Lecture 1. Introduction

Course information

Optimization. 优化 决策 / 选择 / etc ...

Example: from Hongqiao to campus.

Taxi: fast. expensive. waiting in a long queue.

Minhong line 2. cheaper slower. stop far away.

Hongqiao bus line 4. cheapest. much slower. ...

Description via mathematics:

maximize / minimize $f(x)$. objective function.

subject to $x \in \Omega$ feasible set

$x^* = \underset{x \in \Omega}{\operatorname{argmin}} f(x)$ optimal solution

What does "linear and convex" mean?

In general optimization problems are very difficult to solve.

$x \in \mathbb{R}^n$. $f(x): \mathbb{R}^n \to \mathbb{R}$ Ω specified by constraints

$g_1 \cdots g_m$. $g_i(x) = 0$. $g_j(x) \le 0$. $g_k(x) \ge 0$.

Knapsack problem 背包问题. 如果你每天只能去一次教超...

n types of goods. 可乐体积小但重. 薯片轻但占地方.

i-th type has size a_i and weight b_i

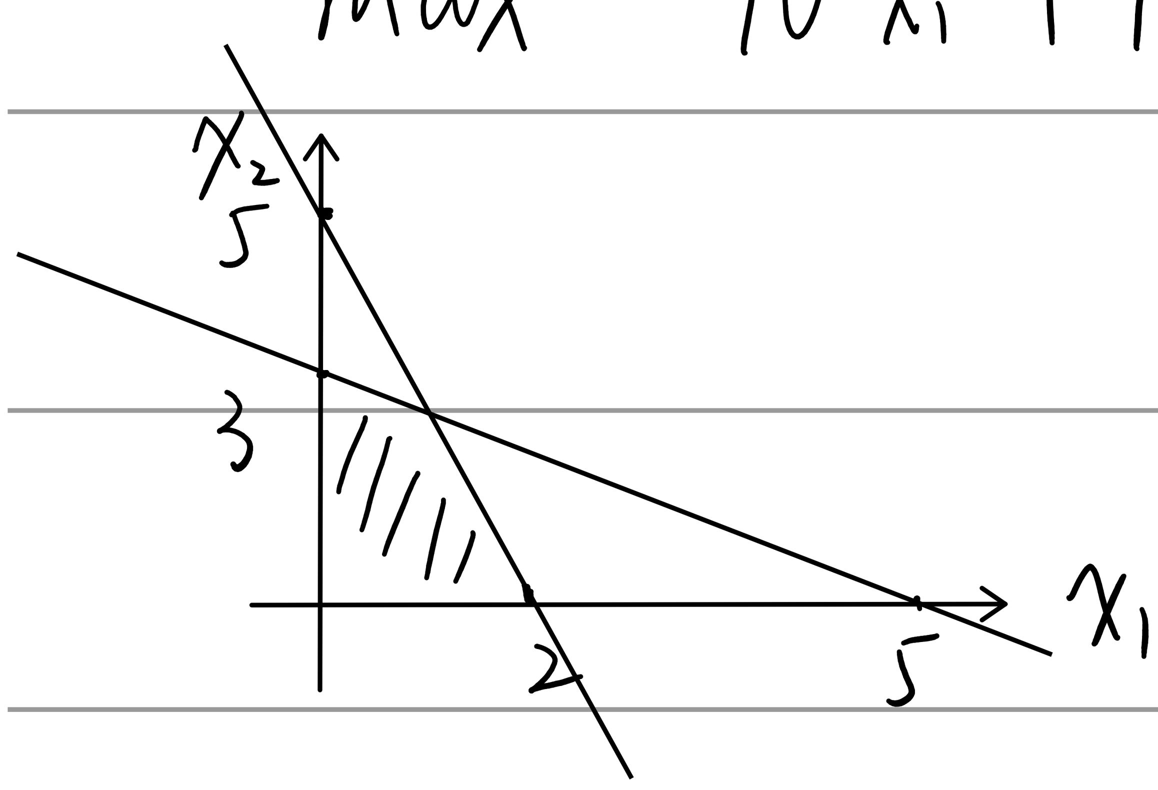
You have a bag of volume capacity A and load capacity B.

