S&DS677: Topics in High-Dimensional Statistics and Information Theory

Spring 2024

Administrivia

- Schedule: Tuesday 4–550pm KT 207
- Instructor: Yihong Wu yihong.wu@yale.edu
  - Office hours: by appointment
- Website:
  http://www.stat.yale.edu/~yw562/teaching/SDS677/index.html
  or just google S&DS677

Administrivia

1. Course prerequisites:

Administrivia

1. Course prerequisites:
   - Maturity with probability theory

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra
   - Prior knowledge on Information Theory (e.g. SDS 364) is NOT required

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra
   - Prior knowledge on Information Theory (e.g. SDS 364) is NOT required
2. Participation (30%):

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra
   - Prior knowledge on Information Theory (e.g. SDS 364) is NOT required
2. Participation (30%):
   - Classroom participation is highly encouraged

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra
   - Prior knowledge on Information Theory (e.g. SDS 364) is NOT required
2. Participation (30%):
   - Classroom participation is highly encouraged
   - Critiques on lecture notes/maybe a few scribes towards the end

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra
   - Prior knowledge on Information Theory (e.g. SDS 364) is NOT required
2. Participation (30%):
   - Classroom participation is highly encouraged
   - Critiques on lecture notes/maybe a few scribes towards the end
3. Homeworks (30%): 2-3 problem sets

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra
   - Prior knowledge on Information Theory (e.g. SDS 364) is NOT required
2. Participation (30%):
   - Classroom participation is highly encouraged
   - Critiques on lecture notes/maybe a few scribes towards the end
3. Homeworks (30%): 2-3 problem sets
4. Final project (40%)

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra
   - Prior knowledge on Information Theory (e.g. SDS 364) is NOT required
2. Participation (30%):
   - Classroom participation is highly encouraged
   - Critiques on lecture notes/maybe a few scribes towards the end
3. Homeworks (30%): 2-3 problem sets
4. Final project (40%)
   - either presenting paper(s) or a standalone research project.

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra
   - Prior knowledge on Information Theory (e.g. SDS 364) is NOT required
2. Participation (30%):
   - Classroom participation is highly encouraged
   - Critiques on lecture notes/maybe a few scribes towards the end
3. Homeworks (30%): 2-3 problem sets
4. Final project (40%)
   - either presenting paper(s) or a standalone research project.
   - topics announced around week 6

Administrivia

1. Course prerequisites:
   - Maturity with probability theory
   - Some linear algebra
   - Prior knowledge on Information Theory (e.g. SDS 364) is NOT required
2. Participation (30%):
   - Classroom participation is highly encouraged
   - Critiques on lecture notes/maybe a few scribes towards the end
3. Homeworks (30%): 2-3 problem sets
4. Final project (40%)
   - either presenting paper(s) or a standalone research project.
   - topics announced around week 6
5. Materials: Lecture notes and additional reading materials will be posted online.

What this course is about?

What this course is about?

Information-theoretic & related methods in high-dimensional statistics

What this course is about?

Information-theoretic & related methods in high-dimensional statistics

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, confidence statements)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X_1, \dots, X_n}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, confidence statements)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X_1, \ldots, X_n}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, confidence statements)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X_1, \ldots, X_n}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:
  - Q1 Characterize statistical optimum: What is possible/impossible?

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, confidence statements)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X_1, \ldots, X_n}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:
  - Q1 Characterize statistical optimum: What is possible/impossible?
  - Q2 How many samples are necessary and sufficient to achieve a prescribed goal?

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, confidence statements)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X_1, \ldots, X_n}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:
  - Q1 Characterize statistical optimum: What is possible/impossible?
  - Q2 How many samples are necessary and sufficient to achieve a prescribed goal?
  - Q3 Can statistical limits be attained comptutationally efficiently, e.g., in $\mathrm{poly}(n,p)$-time? If yes, how? If not, why?

