Information Theory and Coding Methods in Machine Learning and Statistics

Jonathan Scarlett

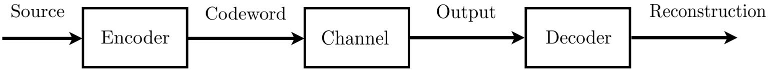
National University of Singapore

CS3236 Optional Lecture
[April 2023]

Information Theory

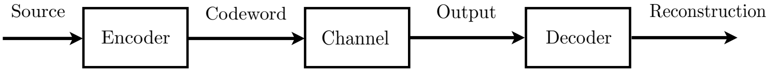
- How do we quantify "information" in data?

- **Information theory** [Shannon, 1948]:
  - Fundamental limits of data communication



Information Theory

- How do we quantify "information" in data?
- **Information theory** [Shannon, 1948]:
  - Fundamental limits of data communication

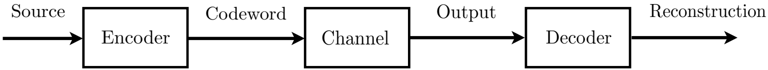


  - Information of source: Entropy
  - Information learned at channel output: Mutual information

Information Theory

- How do we quantify "information" in data?

- **Information theory** [Shannon, 1948]:
  - Fundamental limits of data communication



  - Information of source: Entropy
  - Information learned at channel output: Mutual information

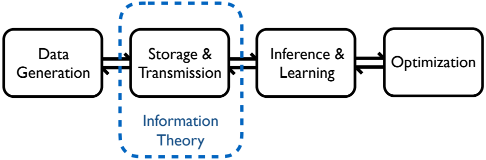
Principles:

- First **fundamental limits** without complexity constraints, then practical methods
- First **asymptotic analyses**, then convergence rates, finite-length, etc.
- Mathematically tractable **probabilistic models**

Information Theory and Data

- Conventional view:

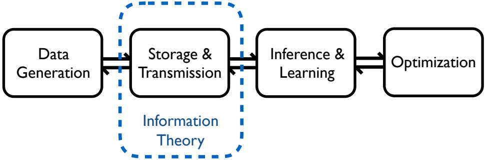
Information theory is a theory of communication



Information Theory and Data

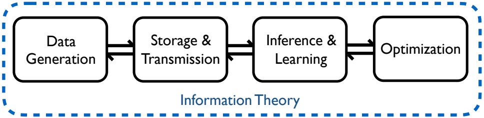
- Conventional view:

Information theory is a theory of communication



- Emerging view:

Information theory is a theory of data



- Extracting information from channel output vs. Extracting information from data

Examples

- **Information theory in machine learning and statistics:**
  - Statistical estimation [Le Cam, 1973]
  - Group testing [Malyutov, 1978]
  - Multi-armed bandits [Lai and Robbins, 1985]
  - Phylogeny [Mossel, 2004]
  - Sparse recovery [Wainwright, 2009]
  - Graphical model selection [Santhanam and Wainwright, 2012]
  - Convex optimization [Agarwal *et al.*, 2012]
  - DNA sequencing [Motahari *et al.*, 2012]
  - Sparse PCA [Birnbaum *et al.*, 2013]
  - Community detection [Abbe, 2014]
  - Matrix completion [Riegler *et al.*, 2015]
  - Ranking [Shah and Wainwright, 2015]
  - Adaptive data analysis [Russo and Zou, 2015]
  - Supervised learning [Nokleby, 2016]
  - Crowdsourcing [Lahouti and Hassibi, 2016]
  - Distributed computation [Lee *et al.*, 2018]
  - Bayesian optimization [Scarlett, 2018]
- **Note:** More than just using entropy / mutual information...

Analogies

Same concepts, different terminology:

Communication Problems	**Data Problems**
Channels with feedback	Active learning / adaptivity
Rate distortion theory	Approximate recovery
Joint source-channel coding	Non-uniform prior
Error probability	Error probability
Random coding	Random sampling
Side information	Side information
Channels with memory	Statistically dependent measurements
Mismatched decoding	Model mismatch
...	...

Cautionary Notes

Some cautionary notes on the information-theoretic viewpoint:

- The simple models we can analyze may be over-simplified (more so than in communication)
- Compared to communication, we often can't get matching achievability/converse (often settle with correct scaling laws)
- Information-theoretic limits not (yet) considered much in practice (to my knowledge) ... but they do guide the algorithm design
- Often encounter gaps between information-theoretic limits and computation limits
- Often information theory simply isn't the right tool for the job

Lecture Plan

Note: The preceding slides are mostly about theoretical results (fundamental performance limits), but practical coding techniques can similarly have a significant impact beyond communication and compression.

Lecture plan:

- Part I: Error-Correcting Codes in Statistical Problems
- Part II: Information-Theoretic Measures in Machine Learning
- Part III: Information-Theoretic Limits of Statistical Problems

Part I: Error-Correcting Codes in Statistical Problems

Warm-Up: A Card Trick

- **A card trick:**
  - Alice and Bob let the audience shuffle a deck and give 5 arbitrary cards to Alice.
  - Alice places 4 of these cards on the table
  - Bob (correctly) guesses the unknown 5th card. How is this possible?

