Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 2: A basic review of probability theory and statistics

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2017)

lions@epfl

MARIE CURIE ACTIONS
FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION
erc
EPFL

License Information for Mathematics of Data Slides

- This work is released under a Creative Commons License with the following terms:
- **Attribution**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- **Non-Commercial**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- **Share Alike**
  - The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- Full Text of the License

- This lecture
  1. Review of probability theory
  2. Learning as an optimization problem
- Next lecture
  1. Basic concepts in convex analysis
  2. Complexity theory review

Recommended reading

- *Probability and Measure*, Patrick Billingsley, Wiley-Interscience, 1995.
- Chapter 7, 8, & 9 in K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.
- V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Inf. Theory*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
- *Chapter 5 in A. W. van der Vaart, *Asymptotic Statistics*, Cambridge Univ. Press, 1998.

Motivation

Example (Gene Mutation)

Each gene has a mutation probability, say μ. We want to find out μ through a series of *independent* experiments.

For instance, Phenylketonuria is a disease caused by mutations of the PAH gene. In Australia, there are roughly 2400 patients out of a 24 million population.

How would you estimate the mutation probability, if you were an Australian doctor?

Motivation

Key questions

- How do we model the problem rigorously?
- How well can we do, after the model is specified?
- How can we solve the problem?

Motivation

Key questions

- How do we model the problem rigorously?
- How well can we do, after the model is specified?
- How can we solve the problem?

(Partial) answers

- How do we model the problem rigorously? — Probability theory
- How well can we do, after the model is specified? — Statistical guarantees
- How can we solve the problem? — Optimization algorithms

Motivation

Formal Setup

We introduce the rigorous framework for probability theory, and discuss several important statistical problems that motivate our subsequent optimization lectures.

Basic concepts in probability theory

Definition (Sample space)

The sample space Ω of an experiment is the set of all possible outcomes of that experiment.

Example

If the experiment is testing whether a gene will mutate, the sample space is the set given by $\{$`mutatation, no mutation`$\}$.

Definition (Event)

An event E corresponds to a subset of the sample space; i.e., $E \subseteq \Omega$.

Definition (Probability measure)

Probability measure $P(E)$ maps event E from Ω onto the interval $[0, 1]$ and satisfies the following Kolmogorov axioms:

- $P(E) \geq 0$,
- $P(\Omega) = 1$ and
- $P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$, where $E_1, ..., E_n$ are mutually exclusive (i.e. $E_i \cap E_j = \emptyset$ for all $i \neq j$). Such events are called *mutually exclusive* or *disjoint*.

The rules of probability

Let A and B denote two events in a sample space Ω, and let $P(B) \neq 0$.

Definition (Marginal probability)

The probability of an event (A) occurring ($P(A)$).

Definition (Joint probability)

$P(A, B)$ is the probability of event A and event B occurring. Symmetry property holds, i.e. $P(A, B) = P(B, A)$.

Definition (Conditional probability)

$P(B|A)$ is the probability that B will occur given that A has occurred.

Rules

- Sum rule: $P(A) = \sum_B P(A, B)$
- Product rule: $P(A, B) = P(B|A)P(A)$.

Bayes' rule

Bayes' rule

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

Constituents:

- $P(A)$, the prior probability, is the probability of A before B is observed.
- $P(A|B)$, the posterior probability, is the probability of A given B, i.e., after B is observed.
- $P(B|A)$ is the probability of observing B given A. As a function of A with B fixed, this is the likelihood.

Union of non-disjoint events

Definition (Principle of inclusion-exclusion)

The probability of the union of n events is

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq n} P(E_{i_1} \cap \ldots \cap E_{i_k}),$$

where the second sum is over all subsets of k events.

Union of non-disjoint events

Example (winning craps or lose big)

Let us go to a casino and play craps. You throw two dices. If the outcome is 7 or 11, you win immediately. If it is 2, 3, or 12, you lose immediately. If the outcome is any other number, say x, you continue throwing until the outcome is x or 7. You win in the former, and lose in the latter case.