High Dimensionality of Contemporary Datasets

Fields	Data
Biomedical Research	microarray, ECG, fMRI, ...
Signal Processing	array sensor data, face recognition, hyper-spectral data, ...
Finance	asset returns, ...
$\vdots$	$\vdots$

- Growth of data outpaced by increasing number of features
- A common feature: large d, but just comparable or smaller n

$$\theta \in \mathbb{R}^d \mapsto X_1, \ldots, X_n$$

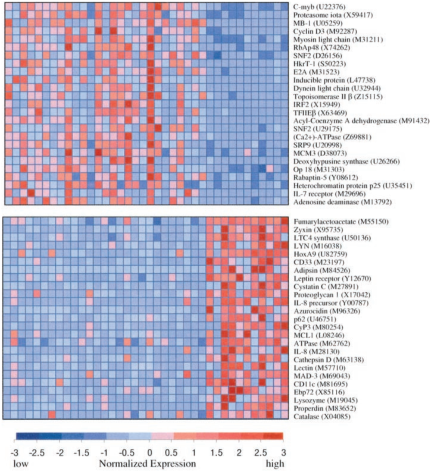
- low-dimensional structure
  - Intrinsic: θ lies in a low-dimensional subset
  - Extrinsic: θ has no structure but we only estimate low-dimensional functional of θ

Classical topics

Example 1: high-dimensional linear regression

Microarray data:

- Leukaemia dataset [Golub et al. '99]: $d = 7129$ genes and $n = 72$ samples
- Typically $d \gg n$
- Interpretability (gene selection)



Ref: [Golub et al. '99, Zou-Hastie '05]

Example 1: high-dimensional linear regression

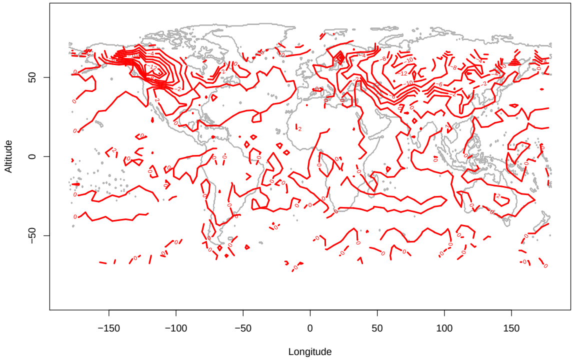
Statistical model

$$y = X\beta + \text{noise}$$

- observation: $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times d}$
- parameter: $\beta \in \mathbb{R}^d$
- goal: estimate β or predict $X\beta$
- assumption: β is sparse

Example 2: Covariance matrix estimation & PCA

Climate Data



One observation: January average temperature in 1969 [$d = 2592$, $n = 157$]

Ref: Bickel & Levina (08)

Example 2: Covariance matrix estimation & PCA

Statistical model

- observation: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0, \Sigma) \in \mathbb{R}^d$
- parameter: $\Sigma = \mathbb{E}[XX'] \in \mathbb{R}^{d \times d}$
- goal: estimate Σ or its principle component (PCA)
- assumption: Σ is sparse/smooth(entrywise decay)/low-rank

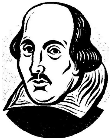
Problems of combinatorial nature

Example 3: How many words did Shakespeare know?

- Linguistics

Estimating the number of unseen species: How many words did Shakespeare know?

By BRADLEY EFRON AND RONALD THISTED
Department of Statistics, Stanford University, California



- Ecology

THE RELATION BETWEEN THE NUMBER OF SPECIES AND THE NUMBER OF INDIVIDUALS IN A RANDOM SAMPLE OF AN ANIMAL POPULATION

By R. A. FISHER (*Galton Laboratory*), A. STEVEN CORBET (*British Museum, Natural History*) AND C. B. WILLIAMS (*Rothamsted Experimental Station*)

Example 3: How many words did Shakespeare know?

```
ACT I
SCENE I. Elsinore. A platform before the castle.

    FRANCISCO at his post. Enter to him BERNARDO

BERNARDO

    Who's there?


.
.
.
.
.
.
.
.
.

PRINCE FORTINBRAS

    Let four captains
    Bear Hamlet, like a soldier, to the stage;
    For he was likely, had he been put on,
    To have proved most royally: and, for his passage,
    The soldiers' music and the rites of war
    Speak loudly for him.
    Take up the bodies: such a sight as this
    Becomes the field, but here shows much amiss.
    Go, bid the soldiers shoot.

    A dead march. Exeunt, bearing off the dead bodies;
    after which a peal of ordnance is shot off
```

