CS 2601 Linear and Convex Optimization

1. Introduction

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

Fall 2022

Outline

- Examples of optimization problems

- Global and local optima

Mathematical optimization problems

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in X \end{array} \qquad \text{or} \qquad \min_{\boldsymbol{x} \in X} f(\boldsymbol{x})$$

- $f : \mathbb{R}^n \to \mathbb{R}$: objective function
- $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$: optimization/decision variables
- $X \subset \mathbb{R}^n$: feasible set or constraint set
  - $\boldsymbol{x}$ is called feasible if $\boldsymbol{x} \in X$ and infeasible if $\boldsymbol{x} \notin X$.

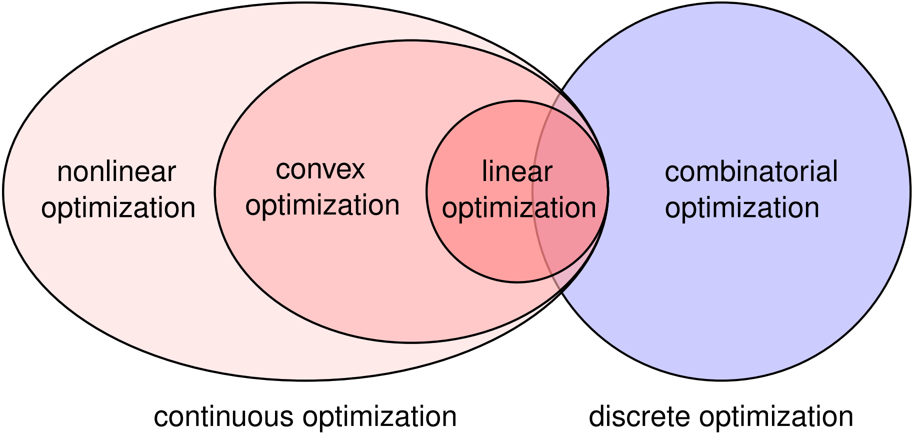
Maximizing f is equivalent to minimizing $-f$; will focus on minimization.

The problem is unconstrained if $X = \mathbb{R}^n$ and constrained if $X \neq \mathbb{R}^n$.

X is often specified by constraint functions,

$$\begin{array}{lll} \min\limits_{\boldsymbol{x}} & f(\boldsymbol{x}) & \\ \text{s. t.} & g_i(\boldsymbol{x}) \leq 0, & i = 1, 2, \ldots, m \\ & h_i(\boldsymbol{x}) = 0, & i = 1, 2, \ldots, k \end{array}$$

Mathematical optimization problems



Data fitting

Recall Hooke's law in physics,

$$F = -k(x - x_0) = -kx + b, \quad \text{where } b = kx_0$$

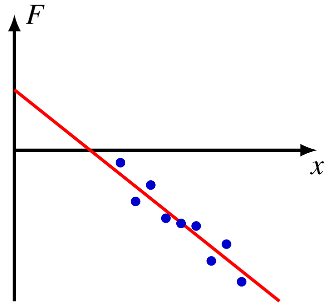
- F : force
- k : spring constant
- x : length
- x_0 : length at rest

Given m measurements $(x_1, F_1), (x_2, F_2), \ldots, (x_m, F_m)$,

$$F_i = -kx_i + b + \epsilon_i$$

- ϵ_i : measurement error

find k, b by fitting a line through data.