Shown:

Hidden:

Warm-Up: A Card Trick

- **A card trick:**
  - Alice and Bob let the audience shuffle a deck and give 5 arbitrary cards to Alice.
  - Alice places 4 of these cards on the table
  - Bob (correctly) guesses the unknown 5th card. How is this possible?

Shown:

Hidden:

- **Initial Approach:**
  - Number of cards from 1 to 52 in some fixed order (known to Alice and Bob)

Warm-Up: A Card Trick

- **A card trick:**
  - Alice and Bob let the audience shuffle a deck and give 5 arbitrary cards to Alice.
  - Alice places 4 of these cards on the table
  - Bob (correctly) guesses the unknown 5th card. How is this possible?

Shown:

Hidden:

- **Initial Approach:**
  - Number of cards from 1 to 52 in some fixed order (known to Alice and Bob)
  - When placing 4 cards, we can order them in $4! = 24$ different ways. But there are $52 - 4 = 48$ possible choices of the hidden card..?

Warm-Up: A Card Trick

- **A card trick:**
  - Alice and Bob let the audience shuffle a deck and give 5 arbitrary cards to Alice.
  - Alice places 4 of these cards on the table
  - Bob (correctly) guesses the unknown 5th card. How is this possible?

Shown:

Hidden:

- **Initial Approach:**
  - Number of cards from 1 to 52 in some fixed order (known to Alice and Bob)
  - When placing 4 cards, we can order them in $4! = 24$ different ways. But there are $52 - 4 = 48$ possible choices of the hidden card..?
  - Need to somehow convey information via the 5 choices of hidden card itself

Warm-Up: A Card Trick

- **A card trick:**
  - Alice and Bob let the audience shuffle a deck and give 5 arbitrary cards to Alice.
  - Alice places 4 of these cards on the table
  - Bob (correctly) guesses the unknown 5th card. How is this possible?

Shown:

Hidden:

- **Initial Approach:**
  - Number of cards from 1 to 52 in some fixed order (known to Alice and Bob)
  - When placing 4 cards, we can order them in $4! = 24$ different ways. But there are $52 - 4 = 48$ possible choices of the hidden card..?
  - Need to somehow convey information via the 5 choices of hidden card itself

- **Elegant Solution:**
  - Find two cards A and B with the same suit (always possible!)

Warm-Up: A Card Trick

- **A card trick:**
  - Alice and Bob let the audience shuffle a deck and give 5 arbitrary cards to Alice.
  - Alice places 4 of these cards on the table
  - Bob (correctly) guesses the unknown 5th card. How is this possible?

Shown:

Hidden:

- **Initial Approach:**
  - Number of cards from 1 to 52 in some fixed order (known to Alice and Bob)
  - When placing 4 cards, we can order them in $4! = 24$ different ways. But there are $52 - 4 = 48$ possible choices of the hidden card..?
  - Need to somehow convey information via the 5 choices of hidden card itself

- **Elegant Solution:**
  - Find two cards A and B with the same suit (always possible!)
  - Either A's number index $+6$ passes B, or vice versa (where $13 + 1$ wraps to 1)

Warm-Up: A Card Trick

- **A card trick:**
  - Alice and Bob let the audience shuffle a deck and give 5 arbitrary cards to Alice.
  - Alice places 4 of these cards on the table
  - Bob (correctly) guesses the unknown 5th card. How is this possible?

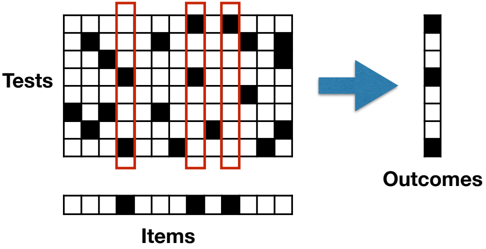
Shown:

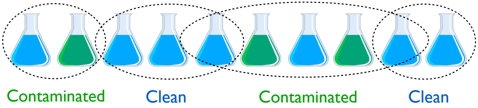
Hidden:

- **Initial Approach:**
  - Number of cards from 1 to 52 in some fixed order (known to Alice and Bob)
  - When placing 4 cards, we can order them in $4! = 24$ different ways. But there are $52 - 4 = 48$ possible choices of the hidden card..?
  - Need to somehow convey information via the 5 choices of hidden card itself

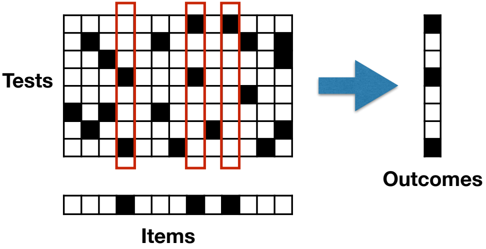
- **Elegant Solution:**
  - Find two cards A and B with the same suit (always possible!)
  - Either A's number index +6 passes B, or vice versa (where $13 + 1$ wraps to 1)
  - Place A (or B) down first, then order the remaining 3 cards to index 6 numbers