Hamlet experiment

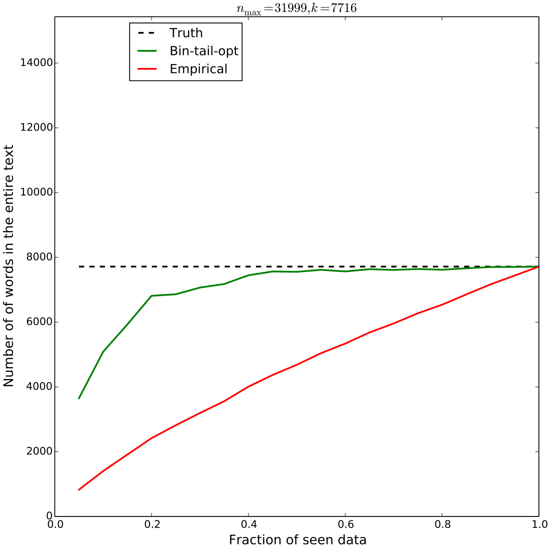
1. Starting from Act I, read a small fraction of the text
2. Stop and estimate the number of distinct words in entire Hamlet

Example 3: How many words did Shakespeare know?

Statistical model: Distinct element problem

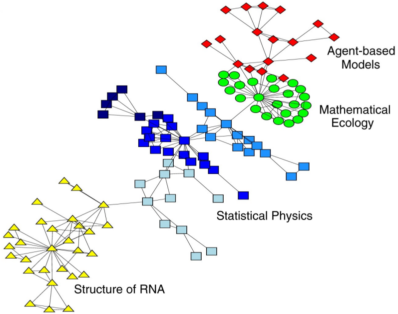
- observation: $X_1, \ldots, X_n$ sampled without replacements from an urn of k colored balls
- parameter: composition of the urn (number of red, blue, etc.)
- goal: number of distinct colors
- assumption: NONE!
- Method: Estimator built from convex/LP duality

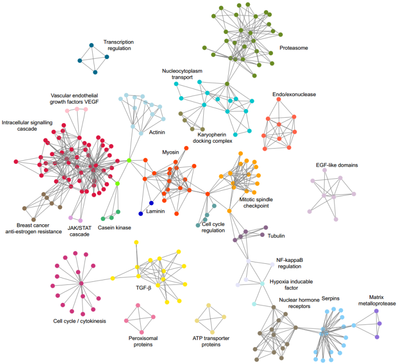
Example 3: How many words did Shakespeare know?



Example 4: Community detection in networks

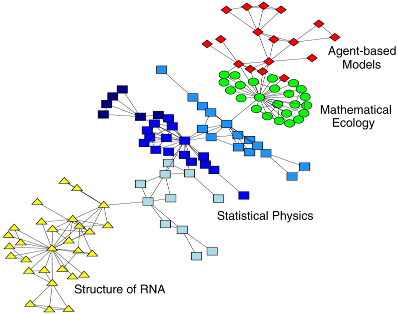
- Networks with community structures arise in many applications