goal: maximize your happiness.

$$\max \sum_i c_i x_i \quad \text{s.t.} \quad \begin{array}{l} \sum_i a_i x_i \le A \\ \sum_i b_i x_i \le B. \end{array} \quad x_i \ge 0.$$

Assume only 2 types (可乐 & 薯片).

$$\max \ 10x_1 + 15x_2. \quad \text{s.t.} \quad \begin{array}{l} 3x_1 + 5x_2 \le 15 \\ 5x_1 + 2x_2 \le 10 \end{array}$$



more types?

simplex algorithm / primal-dual

The goal is specified. We now consider another example.

Data fitting. ideal gas law $pV = nRT$

T	30	40	50	60	...	$V = kT + b$
V	1.011	1.019	1.032	1.041	...	find the possible coefficients.

Now we have a data set $\{(x_1, y_1), \ldots (x_m, y_m)\}$. $x_i \in \mathbb{R}^n$.

Assume $y = w^T x + b$. inconsistent, what is the best approximation?

Question: how to measure the distance between y and estimated $\hat{y}$?

If $y, \hat{y} \in \mathbb{R}$. absolute value of $y - \hat{y}$ is natural. but if $y \in \mathbb{R}^n$?

Norm 范数. a generalization of absolute values.

Definition: Given a vector space V over a field F, a norm $\|\cdot\|$ is a function $\|\cdot\|: V \to \mathbb{R}$ satisfying following properties:

1. (nonnegativity) $\forall v \in V$. $\|v\| \geq 0$
2. (positive definiteness) $\|v\| = 0$ iff $v = 0$.
3. (absolute homogeneity) $\forall r \in \mathbb{R}, v \in V$. $\|r \cdot v\| = |r| \cdot \|v\|$.
4. (triangle inequality) $\forall u, v \in V$, $\|u + v\| \leq \|u\| + \|v\|$.

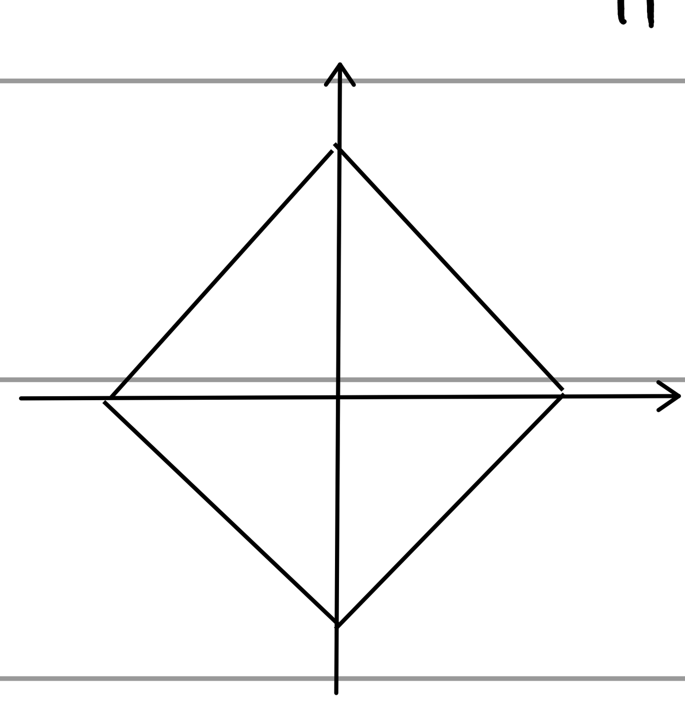
Example. l_p-norm. p-norm, or L^p-norm for $\mathbb{R}^n$.