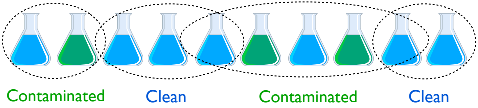
Group Testing





Group Testing





- **Goal:**

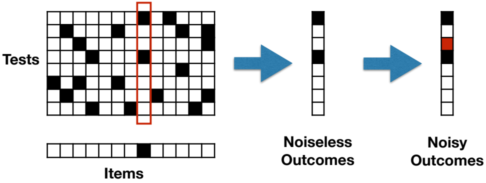
  Given test matrix $\mathbf{X}$ and outcomes $\mathbf{Y}$, recover item vector β

  ...while minimizing the number of tests n

- **Terminology:** The word "defective" replaces "contaminated" or "infected"

1-Sparse Group Testing

- **Simplest case:** Exactly one defective item



- **Noiseless case:** Easy – just let column i be the binary representation of i

- **Noisy case:** Exactly equivalent to channel coding!
  - #items $\Longleftrightarrow$ #messages
  - i-th codeword $\Longleftrightarrow$ i-th column of the test matrix
  - #tests $\Longleftrightarrow$ block length

Of course, having just one defective item is of limited practical interest...

General Group Testing

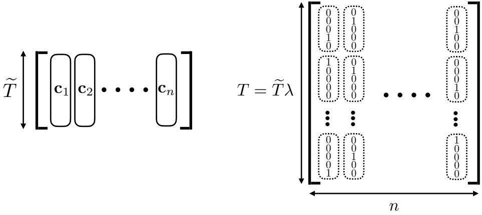
- **Information theory inspired approach:**
  - Much like random coding in channel coding, we can do random testing here
  - Gives strong (and often asymptotically optimal) theoretical guarantees

General Group Testing

- **Information theory inspired approach:**
  - Much like random coding in channel coding, we can do random testing here
  - Gives strong (and often asymptotically optimal) theoretical guarantees
- **Coding based approach #1:**
  - Test designs exists that first identify subsets containing exactly one defective
  - Once this is done, we can get its index using the approach on the previous slide

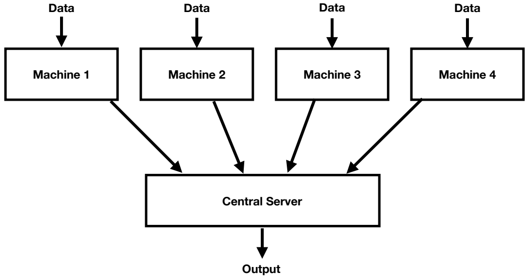
General Group Testing

- **Information theory inspired approach:**
  - Much like random coding in channel coding, we can do random testing here
  - Gives strong (and often asymptotically optimal) theoretical guarantees
- **Coding based approach #1:**
  - Test designs exists that first identify subsets containing exactly one defective
  - Once this is done, we can get its index using the approach on the previous slide
- **Coding based approach #2:** (Kautz-Singleton, 1964; see also arXiv:1808.01457)
  - Step 1: Design a non-binary matrix with Reed-Solomon codewords as columns
  - Step 2: Replace non-binary symbols $A \to 10\ldots0$, $B \to 010\ldots0$, etc.



Coded Computation

- Recently increasing attention has been paid to coding in distributed computation:



- **Motivation:** What if the machines are unreliable and some may not respond?

- **Idea:** Introduce resilience via error-correcting coding (i.e., perform redundant computations to increase resilience to failures)

Simple Strategies

- **Computing summations:** Data set $\mathcal{D}$ consists of N "data points" $(\mathbf{z}_1, \dots, \mathbf{z}_N)$, and we want to compute some summation of the form $\sum_{i=1}^{N} f(\mathbf{z}_i)$.

Simple Strategies

- **Computing summations:** Data set $\mathcal{D}$ consists of N "data points" $(\mathbf{z}_1, \dots, \mathbf{z}_N)$, and we want to compute some summation of the form $\sum_{i=1}^N f(\mathbf{z}_i)$.

- **Strategy 1:** Send all data to all machines and have them return $\sum_{i=1}^N f(\mathbf{z}_i)$
  - For huge data sets, this is likely infeasible

Simple Strategies

- **Computing summations:** Data set $\mathcal{D}$ consists of N "data points" $(\mathbf{z}_1, \dots, \mathbf{z}_N)$, and we want to compute some summation of the form $\sum_{i=1}^N f(\mathbf{z}_i)$.

- **Strategy 1:** Send all data to all machines and have them return $\sum_{i=1}^N f(\mathbf{z}_i)$
  - For huge data sets, this is likely infeasible

- **Strategy 2:** Split the data, say $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2)$; send $\mathcal{D}_1$ (only) to a few machines, and $\mathcal{D}_2$ (only) to a few machines
  - Turns out to be very wasteful of machines as we scale things up

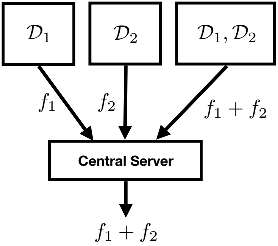
Simple Strategies

- **Computing summations:** Data set $\mathcal{D}$ consists of N "data points" $(\mathbf{z}_1, \dots, \mathbf{z}_N)$, and we want to compute some summation of the form $\sum_{i=1}^N f(\mathbf{z}_i)$.