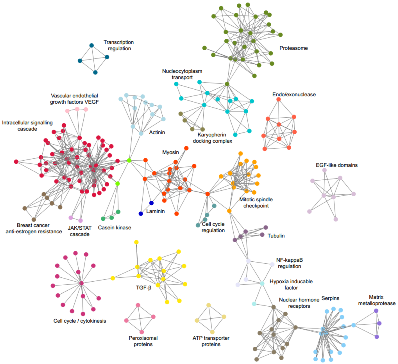




Example 4: Community detection in networks

- Networks with community structures arise in many applications

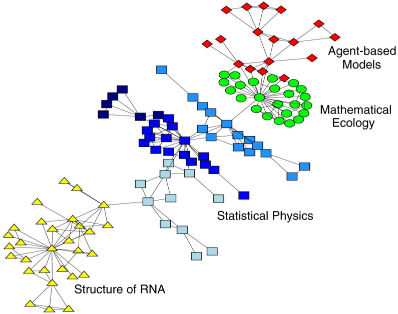


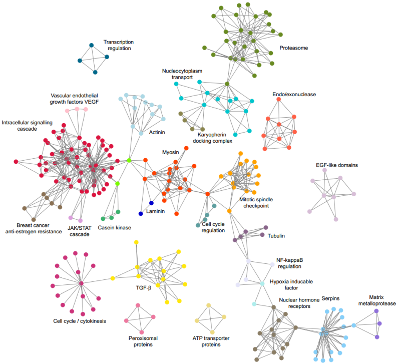


- Task: Discover underlying communities based on the network topology

Example 4: Community detection in networks

- Networks with community structures arise in many applications

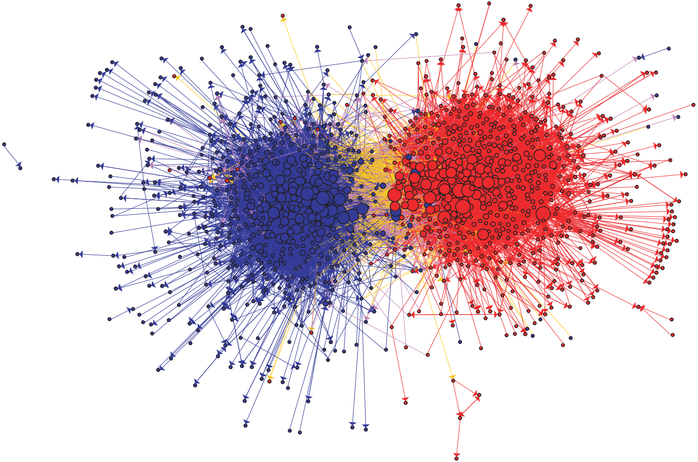




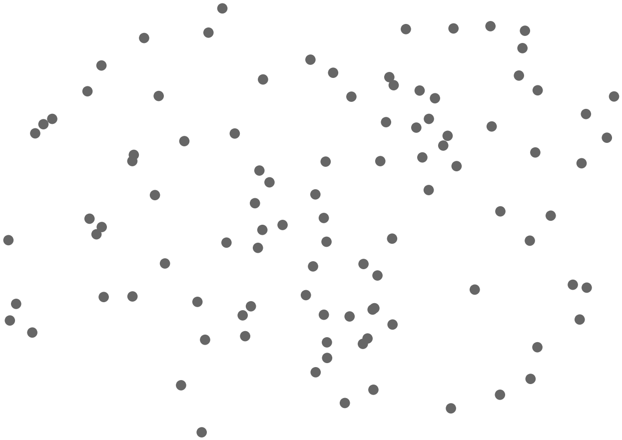
- Task: Discover underlying communities based on the network topology
- Applications: Friend or movie recommendation in online social networks

Political blogosphere

...in the 2004 U.S. election [Adamic-Glance '05]