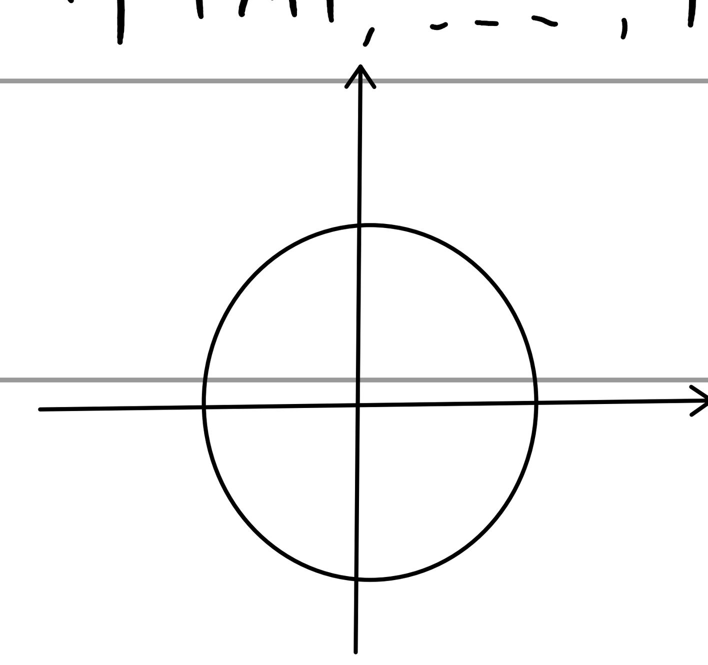
$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$. where $p \geq 1$.

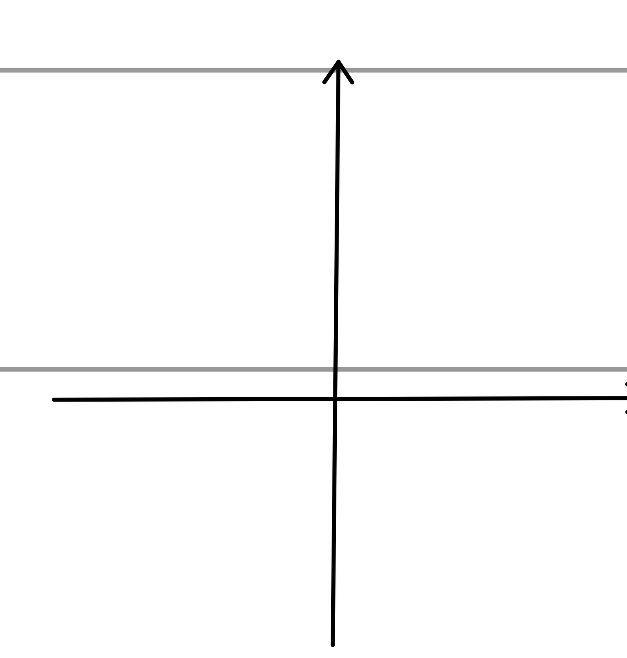
In particular. $\|x\|_1 = |x_1| + \cdots + |x_n|$ (Manhattan distance).

$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$ (Euclidean norm).

$\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$.







Sometimes we use l_0-norm $\|x\|_0 = |x_1|^0 + \cdots + |x_n|^0$. (norm?).

Example. norm induced by inner product. carnonical norm

Definition: an inner product for a vector space V over a field F.

is a function $\langle \cdot, \cdot \rangle : V \times V \to F$ satisfying (assume $F = \mathbb{R}$).

1. (nonnegativity) $\langle x, x \rangle \geq 0$.

2. (positive definiteness) $\langle x, x \rangle = 0$ iff $x = 0$

3. (symmetry) $\langle x, y \rangle = \langle y, x \rangle$. (conjugate if $F = \mathbb{C}$).

