Mathematical Foundations for Computer Science

Probability and Optimization

Chapter 1: Bayesian Probability

Spring 2024

Instructor: Xiaodong Gu

Probability Review

Basic:

Sample Space、Random Experiments、 Probability、Random Variables、Expectation、Variance;

Multiple Variables:

Joint Distribution、Marginal Distribution、Conditional Distribution、Independent

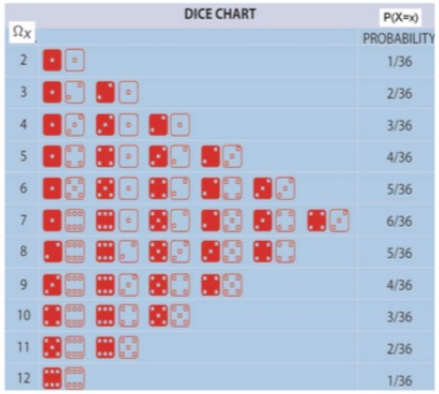
Random Variables

- A **random variable** (informally) is a variable whose value is not initially known

$$\text{Weather} \in \{\text{sunny}, \text{rainy}, \text{cloudy}, \text{snowy}\}$$
$$p(\text{Weather} = \text{sunny}) = 0.3$$
$$p(\text{Weather} = \text{rainy}) = 0.2$$
$$\ldots$$

the variable can take on different values (and it must take on exactly one of these values), each with an associated probability.

- A **random variable** is a function that maps from the sample space to a measurable space (e.g. a real number).



Example:

X = sum of the two results.

$X((2,5))= 7$; $X((3,1))=4$)

Probability Density Function (Distribution)

- A function that describes the probabilities of different outcomes for an experiment

$$P(X): \Omega_X \rightarrow [0, 1]$$

$$P(X=x) = P(\Omega_X=x)$$

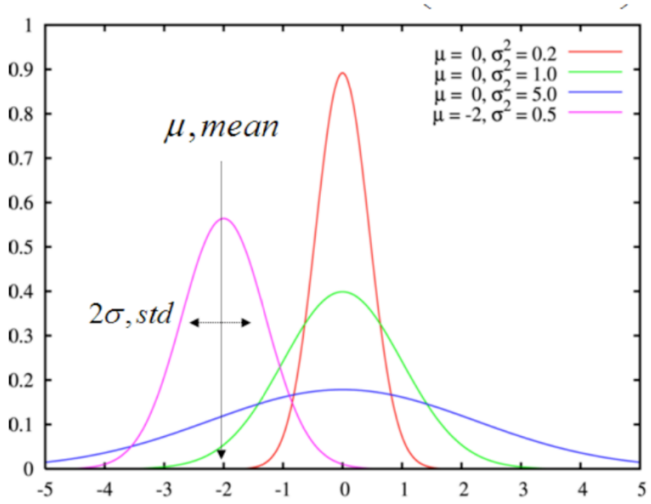
- $P(X = 4) = P(\{(1,3), (2,2), (3,1)\}) = 3/36$.
- If X is continuous, we have a **density function** p(X).

Probability Distributions

Gaussian (Normal) Distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Here $\mu = E[X]$ is the mean (and mode), and $\sigma^2 = var[X]$ is the variance



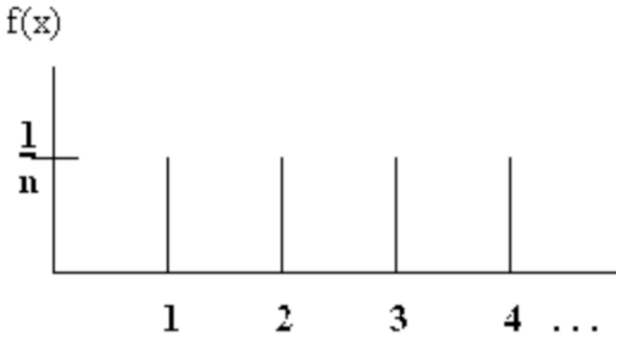
Probability Distributions

Uniform Distribution

Example: outcome of throwing a fair dice

$$P(X=1)=P(X=2)=\ldots=1/6$$

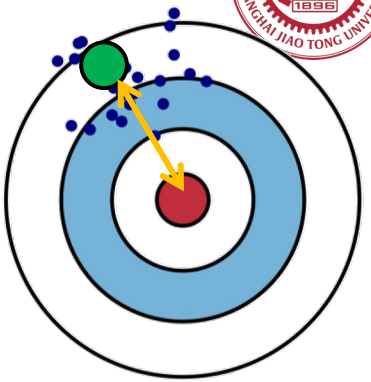
$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x \leq b \\ 0 & \text{otherwise} \end{cases}$$



Discrete

Continuous

Expectation



- The expected value of a random variable X

$$\mu = \boldsymbol{E}[X] = \sum_{x_i \in \chi} x_i P(X = x_i)$$

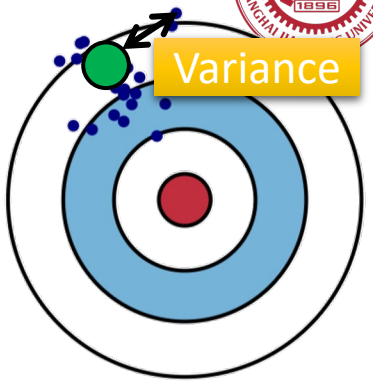
- average value of $X=x_i$, taking into account probability of the various x_i.
- also known as "mean"
- the most common measure of "center" of a distribution

- Estimate mean from actual samples: $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$

- We can also talk about the expected value of functions of x

$$\mathbf{E}[f(x)] = \mathbf{E}_{x \sim \boldsymbol{\chi}} f(x) = \sum_{x_i \in \chi} f(x_i) P(x = x_i)$$

Variance



- The expected deviation from the expectation

$$\text{let } f(x_i) = (x_i - E[X])$$

$$\Longrightarrow \sigma^2 = E[f^2] = \sum_i p(x_i) \cdot (x_i - \boldsymbol{E}[X])^2$$

• average value of squared deviation of from mean , taking into account probability of the various x_i
• the most common measure of "**spread**" of a distribution
• σ is the standard deviation

- Estimate variance from actual samples:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{n} (x_i - \mu)^2$$

https://www.zhihu.com/question/20099757

On Multiple Variables: Joint Probability

- Probability distribution of a random variable X:

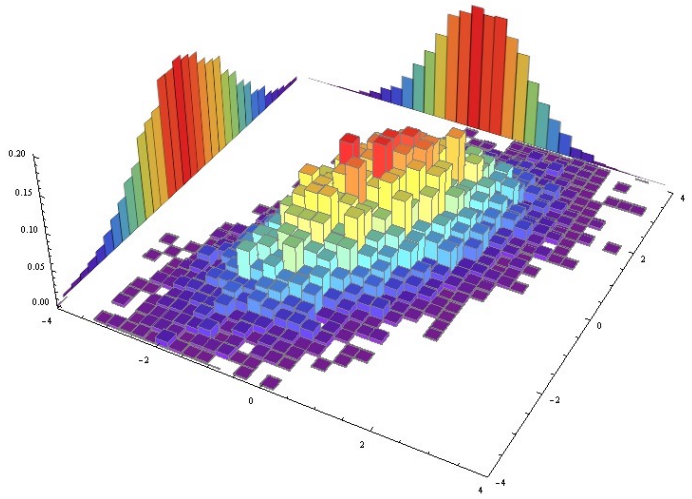
$$P(X): \Omega_X \rightarrow [0, 1]$$

$$P(X=x) = P(\Omega_X=x)$$

- If X and Y are two discrete random variables, we define the **joint probability function** of X and Y by

$$P(X=x, Y=y) = p(x, y)$$

where $p(x, y) \geq 0$ and $\Sigma_x \Sigma_y p(x,y) = 1$



Joint Probability Distribution

- Suppose there are n random variables $X_1, X_2, \ldots, X_n$.
- A **joint probability** $P(X_1, X_2, \ldots, X_n)$, over those random variables is a function defined on the Cartesian product (笛卡尔积) of their state spaces:

$$\prod_{i=1}^{n} \Omega_{X_i} \longrightarrow [0,1]$$

$$P(X_1=x_1, X_2=x_2, \ldots, X_n=x_n) = P(\Omega_{X_1=x_1} \cap \Omega_{X_2=x_2} \cap \cdots \cap \Omega_{n=x_n}).$$

X_1	X_2	…	X_n	Prob
0	0	…	0	0.30
0	1	…	0	0.10
1	0	…	0	0.05
1	1	…	0	0.25

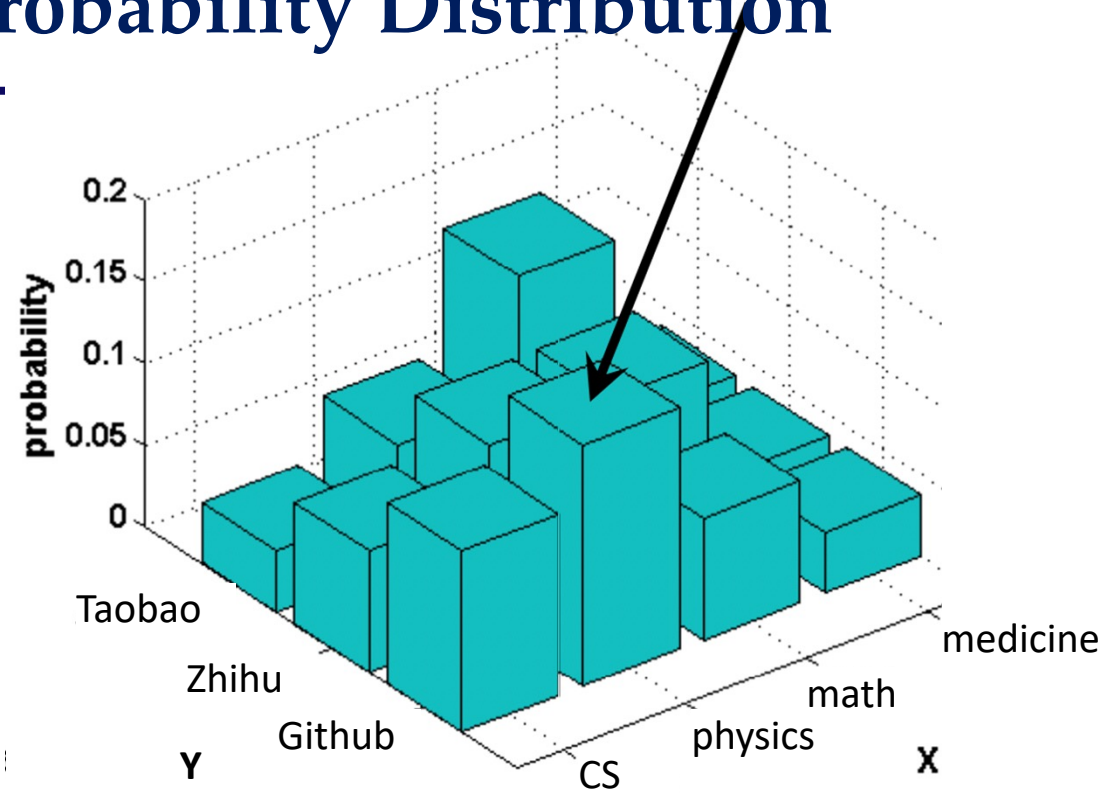
Joint Probability Distribution

Example:

- Population: popular websites browsed by SJTU students.
- Random variables:
  - Major: {CS, Physics, Math, Medicine}
  - Browsing: {GitHub, Zhihu, Taobao}
- Joint probability distribution: P(Major, Browsing):

	GitHub	Zhihu	Taobao
CS	.44	.13	.01
Physics	.17	.01	.02
Math	.09	.07	.01
Medicine	0	0.24	0.1

Joint Probability Distribution



- Joint distribution $P(X_1, X_2, \ldots, X_n)$ contains information about all aspects of the relationships among n random variables.
- In theory, one can answer any query about relationships among the variables based on the joint probability.

Marginal Probability Distribution (边缘分布)

What is the probability of a randomly selected student likes browsing GitHub?

	GitHub	Zhihu	Taobao	P(Major)
CS	.44	.03	.01	.48
Physics	.17	.01	.02	.2
Math	.09	.07	.01	.17
Medicine	0	0.14	0.1	.15
P(Browsing)	.7	.25	.05	

P(Browsing = GitHub) = P(Browsing=GitHub, Major = CS) + P(Browsing=GitHub, Major=Physics) + P(Browsing=GitHub, Major=Math) + P(Browsing=GitHub, Major=Medicine) = .7

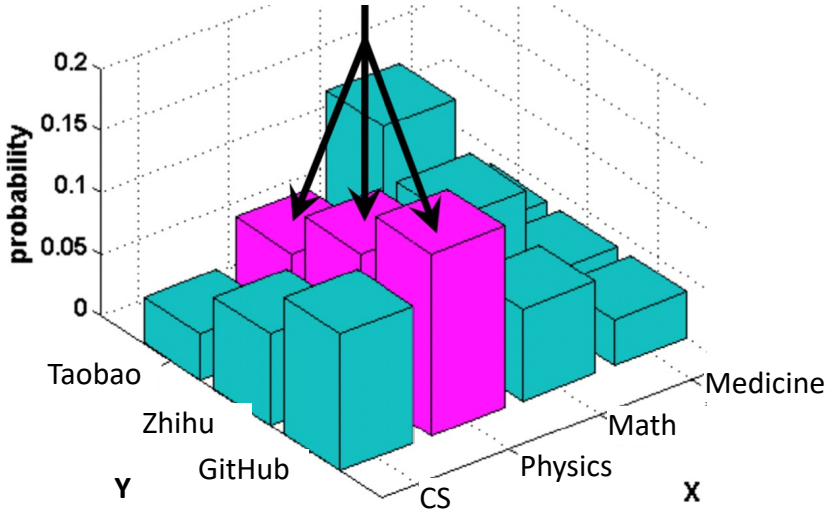
- Called marginal probability because written on the margins.

Marginal Probability Distribution

- Write the equations on the previous slide in a compact form:

$$P(\text{Browsing}) = \sum_{\text{Major}} P(\text{Browsing}, \text{Major})$$

- The operation is called marginalization (边缘化): Variable "*Major*" is marginalized from the joint probability P(Browsing, Major).



Marginal Probability Distribution

Notations for more general cases:

$$P(X,Y) = \sum_{U,V} P(X,Y,U,V)$$

$$\boldsymbol{Y} \subset \{X_1, X_2, \ldots, X_n\},\ \boldsymbol{Z} = \{X_1, X_2, \ldots, X_n\} - \boldsymbol{Y},$$

$$P(\boldsymbol{Y}) = \sum_{\boldsymbol{Z}} P(X_1, X_2, \ldots, X_n)$$

- A **joint probability** gives us a **full picture** about how random variables are related.
- **Marginalization** let us **focus on one aspect** of the picture

Conditional Probability

- For events A and B:

$$P(A|B) = \frac{P(A,B)}{P(B)} (= \frac{P(A \cap B)}{P(B)})$$

- Meaning:
  - $P(A)$: my probability on A (without any knowledge about B)
  - $P(A|B)$: My probability on event A assuming that I know event B is true.

Q: What is the probability of a randomly selected CS student likes browsing GitHub?

