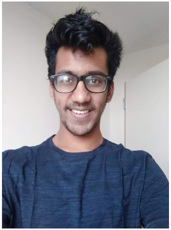
EE378C: Information-theoretic Lower Bounds in Data Science


Yanjun Han
Instructor


Ayfer Özgür
Instructor


Tsachy Weissman
Instructor


Jay Mardia
TA

March 29, 2021

What is this course about: lower bounds

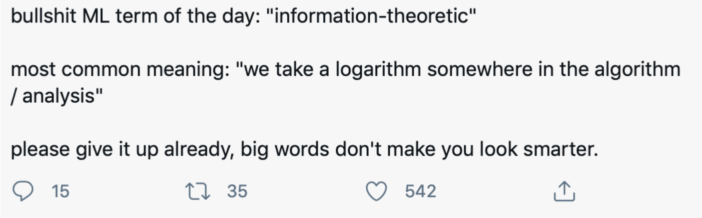
"Grant me the serenity to accept the things I cannot change, courage to change the things I can, and wisdom to know the difference."

—— Reinhold Niebuhr (1892–1971)

Why lower bounds?

- certify the fundamental limits and optimality of procedures
- spot the limitation of the current framework
- understand the tradeoff between real-world resources
- duality: shed light on upper bounds

What is this course about: information-theoretic



- fundamental limit that cannot be breached by any algorithm
- understanding data and problem, not a specific approach
- close to, but not entirely the same as, statistical efficiency

What is this course about: data science

- lots of interdisciplinary problems
- strong ideas from different communities
- however, they typically speak different languages...

> **Not sure what I find the most appealing in TCS, but if I had to pick one thing the concept of reduction would be a strong contender.**
>
> Not sure if other fields have an analogous concept that's so central to the way of thinking. (pointed out that in #Statistics the Central Limit Theorem could be seen as some sort of reduction wrt the Gaussian case, but that seems a bit of a stretch to me.)

The nature of the course

Both science and art:

- science: an extensive set of tools and ideas
- art: what tool to apply, and how to apply

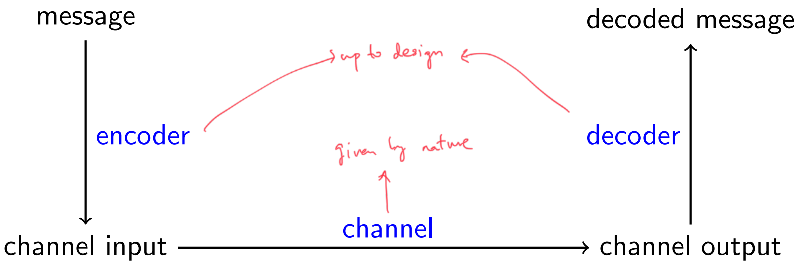
Both breadth and depth:

- breadth: a unified exposition of seemingly disparate ideas
- depth: first principle, deep variants, recent advances

Course style:

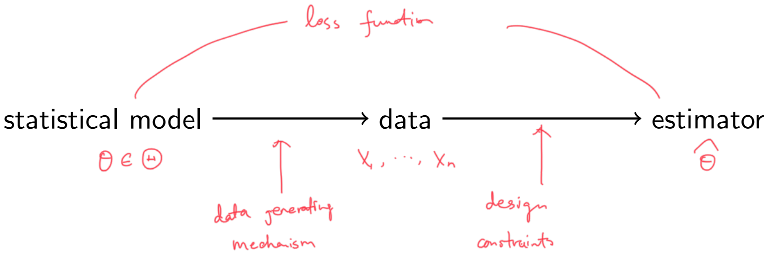
- focus on a diverse set of tools and examples
- involve mathematical proofs
- reproduce a paper in just 15mins

Example I: channel coding



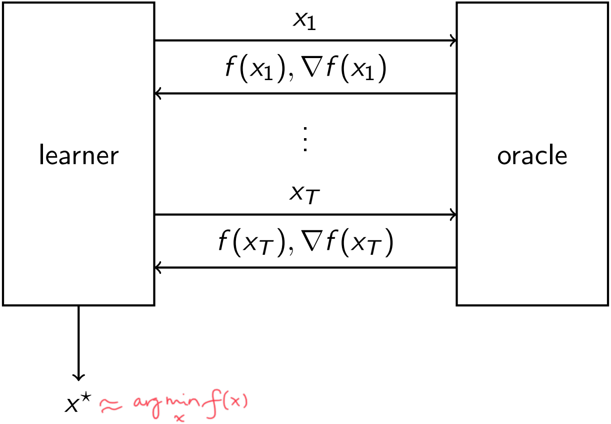
Q: jointly design the encoder and decoder to achieve the optimal tradeoff between the number of channel usage and probability of error

