Elements of convex analysis and convex optimization part II - Convex functions and convex problems

March 15, 2024

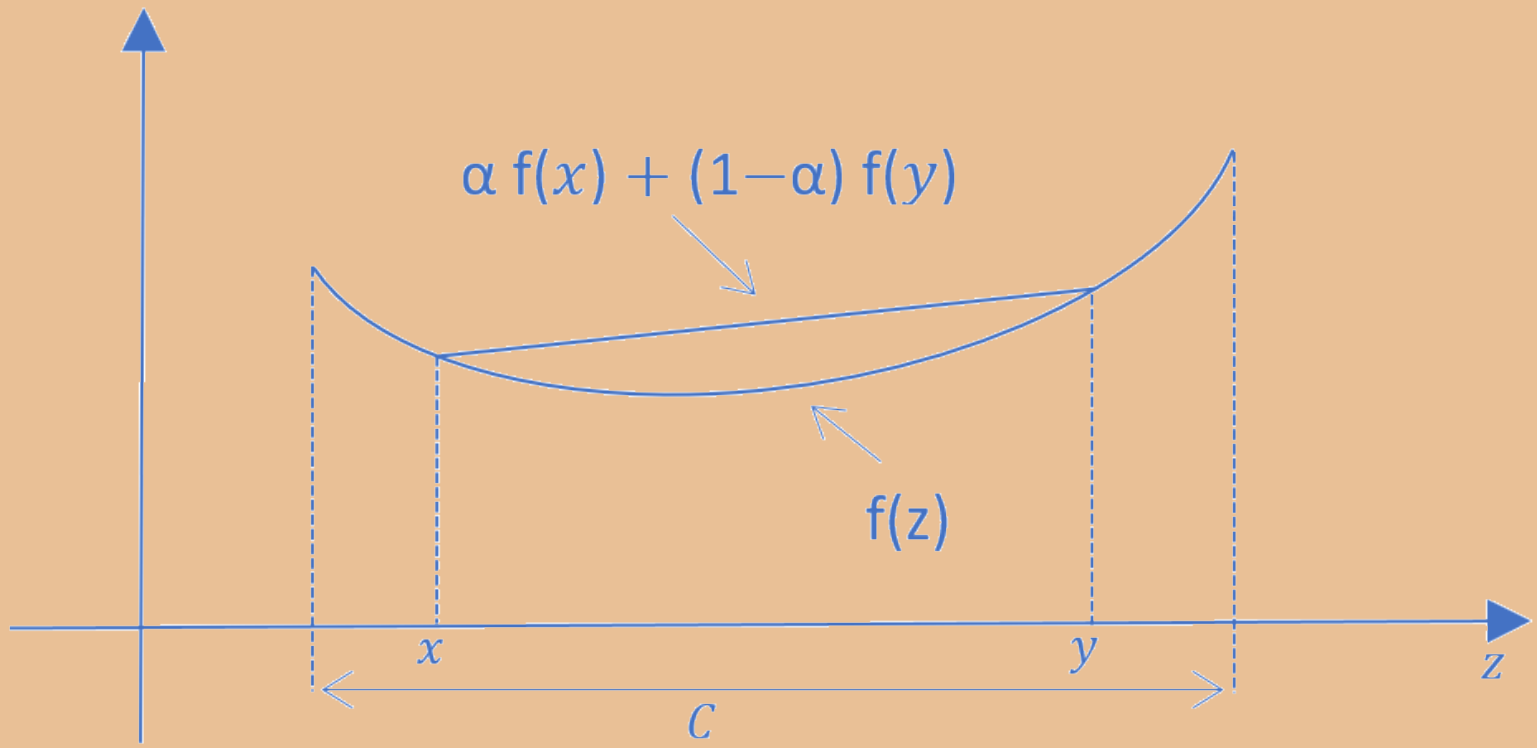
Part II - Convex functions

Definition

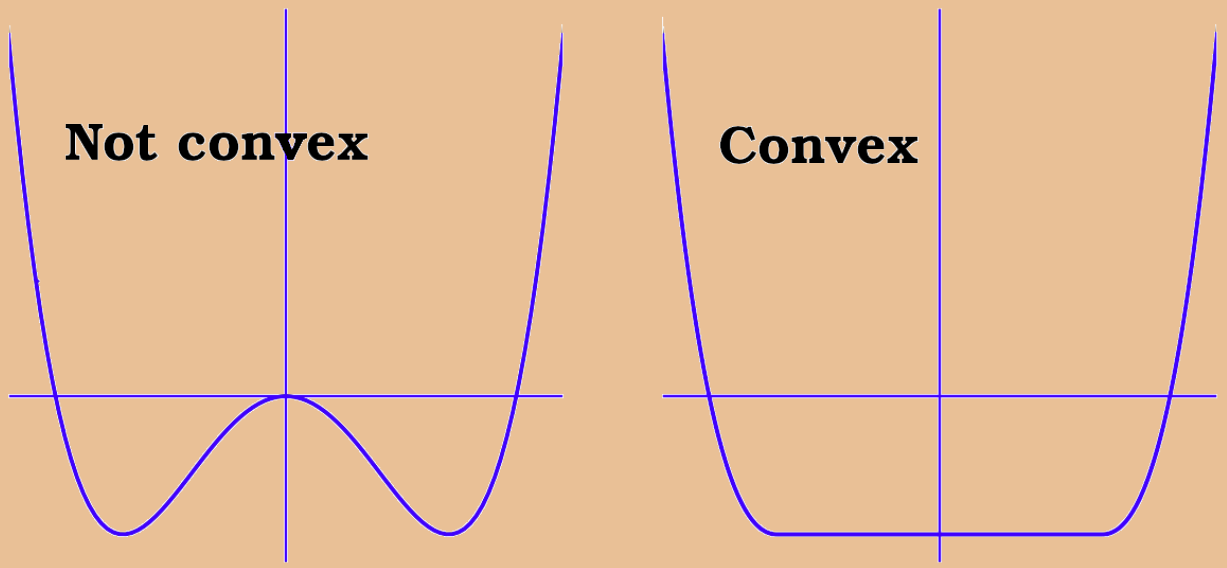
Let $S \subseteq \mathbb{R}^n$ a convex set. A function $f : S \to \mathbb{R}$ is convex over S if for all $x, y \in S$ we have

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

If $<$ instead of $\leq$ then strict convex, in that case $\alpha \in (0,1)$ and $x \neq y$.



Definition



Recall: Convexity is crucial to optimization, because any local minimum is also a global minimum.

Note: Strict convexity implies that the global minimum is unique.

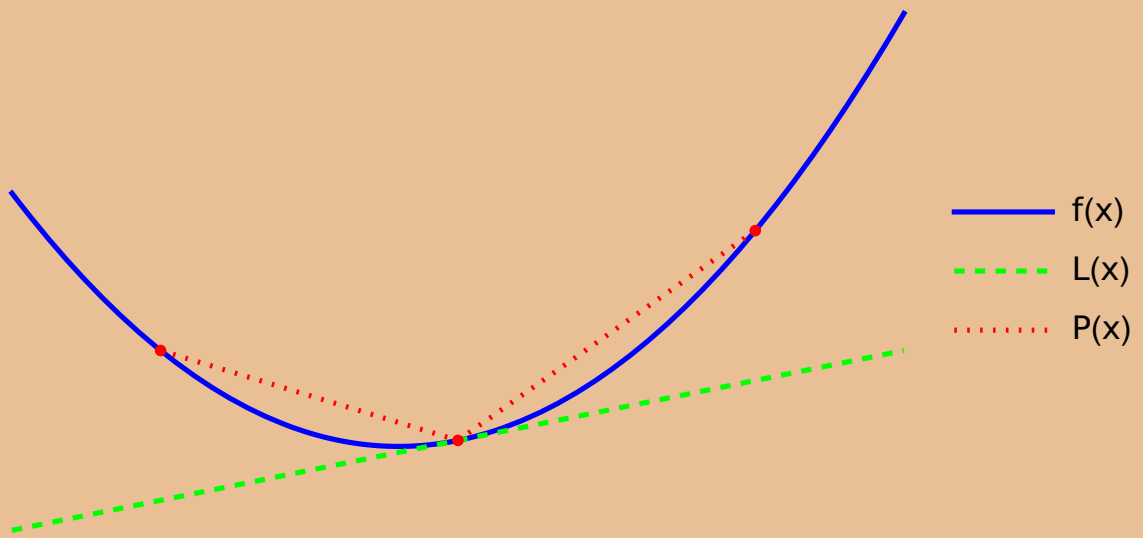
Note 2: There is a stronger concept of convexity called m-strongly convex (more on this later time permitting).

Continuity of Convex Functions

Proposition

Let $f : S \to \mathbb{R}$ convex. Let $x_0 \in \text{int}(S)$. Then f is continuous at x_0.

Proof sketch: Let $f : (a, b) \to \mathbb{R}$ convex. Let $c \in (a, b)$. Let $L(x)$ be a linear function whose graph is a tangent line for f at c, and let P be a piecewise linear function consisting of two chords to the graph of f meeting at c (see figure). Then $L \leq f \leq P$ in a neighborhood of c, and $L(c) = f(c) = P(c)$. Since L and P are continuous at c, it follows from the Squeeze Theorem for functions that f is also continuous at c.



Proposition

A convex function is differentiable almost everywhere (that is, except in a countable number of points)

Examples on $\mathbb{R}$

Convex

- affine: $ax+b$, for any $a, b \in \mathbb{R}$.
- exponential: e^{ax}, for any $a \in \mathbb{R}$.
- powers: x^α on $\mathbb{R}_{>0}$, for $\alpha \geq 1$ or $\alpha \leq 0$.
- powers of absolute value: $|x|^p$ on $\mathbb{R}$, for $p \geq 1$.
- negative entropy: $x \ln(x)$ on $\mathbb{R}_{>0}$.

Concave

- affine: $ax+b$, for any $a, b \in \mathbb{R}$.
- powers: x^α on $\mathbb{R}_{>0}$, for $0 \leq \alpha \leq 1$.
- logarithm: $\ln(x)$ on $\mathbb{R}_{>0}$. Important to recall since it is used many times.

Check Mathematica command:

```
FunctionConvexity[{x^3,x>=0}, x]
```

Examples on $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$

In $\mathbb{R}^n$:

- affine: $f(x) = a^T x + b$, where $x, a \in \mathbb{R}^n$, $b \in \mathbb{R}$.
- any norm: $f(x) = \|x\|$.
- quadratic form $f(x) = x^T Q x$ is convex iff Q is psd.

In $\mathbb{R}^{m\times n}$

- Affine:

$$f(X) = \text{Trace}\left(A^T X\right) + b = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}X_{ij} + b.$$

- Spectral norm: largest singular value

$$f(X) = \|X\|_2 = \sigma_{max}(X) = (\lambda_{max}(X^T X))^{1/2}.$$

Examples on $\mathbb{R}^n$

Why $f(x) = a^T x + b$ is convex? For $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$ we have

$$f(\alpha x + (1-\alpha)y) = a^\top(\alpha x + (1-\alpha)y) + b = \alpha(a^\top x + b) + (1-\alpha)(a^\top y + b)$$

which is $\alpha f(x) + (1-\alpha) f(y)$. The same proof applies to $-f(x)$ and therefore affine functions are convex and concave, and are the only functions in $\mathbb{R}^n$ to be so.

Why $f(x) = \|x\|$ is convex? Apply homogeneous and triangle inequality properties to get

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

By definition of convexity

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

then $f(x) = \|x\|$ is convex.

Examples on $\mathbb{R}^n$

Quadratic form $f(x) = x^T Q x$ is convex iff Q is psd.

Consider $\alpha \in [0,1]$ and let $\tilde{\alpha} = 1-\alpha$. By definition of convex function we have

$$(\alpha x+\tilde{\alpha} y)^T Q(\alpha x+\tilde{\alpha} y) \leq \alpha x^T Q x + \tilde{\alpha} y^T Q y\,.$$

Simplify LHS to get

$$\alpha^2 x^T Q x + \tilde{\alpha}^2 y^T Q y + \alpha\tilde{\alpha} x^T Q y + \tilde{\alpha}\alpha y^T Q x \leq \alpha x^T Q x + \tilde{\alpha} y^T Q y\,.$$

Noting $\tilde{\alpha}\alpha = \alpha-\alpha^2$ and $\tilde{\alpha}\alpha = \tilde{\alpha}-\tilde{\alpha}^2$ we get

$$0 \leq \alpha\tilde{\alpha} x^T Q x + \alpha\tilde{\alpha} y^T Q y - \alpha\tilde{\alpha} x^T Q y - \alpha\tilde{\alpha} y^T Q x$$

Finally,

$$x^T Q x + y^T Q y - x^T Q y - y^T Q x \geq 0$$

or

$$(x-y)^T Q(x-y) \geq 0\,.$$

That is, Q is psd. Note we proved both directions at the same time.

Jensen's inequality

The basic property characterizing a convex function f is that the convex interpolation between two points $f(x)$ and $f(y)$ always overestimates the the function value $f(\alpha x + (1-\alpha)y)$. This property can be generalized to convex combinations of any number of points:

Proposition (Jensen's inequality)

Let $f : S \to \mathbb{R}$ convex, for $S \subseteq \mathbb{R}^n$ convex. Then for any $x_1, x_2, ..., x_m \in S$ and $\alpha \in \Delta_m$ we have

$$f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \leq \sum_{i=1}^{m} \alpha_i f(x_i)\,.$$

Note: that Jensen's inequality can be generalized to infinite sums and integrals.

Proof. By induction on m. Basis case: $m = 2$ is verified since f is convex over S. Next we prove the induction step.

Proof

Induction step:

IH: $f\left(\sum_{i=1}^{k} \alpha_i x_i\right) \leq \sum_{i=1}^{k} \alpha_i f(x_i)$, for $k \leq m-1$.

IT: $f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \leq \sum_{i=1}^{m} \alpha_i f(x_i)$.

By the induction hypothesis, we assume that the inequality holds for $m-1$. For m, we can write:

$$f\left(\sum_{i=1}^{m} \alpha_i x_i\right) = f\left((1-\alpha_m)\sum_{i=1}^{m-1} \frac{\alpha_i}{1-\alpha_m} x_i + \alpha_m x_m\right)$$

Let $y = \sum_{i=1}^{m-1} \frac{\alpha_i}{1-\alpha_m} x_i$. Since f is convex:

$$f\left((1-\alpha_m)y + \alpha_m x_m\right) \leq (1-\alpha_m) f\left(y\right) + \alpha_m f(x_m)$$

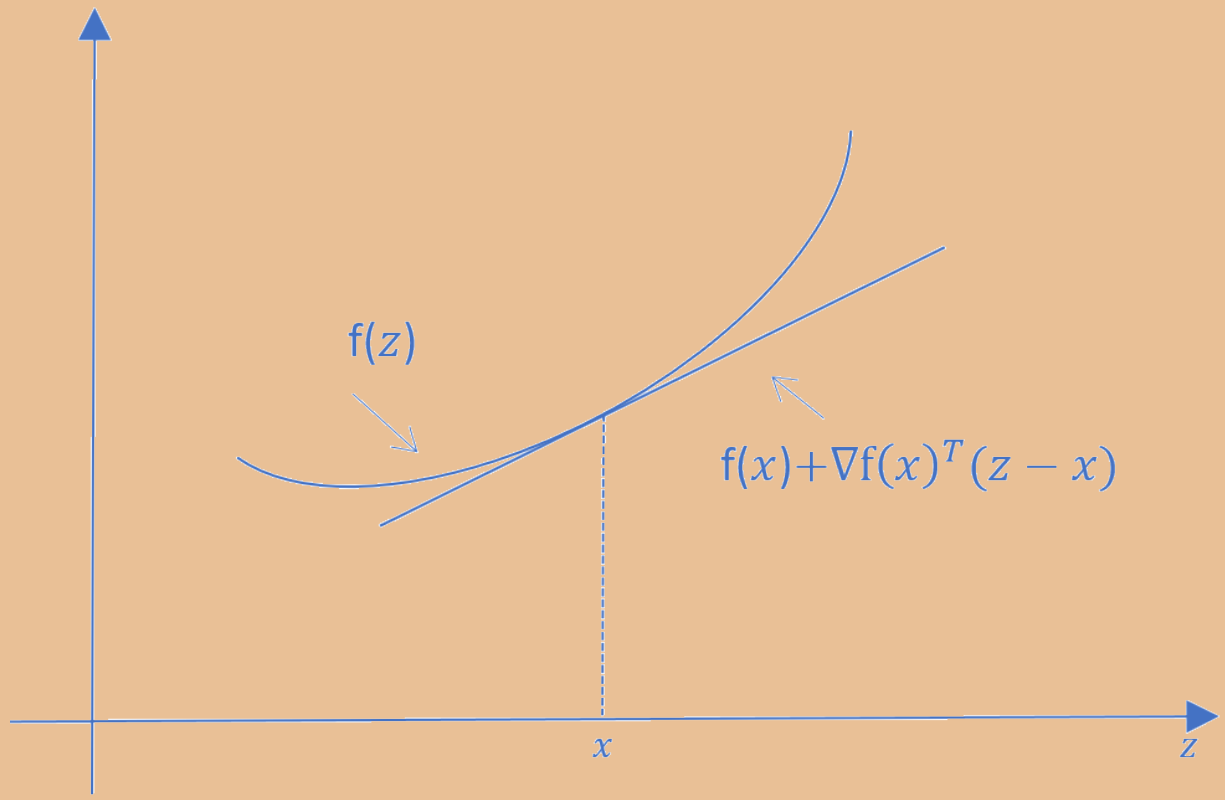
We complete the induction step using the convexity of f:

$$(1-\alpha_m) f\left(\sum_{i=1}^{m-1} \frac{\alpha_i}{1-\alpha_m} x_i\right) + \alpha_m f(x_m) \leq \sum_{i=1}^{m} \alpha_i f(x_i)$$

First order characterization

Convex functions are not necessarily differentiable, but in case they are, we have the following characterization: the tangent hyperplanes of convex functions always underestimate the function.

That is, a convex function f with $\mathrm{dom}\,(f) = S$ has a supporting hyperplane for any $x \in S$.



First order characterization

Proposition (First order inequality)

Let $f : S \to \mathbb{R}$, for $S \subseteq \mathbb{R}^n$ convex and $f \in C^1(S)$. Then f is convex over S iff for all $x, y \in S$ we have

$$f(x) + \nabla f(x)^\top (y - x) \leq f(y)\,.$$

Proof. $\Longrightarrow$: Since f convex we have:

$$\begin{aligned}
f(\alpha x + (1-\alpha)y) &\leq \alpha f(x) + (1-\alpha) f(y) \\
f(y + \alpha(x-y)) &\leq \alpha(f(x) - f(y)) + f(y) \\
f(y + \alpha(x-y)) - f(y) &\leq \alpha(f(x) - f(y)) \\
\frac{f(y + \alpha(x-y)) - f(y)}{\alpha} &\leq f(x) - f(y)\,,
\end{aligned}$$

then take limit when $\alpha \to 0$ and we have the inequality. $\Longleftarrow$: is just algebra (do it).

Note that the function is strictly convex iff the gradient inequality is satisfied with strict inequality for all $z \neq x$.

Note the gradient inequality states that for convex functions, the tangent hyperplane supports the function on every point in the domain.

First order characterization

A direct result of the first order inequality is that the first order optimality condition $\nabla f(x^*) = 0$ is sufficient for x^* to be a global minimum.

Proposition (Sufficiency of stationarity under convexity)

Let $f : S \to \mathbb{R}$ convex, for $S \subseteq \mathbb{R}^n$ convex and $f \in C^1(S)$. Suppose $\nabla f(x^*) = 0$ for some $x^* \in S$. Then x^* is a global minimizer of f over S.

Proof. Indeed, plugging x^* we have

$$f(z) \geq f(x^*) + \nabla f(x^*)^\top (y - x^*), \quad \text{for all } y \in S\,,$$

and using $\nabla f(x^*) = 0$ we get $f(y) \geq f(x^*)$ for all $y \in S$. □

Note the condition is sufficient but not necessary. That is, the fact that x^* is a global minimizer over S does not imply $\nabla f(x^*) = 0$.

However, if $C = \mathbb{R}^n$ the condition is also sufficient (see next slide).

First order characterization

Proposition (Necessity and sufficiency of stationarity under convexity)

Let $f : \mathbb{R}^n \to \mathbb{R}$ convex and $f \in C^1$. Then $\nabla f(x^*) = 0$ iff x^* is a global minimizer of f.

Using the gradient inequality we can now establish the conditions under which a quadratic function is convex or strictly convex.

Convexity and strict convexity of quadratic functions

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = x^\top Q x + 2b^\top x + c$ where Q is symmetric, $x, b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then f is convex (strictly) iff Q is psd (pd).

Proof. Start with first order inequality

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)\,, \quad \forall x, y \in \mathbb{R}^n\,.$$

Plug in f(x) and $\nabla f(x) = 2(Qx + b)$:

$$y^\top Q y + 2b^\top y + \cancel{c} \geq x^\top Q x + 2b^\top x + \cancel{c} + 2(Qx + b)^\top (y - x)\,, \quad \forall x, y \in \mathbb{R}^n$$

iff

$$y^\top Q y - x^\top Q x + \cancel{2b^\top (y - x)} - 2(Qx + \cancel{b})^\top (y - x) \geq 0\,, \quad \forall x, y \in \mathbb{R}^n$$

$$y^\top Q y - x^\top Q x - 2x^\top Q y + 2x^\top Q x \geq 0\,, \quad \forall x, y \in \mathbb{R}^n$$

$$y^\top Q y + x^\top Q x - 2x^\top Q y \geq 0\,, \quad \forall x, y \in \mathbb{R}^n\,.$$

Examples:

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

Side note: Monotonicity of the gradient

Another first order characterization of convexity is the monotonicity property of the gradient. In $\mathbb{R}$, this means that the derivative is nondecreasing. In $\mathbb{R}^n$ we have

Proposition (Monotonicity of the gradient)

Let $f : S \to \mathbb{R}$, for convex $S \subseteq \mathbb{R}^n$ and $f \in C^1(S)$. Then f is convex over S iff

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq 0\,, \quad \forall x, y \in S\,.$$

Proof.

($\Longrightarrow$) by using the gradient inequality.

($\Longleftarrow$) Let $g(t) := f(x + t(y-x))$ for $t \in [0, 1]$.

Note that $x + t(y-x) = ty + (1-t)x$ for $t \in [0, 1]$.

Second order characterization

When $f \in C^2$, convexity can be characterized by the Hessian matrix (recall a function is convex if never "curves down"):

Proposition (Second order characterization of convexity (necessary and sufficient))

Let $f : S \to \mathbb{R}$, for convex $S \subseteq \mathbb{R}^n$ and $f \in C^2(S)$. Then f is convex over S iff $\nabla^2 f(x)$ is psd for all $x \in S$.

We also have the corresponding result for the strict convex case:

Proposition (Sufficient second order condition for strict convexity)

Let $f : S \to \mathbb{R}$, for convex $S \subseteq \mathbb{R}^n$ and $f \in C^2(S)$. Suppose $\nabla^2 f(x)$ is pd for all $x \in S$. Then f is strictly convex over S.

Note that pd of the Hessian is a sufficient condition, but not necessary. For example $f(x) = x^4$ is strictly convex but the second derivative is zero at $x = 0$.

Examples

Show convexity of the following functions using first and second order characterizations:

- **Easy**:
  - x^2 $(x \in \mathbb{R})$;
  - e^{ax} $(x, a \in \mathbb{R})$;
  - $-\ln(x)$ (for $x > 0$);
  - $a\ln(x)$ (for $x > 0$, $a \in \mathbb{R}$).
- **Quadratic function:** $f(x) = x^\top Q x + b^\top x + c$ convex if Q pd.
- **Least-squares objective:** $f(x) = \|Ax - b\|_2^2$,
  $\nabla f(x) = 2A^\top(Ax - b), \quad \nabla^2 f(x) = 2A^\top A$ convex for any A.
- **Log-sum-exp**: $f(x) = \ln(e^{x_1} + e^{x_2} + .. + e^{x_n})$, $x \in \mathbb{R}^n$.
- **Quadratic-over-linear**: $g(x, y) = \frac{x^2}{y}$, $x, y \in \mathbb{R}$, $y > 0$.

  $\nabla^2 f(x, y) = \frac{2}{y^3} \begin{pmatrix} y \\ -x \end{pmatrix} \begin{pmatrix} y \\ -x \end{pmatrix}^\top$.
- **Geometric mean**: $h(x) = (\prod_{k=1}^n x_k)^{1/n}$ on $\mathbb{R}^n_{>0}$ is concave (similar proof as for log-sum-exp). Left as a task.

Log-sum-exp

To show log-sum-exp is convex, we use second order characterization. That is, we prove that $v^\top \nabla^2 f(x) v \geq 0$ for all $v \in \mathbb{R}^n$. We compute the gradient and Hessian:

$$\frac{\partial f}{\partial x_i} = \frac{e^{x_i}}{e^{x_1} + .. + e^{x_n}} = \frac{e^{x_i}}{s}\,.$$

Define $s := e^{x_1} + .. + e^{x_n}$. Moreover,

$$\frac{\partial f}{\partial x_i \partial x_j} = -\frac{e^{x_i} e^{x_j}}{s^2}\,, \quad \text{for } i \neq j\,,$$

$$\frac{\partial f}{\partial x_i \partial x_j} = \frac{e^{x_i} s - e^{x_i} e^{x_i}}{s^2}, \quad \text{for } i = j\,.$$

Now the goal is to write $\nabla^2 f(x)$ in matrix form. Define the vector

$$w := \begin{pmatrix} \frac{e^{x_1}}{s} \\ \vdots \\ \frac{e^{x_n}}{s} \end{pmatrix}\,.$$

Observe w is a vector of positive numbers that sum to 1.

We have

$$\nabla^2 f(x) = \operatorname{diag}(w) - ww^\top .$$

The first term can be also written

$$\operatorname{diag}(u) = \sum_{i=1}^{n} e_i u_i e_i^\top ,$$

where e_i is the i-th component of the canonical $\mathbb{R}^n$ basis.

