CS257 Linear and Convex Optimization

Lecture 1

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

September 7, 2020

Contents

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
- $\boldsymbol{x} = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$: 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,

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

General optimization problems are very difficult; we will focus on convex optimization problems (to be defined later).

Example: Data Fitting

Recall Hooke's law in physics,

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

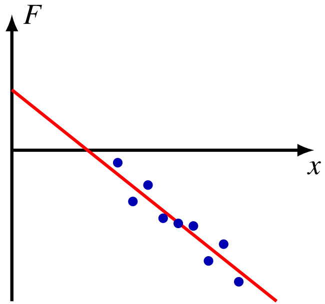
- F : force
- k : spring constant
- x : length
- x_0 : length at rest

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,

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

Example: Linear Least Squares Regression

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

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

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,

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

In a more compact form,

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

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

Example: Shipping 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

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

Example: 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 function $f : \mathbb{R}^n \to \mathbb{R}$, called 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. $f(\boldsymbol{x}) = \boldsymbol{w}^T x + b$

Example: Binary Classification (cont'd)

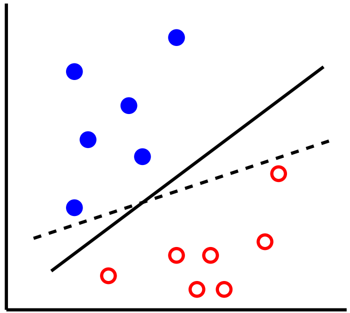
Assume data is linearly separable, i.e. exists hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ s.t.

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

May exist many such hyperplanes.

Want to maximize the minimum distance to the hyperplane

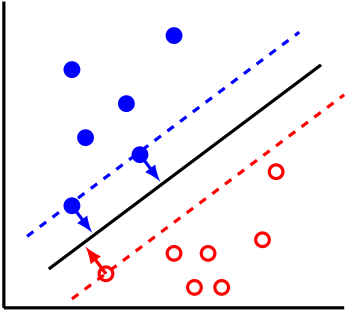
- more robust against noise



Example: Binary Classification (cont'd)

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

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



May exist many such hyperplanes.

Want to maximize the minimum distance to the hyperplane

- more robust against noise

Support vector machine: linear classifier with maximum margin

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

Can be reformulated as equivalent convex optimization problem yielding the same optimal hyperplane.

Example: Binary Classification (cont'd)

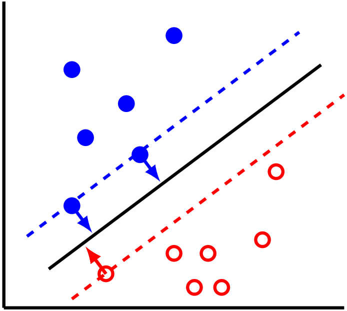
Assume data is linearly separable, i.e. exists hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ s.t.

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

May exist many such hyperplanes.

Want to maximize the minimum distance to the hyperplane

- more robust against noise



Support vector machine: linear classifier with maximum margin

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

We will see this is a convex optimization problem.

SVM

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 $\frac{1}{2}z^2$,

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

Appendix: Distance to Hyperplane

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

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

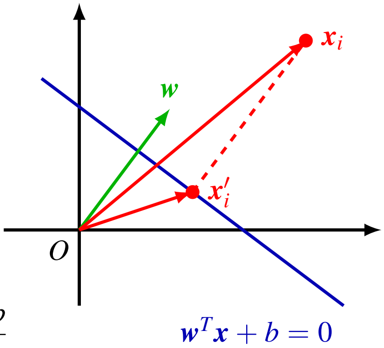
$$\boldsymbol{w}^T\boldsymbol{x}_i' + b = 0$$

- $\boldsymbol{x}_i - \boldsymbol{x}_i' = \gamma_i \boldsymbol{w}$ 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

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



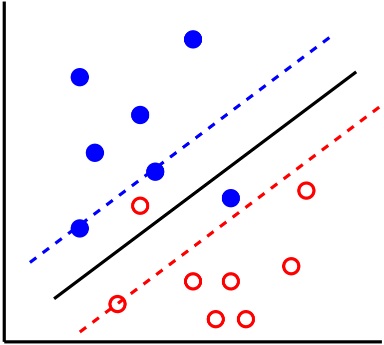
Soft Margin SVM

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 linear separable,

- relax constraints
- penalize deviation



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

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

Contents

Global Optima

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

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

It is also called 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 direction of 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 from 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 $f(\boldsymbol{x}^*)$.