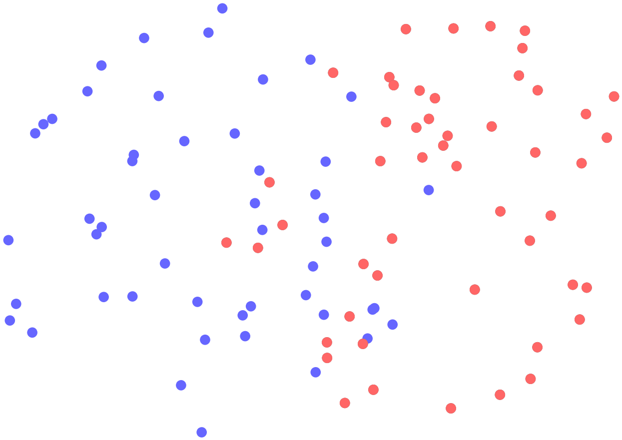


Stochastic block model – graph view



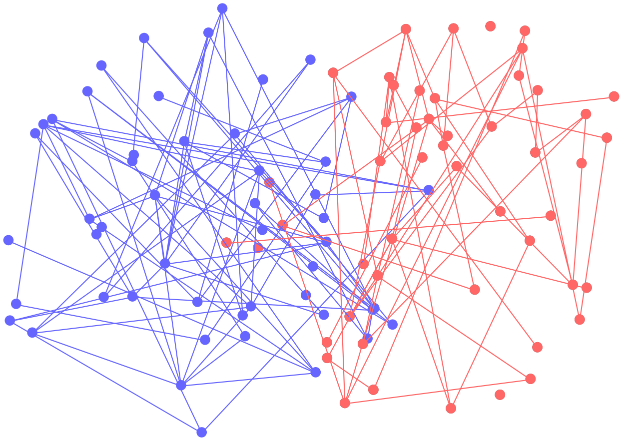
Stochastic block model – graph view

1. n nodes are randomly partitioned into 2 equal-sized communities



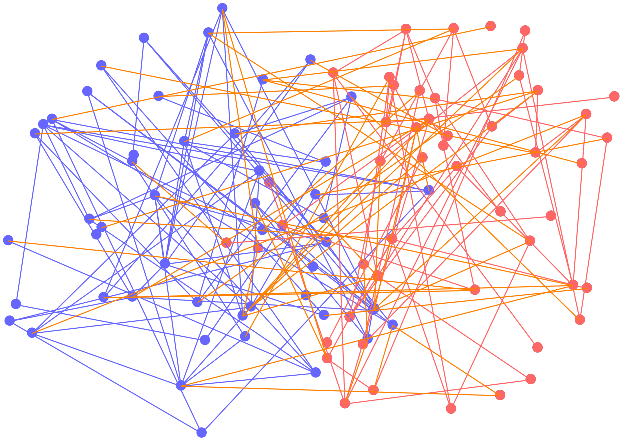
Stochastic block model – graph view

1. n nodes are randomly partitioned into 2 equal-sized communities
2. For every pair of nodes in same community, add an edge w.p. p



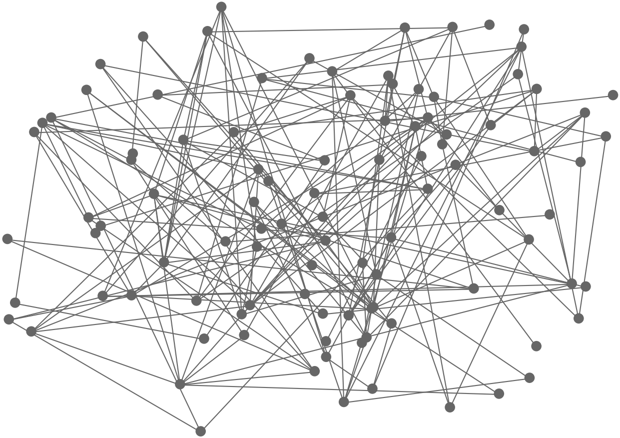
Stochastic block model – graph view

1. n nodes are randomly partitioned into 2 equal-sized communities
2. For every pair of nodes in same community, add an edge w.p. p
3. For every pair of nodes in diff. community, add an edge w.p. q

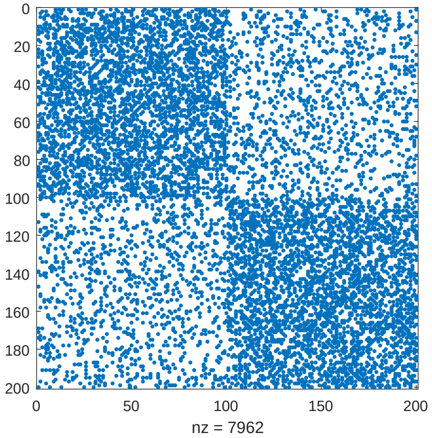


Stochastic block model – graph view

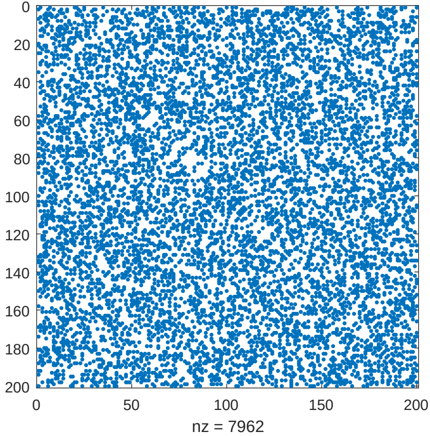
1. n nodes are randomly partitioned into 2 equal-sized communities
2. For every pair of nodes in same community, add an edge w.p. p
3. For every pair of nodes in diff. community, add an edge w.p. q



Stochastic block model – adjacency matrix view



Stochastic block model – adjacency matrix view



Example 4: Community detection

Statistical model: Stochastic block model SBM(n, p, q)

- observation: a single graph G
- parameter: partition of two communities (subsets of $[n]$)
- goal: locate the community (under various criteria)
- assumption: low-rankness of $\mathbb{E}$[adjancency matrix]

Example 5: spiked Wigner model

Noisy observation of rank-one matrix:

$$Y = \lambda x x^\top + Z,$$

where

- signal: x uniform on the hypercube $\{\pm \frac{1}{\sqrt{n}}\}^n$
- noise: Z iid $N(0, \frac{1}{n})$
- goal: recover x better than chance
  - Find unit vector $\hat{x} = \hat{x}(Y)$, s.t. $\mathbb{E}|\langle \hat{x}, x \rangle| = \Omega(1)$