Least squares criterion,

$$\min_{k>0, b>0} \sum_{i=1}^{m} \epsilon_i^2 = \sum_{i=1}^{m} (F_i + kx_i - b)^2$$

Linear least squares regression

A linear model predicts a response/target by a linear combination of predictors/features (plus an intercept/bias),

$$\hat{y} = f(\boldsymbol{x}) = b + \sum_{i=1}^{n} w_i x_i = \boldsymbol{x}^T \boldsymbol{w} + b$$

Given m data points $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)$, linear (least squares) regression finds $\boldsymbol{w}$ and b by minimizing the sum of squared errors,

$$\min_{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}} \sum_{i=1}^{m} (f(\boldsymbol{x}_i) - y_i)^2 = \sum_{i=1}^{m} (\boldsymbol{x}_i^T \boldsymbol{w} + b - y_i)^2$$

In a more compact form,

$$\min_{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}} \|\boldsymbol{X}\boldsymbol{w} + b\boldsymbol{1} - \boldsymbol{y}\|^2$$

- $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)^T \in \mathbb{R}^{m \times n}$, $\boldsymbol{y} = (y_1, \ldots, y_m)^T \in \mathbb{R}^m$
- $\boldsymbol{1} = (1, 1, \ldots, 1)^T \in \mathbb{R}^m$
- $\|\boldsymbol{z}\| = \sqrt{\boldsymbol{z}^T \boldsymbol{z}} = \sqrt{\sum_{i=1}^{n} z_i^2}$ for $\boldsymbol{z} = (z_1, \ldots, z_n)^T \in \mathbb{R}^n$

Optimal transport problem

- need to ship products from n warehouses to m customers
- inventory at warehouse i is a_i, $i = 1, 2, \ldots, n$
- quantity ordered by customer j is b_j, $j = 1, 2, \ldots, m$
- unit shipping cost from warehouse i to customer j is c_{ij}

Let x_{ij} be quantity shipped from warehouse i to customer j

Minimize total cost by solving the following linear program

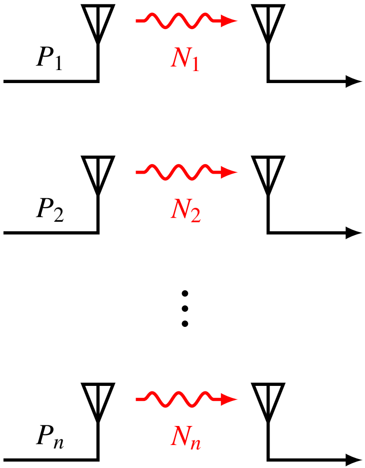
$$
\begin{array}{lll}
\min\limits_{(x_{ij})} & \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} & \\
\text{s.t.} & \displaystyle\sum_{i=1}^{n} x_{ij} = b_j & \text{for} \quad j = 1, 2, \ldots, m \\
& \displaystyle\sum_{j=1}^{m} x_{ij} \leq a_i & \text{for} \quad i = 1, 2, \ldots, n \\
& x_{ij} \geq 0 & \text{for} \quad i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m
\end{array}
$$

Power allocation

We want to transmit information over n communication channels. The capacity of the i-th channel is

$$C_i = W_i \log_2(1 + \frac{P_i}{N_i}) \quad \text{bits/second}$$

- W_i: channel bandwidth in hertz
- P_i: signal power in watts
- N_i: noise power in watts



Given a total power constraint P,

$$\begin{aligned} \max_{P_1,\ldots,P_n} \quad & \sum_{i=1}^{n} W_i \log_2(1 + \frac{P_i}{N_i}) \\ \text{s.t.} \quad & \sum_{i=1}^{n} P_i \leq P, \quad P_i \geq 0 \quad \text{for } i = 1, 2, \ldots, n \end{aligned}$$

Binary classification

vs

Represent an image by a vector $\boldsymbol{x} \in \mathbb{R}^n$, label $y \in \{+1, -1\}$

Given a set of images with labels $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)$, want a function $f : \mathbb{R}^n \to \mathbb{R}$, called a classifier, such that

$$\begin{cases} f(\boldsymbol{x}_i) > 0, & \text{iff } y_i = +1 \\ f(\boldsymbol{x}_i) < 0, & \text{iff } y_i = -1 \end{cases} \iff y_i f(\boldsymbol{x}_i) > 0$$

Once we find f, we can use $\hat{y} = \mathrm{sign}[f(\boldsymbol{x})]$ to classify new images.