Suppose you like big numbers, and you are content with seeing any outcome equal to or larger than 10. You also like winning. What is the probability that you'll be satisfied in the first throw? Let A denote the event of throwing 7 or 11 (winning in the first throw), and let B be the event of outcome equal to or larger than 10. Then, $P(A) = \frac{8}{36}$, $P(B) = \frac{6}{36}$ and $P(A \cap B) = \frac{2}{36}$.

By the inclusion-exclusion principle, $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{3}$.

Random variable

Definition

A real-valued random variable is a **function** that associates a value to the outcome of a randomized experiment $X : \Omega \to \mathbb{R}$.

Example

- Whether a gene will mutate: a function from $\Omega = \{\texttt{mutation, no mutation}\}$ to $\{1, 0\}$.
- Number of mutations in a sequence of n experiments: function from $\Omega = \{\texttt{mutation, no mutation}\}^n \to \mathbb{N} \cup \{0\}$.

Discrete random variable

Probability mass function (pmf)

The probability mass function is the function from values to its probability, $P_X(x) = P(X = x)$ for $x \in \mathcal{X}$ (i.e., a countable subset of the reals) with properties:

- $P_X(x) \geq 0$ for every $x \in \mathcal{X}$,
- $\sum_{x \in \mathcal{X}} P_X(x) = 1$

Discrete random variable

Example

- Bernoulli distribution - distribution of a binary variable $x \in \{0, 1\}$; single parameter $\mu \in [0, 1]$ represents the probability of $x = 1$:

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}.$$

- Binomial distribution - probability of observing m occurrences of 1 in a set of N samples from a Bernoulli distribution:

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}.$$

- Recall the gene mutation example: Let X be the random variable such that $P_X(x = 1) = \mu$, and $P_X(x = 0) = 1 - \mu$. Then $X \sim$ Bern. If we conduct N experiments and let m denote the number of mutations, then the random variable has binomial distribution.

Probability density function (pdf)

- A continuous random variable can have uncountably infinite possible values.

Probability density function (pdf)

The probability density function of a continuous random variable X is an integrable function $p(x)$ satisfying the following:

1. The density is nonnegative: i.e., $p(x) \geq 0$ for any x,
2. Probabilities integrate to 1: i.e., $\int_{-\infty}^{\infty} p(x)dx = 1$,
3. The probability that x belongs to the interval $[a, b]$ is given by the integral of $p(x)$ over that interval: i.e.,

$$P(a \leq X \leq b) = \int_a^b p(x)dx.$$

Basic rules of probability

1. Analog of sum rule: $p(x) = \int p(x, y)dy$
2. Product rule: $p(x, y) = p(y|x)p(x)$.

Expectations and variances

Definition (Expectation (1^{st} moment, mean))

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} x P(X = x) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$

Definition (Variance (2^{nd} moment))

$$\mathbb{V}[X] = \begin{cases} \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 P(X = x) & \text{discrete} \\ \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 p(x) dx & \text{continuous} \end{cases}$$

Definition (Conditional expectation and Covariance)

$$\mathbb{E}[X|Y = y] = \sum_{x \in \mathcal{X}} x P(X = x|Y = y)$$

$$\text{cov}[x, y] = \mathbb{E}\Big[\big(x - \mathbb{E}[X]\big)\big(y - \mathbb{E}[Y]\big)\Big]$$