P(Browsing=GitHub| Major=CS)
= P(Browsing=GitHub, Major=CS) / P(Major=CS) = .44/.48 = .96

By contrast:

P(Browsing=GitHub) = 0.7

Independent

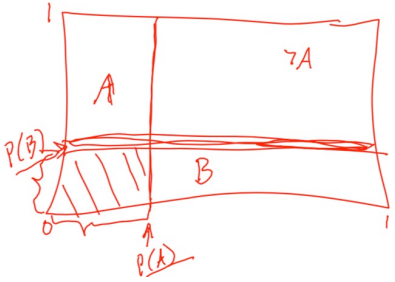
- Two random variables A and B are independent if

$$P(A \cap B) = P(A)P(B), \text{ or } P(B \mid A) = P(B), \text{ or } P(A \mid B) = P(A)$$

Example

A and B are two coin tosses

- The probability of B occurring is **not** affected by the occurrence or non-occurrence of A
- Knowledge about A contains **no** information about B



- If n Boolean variables (A_1, …, A_n) are independent

$$P(A_1 \cap \cdots, \cap A_n) = \prod_{i=1}^{n} P(A_i)$$

Chain Rule of Probability

- For two variables:

  $P(X, Y) = P(X)P(Y|X)$

- For three variables:

  $P(X, Y, Z) = P(X)P(Y|X)P(Z|X,Y)$

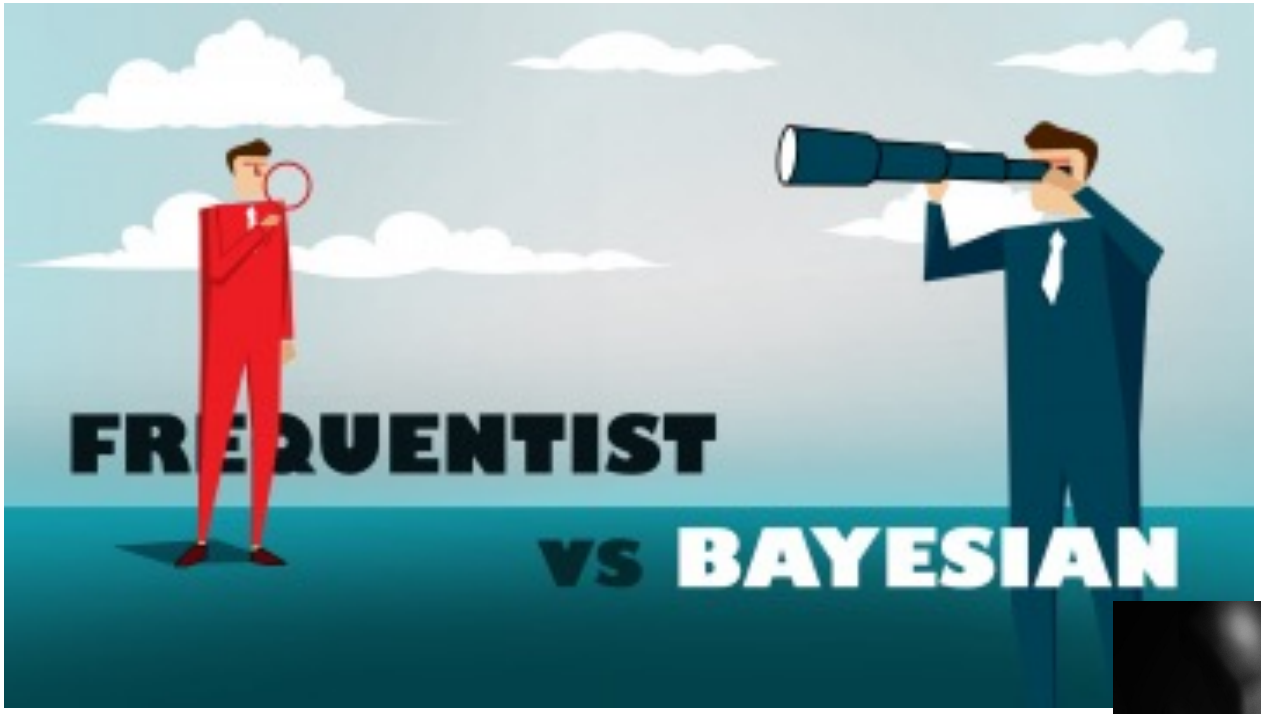
- For ***n*** variables:

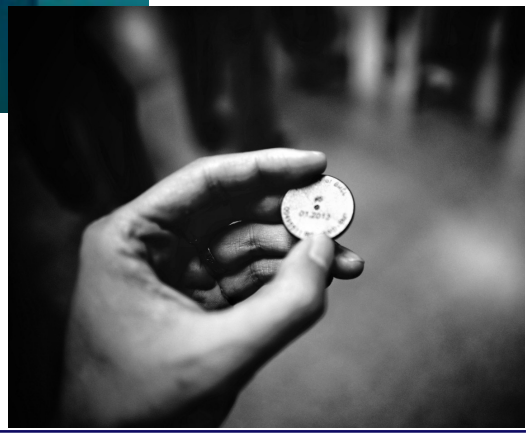
  $P(X_1,\ldots,X_n) = P(X_1)P(X_2|X_1)\ldots P(X_n|X_1,\ldots,X_{n-1})$

Proof:

$$P(X, Y, Z) = \frac{P(X, Y, Z)}{P(X,Y)} \frac{P(X, Y)}{P(X)} P(X)$$

Frequentist vs Bayesian





Frequentist Interpretation

probabilities are long term relative frequencies.

Example: X is the result of coin tossing. $\Omega_X = \{H, T\}$

- $P(X=H) = 1/2$ means that
  - The relative frequency of getting heads will almost surely approach ½ as the number of tosses goes to infinite.
- Proved by the Law of large numbers:
  - X_i: result of the i-th tossing: 1 – H, 0 – T
  - Law of large numbers: $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{2}$ with probability 1

Frequentist interpretation is meaningful only when experiment can be **repeated** under the same condition.

Subjectivist (Bayesian) Interpretation

Probabilities are logically consistent degrees of beliefs.

—applicable when experiments are not repeatable.

—depends on a person's state of knowledge.

Example: "the probability that COVID-19 comes from bats".

- doesn't make sense under frequentist interpretation.
- subjectivist: degree of belief based on state of knowledge
  - primary school student: 0.01
  - me: 0.1
  - biologist: 1 or 0

How to relate such subjective and changing probabilities ?

Prior, Posterior, and Likelihood

- Three important concepts in Bayesian inference, with respect to a hypothesis H and evidence E:

Prior probability P(H): belief about a hypothesis before observing evidence.

Posterior probability P(H|E): belief about a hypothesis after obtaining the evidence.

Example: suppose 0.001% of people suffer from flu. A doctor's Prior probability about a new patient suffering from flu is 0.00001.

If the doctor finds that the patient has fever, his belief about patient suffering from flu would be > 0.00001.

Prior, Posterior, and Likelihood

Suppose a patient is observed got fever (E).

Consider two possible hypothesis:

1. The patient is infected by flu (H_1).
2. The patient is not infected by flu (H_2)

Which is better?

And how to compare?

Obviously, H_1 is a better explanation because $P(E|H_1) > P(E|H_2)$. So $P(E|H_1)$ is a measure of how well H explains E. We call it likelihood.

Likelihood L(H|E):

The **likelihood** of a hypothesis H given evidence E is a measure of how well H explains E. Mathematically,

$$L(H|E) = P(E|H)$$

Bayesian Rule

Bayes' Theorem: relates prior probability, likelihood, and posterior probability:

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} \propto P(H)L(H|E)$$

a normalization constant to ensure $\Sigma_{h\in\Omega_H} P(H=h|E) = 1$

posterior ∝ prior × likelihood

- Usually written in a general form:

Bayes' Rule:

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]}$$

Bayesian Rule

Example: medical diagnosis

given:

- P(Fever |flu) = 0.8
- P(flu) = 0.005
- P(Fever) = 0.05

Question

Find P(flu | Fever)

$$P(\text{flu} \mid \text{Fever})$$

$$= \frac{P(\text{Fever} \mid \text{flu})P(\text{flu})}{P(\text{Fever})}$$

$$= \frac{0.8 \times 0.005}{0.05} = 0.08$$

Mathematical Foundations for Computer Science

Probability and Optimization

Lecture 2: Introduction to Information Theory

Spring 2024

Instructor: Xiaodong Gu

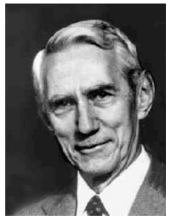
Information and Uncertainty

- The **amount of information** upon observing an event corresponds to how much **surprise** (novelty, uncertain, etc) it brings to us.

  - **Rare event** ⇒ high information (*surprising*).
  - **Common event** ⇒ low information (*unsurprising*).

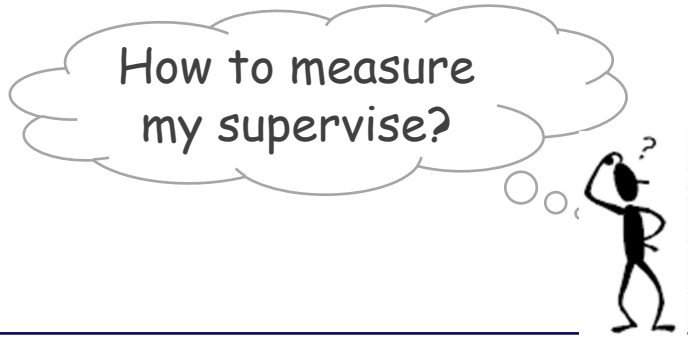
In other words ···

Information is obtained by gaining knowledge about something that was previously uncertain.



Shannon

"Information is the resolution of uncertainty"



Shannon Information Content

- The **amount of information** we receive upon observing an event corresponds to the **probability** of the event.

Information gain upon observing a event with a probability of p

$$= \log_2 \frac{1}{p} \text{ bits.}$$

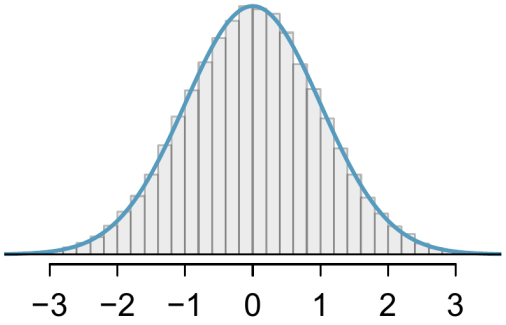
Definition

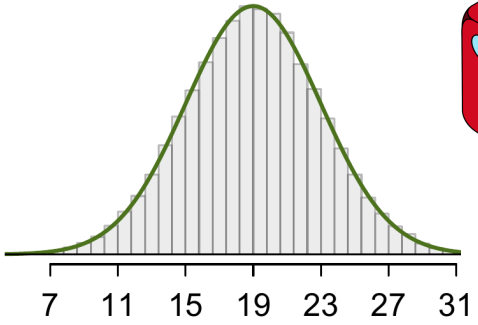
The **Shannon Information Content** (SIC) of an event with probability p is $\log_2 1/p$ bits.

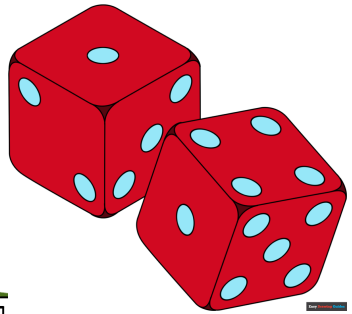
Example 1: coin tossing
- x= [heads; tails], p=[½ ; ½], SIC=[1; 1] bits

Example 2: is it my birthday?
- x= [no; yes], p=[364/365; 1/365], SIC=[0.004; 8.512] bits

What about the information associate with an **probability distribution?**







Entropy (熵)

- The **entropy** is defined as the expected (average) information over the distribution of a random variable.
- Let $\boldsymbol{X}$ be a random variable taking on a finite number of different values $\{x_1, \cdots, x_M\}$ with probabilities $(p_1, \cdots, p_M)$
  (e.g., English letters in a file, results of coin tossing, password, etc)

$$H(X) = \mathbb{E}\left\{\log \frac{1}{p(X)}\right\} = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \text{ bits}$$

- Entropy measures the amount of uncertainty associated with a particular probability distribution.
- The higher the entropy, the less confident we are in the outcome.

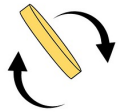
Examples

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

We flip two different coins

Sequence 1:

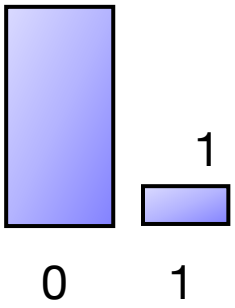
0 0 0 1 0 0 0 0 ... ?



Sequence 2:

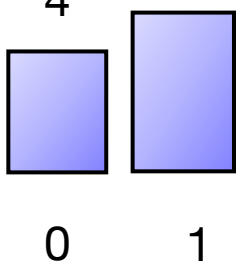
0 1 0 1 1 0 1 0 1... ?

A biased coin flip has 0.5 bit of entropy



$$-\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9} \approx \frac{1}{2}$$

vs



A fair coin flip has 1 bit of entropy

$$-\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} \approx 0.99$$

- Non-uniform probability yields less uncertainty and therefore less entropy
- Uniform probability yields maximum uncertainty and therefore maximum entropy

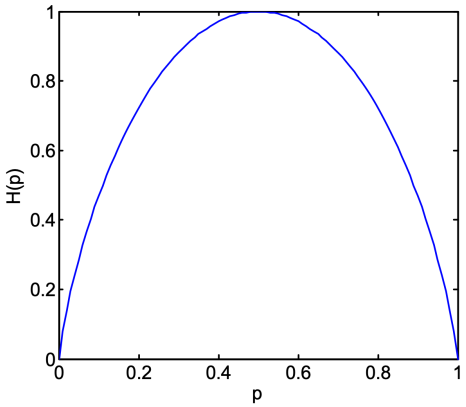
Examples

(1) Bernoulli random variables (e.g., flip a coin)

$x = [0; 1]$, $p(x) = [1\text{-}p; p]$

$H(p) = -(1-p)\log(1-p) - p \log p$

* we often write H(p) to mean H([1-p; p]).



probability p of heads

(2) Four-colored shapes

$x = [\bullet; \blacksquare; \text{✦}; \text{✾}]$, $p_x = [½; ¼; ⅛; ⅛]$

$H(x) = H(p_x) = \Sigma\text{-}\log(p(x))p(x)$

$= 1\times1/2 + 2\times1/4 + 3\times1/8 + 3\times1/8 = 1.75$ bits

Entropy vs Information

$$H(X) = \mathbb{E}\left\{\log \frac{1}{p(X)}\right\} = \mathbb{E}\,(\mathrm{SIC})$$

Entropy is the information that we do not know.

vs

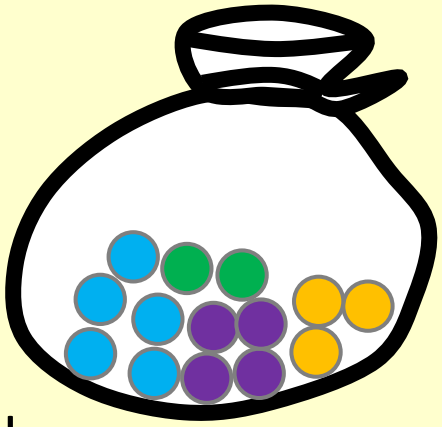
Information is the reduction of entropy.

Some alternative definitions …

- H (X) = the average Shannon information content of X
- H (X) = the average information gained by knowing its value
- H (X) = the average number of bits required to represent or transmit an event drawn from the probability distribution for X
- H (X) = the average number of "yes-no" questions need to find x is in the range [H(x), H(x)+1)
- **H (X) = the average cost of resolving the uncertainty for X**
- …

Game Time

袋子中有四种颜色(橙/紫/青/绿)的球。从中随机抽一只。请你问yes-no问题来确认球的颜色。
例如:“球是黄色吗?”或者
“球是绿色或者青色吗?”



Goal: as fewer questions as possible

Game Time

Information vs Entropy

Information (Certainty)

stage 1: no information

stage 2: having known that ½ orange, ¼ purple, ⅛ cyan, and ⅛ green

stage 3: known that all balls in the bag are orange

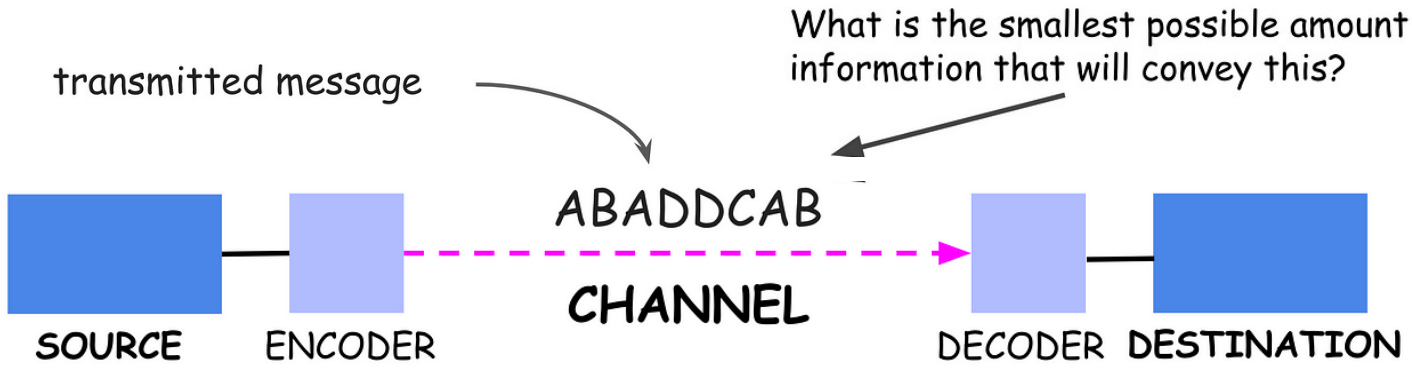
# of questions to ask = 2

# of questions to ask = ½ × 1 + ¼ × 2 + ⅛ ×3 + ⅛ ×3=1.75

# of questions to ask = 0

Entropy (Uncertainty)

Example: Message Compression



Case 1

Probabilities	Encoding
A 25%	'A' = '00'
B 25%	'B' = '01'
C 25%	'C' = '10'
D 25%	'D' = '11'

Can't do better than 2 bits encoding for each letter.

Case 2

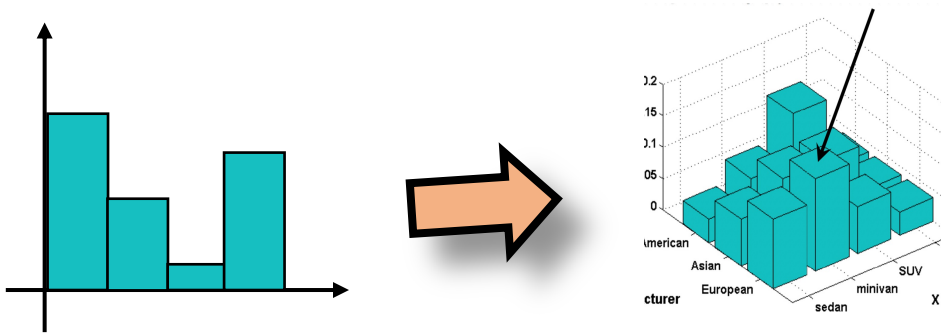
Probabilities	Encoding
A 70%	'A' = '0' (1 bit)
B 26%	'B' = '10' (2 bits)
C 2%	'C' = '110'(3 bits)
D 2%	'D' = '111'(3 bits)

70% of the time only **1 bit** needs to be sent, 26% of the time **2 bits**, and 4% **3 bits**.