Now

$$\begin{aligned} v^\top \nabla^2 f(x) v &= v^\top \operatorname{diag}(w) v - v^\top uu^\top v \\ &= \sum_{i=1}^{n} v_i w_i v_i - (v^\top w)^2 \\ &= \sum_{i=1}^{n} v_i^2 w_i - (v^\top w)^2 \end{aligned}$$

Question:

$$\sum_{i=1}^{n} v_i^2 w_i \geq (v^\top w)^2 \, ?$$

Use C-S inequality. Recall: For any vectors x, y we have

$$\|x\|\|y\| \geq |x^\top y|$$
$$\|x\|\|y\| \geq (x^\top y)^2 .$$

Let

$$s_i := v_i \sqrt{w_i}$$
$$t_i := \sqrt{w_i}$$

Then,

$$\begin{aligned}(v^\top w)^2 = (s^\top t)^2 &\leq \|s\|^2 \|t\|^2 \\ &= (\sum_i v_i^2 w_i)(\sum_i w_i) \\ &= (\sum_i v_i^2 w_i)\, 1 \\ &= \sum_i v_i^2 w_i\end{aligned}$$

Since $\sum_i v_i^2 w_i \geq (v^\top w)^2$ we have $\nabla^2 f(x)$ is psd.

Example - softmax/softargmax

An interesting fact about the log-sum-exp function (denoted LSE(x)) is that it can be used as a "softmax", and its gradient is called the "softargmax" function which is also convex (Task: prove it):

$$\nabla \mathrm{LSE}(x) = \mathrm{softargmax}(x) = \begin{pmatrix} \frac{e^{x_1}}{e^{x_1}+\cdots+e^{x_n}} \\ \vdots \\ \frac{e^{x_n}}{e^{x_1}+\cdots+e^{x_n}} \end{pmatrix}.$$

The softmax function converges to a one-hot representation of the output (assuming there is a unique maximum arg):

$$\arg\max(z_1, \ldots, z_n) = (y_1, \ldots, y_n) = (0, \ldots, 0, 1, 0, \ldots, 0),$$

where the output coordinate $y_i = 1$ iff z_i is the arg max of $(z_1, \ldots, z_n)$, meaning z_i is the unique maximum value of $(z_1, \ldots, z_n)$. For example, in this encoding

$$\arg\max(1, 5, 10) = (0, 0, 1),$$

Example - softmax/softargmax

since the third argument is the maximum. Formally, each maximum value is assigned a value of $\frac{1}{k}$, where k is the count of arguments that assume the maximum value. For instance,

$$\text{argmax}(1, 5, 5) = (0, \frac{1}{2}, \frac{1}{2}),$$

since there are two arguments which are equal to the maximum value of 5. Suppose we have a set of n numbers $x_1, x_2, ..., x_n$ and we want to compute the maximum of these numbers. One way to do this is to use the LSE(x) function with a scaling parameter t:

$$\max(x_1, .., x_n) = \lim_{t \to \infty} \frac{1}{t} \ln \left(\sum_{i=1}^{n} \exp(t\, x_i) \right) .$$

See the following example (for $n = 2$) `https://www.wolframcloud.com/env/pguerra0/SoftMax.nb`

Example - softmax/softargmax

The following extension to log-sum-exp is also convex and smoothly approximates $\max_{i=1,\ldots,k}(a_i^T x + b_i)$, where $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$:

$$g(x) = \log\left(\sum_{i=1}^{k} e^{a_i^\top x + b_i}\right).$$

To show that g is convex, we just observe that g is an affine composition involving f.

Operations preserving convexity

Proposition (Linear combinations of convex functions are convex)

1. *Let $f : S \to \mathbb{R}$ be convex defined over a convex $S \subseteq \mathbb{R}^n$ and let $\alpha \geq 0$. Then αf is convex over S.*
2. *Let $f_1, f_2, ..., f_p$ be convex over a convex set $S \subseteq \mathbb{R}^n$. Then $f_1 + f_2 + ... + f_p$ is convex over S.*

Proof. next slide.

Note: In the second property the functions f_i may be defined over different convex sets S_i. Then the domain of f is taken to be $S := \bigcap_i S_i$.

Operations preserving convexity

1. Define $g(x) := \alpha f(x)$. Let $x, y \in S$ and $\lambda \in [0,1]$. Then

$$\begin{aligned} g(\lambda x + (1-\lambda)y) &= \alpha f(\lambda x + (1-\lambda)y) \\ &\leq \alpha\lambda f(x) + \alpha(1-\lambda)f(y) \\ &= \lambda g(x) + (1-\lambda)g(y) \end{aligned}$$

2. Let $x, y \in S$ and $\lambda \in [0,1]$. Since f_i is convex, we have for each i=1, .., p

$$f_i(\lambda x + (1-\lambda)y) \leq \lambda f_i(x) + (1-\lambda)f_i(y)\,.$$

Summing over i we get

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$$

where $g = f_1 + f_2 + ... + f_p$.

Operations preserving convexity

Proposition (Affine composition)

Let $f : S \to \mathbb{R}$ convex, for $S \subseteq \mathbb{R}^n$ convex. Let $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$. Then the function $g : \mathbb{R}^m \to \mathbb{R}$ defined by

$$g(x) := f(Ax + b)$$

is convex over the convex set $S = \{x \in \mathbb{R}^m : Ax + b \in S\}$.

Proof. First step is to notice that D is a convex set since it is the inverse map of a convex set $D = A^{-1}(S - b)$ (see slides about convex sets).

Note: Affine composition can be seen as a change of variables.

Operations preserving convexity

Example

Consider

$$\varphi(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^2 + x_1 - x_2 + e^{x_1+x_2}\,.$$

To prove φ is convex over $\mathbb{R}^2$, note that $\varphi = \varphi_1 + \varphi_2$, where

$$\begin{aligned}\varphi_1(x_1, x_2) &= x_1^2 + 2x_1x_2 + 3x_2^2 + x_1 - x_2\,,\\ \varphi_2(x_1, x_2) &= e^{x_1+x_2}\end{aligned}$$

The function φ_1 is convex since it is a quadratic function with matrix $Q = \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}$ which is pd since $Tr(A) = 4 > 0$, $det(A) = 2 > 0$. The function φ_2 is also convex since it is constructed by making the linear change of variables $t = x_1 + x_2$ in the one-dimensional $f(t) = e^t$. Here $A = (1, 1)$, $x = (x_1, x_2)^\top$, and $b = 0$.

Operations preserving convexity

Example

The function $\varphi(x_1, x_2) = x_1 + e^{x_2}$ is convex over $\mathbb{R}^2$ as a sum of convex functions: The function x_1 is convex since it is constructed by making the linear change of variables $t = x_1$ in the one-dimensional $\varphi(t) = t$. Same for the function e^{x_2}.

Note that in this example as in the previous one, the affine composition result is required to prove that convexity is preserved. It is not immediate that a $\mathbb{R}$ to $\mathbb{R}$ function is convex when the argument is in $\mathbb{R}^2$.

Example

The function $\varphi(x_1, x_2) = -ln(x_1 x_2)$ is convex over $\mathbb{R}^2_{>0}$ since it can be written as

$$f(x_1, x_2) = -ln(x_1) - ln(x_2),$$

and the convexity of $-ln(x_1)$ and $-ln(x_2)$ follows from the convexity of $\varphi(t) = -ln(t)$ over $\mathbb{R}_{>0}$.

Operations preserving convexity

Examples

- *log barrier for linear inequalities:*

$$f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^\top x), \quad \operatorname{dom} f = \{x \mid a_i^\top x < b_i,\ i = 1, \ldots, m\}$$

- *Any norm of affine function:* $f(x) = \|Ax + b\|$.

Operations preserving convexity

In general, convexity is not preserved under composition of convex functions. For example, let $f(z) = z^2$ and $h(t) = t^2 - 4$. Then f and h are convex. However, their composition

$$g(t) = f(h(t)) = (t^2 - 4)^2$$

is not convex, because $g''(t) = 12t^2 - 16$ takes positive and negative values.

Despite of this, under certain conditions, we can convexity is preserved (see next slide).

Operations preserving convexity

Proposition (Composition of a nondecreasing convex function with a convex function)

Let $h : S \to \mathbb{R}$ convex, for $S \subseteq \mathbb{R}^n$ convex. Let $f : I \to \mathbb{R}$ be a one-dimensional nondecreasing convex function over the interval $I \subseteq \mathbb{R}$. Assume that the image of SS under h is contained in I, $h(S) \subseteq I$. Then the composition of f with h defined by

$$g(x) := f(h(x)), \quad x \in S,$$

is a convex function over S.

Note nondecreasing means $x > y \implies f(x) \geq f(y)$ for all $x, y \in I$.

Proof. Let $x, y \in S$ and $\alpha \in [0, 1]$. Then,

$$\begin{aligned} g(\alpha x + (1-\alpha)y) &= f(h(\alpha x + (1-\alpha)y)) \\ &\leq f(\alpha h(x) + (1-\alpha)h(y)) \quad \text{(convexity of h and nondecreasing f)} \\ &\leq \alpha f(h(x)) + (1-\alpha)f(h(y)) \quad \text{(convexity of f)} \\ &= \alpha g(x) + (1-\alpha)g(y)\,. \quad \text{(definition of g)} \end{aligned}$$

Operations preserving convexity

Example

The function $g(x) = e^{\|x\|^2}$ *is convex since it can be represented as* $g(x) = f(h(x))$, *where* $f(t) = e^t$ *is a nondecreasing convex function and* $h(x) = \|x\|^2$ *is a convex function.*

Example

The function $g(x) = (\|x\|^2 + 1)^2$ *is a convex function over* $\mathbb{R}^n$ *since it can be represented as* $g(x) = f(h(x))$, *where* $f(t) = t^2$ *and* $h(x) = \|x\|^2 + 1$. *Both* f *and* h *are convex, but note that* h *is not a nondecreasing function on* $\mathbb{R}^n$. *However, the image of* $\mathbb{R}^n$ *under* f *is the interval* $[1, \infty)$ *on which the function* h *is nondecreasing. Consequently, the composition* $g(x) = f(h(x))$ *is convex.*

General composition with scalar functions

Composition of $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$

$$f(x) = h(g(x)) .$$

Then f is convex if $\begin{cases} g \text{ convex}, h \text{ convex and nondecreasing} \\ g \text{ concave}, h \text{ convex and nonincreasing} . \end{cases}$

Examples

- $h(x) = x^2$ *is convex but not decreasing for all* x*, so the composition with any convex* g *is not guaranteed to be convex. Note that* $h(ax+b)$ *is convex since is the composition with affine function.*
- $\exp(g(x))$ *is convex if* g *is convex.*
- $1/g(x)$ *is convex if* g *is concave and positive.*

General composition with vector functions

Composition of $g : \mathbb{R}^n \to \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}$:

$$f(x) = h(g(x)) = h(g_1(x), g_2(x), \cdots, g_k(x)) .$$

Then f is convex if

$$\begin{cases} g_i \text{ convex}, h \text{ convex and nondecreasing in each argument} \\ g_i \text{ concave}, h \text{ convex and nonincreasing in each argument} . \end{cases}$$

(Same as before but componentwise).

Example

- $\ln(\sum_{i=1}^m \exp(g_i(x)))$ *is convex if* g_i *are convex.*

Check `https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf` for more on composition (not required).

Operations preserving convexity

Proposition (Pointwise maximum of convex functions)

Let $f_1, .., f_p : S \to \mathbb{R}$ be p convex functions over the convex set $S \subseteq \mathbb{R}^n$. Then the maximum function

$$f(x) := \max_{i=1,..p} f_i(x)$$

is a convex function over S.

Proof. Let $x, y \in S$ and let $\alpha \in [0, 1]$. Then,

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &= \max_{i=1,..p} f_i(\alpha x + (1-\alpha)y) \\ &\leq \max_{i=1,..p} \{\alpha f_i(x) + (1-\alpha) f_i(y)\} \quad \text{(convexity of } f_i\text{)} \\ &\leq \max_{i=1,..p} f_i(x)) + (1-\alpha) \max_{i=1,..p} f_i(y)) \quad (*) \\ &= \alpha f(x) + (1-\alpha) f(y) \quad \text{(definition of f).} \end{aligned}$$

Operations preserving convexity

The inequaity $(*)$ follows from $\{a_i\}_{i=1}^p$, $\{b_i\}_{i=1}^p$ one has

$$\max_{i=1,..p}(a_i + b_i) \leq \max_{i=1,..p} a_i + \max_{i=1,..p} b_i \,.$$

Note if f_i has different convex domains S_i then f has convex domain on $\bigcap_i S_i$.

Example

Let $f(x) = \max\{x_1, x_2, ..., x_n\}$.
Then f *is convex since it is the maximum of* n *linear (hence convex) functions* $f_i(e_i^\top x) = x_i$, *for* $i = 1, .., p$. *In each* f_i *we use the property of affine composition.*

What about $g(x) = \min\{x_1, x_2, ..., x_n\}$? It is very easy to construct a counterexample so in general is not convex.

Operations preserving convexity

Example (sum of the k largest values)

Given a vector $x = (x_1, x_2, ..., x_n)^\top$. Let $x_{(i)}$ denote the i-th largest value in x. In particular $x_{(1)} = \max\{x_1, x_2, ..., x_n\}$ and $x_{(n)} = \min\{x_1, x_2, ..., x_n\}$.

We proved before that $h(x) = x_{(1)}$ is convex and we saw $h(x) = x_{(n)}$ is not convex in general. The same happens with the function $x_{(i)}$, for $i = 2, .., n$ is is not convex in general.

Despite of this, the sum of the k largest

$$h_k(x) = x_{(1)} + x_{(2)} + ... + x_{(k)}\,,$$

is convex. To show this, note that h_k can be rewritten as

$$h_k(x) = \max\{x_{i_1} + x_{i_2} + ... + x_{i_k} : i_1, i_2, ..., i_k \in \{1, 2, ..., n\} \text{ are different}\}\,,$$

so that h_k, as a maximum of linear (hence convex) functions, is a convex function.

Operations preserving convexity

When you minimize a convex function over one of its arguments, the resulting function remains convex with respect to the other argument.

Proposition (Partial minimization)

Let $f : S \times D \to \mathbb{R}$ be a convex function defined over the set $S \times D$ where $S \subseteq \mathbb{R}^m$ and $D \subseteq \mathbb{R}^n$ are convex sets. Let

$$g(x) = \min_{y \in D} f(x, y), \quad x \in S,$$

where we assume that the minimum in the above definition is attained. Then g is convex over S.

Note: We will use this property later in duality theory. The dual problem often arises from partial minimization of the Lagrangian of the primal problem.

Note 2: This property is also useful in large-scale distributed optimization, where decomposing a problem into smaller subproblems that can be solved more easily or in parallel.

Partial minimization

Example 1: Quadratic function. Let $f(x, y) = x^2 + y^2$ convex over $\mathbb{R}^2$. For each x, we want to minimize over y:

$$g(x) = \min_{y \in \mathbb{R}} (x^2 + y^2)$$

For a fixed x, this function is minimized when $y = 0$. Therefore,

$$g(x) = x^2$$

which is also convex.

Example 2: Linear function. Let $f(x, y) = x + 2y$. This is a linear function and hence convex over $\mathbb{R}^2$. For each x, minimize over y:

$$g(x) = \min_{y \in \mathbb{R}} (x + 2y)$$

For any fixed x, the value decreases without bound as y goes to negative infinity. This means that the minimum is not attained so the proposition cannot be applied.

Example 3: Piecewise function. Let

$$f(x,y) = \begin{cases} x+y & \text{if } y \geq 0 \\ x-y & \text{if } y < 0 \end{cases}$$

This function is convex since both pieces are convex and the function is continuous. For each x, minimize over y:

$$g(x) = \min_{y \in \mathbb{R}} f(x,y)$$

For $y \geq 0$, the function is increasing in y, and for $y < 0$, it's decreasing. Hence for any fixed x, the function is minimized when $y = 0$. Therefore,

$$g(x) = x$$

which is a linear function and thus convex.

Partial minimization

In general the quadratic case:

$$f(x, y) = x^\top A x + 2x^\top B y + y^\top C y\,,$$

with

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \text{ psd}\,, \quad \text{and} \quad C \text{ pd}\,.$$

minimizing over y gives $g(x) = \min_y f(x, y) = x^\top (A - BC^{-1}B^\top)x$.

We have that g is convex, and hence we have $A - BC^{-1}B^\top$ is psd. (Schur complement).

Partial minimization

Another interesting example of partial minimization is:

Example (Distance from a point to a convex set)

Let $S \subseteq \mathbb{R}^n$ be a convex set. The distance function defined by

$$d(x, S) := \min_{y \in S} \ \|x - y\|$$

is convex since norm is convex over $\mathbb{R}^n \times S$, and thus by the previous property it follows that $d(\cdot, S)$ is convex.

A more intuitive notation would be $d_S(x) : \mathbb{R}^n \to \mathbb{R}$.

Level sets of convex functions

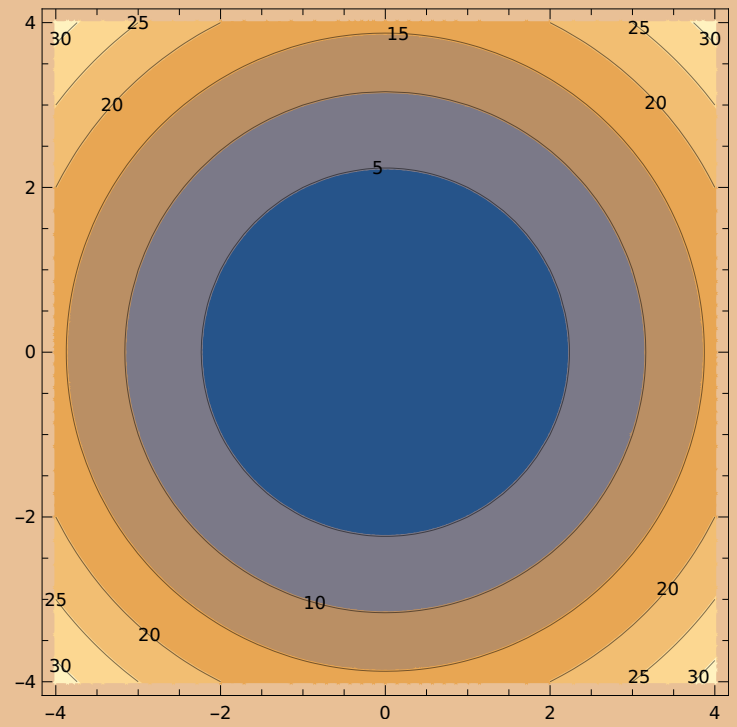
A fundamental property of convex functions is that their level sets are convex.

Let $f : S \to \mathbb{R}$ convex over convex S, then for any $\beta \in \mathbb{R}$ the level set

$$\mathrm{Lev}\,(f, \beta) := \{x \in S : f(x) \leq \beta\}$$

is convex.

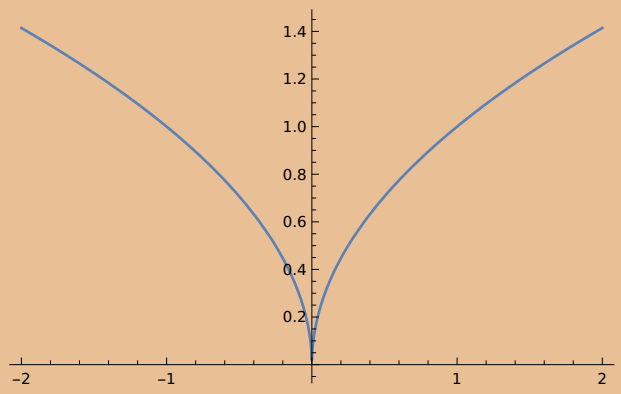
Note: The converse is not true.



Example: not every function that has convex level sets is convex:

$$f(x) = \sqrt{|x|}\,.$$

Levels sets in this case are intervals.



Level sets of convex functions

Example

Consider the following subset of $\mathbb{R}^n$*:*

$$D = \{x \in \mathbb{R}^n : (x^\top Q x + c)^2 + \ln\left(\sum_{i=1}^n e^{x_i}\right) \leq \beta\},$$

where Q *is a given psd matrix,* $c, \beta \in \mathbb{R}$*. The set* D *is convex because is a level set of a convex function. Specifically,* $D = \mathrm{Lev}\,(f, \beta)$*, where*

$$f(x) = (x^\top Q x + c)^2 + \ln\left(\sum_{i=1}^n e^{x_i}\right) .$$

The function f is the sum of two convex functions: the log-sum-exp function, and the function $g(x) = (x^\top Q x + c)^2$, which is convex as a composition of the nondecreasing convex function $\varphi(t) = t^2$ defined on $\mathbb{R}_{\geq 0}$, with the convex quadratic function $x^\top Q x + c$.

Level sets of convex functions

Example (Quotient of affine functions)

Consider

$$f(x) = \frac{a^\top x + b}{c^\top x + d},$$

where $a, c \in \mathbb{R}^n$, $c \neq 0$ *and* $b, d \in \mathbb{R}$ *defined on the convex set*

$$C = \{x \in \mathbb{R}^n : c^\top x + d > 0\}.$$

Set C *ensures we are not dividing by zero. In general, this is not a convex function, but the levels sets are convex:*

$$\begin{aligned} \mathrm{Lev}(f, \alpha) &= \{x \in C : f(x) \leq \alpha\} \\ &= \{x \in \mathbb{R}^n : c^\top x + d > 0,\ (a - \alpha c)^\top x + (b - \alpha d) \leq 0\} \end{aligned}$$

which is an intersection of halfspaces when $a \neq \alpha c$ *(or emptyset otherwise).*

Directional derivatives of convex functions

Convex functions are not necessarily differentiable, but all the directional derivatives at interior points exist:

Proposition (Existence of directional derivatives for convex functions)

Let $f : S \to \mathbb{R}$ convex, for $S \subseteq \mathbb{R}^n$ convex. Let x_0 interior point of S. Then, for any $v \neq 0$, the directional derivative $\partial_v f(x_0)$ exists.

Local Lipschitz continuity of convex functions

Convex functions are local Lipschitz continuous at interior points of their domain:

Proposition

Let $f : S \to \mathbb{R}$ be convex over convex $S \subseteq \mathbb{R}^n$. Let x_0 interior point of S. Then there exist $\epsilon > 0$ and $L > 0$ such that $B(x_0, \epsilon) \subseteq S$ and

$$|f(x) - f(x_0)| \leq L\|x - x_0\|,$$

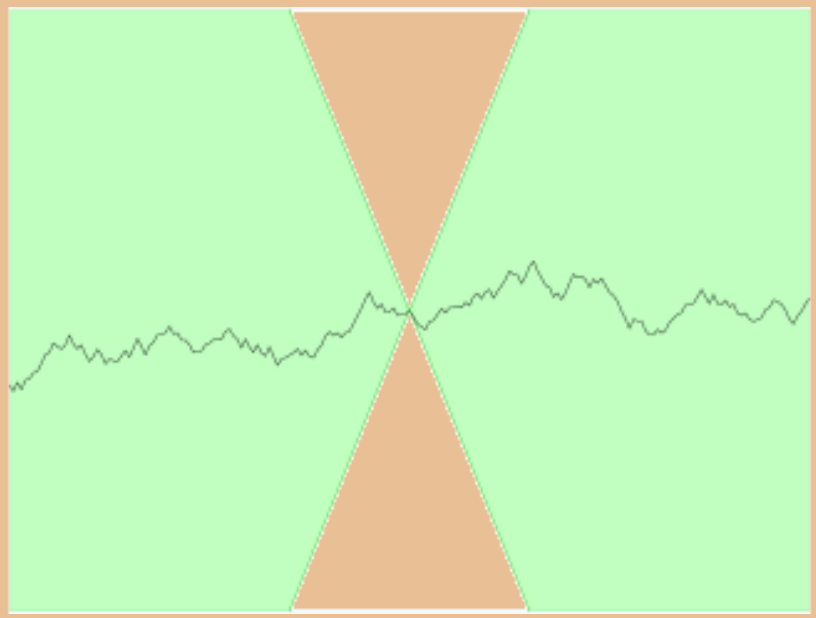
for all $x \in B(x_0, \epsilon)$.

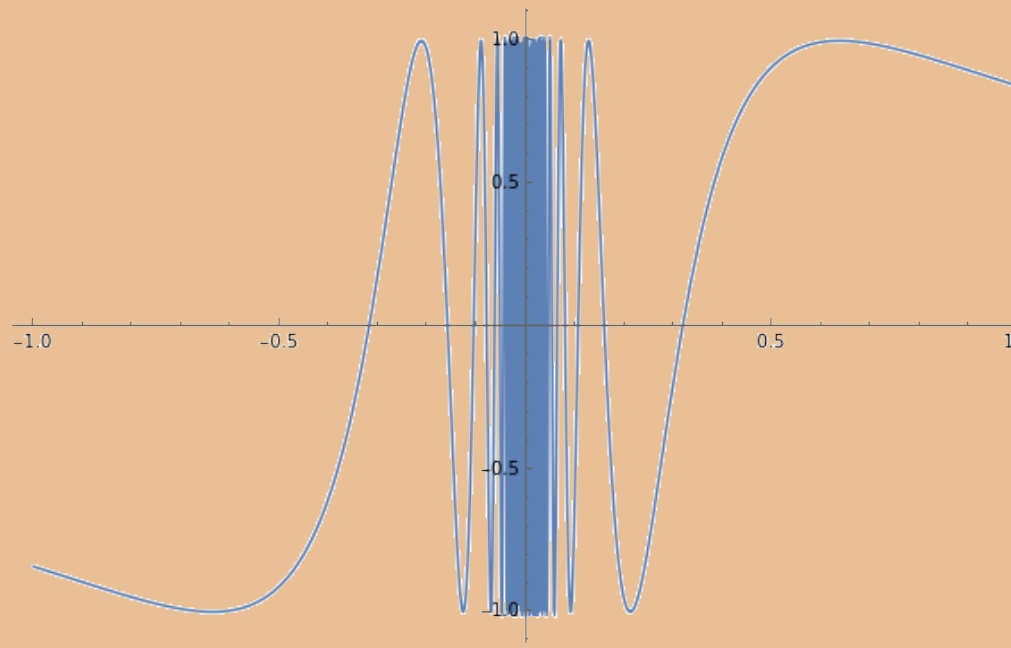
Proof by applying Krein-Milman and Jensen's inequality (not presented here).

Intuitively, this means that the function cannot become excessively steep in the neighborhood.

Important property for optimization because it guarantees the existence and uniqueness of solutions, and enables the use of efficient numerical methods.

Local Lipschitz continuity of convex functions





Left: Lipschitz continuous (function remains inside the green cones).
Right: Bounded but not local Lipschitz at 0.

A bounded function does not mean it is locally Lipschitz continuous.
Example: $\sin(1/x)$. This function becomes infinitely steep at 0 and therefore run outside the green cone.