Normal (Gaussian) Distribution

Gaussian distribution

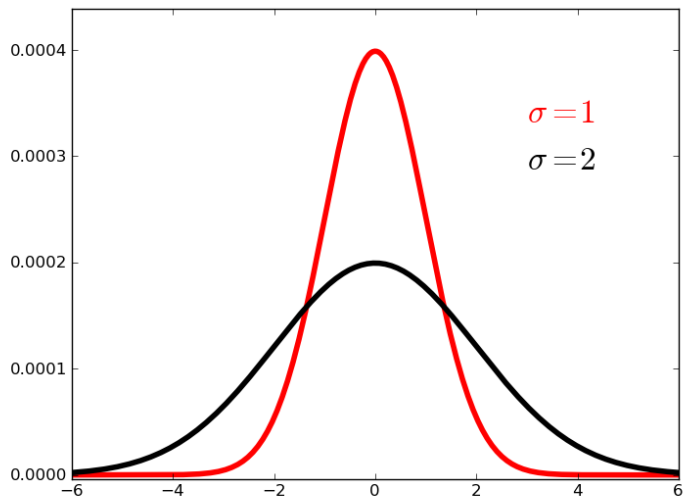
For $\mathbf{x} \in \mathbb{R}^d$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right),$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is the covariance matrix and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

- In the case of a single variable

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



Basic statistics

Parametric estimation model

A parametric estimation model consists of the following four elements:

1. A *parameter space*, which is a subset $\mathcal{X}$ of $\mathbb{R}^p$
2. A *parameter* $\mathbf{x}^\natural$, which is an element of the parameter space
3. A class of probability distributions $\mathcal{P}_\mathcal{X} := \{\mathbb{P}_\mathbf{x} : \mathbf{x} \in \mathcal{X}\}$, parametrized by $\mathbf{x} \in \mathcal{X}$
4. A *sample* $\mathbf{b}$, which follows the probability distribution $\mathbf{b} \sim \mathbb{P}_{\mathbf{x}^\natural} \in \mathcal{P}_\mathcal{X}$

Statistical estimation seeks to approximate the value of $\mathbf{x}^\natural$, given $\mathcal{X}$, $\mathcal{P}_\mathcal{X}$, and $\mathbf{b}$.

Definition (Estimator)

An estimator $\hat{\mathbf{x}}$ is a mapping that takes $\mathcal{X}$, $\mathcal{P}_\mathcal{X}$, and $\mathbf{b}$ as inputs, and outputs a value in $\mathbb{R}^p$.

- The output of an estimator depends on the sample, and hence, is random.
- The output of an estimator is not necessarily equal to $\mathbf{x}^\natural$.

Ordinary least-squares estimator

Ordinary least-squares estimator (OLS)

The ordinary least-squares estimator is given by

$$\hat{\mathbf{x}}_{\text{OLS}} \in \arg\min_{\mathbf{x}} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}.$$

Ordinary least-squares estimator: An intuitive model

Gaussian linear model

Let $\mathbf{x}^\natural \in \mathbb{R}^p$. Let $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w} \in \mathbb{R}^n$ for some matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$, where $\mathbf{w}$ is a Gaussian vector with zero mean and covariance matrix $\sigma^2 I$.

The probability density function $p_{\mathbf{x}}(\cdot)$ is given by

$$p_{\mathbf{x}}(\mathbf{b}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{b}-\mathbf{A}\mathbf{x}\right\|_2^2\right).$$

Therefore, the maximum likelihood (ML) estimator is defined as

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \left\{ -\log p_{\mathbf{x}}(\mathbf{b}) = -\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left\|\mathbf{b}-\mathbf{A}\mathbf{x}\right\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\},$$

which is equivalent to

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \left\{ \left\|\mathbf{b}-\mathbf{A}\mathbf{x}\right\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}.$$

OLS is the ML estimator for the Gaussian linear model.

Maximum-likelihood estimator

Recall the general setting.

Parametric estimation model

A parametric estimation model consists of four elements:

1. A *parameter space*, which is a subset $\mathcal{X}$ of $\mathbb{R}^p$,
2. A *parameter* $\mathbf{x}^\natural$, which is an element of the parameter space,
3. A class of probability distributions $\mathcal{P}_\mathcal{X} := \{\mathbb{P}_\mathbf{x} : \mathbf{x} \in \mathcal{X}\}$, parametrized by $\mathbf{x} \in \mathcal{X}$,
4. A *sample* $\mathbf{b}$, which follows the probability distribution $\mathbb{P}_{\mathbf{x}^\natural} \in \mathcal{P}_\mathcal{X}$.