How many bits are required *on average* in these two scenarios?

Joint Entropy (联合熵)

- Extend the notion to a **pair** of discrete random variables (X, Y)



$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

$$H(X,Y) = -\mathbb{E}\log p(X,Y)$$

- Nothing new: can be considered as a single vector-valued random variable
- Useful to measure dependence of two random variables

Joint Entropy

Example: X = {Raining, Not raining}, Y = {Cloudy, Not cloudy}

P(X,Y)	Cloudy	Not Cloudy
Raining	24/100	1/100
Not Raining	25/100	50/100

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x,y) \\
&= -\frac{24}{100}\log_2\frac{24}{100} - \frac{1}{100}\log_2\frac{1}{100} - \frac{25}{100}\log_2\frac{25}{100} - \frac{50}{100}\log_2\frac{50}{100} \\
&\approx 1.56\text{bits}
\end{aligned}
$$

Conditional Entropy (条件熵)

Definition

Conditional Entropy H(Y|X): entropy of a random variable Y given the knowledge of another random variable X.

- For $(X,Y) \sim p(x,y)$

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X=x) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(y|x)$$

Conditional entropy can be computed as the expectation of entropy over different conditions $E_{P(X=x)}[H(Y \mid X = x)]$

- $H(Y|X) \neq H(X|Y)$

Conditional Entropy – Example

Example: X = {Raining, Not raining}, Y = {Cloudy, Not cloudy}

P(X,Y)	Cloudy	Not Cloudy
⇨ Raining	24/100	1/100
Not Raining	25/100	50/100

What is the entropy of cloudiness Y , **given that it is raining**?

$$
\begin{aligned}
H(Y|X=x) &= -\sum_{y\in Y} p(y|x)\log_2 p(y|x) \\
&= -\frac{24}{25}\log_2\frac{24}{25} - \frac{1}{25}\log_2\frac{1}{25} \\
&\approx 0.24\text{bits}
\end{aligned}
$$

We used: $p(y|x) = p(x, y)/p(x)$, and $p(x) = \Sigma_y p(x, y)$ （sum in a row)

Exercise

Joint Entropy: **H(X, Y)**

$H(X, Y) = \mathbb{E}_{X,Y} -\log p(X, Y)$

$= -p(X=0,Y=0) \log p(X=0,Y=0) - \ldots$

$= -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - 0 \log 0 - \frac{1}{4} \log \frac{1}{4}$

$= 1.5$ bits

$P(X,Y)$	Y=0	Y=1
X=0	½	¼
X=1	0	¼

Note: 0 log 0 = 0

Conditional Entropy: **H(Y|X)**

$H(Y|X) = \mathbb{E}_{X,Y} -\log p(Y|X)$

$= -p(X=0, Y=0) \log p(Y=0|X=0) - \ldots$

$= -\frac{1}{2} \log \frac{2}{3} - \frac{1}{4} \log \frac{1}{3} - 0 \log 0 - \frac{1}{4} \log 1$

$= 0.689$ bits

< 1.5 bits

$P(Y|X)$	Y=0	Y=1
X=0	⅔	⅓
X=1	0	1

Note: rows sum to 1

Conditional Entropy – View 1

Average Row Entropy:

P(X,Y)	Y=0	Y=1	P(Y|X=x)	P(X)
X=0	½	¼	H(⅓)	¾
X=1	0	¼	H(1)	¼

$$
\begin{aligned}
H(Y|X) &= \mathbb{E}_{X,Y} -\log p(Y|X) \\
&= \Sigma_{X,Y} - p(X,Y) \log p(Y|X) \\
&= \Sigma_{X,Y} - \mathrm{p(X)p}(Y|X) \log \mathrm{p}(Y|X) \\
&= \Sigma_X \mathrm{p}(X) \Sigma_Y - \mathrm{p}(Y|X) \log p(Y|X) \\
&= \Sigma_x \mathrm{p(x)H}(Y|X = x) \\
&= ¾ \times \mathrm{H}(⅓) + ¼ \times \mathrm{H}(1) = 0.689 \text{ bits}
\end{aligned}
$$

The average entropy of all rows weighed by their probabilities p(x)

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) = \frac{1}{4} H(\text{cloudy}|\text{is raining}) + \frac{3}{4} H(\text{cloudy}|\text{not raining})$$

Conditional Entropy – View 2

Additional Entropy:

$P(X,Y)$	$Y=0$	$Y=1$	$P(X)$
$X=0$	½	¼	¾
$X=1$	0	¼	¼

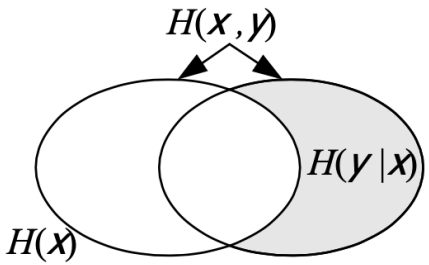
$$P(Y|X) = P(X,Y)/P(X)$$

$$\Downarrow$$

$$H(Y|X) = \mathbb{E}_{X,Y}\,\text{-log}\, p(Y|X) = \mathbb{E}\{\text{-log}\, p(X,Y)\} - \mathbb{E}\{\text{-log}\, P(X)\}$$

$$= H(X,Y) - H(X) = H(½, ¼, 0, ¼) - H(¼) = 0.689\text{bits}$$

H(Y|X) is the remaining uncertainty in Y given the knowledge of X



Chain Rules

- **Probability**

$$P(X, Y, Z) = P(Z|X,Y)\ P(Y|X)\ P(X)$$

- **Entropy**

$$H(X, Y, Z) = H(Z|X, Y) + H(Y|X) + H(X)$$

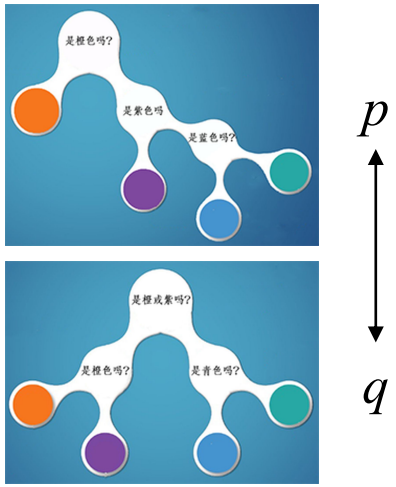
$$H(X_{1:n}) = \Sigma^{n}_{i=1} H(X_i|X_{1:i-1})$$

The log operator in the definition of entropy converts products of probability probability into sums of entropy.

Cross Entropy (交叉熵)

Recall the ball verification game:
what if the true distribution is p= (½, ¼, ⅛, ⅛) while our knowledge level remains at stage 1 (i.e., q=(¼, ¼, ¼, ¼) and we persist in strategy1)?

Now the # of questions to ask is ½ ×2+¼ × 2 + ⅛ ×2 + ⅛ ×2 = 2 > 1.75



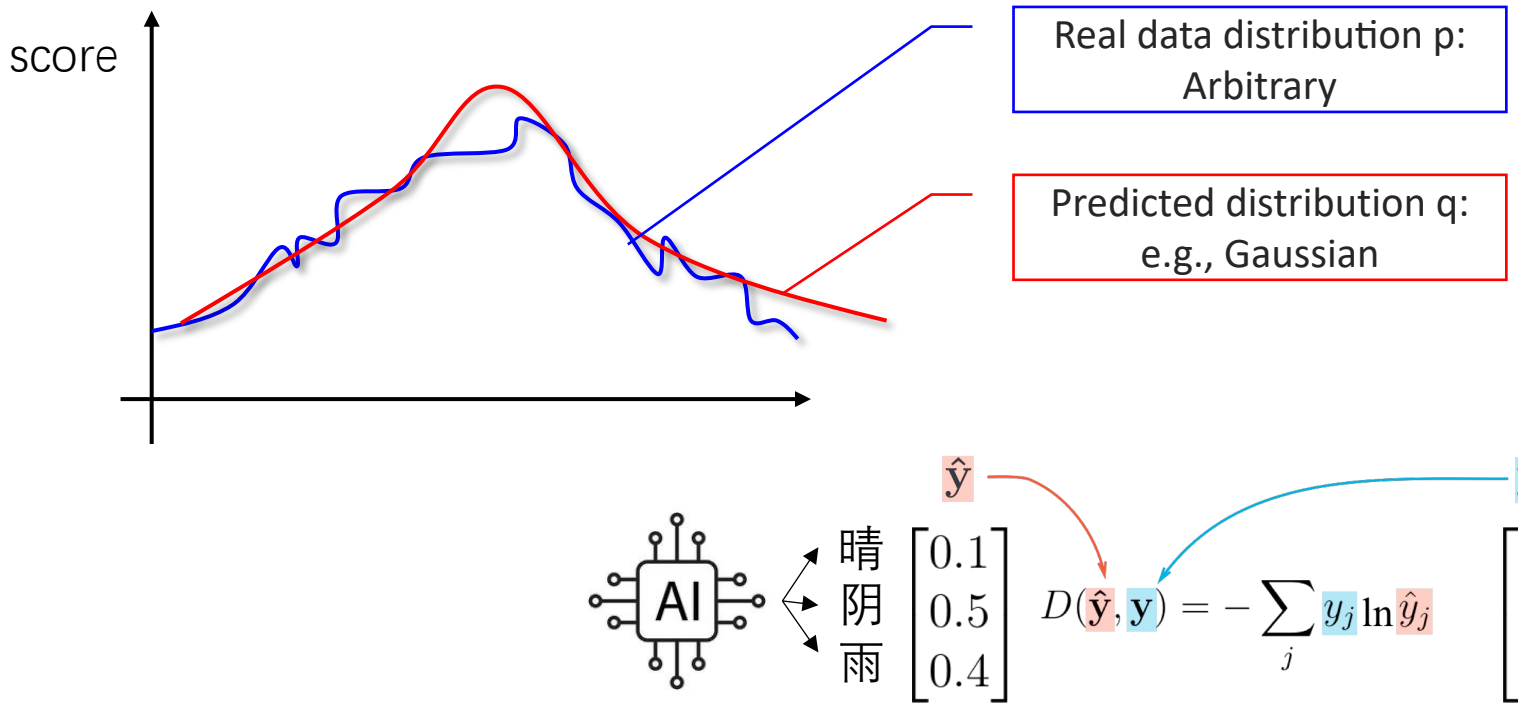
Definition

Cross Entropy: cost of resolving uncertainty under assumption of a distribution q while the real distribution is p.

$$CE(p, q) = \boldsymbol{E}_{x\sim p(x)}\left\{\log \frac{1}{q(x)}\right\} = -\sum_{x\in X} p(x)\log_2 q(x)$$

Cross Entropy

Cross entropy is widely used in **machine learning** to measure the quality of the estimated distribution $\hat{y}$ with respect to the real data distribution y.

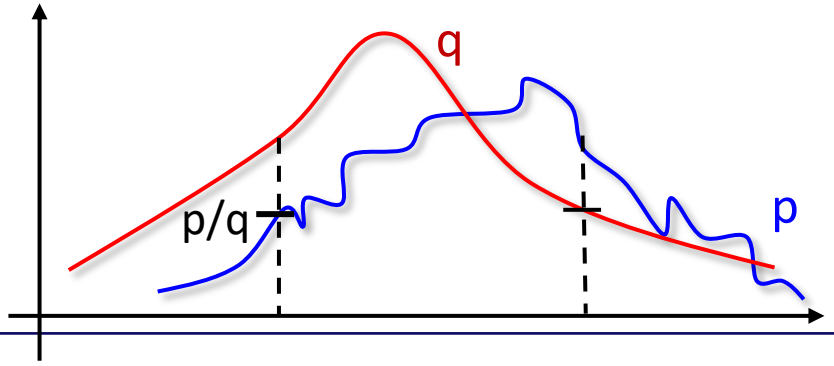


Relative Entropy (KL divergence)

- A measure of inefficiency under assumption of a distribution q while the true distribution is p.

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

- also known as Kullback-Leibler (KL) divergence in statistics: expected log-likelihood ratio.
- can be used to measure the "distance" between two distributions
- KL=0 if and only if p=q



Let $f(t)=-\ln t$. Then f is strictly convex. We have $0=\int f(q(x)/p(x))p(x)dx \geq f(\int q(x)/p(x)\ p(x)dx)=f(1)=0$. Since equality holds in Jensen's inequality and f is strictly convex it follows that $q(x)/p(x)$ is a constant, hence $=1$ almost everywhere. Hence $P=Q$.

Relative Entropy (KL divergence)

- **Cross Entropy vs Relative Entropy**:

$$D_{KL}(p||q) = CE(p, q) - H(p)$$

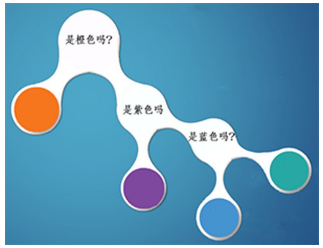
Suppose we construct code using q distribution, we need $D_{KL}(p||q)$ additional bits on average to describe the random variable.

Relative Entropy (KL divergence)

Example: the ball color verification game again

Q: what exactly is the difference of using two strategies (strategy 1, or strategy 2) at stage 2?

A: $D_{KL}(p||q) = CE(p, q) - H(p) = \mathbb{E}_p(-\log q_x) - H(p) = 2 - 1.75 = 0.25$ bits



p

q

Practice

$X = [1\ 2\ 3\ 4\ 5\ 6]^T$

$p = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6]$ $H(p) = 2.585$

$q = [1/10, 1/10, 1/10, 1/10, 1/10, ½]$ $H(q) = 2.161$

$D_{KL}(p\|q) = \mathbb{E}_p(-\log q_x) - H(p) = 2.935 - 2.585 = 0.35$

$D_{KL}(q\|p) = \mathbb{E}_q(-\log p_x) - H(q) = 2.585 - 2.161 = 0.424$

Mutual Information

$P(X,Y)$	$Y=0$	$Y=1$
$X=0$	½	¼
$X=1$	0	¼

- If you try to guess Y, you have a 50% chance of being correct.
- However, what if you know X?

Best guess: choose Y=X

— If X=0 (p=0.75) then 66% correct prob
— If X=1 (p=0.25) then 100% correct prob
— Overall 75% correct probability

Definition

Mutual Information: the amount of information that one random variable carries about another one.

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y))$$

Entropy and Mutual Information

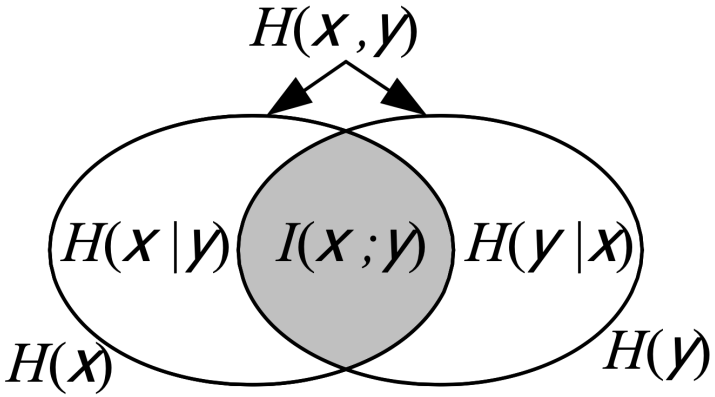
Uncertainty in x

Uncertainty in x having known y

$$I(X;Y) = H(X) - H(X|Y)$$

$$= H(X) + H(Y) - H(X,Y)$$

I (X; Y) is the intersection of information in X with information in Y

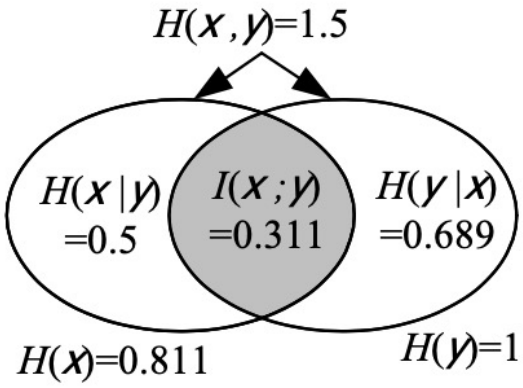


- I(X;Y) = I(Y;X)

Mutual Information

Example:

$P(X,Y)$	$Y=0$	$Y=1$
$X=0$	½	¼
$X=1$	0	¼



$$I(X; Y) = H(X) - H(X|Y)$$
$$= H(X) + H(Y) - H(X, Y)$$
$$H(X) = 0.811,\ H(Y) = 1,\ H(X,Y) = 1.5$$
$$I(X; Y) = 0.311$$

Mutual Information

Example: relationship between skin cancer and blood type

X: blood type

Y: chance for skin cancer

	A	B	AB	O
Very Low	1/8	1/16	1/32	1/32
Low	1/16	1/8	1/32	1/32
Medium	1/16	1/16	1/16	1/16
High	1/4	0	0	0