- **Strategy 1:** Send all data to all machines and have them return $\sum_{i=1}^N f(\mathbf{z}_i)$
  - For huge data sets, this is likely infeasible

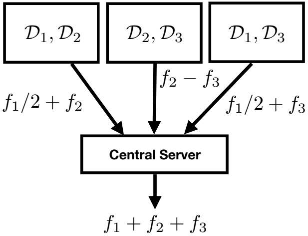
- **Strategy 2:** Split the data, say $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2)$; send $\mathcal{D}_1$ (only) to a few machines, and $\mathcal{D}_2$ (only) to a few machines
  - Turns out to be very wasteful of machines as we scale things up

- **Very simple coding example:**



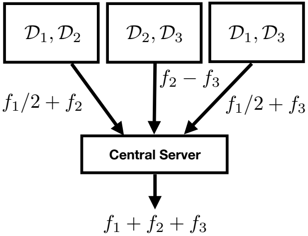
Coding Strategies

- **Less simple coding example:**



Coding Strategies

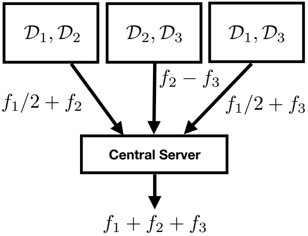
- **Less simple coding example:**



- **Generalized version:**
  - Split data into parts, design allocation of parts to machines
  - Use linear algebra techniques to design weighting coefficients
  - Trade-off between (i) total #machines needed, (ii) #machines that can fail, and (iii) amount of data per machine

Coding Strategies

- **Less simple coding example:**



- **Generalized version:**
  - Split data into parts, design allocation of parts to machines
  - Use linear algebra techniques to design weighting coefficients
  - Trade-off between (i) total #machines needed, (ii) #machines that can fail, and (iii) amount of data per machine

Notes:

- Key difference to regular codes is using real arithmetic instead of modulo-2
- For details, see `https://arxiv.org/abs/1612.03301`
- Other computation tasks include matrix multiplication, Fourier transform, etc.

Other Uses of Error-Correcting Codes

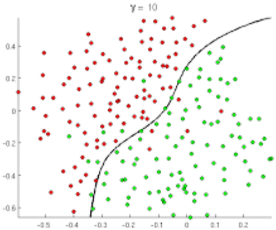
Other non-standard applications of error-correcting codes:

- Distributed storage
- Statistical inverse problems (e.g., compressive sensing)
- Cryptography
- Hashing
- Theoretical computer science proofs (and algorithms)

Part II: Information-Theoretic Measures in Machine Learning

Binary Classification

- **Illustration of binary classification problem:**



  - Features $\mathbf{x} \in \mathbb{R}^d$ (e.g., age, income, #years working)
  - Label $y \in \{-1, 1\}$ (e.g., is this person going to repay their loan?)

- Learning is done via training data, i.e., a collection $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of pairs that we believe to be representative of the population (e.g., historical data)

Feature Selection

• Suppose that in the dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, each input $\mathbf{x}$ has a large number of mostly-irrelevant features. How to find which are relevant?

• **A popular approach:** Seek features such that (an empirical estimate of) the mutual information is as high as possible:

$$\text{maximize}_{S\,:\,|S|\le k} I(\mathbf{X}_S; Y),$$

where $\mathbf{x}_S$ is the subset of $\mathbf{x}$ containing only the features indexed by S.

- **Intuition:** Find the features that are most informative about Y

Compact Representations

- Building on the previous slide, researchers have used mutual information to measure the compactness and informativeness of features $\{u_i\}_{i=1}^n$ produced by an algorithm:
  - **Informativeness:** $I(\mathbf{U}; Y)$ is large (motivated by channel coding)
  - **Compactness:** $I(\mathbf{U}; \mathbf{X})$ is small (motivated by rate-distortion theory)

(e.g., see arXiv:1703.00810)

Compact Representations

- Building on the previous slide, researchers have used mutual information to measure the compactness and informativeness of features $\{u_i\}_{i=1}^n$ produced by an algorithm:
  - **Informativeness:** $I(\mathbf{U}; Y)$ is large (motivated by channel coding)
  - **Compactness:** $I(\mathbf{U}; \mathbf{X})$ is small (motivated by rate-distortion theory)

(e.g., see arXiv:1703.00810)

- **Problems/limitations:** (e.g., see arXiv:1802.09766, arXiv:1810.05728)
  - Mutual information is one of many choices, unclear whether it's the "best"
  - Can be unclear whether these quantities actually translate to the ultimate goal (e.g., classification prediction accuracy)
  - May fail to capture important aspects (e.g., learnability, robustness)
  - In continuous-valued settings, the mutual information can trivially be ∞, or exhibit other trivial behavior

- **General principle:** Ideally (*in my opinion*), measures like entropy, mutual information, and KL divergence are most powerful when they are not introduced manually, but instead naturally arise as the answer to a fundamental problem

Generalization Bounds

- One of the most fundamental concepts in learning theory is generalization:
  - Training accuracy: Measure of accuracy on training data
  - Test accuracy: Measure of accuracy on (unseen) test data
  - Generalization error: The difference between the two

Generalization Bounds

- One of the most fundamental concepts in learning theory is generalization:
  - Training accuracy: Measure of accuracy on training data
  - Test accuracy: Measure of accuracy on (unseen) test data
  - Generalization error: The difference between the two

