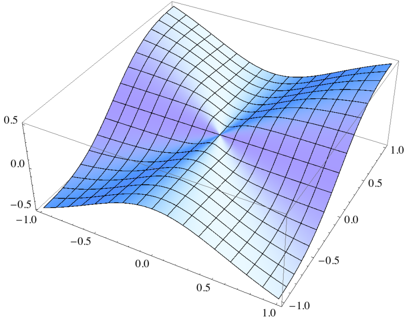
Exercise: where else f is non differentiable? Prove it is not

- $f(x_1, x_2) = \dfrac{x_1^2 x_2}{x_1^2 + x_2^2}$
- Assume $f(0, 0) = 0$ since $\lim_{\alpha \to 0} f(\alpha d_1, \alpha d_2) = 0$
- $\exists \frac{\partial f}{\partial d} \;\; \forall d \in \mathbb{R}^n \setminus \{\, [\, 0 \, , \, 0 \,] \,\}$
- f non differentiable in $[\, 0 \, , \, 0 \,]$



Exercise: prove all this (hint: compute $\lim_{t \to 0} f(td_1, td_2)/t$, prove it cannot ever have the form $\langle \, v \, , \, d \, \rangle$ for any v)

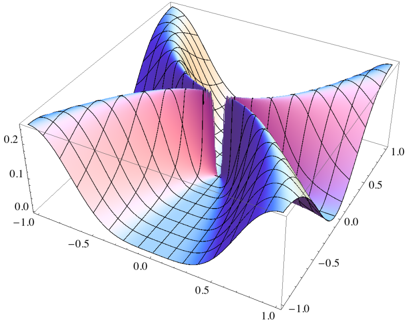
- $f(x_1, x_2) = \dfrac{x_1^2 x_2}{x_1^2 + x_2^2}$
- Assume $f(0,0) = 0$ since $\lim_{\alpha \to 0} f(\alpha d_1, \alpha d_2) = 0$
- $\exists \frac{\partial f}{\partial d} \;\; \forall d \in \mathbb{R}^n \setminus \{ [0\,,\,0] \}$
- f non differentiable in $[\,0\,,\,0\,]$



Exercise: prove all this (hint: compute $\lim_{t \to 0} f(td_1, td_2)/t$, prove it cannot ever 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 $\frac{\partial f}{\partial x_2}$ for directions where the limit is $\neq$)

- $f(x_1, x_2) = \left(\frac{x_1^2 x_2}{x_1^4 + x_2^2} \right)^2$ (again, $f(0,0) = 0$)
- f not continuous $\Longrightarrow$
  not differentiable at $[\,0\,,\,0\,]$
- Yet $\frac{\partial f}{\partial d}(0,0) = 0 \ \ \forall d \in \mathbb{R}^n$
- Directional derivatives
  $\exists$ and are all equal,
  yet ∇f does not



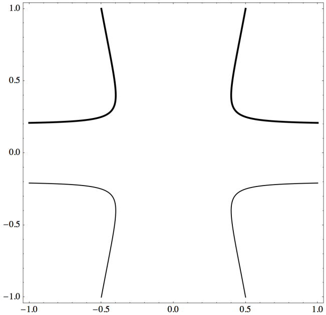
- Trick: f does nasty things on curved lines, not straight ones

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

Exercise: prove f not continuous at $[\,0\,,\,0\,]$ (check $\lim_{k\to\infty} f(1/k, 1/k^2)$)

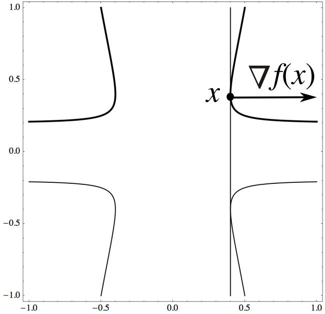
- In $\mathbb{R}^n$, $S(\, L_x \,,\, f(\, x\,)\,)$ is a line passing by x and $\nabla f(\, x\,) \perp S(\, L_x \,,\, f(\, x\,)\,)$

