Information-Theoretic Limits for Inference, Learning, and Optimization (Learnt from workshop CSCIT2019, Prof. Jonathan Scarlett)

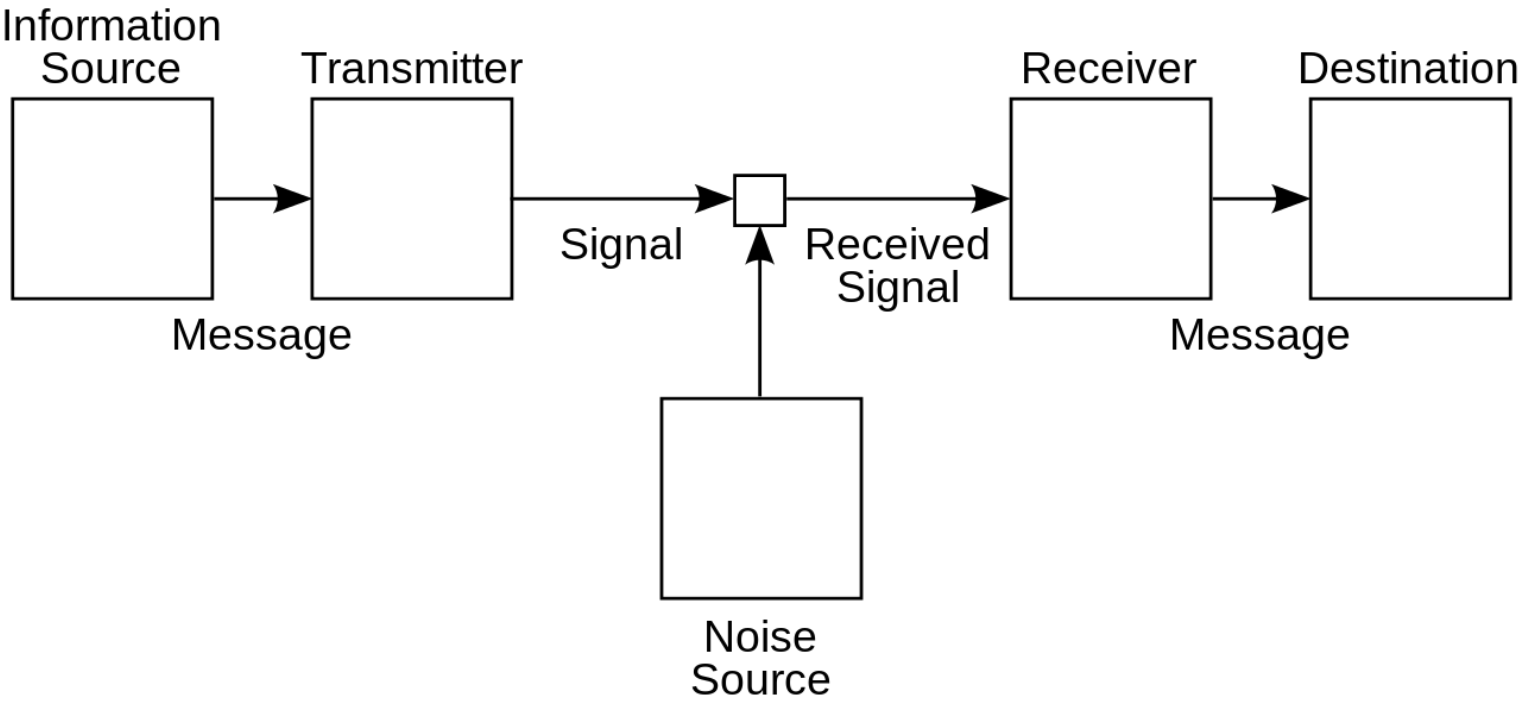
Jie Wang

August 6, 2019

Table of Content

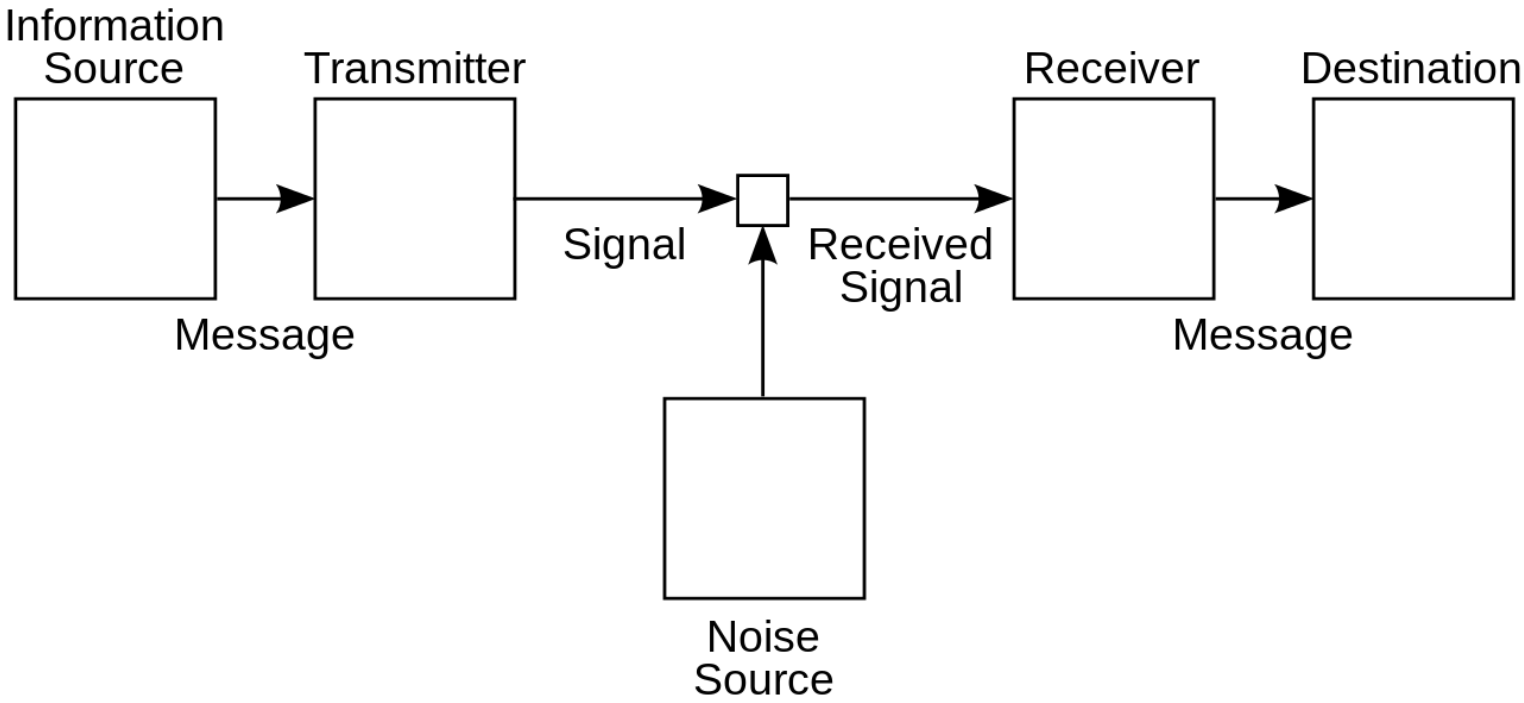
Classical Information Thoery

- How to measure the “information” in data?
- Information Theory [Shannon, 1948]: A Theory of communication.



  - Fundamental Limits of data communication
  - Information of Source: Entropy
  - Information Learned at Channel Output: Mutual Information