How to find f? Let's consider linear classifiers, i.e. $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$, and two methods to determine $\boldsymbol{w}, b$, i.e. logistic regression and support vector machine.

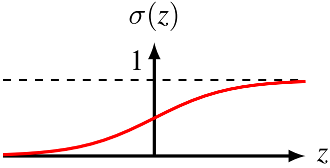
Logistic regression

Logistic regression assumes the probability for an image $\boldsymbol{x}$ to have label y is modeled by

$$p(y \mid x) = \sigma(yf(x)) = \sigma(y(\boldsymbol{w}^T\boldsymbol{x} + b))$$

where σ is the sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



To determine $\boldsymbol{w}, b$, maximize the likelihood

$$\max_{\boldsymbol{w},b} \quad L(\boldsymbol{w}, b) = \prod_{i=1}^{m} p(y_i \mid x_i)$$

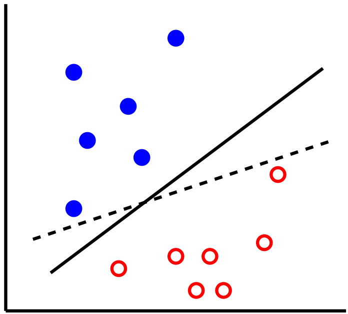
or equivalently, minimize the negative log likelihood

$$\min_{\boldsymbol{w},b} \quad NLL(\boldsymbol{w}, b) = -\log L(\boldsymbol{w}, b) = \sum_{i=1}^{m} \log(1 + e^{-y_i(\boldsymbol{x}_i^T\boldsymbol{w}+b)})$$

Support vector machine

Assume data is linearly separable, i.e. exists hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ s.t.

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 0, \quad \forall i$$



There may exist many such hyperplanes with different margins, i.e. minimum distance from data points to the hyperplane.

- A classifier with a larger margin is more robust against noise.

Support vector machine

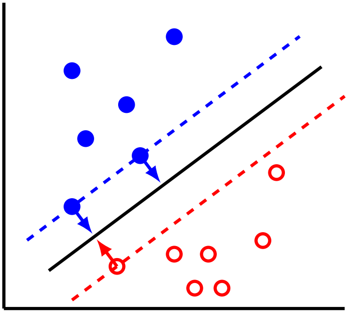
Given a hyperplane $P : \boldsymbol{w}^T\boldsymbol{x} + b = 0$,

- distance from $\boldsymbol{x}_i$ to P,

$$\operatorname{dist}(\boldsymbol{x}_i, P) = \frac{|\boldsymbol{w}^T\boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|}$$

- margin

$$\min_{1\le i\le m} \frac{|\boldsymbol{w}^T\boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|}$$



Support vector machine (SVM) maximizes the margin

$$\begin{aligned} &\max_{\boldsymbol{w},b} \quad \min_{1\le i\le m} \frac{|\boldsymbol{w}^T\boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|} \\ &\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 0, \quad i = 1, 2, \ldots, m \end{aligned}$$

Not easy to solve in this form.

Support vector machine

Problem reformulation

- Note $|\boldsymbol{w}^T\boldsymbol{x}_i + b| = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$, as $y_i = \operatorname{sgn}(\boldsymbol{w}^T\boldsymbol{x}_i + b)$.
- For $\alpha > 0$, $\tilde{\boldsymbol{w}} = \alpha\boldsymbol{w}$ and $\tilde{b} = \alpha b$ determine the same hyperplane P,

$$\boldsymbol{x} \in P \iff \boldsymbol{w}^T\boldsymbol{x} + b = 0 \iff \tilde{\boldsymbol{w}}^T\boldsymbol{x} + \tilde{b} = 0$$

- Choosing α properly, we can assume $\min\limits_{1\le i\le m} y_i(\tilde{\boldsymbol{w}}^T\boldsymbol{x}_i + \tilde{b}) = 1$,

$$\begin{aligned} \max_{\tilde{\boldsymbol{w}},\tilde{b}} \quad & \frac{1}{\|\tilde{\boldsymbol{w}}\|} \\ \text{s. t.} \quad & y_i(\tilde{\boldsymbol{w}}^T\boldsymbol{x}_i + \tilde{b}) \ge 1, \quad i = 1, 2, \ldots, m \end{aligned}$$