Definition (Maximum-likelihood estimator)

The maximum-likelihood (ML) estimator is given by

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \{-\log p_\mathbf{x}(\mathbf{y})\},$$

where $p_\mathbf{x}(\cdot)$ denotes the probability density function or probability mass function of $\mathbb{P}_\mathbf{x}$, for $\mathbf{x} \in \mathcal{X}$.

Gene mutation

Gene mutation

Suppose the mutation probability is $P(\texttt{mutation}) = \mu$, and you want to estimate μ. Suppose you have observed m mutations in N experiments.

The probability mass function is given by the binomial distribution

$$p(\#\text{ mutations} = m|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}.$$

The maximum-likelihood estimator is

$$\mu_{\text{ML}} = \arg \min_{\mu \in [0,1]} -m \log \mu - (N-m) \log(1-\mu).$$

It is easy to see that $\mu_{\text{ML}} = \frac{m}{N}$.

Logistic regression

Logistic regression [1]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$. Let $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^p$ be given. The sample is given by $\mathbf{b} := (b_1, \dots, b_n) \in \{-1, 1\}^n$, where each b_i is a Bernoulli random variable satisfying

$$\mathbb{P}\left\{b_i = 1\right\} = 1 - \mathbb{P}\left\{b_i = -1\right\} = \left[1 + \exp\left(-\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle\right)\right]^{-1},$$

and $b_1, \dots, b_n$ are independent.

The probability mass function $p_{\mathbf{x}}(\cdot)$ is given by

$$p_{\mathbf{x}}(\mathbf{b}) = \Pi_{i=1}^n \left[1 + \exp\left(-b_i \left\langle \mathbf{a}_i, \mathbf{x} \right\rangle\right)\right]^{-1}.$$

Therefore, the maximum-likelihood estimator is defined as

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}} \left\{ -\log p_{\mathbf{x}}(\mathbf{b}) = \sum_{i=1}^n \log\left[1 + \exp\left(-b_i \left\langle \mathbf{a}_i, \mathbf{x} \right\rangle\right)\right] : \mathbf{x} \in \mathbb{R}^p \right\}.$$

- $\hat{\mathbf{x}}_{\mathsf{ML}}$ defines a *linear classifier*. For any new $\mathbf{a}_i$, $i \geq n+1$, we can predict the corresponding b_i by predicting $b_i = 1$ if $\langle \mathbf{a}_i, \hat{\mathbf{x}}_{\mathsf{ML}} \rangle \geq 0$, and $b_i = -1$ otherwise.

Graphical model learning (self-study)

Graphical model selection

Let $\mathbf{\Theta}^\natural \in \mathbb{R}^{p\times p}$ be a positive-definite matrix. The sample is given by $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, which are i.i.d. random vectors with zero mean and covariance matrix $\mathbf{\Theta}^{\natural\,-1}$.

When $\mathbf{x}_i$'s are Gaussian random vectors, the probability density function $p_{\mathbf{\Theta}}(\cdot)$ is given by

$$p_{\mathbf{\Theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \Pi_{i=1}^n \left[(2\pi)^{-p/2} \det\left(\mathbf{\Theta}^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}_i^T \mathbf{\Theta} \mathbf{x}_i\right)\right]$$

$$= (2\pi)^{-np/2} \det(\mathbf{\Theta})^{n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^n \left(\mathbf{x}_i^T \mathbf{\Theta} \mathbf{x}_i\right)\right]$$

Therefore, the ML estimator is defined as

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{\Theta}} \left\{ -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det\left(\mathbf{\Theta}\right) + \frac{n}{2}\mathrm{Tr}\left(\widehat{\mathbf{\Sigma}}\mathbf{\Theta}\right) : \mathbf{\Theta} \in \mathbb{S}^p_{++} \right\}.$$

Checking the fidelity

Given an estimator $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \{F(\mathbf{x})\}$, we need to address two key questions:

1. Is the formulation reasonable?
2. What is the role of the data size?

Standard approach to checking the fidelity

Standard approach

1. Specify a performance criterion $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural)$ that should be small if $\hat{\mathbf{x}} = \mathbf{x}^\natural$.
2. Show that $\mathcal{L}$ is actually *small in some sense* when *some condition* is satisfied.