Example 5: spiked Wigner model

Noisy observation of rank-one matrix:

$$Y = \lambda x x^\top + Z,$$

where

- signal: x uniform on the hypercube $\{\pm \frac{1}{\sqrt{n}}\}^n$
- noise: Z iid $N(0, \frac{1}{n})$
- goal: recover x better than chance
  - Find unit vector $\hat{x} = \hat{x}(Y)$, s.t. $\mathbb{E}|\langle \hat{x}, x \rangle| = \Omega(1)$
- Random matrix theory: PCA works iff $\lambda > 1$ [Baik-Ben Arous-Peche '04]

Example 5: spiked Wigner model

Noisy observation of rank-one matrix:

$$Y = \lambda x x^\top + Z,$$

where

- signal: x uniform on the hypercube $\{\pm\frac{1}{\sqrt{n}}\}^n$
- noise: Z iid $N(0, \frac{1}{n})$
- goal: recover x better than chance
  - Find unit vector $\hat{x} = \hat{x}(Y)$, s.t. $\mathbb{E}|\langle \hat{x}, x\rangle| = \Omega(1)$
- Random matrix theory: PCA works iff $\lambda > 1$ [Baik-Ben Arous-Peche '04]
- We will show $\lambda > 1$ is needed by any algo (information-percolation method)

What is information theory

Information theory: theory of fundamental limits

I. Information measures: How to measure randomness, dependency, dissimilarity (entropy, mutual information, divergence...)

II. Coding theorems: Operational problems (data compression, data transmission, etc)

$$\text{information measures} \; \underset{\text{operational meaning}}{\overset{\text{coding theorems}}{\rightleftharpoons}} \; \text{fundamental limits}$$

What is information theory

Information theory: theory of fundamental limits

I. Information measures: How to measure randomness, dependency, dissimilarity (entropy, mutual information, divergence...)

II. Coding theorems: Operational problems (data compression, data transmission, etc)

$$\text{information measures} \underset{\text{operational meaning}}{\overset{\text{coding theorems}}{\rightleftharpoons}} \text{fundamental limits}$$

Information-theoretic methods

- Negative results (converse, impossibility results, lower bound):
  - Conceptually: quantify "information" and "dissimilarity"
    - two distributions too "close" $\Rightarrow$ impossible to distinguish
    - $I(\text{observation}; \text{parameter})$ too "small" $\Rightarrow$ impossible to estimate
    - dimension/entropy too "high" $\Rightarrow$ need large sample size

Information-theoretic methods

- Negative results (converse, impossibility results, lower bound):
  - Conceptually: quantify "information" and "dissimilarity"
    - two distributions too "close" $\Rightarrow$ impossible to distinguish
    - $I(\text{observation}; \text{parameter})$ too "small" $\Rightarrow$ impossible to estimate
    - dimension/entropy too "high" $\Rightarrow$ need large sample size
  - More advanced techniques:
    - area theorem
    - strong data processing inequality and information-percolation method (Broadcasting on trees, spiked Wigner model...)
    - (truncated) second moment method

Information-theoretic methods

- Negative results (converse, impossibility results, lower bound):
  - Conceptually: quantify "information" and "dissimilarity"
    - two distributions too "close" $\Rightarrow$ impossible to distinguish
    - $I(\text{observation}; \text{parameter})$ too "small" $\Rightarrow$ impossible to estimate
    - dimension/entropy too "high" $\Rightarrow$ need large sample size
  - More advanced techniques:
    - area theorem
    - strong data processing inequality and information-percolation method (Broadcasting on trees, spiked Wigner model...)
    - (truncated) second moment method

- Positive results (achievability, constructive results, upper bound):
  - maximal likelihood estimate
  - entropy method (estimators based on pairwise comparison)
  - duality method
  - aggregation
  - efficient procedures/algorithms

What's new in this iteration

- Applications of variational representation
  - PAC Bayes method
  - Application in probability: concentration of sample covariance, small ball
  - Application in ML: variational autoencoder
- Functional estimation and composite hypothesis testing
  - Sparse detection (Ingster-Donoho-Jin)
  - Uniformity testing
- Risk bound based on duality
  - Dualizing two-point lower bound
  - Compund decision and large alphabet problems
- Advanced^2 topics
  - Universal compression and prediction
  - Area theorem (I-MMSE identity) and statistical lower bound
  - Aggregation and exponential weighting (Leung-Barron)