Y: marginal (¼, ¼, ¼, ¼)  X: marginal (½, ¼, ⅛, ⅛) 

$H(Y) = 2$ bits, $H(X) = 1.75$ bits

Conditional entropy: $H(X|Y) = 1.375$ bits, $H(Y|X) = 1.625$ bits

Mutual Information: $I(X; Y) = H(X) + H(Y) - H(X, Y) = 0.375$ bits

Mathematical Foundations for Computer Science

Probability and Optimization

Lecture 3: Linear Programming

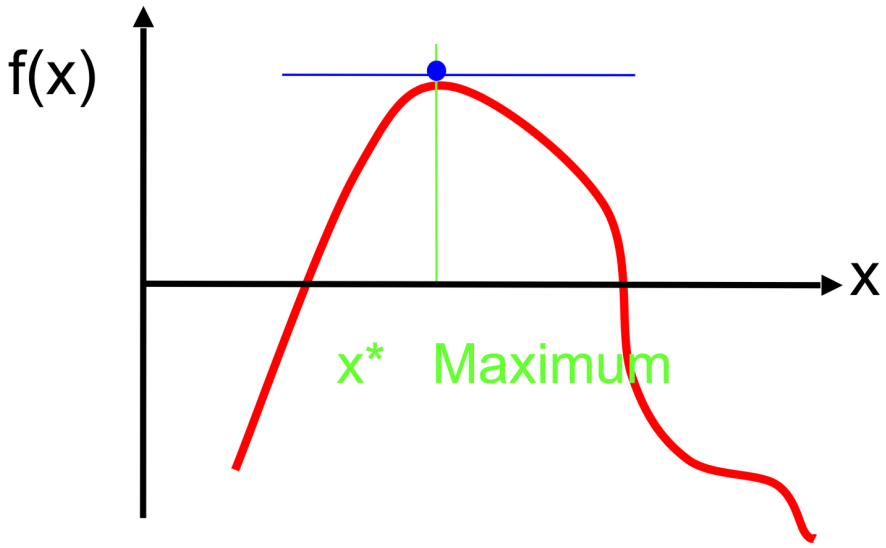
Spring 2024

Instructor: Xiaodong Gu

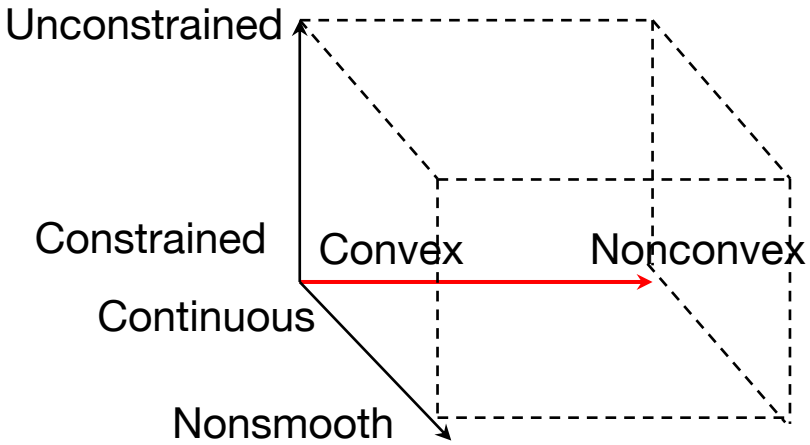
What is Optimization?

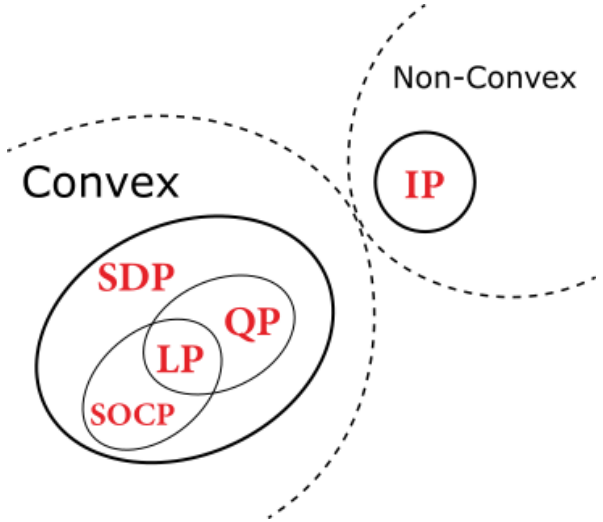
- **Optimization = search for the best solution !**
- In mathematical terms:

> **minimization** or **maximization** of an **objective function** $f(x)$ depending on variables x subject to constraints



Overview of Optimization Problems





Our focus in this course:

— **C**onstrained, **C**ontinuous, **C**onvex

Linear Algebra Warmup

Vector, Matrix, Positive (semi-) definite

Vector Arithmetic

- **Addition of two vectors**

$$\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y} = (x_1 + y_1, \ \ldots \ , x_n + y_n)^T$$

- **Scalar multiplication of a vector**

$$\boldsymbol{y} = a\boldsymbol{x} = (ax_1, \ldots, ax_n)^T$$

Inner Product

- The inner product (dot product) of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{R}^n$:

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x} = \Sigma_i \, x_i y_i$$

Example: $\boldsymbol{x} = (4, 6, 1)^T$, $\boldsymbol{y} = (1, 3, 1)^T$, $\boldsymbol{x}^T\boldsymbol{y} = 4\times1+6\times3+1\times1 = 23$

- If $\boldsymbol{x}^T\boldsymbol{y} = 0$, then $\boldsymbol{x}$ and $\boldsymbol{y}$ are **orthogonal**.

Vector Norms (范数)

For a vector $\boldsymbol{x} \in \mathbb{R}^n$ with elements $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$:

- The l_2 norm, or Euclidean norm:

$$\|x\|_2 = \|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

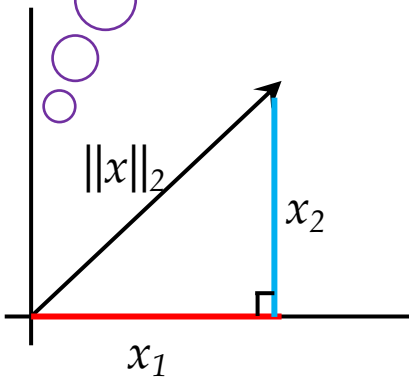
- The l_1 norm:

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

- The l_p-norm:

$$\|x\|_p = \sqrt[p]{|x|_1^p + |x|_2^p + \cdots + |x|_n^p}$$

Can generally think of as the vector "**length**"



https://zhuanlan.zhihu.com/p/26884695

Multiply a Vector by a Matrix

- Multiplying a matrix $\mathbf{A} = [a_{ij}]_{m \times n}$ with a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ can be written as a weighted sum of $\mathbf{A}$'s column vectors

$$\mathbf{Ax} = \mathbf{y}$$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

column vector

$$y_i = \sum_{j=1}^{n} a_{ij} x_i$$

Matrix Multiply Vector (Left)

$x^T A = y$

$$[x_1, x_2, \ldots, x_n] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = [y_1, y_2, \ldots, y_n]$$

row vector

Matrix Multiply Vector (Left & Right)

$\mathbf{x^T A x = z}$

$$[x_1, x_2, \ldots, x_n] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= [y_1, y_2, \ldots, y_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} y_i x_i = z$$

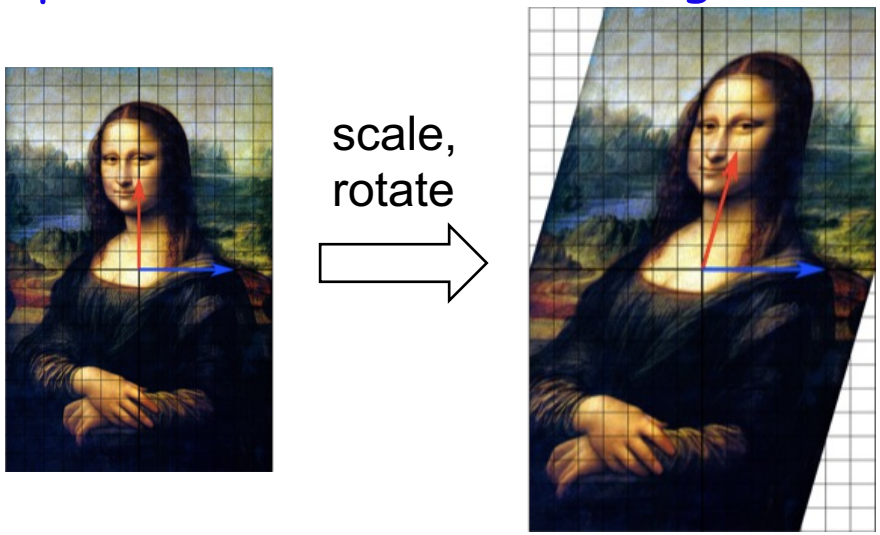
salar

Matrix Multiplication

If $\mathbf{C}_{m\times n} = \mathbf{A}_{m\times p}\mathbf{B}_{p\times n}$, then $[c_{ij}] = \sum_{k=1}^{p} a_{ik}b_{kj}$

- in general, non-commutative: $\mathbf{AB} \neq \mathbf{BA}$
- associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- distributive: $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$

Matrix multiplication scales/rotates a geometric plane

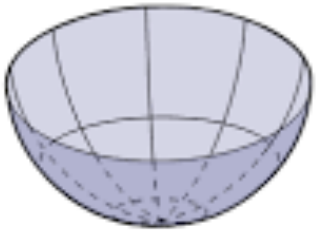


Positive (semi-)definite Matrices

- A symmetric matrix $\mathbf{A}$ is positive semi-definite (PSD) if for all $\boldsymbol{x} \in \mathrm{R}^n$, $\boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} \geq 0$

  Example:

$$z^T I z = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x^2 + y^2 \geq 0$$



  - The identity matrix $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is positive–definite

- A symmetric matrix $\mathbf{A}$ is positive definite (PD) if for all $\boldsymbol{x} \in \mathrm{R}^n$, $\boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} > 0$

Notation: $\mathbf{A} \succeq 0$ if A is PSD, $\mathbf{A} \succ 0$ if A is PD

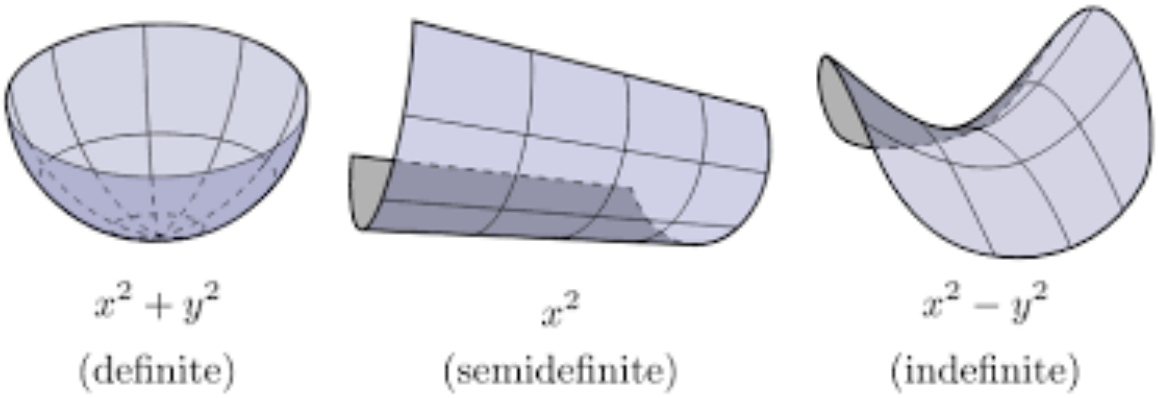
Positive (semi-)definite Matrices

Examples:

$$z^T Az = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x^2 + y^2$$

$$z^T Az = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x^2$$

$$z^T Az = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x^2 - y^2$$



$x^2 + y^2$ (definite)

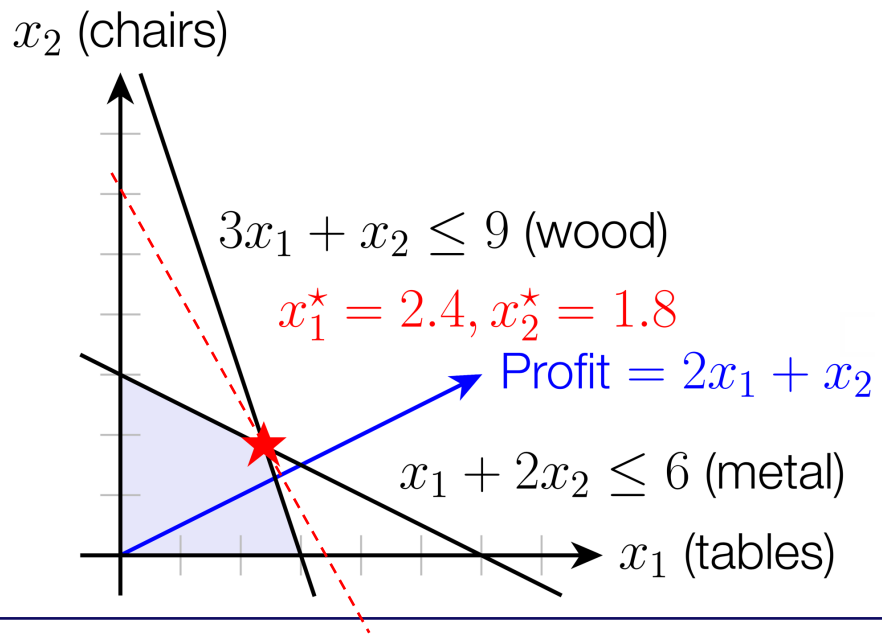
x^2 (semidefinite)

$x^2 - y^2$ (indefinite)

Linear Programming

Example: Optimal Manufacturing

A large factory makes tables and chairs. Each table returns a profit of $200 and each chair a profit of $100. Each table takes 1 unit of metal and 3 units of wood and each chair takes 2 units of metal and 1 unit of wood. The factory has 6K units of metal and 9K units of wood. How many tables and chairs should the factory make to maximize profit?



$$\begin{aligned} \underset{x_1,x_2}{\text{maximize}} \quad & 2x_1 + x_2 \\ \text{subject to} \quad & x_1 + 2x_2 \leq 6 \\ & 3x_1 + x_2 \leq 9 \\ & x_1, x_2 \geq 0 \end{aligned}$$

$$c = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \Downarrow \quad A = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix} \quad b = \begin{pmatrix} 6 \\ 9 \end{pmatrix}$$

$$\begin{aligned} \text{maximize} \quad & c^T x \\ \text{subject to} \quad & Ax \preceq b \\ & x \succeq 0 \end{aligned}$$

Example: Manufacturing

The General Form

- n products, m raw materials
- Every unit of product j uses a_{ij} units of raw material i
- There are b_i units of material i available
- Product j yields profit c_j per unit
- Facility wants to maximize profit subject to available raw materials

x_1	x_2	x_3	x_4	
a_{11}	a_{12}	a_{13}	a_{14}	b_1
a_{21}	a_{22}	a_{23}	a_{24}	b_2
a_{31}	a_{32}	a_{33}	a_{34}	b_3
c_1	c_2	c_3	c_4	

maximize $z = c_1x_1+c_2x_2+\ldots+c_nx_n$

subject to $a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n \leq b_1$

$\ldots$

$a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n \leq b_m$

and $x_i \geq 0$, for $i=1,\ldots,n$

Linear Programming: Canonical Form

$$\begin{aligned} \text{maximize} \quad & c^T x \\ \text{subject to} \quad & a_i^T x \leq b_i, \quad \text{for } i = 1, \ldots, m \\ & x_j \geq 0, \quad \text{for } j = 1, \ldots, n \end{aligned}$$

- Every LP can be **transformed** to the canonical form (规范形式):
  - Minimizing $c^T x$ is equivalent to maximizing $-c^T x$
  - ≥ constraints can be flipped by multiplying by -1
  - Each equality constraint can be replaced by two inequalities ($a_i^T x \leq b_i$ and $-a_i^T x \leq b_i$)
  - Unconstrained variables x_j can be replaced by $x_j^+ - x_j^-$, where both x_j^+ and x_j^- are constrained to be nonnegative.