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



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

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

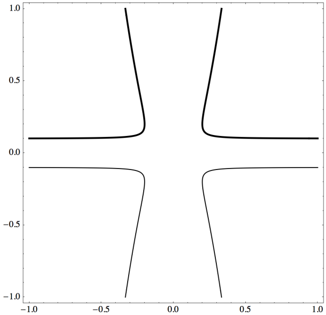


- f differentiable at $x \Longrightarrow$

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

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

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

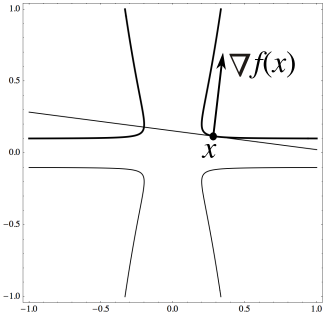


- f differentiable at $x \Longrightarrow$

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

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

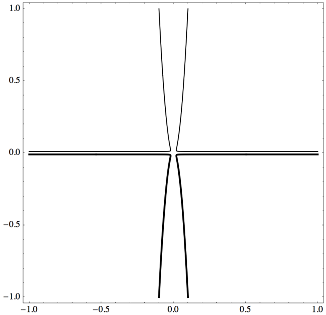
$$f(x_1, x_2) = \frac{x_1^2 x_2}{x_1^2 + x_2^2} \quad , \quad \nabla f(\,x\,) = \left[\, \frac{2x_1 x_2^3}{(x_1^2 + x_2^2)^2}\,,\, \frac{x_1^2(x_1^2 - x_2^2)}{(x_1^2 + x_2^2)^2} \,\right] \quad \text{(cf. Ex. 2)}$$



- f differentiable at $x \Longrightarrow$
  $S(\,L_x\,,\,f(\,x\,)\,) \perp S(\,f\,,\,f(\,x\,)\,) \perp \nabla f(\,x\,)$
- f differentiable at $x \Longrightarrow$
  $S(\,f\,,\,f(\,x\,)\,)$ "smooth"

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

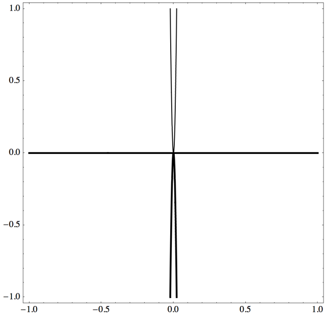
$$f(x_1, x_2) = \frac{x_1^2 x_2}{x_1^2 + x_2^2} \quad , \quad \nabla f(\,x\,) = \left[\frac{2x_1 x_2^3}{(x_1^2 + x_2^2)^2} \, , \, \frac{x_1^2(x_1^2 - x_2^2)}{(x_1^2 + x_2^2)^2} \right] \quad \text{(cf. Ex. 2)}$$



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

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

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



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

- f differentiable $\Longrightarrow$ all relevant objects in $\mathbb{R}^{n+1}$ and $\mathbb{R}^n$ are smooth
- f non differentiable $\Longrightarrow$ 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 partial derivatives $\in \mathbb{R}^{m,n}$

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

- Jacobian = matrix with gradients as rows
- As usual, much better if continuous $\equiv$ every f_j differentiable
- A special case of vector-valued function is particularly important

- $\frac{\partial f}{\partial x_i} : \mathbb{R}^n \to \mathbb{R}$, hence has partial derivatives
- Second order partial derivative (just do it twice) $$\frac{\partial^2 f}{\partial x_j \partial x_i} \qquad \frac{\partial^2 f}{\partial x_i \partial x_i} = \frac{\partial^2 f}{\partial x_i^2}$$
- $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n$, hence has a Jacobian: Hessian of f

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

- $\nabla^2 f$ is f'': much more complex object, but somes things generalise nicely
- Very important concept: second-order model = first-order model plus second-order term ($\equiv$ better) $$Q_x(y) = L_x(y) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)$$

- $\exists \delta > 0$ s.t. $\forall y \in \mathcal{B}(\, x \,,\, \delta \,)$

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

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

- Symmetry is important: all eigenvalues of $\nabla^2 f$ are real, useful geometric characterization, special cases (all $\geq 0 / \leq 0, \dots$)

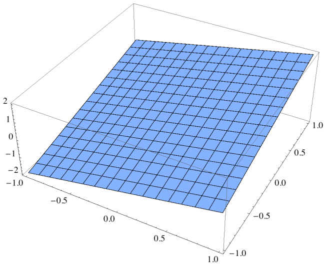
- The very good class: $C^2 := \nabla^2 f(\, x \,)$ continuous $\Longrightarrow$
  - $\nabla^2 f(\, x \,)$ symmetric
  - $\nabla f(\, x \,)$ continuous (why?)