- Maximizing $1/|z|$ is equivalent to minimizing $\frac{1}{2}z^2$,

$$\begin{aligned} \min_{\tilde{\boldsymbol{w}},\tilde{b}} \quad & \frac{1}{2}\|\tilde{\boldsymbol{w}}\|^2 \\ \text{s. t.} \quad & y_i(\tilde{\boldsymbol{w}}^T\boldsymbol{x}_i + \tilde{b}) \ge 1, \quad i = 1, 2, \ldots, m \end{aligned}$$

We will see this reformulation is a convex problem and easy to solve.

Appendix: Distance to hyperplane

- $\boldsymbol{w} \perp$ hyperplane $P : \boldsymbol{w}^T\boldsymbol{x} + b = 0$
- $\boldsymbol{x}_i'$ is the orthogonal projection of $\boldsymbol{x}_i$ onto P, i.e.

$$\boldsymbol{x}_i - \boldsymbol{x}_i' \perp P$$

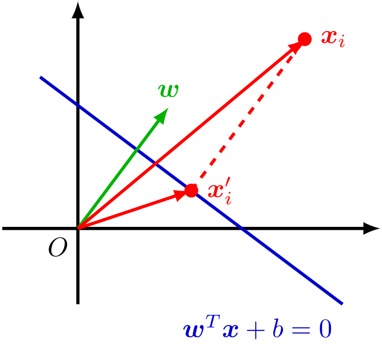
$$\boldsymbol{w}^T\boldsymbol{x}_i' + b = 0$$

- $\boldsymbol{x}_i - \boldsymbol{x}_i' = \gamma_i \boldsymbol{w}$ for some $\gamma_i \in \mathbb{R}$,

$$\boldsymbol{w}^T(\boldsymbol{x}_i - \gamma_i \boldsymbol{w}) + b = 0 \implies \gamma_i = \frac{\boldsymbol{w}^T\boldsymbol{x}_i + b}{\boldsymbol{w}^T\boldsymbol{w}}$$

- distance from $\boldsymbol{x}_i$ to P is

$$\min_{\boldsymbol{x}\in P} \|\boldsymbol{x}_i - \boldsymbol{x}\| = \|\boldsymbol{x}_i - \boldsymbol{x}_i'\| = \|\gamma_i \boldsymbol{w}\| = \frac{|\boldsymbol{w}^T\boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|}$$



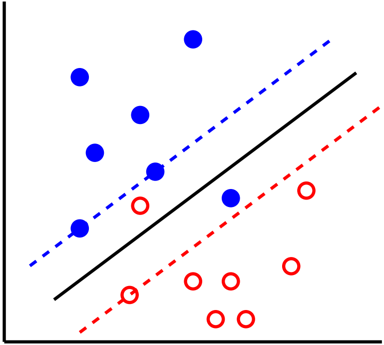
Soft margin SVM

Hard margin SVM requires linear separability

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \quad \forall i$$

When not linearly separable,

- relax constraints
- penalize deviation



Soft margin SVM: introduce slack variables $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \quad (C > 0 \text{ is a hyperparameter})$$
$$\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, m$$
$$\boldsymbol{\xi} \geq \boldsymbol{0}, \quad (\text{i.e.} \quad \xi_i \geq 0, \quad i = 1, 2, \ldots, m)$$

Outline

- Examples of optimization problems

- Global and local optima

Global optima

$\boldsymbol{x}^* \in X$ is a global minimum[1] of f over X if

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in X$$

It is an optimal solution of the minimization problem

$$\min_{\boldsymbol{x} \in X} f(\boldsymbol{x}) \tag{P}$$

and $f(\boldsymbol{x}^*)$ is the optimal value of (P).

Global maximum is defined by reversing the direction of the inequality.