Lipschitz continuity of convex functions

We say the function is Lipschitz (not only local) if there exists $L > 0$ such that

$$|f(x) - f(y)| \leq L\|x - y\|\,,$$

is valid for all $x, y \in S$.

Convex functions are not necessarily differentiable, but in the case the function is differentiable, then we have the following relation between Lipschitz continuity and mean value theorem:

Proposition (Lipschitz iff bounded first derivative)

A function $f : S \to \mathbb{R}$, $S \subseteq \mathbb{R}^n$ and $f \in C^1$ is Lipschitz continuous iff it has bounded first derivative. In this case, $L = \sup_{x \in S} \|\nabla f(x)\|$.

Proof ($\Longleftarrow$): Follows from the mean value inequality:

$$|f(x) - f(y)| \leq \sup_z \|\nabla f(z)\|\|x - y\| \leq L\|x - y\|\,,$$

where $\nabla f(z)$ is the gradient of at some point z on the line segment between x and y.

Task: This motivates a method for finding L as an optimization problem.

Lipschitz continuity of convex functions - Examples

From $\mathbb{R}^2$ to $\mathbb{R}$:

1. $f(x_1, x_2) = ax_1 + bx_2 + c$, where $a, b, c \in \mathbb{R}$ are constants. This function is Lipschitz continuous with Lipschitz constant $L = \sqrt{a^2 + b^2}$.
2. $f(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$, the Euclidean norm. This function is Lipschitz continuous with $L = 1$.
3. $f(x_1, x_2) = \sin(x_1) + \cos(x_2)$. This function is Lipschitz continuous with $L = \sqrt{2}$.
4. $f(x_1, x_2) = \max\{x_1, x_2\}$. This function is Lipschitz continuous with $L = 1$.
5. $f(x_1, x_2) = e^{x_1 x_2}$. This function is Lipschitz continuous on any bounded set with $L = e^{M^2}$, where M is a bound on $|x_1|$ and $|x_2|$.

Lipschitz continuity of convex functions - Examples

To prove 1): We have:

$$|f(x_1, x_2) - f(y_1, y_2)| = |(ax_1 + bx_2 + c) - (ay_1 + by_2 + c)| = |a(x_1 - y_1) + b(x_2 - y_2)|$$

We want to bound this by multiplying $\|(x_1, x_2) - (y_1, y_2)\|$ with some constant. Notice that by the triangle inequality and the Cauchy-Schwarz inequality:

$$\begin{aligned} |a(x_1 - y_1) + b(x_2 - y_2)| &\leq \sqrt{a^2 + b^2}\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \\ &= \sqrt{a^2 + b^2}\|(x_1, x_2) - (y_1, y_2)\| \end{aligned}$$

Therefore, the Lipschitz constant L of the function f is given by:

$$L = \sqrt{a^2 + b^2}\,.$$

In general, for $f(x) = c^\top x$, where $x, c \in \mathbb{R}^n$ we have $L = \|c\|$.

Lipschitz continuity of convex functions - Examples

To prove 2): Compute the gradient

$$\frac{\partial f}{\partial x_1} = \frac{x_1}{\sqrt{x_1^2 + x_2^2}} \quad \text{and} \quad \frac{\partial f}{\partial x_2} = \frac{x_2}{\sqrt{x_1^2 + x_2^2}}$$

Then,

$$\|\nabla f\| = \sqrt{\left(\frac{x_1}{\sqrt{x_1^2 + x_2^2}}\right)^2 + \left(\frac{x_2}{\sqrt{x_1^2 + x_2^2}}\right)^2} = \sqrt{\frac{x_1^2 + x_2^2}{x_1^2 + x_2^2}} = 1$$

In case, the magnitude of the gradient is 1 for $(x_1, x_2) \neq (0, 0)$, and thus $L = 1$ since this is the supremum of the gradient magnitude over $\mathbb{R}^2$.

Note that, at $(x_1, x_2) = (0, 0)$, the partial derivatives are not defined, which is a typical situation when dealing with fractions. Consider the limit of $\|\nabla f\|$ as $(x_1, x_2) \to (0, 0)$, which is 1.

Finding the Lipschitz constant

To find the Lipschitz constant L of a Lipschitz continuous function $f : C \to \mathbb{R}$, we can solve the following optimization problem:

$$\underset{x,y \in C}{\text{maximize}} \quad \frac{|f(x) - f(y)|}{\|x - y\|} .$$

If the function is differentiable,

$$\underset{x \in C}{\text{maximize}} \quad \|\nabla f(x)\| .$$

Restriction of a convex function to a line

The function $f : S \to \mathbb{R}$, $S \subseteq \mathbb{R}^n$ is convex iff the function $g : D \to \mathbb{R}$, defined as

$$g(t) = f(x+tv), \quad D = \{t \in \mathbb{R} : \ x+tv \in S\},$$

is convex in t for any $x + tv \in S$.

Proof: By composition with an affine mapping.

Example: Let $f : S^n_{++} \to \mathbb{R}$ with $f(X) = \log\det(X)$, recall S^n_{++} is the set of pd matrices in $\mathbb{R}^{n\times n}$.

$$\begin{aligned} g(t) &= \log\det(X + tV) = \log\det(X) + \log\det(I + tX^{-1/2}VX^{-1/2}) \\ &= \log\det(X) + \sum_{i=1}^{n} \log(1 + t\lambda_i) \end{aligned}$$

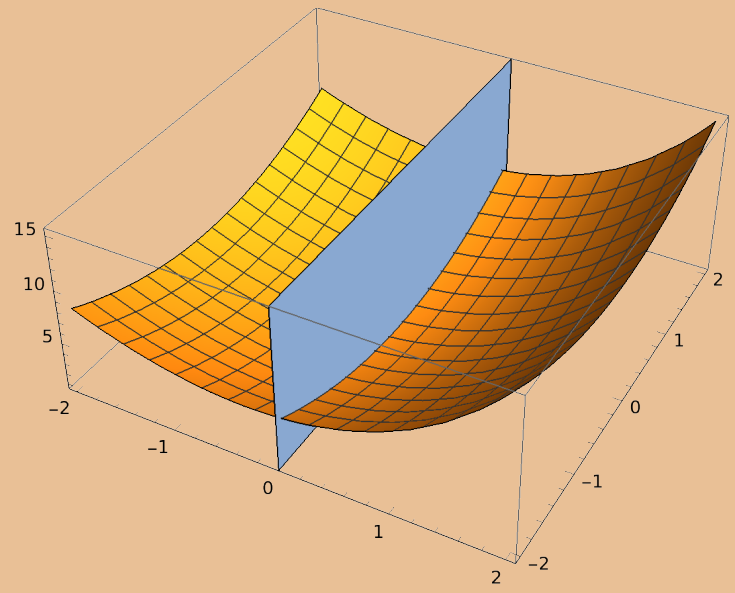
where λ_i are the eigenvalues of $X^{-1/2}VX^{-1/2}$.

Now note that log is a concave function and the last term is an affine composition with respect to t. The first term is a constant.
Thus g is concave in t (for any choice of X pd and V), and hence f is concave.
Note: To apply convexity/concavity definition in f is too difficult.

Restriction of a convex function to a line

We can check convexity of f by checking convexity of functions of one variable.

Many of the algorithms we will see work iteratively minimizing a function over lines, so it is useful that the restriction of a convex function to a line remains convex.

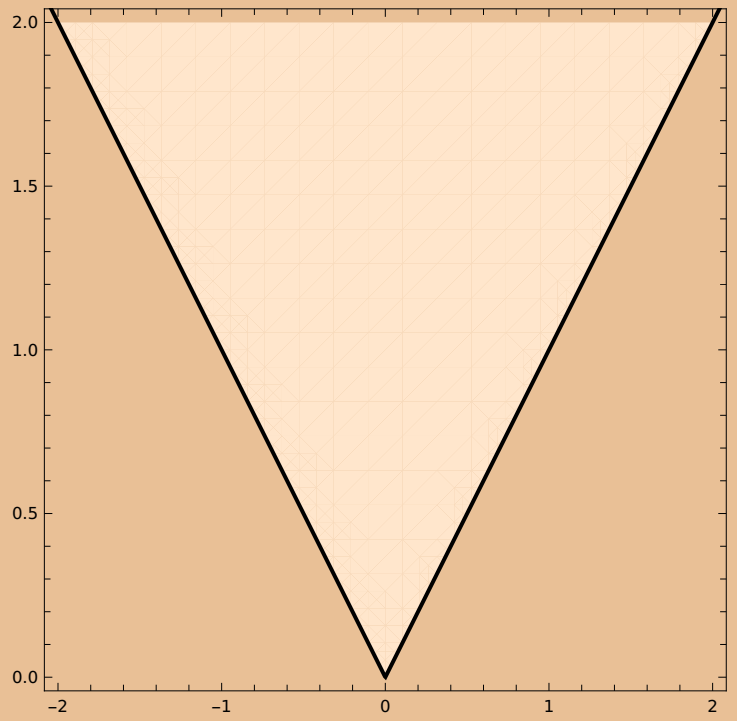


Note plane is parallel to z-axis.

Nonsmooth convex function

Let $f(x) = |x|$. Not differentiable at 0. This function is a norm so we know it is convex:

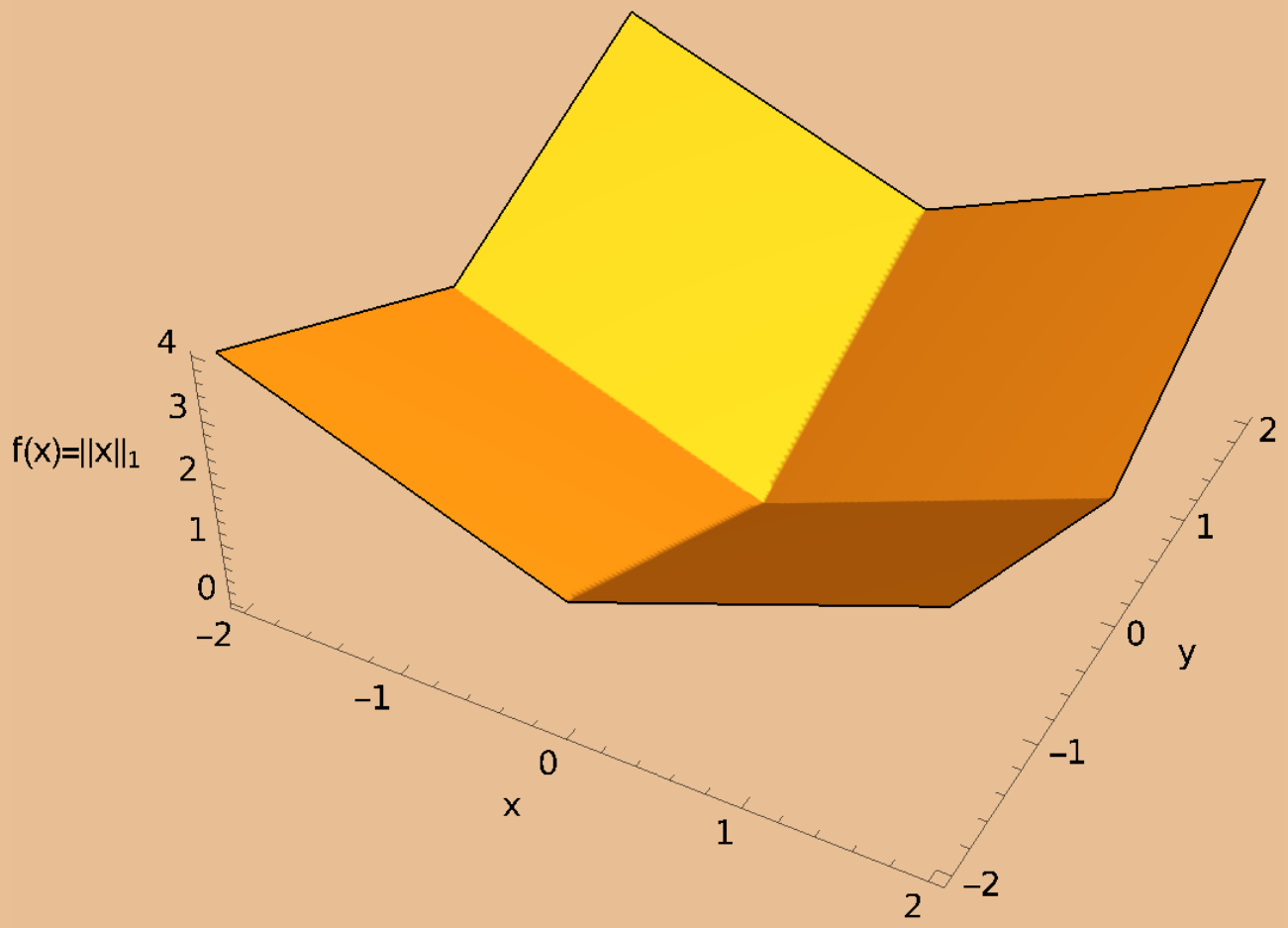
$$\begin{aligned} f(\lambda x_1 + (1-\lambda)x_2) &= |\lambda x_1 + (1-\lambda)x_2| \\ &\leq |\lambda x_1| + |(1-\lambda)x_2| \quad \text{(triangular inequality)} \\ &= \lambda|x_1| + (1-\lambda)|x_2| = \lambda f(x_1) + (1-\lambda)f(x_2) \end{aligned}$$



Note: In general, no norm is differentiable at the origin.

Nonsmooth convex function

In $\mathbb{R}^n$, we have $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$.



1-norm is useful in applications (see next slides).

Ridge regression vs. LASSO

Recall the regularized least squares problem:

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \|A\beta - b\|^2 + \lambda \|\beta\|^2 \,.$$

This is called ridge regression in the context of data fitting. Here A and b are given by the data and β is a parameter vector. This can be also written as the constrained problem:

$$\begin{aligned} \text{minimize} \quad & \|A\beta - b\|^2 \\ \text{s.t.} \quad & \|\beta\|^2 \leq t \,, \end{aligned}$$

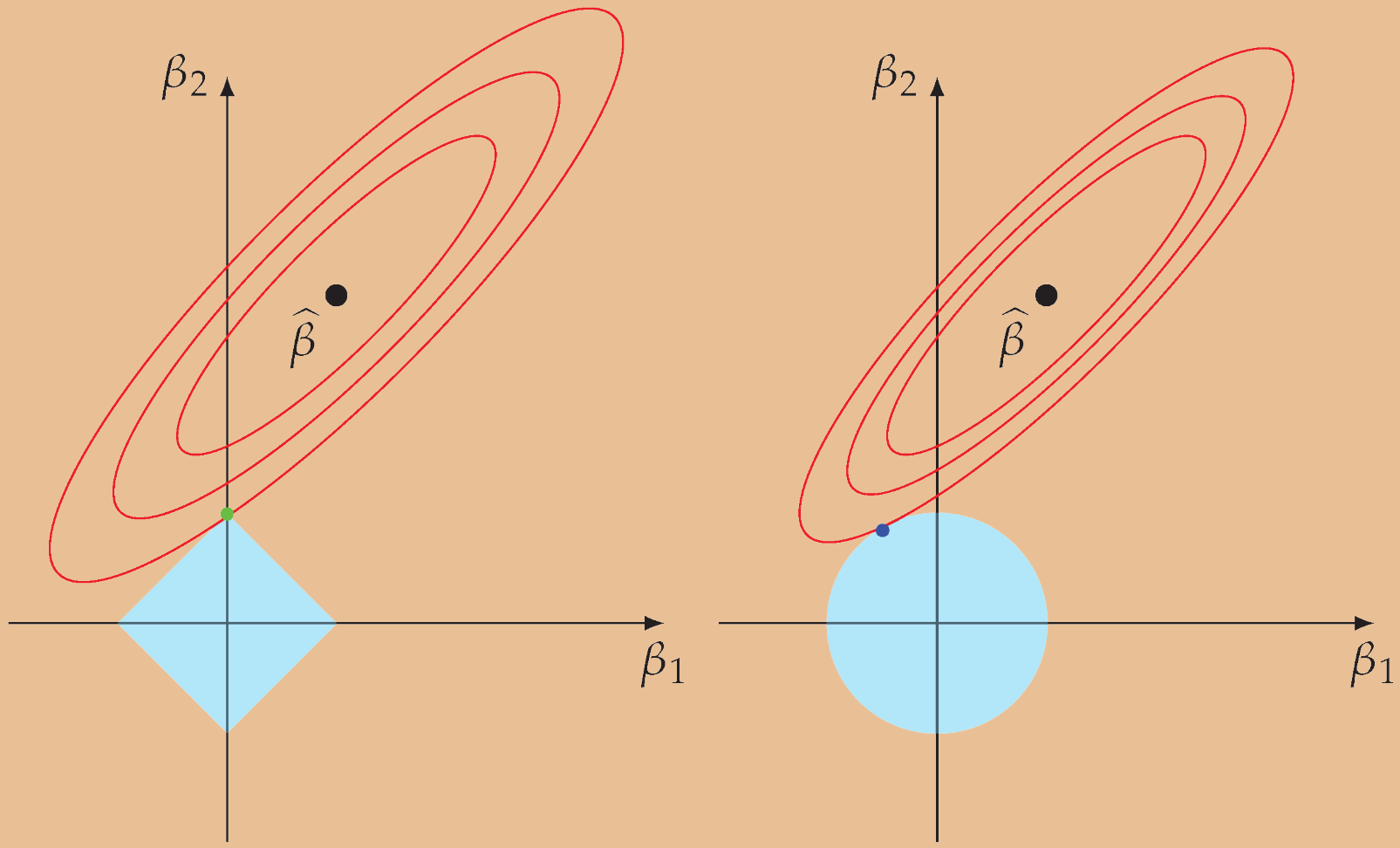
where t is a parameter for example $t = 1$.

LASSO regression penalizes the size of the solution β using the 1-norm:

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \|A\beta - b\|^2 + \lambda \|\beta\|_1 \,.$$

`https://colab.research.google.com/drive/13beHomPWnvz5XZ_E3Mnutsh3eZdh2HD4`

Ridge regression vs. LASSO



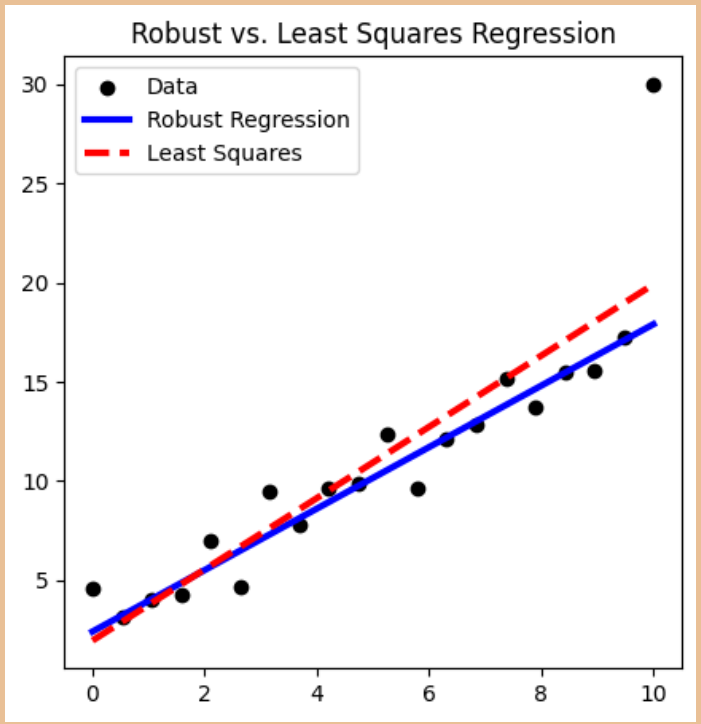
Left: LASSO. Observe $\hat{\beta}_1 = 0$, right: Ridge. Observe $\hat{\beta}_1 \neq 0$ and $\hat{\beta}_2 \neq 0$.

Robust regression

Another interesing use of $\|x\|_1$ is robust regression: aims to fit a regression model in the presence of outliers or when there are violations of the assumptions underlying least squares regression.

That is, solve

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \|A\beta - b\|_1 \, .$$



Robust regression

In general,

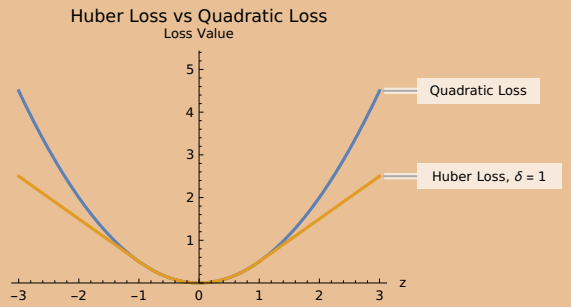
$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \ell(a_i^\top \beta - b_i)\,.$$

where ℓ is a loss function (or a weighting function) that is less sensitive to outliers than the squared error loss used in standard least squares regression.

Some popular loss function choices:

- Huber loss: This is a combination of the squared loss for small residuals and the absolute loss for large residuals:

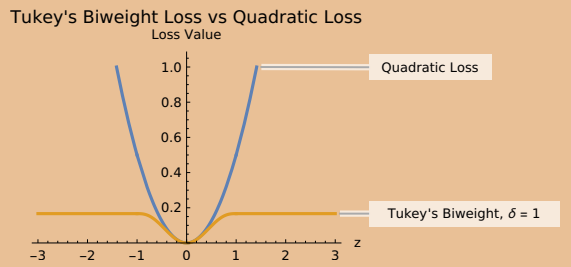
$$\rho(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \le \delta \\ \delta|z| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$



  Here, δ is a threshold that determines the transition between the squared and absolute loss.

- Bisquare (or Tukey's biweight) loss: Stop loss for residuals that are "too large":

$$\rho(z) = \begin{cases} \frac{1}{6}(1 - (1 - (z/\delta)^2)^3) & \text{if } |z| \le \delta \\ \frac{1}{6} & \text{otherwise} \end{cases}$$



  Again, δ is a threshold.

Side note: Extended functions

Up to this point, we worked with functions $f : \mathbb{R}^n \to \mathbb{R}$. It is sometimes useful to consider functions that take values on

$$\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}.$$

By the use of $\bar{\mathbb{R}}$, given any function $g : X \to \mathbb{R}$, $X \subseteq \mathbb{R}^n$ we can define the extended function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$

$$f(x) = \begin{cases} g(x) & \text{if } x \in X \\ +\infty & \text{otherwise} \end{cases}.$$

The domain of an extended function f is

$$\operatorname{dom}(f) = \{x \in \mathbb{R}^n : f(x) < +\infty\}.$$

In advanced optimization books, like Nonlinear Optimization by Ruszczyński, this concept is extensively used.

Side note: Extended functions

By the use of extended functions, many definitions and proofs can be simplified. For example the definition of convex function:

Definition

A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ *is convex if for all* $x, y \in \mathbb{R}^n$ *we have that*

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

where we assume the following arithmetic with $+\infty$

- $x + \infty = +\infty$ *for all* $x \in \mathbb{R}$.
- $x \cdot +\infty = +\infty$ *for all* $x \in \mathbb{R}_{>0}$.
- $0 \cdot +\infty = 0$.

Side note: Indicator function

A useful example of extended function.

Let $S \subseteq \mathbb{R}^n$. The indicator function of S is a function $\delta_S : \mathbb{R}^n \to \bar{\mathbb{R}}$

$$\delta_S(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{otherwise.} \end{cases}$$

It can be shown that it is a convex function iff C is convex.

Suppose S is convex. Then to check $\delta_S(x)$ is convex we apply the definition. We have 4 cases:

- Let $x, y \in S$: by convexity LHS=0 and RHS =0 (if not convex then LHS may not be 0).
- Let x and y not in S: LHS=0 or ∞ and RHS always ∞.
- the other two cases similar than the last one (do it).

By the use of this function we can represent the constrained problem minimize $f(x)$ s.t. $x \in X$ as the unconstrained minimize $f(x) + \delta_X(x)$.

Side note: The pointwise supremum

We can now generalize the pointwise maximum. Let I be an arbitrary index set. Then, if $f_i : S_i \to \bar{\mathbb{R}}$, $i \in I$, is a family of convex functions, then

$$f(x) := \sup_{i \in I} f_i(x)$$

is convex over $S := \bigcap_i S_i$.

I is an arbitrary set and not a finite one. The maximum can now take the value $+\infty$.

Examples:

- Distance to farthest point in a set C:

$$f(x) = \sup_{y \in C} \|x - y\|.$$

- Largest eigenvalue of symmetric matrix: for $A \in S^n$,

$$f(A) \equiv \lambda_{\max}(A) = \sup_{\|y\|_2 = 1} y^T A y.$$

  Note the domain of this function is contained in $\mathbb{R}^{n \times n}$. Note for a symmetric matrix, its maximum eigenvalue is well defined.

Epigraph characterization

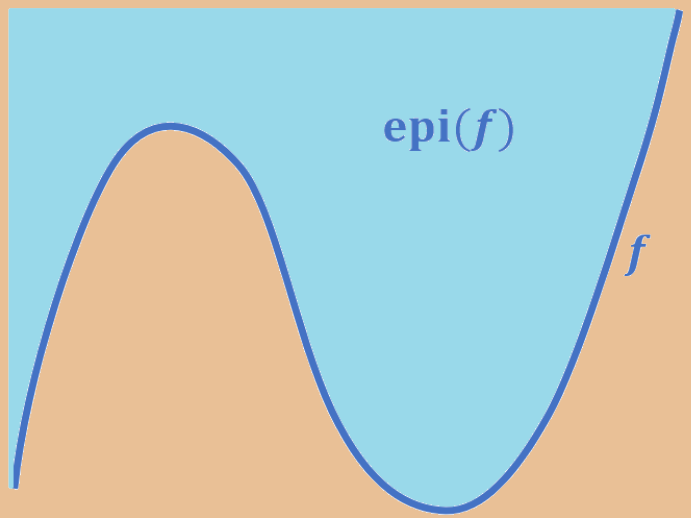
Definition (Epigraph)

Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^n$ convex. The epigraph is the set given by

$$\operatorname{epi}(f) = \{(x, v)^\top \in \mathbb{R}^{n+1} : x \in D, \;\; f(x) \leq v\}.$$

Proposition (Characterization of a convex function)

Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^n$ convex. The function f is convex iff $\operatorname{epi}(f)$ *is a convex set.*



This characterization can be used as the definition of a convex function.

Epigraph characterization

The epigraph is not necessarily closed, but if D is closed and f is continuous then $\text{epi}(f)$ is closed.

Because of the previous observation, we can focus on convex sets only. This is convenient since we can characterize convex sets by its supporting hyperplane.