- C^2 (strictly speaking C^3) is the best class ever for optimization

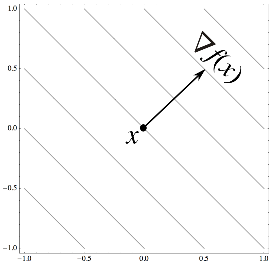
- Second-order information very useful, we will see why

Exercise: $f(x_1, x_2) = x_2^3 / (x_1^2 + x_2^2)$ is non differentiable in $[\, 0 \,,\, 0 \,]$; check that $\nabla^2 f$ is not symmetric there

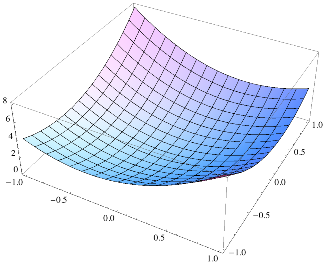
Outline



- Linear function: $f(x) = cx$, fixed $c \in \mathbb{R}^n$
- $\nabla f(x) = c \quad , \quad \nabla^2 f(x) = 0$



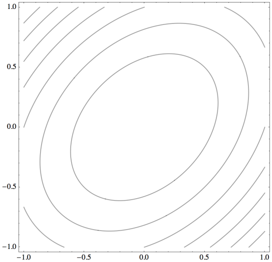
- Linear function: $f(x) = cx$, fixed $c \in \mathbb{R}^n$
- $\nabla f(x) = c \quad , \quad \nabla^2 f(x) = 0$
- Level sets are parallel hyperplanes orthogonal to c $(= [1, 1]$ here)



- Quadratic function: $f(x) = \frac{1}{2}x^T Qx + qx$ fixed $Q \in \mathbb{R}^{n\times n}$, $q \in \mathbb{R}^n$; here

$$Q = \begin{bmatrix} 6 & -2 \\ -2 & 6 \end{bmatrix} \quad , \quad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

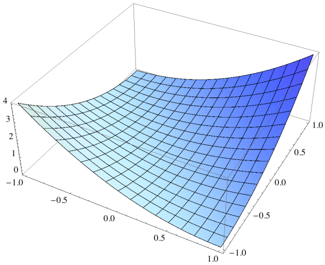
- $\nabla f(x) = Qx + q \quad , \quad \nabla^2 f(x) = Q$



- Quadratic function: $f(x) = \frac{1}{2}x^T Qx + qx$
  fixed $Q \in \mathbb{R}^{n\times n}$, $q \in \mathbb{R}^n$; here

$$Q = \begin{bmatrix} 6 & -2 \\ -2 & 6 \end{bmatrix} \quad , \quad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

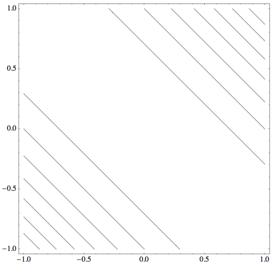
- $\nabla f(x) = Qx + q \quad , \quad \nabla^2 f(x) = Q$
- Level sets are ellipsoids



- Quadratic function: $f(x) = \frac{1}{2}x^T Qx + qx$
  fixed $Q \in \mathbb{R}^{n\times n}$, $q \in \mathbb{R}^n$; here

$$Q = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \quad , \quad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

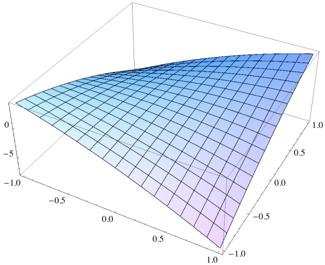
- $\nabla f(x) = Qx + q$, $\nabla^2 f(x) = Q$



- Quadratic function: $f(x) = \frac{1}{2}x^T Qx + qx$
  fixed $Q \in \mathbb{R}^{n\times n}$, $q \in \mathbb{R}^n$; here

$$Q = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \quad , \quad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