Example II: statistical efficiency



Q: find the best estimator (and possibly the data generating mechanism) which leads to the smallest expected loss

Example III: oracle-based optimization



Q: find the information-based complexity of optimization

Lecture plan

- Lecture 1: Introduction and puzzles
- Lecture 2: Importance of problem structure
- Lecture 3-5: The general idea of reduction
  - Lecture 3: deficiency, Le Cam's distance, asymptotic equivalence
  - Lecture 4: limits of experiments, Hájek-Le Cam classical asymptotics
  - Lecture 5: statistical-computational tradeoff (Jay)
- Lecture 6-12: The general idea of testing
  - Lecture 6: f-divergence and joint range
  - Lecture 7: Le Cam's two-point method
  - Lecture 8: one point vs. a mixture, Ingster-Suslina method
  - Lecture 9: two mixtures, orthogonal polynomials, moment matching
  - Lecture 10: testing multiple hypotheses, Fano and Assouad
  - Lecture 11: applications of Fano and Assouad in various domains
  - Lecture 12: global Fano, packing and covering

Lecture plan (cont'd)

- Lecture 13-18: special topics (tentative schedule)
  - geometric arguments in information theory (Tsachy)
  - converse results via optimal transport (Ayfer)
  - compression-based arguments
  - communication/privacy constrained estimation
  - min-max vs. max-min formulations
  - penalty of adaptation
- Final week: project presentations (time TBD)

Course logistics

- Instructors: (OHs start next week)
  - Yanjun Han (OH: Wed 11:30am – 1:00pm)
  - Ayfer Özgür (OH: TBD)
  - Tsachy Weissman (OH: TBD)
- TA: Jay Mardia (OH: TBD)
- Time: Mon and Wed, 10:00 – 11:20am
- Website: `web.stanford.edu/class/ee378c`
- Piazza: `piazza.com/stanford/spring2021/ee378c/home`
- Gradescope: entry code 3YD8J7
- Grading: 10% scribe + 60% homework + 30% project
  - C/NC students: only project required

Scribe

- half of the materials already covered in https://theinformaticists.com/category/blog/online-lectures/
- only need to scribe the remaining materials
- sign-up sheet: https://bit.ly/31quUIb
- group size: 1-3 students per lecture
- template available on course website
- email to us within 72 hours after the lecture

Homework

- around 3 homework sets
- mainly filling in the omitted proofs in class, and working out generalizations of examples covered in class
- HW1 will be released next week, and due two weeks after

Project

Default option: literature review

- choose one paper from the reading list on website, write a literature review, and give a presentation in the last week
- requirement: present main ideas and technical details of lower bound
- group size: 1-2 students
- ddl of choosing paper: end of Week 6 (May 9, 2021)
- email us your choice, first come first serve

Another option: propose your own research

- find your own problem, and work out the lower bounds
- proposal (ddl: April 18, 2021) + presentation + report
- group size: 1-4 students
- feel free to approach us for feedbacks and topic suggestions
- failed attempts welcomed

What you can expect from us

- we're here to guide your learning and try to challenge you to engage in the learning process
- we'll do our best to give you tools, feedback, and support to succeed, and please let us know if we can do anything more
- learning is a never-ending process, so we hope to motivate you to seek out more information on topics we dont have time to cover
- we encourage you to visit us in office hours or to email us for meeting individually. we want to get to know you and support you in this learning experience!

This lecture

Lower bounds in puzzles

Puzzle I: card guessing game

Problem:

- Alice draws five cards from a deck of n cards
- Alice reveals four of her cards to Bob, one after one
- Bob correctly guesses what is the remaining card
- For which n is this possible?

A (simple?) scheme for $n = 52$ (due to William Fitch Cheney):

two cards w. the same suit (spade) $x, y \in \{1, 2, \cdots, 13\}$

$\mathbb{F}_{13}$ or $\mathbb{Z}/13\mathbb{Z}$.