Example:

minimize $3x_1+x_2$
subject to $x_1 > x_2+5$
$x_1+3x_2=10$

⇒

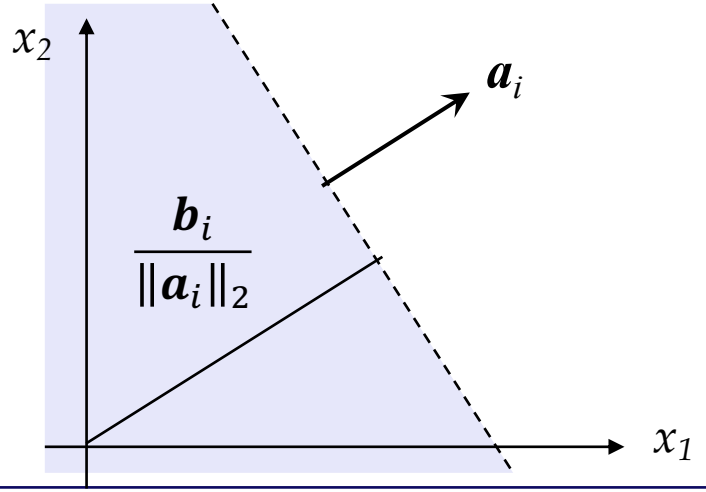
maximize $-3x_1-x_2$
subject to $-x_1+x_2 < -5$
$x_1+3x_2 \leq 10$
$-x_1-3x_2 \leq 10$

Geometric View

- Consider the inequality constraints of the linear program written out explicitly:

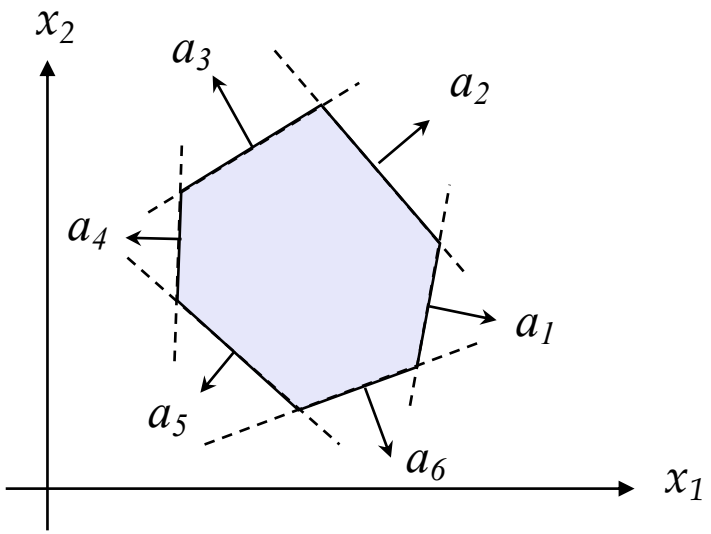
$$\begin{aligned} \underset{x}{\text{maximize}} \quad & c^T x \\ \text{subject to} \quad & a_i^T x \le b_i, \ i = 1,\ldots,m \end{aligned}$$

- Each linear inequality constraint $a_i^{\mathrm{T}} x \le b_i$ defines a half-space, which is orthogonal to the coefficient vector $\boldsymbol{a}$.



Geometric View

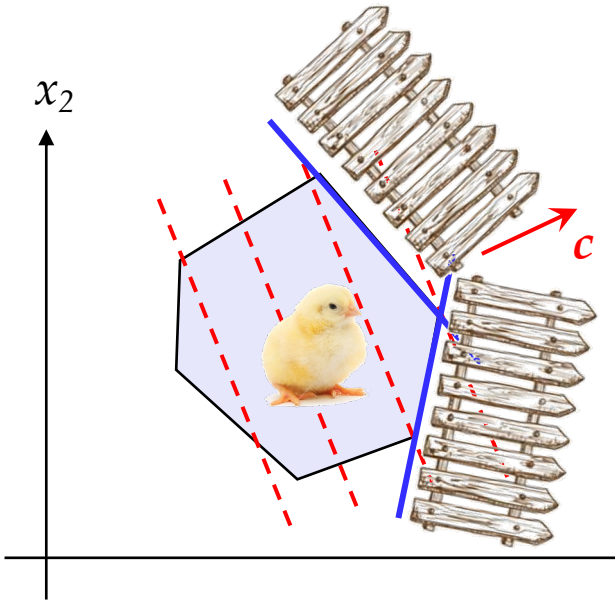
- Multiple halfspace constraints, $\boldsymbol{a}_i^T\boldsymbol{x} \le \boldsymbol{b}_i$, $i = 1, \ldots, m$ (or equivalently $\boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}$), define what is called a polyhedron (多面体).



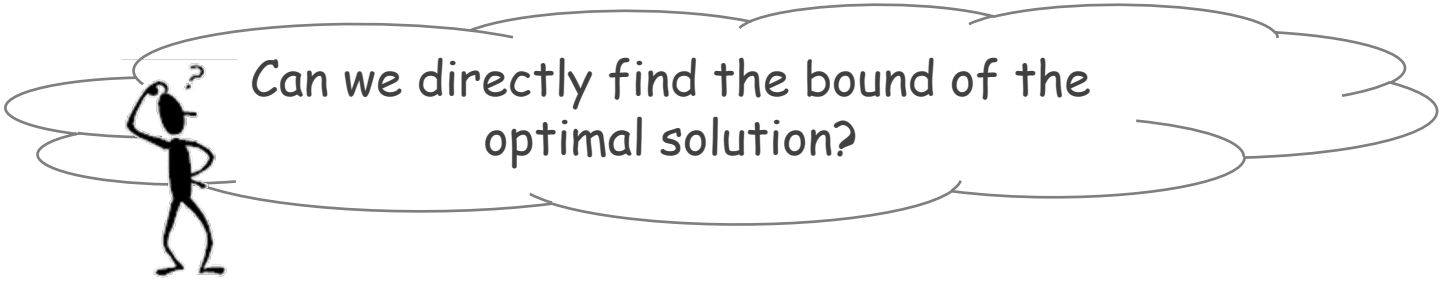
The feasible region of an LP is a polyhedron.

Duality in LP (对偶性)

Rethinking LP



Only the two constraints in the right corner are used, while the others not?



An Alternative View of LP

- Find the **upper bound** of the original objective

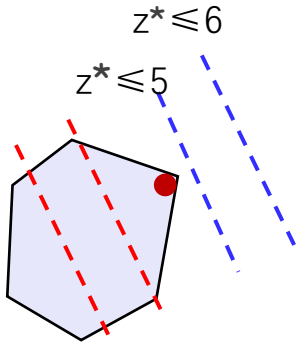
maximize $2x_1+3x_2 = z^*$
subject to $4x_1+8x_2 \le 12$,
$2x_1+x_2 \le 3$,
$3x_1+2x_2 \le 4$,
$x_1 \ge 0, x_2 \ge 0$

let z^* denote the original objective

What is the upper bound of z*?

Let's try to bound z^* from the original constrains:

$$2x_1+3x_2 \le 4x_1+8x_2 \le 12$$
$$2x_1+3x_2 \le \tfrac{1}{2}(4x_1+8x_2) \le \tfrac{1}{2}(12) = 6$$



So we have improved our upper bound to 6!

$$2x_1+3x_2 = \tfrac{1}{3}[(4x_1+8x_2)+(2x_1+x_2)] \le \tfrac{1}{3}(12+3) = 5$$

We further improved our upper bound to 5!

By continuously improving the upper bound, can we reach the original optimal?

An Alternative View of LP

maximize $2x_1+3x_2$

s.t. $y_1 \times (\ 4x_1+8x_2 \le 12\ )$

$y_2 \times (\ 2x_1+x_2 \le 3\ )$

$y_3 \times (\ 3x_1+2x_2 \le 4\ )$

minimize $12y_1+3y_2+4y_3$

subject to $4y_1+2y_2+3y_3 \ge 2$

$8y_1+y_2+2y_3 \ge 3$

$y_1,y_2,y_3 \ge 0$

We are finding non-negative multipliers y_1, y_2, y_3 to the original constrains, which yields

equivalent to the dual problem:

$z^*=2x_1+3x_2 \le (4y_1+2y_2+3y_3)x_1+(8y_1+y_2+3y_3)x_2 \le 12y_1+3y_2+4y_3$

We want it to be an upper bound of the original optimality

Linear Programming Duality

	x_1	x_2	x_3	x_4	
y_1	a_{11}	a_{12}	a_{13}	a_{14}	b_1
y_2	a_{21}	a_{22}	a_{23}	a_{24}	b_2
y_3	a_{31}	a_{32}	a_{33}	a_{34}	b_3
	c_1	c_2	c_3	c_4	

Primal LP

$$\begin{aligned} \text{maximize} \quad & c^T x \\ \text{subject to} \quad & Ax \preccurlyeq b \\ & x \succcurlyeq 0 \end{aligned}$$

Dual LP

$$\begin{aligned} \text{minimize} \quad & b^T y \\ \text{subject to} \quad & A^T y \succcurlyeq c \\ & y \succcurlyeq 0 \end{aligned}$$

- $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$
- y_i is the dual variable corresponding to primal constraint $A_i x \leq b_i$
- $A_j^T y \geq c_j$ is the dual constraint corresponding to primal variable x_j
- Every feasible solution y of dual LP provides an upper bound on the maximum of the objective function of primal LP

Interpretation 1: Economic Interpretation

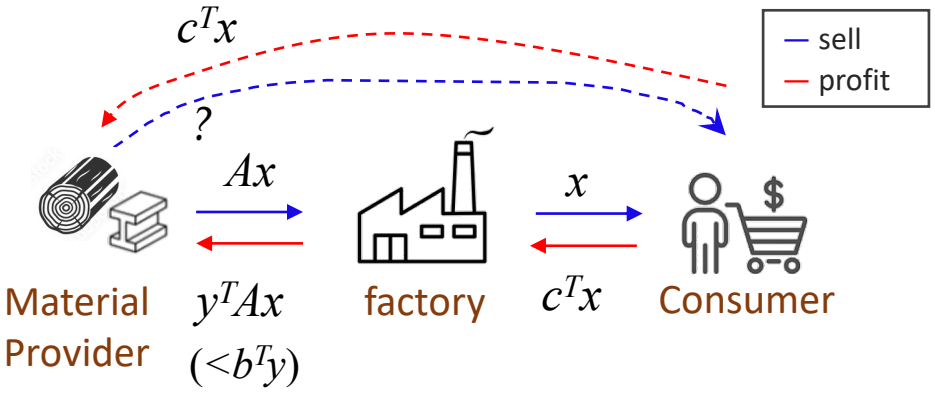
Primal LP

$$\begin{aligned} \max \quad & \sum_{j=1}^{n} c_j x_j \\ \text{s.t.} \quad & \sum_{j=1}^{n} a_{ij} x_j \le b_i, \quad \text{for } i \in [m] \\ & x_j \ge 0, \quad \text{for j} \in [n] \end{aligned}$$

Dual LP

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} b_i y_i \\ \text{s.t.} \quad & \sum_{i=1}^{m} a_{ij} y_i \ge c_j, \quad \text{for j} \in [n] \\ & y_i \ge 0, \quad \text{for i} \in [m] \end{aligned}$$

	x_1	x_2	x_3	x_4	
y_1	a_{11}	a_{12}	a_{13}	a_{14}	b_1
y_2	a_{21}	a_{22}	a_{23}	a_{24}	b_2
y_3	a_{31}	a_{32}	a_{33}	a_{34}	b_3
	c_1	c_2	c_3	c_4	



- Dual variable y_i is a proposed **price** per unit of raw material i
- Dual price vector is feasible if facility has incentive to sell materials
- Buyer wants to spend as little as possible to buy materials

Interpretation 2: Finding the Best Upper bound

	x_1	x_2	x_3	x_4	
y_1	a_{11}	a_{12}	a_{13}	a_{14}	b_1
y_2	a_{21}	a_{22}	a_{23}	a_{24}	b_2
y_3	a_{31}	a_{32}	a_{33}	a_{34}	b_3
	c_1	c_2	c_3	c_4	

Primal LP

maximize $2x_1+3x_2$

$y_1 \times (4x_1+8x_2 \le 12)$

$y_2 \times (2x_1+x_2 \le 3)$

$y_3 \times (3x_1+2x_2 \le 4)$

Dual LP

minimize $12y_1+3y_2+4y_3$

subject to $4y_1+2y_2+3y_3 \ge 2$

$8y_1+y_2+2y_3 \ge 3$

$y_1, y_2, y_3 \ge 0$

$z^*=2x_1+3x_2 \le (4y_1+2y_2+3y_3)x_1+(8y_1+y_2+3y_3)x_2 \le 12y_1+3y_2+4y_3$ ← Upper Bound

- Multiplying each row i by y_i and summing gives the inequality

$$y^T Ax \le y^T b$$

- When $y^T A \ge c^T$, the right hand side of the inequality is an upper bound on $c^T x$ for every feasible x.

$$c^T x \le y^T Ax \le y^T b$$

- The dual LP can be thought of as trying to find the best upper bound on the primal that can be achieved this way.

Mathematic Foundations for Computer Science

Probability and Optimization

Chapter 4: Convex Optimization

Spring 2024

Instructor: Xiaodong Gu

Standard Form

- Find the minimizer of a function subject to constraints:

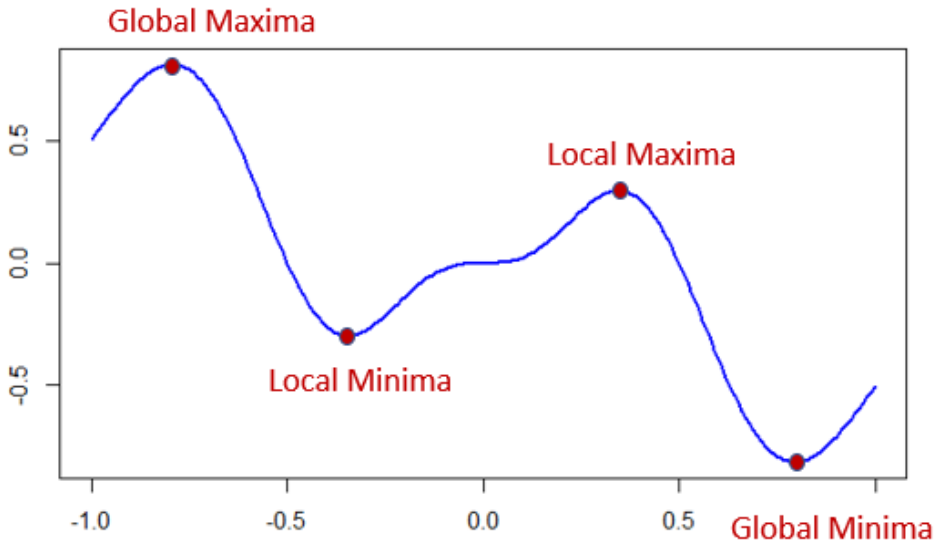
$$\begin{array}{lll} \underset{x}{\text{minimize}} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0 & i = 1, \ldots, m \\ & h_i(x) = 0 & i = 1, \ldots, p \end{array}$$

where

$x = (x_1, \ldots, x_n)$ is the optimization variable
$f_0 : \mathbf{R}^n \longrightarrow \mathbf{R}$ is the objective function
$f_i : \mathbf{R}^n \longrightarrow \mathbf{R}, \quad i = 1, \ldots, m$ are inequality constraint functions
$h_i : \mathbf{R}^n \longrightarrow \mathbf{R}, \quad i = 1, \ldots, p$ are equality constraint functions.

Goal: find an optimal solution $x^\star$ that minimizes f_0 while satisfying all the constraints.

Local Minima and Global Minima



Local Minima:
A solution that is optimal within a neighboring set.

Global Minima:
The optimal solution among all possible solutions.

What kind of functions have local minima == global minima?

Convex Optimization

Convex Optimization

A problem of minimizing a convex function (or maximizing a concave function) over a convex set.