- **Information-theoretic approach:** Under certain conditions, it can be shown that the generalization error is small when the learning algorithm output doesn't depend overly strongly on the training data. Mathematically,

$$\text{Generalization error} \lesssim \sqrt{I(\mathcal{D}; W)/n}, \tag{1}$$

where $\mathcal{D}$ is the training data (of size n), and W is the learning algorithm's output

- Here mutual information appears in the result but not in the problem formulation

- **Further details:** arXiv:1511.05219, arXiv:1705.07809

Part III: Information-Theoretic Limits of Statistical Problems

Statistical Estimation

Statistical estimation problems:

- Seek to estimate an unknown quantity θ (may be discrete, continuous, or some abstract type)
- We have access to data samples $Y_1, \dots, Y_n$ drawn independently from some P_θ
- (In some cases, each Y_i has an associated "input" $\mathbf{X}_i$)

Statistical Estimation

Statistical estimation problems:

- Seek to estimate an unknown quantity θ (may be discrete, continuous, or some abstract type)
- We have access to data samples $Y_1, \dots, Y_n$ drawn independently from some P_θ
- (In some cases, each Y_i has an associated "input" $\mathbf{X}_i$)

Example 1: Gaussian mean estimation

- $Y_i = \theta + Z_i$ where $\theta \in \mathbb{R}^d$ and Z_i is i.i.d. Gaussian noise
- Estimation error: $\|\hat{\theta} - \theta\|^2 = \sum_{i=1}^d (\hat{\theta}_i - \theta_i)^2$

Statistical Estimation

Statistical estimation problems:

- Seek to estimate an unknown quantity θ (may be discrete, continuous, or some abstract type)
- We have access to data samples $Y_1, \dots, Y_n$ drawn independently from some P_θ
- (In some cases, each Y_i has an associated "input" $\mathbf{X}_i$)

Example 1: Gaussian mean estimation

- $Y_i = \theta + Z_i$ where $\theta \in \mathbb{R}^d$ and Z_i is i.i.d. Gaussian noise
- Estimation error: $\|\hat{\theta} - \theta\|^2 = \sum_{i=1}^d (\hat{\theta}_i - \theta_i)^2$

Example 2: Group testing

- θ is the defective set, Y_i is the i-th test outcome, $\mathbf{X}_i$ is the i-th test design
- Probability of error: $\mathbb{P}[\hat{\theta} \neq \theta]$

Terminology: Achievability and Converse

Achievability result (example): Given $\overline{n}(\epsilon)$ data samples, there exists an algorithm achieving an "error" of at most ϵ

- Discrete estimation error: $\mathbb{P}[\hat{\theta} \neq \theta] \le \epsilon$
- Continuous estimation error: $\|\hat{\theta} - \theta_{\rm true}\|^2 \le \epsilon$
- Optimization error: $f(x_{\rm selected}) \le \min_x f(x) + \epsilon$

(The latter two may be either *on average* or *with high probability*)

Terminology: Achievability and Converse

Achievability result (example): Given $\overline{n}(\epsilon)$ data samples, there exists an algorithm achieving an "error" of at most ϵ

- Discrete estimation error: $\mathbb{P}[\hat{\theta} \neq \theta] \leq \epsilon$
- Continuous estimation error: $\|\hat{\theta} - \theta_{\rm true}\|^2 \leq \epsilon$
- Optimization error: $f(x_{\rm selected}) \leq \min_x f(x) + \epsilon$

(The latter two may be either *on average* or *with high probability*)

Converse result (example): In order to achieve an "error" of at most ϵ, any algorithm requires at least $\underline{n}(\epsilon)$ data samples

Converse results tend to be where information theory plays a larger role in statistical problems

High-Level Steps

Example steps in attaining a converse bound:

1. Reduce estimation problem to multiple hypothesis testing
2. Apply a form of Fano's inequality
3. Bound the resulting mutual information term

(*Multiple hypothesis testing*: Given samples $Y_1, \dots, Y_n$, determine which distribution among $P_1(\mathbf{y}), \dots, P_M(\mathbf{y})$ generated them. $M = 2$ gives binary hypothesis testing.)

Fano's Inequality

- Fano's inequality as stated in textbooks:

$$H(V|\hat{V}) \le H_2(P_{\mathrm{e}}) + P_{\mathrm{e}} \log_2(M-1)$$

where M is the number of values that V can take, and $P_{\mathrm{e}} = \mathbb{P}[\hat{V} \neq V]$

Fano's Inequality

- Fano's inequality as stated in textbooks:

$$H(V|\hat{V}) \leq H_2(P_\mathrm{e}) + P_\mathrm{e} \log_2(M-1)$$

where M is the number of values that V can take, and $P_\mathrm{e} = \mathbb{P}[\hat{V} \neq V]$

- Useful form for M-ary hypothesis testing and uniform V:

$$\mathbb{P}[\hat{V} \neq V] \geq 1 - \frac{I(V;\hat{V}) + \log 2}{\log M}.$$

  - Intuition: Need learned information $I(V;\hat{V})$ to be close to prior uncertainty $\log M$, otherwise the error probability will be significant

Fano's Inequality

- Fano's inequality as stated in textbooks:

$$H(V|\hat{V}) \le H_2(P_{\rm e}) + P_{\rm e} \log_2(M-1)$$

where M is the number of values that V can take, and $P_{\rm e} = \mathbb{P}[\hat{V} \ne V]$