Maximum and minimum are called extremum.

Note. Global extrema may not exist.

- $f(x) = x$, $X = \mathbb{R}$, $\inf_{x \in X} f(x) = -\infty$ unbounded below
- $f(x) = x$, $X = (0, 1)$, $\inf_{x \in X} f(x) = 0$, but not achievable

[1] Global minimum often also refers to the minimum value $f(\boldsymbol{x}^*)$.

Review: Inner product and norm

Euclidean inner product on $\mathbb{R}^n$: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^n x_i y_i$

Euclidean norm (2-norm, ℓ_2-norm): $\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^T \boldsymbol{x}} = \sqrt{\sum_{i=1}^n x_i^2}$

A norm on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ satisfying

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

Example.

- p-norm (ℓ_p-norm): $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, $p \geq 1$
- ∞-norm (ℓ_∞-norm): $\|\boldsymbol{x}\|_\infty = \max\limits_{1 \leq i \leq n} |x_i| = \lim\limits_{p \to \infty} \|\boldsymbol{x}\|_p$

Property 4 of these norms is given by Minkowski's inequality.

By default, $\|\boldsymbol{x}\|$ means $\|\boldsymbol{x}\|_2$.

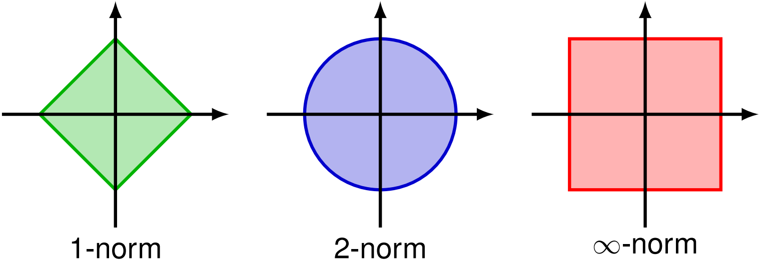
Review: Open and closed balls

Open ball of radius r centered at $\boldsymbol{x}_0$

$$B(\boldsymbol{x}_0, r) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}_0\| < r\}$$

Closed ball of radius r centered at $\boldsymbol{x}_0$

$$\bar{B}(\boldsymbol{x}_0, r) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq r\}$$



unit balls in $\mathbb{R}^2$ with different norms

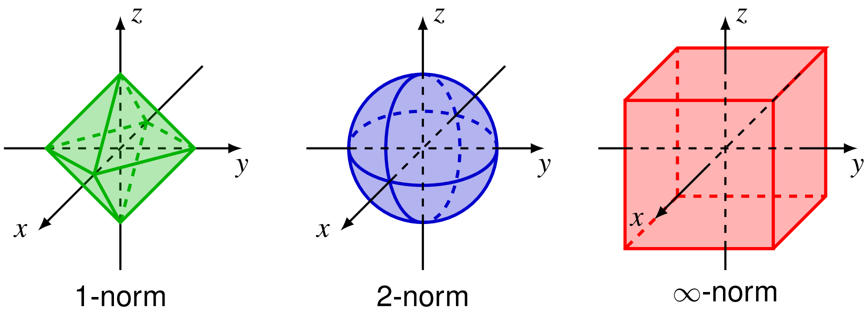
Review: Open and closed balls

Open ball of radius r centered at $\boldsymbol{x}_0$

$$B(\boldsymbol{x}_0, r) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}_0\| < r\}$$

Closed ball of radius r centered at $\boldsymbol{x}_0$

$$\bar{B}(\boldsymbol{x}_0, r) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}_0\| \le r\}$$



unit balls in $\mathbb{R}^3$ with different norms

Review: Open and closed sets

A set S is open if for any $\boldsymbol{x} \in S$, there exists $\epsilon > 0$ s.t. $B(\boldsymbol{x}, \epsilon) \subset S$.

A set S is closed if its complement S^c is open.

Examples in $\mathbb{R}$.