Math Review

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): $\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^T \boldsymbol{x}} = \sqrt{\sum_{i=1}^n x_i^2}$

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 $\boldsymbol{x} = \mathbf{0}$
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.

- 1-norm: $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$
- p-norm: $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, $p \geq 1$
- ∞-norm: $\|\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$

Property 4 is given by Minkowski's inequality.

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

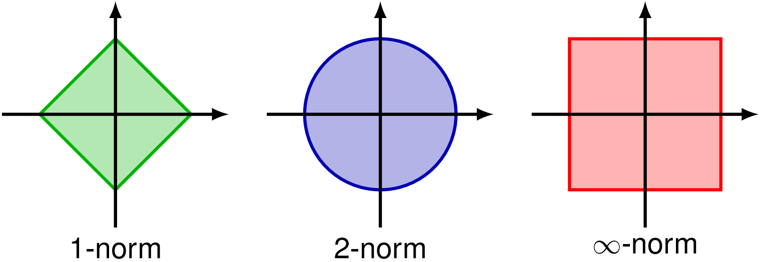
Math Review

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

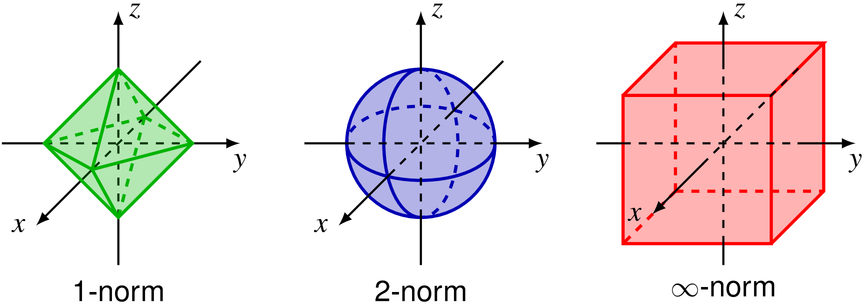
Math Review

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 $\mathbb{R}^3$ with different norms

Math Review

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.

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

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

Math Review

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
- $(0, 1)$, $(0, 1]$ and $[1, \infty)$ are not compact

A function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ is continuous at $\boldsymbol{x}$ 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 $\boldsymbol{x} \in X$.

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 $f(0) \leq f(x) \leq f(2)$ on $[-1, 2]$.

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 $[0, 1)$ 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 $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ exists, i.e. there exists $\boldsymbol{x}^*$ s.t. $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})$, $\forall \boldsymbol{x}$.

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 $\bar{B}(\mathbf{0}, M)$ is compact
- By the Extreme Value Theorem, there exists $\boldsymbol{x}^* \in X$ s.t.

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

- For $\boldsymbol{x} \notin \bar{B}(\mathbf{0}, M)$, $f(\boldsymbol{x}^*) \le f(\mathbf{0}) < f(\boldsymbol{x})$.

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.

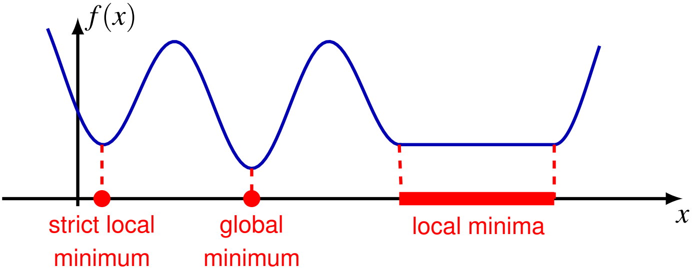
Local Minimum

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

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

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

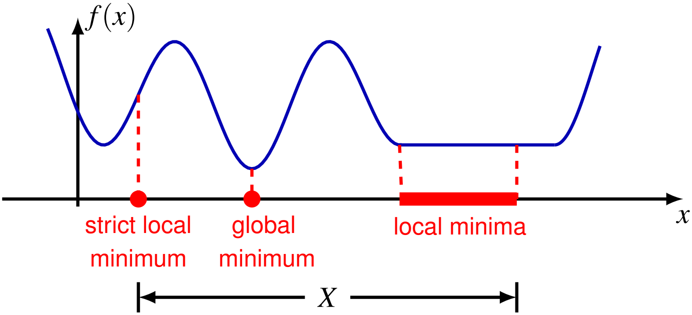
Local Minimum

$\boldsymbol{x}^* \in X$ is a local minimum of f 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