Proposition

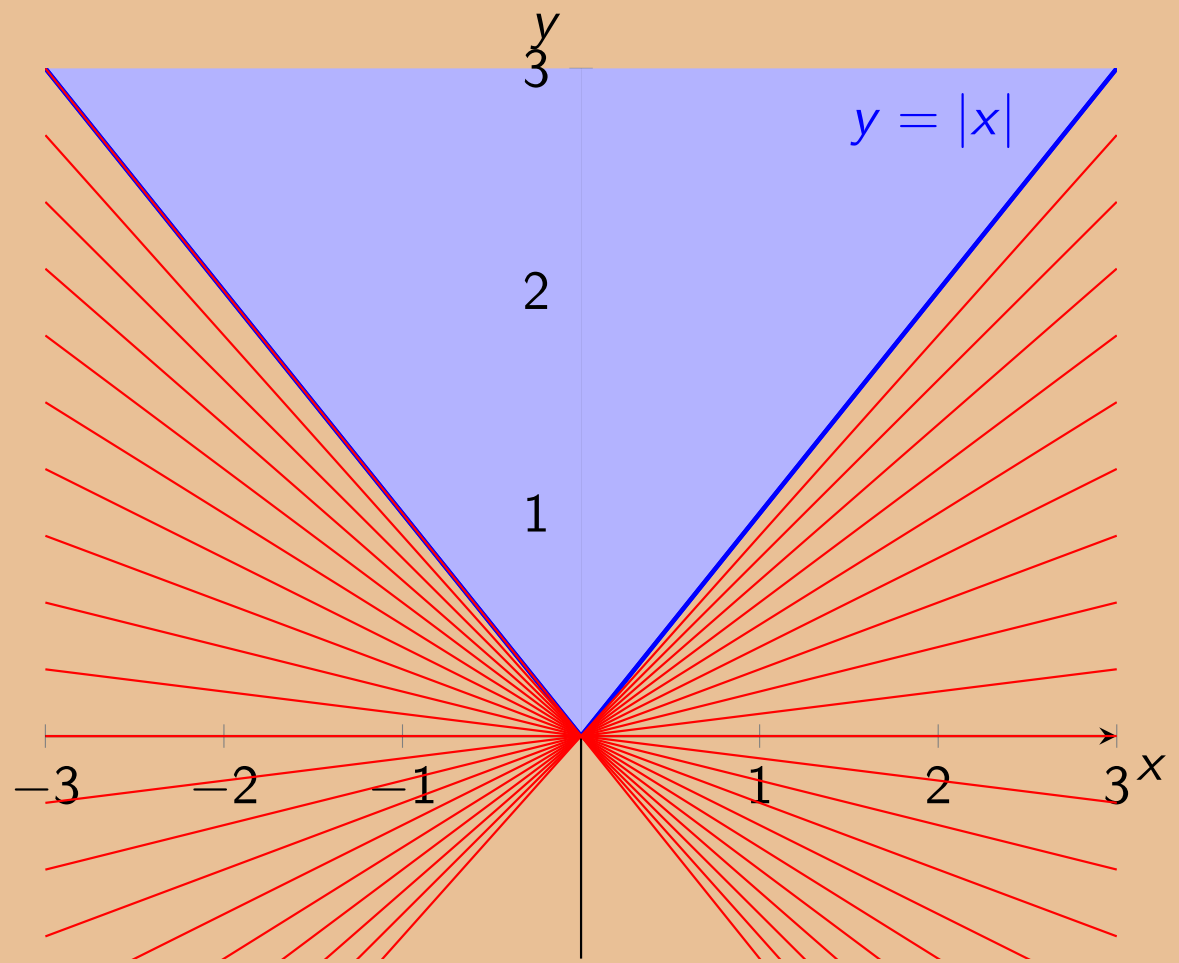
Let $f : D \to \mathbb{R}$ be a convex function, where $D \subseteq \mathbb{R}^n$ convex. Then, for any $\bar{x} \in int(D)$, there exists a hyperplane supporting $cl(epi(f))$ at $\bar{x}$.

Furthermore, if f is differentiable at $\bar{x}$, then the supporting hyperplane is unique and defined by $f(\bar{x}) + \nabla f(x)^\top (x - \bar{x})$, $\forall x \in D$.

It can be shown that this supporting hyperplane is non-vertical.

Epigraph characterization

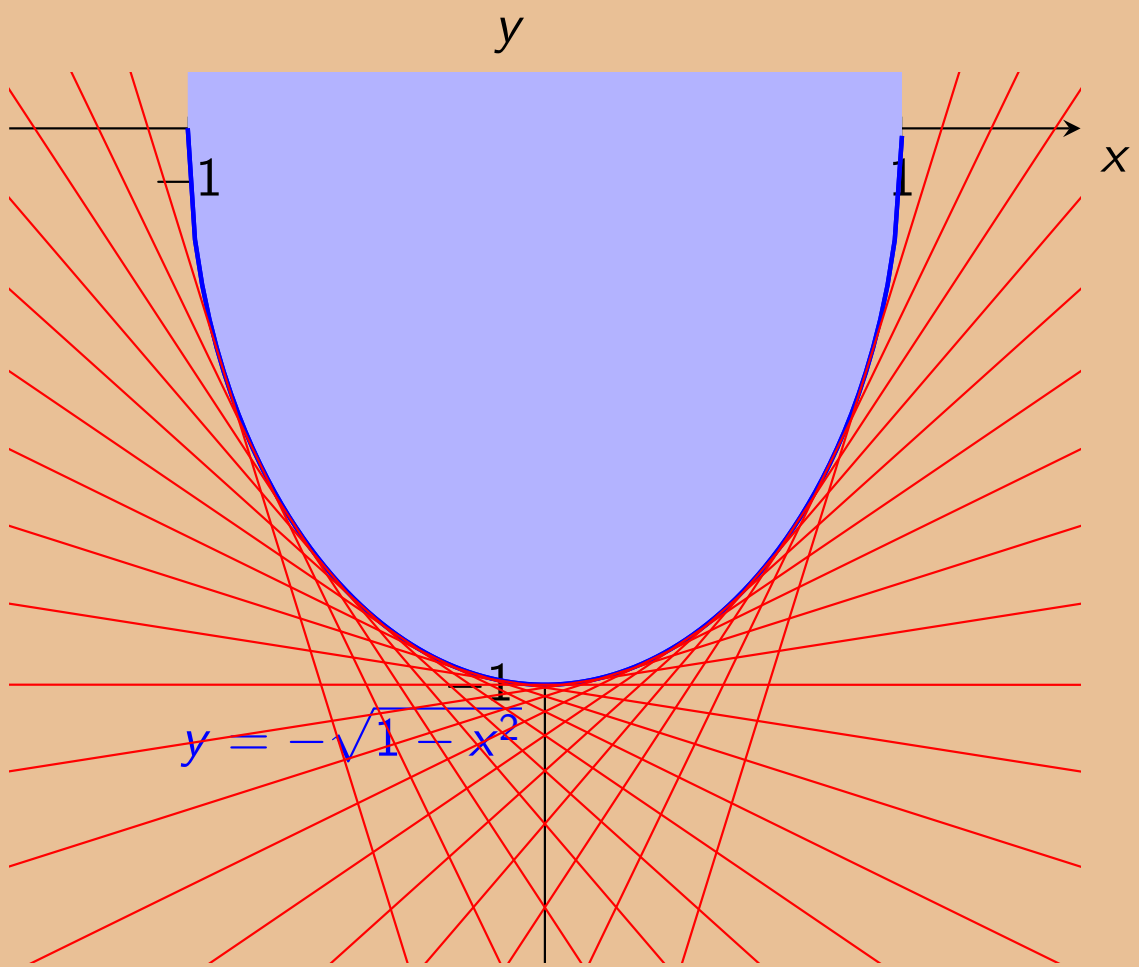
Example 1: The continuous and convex function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = |x|$ is differentiable at every point except $x = 0$. By the previous property, there exists a nonvertical hyperplane supporting $\text{epi}(f)$ at every point in the graph. The hyperplane is not unique at the origin.



Epigraph characterization

Example 2: The same is true for $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) = \|x\|_1$.

Example 3: The epigraph of the function $f(x) = -\sqrt{1 - x^2}$ has supporting hyperplanes for $x \in (-1, 1)$ but vertical supporting hyperplanes at -1 and 1.



Summary: Characterizations of convex functions depending on differentiability [1]

- A C^0 function is convex if the area above the function is a convex set. (epigraph).
- A C^0 function is convex if the function is always below its convex interpolation between points. (first definition of convexity we studied).
- A C^1 function is convex if the function is always above its tangent planes. (gradient inequality).
- A C^1 function is convex if the function if its derivative is non-decreasing. (monotonicity of the gradient).
- A C^2 function is convex if it is not curved downwards at any point. (psd Hessian).

[1] C^0: continuous functions. C^1: differentiable functions whose derivative is continuous (or continuously differentiable).

Establishing the convexity of a function

Summary of methods for establishing the convexity of a given function $f : S \to \mathbb{R}^n$:

- By definition (often simplified by restricting to a line). Show epigraph is a convex set.
- Use a first order characterization if $f \in C^1$.
- Show $\nabla^2 f(x)$ is psd if $f \in C^2$.
- Show that f is obtained from simple convex functions by operations that preserve convexity.
- Helper: Mathematica command `FunctionConvexity`

Example: Show the following function $f : \mathbb{R}^n \to \mathbb{R}$ is convex:

$$f(x) = \max\{\ln\left(\frac{1}{(c^T x + d)^2}\right), \|Ax - b\|^3\}.$$

Part III - Convex optimization problems

Maximal points of convex functions

We now explore an important property useful for maximizing a convex function over a convex set.

Theorem

Let $f : S \to \mathbb{R}$ be a convex function over the compact and convex $S \subseteq \mathbb{R}^n$. Then, there exists at least one maximizer of f over S that is an extreme point of S.

Example

Let

$$f(x) = x^\top Q x\,, \quad Q = \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}\,,$$

and consider

$$\text{maximize } f(x) \quad \textit{s.t.} \quad x_1 + 2x_2 + 3x_3 = 1,\ x \geq 0\,.$$

Maximal points of convex functions

Example

Let

$$f(x) = x^\top Q x + 2b^\top x + c\,, \quad Q \text{ is psd}\,,$$

with $x, b \in \mathbb{R}^n$, $c \in \mathbb{R}$, *and consider*

$$\text{maximize } f(x) \quad \text{s.t.} \quad \|x\|_\infty \leq 1\,.$$

Since the objective function is convex, and the feasible set is convex and compact, it follows that there exists a maximizer at an extreme point of the feasible set $\{-1, 1\}^n$ *thus* $x_i^* = 1$ *or* $x_i^* = -1$ *for* i=$1, .., n$.

A particular example, $f(x_1, x_2) = 3x_1^2 - 2x_1x_2 + 2x_2^2 + 2x_1 + 1$.

See GeoGebra: `https://www.geogebra.org/3d/n5hdun9p`

Example

Basic feasible solutions in linear programs.

Constrained optimization problem

Before defining what is a convex optimization problem we recall the main goal of this course:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{s.t.} \quad & x \in \mathcal{X}, \end{aligned}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$.

A more explicit formulation (replacing $\mathcal{X}$ with equalities and inequalities), called standard form, is

$$\begin{aligned} \text{minimize} \quad & f(x) & \\ \text{s.t.} \quad & g_i(x) \leq 0 & i{=}1, ..., m, \\ & h_j(x) = 0 & j{=}1, ..., p, \end{aligned}$$

where $f, g_1, ..., g_m, h_1, ..., h_p : \mathbb{R}^n \to \mathbb{R}$.

Constrained optimization problem

This problem has the implicit constraint

$$D = \operatorname{dom}(f) \cap \bigcap_{i=1}^{m} \operatorname{dom}(g_i) \cap \bigcap_{j=1}^{p} \operatorname{dom}(h_j)\,.$$

We call D the domain of the problem.

The constraints $g_i(x) \leq 0$, $h_j(x) = 0$, are the explicit constraints.

The feasible region is

$$\mathcal{F} = \{x \in D : g_i(x) \leq 0,\ \ i{=}1, \ldots, m,\ \ h_j(x) = 0,\ \ j{=}1, \ldots, p\}.$$

It is within this region that the solution to the optimization problem lies.

Constrained optimization problem

Optimal value:

- $f^* = \inf\{f(x) : x \in \mathcal{F}\}$.
- $f^* = -\infty$ if problem is unbounded below.
- $f^* = \infty$ if $\mathcal{F}$ is empty.
- a feasible $x \in \mathcal{F}$ is optimal if $f(x) = f^*$.
- $\mathcal{X}^*$ is the set of optimal points.

Constrained optimization problem

A problem is unconstrained if it has no explicit constraints $j = p = 0$.

Examples with $n = 1$, $j = p = 0$:

- $f(x) = \frac{1}{x}$, $\mathrm{dom}\,(f) = \mathbb{R}_{>0} : f^* = 0$, no optimal point.
- $f(x) = -\log x$, $\mathrm{dom}\,(f) = \mathbb{R}_{>0} : f^* = -\infty$.
- $f(x) = x \log x$, $\mathrm{dom}\,(f) = \mathbb{R}_{>0} : f^* = -\frac{1}{e}$, $\quad x^* = \frac{1}{e}$ is optimal.

Convex optimization problems

A convex optimization problem is a problem consisting of minimizing a convex function over a closed and convex set:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{s.t.} \quad & x \in C, \end{aligned} \tag{CP}$$

where C is a closed and convex set and f is a convex function over C.

A more explicit formulation is

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \quad i = 1, ..., m, \\ & a_j^\top x = b_j \quad j = 1, ..., p, \end{aligned}$$

where $f, g_1, ..., g_m : \mathbb{R}^n \to \mathbb{R}$ are convex functions and $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$.

Convex optimization problems

Example 1: The problem

$$\begin{aligned} \text{minimize} \quad & x_1 + x_2 \\ \text{s.t.} \quad & x_1^2 + x_2^2 \leq 1 \end{aligned}$$

is convex since the objective function is linear, and thus convex, and the inequality constraint is the cuadratic convex function
$f(x_1, x_2) = x_1^2 + x_2^2 - 1$.

Example 2: The problem

$$\begin{aligned} \text{minimize} \quad & x_1^2 - x_2 \\ \text{s.t.} \quad & x_1^2 + x_2^2 = 1 \end{aligned}$$

is nonconvex. The objective function is convex, but the constraint is a nonlinear equality constraint and therefore nonconvex.

Example 3:

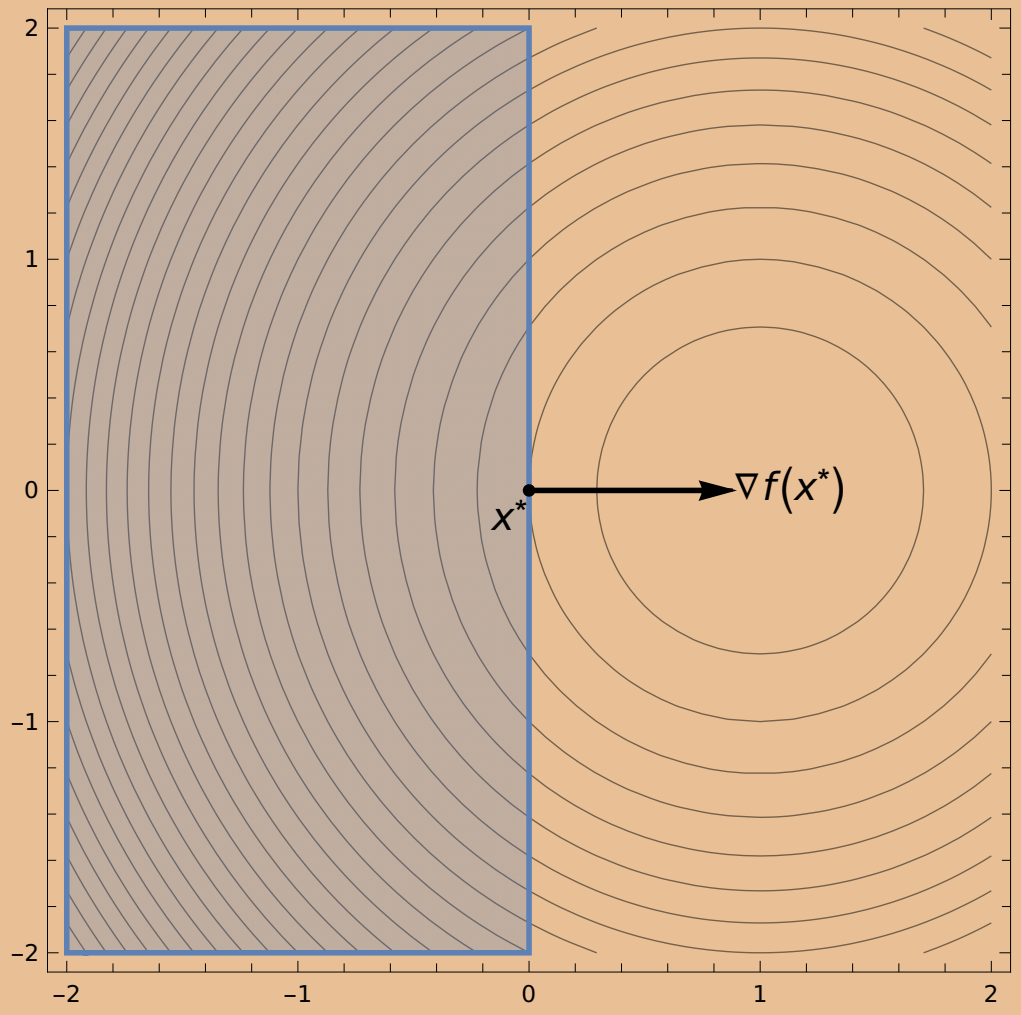
$$\begin{aligned} \text{minimize} \quad & f(x) = (x_1 - 1)^2 + x_2^2 \\ \text{s.t.} \quad & \frac{x_1}{(1 + x_2^2)} \leq 0 \\ & (x_1 + x_2)^2 = 0\,. \end{aligned}$$

- f is convex.
- Feasible set $\{(x_1, x_2) : x_1 = -x_2 \leq 0\}$ is convex.
- Not a convex problem: first equation is not convex, second is not affine.

Equivalent (but not identical) to the convex problem:

$$\begin{aligned} \text{minimize} \quad & (x_1 - 1)^2 + x_2^2 \\ \text{s.t.} \quad & x_1 \leq 0 \\ & x_1 + x_2 = 0\,. \end{aligned}$$

Convex optimization problems



Convex optimization problems

Example 4: minimize $f(x) = -\sum_{i=1}^{k} \log(b_i - a_i^\top x)$ is an unconstrained problem with implicit constraints $a_i^\top x < b_i$.

Example 5: Least squares is a family of convex problems we worked out extensively.

Note: In a convex optimization problem the feasible set is a convex set since

$$C = (\cap_{i=1}^{m} \text{Lev}\,(g_i, 0)) \cap \left(\cap_{j=1}^{p} \{x : h_j(x) = 0\}\right)$$

which implies that C is a convex set as an intersection of level sets of convex sets and hyperplanes.

Theorem (local is global in convex optimization)

Let $f : C \to \mathbb{R}$ be a convex function defined on the convex set C. Let $x^* \in C$ be a <u>local</u> minimum of f over C. Then x^* is a <u>global</u> minimum of f over C.

Proof.

Since x^* is a local minimum of f over C, then exists $r > 0$ such that $f(x^*) \leq f(x)$ for all $x \in C$ satisfying $x \in B[x^*, r]$, that is $x \in B[x^*, r] \cap C$.

Let $y \in C$, $y \neq x^*$. We now show that

$$f(x^*) \leq f(y)\,.$$

Let $\alpha \in (0, 1]$ be such that $x^* + \alpha(y - x^*) \in B[x^*, r]$ (an example of such α is $\alpha = \frac{r}{\|y - x^*\|}$).

Since $x^* + \alpha(y - x^*) \in B[x^*, r] \cap C$, it follows that

$$f(x^*) \leq f(x^* + \alpha(y - x^*))\,,$$

and hence by convexity

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

Thus, $\alpha f(x^*) \leq \alpha f(y)$, and hence $f(x^*) \leq f(y)$. □

Convex optimization problems

A slight modification of the previous result shows that any local minimum of a *strictly* convex function over a convex set is a *strict* global minimum of the function over the set.

Theorem

Let $f : C \to \mathbb{R}$ be a strictly convex function defined over the convex set $C \subseteq \mathbb{R}^n$. Let $x^* \in C$ be a local minimum of f over C. Then x^* is a strict global minimum of f over C.

Convex optimization problems

Definition (The optimal set of a convex problem)

The optimal set of (CP) *is the set of all global minimizers, that is,*

$$C^* := \{x \in C : x \text{ minimizes } f(x)\} = \arg\min \{f(x) : x \in C\}.$$

This definition of an optimal set is also valid for nonconvex problems.

Recall that f^* is naturally unique (why?).

A remarkable property of convex problems is that their optimal sets are also convex.

Theorem (Convexity of the optimal set)

Let $f : C \to \mathbb{R}$ be a convex function defined over the convex set $C \subseteq \mathbb{R}^n$. Then the set of optimal solutions of the problem (CP) *is convex. If, in addition, f is strictly convex over C, then C^* has only one element.*

Proof.

If $C^* = \emptyset$, the result follows trivially.

Suppose that $C^* \neq \emptyset$ and denote an optimal value by f^*. Let $x, y \in C^*$ and $\alpha \in [0, 1]$. Then by convexity we have

$$f(\alpha x + (1-\alpha)y) \leq \alpha f^* + (1-\alpha)f^* = f^*\,,$$

and hence $\alpha x + (1-\alpha)y$ is also optimal and therefore belongs to C^*. Thus C^* is convex.

Suppose now that f is strictly convex and $C^* \neq \emptyset$. To show that C^* contains one element, suppose by contradiction that there exist $x, y \in C^*$ such that $x \neq y$. Then $\frac{1}{2}x + \frac{1}{2}y \in C$, and by the strict convexity of f we have

$$f\left(\frac{1}{2}x + \frac{1}{2}y\right) < \frac{1}{2}f(x) + \frac{1}{2}f(y) = \frac{1}{2}f^* + \frac{1}{2}f^* = f^*,$$

which is a contradiction to the fact that f^* is the optimal value. □

First order condition

We now show that the first-order necessary condition in the differentiable convex case for a point to be a minimizer is also sufficient.

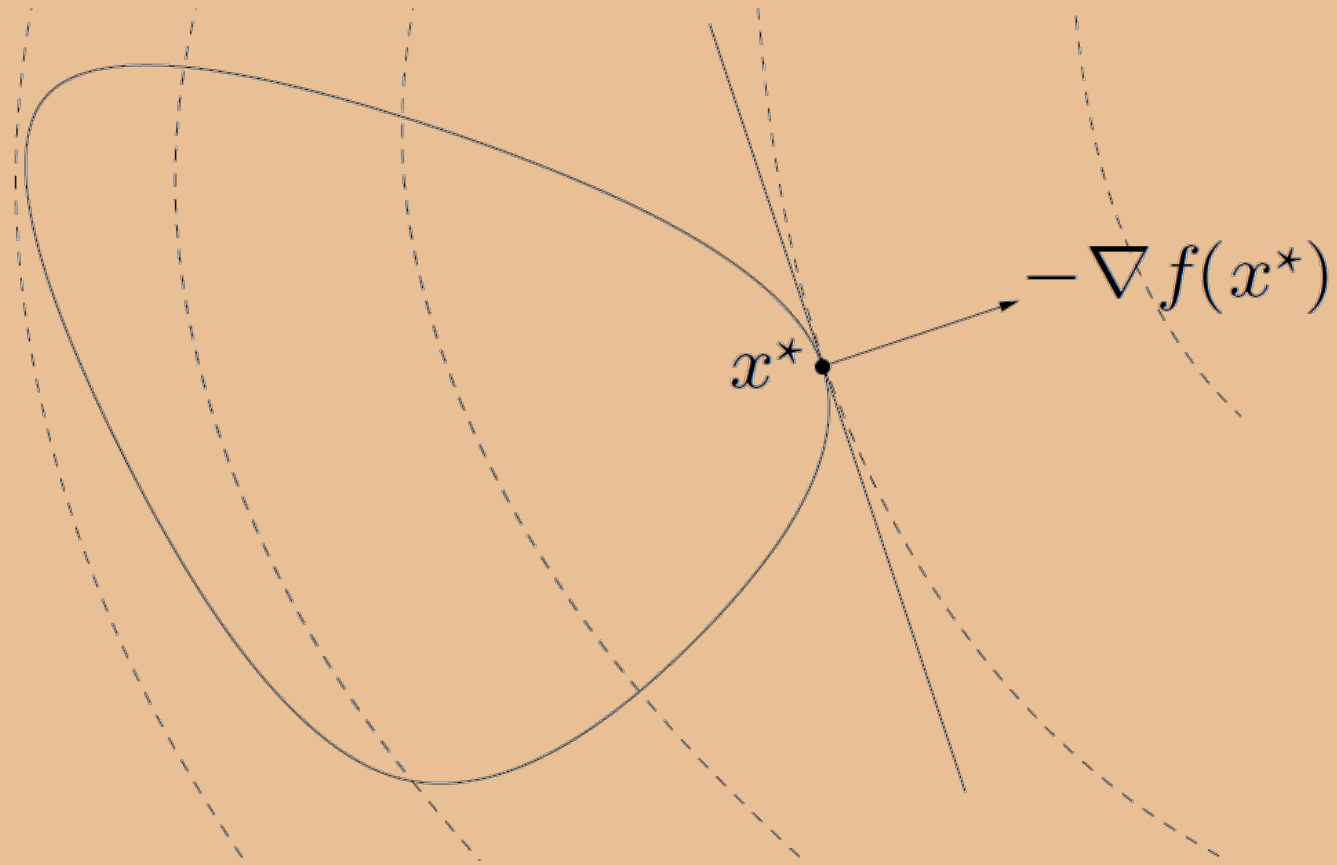
Proposition (FD-FONSC convex case)

Let $f : S \to \mathbb{R}$ be a convex function defined on the convex set $S \subseteq \mathbb{R}^n$, and $f \in C^1$ on an open convex set containing S. Then, x^* is a global minimizer of f over S iff we have

$$\nabla f(x^*)^\top (x - x^*) \geq 0\,, \qquad \textit{for all } y \in S\,.$$

Note: Any feasible direction in a convex set S can be written $d = x - x^*$ for $x \in S$.