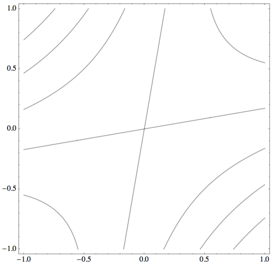
- $\nabla f(x) = Qx + q$, $\nabla^2 f(x) = Q$
- Level sets are degenerate ellipsoids



- Quadratic function: $f(x) = \frac{1}{2}x^T Qx + qx$
  fixed $Q \in \mathbb{R}^{n\times n}$, $q \in \mathbb{R}^n$; here

$$Q = \begin{bmatrix} -2 & 6 \\ 6 & -2 \end{bmatrix} \quad , \quad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

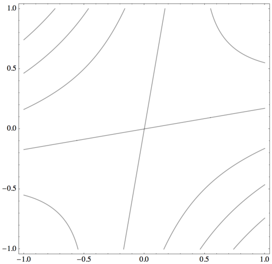
- $\nabla f(x) = Qx + q$, $\nabla^2 f(x) = Q$



- Quadratic function: $f(x) = \frac{1}{2}x^T Qx + qx$ fixed $Q \in \mathbb{R}^{n\times n}$, $q \in \mathbb{R}^n$; here

$$Q = \begin{bmatrix} -2 & 6 \\ 6 & -2 \end{bmatrix} \quad , \quad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- $\nabla f(x) = Qx + q$, $\nabla^2 f(x) = Q$
- Level sets are hyperboloids



- Quadratic function: $f(x) = \frac{1}{2}x^T Qx + qx$
  fixed $Q \in \mathbb{R}^{n \times n}$, $q \in \mathbb{R}^n$; here

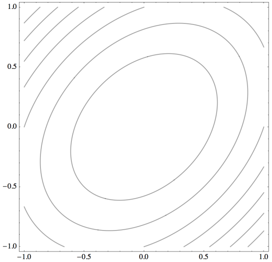
$$Q = \begin{bmatrix} -2 & 6 \\ 6 & -2 \end{bmatrix} \quad , \quad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- $\nabla f(x) = Qx + q$, $\nabla^2 f(x) = Q$
- Level sets are hyperboloids
- Several different cases, let's try to work them out
- Can always assume Q symmetric $\Longrightarrow$ has spectral decomposition

$$x^T Qx = [(x^T Qx) + (x^T Qx)^T] / 2 = x^T [(Q + Q^T) / 2] x = H \Lambda H^T$$

  - H^i eigenvectors, orthonormal ($H_i \perp H_j$, $\| H_i \| = 1$)
  - Λ diagonal, λ_i corresponding real eigenvalues
- Sign of eigenvalues (positive/negative definiteness) $\to$ shape of level sets

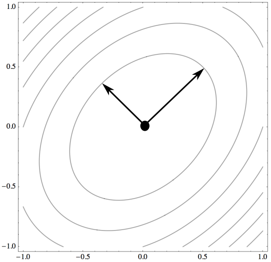
- Easy case: Q nonsingular $\equiv \lambda_i \neq 0 \, \forall i$ (regardless of the sign)
- Then $f(x) = \frac{1}{2}(x - \bar{x})^T Q(x - \bar{x})$ [+ constant] for $\bar{x} = -Q^{-1}q$ (**check**)
- $\bar{x}$ center of the ellipsoid, $y = x - \bar{x}$, $f_{\bar{x}}(y) = y^T Q y$ [+ constant]



- Along H_i: $f_i(\alpha) = f_{\bar{x}}(\alpha H_i) = \alpha^2 \lambda_i$ (**check**)
- $S(f_{\bar{x}}, 1) \equiv f_i(\alpha) = 1 \equiv \alpha = \sqrt{1/\lambda_i} \Longrightarrow$

$$Q = \begin{bmatrix} 6 & -2 \\ -2 & 6 \end{bmatrix} , \quad H = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} , \quad \lambda = \begin{bmatrix} 8 \\ 4 \end{bmatrix}$$