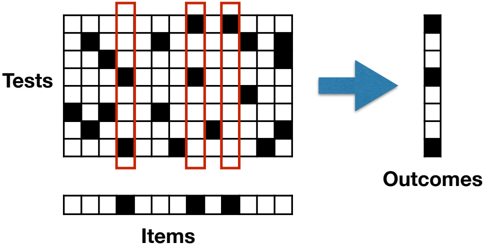
- Useful form for M-ary hypothesis testing and uniform V:

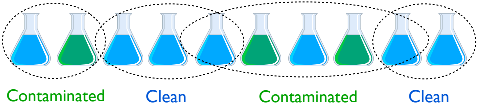
$$\mathbb{P}[\hat{V} \ne V] \ge 1 - \frac{I(V;\hat{V}) + \log 2}{\log M}.$$

  - Intuition: Need learned information $I(V;\hat{V})$ to be close to prior uncertainty $\log M$, otherwise the error probability will be significant

- **Variations:**
  - Non-uniform V
  - Approximate recovery
  - Conditional version

Group Testing





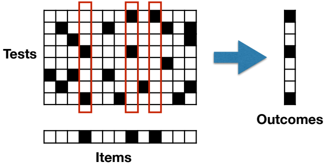
- **Goal:**

  Given test matrix $\mathbf{X}$ and outcomes $\mathbf{Y}$, recover item vector β

  ...while minimizing the number of tests n

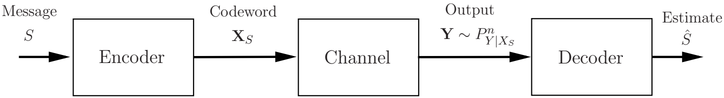
- **Terminology:** The word "defective" replaces "contaminated" or "infected"

Information Theory and Group Testing



- **Information-theoretic viewpoint:**

  S : Defective set
  $\mathbf{X}_S$: Columns indexed by S



- **Example formulation of general result:**

Sample complexity

Entropy

(Model uncertainty)

$$n^* \sim \frac{H(S)}{I(P_{Y|X_S})}$$

Mutual Information

(Information learned from measurements)

Converse via Fano's Inequality

- **Reduction to multiple hypothesis testing:** Trivial! Set $V = S$.

Converse via Fano's Inequality

- **Reduction to multiple hypothesis testing:** Trivial! Set $V = S$.

- **Application of Fano's Inequality:**

$$\mathbb{P}[\hat{S} \neq S] \geq 1 - \frac{I(S; \hat{S} | \mathbf{X}) + \log 2}{\log \binom{p}{k}}$$

Converse via Fano's Inequality

- **Reduction to multiple hypothesis testing:** Trivial! Set $V = S$.

- **Application of Fano's Inequality:**

$$\mathbb{P}[\hat{S} \neq S] \geq 1 - \frac{I(S;\hat{S}|\mathbf{X}) + \log 2}{\log \binom{p}{k}}$$

- **Mutual information bound:** $I(S;\hat{S}|\mathbf{X}) \leq nC$ where C is the capacity of the "channel" that introduces noise to the test outcomes

Converse via Fano's Inequality

- **Reduction to multiple hypothesis testing:** Trivial! Set $V = S$.

- **Application of Fano's Inequality:**

$$\mathbb{P}[\hat{S} \neq S] \geq 1 - \frac{I(S; \hat{S} | \mathbf{X}) + \log 2}{\log \binom{p}{k}}$$

- **Mutual information bound:** $I(S; \hat{S} | \mathbf{X}) \leq nC$ where C is the capacity of the "channel" that introduces noise to the test outcomes

- **Final result:** With p items, k defectives, and n tests, we have

$$n \leq \frac{k \log \frac{p}{k}}{C}(1 - \epsilon) \implies \mathbb{P}[\hat{S} \neq S] \not\to 0.$$

where the $k \log \frac{p}{k}$ numerator comes from an asymptotic simplification of $\log \binom{p}{k}$

Further Results

Further uses of information theory in group testing:

- Information-theoretic achievability (much more technically challenging, but the final result often matches the above converse)
- Practical algorithms inspired by information-theoretic analyses
- Coding-based test designs

Survey article: arXiv:1902.06002

What About Continuous-Valued Estimation?

Running Example: Gaussian Mean Estimation

- To simplify the discussion, let's focus on the problem of Gaussian mean estimation
- **Gaussian mean estimation:**
  - There exists an unknown vector $\theta \in \mathbb{R}^p$ we would like to estimate
  - The data given to us is $Y_1, \dots, Y_n$, where

$$Y_i = \theta + Z_i$$

    with $Z_i \in \mathbb{R}^p$ being i.i.d. $N(0, \sigma^2)$ additive noise
  - In other words, estimate θ from independent $N(\theta, \sigma^2 I_p)$ samples

- **Algorithmic goal:** Design an estimation algorithm to obtain an estimate $\hat{\theta}$ such that $\|\theta - \hat{\theta}\| \le \epsilon$ for some target accuracy ϵ (either in expectation or with high probability – we will not worry so much about the details)