First order condition



Note: If nonzero, $\nabla f(x^*)$ defines a supporting hyperplane to feasible set S at x^*.

First order condition

We also have the particular case:

Proposition (FONSC convex case)

Let $f : S \to \mathbb{R}$, $f \in C^1$, be a convex function defined on the convex set $S \subseteq \mathbb{R}^n$. Then, x^* is a global minimizer of f over S iff the point $x^* \in int(S)$ is such that $\nabla f(x^*) = 0$.

Example: Consider

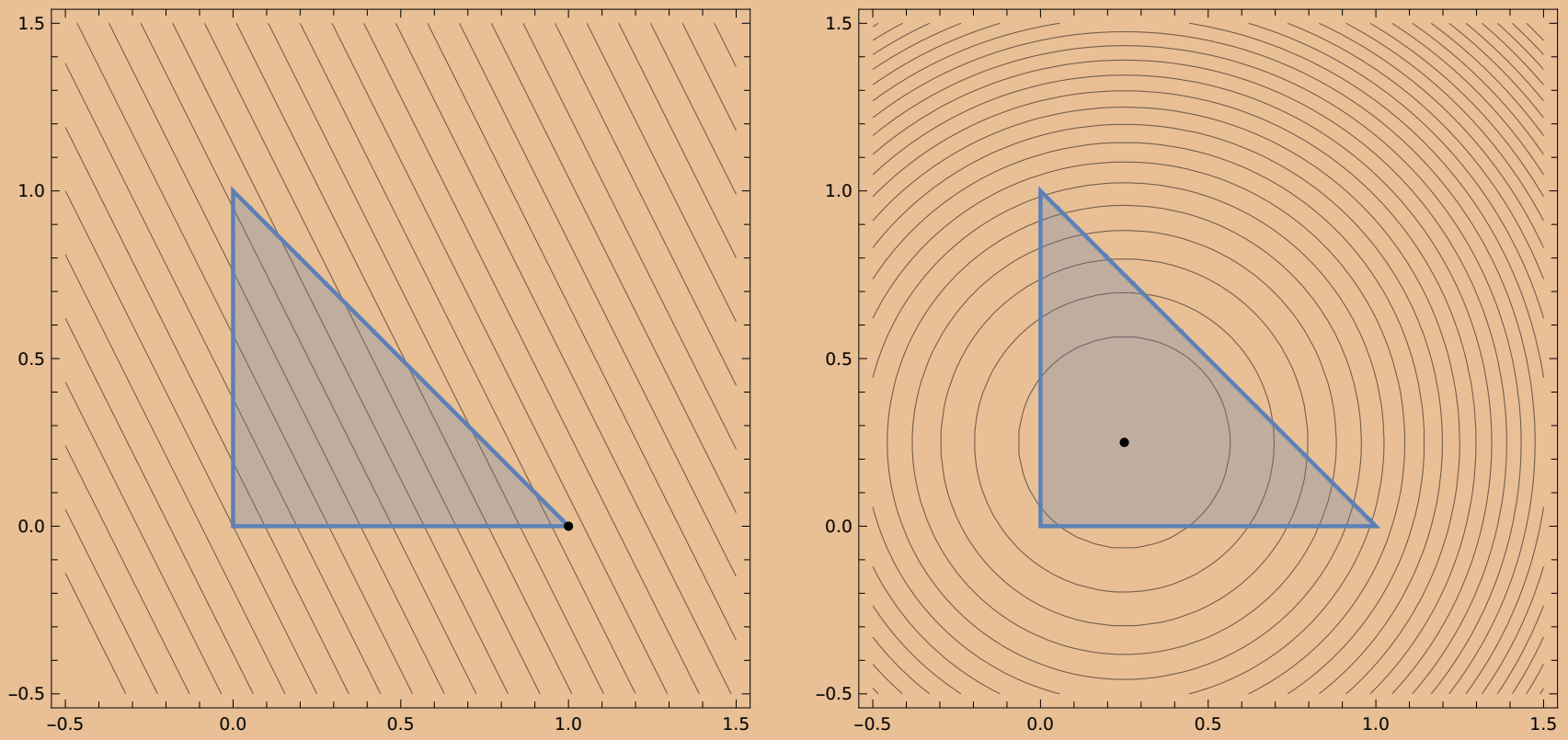
$$\begin{aligned} \text{minimize} \quad & -2x_1 - x_2 \\ \text{s.t.} \quad & (x_1, x_2) \in \mathcal{X}\,, \end{aligned}$$

and

$$\begin{aligned} \text{minimize} \quad & (x_1 - \frac{1}{4})^2 + (x_2 - \frac{1}{4})^2 \\ \text{s.t.} \quad & (x_1, x_2) \in \mathcal{X}\,. \end{aligned}$$

where $\mathcal{X} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq 1,\ x_1 \geq 0,\ x_2 \geq 0\}$.

First order sufficient condition



Left: $\nabla f \neq 0$. The point $(1,0)^\top$ satisfies $\nabla f(1,0)^\top d \geq 0$ for all d feasible direction.

Right: $\nabla f(1/4, 1/4) = 0$. In both cases we have convex $f \in C^1$ and convex $\mathcal{X}$. Both points are global min in $\mathcal{X}$.

Examples

Equality constrained problem:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{s.t.} \quad & Ax = b\,. \end{aligned}$$

x^* is optimal iff there exists a v such that

$$x^* \in \operatorname{dom}(f)\,, \quad Ax^* = b\,, \quad \nabla f(x^*) + A^\top v = 0\,.$$

Minimization over nonnegative orthant:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{s.t.} \quad & x \geq 0\,. \end{aligned}$$

x^* is optimal iff $x^* \in \operatorname{dom}(f)$, $x^* \geq 0$, and $\begin{cases} \nabla f(x^*)_i \geq 0 & x_i^* = 0 \\ \nabla f(x^*)_i = 0 & x_i^* > 0 \end{cases}$

We will work out these problems in the next section.

Equivalent convex problems

Two problems are (informally) equivalent if the solution of one is readily obtained from the solution of the other, and vice-versa.

Some common problem transformations that preserve convexity:

i) Eliminating equality constraints:

$$\begin{aligned} \text{minimize}_{\,x} \;\; & f(x) \\ \text{s.t.} \;\; & g_i(x) \leq 0, \;\; i{=}1, \ldots, m \\ & Ax = b\,. \end{aligned}$$

is equivalent to (solve the linear system and replace the set of solutions)

$$\begin{aligned} \text{minimize}_{\,z} \;\; & f(Fz + x_0) \\ \text{s.t.} \;\; & g_i(Fz + x_0) \leq 0, \;\; i{=}1, \ldots, m\,. \end{aligned}$$

where F and x_0 are such that $Ax = b \iff x = Fz + x_0$ for some z. Here F is a basis matrix of the null space.

ii) The opposite can be done (introduce linear constraints).

Equivalent convex problems

iii) Introducing slack variables for linear inequalities (we studied this technique already).

iv) Standard form convex problem is equivalent to

$$\begin{array}{ll} \text{minimize}_{(x,t)} & t \\ \text{s.t.} & f(x) - t \leq 0 \\ & g_i(x) \leq 0, \quad i=1,\ldots,m \\ & Ax = b\,. \end{array}$$

Abstracts away the specific form of the objective function.

v) Minimizing over some variables:

$$\begin{array}{ll} \text{minimize} & f(x_1, x_2) \\ \text{s.t.} & g_i(x_1) \leq 0, \quad i=1,\ldots,m\,. \end{array}$$

is equivalent to

$$\begin{array}{ll} \text{minimize} & \tilde{f}(x_1) \\ \text{s.t.} & \tilde{g}_i(x_1) \leq 0, \quad i=1,\ldots,m\,. \end{array}$$

where $\tilde{f}(x_1) = \inf_{x_2} f(x_1, x_2)$.

Linear optimization problems

A fundamental example of convex optimization problem is the linear program:

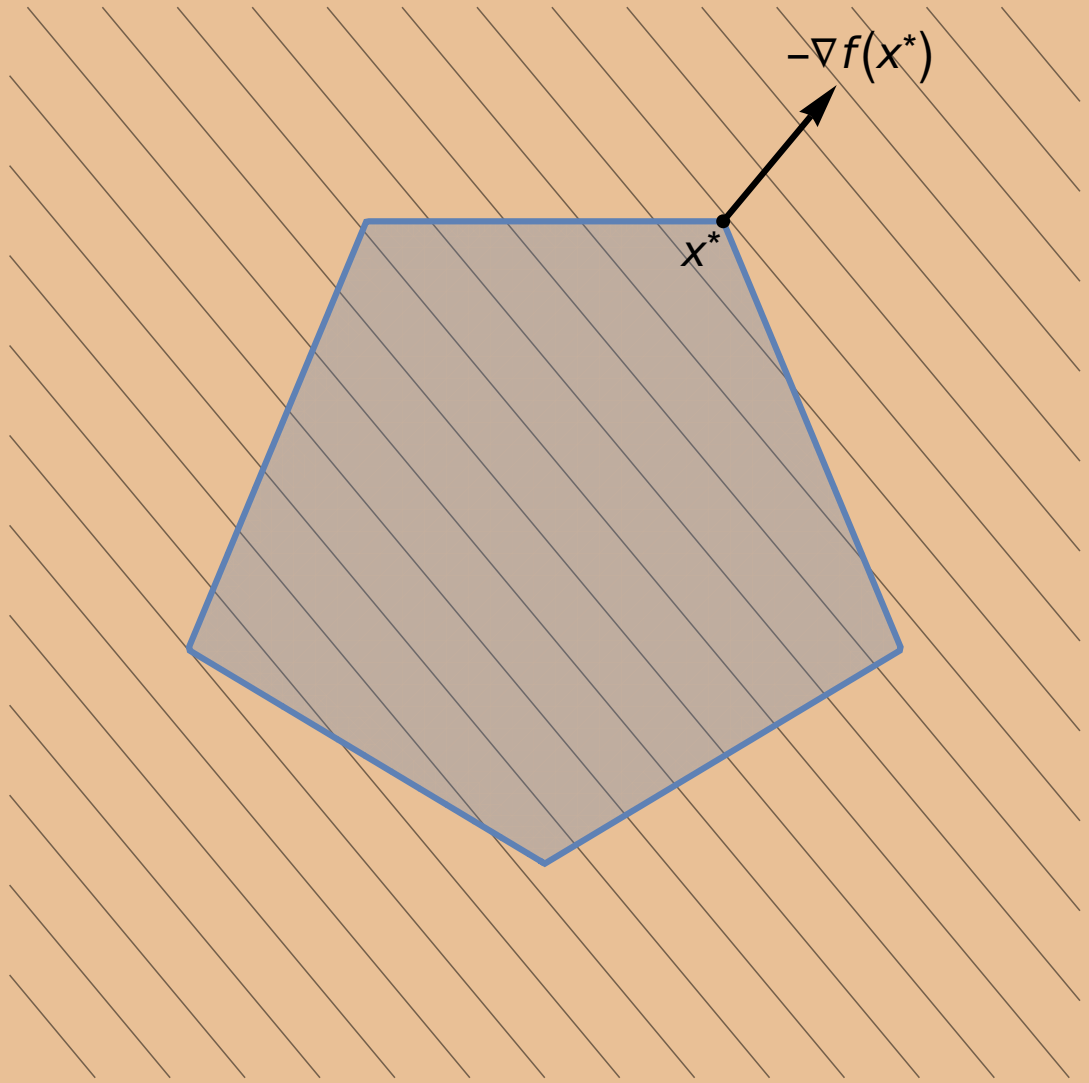
$$\begin{aligned} \text{minimize} \quad & c^\top x \\ \text{s.t.} \quad & Ax \leq b \\ & Bx = g\,, \end{aligned} \tag{LP}$$

where $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times n}$.
In standard form:

$$\begin{aligned} \text{minimize} \quad & c^\top x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0\,. \end{aligned}$$

Any LP can be converted to standard form.
Application: Arbitrage detection using LP (very interesting, but time permitting).

Linear optimization problems



Constraint set is a polytope.

Linear optimization problems

Example: Piecewise-linear minimization:

$$\text{minimize} \quad \max_{i=1,\ldots,m} \left(a_i^\top x + b_i\right)$$

is equivalent to the LP

$$\begin{aligned} \text{minimize} \quad & t \\ \text{s.t.} \quad & a_i^\top x + b_i \leq t, \quad i=1,\ldots,m\,. \end{aligned}$$

Piecewise-linear minimization can be used in i) portfolio optimization problems where the goal is to minimize the worst-case scenario of portfolio losses; ii) Support vector machines (more of this later); iii) in control theory, used to ensure stability or performance under the worst-case disturbances or uncertainties; etc

Linear optimization problems

Example: Chebyshev center of the polytope

$$P = \{x : a_i^\top x \leq b_i,\ i = 1, \ldots, m\}\,.$$

is the center of the largest inscribed ball (inflated euclidean ball)

$$B = \{x_c + u : \|u\|_2 \leq r\}\,.$$

Observe: For the ball B to be entirely contained within P, the farthest point from the center to the surface of the ball, in the direction of a_i, must not exceed the boundary defined by b_i.

Thus $a_i^\top x \leq b_i$ for all $x \in B$ if and only if

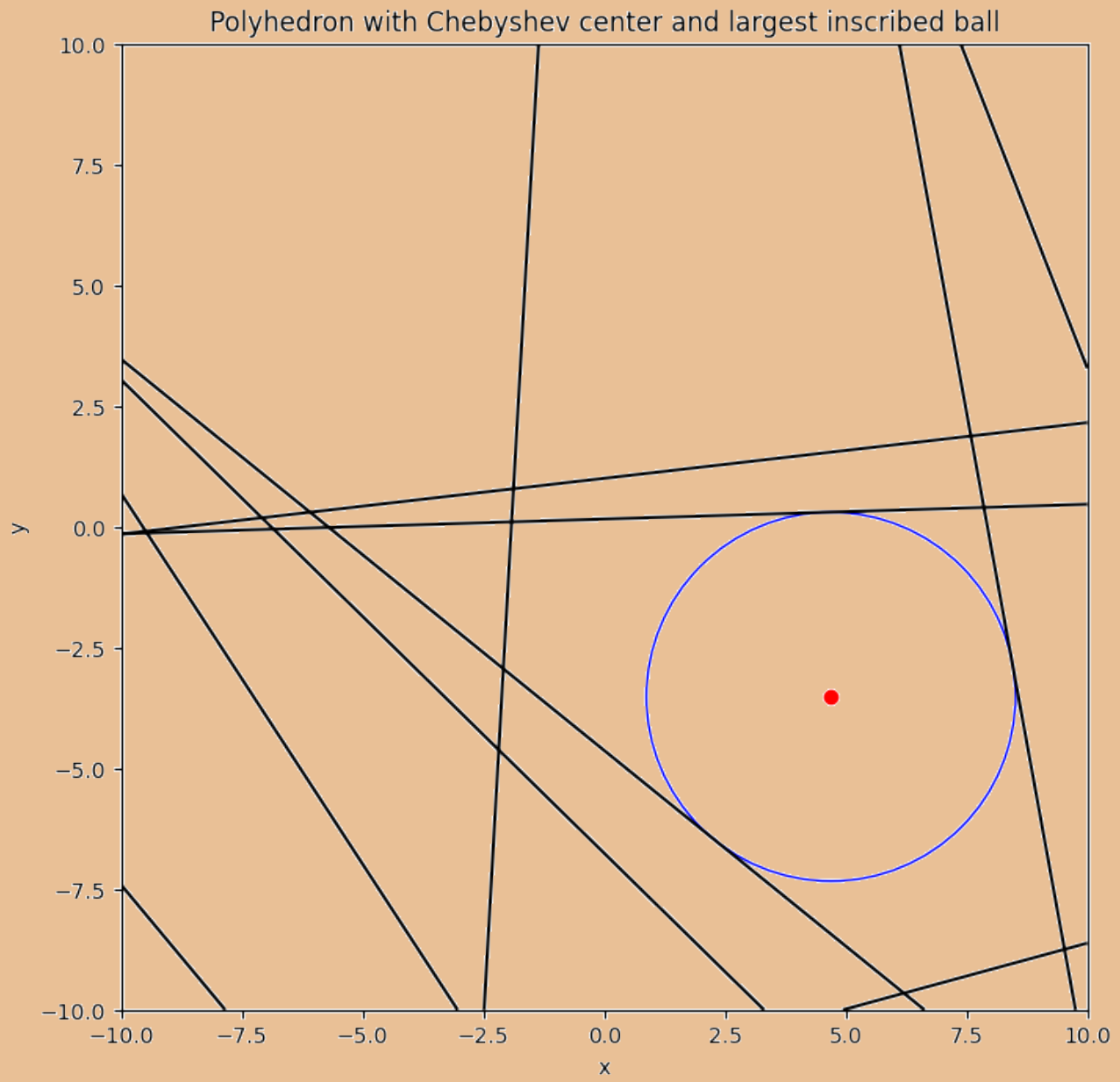
$$\sup_u\ \{a_i^\top (x_c + u) : \|u\|_2 \leq r\} = a_i^\top x_c + r\|a_i\|_2 \leq b_i\,.$$

The equality is because we are maximizing $a_i^\top u$ and thus use C-S inequality.

Hence, x_c, r can be determined by solving the LP

$$\begin{aligned} &\text{maximize} && r \\ &\text{s.t.} && a_i^\top x_c + r\|a_i\|_2 \leq b_i, \quad i = 1, \ldots, m \end{aligned}$$

Linear optimization problems



Linear optimization problems

Note: To randomly generate a polytope in low dimensions $n = 2$ or 3 it is enough to randomly generate A and b. Given a sufficiently large but not excessive number of inequalities relative to n, the generated half-spaces will likely intersect in a way that forms a non-empty polyhedron. The chance that all these half-spaces will point away from a common center and result in an empty intersection is low.

Quadratic with affine constraints

Another essential example of convex problem is the class

$$\begin{aligned} \text{minimize} \quad & x^\top Q x + 2b^\top x + c \\ \text{s.t.} \quad & Ax \leq g\,, \end{aligned} \tag{QP}$$

where $Q \in \mathbb{R}^{n\times n}$ pd, $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $g \in \mathbb{R}^m$.

Examples: We already worked out extensively the problems

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \, \|Ax - b\|^2 \qquad \text{(least squares)}$$

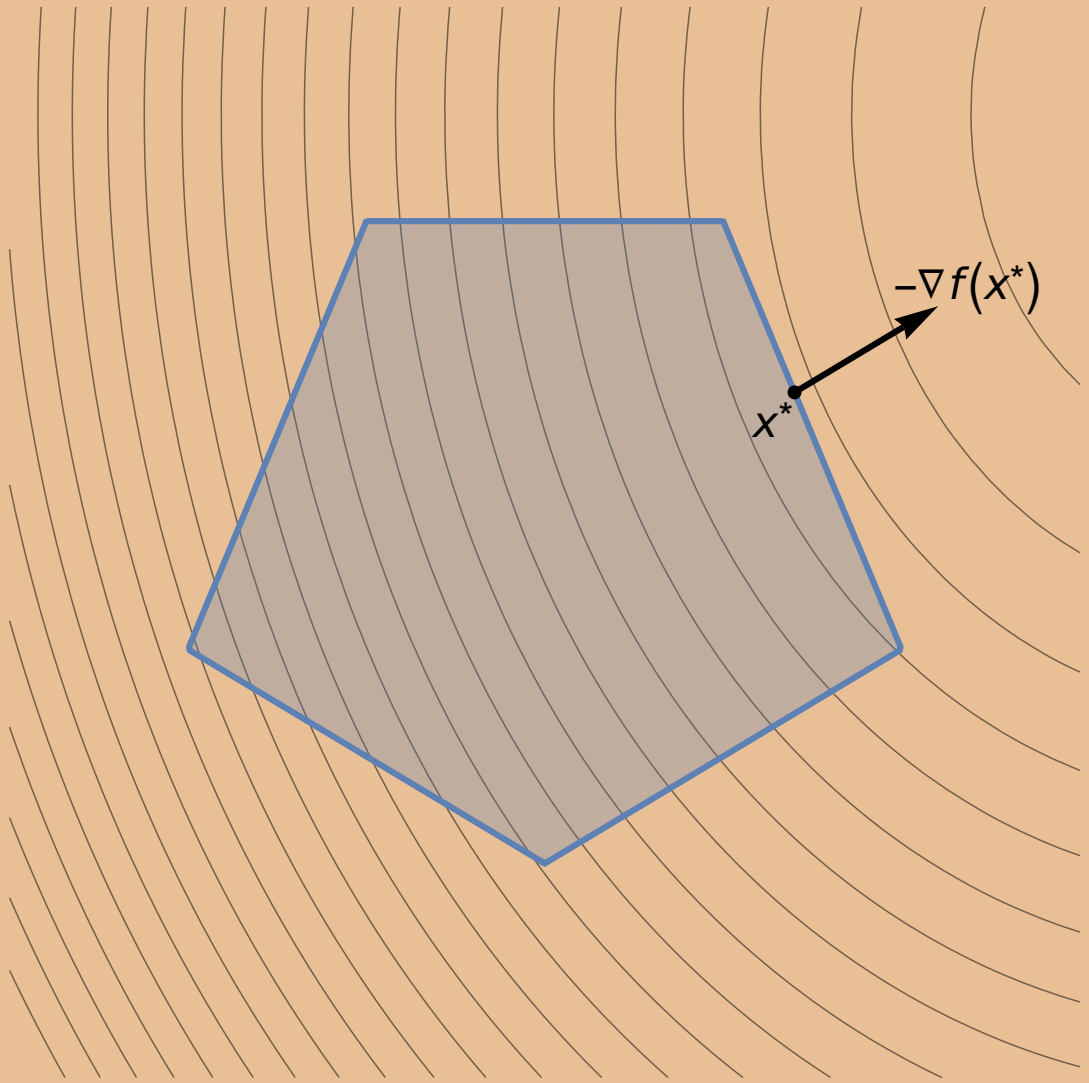
and

$$\begin{aligned} \text{minimize} \quad & \|x\|^2 \\ \text{s.t.} \quad & Ax = b\,. \end{aligned}$$

Both problems with analytical solution. If we add affine constraints to the LS problem then no more.

Application: Portfolio selection (more on this later).

Quadratic with affine constraints

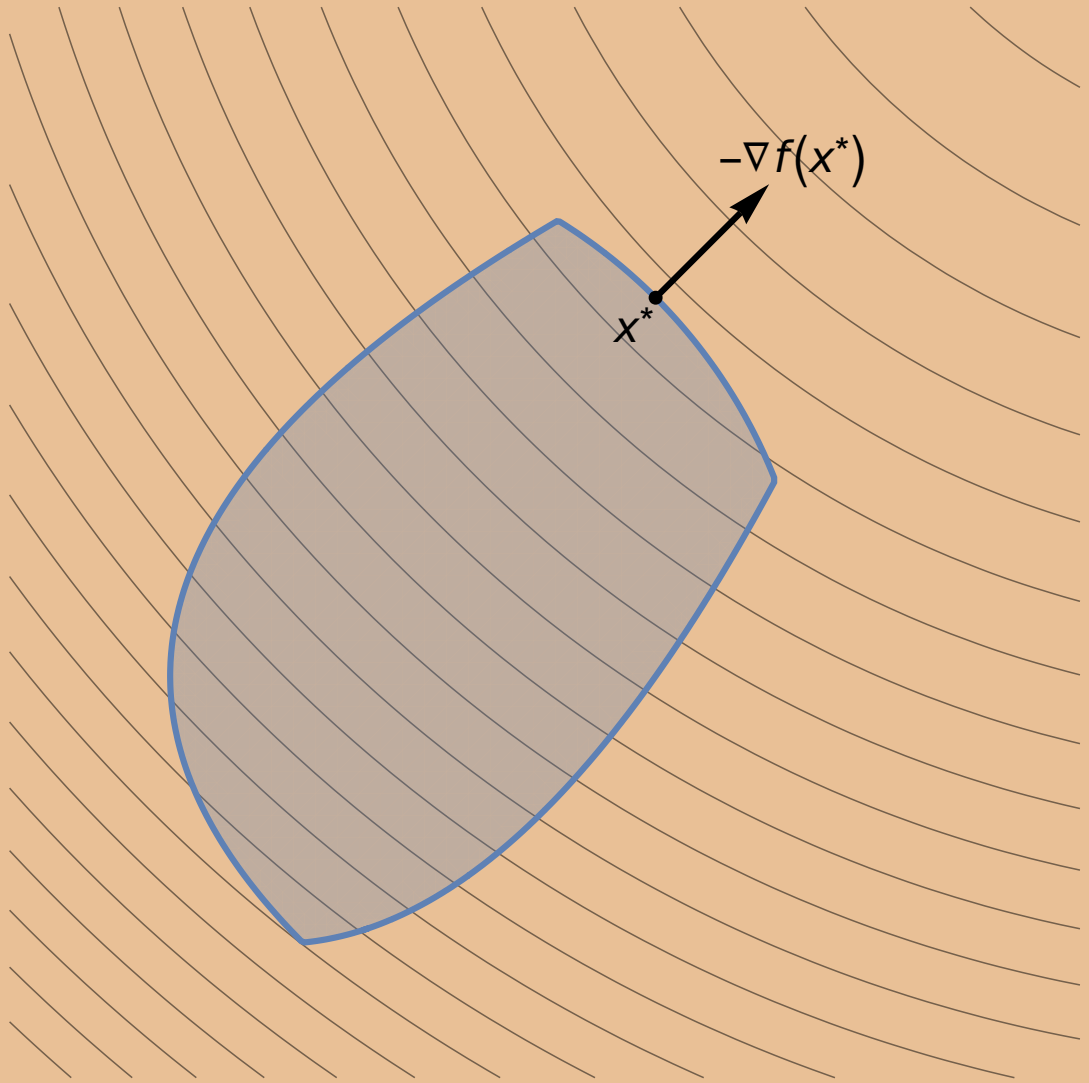


Quadratically constrained quadratic program (QCQP)

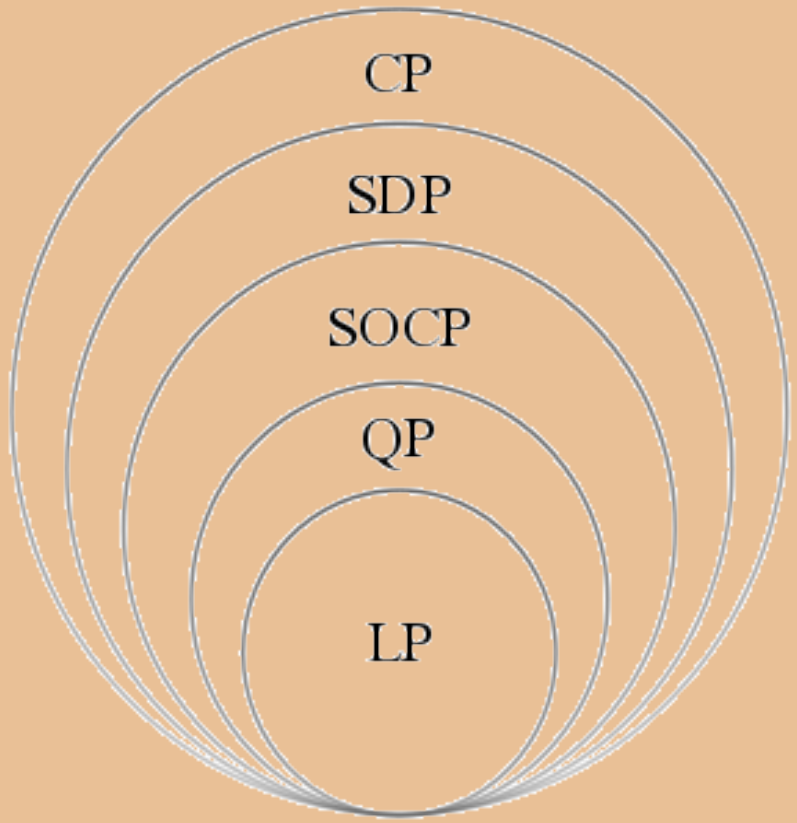
$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}x^\top Q_0 x + q_0^\top x + c_0 \\ \text{s.t.} \quad & \frac{1}{2}x^\top Q_i x + q_i^\top x + c_i \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

- If Q_i for i=0, 1, ..., m are psd then objective and constraints are convex quadratic.
- If $Q_1, \ldots, Q_m$ are pd, feasible region is the intersection of m ellipsoids and an affine set.