- The standard form:

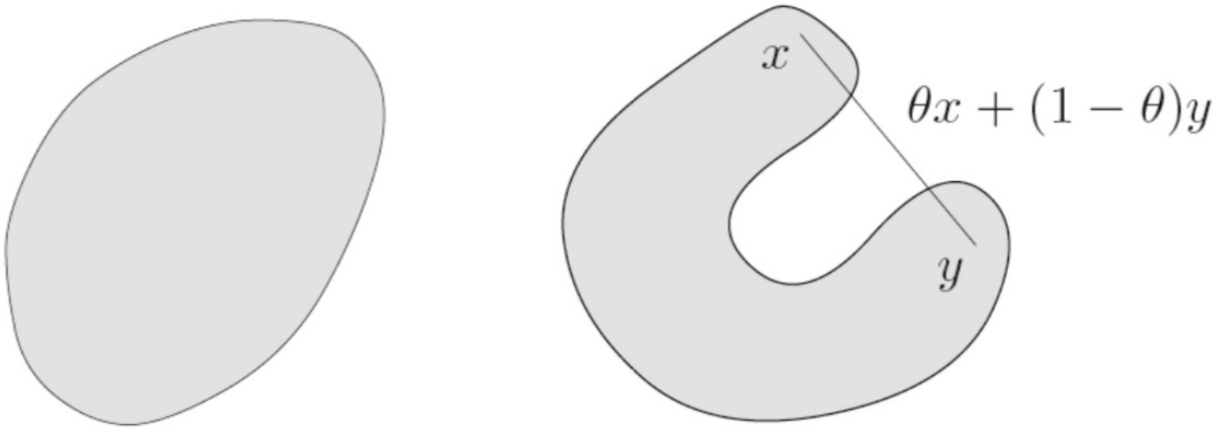
$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && g_i(x) \le 0, \text{ for i} \in C_1. \\ & && h_i(x) = 0, \text{ for i} \in C_2. \end{aligned}$$

where f_0, g_i, h_i are **convex**

Convex optimization problems have **local minima == global minima**

Convex Set

- A set $C \in \mathbb{R}^n$ is said to be **convex** if the line segment between any two points is in the set: for any $x, y \in C$ and $0 \leq \theta \leq 1$,



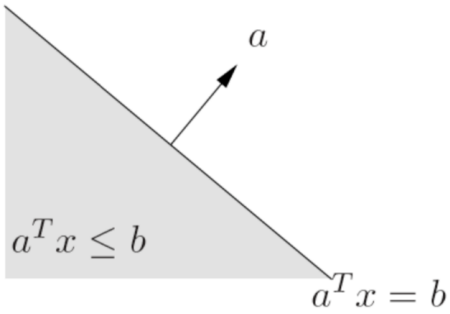
convex

non-convex

Examples

- **Trivial**: $\emptyset$, point, line, etc.
- **Hyperplane**:

  $C = \{x \mid a^T x = b\}$ where $a \in \mathbb{R}^n$, $b \in \mathbb{R}$



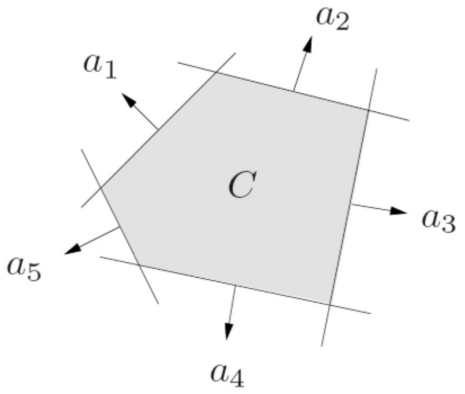
- **Halfplane**:

  $C = \{x \mid \boldsymbol{a}^T x \leq b\}$

- **Polyhedron**:

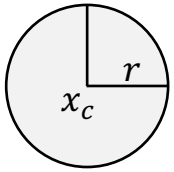
  $C = \{x \mid \boldsymbol{A}x \leq b, \boldsymbol{C}x = d\}$

  where $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, C \in \mathbb{R}^{p \times n}, d \in \mathbb{R}^p$



- **Euclidean ball:**

  $B(x_c, r) = \{x \mid \|x - x_c\|_2 \leq r\}$

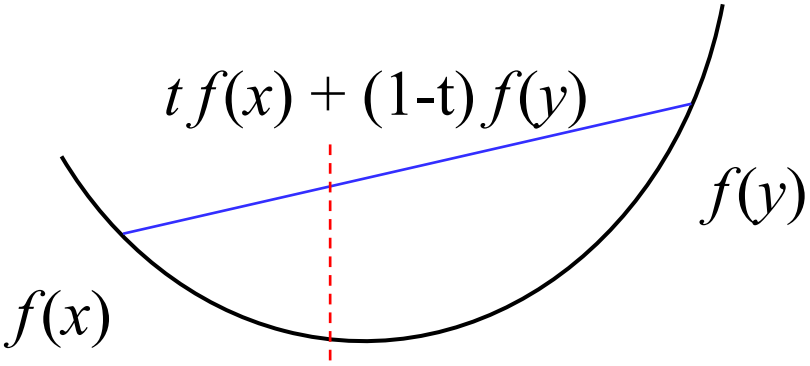


Convex Functions

- A function $f: \mathbb{R}^n \to \mathbb{R}$ is convex if for $x, y \in \mathbf{dom}\, f$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y),$$

for all $t \in [0, 1]$.



Example of Convex Functions

- Exponential function: e^{ax}
- Logarithmic function $\log(x)$ is concave
- Affine function: $A^{\mathrm{T}}x+b$
- Quadratic function: $x^{\mathrm{T}}Qx + b^{\mathrm{T}}x + c$ is convex if Q is positive semidefinite (PSD)
- Least squares loss: $\|y - Ax\|^2_2$
- Norm: $\| x \|$ is convex for any norm

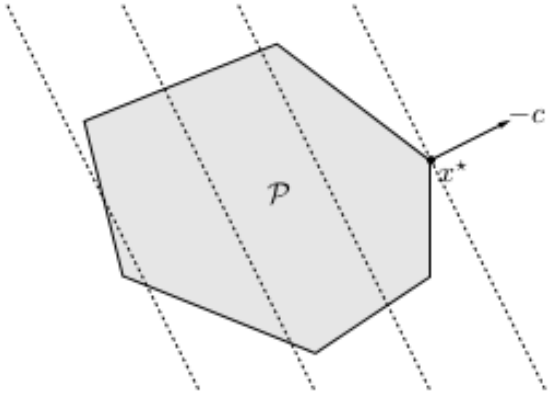
$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \qquad \|x\|_1 = \sum_{i=1}^{n} |x_i|$$

Convex Optimization Problems

Linear Programming

- We have already seen linear programming

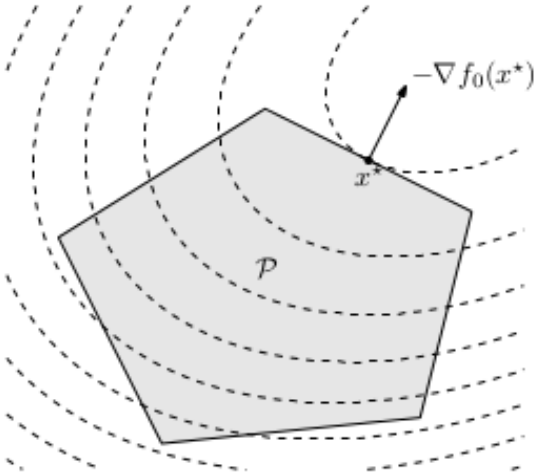
$$\begin{aligned} \text{minimize} \quad & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \geq 0 \end{aligned}$$



Quadratic Programming (二次规划)

- Minimize a convex quadratic function (二次函数) over a polyhedron. Require $P \succcurlyeq 0$.

$$\begin{array}{ll} \text{minimize} & x^\intercal P x + c^\intercal x + d \\ \text{subject to} & Ax \leq b \end{array}$$



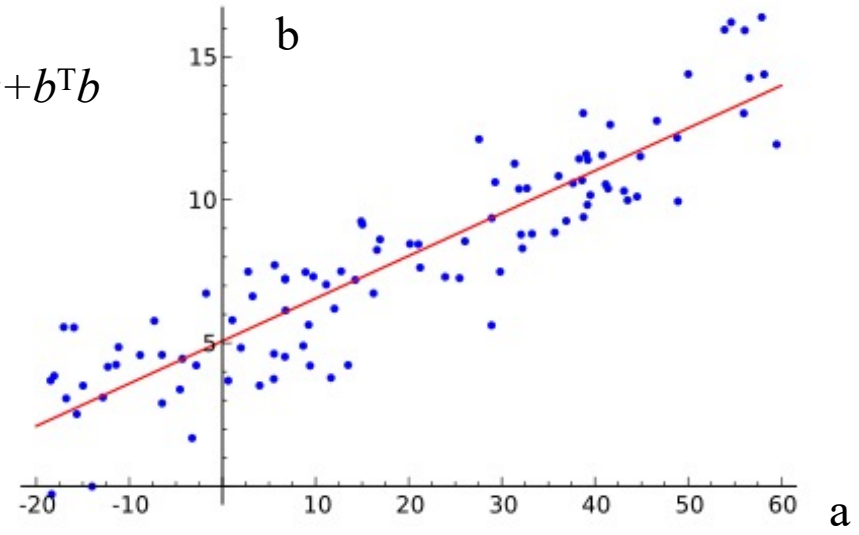
$P \succcurlyeq 0$: positive semi-definite, $P \in S^n_+$
symmetric & all eigenvalues are nonnegative,
$x^T \boldsymbol{P} x \geq 0$ for all x

Example: QP

Constrained Least Squares

Given a set of measurements $(a_1, b_1), \ldots, (a_m, b_m)$, where $a_i \in \mathrm{R}^n$ is the i-th input and $b_i \in \mathrm{R}$ is the i-th output, fit a linear function minimizing mean square error, subject to known bounds on the linear coefficients.

minimize $||\boldsymbol{A}x - b||^2_2 = x^T\boldsymbol{A}^T\boldsymbol{A}x\text{-}2b^T\boldsymbol{A}x+b^Tb$
subject to $l_i \le x_i \le u_i$, for $i = 1,\ldots,n$.

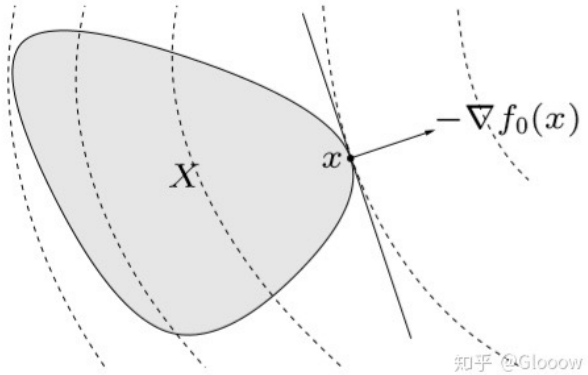


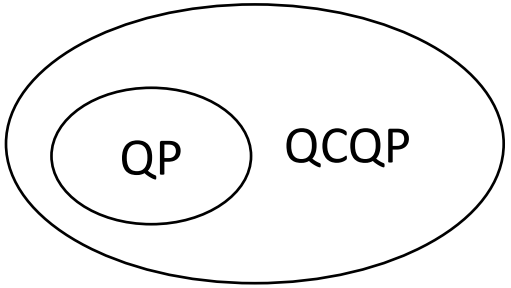
Quadratically Constrained Quadratic Programming (QCQP) 二次约束二次规划

- Minimize a convex quadratic function (二次函数) over a quadratic constraint.

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P_0 x + q_0^T x + r_0 \\ \text{subject to} & (1/2)x^T P_i x + q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

- $P_i \in \mathbf{S}_+^n$; objective and constraints are convex quadratic
- if $P_1, \ldots, P_m \in \mathbf{S}_{++}^n$, feasible set is intersection of m ellipsoids and an affine set





Geometric Programming

- **Geometric programming** is an optimization problem of the following form

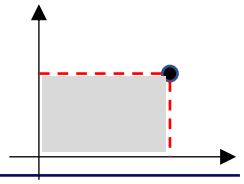
$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le b_i, \quad \text{for } i \in \mathcal{C}_1. \\ & h_i(x) = b_i, \quad \text{for } i \in \mathcal{C}_2. \\ & x \succeq 0 \end{array}$$

where f_i's are **posynomials**, h_i's are **monomials**, and $b_i > 0$ ($w \log 1$).

Definition

- A monomial (单项式) is a function $f: \mathrm{R}^n_+ \to \mathrm{R}_+$ of the form
$$f(x) = c x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}$$
where $c \ge 0$, $a_i \in \mathrm{R}$.
- A posynomial (正多项式) is a sum of monomials.

Interpretation: GP minimizes volume/area, subject to constraints.



Example: GP

A manufacturer is designing a suitcase (手提箱)

- variables: h, w, d
- want to minimize surface area: $2(hw + hd + wd)$ (i.e., amount of material used)
- have a target volume: $hwd \geq 5$
- practical/aesthetic (美学) constraints limit aspect ratio:

$$h/w \leq 2,\ h/d \leq 3$$

- constrained by airline to $h + w + d \leq 7$

$$\begin{array}{ll} \text{minimize} & 2hw + 2hd + 2wd \\ \text{subject to} & h^{-1}w^{-1}d^{-1} \leq \frac{1}{5} \\ & hw^{-1} \leq 2 \\ & hd^{-1} \leq 3 \\ & h + w + d \leq 7 \\ & h, w, d \geq 0 \end{array}$$

$$\begin{aligned} \widetilde{h} &= \log h \\ \widetilde{w} &= \log w \\ \widetilde{d} &= \log d \end{aligned}$$

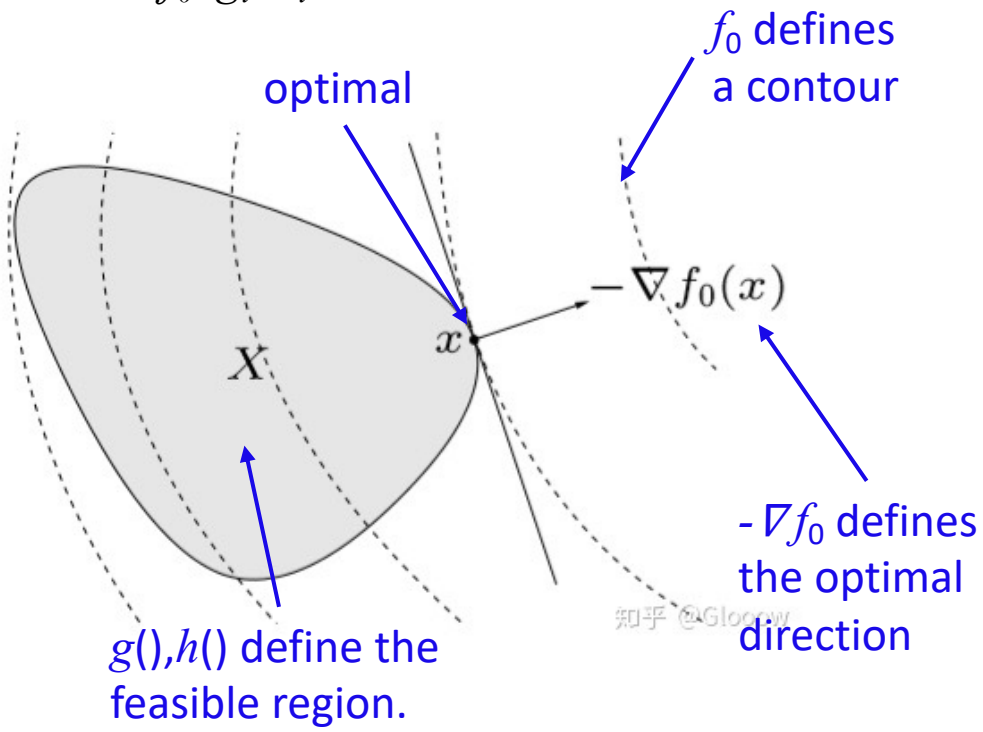
$\Longrightarrow$

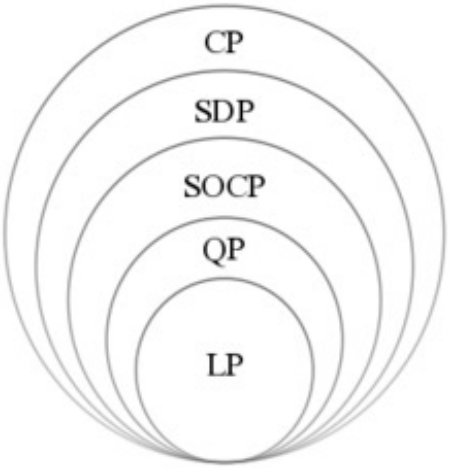
$$\begin{array}{ll} \text{minimize} & 2e^{\widetilde{h}+\widetilde{w}} + 2e^{\widetilde{h}+\widetilde{d}} + 2e^{\widetilde{w}+\widetilde{d}} \\ \text{subject to} & e^{-\widetilde{h}-\widetilde{w}-\widetilde{d}} \leq \frac{1}{5} \\ & e^{\widetilde{h}-\widetilde{w}} \leq 2 \\ & e^{\widetilde{h}-\widetilde{d}} \leq 3 \\ & e^{\widetilde{h}} + e^{\widetilde{w}} + e^{\widetilde{d}} \leq 7 \end{array}$$

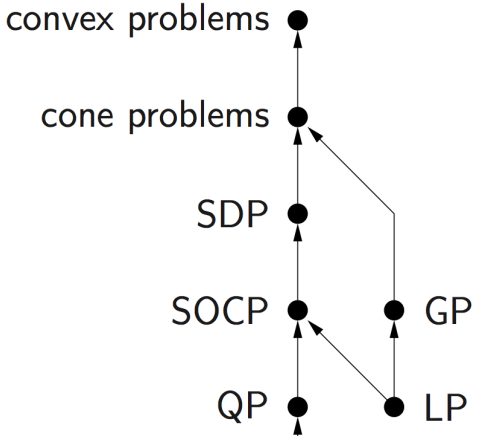
Summary

minimize $f_0(x)$
subject to $g_i(x) \le 0$, for i $\in C_1$.
$h_i(x) = 0$, for i $\in C_2$.

where f_0, g_i, h_i are **convex**







Optimization Algorithms

Unconstrained Minimization

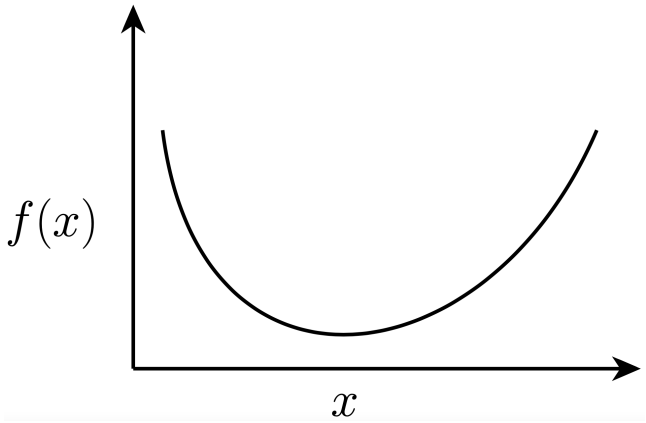
- Gradient descent, Newton's method

Constrained Minimization

- Interior-point Methods

Unconstrained Minimization

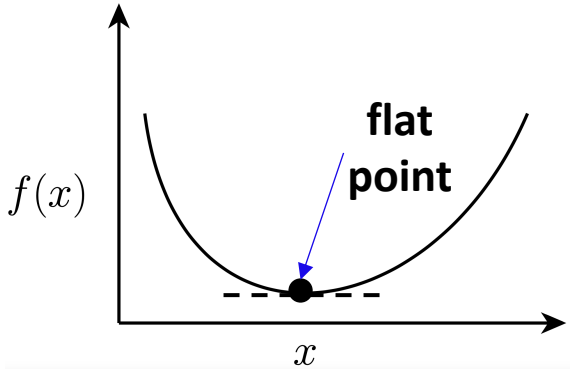
$$\text{minimize } f(x)$$



- f convex, twice continuously differentiable (hence **dom** f open)

How to Solve Unconstrained Minimization?