- $(0, 1)$ is open.
- $[0, 1]$ is closed.
- $(0, 1]$ is neither open nor closed.
- $[1, \infty)$ is closed.

A sequence $\{\boldsymbol{x}_n\}$ converges to $\boldsymbol{x}$, denoted $\boldsymbol{x}_n \to \boldsymbol{x}$ or $\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x}$ if

$$\lim_{n\to\infty} \|\boldsymbol{x} - \boldsymbol{x}_n\| = 0$$

Note. In $\mathbb{R}^n$, if $\boldsymbol{x}_n \to \boldsymbol{x}$ in one norm, it converges in any norm. The convergence in norm is equivalent to coordinate-wise convergence.

Theorem. S is closed iff for any sequence $\{\boldsymbol{x}_n\} \subset S$,

$$\boldsymbol{x}_n \to \boldsymbol{x} \implies \boldsymbol{x} \in S.$$

Review: Compactness

A set S is bounded if there exists $M < \infty$ s.t. $\|\boldsymbol{x}\| \leq M, \forall \boldsymbol{x} \in S$.

A set $S \subset \mathbb{R}^n$ is compact if it is closed and bounded.

Examples in $\mathbb{R}$.

- $[0, 1]$ is compact
- $(0, 1)$, $(0, 1]$ are bounded but not closed, so not compact
- $[1, \infty)$ is closed but not bounded, so not compact

Examples in $\mathbb{R}^n$.

- For $0 < r < \infty$, the open ball $B(0, r) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| < r\}$ is not compact.
- For $r < \infty$, the closed ball $\bar{B}(0, r) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq r\}$ is compact.
- $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \geq \boldsymbol{0}\}$ is closed but unbounded, so not compact.

Review: Continuity

A function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ is continuous at $\boldsymbol{x} \in X$ if for any $\epsilon > 0$, there exists $\delta > 0$ s.t.

$$\boldsymbol{y} \in X \cap B(\boldsymbol{x}, \delta) \implies |f(\boldsymbol{y}) - f(\boldsymbol{x})| < \epsilon$$

Equivalently, f is continuous at $\boldsymbol{x} \in X$ if

$$\forall \{\boldsymbol{x}_n\} \subset X, \quad \boldsymbol{x}_n \to \boldsymbol{x} \implies f(\boldsymbol{x}_n) \to f(\boldsymbol{x})$$

f is continuous on X if it is continuous at every $\boldsymbol{x} \in X$.

Theorem. If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous on $\mathbb{R}^n$, then for any $c \in \mathbb{R}$,

1. $\{\boldsymbol{x} : f(\boldsymbol{x}) < c\}$ is open;
2. $\{\boldsymbol{x} : f(\boldsymbol{x}) \leq c\}$ is closed.

Existence of Global Optima

Extreme Value Theorem. If f is continuous on a compact set X, then f attains its maximum and minimum on X, i.e. there exist $\boldsymbol{x}_1, \boldsymbol{x}_2 \in X$ (not necessarily unique) s.t.

$$f(\boldsymbol{x}_1) \leq f(\boldsymbol{x}) \leq f(\boldsymbol{x}_2), \quad \forall \boldsymbol{x} \in X.$$

Example. $f(x) = x^2$ satisfies $f(0) \leq f(x) \leq f(2) = f(-2)$ on $[-2, 2]$.

The Extreme Value Theorem gives sufficient conditions for the existence of global optima, but they are not necessary.

Example. $f(x) = x^2$.

- $\inf_{x \in (0,1)} f(x) = 0$, but $f(x) > 0$ for all $x \in (0, 1)$, no global min.
- $\min_{x \in [0,1)} f(x) = f(0)$, $x^* = 0$ is global min, but $[0, 1)$ is not closed.
- $\min_{x \in \mathbb{R}} f(x) = f(0)$, $x^* = 0$ is global min, but $\mathbb{R}$ unbounded.