High-Level Steps

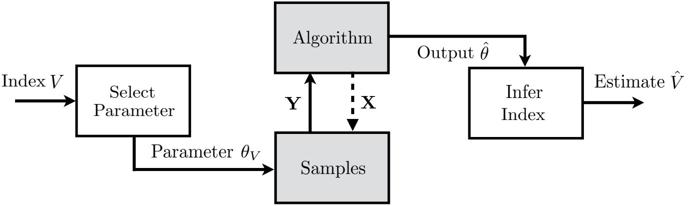
Steps in attaining a converse bound:

1. Reduce estimation problem to multiple hypothesis testing
2. Apply a form of Fano's inequality
3. Bound the resulting mutual information term

(*Multiple hypothesis testing*: Given samples $Y_1, \dots, Y_n$, determine which distribution among $P_1(\mathbf{y}), \dots, P_M(\mathbf{y})$ generated them. $M = 2$ gives binary hypothesis testing.)

Reduction to Multiple Hypothesis Testing (I)

- Lower bound worst-case error by average over <u>hard subset</u> $\theta_1, \ldots, \theta_M$:

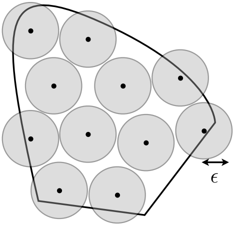


Idea:

- Show "successful" algorithm $\hat{\theta} \implies$ Correct estimation of V (When is this true?)
- <u>Equivalent statement</u>: *If V can't be estimated reliably, then $\hat{\theta}$ can't be successful.*

Reduction to Multiple Hypothesis Testing (II)

- **Example:** Suppose algorithm is claimed to return $\hat{\theta}$ such that $\|\hat{\theta} - \theta\|_2 \leq \epsilon$



- If $\theta_1, \ldots, \theta_M$ are separated by 2ϵ, then we can identify the correct $V \in \{1, \ldots, M\}$

- **Note:** Tension between number of hypotheses, difficulty in distinguishing them, and sufficient separation. *Choosing a suitable set* $\{\theta_1, \ldots, \theta_M\}$ *can be challenging.*

Mutual Information Bound

- For simplicity, first consider the 1D case, i.e., $\theta \in \mathbb{R}$ and $Y = \theta + Z$

- In this case, a suitable choice is $\theta_1 = +C$ and $\theta_2 = -C$ for some constant C
  - Mutual information essential reduces to $D\big(N(+C, \sigma^2) \| N(-C, \sigma^2)\big)$, which is easily computed to equal $\frac{2C^2}{\sigma^2}$
  - C can be optimized at the end of the analysis to give the best bound

Mutual Information Bound

- For simplicity, first consider the 1D case, i.e., $\theta \in \mathbb{R}$ and $Y = \theta + Z$
- In this case, a suitable choice is $\theta_1 = +C$ and $\theta_2 = -C$ for some constant C
  - Mutual information essential reduces to $D\big(N(+C,\sigma^2)\|N(-C,\sigma^2)\big)$, which is easily computed to equal $\frac{2C^2}{\sigma^2}$
  - C can be optimized at the end of the analysis to give the best bound
- **General d-dimensional case:** Instead consider vectors of the form

$$\boldsymbol{\theta}_i = (C, -C, -C, C, C, \dots, -C, C)$$

and using tools from coding theory to ensure the signs keep them well-separated

Beyond Fano's Inequality

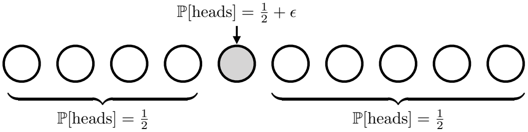
Limitations and Generalizations

- **Limitations of Fano's Inequality.**
  - Non-asymptotic weakness
  - Often hard to tightly bound mutual information in adaptive settings
  - Closely tied to KL divergence (relative entropy) which is not always the ideal measure
- **Generalizations of Fano's Inequality.**
  - Non-uniform V [Han/Verdú, 1994]
  - More general divergences measures [Guntuboyina, 2011]
  - Continuous V [Duchi/Wainwright, 2013]

(This list is certainly incomplete!)

Example: Difficulties in Adaptive Settings

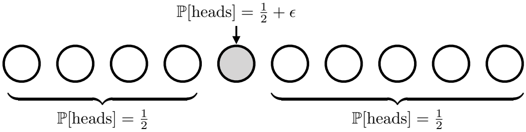
- **A simple search problem:** Find the (only) biased coin using few flips



  - Heavy coin $V \in \{1, \dots, M\}$ uniformly at random
  - Selected coin at time $i = 1, \dots, n$ is X_i, observation is $Y_i \in \{0, 1\}$ (1 for heads)

Example: Difficulties in Adaptive Settings

- **A simple search problem:** Find the (only) biased coin using few flips

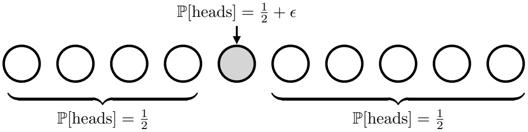


  - Heavy coin $V \in \{1, \dots, M\}$ uniformly at random
  - Selected coin at time $i = 1, \dots, n$ is X_i, observation is $Y_i \in \{0, 1\}$ (1 for heads)