Example

Take the ℓ_2-error $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural) := \left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2^2$ as an example. Then we may verify the fidelity via one of the following ways, where ε denotes a small enough number:

1. $\mathbb{E}\left[\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural))\right] \leq \varepsilon$ (expected error),
2. $\mathbb{P}\left(\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural) \geq \epsilon\right) \leq \delta$ for some δ depending on ϵ (consistency),
3. $\sqrt{n}(\hat{\mathbf{x}} - \mathbf{x}^\natural)$ converges in distribution to $\mathcal{N}(0, \mathbf{I})$ (asymptotic normality),
4. $\sqrt{n}(\hat{\mathbf{x}} - \mathbf{x}^\natural)$ converges in distribution to $\mathcal{N}(0, \mathbf{I})$ in a local neighborhood (local asymptotic normality).

if *some condition* is satisfied. Such conditions typically revolve around the data size.

Standard approach: Expected error

Gaussian linear model

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and let $\mathbf{A} \in \mathbb{R}^{n \times p}$. The samples are given by $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is a sample of a Gaussian random vector $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

What is the performance of the ML estimator

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \right\}?$$

Theorem (Performance of the LS estimator [2])

If $\mathbf{A}$ is a matrix of independent and identically distributed (i.i.d.) standard Gaussian distributed entries, and if $n > p + 1$, then

$$\mathbb{E}\left[\left\| \hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural \right\|_2^2 \right] = \frac{p}{n - p - 1} \sigma^2 \to 0 \text{ as } \frac{n}{p} \to \infty.$$

⋆Approach 2: Consistency

Graphical model learning (self-study)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be samples of a sub-Gaussian random vector with zero mean and some unknown positive-definite covariance matrix $\mathbf{\Sigma}^\natural \in \mathbb{R}^{p \times p}$. (Sub-Gaussian random variables will be defined in recitation.)

What is the performance of the M-estimator $\widehat{\mathbf{\Sigma}} := \widehat{\mathbf{\Theta}}^{-1}$, where

$$\widehat{\mathbf{\Theta}}_{\mathsf{ML}} \in \arg \min_{\mathbf{\Theta} \in \mathbb{S}^p_{++}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[-\log \det \left(\mathbf{\Theta} \right) + \mathbf{x}_i^T \mathbf{\Theta} \mathbf{x}_i \right] \right\} ?$$

- If $\mathbf{y} = f(\mathbf{x})$, then $\hat{\mathbf{y}}_{\mathsf{ML}} = f(\hat{\mathbf{x}}_{\mathsf{ML}})$. This is called the *functional invariance* property of ML estimators.

Theorem (Performance of the ML estimator [3])

Suppose that the diagonal elements of $\mathbf{\Sigma}^\natural$ *are bounded above by* $\kappa > 0$*, and each* $X_i / \sqrt{\left(\Sigma^\natural\right)_{i,i}}$ *is sub-Gaussian with parameter* c*. Then*

$$\mathbb{P}\left(\left\{ \left| \left(\widehat{\mathbf{\Sigma}}_{ML}\right)_{i,j} - \left(\mathbf{\Sigma}^\natural\right)_{i,j} \right| > t \right\} \right) \leq 4 \exp \left[-\frac{n t^2}{128 \left(1 + 4c^2\right) \kappa^2} \right] \to 0 \text{ as } n \to \infty$$

for all $t \in \left(0, 8\kappa \left(1 + 4c^2\right)\right)$.

Basic statistical learning

Statistical Learning Model [4]

A statistical learning model consists of the following three elements.