4. (linearity) $\langle rx, y \rangle = r \langle x, y \rangle$ (homogeneity).

$\langle x+y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ (additivity).

In particular. x, y are orthogonal if $\langle x, y \rangle = 0$.

Canonical norm: induced by inner product $\|x\| = \sqrt{\langle x, x \rangle}$.

Cauchy-Schwarz inequality: $|\langle x, y \rangle| \leq \|x\| \, \|y\|$

Euclidean inner product space: $\langle x, y \rangle = x^T y = \sum x_i y_i$

Back to data fitting: least squares solution. using l_2-norm.

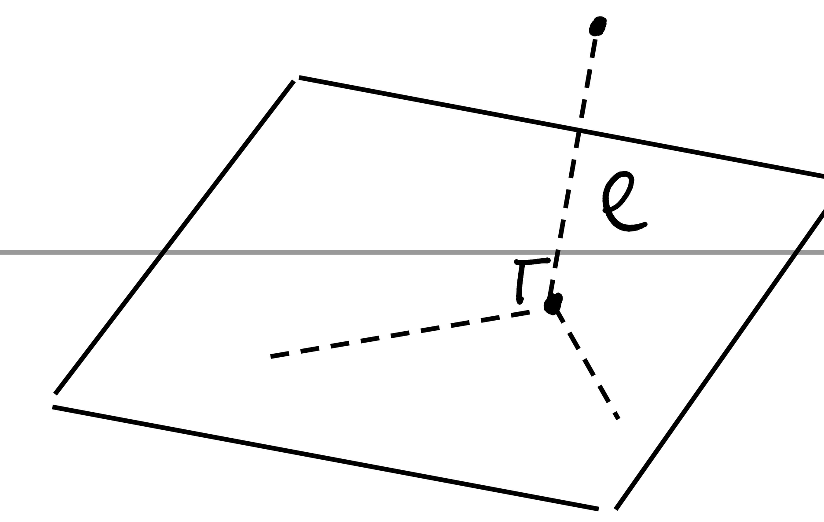
$y = w^T x + b$. let $X = (x_1, \dots x_m)^T \in \mathbb{R}^{m \times n}$, $Y = (y_1, \dots y_m) \in \mathbb{R}^m$.

$\min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \sum (w^T x_i + b - y_i)^2 = \min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \| Xw + b \cdot \mathbf{1} - Y \|_2^2$.

b is not important. let $x_i' = (x_{i1}, \dots x_{in}, 1)$. $w' = (w_1, \dots w_n, b)$

$\min_{w \in \mathbb{R}^n} \| Xw - y \|_2^2$. Xw: column space of X. $R(X)$ or $\mathrm{im}(X)$.

(orthogonal) projection of y onto $R(X)$.



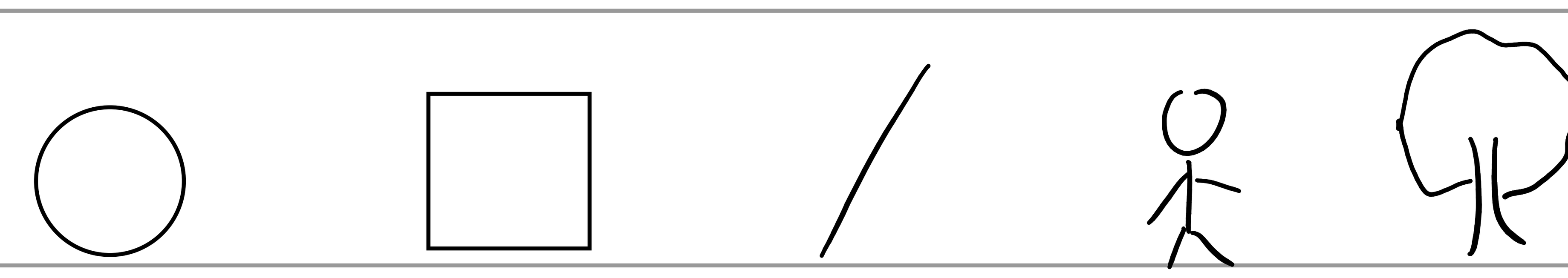
assume $e \perp R(X)$. $\Rightarrow X^T e = 0$.