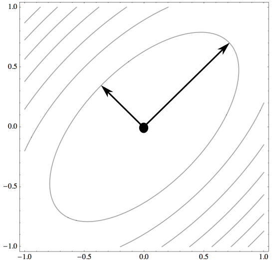
- Easy case: Q nonsingular $\equiv \lambda_i \neq 0 \, \forall i$ (regardless of the sign)
- Then $f(\, x\,) = \frac{1}{2}(\, x - \bar{x}\,)^T Q(\, x - \bar{x}\,)$ [+ constant] for $\bar{x} = -Q^{-1}q$ (**check**)
- $\bar{x}$ center of the ellipsoid, $y = x - \bar{x}$, $f_{\bar{x}}(y) = y^T Qy$ [+ constant]



- Along H_i: $f_i(\,\alpha\,) = f_{\bar{x}}(\,\alpha H_i\,) = \alpha^2 \lambda_i$ (**check**)
- $S(\, f_{\bar{x}}\,, 1\,) \equiv f_i(\,\alpha\,) = 1 \equiv \alpha = \sqrt{1/\lambda_i} \Longrightarrow$ $H_i \perp$ axes of $S(\, f_{\bar{x}}\,, 1\,)$, length $\sqrt{1/\lambda_i}$

$$Q = \begin{bmatrix} 6 & -2 \\ -2 & 6 \end{bmatrix} \;, \; H = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \;, \; \lambda = \begin{bmatrix} 8 \\ 4 \end{bmatrix}$$

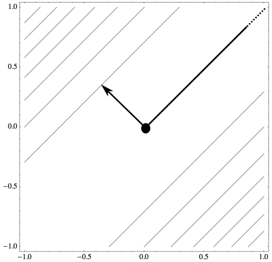
- Easy case: Q nonsingular $\equiv \lambda_i \neq 0 \, \forall i$ (regardless of the sign)
- Then $f(x) = \frac{1}{2}(x - \bar{x})^T Q(x - \bar{x})$ [+ constant] for $\bar{x} = -Q^{-1}q$ (**check**)
- $\bar{x}$ center of the ellipsoid, $y = x - \bar{x}$, $f_{\bar{x}}(y) = y^T Q y$ [+ constant]



- Along H_i: $f_i(\alpha) = f_{\bar{x}}(\alpha H_i) = \alpha^2 \lambda_i$ (**check**)
- $S(f_{\bar{x}}, 1) \equiv f_i(\alpha) = 1 \equiv \alpha = \sqrt{1/\lambda_i} \Longrightarrow$ $H_i \perp$ axes of $S(f_{\bar{x}}, 1)$, length $\sqrt{1/\lambda_i}$
- The smaller λ_i, the longer the axis

$$Q = \begin{bmatrix} 5 & -3 \\ -3 & 5 \end{bmatrix} \;,\; H = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \;,\; \lambda = \begin{bmatrix} 8 \\ 2 \end{bmatrix}$$

- Easy case: Q nonsingular $\equiv \lambda_i \neq 0 \, \forall i$ (regardless of the sign)
- Then $f(x) = \frac{1}{2}(x - \bar{x})^T Q(x - \bar{x})$ [+ constant] for $\bar{x} = -Q^{-1}q$ (**check**)
- $\bar{x}$ center of the ellipsoid, $y = x - \bar{x}$, $f_{\bar{x}}(y) = y^T Q y$ [+ constant]

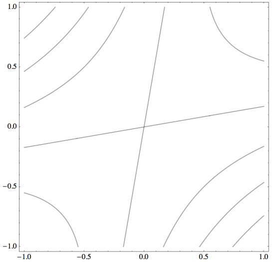


- Along H_i: $f_i(\alpha) = f_{\bar{x}}(\alpha H_i) = \alpha^2 \lambda_i$ (**check**)
- $S(f_{\bar{x}}, 1) \equiv f_i(\alpha) = 1 \equiv \alpha = \sqrt{1/\lambda_i} \Longrightarrow$ $H_i \perp$ axes of $S(f_{\bar{x}}, 1)$, length $\sqrt{1/\lambda_i}$
- The smaller λ_i, the longer the axis
- With $\lambda_i = 0$, "axis $\to \infty$" (but Q singular)

$$Q = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix} \ , \ H = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \ , \ \lambda = \begin{bmatrix} 8 \\ 0 \end{bmatrix}$$