Quadratically constrained quadratic program (QCQP)



A hierarchy of convex optimization problems



(LP: linear programming, QP: quadratic programming, SOCP second-order cone program, SDP: semidefinite programming, CP: cone optimization.)

source: wikipedia

CVXpy modeling language

CVXpy is an open source Python-embedded modeling language for convex optimization problems. It lets you express your problem in a math way. Essentially no programming is required.

Here we present some of its basic features. A more comprehensive and complete guide can be found at the CVXpy website: `https://www.cvxpy.org/tutorial/functions/index.html` .

Atoms: CVXpy accepts only convex functions as objective and constraint functions. There are several basic convex functions, called "atoms," which are embedded in CVX. Some of these atoms are given in the following table.

Disciplined convex programming (DCP) is a system for constructing mathematical expressions with known curvature from a given library of base functions. CVXPY uses DCP to ensure that the specified optimization problems are convex. Visit dcp.stanford.edu for a more interactive introduction to DCP.

See colab: `https://colab.research.google.com/drive/1YNhCd9Xv8IIeUp6bnpdxHUn8lv9AIU7t`

CVXpy atoms

function	meaning	attributes
norm(x, p)	$\sqrt[p]{\sum_{i=1}^{n} \lvert x_i \rvert^p} (p \geq 1)$	convex
sum_squares(x)	$\sum_{i=1}^{n} x_i^2$	convex
pos(x)	$[x]^+$	convex, nonincreasing
sqrt(x)	$\sqrt{x}$	concave, nondecreasing
inv_pos(x)	$\frac{1}{x}$ $(x > 0)$	convex, nonincreasing
max(x)	$\max\{x_1, x_2, ..., x_n\}$	convex, nondecreasing
quad_over_lin(x, y)	$\frac{\lVert x \rVert^2}{y}$ $(y > 0)$	convex
quad_form(x, P)	$x^\top P x$ (P is psd)	convex
log_sum_exp(x)	$\ln(\sum_{i=1}^{n} e^{x_i})$	convex, nondecreasing
sum_largest(x, k) k=1,2,..	sum of k largest x	convex, nondecreasing

In addition, CVXpy is aware that the function x^p for an even integer p is a convex function and that affine functions are both convex and concave.

CVXpy Operations

Atoms can be incorporated by several operations which preserve convexity:

- addition,
- multiplication by a nonnegative scalar,
- composition of a nondecreasing convez function with a convex function,
- composition of a convex function with an affine transformation.

CVXpy Operations

CVXpy is also aware that minus a convex function is a concave function. The constraints that CVXpy is willing to accept are inequalities of the form

$$f(x) <= g(x)$$
$$g(x) >= f(x)$$

where f is convex and g is concave. Equality constraints must be affine, and the syntax is (h and s are affine functions)

$$h(x) == s(x)$$

Note that the equality must be written in the format $==$. Otherwise, it will be interpreted as a substitution operation.

Example - Center of a set of points

Suppose that we are given m points $a_1, a_2, ..., a_m$ in $\mathbb{R}^n$. The objective is to find the center of the minimum radius closed ball containing all the points.

In mathematical terms, the problem can be written as (r denotes that radius and x is the center)

$$\begin{aligned} \text{minimize}_{x,r} \quad & r \\ \text{s.t.} \quad & a_i \in B[x, r], \quad i{=}1, 2, ..., m\,. \end{aligned}$$

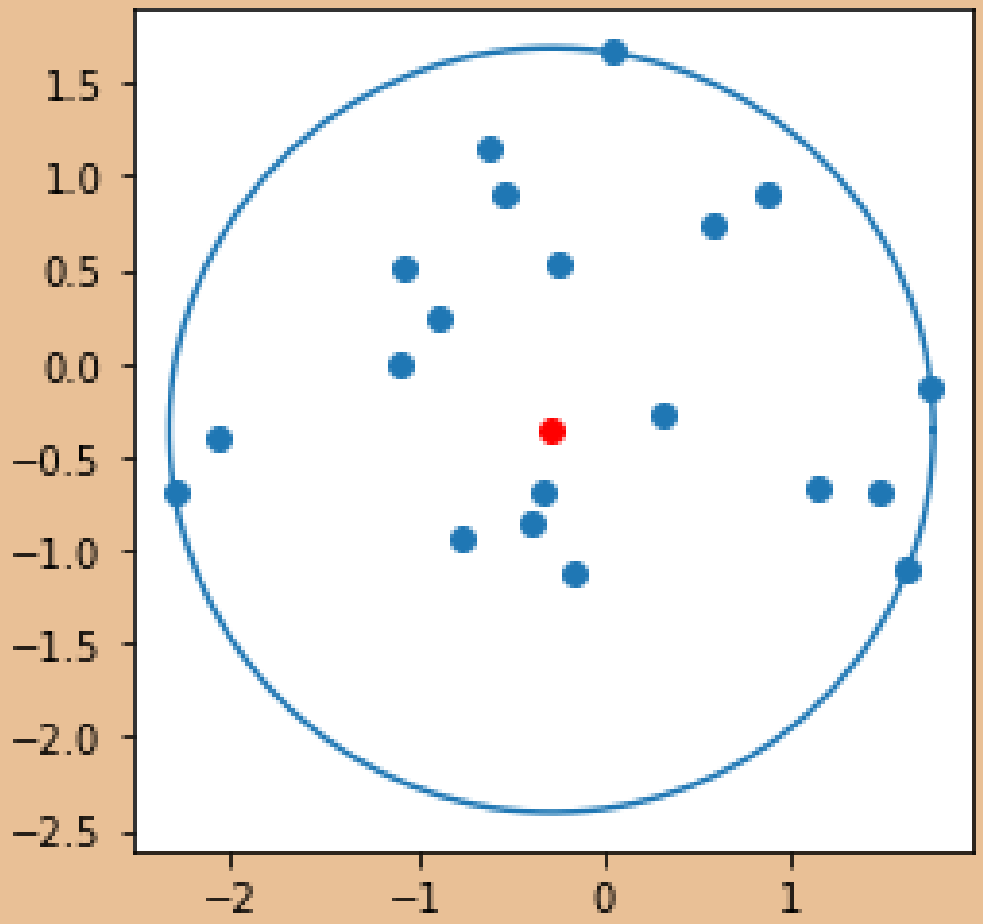
Recalling that $B[x, r] = \{y : \|y - x\| \leq r\}$, the problem can be written as

$$\begin{aligned} \text{minimize}_{x,r} \quad & r \\ \text{s.t.} \quad & \|x - a_i\| \leq r, \quad i{=}1, 2, ..., m\,. \end{aligned} \qquad \text{(Center)}$$

See colab: `https://colab.research.google.com/drive/1ozlrLwnUigYl-t75KrcF6Rue7b1lI3ik?usp=sharing`

Example - Center of a set of points

Convex problem since minimize a linear function subject to convex inequality constraints: the function $\|x - a_i\| - r$ is convex as a sum of a translation of the norm function and the linear function $-r$.



Classification via linear separators (SVMs)

Suppose that we are given two types of points in $\mathbb{R}^n$: type A and type B. The type A points are given by

$$x_1, x_2, ..., x_m \in \mathbb{R}^n$$

and the type B points are given by

$$x_{m+1}, x_{m+2}, ..., x_{m+p} \in \mathbb{R}^n.$$

The objective is to find a linear separator, which is a hyperplane of the form

$$H(a, b) = \{x \in \mathbb{R}^n : a^\top x + b = 0\}$$

for which the type A and type B points are in its opposite sides:

$$a^\top x_i + b < 0, \quad i=1, 2, ..., m,$$
$$a^\top x_i + b > 0, \quad i=m+1, m+2, ..., m+p\,.$$

Classification via linear separators (SVMs)

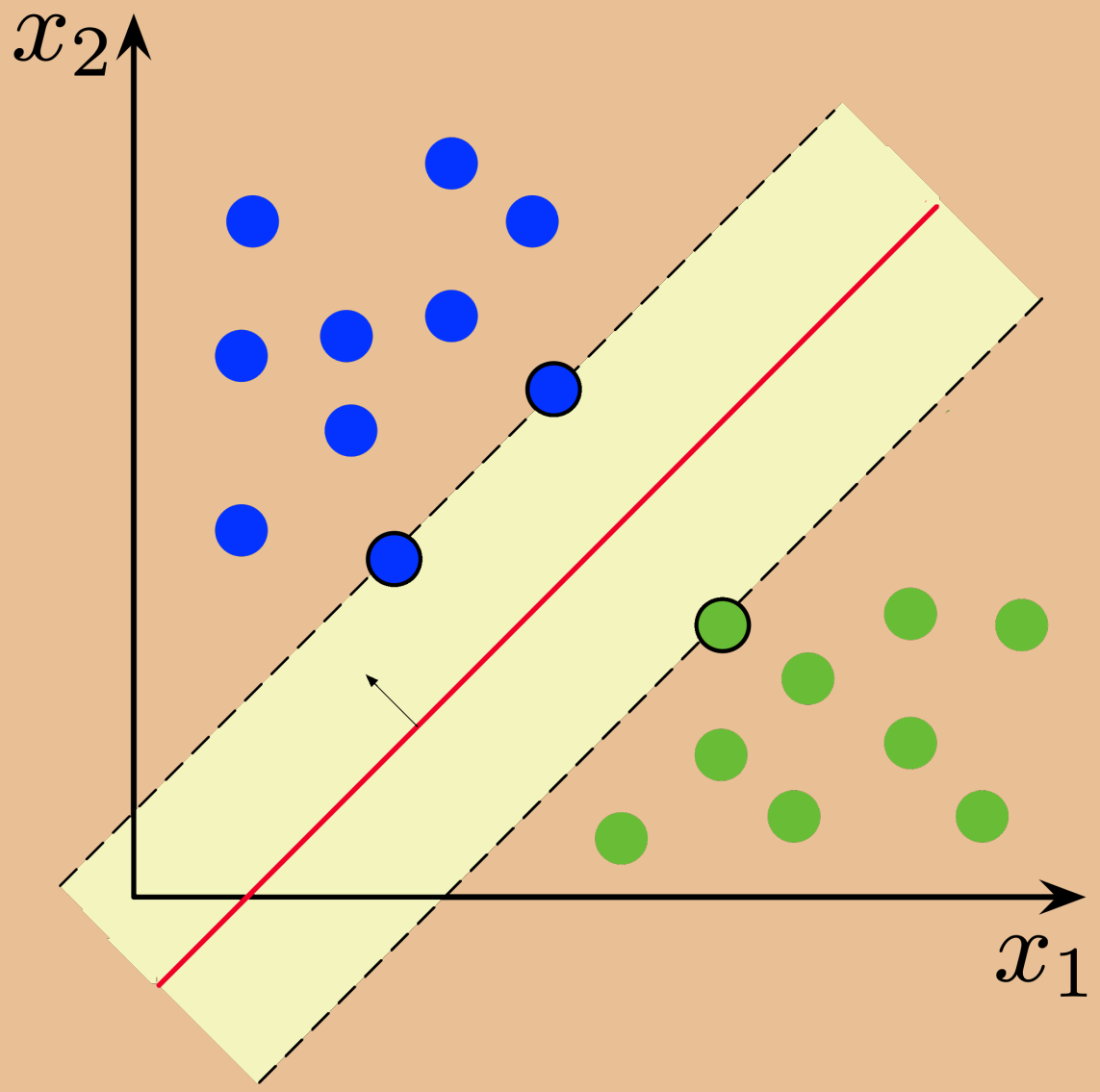


Figure: wikipedia

Classification via linear separators (SVMs)

- The underlying assumption is that the two sets of points are linearly separable, meaning that the latter set of inequalities has a solution.
- There are many linear separators, so an additiona requirement must be imposed: the line farthest as possible from all the points.
- The distance from the separator line to the closest point is called the margin.
- To compute the margin, we need to have a formula for the distance between a point and a hyperplane. The next property provides such a formula (we will see the proof later).

Classification via linear separators (SVMs)

Proposition

Let $H(\alpha, \beta) = \{x \in \mathbb{R}^n : \alpha^\top x = \beta\}$, *where* $0 \neq \alpha \in \mathbb{R}^n$ *and* $\beta \in \mathbb{R}$. *Let* $y \in \mathbb{R}^n$. *Then the distance between* y *and the set* H *is given by*

$$d(y, H) = \frac{|\alpha^\top y - \beta|}{\|\alpha\|}.$$

We therefore conclude that the margin corresponding to a hyperplane $H(a, -b)$, $a \neq 0$ is

$$\min_{i=1,2,\ldots,m+p} \frac{|a^\top x_i + b|}{\|a\|}.$$

Classification via linear separators (SVMs)

So far, the problem is

$$\begin{aligned} \text{maximize} \quad & \left\{ \min_{i=1,2,\ldots,m+p} \frac{|a^\top x_i + b|}{\|a\|} \right\} \\ \text{s.t.} \quad & a^\top x_i + b < 0, \quad i=1,2,\ldots,m, \\ & a^\top x_i + b > 0, \quad i=m+1, m+2, \ldots, m+m\,. \end{aligned}$$

This problem is not convex. The goal is to find a convex reformulation.

Classification via linear separators (SVMs)

First note that, when finding the best linear separator, the scale of a and b is free. That is, if (a, b) is an optimal solution, then so is $(\gamma a, \gamma b)$ for $\gamma \neq 0$.

One alternative is to arbitrarily choose

$$\min_{i=1,2,\ldots,m+p} |a^\top x_i + b| = 1,$$

and the problem can then be rewritten as

$$\begin{aligned} \text{maximize} \quad & \frac{1}{\|a\|} \\ \text{s.t.} \quad & \min_{i=1,2,\ldots,m+p} |a^\top x_i + b| = 1, \\ & a^\top x_i + b < 0, \quad i{=}1, 2, \ldots, m, \\ & a^\top x_i + b > 0, \quad i{=}m{+}1, m{+}2, \ldots, m{+}p\,. \end{aligned}$$

Classification via linear separators (SVMs)

The combination of the first equality and the other inequality constraints implies that a valid reformulation is

$$\begin{aligned} \text{minimize} \quad & \|a\| \\ \text{s.t.} \quad & a^\top x_i + b \leq -1, \quad i{=}1, 2, ..., m, \\ & a^\top x_i + b \geq 1, \quad i{=}m{+}1, m{+}2, ..., m{+}p\,. \end{aligned} \tag{SVM}$$

where we also used the fact that maximizing $\frac{1}{\|a\|}$ is the same as minimizing $\|a\|$ in the sense that the optimal set stays the same.

Classification via linear separators (SVMs)

The removal of the "min" constraint is valid since any feasible solution of problem (SVM) satisfies $\min_{i=1,2,\ldots,m+p} |a^\top x_i + b| \geq 1$.

If (a, b) is in additon optimal, then equality must be satisfied.

Otherwise, if $\min_{i=1,2,\ldots,m+p} |a^\top x_i + b| > 1$, then a solution with lower objective function value can be achieved.

Mathematica implementation:

`https://www.wolframcloud.com/obj/pguerra0/Published/SVM.nb`

Notice that the number of total points affects the quality of the solution.

Task: Implement SVM in Python.

Example - Portfolio selection

Suppose that an investor wishes to construct a portfolio out of n assets.

Let Y_j, j=1, 2, ..., n, be the random variable representing the return (or payoff) from asset j. We assume that the expected returns and covariances are known, that is,

$$\mu_j = \mathbb{E}[Y_j], \quad j=1, 2, ..., n,$$

and

$$\sigma_{i,j} = \mathrm{Cov}[Y_i,\ Y_j] = \mathbb{E}[(Y_i-\mu_i)(Y_j-\mu_j)], \quad i,j=1, 2, ..., n.$$

The problem has n decision variables $x_1, x_2, ..., x_n$, where x_j denotes the proportion of budget invested in asset j such that $x \in \Delta_n$.

Portfolio selection

The portfolio return is

$$R = \sum_{j=1}^{n} x_j Y_j \,,$$

whose expectation and variance are given by

$$\mathbb{E}[R] = \mu^\top x \,, \quad \mathrm{Var}[R] = x^\top C x \,,$$

where $\mu = (\mu_1, \mu_2, ..., \mu_n)^\top$ and C is the covariance matrix whose elements are given by $C_{i,j} = \sigma_{i,j}$ for $1 \leq i, j \leq n$. It is important to note that the covariance matrix is at least psd.

The variance of the portfolio, $x^\top C x$, is interpreted as the risk of the allocation x.

Next, we present 3 formulations of the portfolio optimization problem, originally proposed by Markowitz in 1952.

Portfolio selection

Recall the variance of a sum of random variables:

$$\begin{aligned}
\mathrm{Var}\left[a_1 Y_1 + a_2 Y_2\right] &= a_1 a_1 \mathrm{Var}\left[Y_1\right] + a_2 a_2 \mathrm{Var}\left[Y_2\right] + 2 a_1 a_2 \mathrm{Cov}\left[Y_1,\ Y_2\right] \\
&= a_1 a_1 \mathrm{Cov}\left[Y_1,\ Y_1\right] + a_2 a_2 \mathrm{Cov}\left[Y_2,\ Y_2\right] + 2 a_1 a_2 \mathrm{Cov}\left[Y_1,\ Y_2\right] \\
&= \sum_{i=1}^{2} \sum_{j=1}^{2} a_i a_j \mathrm{Cov}\left[Y_i,\ Y_j\right] \\
&= (a_1, a_2) \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} .
\end{aligned}$$

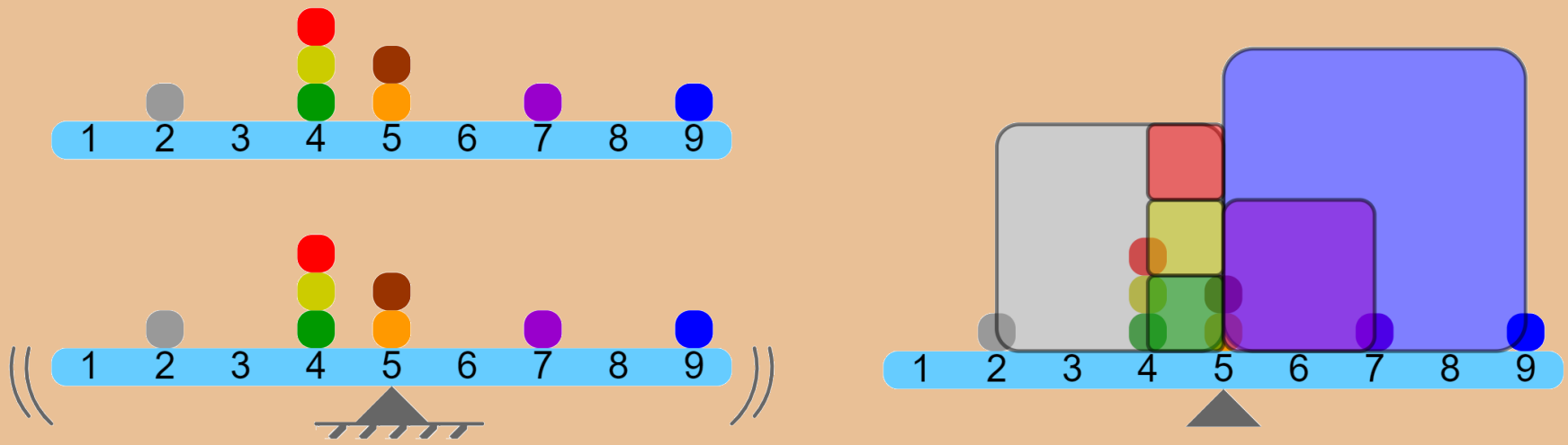


Figure: wikipedia

Markowitz portfolio models

One formulation of the problem is to find a portfolio minimizing the risk under the constraint that a minimal return level is guaranteed:

$$\begin{aligned} \text{minimize} \quad & x^\top C x \\ \text{s.t.} \quad & \mu^\top x \geq \alpha, \\ & \mathbf{1}^\top x = 1, \\ & x \geq 0, \end{aligned} \tag{1}$$

where $\mathbf{1}$ is the vector of all ones and α is the minimal required return value. Another option is to maximize the expected return subject to a bounded risk constraint:

$$\begin{aligned} \text{maximize} \quad & \mu^\top x \\ \text{s.t.} \quad & x^\top C x \leq \beta, \\ & \mathbf{1}^\top x = 1, \\ & x \geq 0, \end{aligned} \tag{2}$$

where β is the upper bound on the risk.

Markowitz portfolio models

A third option is to write an objective function which is a combination of the expected return and the risk:

$$\begin{aligned} \text{minimize} \quad & \gamma x^\top C x - \mu^\top x \\ \text{s.t.} \quad & \mathbf{1}^\top x = 1, \\ & x \geq 0, \end{aligned} \tag{3}$$

where $\gamma > 0$ is a penalty parameter for the assumed risk.

Note each of the three models (1), (2), and (3) depends on a certain paramenter ($\alpha, \beta,$ or γ) whose value dictates the tradeoff level between profit and risk.

Determining the value of each of these parameters is not necessarily an easy task, and it also depends on the subjective preferences of the investors.

Those 3 models are all convex optimization problems since $x^\top C x$ is convex in this case. The model (3) is a convex quadratic problem.

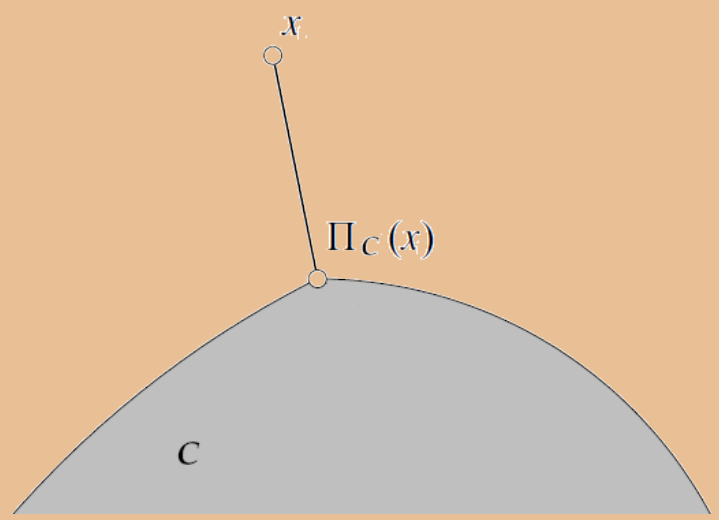
Orthogonal projection operator

In the following slides, we recall the orthogonal projection problem. This is an important convex optimization problem.

Given a nonempty closed convex set C, the orthogonal projection operator $\Pi_C : \mathbb{R}^n \to C$ is defined by

$$\Pi_C(x) := \arg\min\{\|y - x\|^2 : y \in C\}. \qquad \text{(OPO)}$$

Given the argument x, the image $\Pi_C(x)$ is the point in C that is closest to x.



Note that the orthogonal projection operator is defined as a solution of a convex optimization problem, specifically, a minimization of a convex quadratic function subject to a convex feasibility set.

Orthogonal projection revisited

The following theorem states that the orthogonal projection operator is in fact well-defined, meaning that the optimization problem in (OPO) has a unique optimal solution.

Theorem (Projection theorem)

Let C be a nonempty closed convex set. Then the problem (OPO) has a unique optimal solution.

Proof.

Since $\|y - x\|^2$ is a quadratic function with a pd matrix, it is coercive. Then, since C is convex, the minimum is attained and therefore the problem has an optimal solution. Also, since the matrix is pd the function is strictly convex, therefore the minimum is strict (unique). $\square$

Orthogonal projection revisited

Recall the distance function from a point to a convex set:

$$d_C(x) = \min_{y \in C} \|x - y\| .$$

We proved that function is convex. This function can be written in terms of the orthogonal projection:

$$d_C(x) = \|x - \Pi_C(x)\| .$$

Computing the orthogonal projection operator might be a difficult task, but there are some examples of simple sets on which the orthogonal projection can be easily computed.

Next, we solve some particular examples of this problem (that is, some particular sets C).