$\exists \hat{w}$. s.t. $y - e = X\hat{w}$. $\Rightarrow X^T(y - X\hat{w}) = 0$.

$\Rightarrow X^T y = X^T X \hat{w} \quad \Rightarrow \hat{w} = (X^T X)^{-1} X^T y$ if $\mathrm{rank}(X) = n$. $\square$.

Classification and support vector machine.

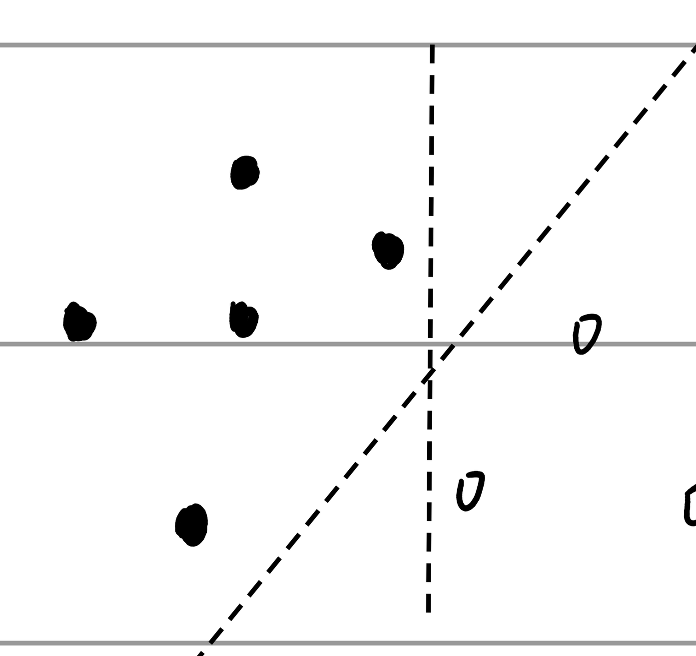
classify



Given a data set $\{x_1 \dots x_m\}$, $x_i \in \mathbb{R}^n$. classify by $y_i \in \{\pm 1\}$.

Goal: separate $\{x_1 \dots x_m\}$ by a hyperplane $w^T x + b = 0$.

maximize the distance to the hyperplane.



constraints: $y_i(w^T x_i + b) > 0$.

Distance to the hyperplane P: $w^T x + b = 0$. $\quad w \perp P$.

orthogonal projection $x \to x'$ on the hyperplane P.

$(x - x') = r \cdot w$ for some $r \in \mathbb{R}$. and $w^T x' + b = 0$.

$w^T(x - rw) + b = 0 \quad \Rightarrow \quad r = \dfrac{w^T x + b}{w^T w} \leftarrow \|w\|^2$

distance from x to P is $\min_{y \in P} \|x - y\| = \|x - x'\| = \dfrac{|w^T x + b|}{\|w\|}$.

Back to classification: linear classifier with maximum margin:

$$\max_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \min_i \frac{|w^T x_i + b|}{\|w\|} \quad \text{subject to} \quad y_i(w^T x_i + b) > 0.$$

So $|w^T x_i + b| = y_i(w^T x_i + b)$. choose proper w and b

such that $\min y_i(w^T x_i + b) = 1$. $\min_i \frac{|w^T x_i + b|}{\|w\|} = \frac{1}{\|w\|}$.

$$\min_{w \in \mathbb{R}^n\, b \in \mathbb{R}} \|w\|^2. \quad \text{subject to} \quad y_i(w^T x_i + b) \geq 1. \qquad \square.$$