1. A sample of i.i.d. random variables $(\mathbf{a}_i, b_i) \in \mathcal{A} \times \mathcal{B}$, $i = 1, \ldots, n$, following an *unknown* probability distribution $\mathbb{P}$.
2. A class (set) $\mathcal{F}$ of functions $f : \mathcal{A} \to \mathcal{B}$.
3. A loss function $L : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$.

Definition

Let $(\mathbf{a}, b)$ follow the probability distribution $\mathbb{P}$ and be independent of $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_n, b_n)$. Then, the risk corresponding to any $f \in \mathcal{F}$ is its expected loss:

$$R(f) := \mathbb{E}_{(\mathbf{a},b)} \left[L(f(\mathbf{a}), b) \right].$$

Statistical learning seeks to find a $f^\star \in \mathcal{F}$ that minimizes the risk, i.e., it solves

$$f^\star \in \arg\min_f \{ R(f) : f \in \mathcal{F} \}.$$

- Since $\mathbb{P}$ is unknown, the optimization problem above is intractable.

Empirical risk minimization (ERM)

By the law of large numbers, we can expect that for each $f \in \mathcal{F}$,

$$R(f) := \mathbb{E}\left[L(\mathbf{a}, b)\right] \approx \frac{1}{n} \sum_{i=1}^{n} L(f(\mathbf{a}_i), b_i)$$

when n is large enough, with high probability.

Empirical risk minimization (ERM) [4]

We approximate $f^\star$ by minimizing the *empirical average of the loss* instead of the risk. That is, we consider the optimization problem

$$\hat{f}_n \in \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} L(f(\mathbf{a}_i), b_i) : f \in \mathcal{F} \right\}.$$

Least squares revisited

Recall that the LS estimator is given by

$$\hat{\mathbf{x}}_{\mathsf{LS}} \in \arg\min \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\} = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} (b_i - \langle \mathbf{a}_i, \mathbf{x} \rangle)^2 : \mathbf{x} \in \mathbb{R}^p \right\},$$

where we define $\mathbf{b} := (b_1, \ldots, b_n)$ and $\mathbf{a}_i$ to be the i-th row of $\mathbf{A}$.

A statistical learning view of least squares

This corresponds to a statistical learning model, for which

- the sample is given by $(\mathbf{a}_i, b_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$,
- the function class $\mathcal{F}$ is given by $\mathcal{F} := \{f_{\mathbf{x}}(\cdot) := \langle \cdot, \mathbf{x} \rangle : \mathbf{x} \in \mathbb{R}^p\}$, and
- the loss function is given by $L(f_{\mathbf{x}}(\mathbf{a}), b) := (b - f_{\mathbf{x}}(\mathbf{a}))^2$.

The corresponding ERM solution is

$$\hat{f}_n(\cdot) := \langle \cdot, \hat{\mathbf{x}}_{\mathsf{LS}} \rangle .$$

- Thus the LS estimator also seeks to, given $\mathbf{a}$, minimize the error of predicting the corresponding b by a linear function in terms of the squared error.

Practical Issues

Given an estimator $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \{F(\mathbf{x})\}$ of $\mathbf{x}^\natural$, we discussed two key questions:

1. Is the formulation reasonable?
2. What is the role of the data size?

Consider the estimation error in the ℓ_2-norm: $\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2$.

- Is $\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2$ enough to evaluate the performance of the estimator $\hat{\mathbf{x}}$?

Practical Issues

No, because in general we can only *numerically approximate* the solution of

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x})\}.$$

Implementation

How do we *numerically approximate* $\hat{\mathbf{x}}$?

Practical performance

Denote the numerical approximation by $\mathbf{x}_\epsilon^\star$. The practical performance is governed by

$$\left\|\mathbf{x}_\epsilon^\star - \mathbf{x}^\natural\right\|_2 \leq \underbrace{\|\mathbf{x}_\epsilon^\star - \hat{\mathbf{x}}\|_2}_{\text{approximation error}} + \underbrace{\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}}.$$

How do we evaluate $\|\mathbf{x}_\epsilon^\star - \hat{\mathbf{x}}\|_2$?