Orthogonal projection revisited

Example (Projection on the nonnegative orthant)

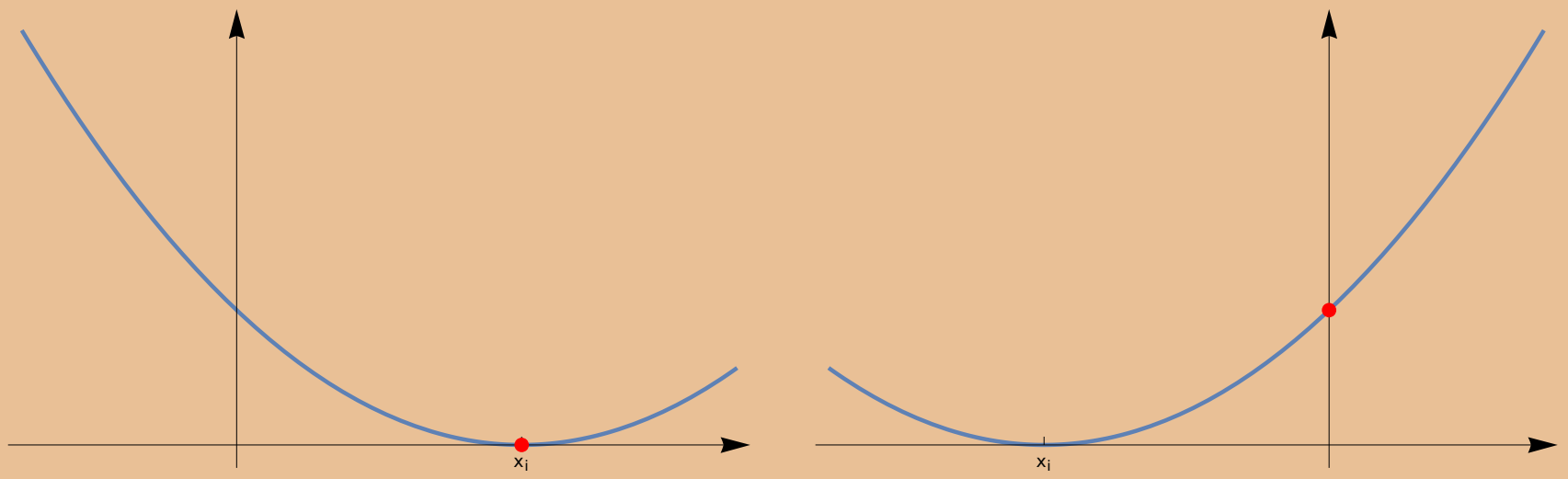
Let $C = \mathbb{R}^n_{\geq 0}$. To compute the orthogonal projection of $x \in \mathbb{R}^n$ onto $\mathbb{R}^n_{\geq 0}$, we need to solve the convex optimization problem

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & \|y - x\|^2 = \sum_{i=1}^{n} (y_i - x_i)^2 \\ s.t. \quad & y_1, y_2, ..., y_n \geq 0\,. \end{aligned} \tag{4}$$

Since this problem is separable, in the sense that the i-th component of the optimal solution y^* of problem (4) is the optimal solution of the univariate problem

$$\begin{aligned} \underset{y_i}{\text{minimize}} \quad & (y_i - x_i)^2 \\ s.t. \quad & y_i \geq 0\,. \end{aligned}$$

Projection on the nonnegative orthant



Solution of

$$\begin{aligned} \underset{y_i}{\text{minimize}} \quad & (y_i - x_i)^2 \\ \text{s.t.} \quad & y_i \geq 0\,. \end{aligned}$$

is

$$y_i^* = \max\{0, x_i\}\,.$$

Orthogonal projection operator

Example (Continued)

The solution of the univariate problem is given by $y_i^* = [x_i]_+$, *where*

$$[a]_+ := \max\{0, a\} = \begin{cases} a, & a \geq 0 \\ 0, & a < 0 \end{cases}.$$

Extending the above definition to a vector $v \in \mathbb{R}^n$ *is defined by*

$$[v]_+ := ([v_1]_+, [v_2]_+, ..., [v_n]_+)^\top .$$

Then the orthogonal projection operator onto $\mathbb{R}^n$ *is given by*

$$\Pi_{\mathbb{R}^n_{\geq 0}}(x) = [x]_+ .$$

Orthogonal projection operator

Example (Projection on boxes)

A box is a subset of $\mathbb{R}^n$ of the form

$$B = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n] = \{x \in \mathbb{R}^n : a_i \leq x_i \leq b_i\},$$

where $a_i \leq b_i$, for all i=1, 2, ..., n.

Some of the b_i's may be equal to ∞ and some of the a_i's $-\infty$.

A similar separability argument as used before shows that the orthogonal projection is given by

$$y = \Pi_B(x),$$

where

$$y_i = \begin{cases} b_i & \text{if } x_i \geq b_i \\ x_i & \text{if } a_i < x_i < b_i \\ a_i & \text{if } x_i \leq a_i \,, \end{cases}$$

for any i=1, 2, ..., n.

Orthogonal projection revisited

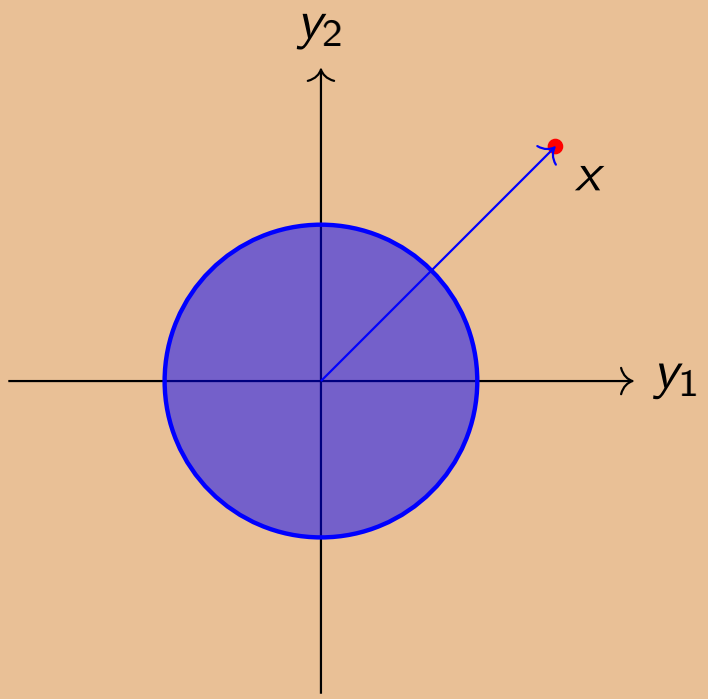
Example (Projection on 2-norm balls)

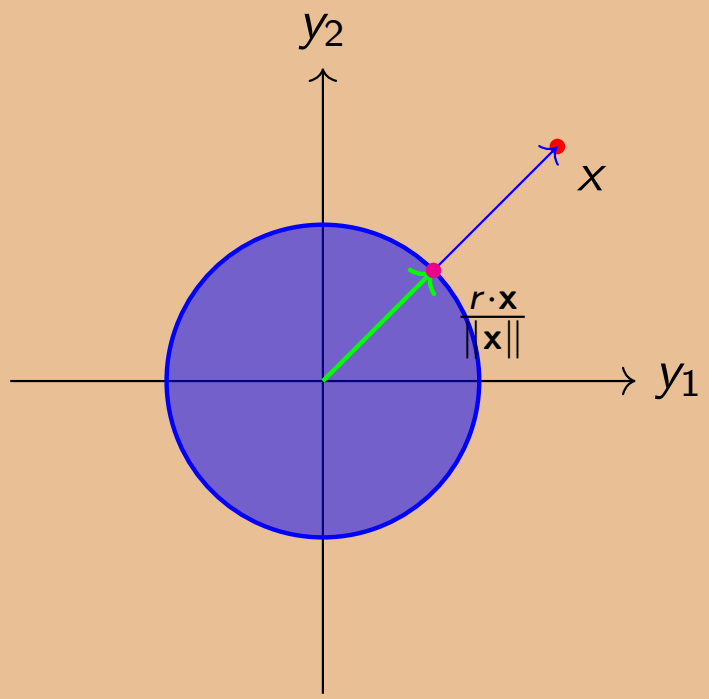
Let $C = B[0, r] = \{y \in \mathbb{R}^n : \|y\| \leq r\}$, *for* $r > 0$. *The optimization problem associated with the computation of* $\Pi_C(x)$ *is given by*

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & \|y - x\|^2 \\ \text{s.t.} \quad & \|y\|^2 \leq r^2 \,. \end{aligned}$$

If $\|x\| \leq r$, *then obviously* $y - x$ *is the optimal solution since it corresponds to the optimal value* 0. *When* $\|x\| > r$, *then the optimal solution must belong to the boundary of the ball since otherwise, by the first order necessary condition, it would be a stationary point of the objective function, that is,* $2(y - x) = 0$, *and hence* $y = x$, *which is impossible since* $x \notin C$.

Projection on balls





Orthogonal projection revisited

Example

We thus conclude that the problem in this case is equivalent to

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & \|y - x\|^2 \\ s.t. \quad & \|y\|^2 = r^2 \,, \end{aligned}$$

which can be equivalently written as

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & -2x^\top y + r^2 + \|x\|^2 \\ s.t. \quad & \|y\|^2 = r^2 \,. \end{aligned}$$

Orthogonal projection revisited

Example (Continued)

The optimal solution y^* of the above problem is the same as the optimal solution of

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & -2x^\top y \\ \text{s.t.} \quad & \|y\|^2 = r^2 \,, \end{aligned}$$

By the Cauchy-Schwarz inequality, the objective function can be bounded by

$$-2x^\top y \geq -2\|x\|\|y\| = -2r\|x\| \,,$$

and on the other hand, this lower bound is attained at $y = r\frac{x}{\|x\|}$, and hence the orthogonal projection is given by

$$\Pi_{B[0,r]} = \begin{cases} x & \text{if } \|x\| \leq r \\ r\frac{x}{\|x\|} & \text{if } \|x\| > r \,. \end{cases}$$

Task: what if the center is c?

Subgradient optimality condition

Introduction [2]

"The classical theory of optimization has always been connected to differentiation and strong regularity assumptions. However, these assumptions are often too demanding for practical applications, due to the nonsmoothness of natural phenomena".

"Everyone who has dealt with optimization knows that convexity is a pleasant property and that optimization has a very clear geometrical interpretation. Thus it was natural that the theory was first developed for convex functions and the treatment was quite geometrical".

"In optimization theory the meaning of differentiation is to locally linearize the given differentiable function in the sense that the hyperplane generated by the gradient is the tangent plane of the graph of the function. For convex functions these linearizations are always lower approximations".

---

[2] Extracted from the introduction of the book *"Nonsmooth optimization"* by Makela & Neittaanmaki.

Introduction

"These ideas were generalized for nonsmooth convex functions by defining the concepts of subgradient and subdifferential".

"A subgradient at a fixed point is a vector which has the property that the hyperplane at that point generated by the vector is a lower approximation to the function; the set of all subgradients at that point is called sub differential. These new concepts made it possible to obtain the same approximation properties as in the smooth case".

Subgradients

Recall for $f \in C^1$ convex we have the characterization

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \quad \forall x, y \in \operatorname{dom}(f).$$

We can generalize this idea.

Definition

A subgradient of a convex function f at an interior point x is any vector $g \in \mathbb{R}^n$ such that

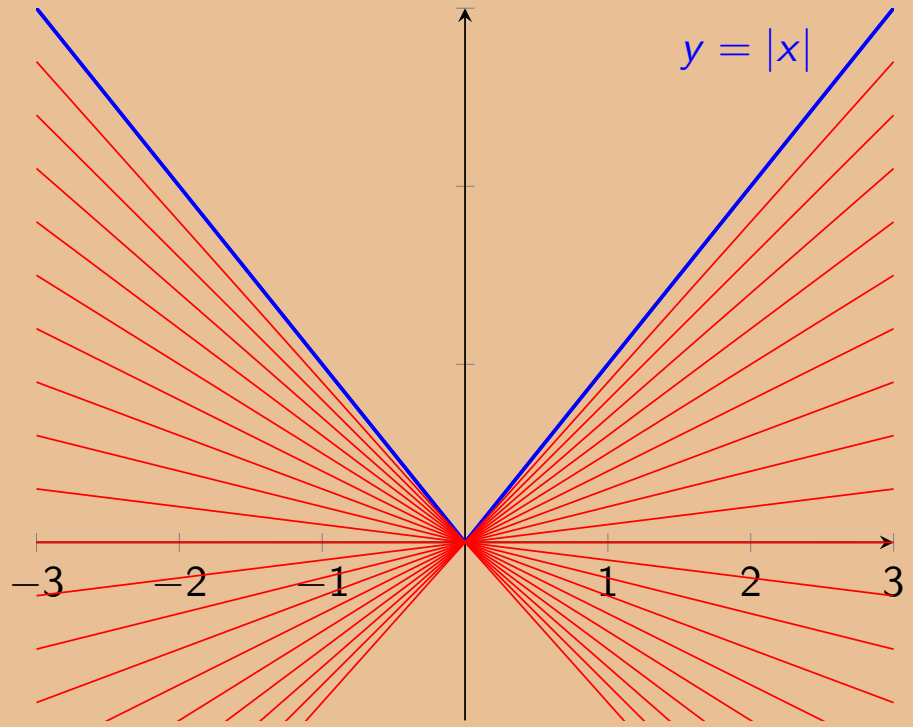
$$f(y) \geq f(x) + g^\top (y - x) \quad \forall y \in \operatorname{dom}(f).$$

Notes:

- If f differentiable at x, then $g = \nabla f(x)$.
- It can be generalized for nonconvex functions.
- Relevance: optimality characterization for nondifferentiable convex functions.

Subgradients - Example

Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$.



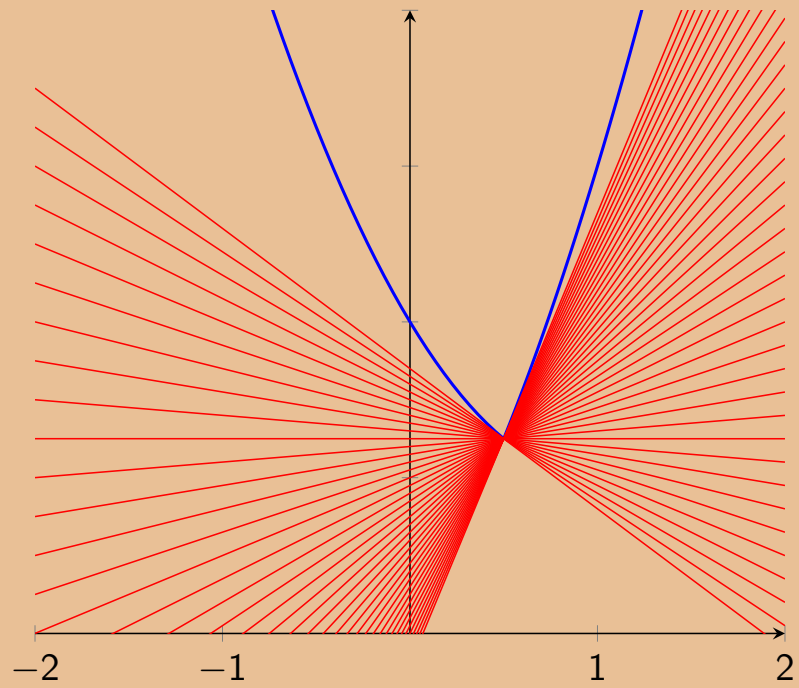
- For $x > 0$, $g = f'(x) = 1$.
- For $x < 0$, $g = f'(x) = -1$.
- In summary, for $x \neq 0$, $g = \text{sign}(x)$.
- For $x = 0$, g is any point on the interval $[-1, 1]$.

Subgradients - Example

Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable, and define

$$f(x) = \max\{f_1(x), f_2(x)\}\,.$$

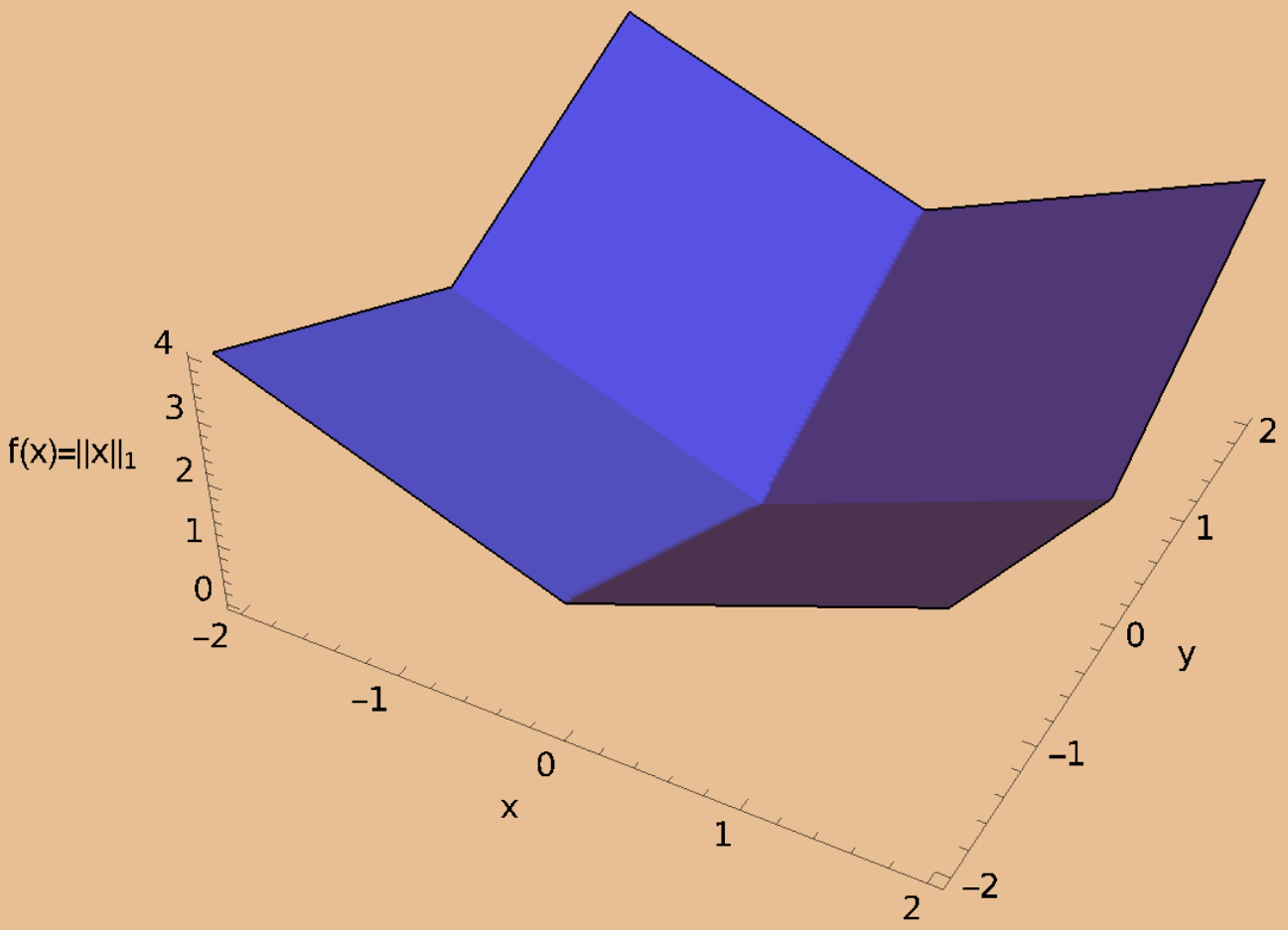
Illustration in $\mathbb{R}$:



- For x such that $f_1(x) > f_2(x)$, $g = \nabla f_1(x)$.
- For x such that $f_2(x) > f_1(x)$, $g = \nabla f_2(x)$.
- For $f_1(x) > f_2(x)$, g is any point on the interval $[\nabla f_1(x), \nabla f_2(x)]$.

Subgradients - Example

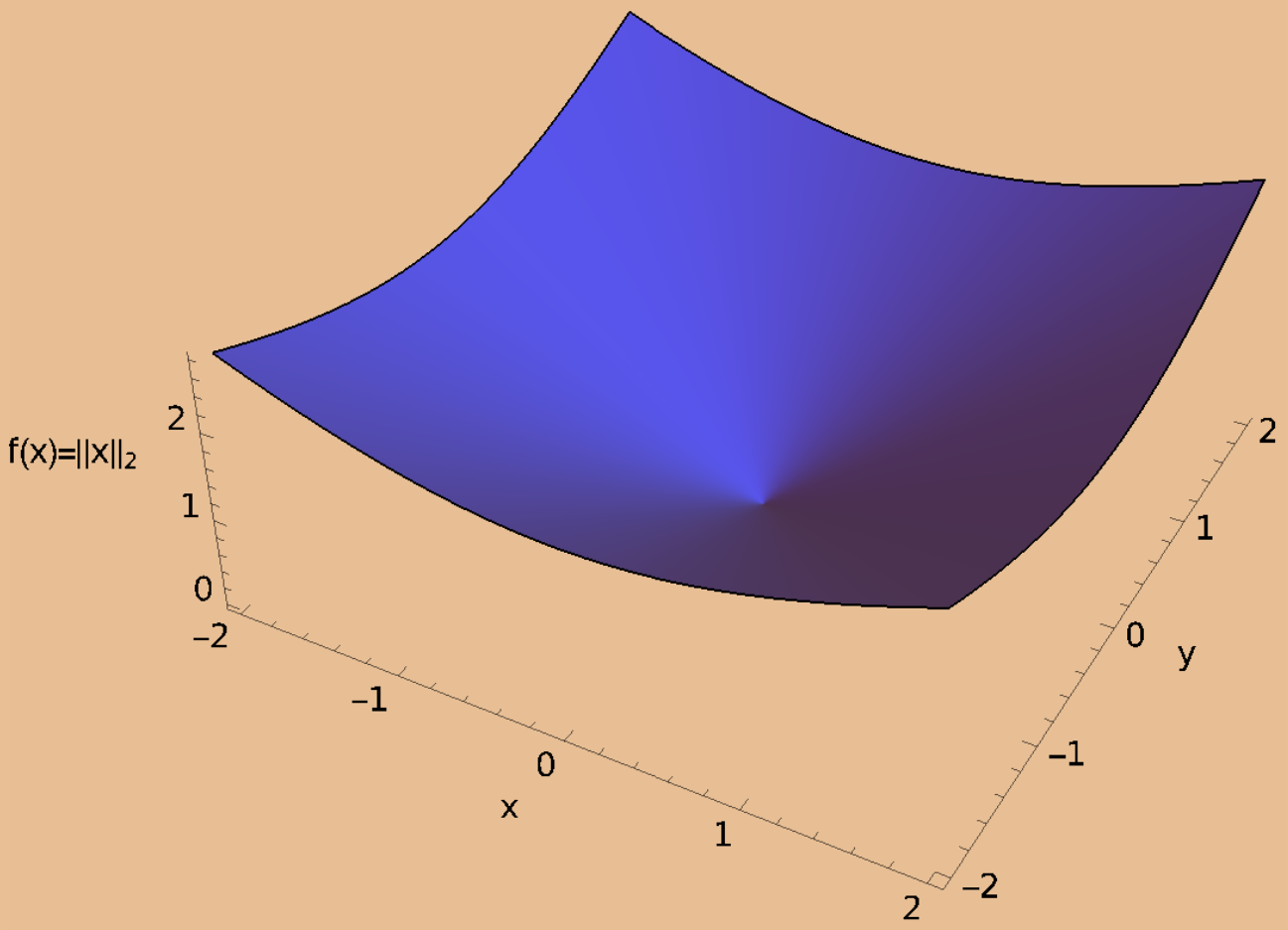
Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$.



- For $x \neq 0$, unique subgradient $g_i = \text{sign}(x_i)$.
- For $x = 0$, i-th component g_i is any element of $[-1, 1]$.

Subgradients - Example

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$.



- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$.
- For $x = 0$, subgradient g is any element of $\{z : \|z\|_2 \leq 1\}$.

Subdifferential

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ convex. The set

$$\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x\},$$

is called the subdifferential of f at x.

Notes:

- $\partial f(x)$ is closed and convex.
- Nonempty.
- If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$. Conversely, if $\partial f(x) = \{g\}$, then f is differentiable at x and $\nabla f(x) = g$.