- **Non-adaptive setting:**
  - Since X_i and V are independent, can show $I(V; Y_i | X_i) \lesssim \frac{\epsilon^2}{M}$
  - Substituting into Fano's inequality gives the requirement $n \gtrsim \frac{M \log M}{\epsilon^2}$

Example: Difficulties in Adaptive Settings

- **A simple search problem:** Find the (only) biased coin using few flips



  - Heavy coin $V \in \{1, \dots, M\}$ uniformly at random
  - Selected coin at time $i = 1, \dots, n$ is X_i, observation is $Y_i \in \{0, 1\}$ (1 for heads)

- **Non-adaptive setting:**
  - Since X_i and V are independent, can show $I(V; Y_i | X_i) \lesssim \frac{\epsilon^2}{M}$
  - Substituting into Fano's inequality gives the requirement $n \gtrsim \frac{M \log M}{\epsilon^2}$

- **Adaptive setting:**
  - Nuisance to characterize $I(V; Y_i | X_i)$, as X_i depends on V due to adaptivity!
  - Worst-case bounding only gives $n \gtrsim \frac{\log M}{\epsilon^2}$

Additive Change of Measure

- Let $P(\mathbf{y})$ and $Q(\mathbf{y})$ be two distributions on the observations

Additive Change of Measure

- Let $P(\mathbf{y})$ and $Q(\mathbf{y})$ be two distributions on the observations
- A very basic inequality (essentially by definition):

$$|\mathbb{P}_P[A] - \mathbb{P}_Q[A]| \le \|P - Q\|_{\mathrm{TV}}$$

for any event A

  - Total variation (TV) distance: A measure of the difference between two distributions (KL divergence is another such measure)
  - Intuition:
    - Let Q be a distribution where nothing can reasonably be learned (e.g., pure noise)
    - Then "learning" on Q is doomed to fail
    - So if P is too close to Q, then learning on P is also likely to fail

Additive Change of Measure

- Let $P(\mathbf{y})$ and $Q(\mathbf{y})$ be two distributions on the observations

- A very basic inequality (essentially by definition):

$$|\mathbb{P}_P[A] - \mathbb{P}_Q[A]| \leq \|P - Q\|_{\rm TV}$$

for any event A

  - Total variation (TV) distance: A measure of the difference between two distributions (KL divergence is another such measure)
  - Intuition:
    - Let Q be a distribution where nothing can reasonably be learned (e.g., pure noise)
    - Then "learning" on Q is doomed to fail
    - So if P is too close to Q, then learning on P is also likely to fail

- **Applications:**
  - Statistical estimation [Le Cam, 1973]
  - Multi-armed bandits [Auer *et al.*, 1995]

Multiplicative Change of Measure

- **Multiplicative change of measure:** Relate the probability of a success event $\mathcal{A}$ under two different distributions $P(\mathbf{y}), Q(\mathbf{y})$ as follows

$$\mathbb{P}_P[\mathcal{A}] \le \mathbb{P}_P\left[\frac{P(\mathbf{Y})}{Q(\mathbf{Y})} > \gamma\right] + \gamma \mathbb{P}_Q[\mathcal{A}],$$

where γ is an arbitrary threshold

  - Intuition: Again, Q could be a distribution under which nothing can be learned

Multiplicative Change of Measure

- **Multiplicative change of measure:** Relate the probability of a success event $\mathcal{A}$ under two different distributions $P(\mathbf{y}), Q(\mathbf{y})$ as follows

$$\mathbb{P}_P[\mathcal{A}] \le \mathbb{P}_P\left[\frac{P(\mathbf{Y})}{Q(\mathbf{Y})} > \gamma\right] + \gamma \mathbb{P}_Q[\mathcal{A}],$$

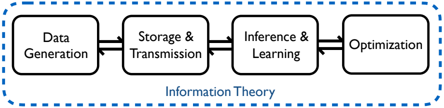
where γ is an arbitrary threshold

  - Intuition: Again, Q could be a distribution under which nothing can be learned

- **Applications:**
  - Channel coding [Wolfowitz, 1957] [Verdú and Han, 1994]
  - Multi-armed bandits [Lai and Robbins, 1985]
  - Statistical estimation [Tsybakov, 2009] [Venkataramanan and Johnson, 2018]
  - Group testing and sparse recovery [Scarlett and Cevher, 2017]

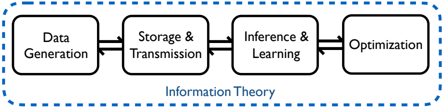
Conclusion

- **Information theory as a theory of data**:



Conclusion

- **Information theory as a theory of data**:



- **Aspects covered in this talk:**
  - Non-standard applications of error correcting codes
  - Information measures in machine learning
  - Information-theoretic limits of statistical problems

Many useful applications of information theory / coding, and more to come!

Tutorial Material

- **Tutorial Chapter:** "An Introductory Guide to Fano's Inequality with Applications in Statistical Estimation" [Scarlett/Cevher, 2021]

https://arxiv.org/abs/1901.00555

(Chapter in book *Information-Theoretic Methods in Data Science*, Cambridge University Press)

- **Group Testing Survey:** "Group Testing: An Information Theory Perspective" [Aldridge/Johnson/Scarlett, 2019]

https://arxiv.org/abs/1902.06002