- Starting with the one dimensional case

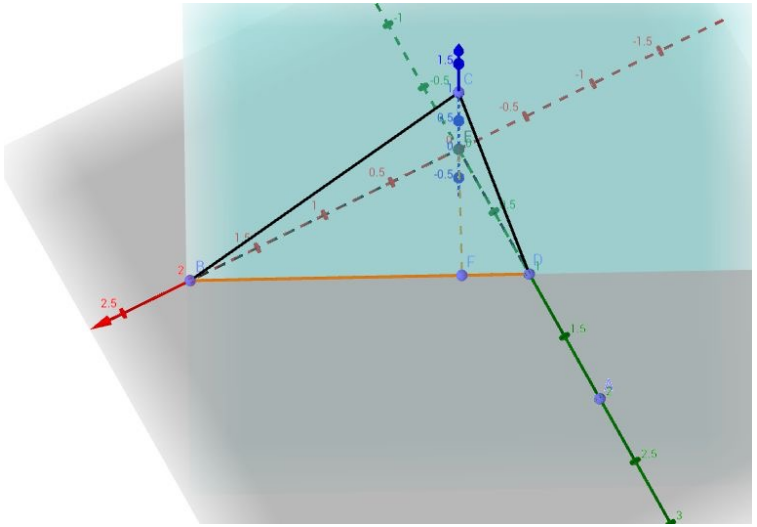


- To find minimum point $x^\star$, we can look at the derivative of the function $f'(x)$: any location where $f'(x) = 0$ will be a **"flat" point** in the function
- For convex problems, this is guaranteed to be a minimum.

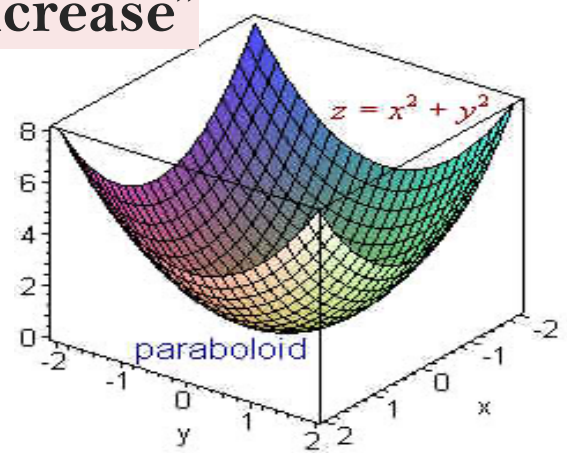
What is the "flat" point in the multi-dimensional case?

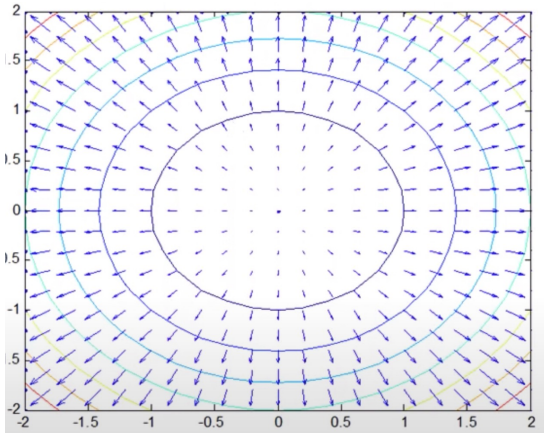
Gradient (梯度)

"direction and rate of the fastest increase"



https://www.zhihu.com/question/36301367





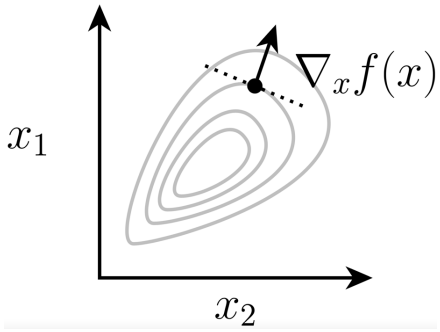
Contours for x^2+y^2 with gradients

Gradient

Definition

For f: $R^n \rightarrow R$, its gradient $\nabla_x f$: $R^n \rightarrow R^n$ at point $x = (x_1, \ldots, x_n)$ is defined as a vector containing partial derivatives with respect to each dimension.

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$



- For continuously differentiable f and unconstrained optimization, optimal point must have $\nabla_x f(x^\star) = 0$

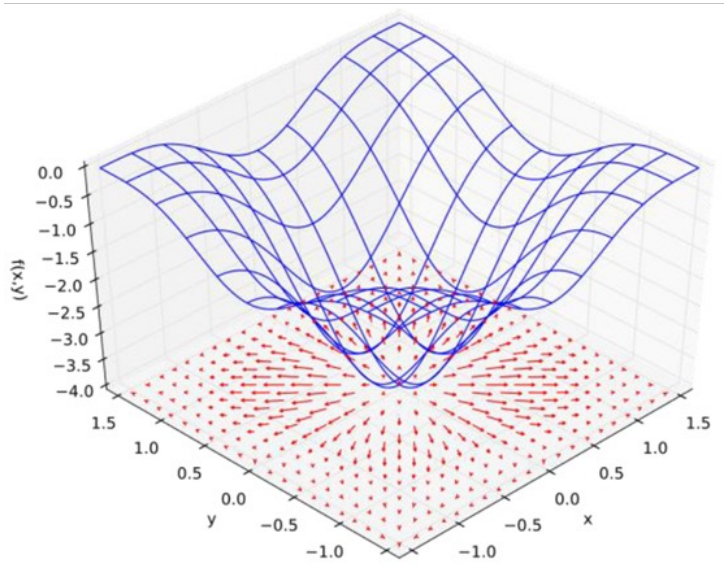
Gradient

Examples:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

$$f(x) = 2x_1^2 x_2 - x_1 x_3^3$$

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \frac{\partial}{\partial x_3} f(x) \end{bmatrix} = \begin{bmatrix} 4x_1 x_2 - x_3^3 \\ 2x_1^2 \\ -3x_1 x_3^2 \end{bmatrix}$$

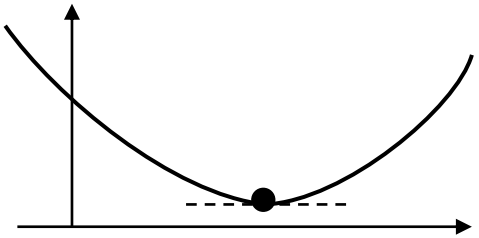


gradients for arbitrary function $f(x,y)$

How to Solve Unconstrained Minimization?

Direct Solution

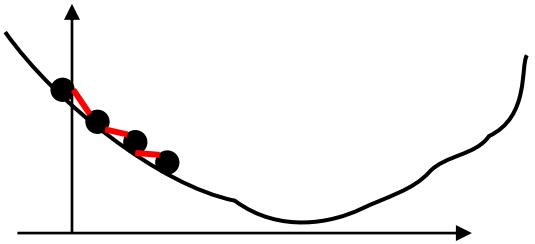
- In some cases, it is possible to analytically compute the $\boldsymbol{x}^\star$ such that $\nabla_x f(x^\star) = 0$



Iterative methods

- More commonly the condition $\nabla_x f(x^\star) = 0$ will not have an analytical solution, require iterative methods
- Produce sequence of points $x^{(k)} \in \mathbf{dom}\, f$, k=0,1,… with

$$f(x^{(k)}) \rightarrow \mathrm{OPT(primal)}$$



Iterative Methods

Algorithms

- gradient descend
- steepest descend
- Newton's method
- …

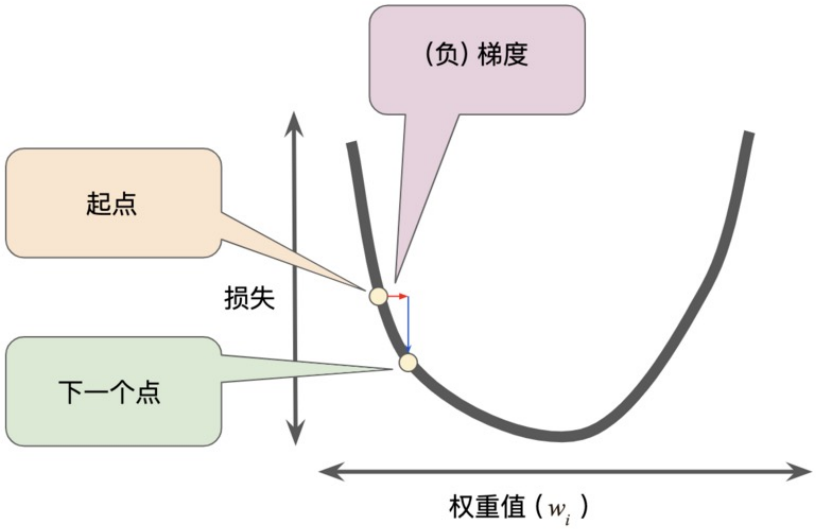
Gradient Descend

- The gradient doesn't just give us the optimality condition, it also points in the direction of "**steepest ascent**" for the function f.
- Motivates the gradient descent algorithm, which repeatedly takes steps in the direction of the negative gradient.

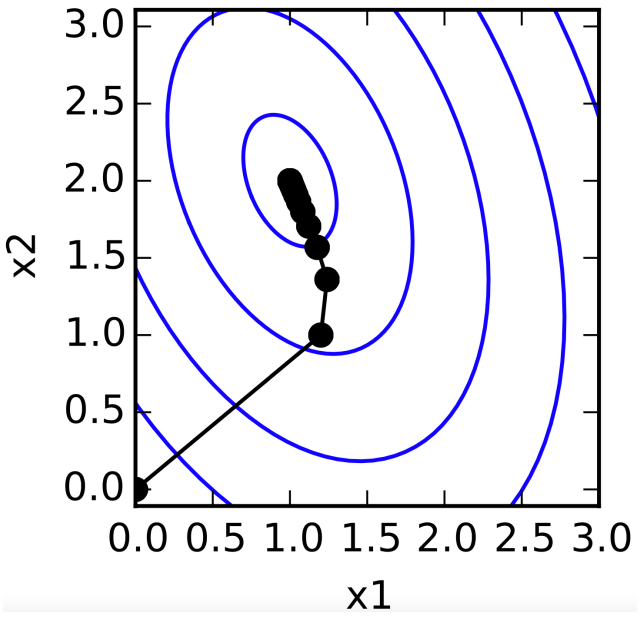
- Goal: $\min_x f(x)$
- Iteration:

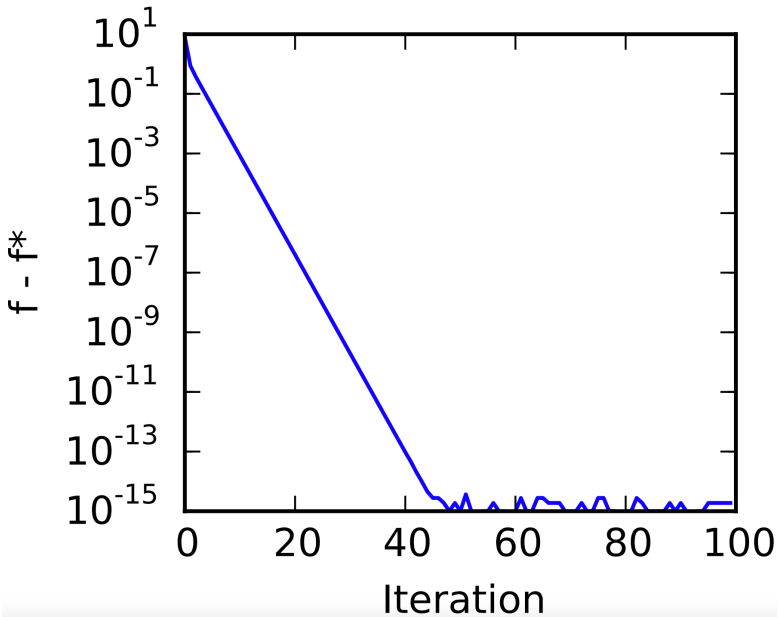
$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

- η_t is the step size.



Gradient Descend





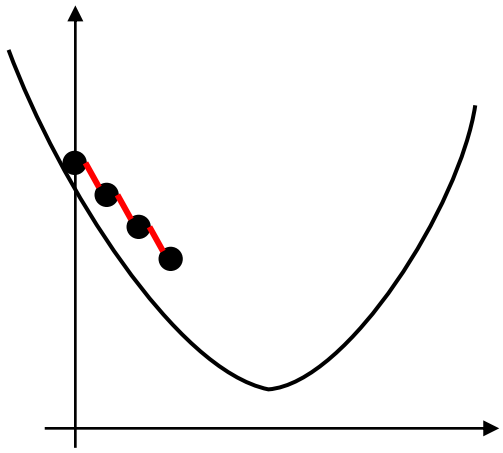
100 iterations of gradient descent on function

$$f(x) = 2x_1^2 + x_2^2 + x_1x_2 - 6x_1 - 5x_2$$

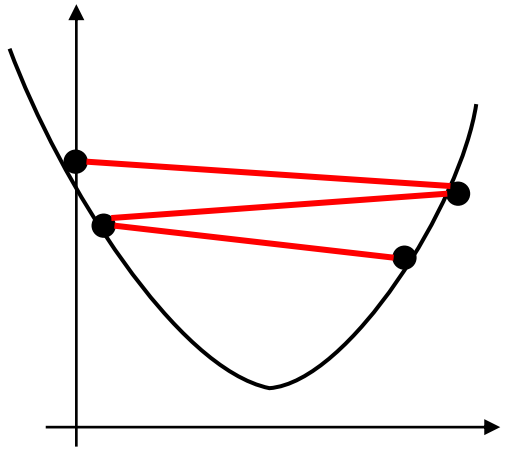
How to Choose Step Size?

$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

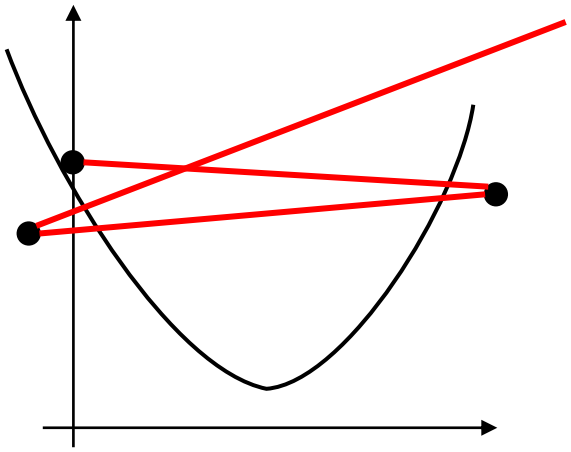
- If the step size is too big, the value of function can diverge.
- If the step size is too small, the convergence is too slow.