Level sets of quadratic functions, easy case

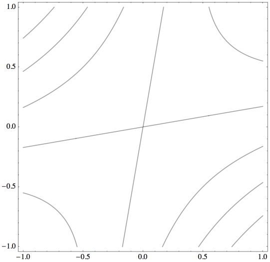
- Easy case: Q nonsingular $\equiv \lambda_i \neq 0 \, \forall i$ (regardless of the sign)
- Then $f(x) = \frac{1}{2}(x - \bar{x})^T Q(x - \bar{x})$ [+ constant] for $\bar{x} = -Q^{-1}q$ (**check**)
- $\bar{x}$ center of the ellipsoid, $y = x - \bar{x}$, $f_{\bar{x}}(y) = y^T Q y$ [+ constant]



- Along H_i: $f_i(\alpha) = f_{\bar{x}}(\alpha H_i) = \alpha^2 \lambda_i$ (**check**)
- $S(f_{\bar{x}}, 1) \equiv f_i(\alpha) = 1 \equiv \alpha = \sqrt{1/\lambda_i} \Longrightarrow$ $H_i \perp$ axes of $S(f_{\bar{x}}, 1)$, length $\sqrt{1/\lambda_i}$
- The smaller λ_i, the longer the axis
- With $\lambda_i = 0$, "axis $\rightarrow \infty$" (but Q singular)
- With $\lambda_i < 0$ sign reverses, no longer "axes"

$$Q = \begin{bmatrix} 3 & -5 \\ -5 & 3 \end{bmatrix} \quad , \quad H = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \quad , \quad \lambda = \begin{bmatrix} 8 \\ -2 \end{bmatrix}$$

- Easy case: Q nonsingular $\equiv \lambda_i \neq 0 \, \forall i$ (regardless of the sign)
- Then $f(\,x\,) = \frac{1}{2}(\,x - \bar{x}\,)^T Q(\,x - \bar{x}\,)$ [+ constant] for $\bar{x} = -Q^{-1}q$ (**check**)
- $\bar{x}$ center of the ellipsoid, $y = x - \bar{x}$, $f_{\bar{x}}(y) = y^T Qy$ [+ constant]



- Along H_i: $f_i(\,\alpha\,) = f_{\bar{x}}(\,\alpha H_i\,) = \alpha^2 \lambda_i$ (**check**)
- $S(\,f_{\bar{x}}\,,\,1\,) \equiv f_i(\,\alpha\,) = 1 \equiv \alpha = \sqrt{1/\lambda_i} \Longrightarrow$
  $H_i \perp$ axes of $S(\,f_{\bar{x}}\,,\,1\,)$, length $\sqrt{1/\lambda_i}$
- The smaller λ_i, the longer the axis
- With $\lambda_i = 0$, "axis $\rightarrow \infty$" (but Q singular)
- With $\lambda_i < 0$ sign reverses, no longer "axes"

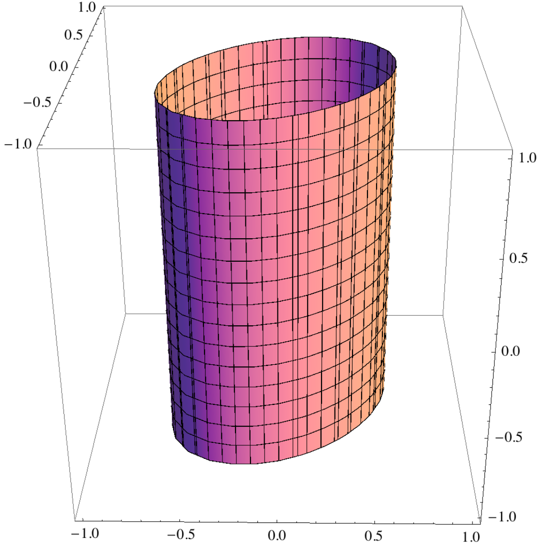
$$Q = \begin{bmatrix} 3 & -5 \\ -5 & 3 \end{bmatrix} \;,\; H = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \;,\; \lambda = \begin{bmatrix} 8 \\ -2 \end{bmatrix}$$

- $\forall i \, \lambda_i > 0 \equiv Q \succ 0 \Longrightarrow \bar{x}$ minimum of f (why?)
- $\exists \, i \, \lambda_i < 0 \equiv Q \not\succeq 0 \Longrightarrow f$ unbounded below (why?)