Existence of Global Optima (cont'd)

Corollary. If f is continuous on $\mathbb{R}^n$ and $f(\boldsymbol{x}) \to +\infty$ as $\|\boldsymbol{x}\| \to \infty$, then the global min exists, i.e. there exists $\boldsymbol{x}^*$ s.t. $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$, $\forall \boldsymbol{x}$.

Proof.

- Since $f(\boldsymbol{x}) \to +\infty$ as $\|\boldsymbol{x}\| \to \infty$, there exists $M > 0$ s.t. $f(\boldsymbol{x}) > f(\mathbf{0})$ when $\|\boldsymbol{x}\| > M$.
- The closed ball $\bar{B}(\mathbf{0}, M) = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq M\}$ is compact.
- By the Extreme Value Theorem, there exists $\boldsymbol{x}^* \in \bar{B}(\mathbf{0}, M)$ s.t.

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}), \quad \forall x \in \bar{B}(\mathbf{0}, M)$$

- For $\boldsymbol{x} \notin \bar{B}(\mathbf{0}, M)$, $f(\boldsymbol{x}^*) \leq f(\mathbf{0}) < f(\boldsymbol{x})$, so $\boldsymbol{x}^*$ is a global min on $\mathbb{R}^n$.

A function f is called coercive if $f(\boldsymbol{x}) \to +\infty$ as $\|\boldsymbol{x}\| \to \infty$.

Example. $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$ coercive, $\boldsymbol{x}^* = \mathbf{0}$ is global minimum.

Example. $f(\boldsymbol{x}) = e^{-\|\boldsymbol{x}\|}$ not coercive, no global minimum.

Example. $f(x) = \sin x$ not coercive, $x^* = -\frac{\pi}{2}$ is global minimum.

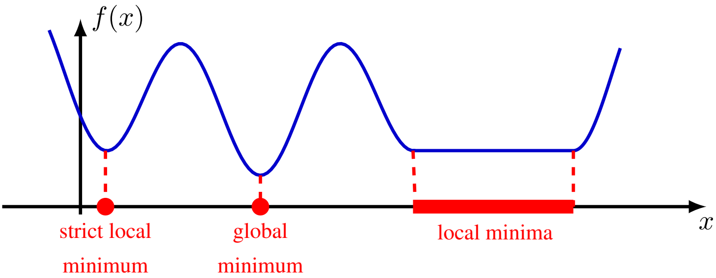
Local Minimum

$\boldsymbol{x}^* \in X$ is a local minimum of f on X if there exists $\epsilon > 0$ s.t.

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in X \cap B(\boldsymbol{x}^*, \epsilon)$$

x^* is a strict local minimum if strict inequality holds for $x \neq x^*$.

Local maximum is defined by reversing direction of inequality.



Global minimum is always local minimum, but not vice versa.

- We will see local min is global min for convex problems

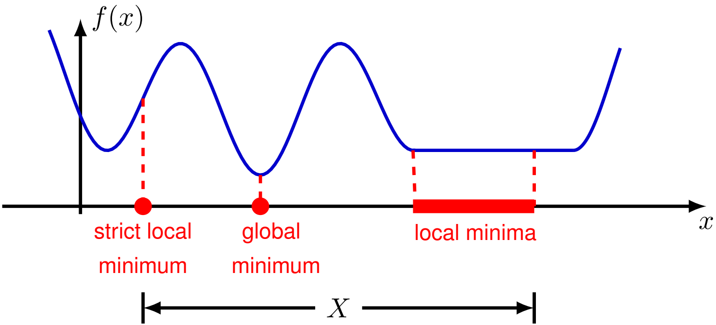
Local Minimum

$x^* \in X$ is a local minimum of f on X if there exists $\epsilon > 0$ s.t.

$$f(x^*) \leq f(x), \quad \forall x \in X \cap B(x^*, \epsilon)$$

x^* is a strict local minimum if strict inequality holds for $x \neq x^*$.

Local maximum is defined by reversing direction of inequality.



Global minimum is always local minimum, but not vice versa.

- We will see local min is global min for convex problems