η too small:
slow progress

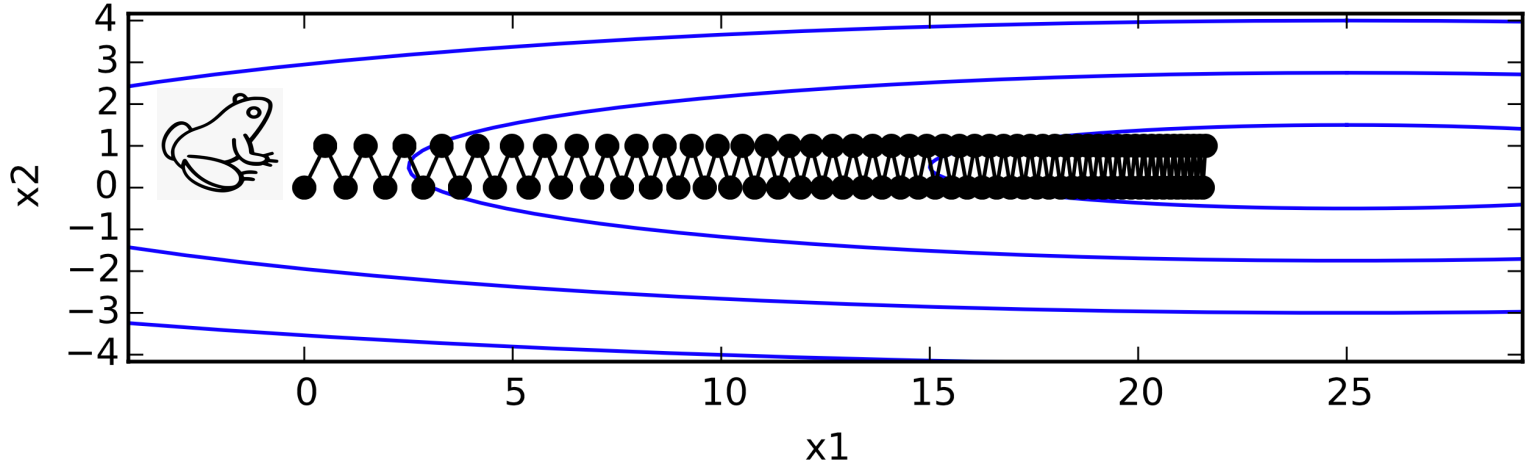


η too large:
oscillations



η much too large:
instability

Trouble with Gradient Descent



Gradient descent with backtracking line search on function

$$f(x) = 0.01x_1^2 + x_2^2 - 0.5x_1 - x_2$$

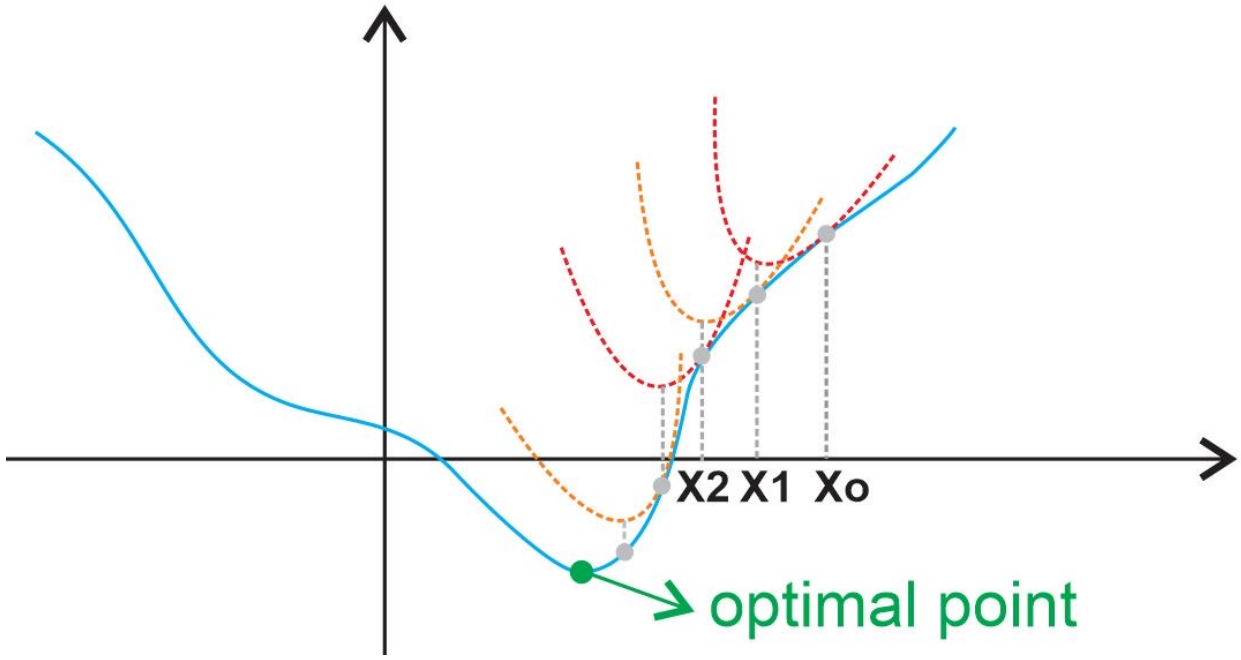
Gradient is given by $(0.02x_1 - 0.5,\ 2x_2 - 1)$ which is very poorly scaled ($x_2 >> x_1$)

- Motivates approaches that "automatically" find the right scaling.

Newton's Method

- The next step moves to the minimal of the second order approximation of f

$$x_{t+1} = x_t - [\nabla^2 f(x_t)]^{-1} \nabla f(x_t)$$



Newton's Method

Interpretations

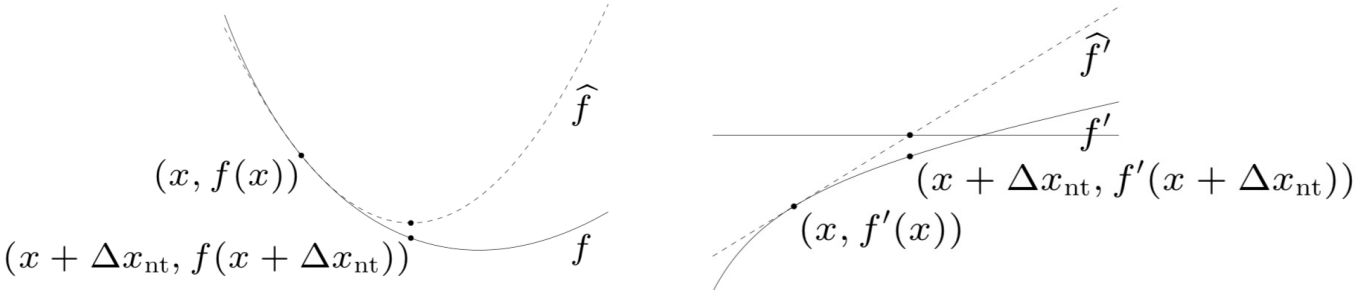
- $x+\Delta x_{\text{nt}}$ minimizes second order approximation

$$\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

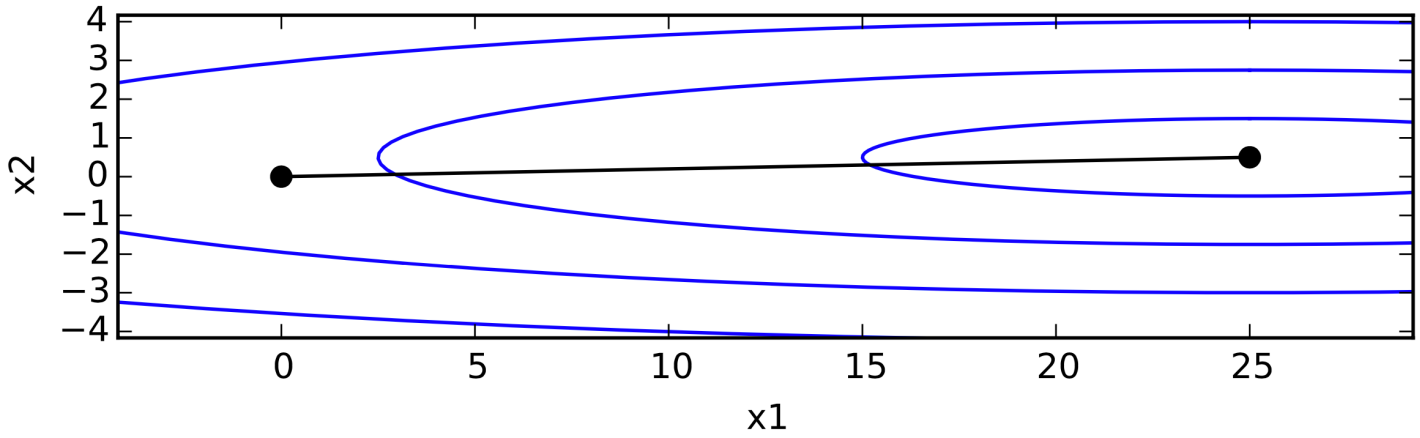
Let $\nabla \hat{f}(x+v) = 0$:

$$\nabla f(x+v) \approx \nabla \hat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$

$$\Longrightarrow v^\star = \Delta x_{\text{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$



Newton's Method



$$f(x) = 0.01x_1^2 + x_2^2 - 0.5x_1 - x_2$$

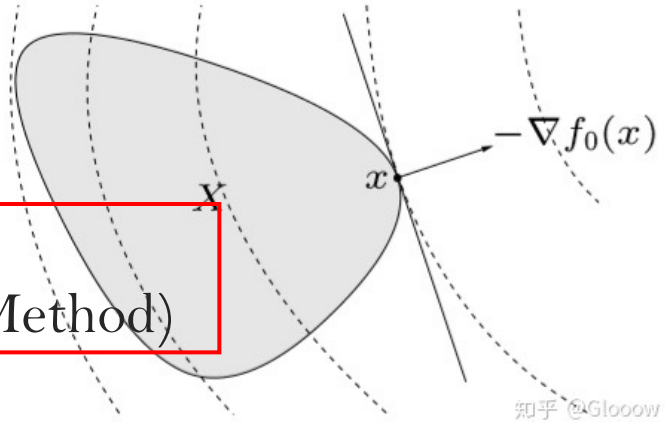
For our previous example, Newton's method finds the exact solution in a single step (holds true for any convex quadratic function)

- Newton's method is usually **much** faster than gradient descent

Constrained Minimization

Penalty Function Method (罚函数法)

- Quadratic Penalty Method
- Interior Point Method (内点法)
  (Logarithmic Penalty Method, Barrier Method)
- …



Augmented Lagrangian method

…

Inequality Constrained Minimization

- What about minimization with inequality constrains ?

$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && f_i(x) \le 0,\ i = 1,\ldots,m \\ & && Ax = b \end{aligned}$$

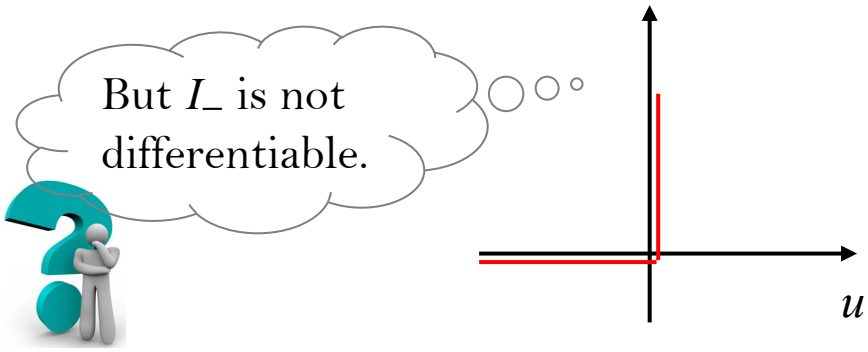
where f_i's are assumed to be convex & twice continuously differentiable

Trapping Points within the Feasible Set

- The inequality constraints can be made implicit by rewriting the objective with a penalty function $I_{-}()$.

$$\begin{aligned} &\min f_0(x) \\ &\text{s.t. } f_i(x) \le 0,\ i = 1,\ldots,m \\ &\quad Ax = b \end{aligned} \quad \Longrightarrow \quad \begin{aligned} &\min f_0(x) + \sum_{i=1}^{m} I_{-}(f_i(x)) \\ &\text{s.t. } Ax = b \end{aligned}$$

$$I_{-}(u) = \begin{cases} 0 & u \le 0 \\ \infty & \text{otherwise} \end{cases}$$



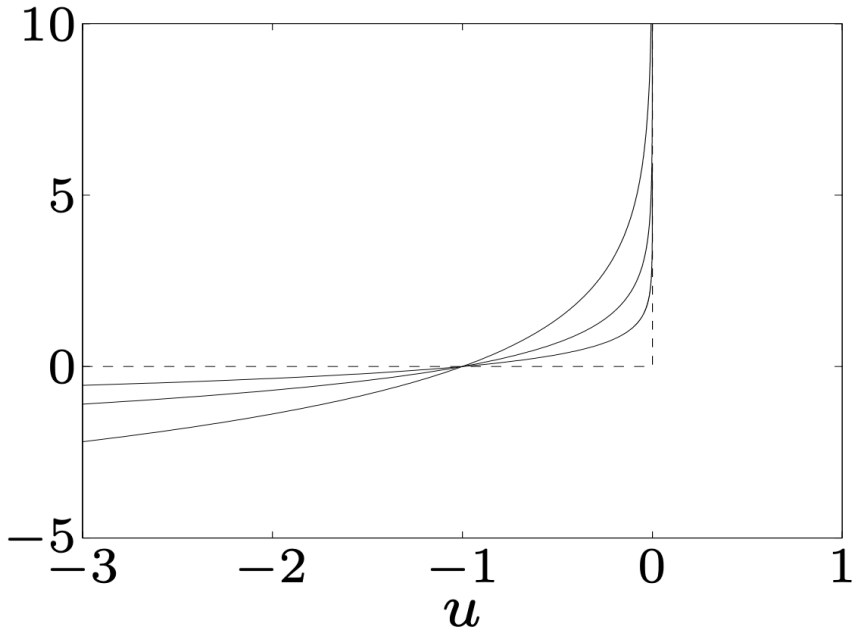
Idea: approximate I_{-} by some differentiable function.

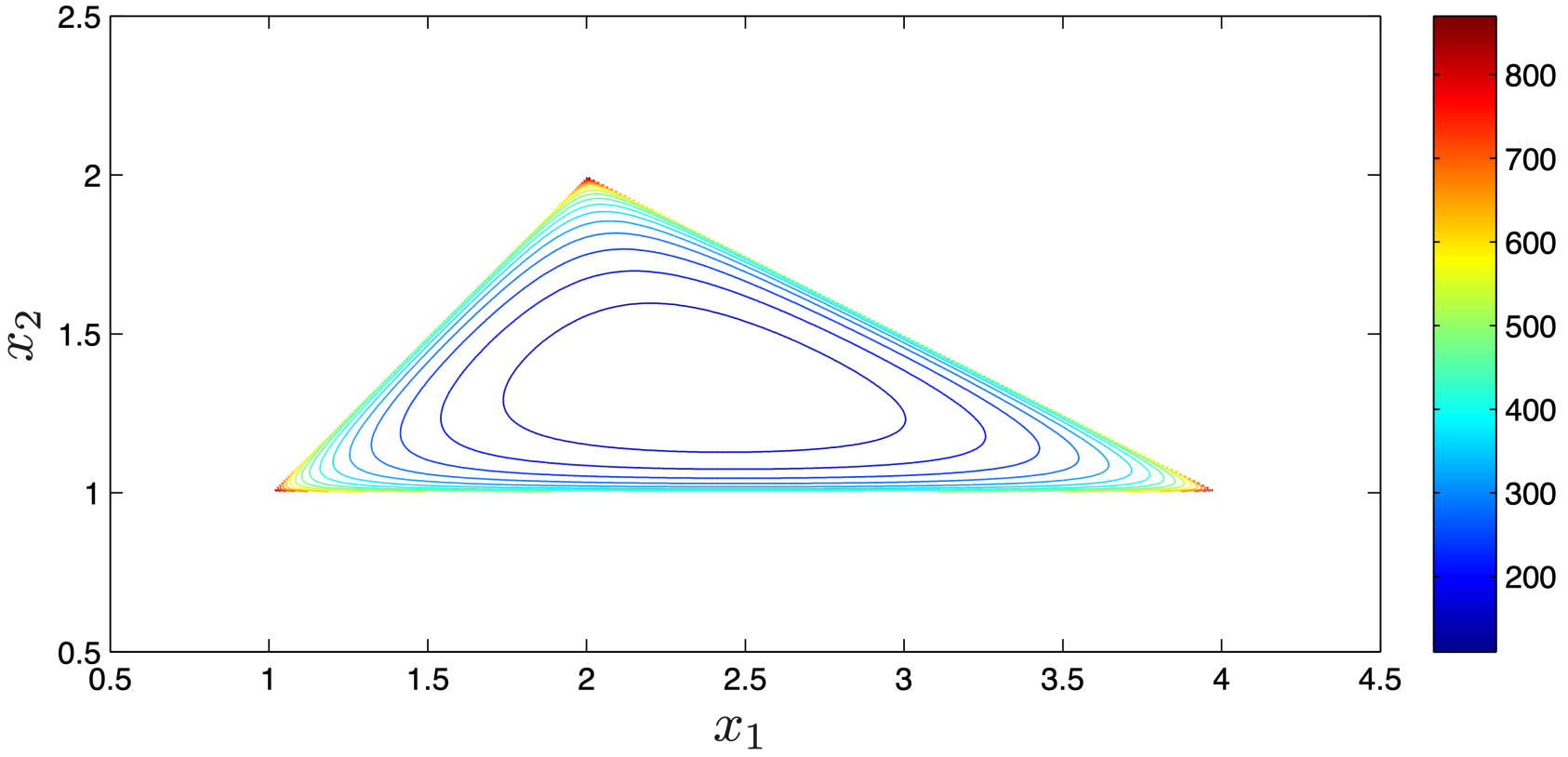
Logarithmic Barrier Function (对数闸函数)

- Approximate I_- by

$$I_-(u) = -\mu \log(-u), \quad \textbf{dom}\, I_- = \{x \in \mathbb{R} \mid x < 0\}$$

where $\mu > 0$ is a parameter that controls the accuracy of the approx.





Example: A set of linear inequalities $b_i - a_i^T x \leq 0$, $i = 1,...,m$.

The corresponding approx. is

$$-\mu \sum_{i=1}^{m} \log(a_i^T x - b_i)$$

Logarithmic Barrier Function

- The original problem can be approximated as:

$$\begin{aligned} \min \ & f_0(x) + \mu\phi(x) \\ \text{s.t.} \ & Ax = b \end{aligned}$$

with $$\phi(x) = -\sum_{i=1}^{m} \log(-f_i(x)), \quad \mathbf{dom}\,\phi = \{x \mid f_1(x) < 0, \ldots, f_m(x) < 0\}$$

ϕ is called the **logarithmetic barrier function**. Some nice properties:
– ϕ is convex (by composition).
– ϕ is twice differentiable:

$$\nabla\phi(x) = \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla f_i(x)$$

$$\nabla^2\phi(x) = \sum_{i=1}^{m} \frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^T + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla^2 f_i(x)$$

This means that the objective function is convex and twice differentiable.