$x - y \in \{1, 2, \cdots, 6\} \mod 13.$

First card: y use remaining c_1, c_2, c_3 to encode 6 possibilities.

Solution to Puzzle I

Proof of $n \leq 124$:

A = set of 5-tuples unordered

B = set of 4-tuples ordered

Alice: $f: \mathcal{A} \to \mathcal{B}$, $A \mapsto B$ ($B \not\ni A$)

Bob: $g: \mathcal{B} \to \mathcal{A}$, $B \mapsto A$

$g(f(A)) = A$.

f injective: $|\mathcal{A}| \leq |\mathcal{B}| \implies \binom{n}{5} \leq \binom{n}{4} \cdot 4! \implies n \leq 124$.

Scheme for $n = 124$:

$c_1 < c_2 < c_3 < c_4 < c_5$

1) decide which card to keep → $s = c_1 + c_2 + \cdots + c_5 \pmod 5$; Alice keeps c_s

2) use the order to encode 24 possibilities ($24 = 4!$)

Bob: $t = d_1 + d_2 + \cdots + d_4 \pmod 5$

$$\begin{cases} t + x \equiv 1 \pmod 5 \ \& \ x < d_1 \\ t + x \equiv 2 \pmod 5 \ \& \ d_1 < x < d_2 \\ \vdots \end{cases}$$

Puzzle II: coin flipping game

Problem:

- An audience flips n coins on the table
- The audience chooses one of the coin and secretly tells Alice
- Alice flips one of the coin
- Bob comes to the table, and identify which coin is chosen by the audience
- For which n is this possible?

Structure of this problem:

$s \in \{0,1\}^n$ initial state

$k \in [n]$: chosen by the audience

Alice's strategy:

$f(s,k) \in [n]$

Bob: recover k from

$s \oplus f(s,k) \longleftrightarrow k$

Solution to Puzzle II

Claim: it is equivalent to have an n-coloring on the vertices of a binary cube $\{0,1\}^n$, such that the n neighbors of each vertex exactly have n different colors

$\Longrightarrow$ we color s as k-th color if Bob outputs k when seeing s

$\Longleftarrow$ similar.

Impossibility result: n must be a power of 2

$n \mid 2^n \Rightarrow$ ↗

Possibility result: $n = 2^m$ suffices

$s \in \{0,1\}^n \rightarrow$ color $s_1 g_1 + \cdots + s_n g_n \in G$

$G = \{g_1, \cdots, g_n\}$ $\quad 2g = 0 \ (g + g = 0) \quad \forall g \in G.$

$G = \mathbb{F}_2^m \quad (0, 1, \cdots, 0) \in \mathbb{F}_2^m$

Puzzle III: hat guessing game

Problem:

- There are 15 people around the table, each of whom wears a red or blue hat independently with probability 1/2
- Everyone can see all others' hats, but not their own
- Everyone simultaneously chooses to guess the color of his/her hat, or chooses to pass
- They win if there is at least one person choosing to guess, and all guesses are correct
- What is the maximum winning probability?

$\frac{15}{16}$

Structure of this problem:

$(s_1, \cdots, s_{15})$

each $k \in [15]$

$f_k(s_1, s_2, \cdots, s_{k-1}, s_{k+1}, \cdots, s_{15}) \in \{R, B, \text{Pass}\}$.

Solution to Puzzle III

Impossibility result: $p \leq 15/16$

$S \qquad (s_1, \cdots, s_{15}) \in S \xrightarrow{f} (s_1, \cdots, s_{k-1}, \tilde{s}_k, s_{k+1}, \cdots, s_n) \in F.$

$\Downarrow$

$\exists\, k \in [15] \quad f_k \in \{R, B\}$

$\forall s \in F, \quad |f^{-1}(s)| \leq 15 \Rightarrow |S| \leq 15|F|$

$p = \frac{|S|}{|S|+|F|} \leq \frac{15}{16}.$

Possibility result: $p = 15/16$ is achievable

Make $|f^{-1}(\{s\})| = 15$ always hold!

Next lecture

Importance of problem structure:

- example in communication complexity
- slight change in the problem leads to dramatic change in the final answer