- Q singular $\equiv \exists\, \lambda_i = 0 \equiv \ker(\,Q\,) \neq \{\,0\,\}$
- $\mathbb{R}^n = row(\,Q\,) + ker(\,Q\,)$, $row(\,Q\,) \perp ker(\,Q\,)$
  $row(\,Q\,) \equiv$ subspace spanned by rows
  $ker(\,Q\,) \equiv$ subspace spanned by H_i with $\lambda_i = 0$
- $q = q_+ + q_0$, $q_+ \perp q_0$, where
  $q_+ \in row(\,Q\,) = row(\,-Q\,) \equiv \bar{x}^T(-Q) = q_+^T$ and $q_0 \in ker(\,Q\,) \equiv Qq_0 = 0$
- Then $f(\,x\,) = \frac{1}{2}(\,x - \bar{x}\,)^T Q(\,x - \bar{x}\,) + q_0 x$ [+ constant] (**check**)
- f is "truly quadratic" on $row(\,Q\,)$ and linear on $ker(\,Q\,)$
- Assume $Q \succeq 0$: f has minimum $\iff$ $q_0 = 0 \;\equiv$
  $\bar{x}^T(-Q) = q^T \;\equiv\; Q\bar{x} + q = 0 \;\equiv\; \nabla f(\,\bar{x}\,) = 0$ has solution
- First example of first-order (global) optimality condition, more to come
- Linear algebra is crucial for optimization

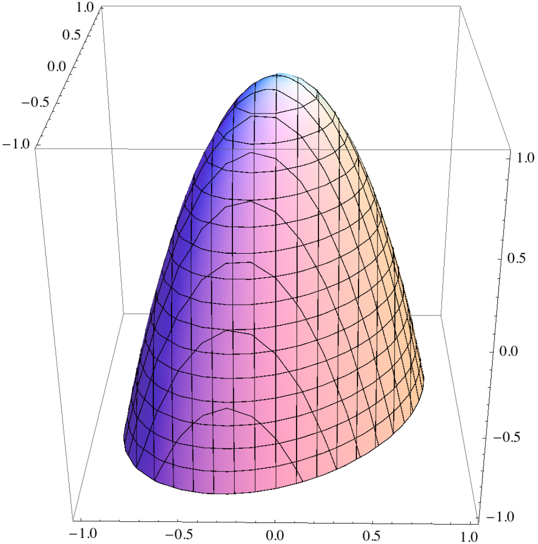
$$Q = \begin{bmatrix} 6 & -2 & 0 \\ -2 & 6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad , \quad q = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} H = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad , \quad \lambda = \begin{bmatrix} 8 \\ 4 \\ 0 \end{bmatrix}$$

$S(f, 1)$



$$Q = \begin{bmatrix} 6 & -2 & 0 \\ -2 & 6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad , \quad q = \begin{bmatrix} 0 \\ 0 \\ {\color{red}1} \end{bmatrix} H = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad , \quad \lambda = \begin{bmatrix} 8 \\ 4 \\ 0 \end{bmatrix}$$

$S(f, 1)$



Outline

- Optimization difficult/impossible in general
- Need conditions to make it possible:
  - X closed, otherwise x_* may be on the unreachable boundary
  - possibly X compact $\Longrightarrow$ every sequence has accumulation point (but not always possible, $f_* = -\infty$ happens)
  - f (lower semi-)continuous, otherwise can "jump away" on would-be x_*
  - some sort of derivative information to tell the way to x_*
- The more derivatives you have, the better
- Derivatives $\Longrightarrow$ first- and second-order model
- f "complicated", model looks like f (close to x) and simple
- Fundamental concept we will use all the time

- Boyd, Vandenberghe “Convex optimization” Appendix A
- Bazaraa, Sherali, Shetty “Nonlinear programming” Appendix A1, A3, A4
- Nocedal, Wright “Numerical Optimization” Appendix A2
- Google + Wikipedia; e.g.
  `https://mathinsight.org/differentiability_multivariable_theorem`