- Recall that an ϵ-approximation solution is any point $\mathbf{x}_\epsilon^\star$ such that

$$F(\mathbf{x}_\epsilon^\star) - F(\hat{\mathbf{x}}) \leq \varepsilon.$$

⋆Time-data trade-off [5]

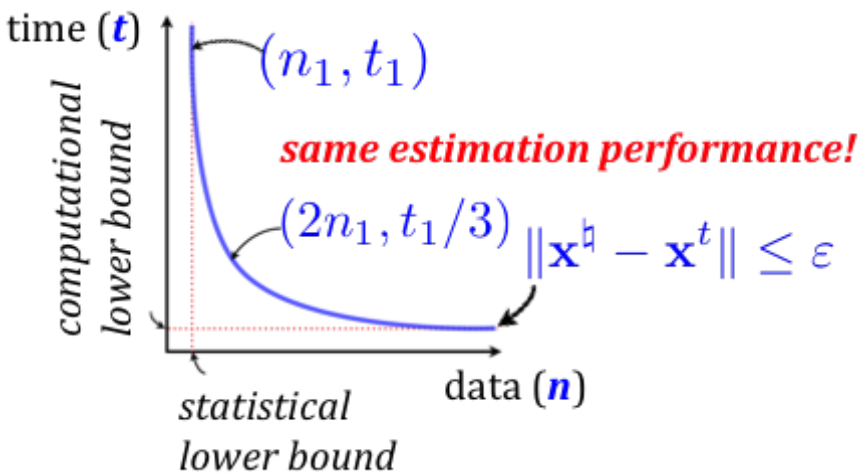
An alternative view: Joint study of approximation/statistical errors

$$\left\|\mathbf{x}_\epsilon^\star - \mathbf{x}^\natural\right\|_2 \leq \underbrace{\|\mathbf{x}_\epsilon^\star - \hat{\mathbf{x}}\|_2}_{\text{approximation error}} + \underbrace{\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}} .$$

How do we evaluate $\|\mathbf{x}_\epsilon^\star - \hat{\mathbf{x}}\|_2 + \left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2$?

We may fix a precision ε, and consider the approximation and statistical error jointly:

$$\left\|\mathbf{x}_\epsilon^\star - \mathbf{x}^\natural\right\|_2 \leq \underbrace{\|\mathbf{x}_\epsilon^\star - \hat{\mathbf{x}}\|_2}_{\text{approximation error}} + \underbrace{\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}} \leq \varepsilon.$$



Practical Issues

How do we *numerically approximate* $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x})\}$ for a given F?

General idea of an optimization algorithm

Guess a solution, and then *refine* it based on *oracle information*.

Repeat the procedure until the result is *good enough*.

How do we evaluate the approximation error $\|\mathbf{x}_\epsilon^\star - \hat{\mathbf{x}}\|_2$?

General concept about the approximation error

It depends on the *characteristics* of the function F and the chosen numerical *optimization algorithm*.

Need for convex analysis

General idea of an optimization algorithm

Guess a solution, and then *refine* it based on *oracle information*.

Repeat the procedure until the result is *good enough*.

General concept about the approximation error

It depends on the *characteristics* of the function F and the chosen numerical *optimization algorithm*.

Role of convexity

Convexity provides a key optimization framework in obtaining numerical approximations at theoretically well-understood computational costs.

To precisely understand these ideas, we need to understand basics of *convex analysis*.

References I

[1] M. I. Jordan *et al.*, "Why the logistic function? a tutorial discussion on probabilities and neural networks," 1995.

[2] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized LASSO: A precise analysis," 2013, arXiv:1311.0830v2 [cs.IT].

[3] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Stat.*, vol. 5, pp. 935–980, 2011.

[4] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Inf. Theory*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[5] J. J. Bruer, J. A. Tropp, V. Cevher, and S. R. Becker, "Designing statistical estimators that balance sample size, risk, and computational cost," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 612–624, 2015.