Subdifferential - Geometric view

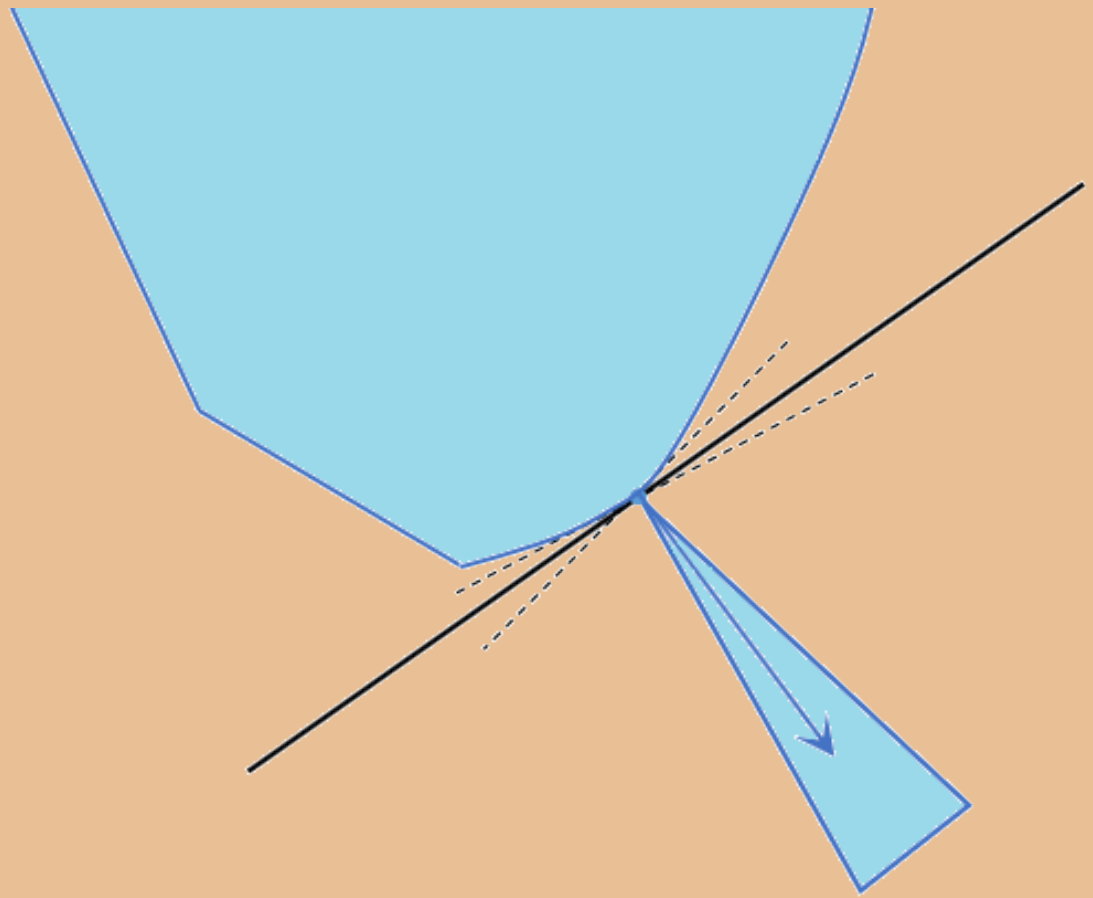
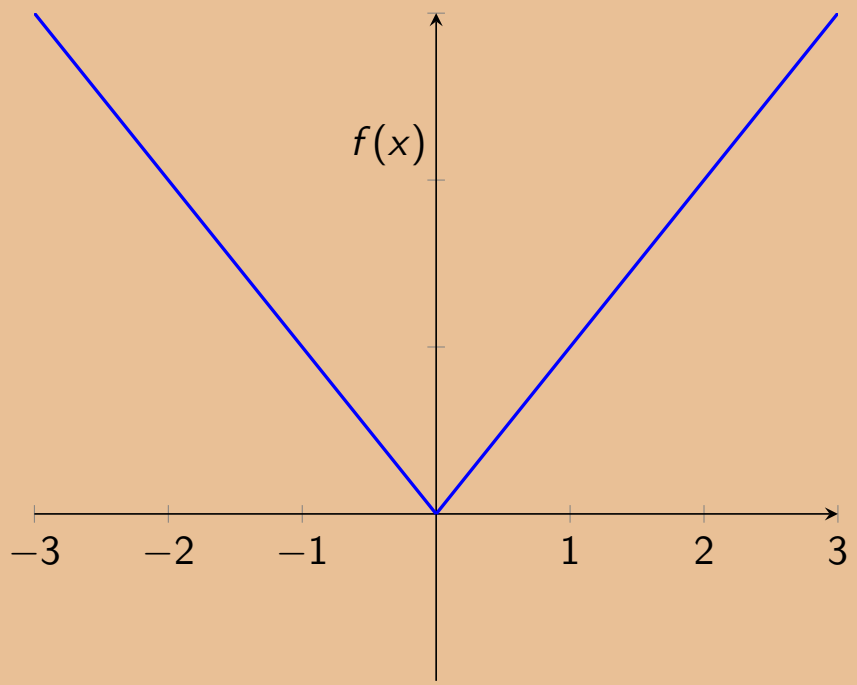


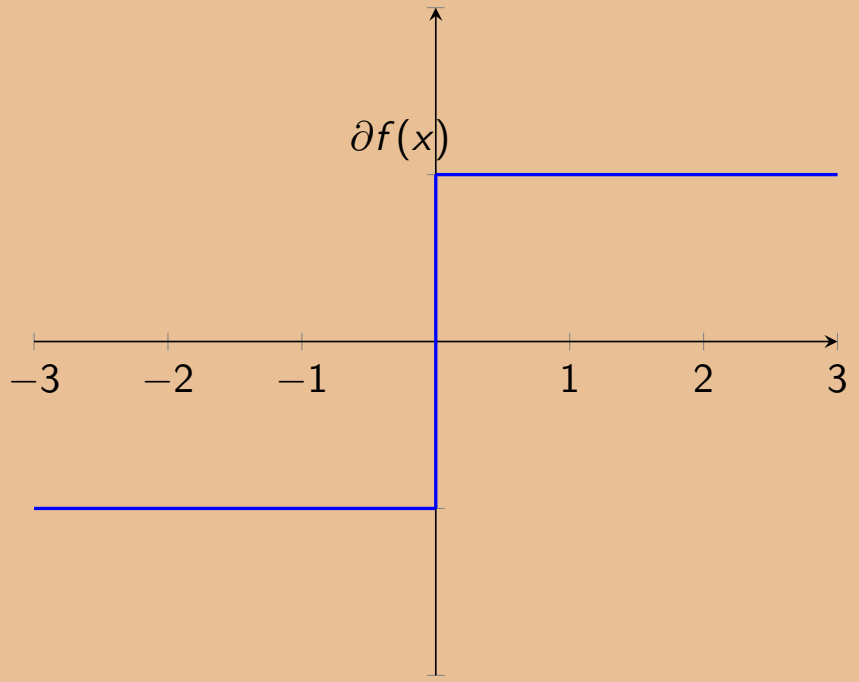
Illustration: Suppose $g \in \partial f(x)$. The definition of subgradient means that the epigraph of f is located on or above the graph of the affine function $h(y) = f(x) + g^\top (y - x)$.

Subdifferential - Examples

Let $f(x) = |x|$. Then

$$\partial f(x) = \begin{cases} \{-1\}, & \text{if } x < 0 \\ [-1, 1], & \text{if } x = 0 \\ \{1\}, & \text{if } x > 0 \end{cases}.$$

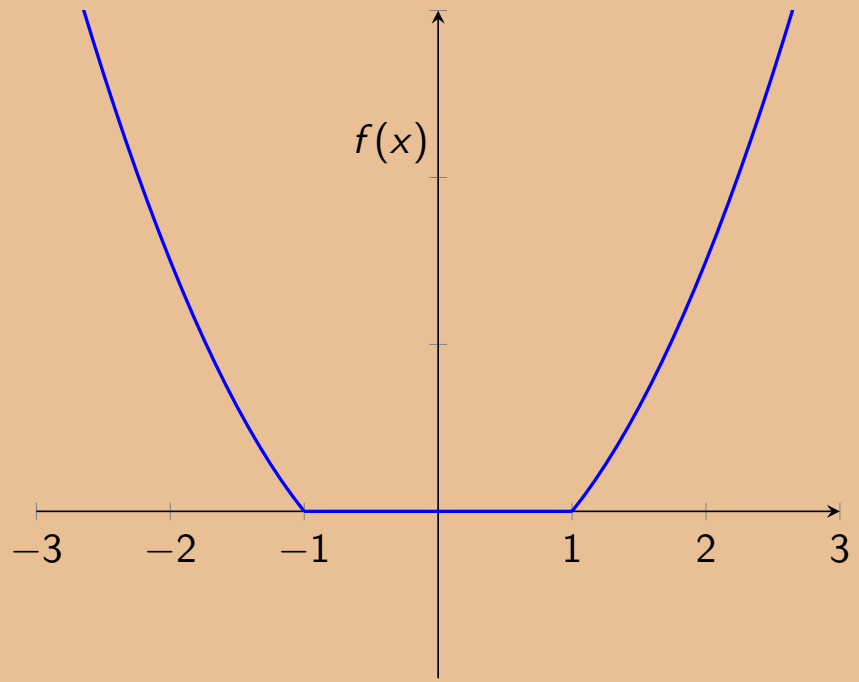


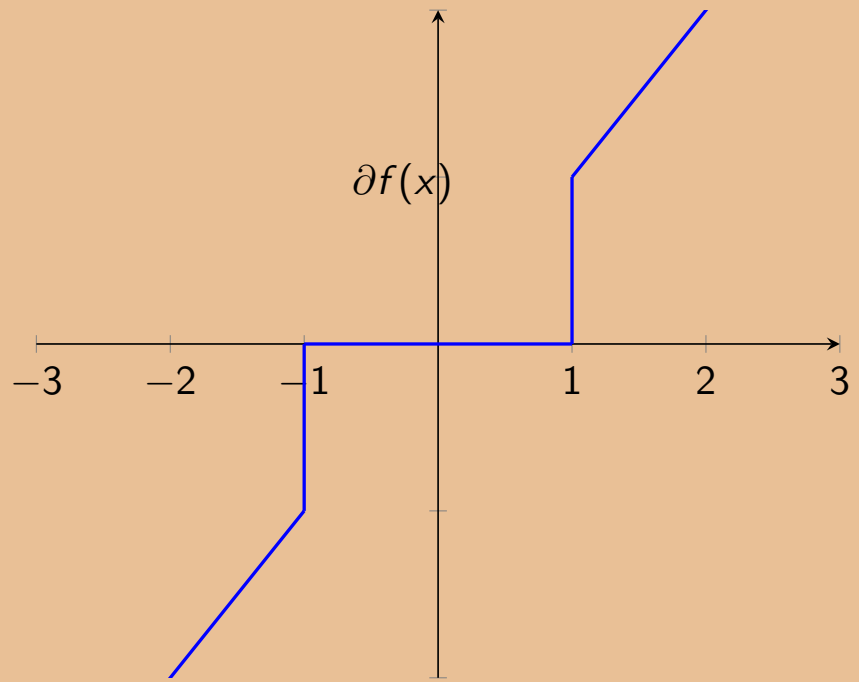


Subdifferential - Examples

Let $f(x) = \max\{0, \frac{1}{2}(x^2 - 1)\}$. Then

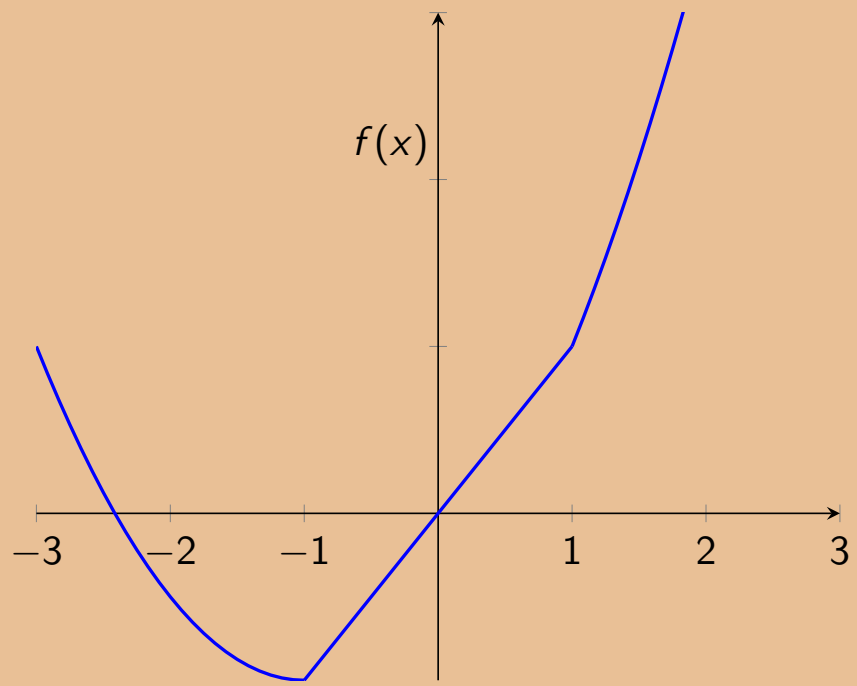
$$\partial f(x) = \begin{cases} \{x\}, & \text{if } x < -1 \\ [-1, 0], & \text{if } x = -1 \\ \{0\}, & \text{if } -1 < x < 1 \\ [0, 1], & \text{if } x = 1 \\ \{x\}, & \text{if } x > 1 \end{cases}.$$

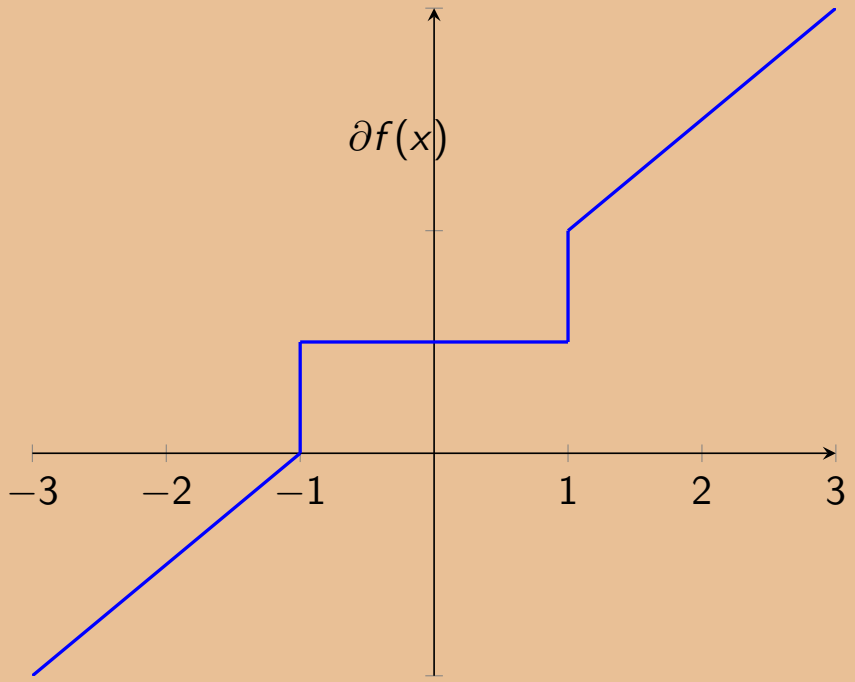




Subdifferential - Examples

Let $f(x) = x + \max\{0, \frac{1}{2}(x^2 - 1)\}$.





Subdifferential calculus

Basic properties (recall ∂f is a set): Let f be convex.

- Scaling: $\partial(af) = a\partial f$ provided $a > 0$.
- Addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$.
- Affine composition: If $g(x) = f(Ax + b)$, then

$$\partial g(x) = A^\top \partial f(Ax + b)\,.$$

- Finite pointwise maximum: If $f(x) = \max_{i=1,\dots,m} f_i(x)$, then

$$\partial f(x) = \operatorname{Conv}\left(\bigcup_{i:f_i(x)=f(x)} \partial f_i(x)\right),$$

  the convex hull of union of subdifferentials of all functions that achieve the maximum at x. Example: if only f_j achieves the max at x then $\partial f(x) = \partial f_j(x)$.

- There are some more but we postpone for later.

Subgradient optimality condition

Proposition (Subgradient optimality condition)

Let f convex. Then

$$x^* \in \arg\min_{x \in \mathbb{R}^n} f(x) \quad \textit{iff} \quad 0 \in \partial f(x^*)\,.$$

That is, x^* is a minimizer of f iff the zero vector 0 is a subgradient of f at x^*.

Why? $g = 0$ being a subgradient means that $\forall y$

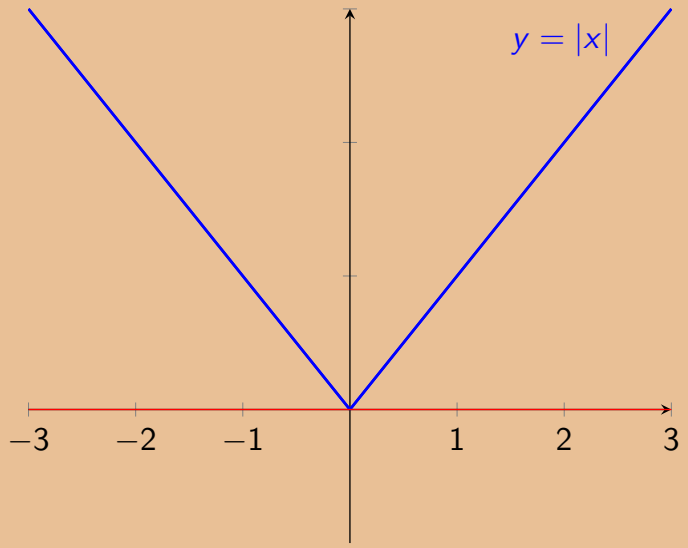
$$f(y) \geq f(x^*) + 0^\top (y - x^*)\,.$$

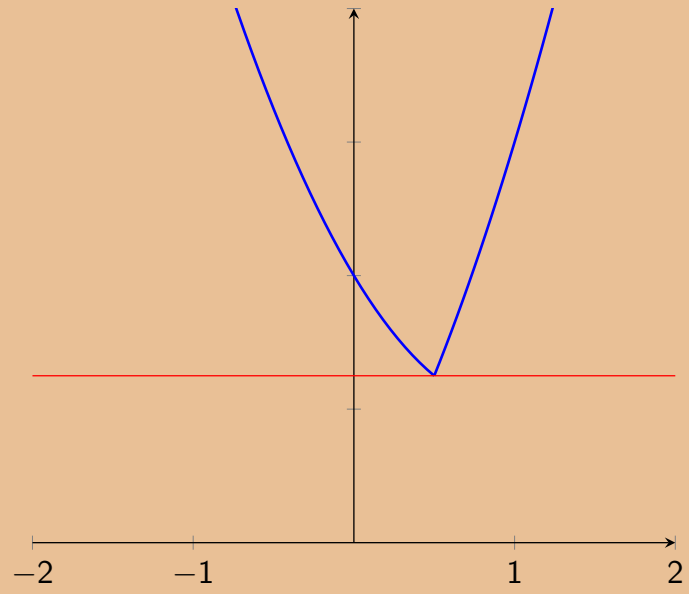
Note that, if f is convex and differentiable, then $\partial f(x) = \{\nabla f(x)\}$ and thus we recover the basic optimality condition

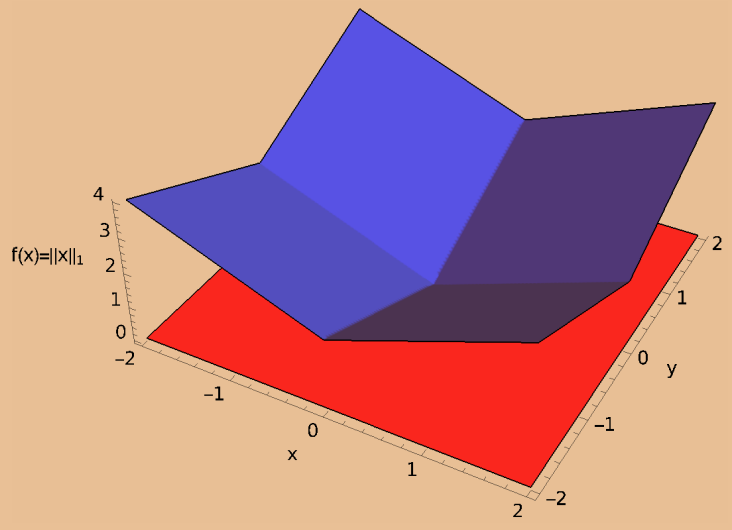
$$x^* \in \arg\min_{x \in \mathbb{R}^n} f(x) \quad \text{iff} \quad \nabla f(x^*) = 0\,.$$

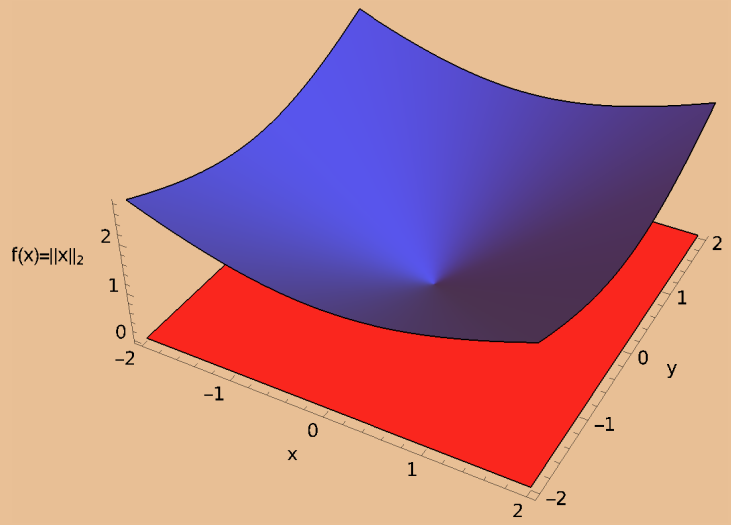
Subgradient optimality condition

In the previous examples:









Subgradient optimality condition - Example

Example: Let

$$f(x, y) = x + y + \max(0, x^2 + y^2 - 1)\,.$$

This function is not C^1.

We have: Region A: $x^2 + y^2 - 1 < 0$. Here, the function simplifies to $f_A(x, y) = x + y$. Is easy to prove that, in this open set, there are no points that satisfy $\nabla f_A = 0$.

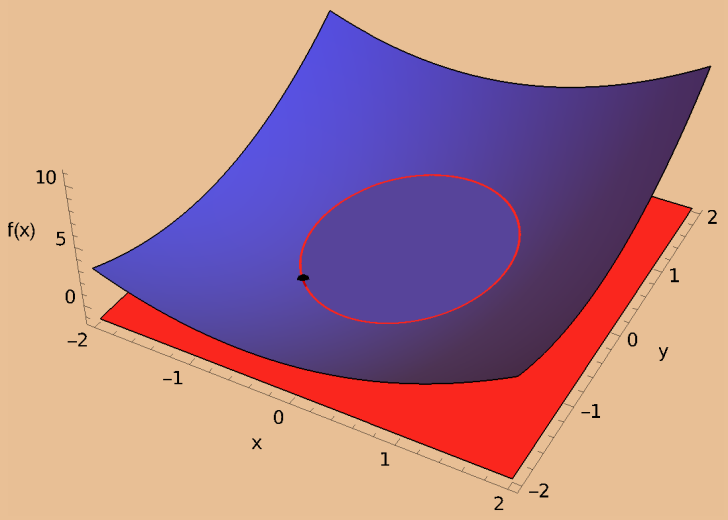
Region B: $x^2 + y^2 - 1 > 0$. Here, the function is $f_B(x, y) = x + y + (x^2 + y^2 - 1)$. In this open set there are no points that satisfy $\nabla f_B = 0$.

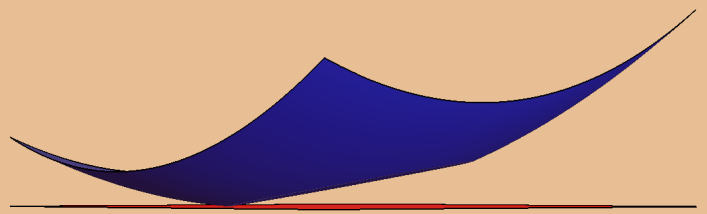
So we are left with the boundary $x^2 + y^2 - 1 = 0$, which is the unit circle. On this circle, $f(x, y) = x + y$.

Subgradient optimality condition - Example

We solved this problem before. (minimize $c^\top x$ s.t. $x^2+y^2=1$). The solution is:

$$x^* = (-1/\sqrt{2}, -1/\sqrt{2})\,.$$





Subgradient optimality condition - Example

We can visualize that the minimizer is achieved at the point c where $0 \in \partial f(c)$. We can also use subdifferential properties. We start with the inclusion

$$0 \in \partial f \quad \Longleftrightarrow \quad 0 \in \partial(x + y + \max(0, x^2 + y^2 - 1)).$$

Using addition property we have

$$0 \in \partial(x + y) + \partial(\max(0, x^2 + y^2 - 1)).$$

Let $g(x, y) = x + y$ and $h(x, y) = \max(0, x^2 + y^2 - 1)$. Since $g(x, t) \in C^1$, the subdifferential $\partial(x + y) = \{\nabla(x + y)\} = \{(1, 1)^\top\}$ is a singleton set.

On the other hand, we have that $\partial(\max(0, x^2 + y^2 - 1))$ is the subdifferential of the pointwise maximum of two C^1 functions. That is, the convex hull of $\{\nabla(0), \nabla(x^2 + y^2 - 1)\} = \{(0, 0)^\top, (2x, 2y)^\top\}$.

Subgradient optimality condition - Example

The subgradient optimality condition $0 \in \partial f(x, y)$ means there exist subgradients $v_g \in \partial g(x, y)$ and $v_h \in \partial h(x, y)$ such that $v_g + v_h = 0$. Given that $v_g = (1, 1)^T$, we want to find $v_h \in \partial h(x, y)$ such that:

$$(1, 1)^T + v_h = (0, 0)^T$$

This means $v_h = (-1, -1)^T$.
If $x^2 + y^2 - 1 = 0$, v_h can be any convex combination of $(0, 0)^T$ and $(2x, 2y)^T$. Therefore, we set $v_h = \lambda(0, 0)^T + (1 - \lambda)(2x, 2y)^T$ and solve for λ and (x, y) such that $v_h = (-1, -1)^T$. We have

$$(1 - \lambda)(2x, 2y)^T = (-1, -1)^T .$$

The system implies $x = y$. We have $x^2 + y^2 = 1$ so $2x^2 = 1$ and $x = y = -1/\sqrt{2}$ (the negative root is chosen because the subgradient $(2x, 2y)$ needs to be in the opposite direction of $(1, 1)$).

Subgradient optimality condition - Example

Now, we check if $(x, y) = (-1/\sqrt{2}, -1/\sqrt{2})$ satisfies the optimality condition with an appropriate λ:

$$(1-\lambda)(2(-1/\sqrt{2}), 2(-1/\sqrt{2})) = (-1, -1)$$

$$(1-\lambda)(-2/\sqrt{2}, -2/\sqrt{2}) = (-1, -1)$$

$$(1-\lambda)(-\sqrt{2}, -\sqrt{2}) = (-1, -1)$$

$$(1-\lambda) = 1/\sqrt{2}$$

$$\lambda = 1 - 1/\sqrt{2}$$

So, for $\lambda = 1 - 1/\sqrt{2}$, $(x, y) = (-1/\sqrt{2}, -1/\sqrt{2})$ satisfies the subgradient optimality condition, which means $(-1/\sqrt{2}, -1/\sqrt{2})$ is indeed a solution to the minimization problem.

Example - Particular case of the finite pointwise maximum

Example: The following result shows the finite pointwise maximum property in a particular case:

Let

$$f(x) = \max_{i \in I} f_i(x),$$

where I is a finite set, and f_i are convex and differentiable functions for each $i \in I$.

Define $\mathcal{I}(x) = \{f_i \in I : f_i(x) = f(x)\}$, then

$$\partial f(x) = \text{Conv}\left(\{\nabla f_i(x) : i \in \mathcal{I}(x)\}\right).$$

References

- Axler, S. (2014).
  Linear Algebra Done Right.
  Springer.
- Bazaraa, M. S. and Shetty, C. M. (1979).
  Nonlinear Programming: Theory and Algorithms.
  Wiley, 1 edition.
- Beck, A. (2014).
  Introduction to Nonlinear Optimization.
  MOS-SIAM Series on Optimization. SIAM.
- Bertsekas, D. P. (1999).
  Nonlinear Programming.
  Athena Scientific, 2 edition.
- Boyd, S. and Vandenberghe, L. (2004).
  Convex Optimization.
  Cambridge University Press.
- Chong, E. K. P. and Żak, S. H. (2013).
  An Introduction to Optimization.
  Wiley, 4 edition.
- Fletcher, R. (1987).
  Practical Methods of Optimization.
  John Wiley & Sons, New York, NY, USA, second edition.
- Nocedal, J. and Wright, S. (2006).
  Numerical Optimization.
  Springer.
- Ruszczynski, A. (2006).
  Nonlinear Optimization.
  Princeton University Press.