Contributions of Information Theory

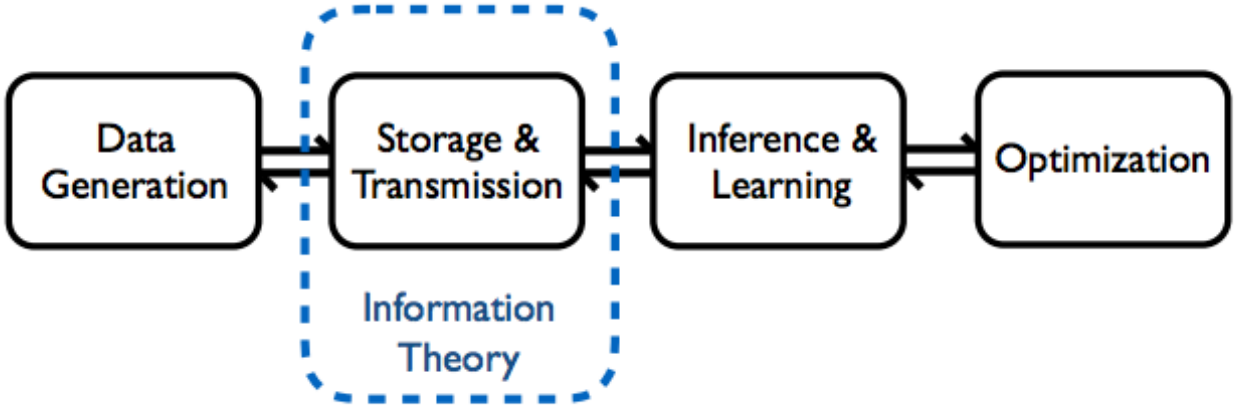


High-Level Contributions:

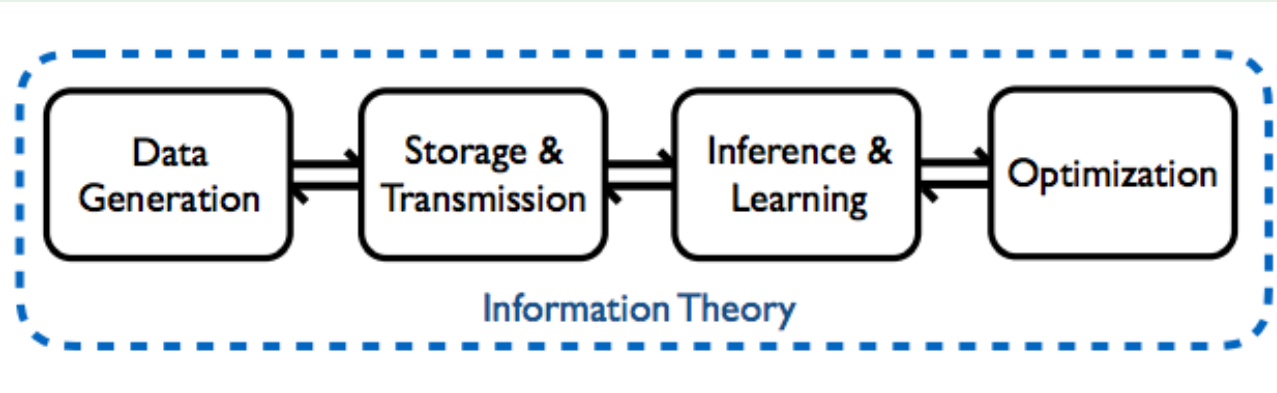
- First fundamental limits. Hard for practial design
- First asymptotic analysis. Hard for fintie block-length analysis.
- Build mathematically tractable *probabilistic models*

Information Theory and Data

Conventional View: Information Theory is a theory of communication.



Emerging View: Information Theory is a theory of data.



Examples

- Information theory in machine learning and statistics:
  - DNA sequencing

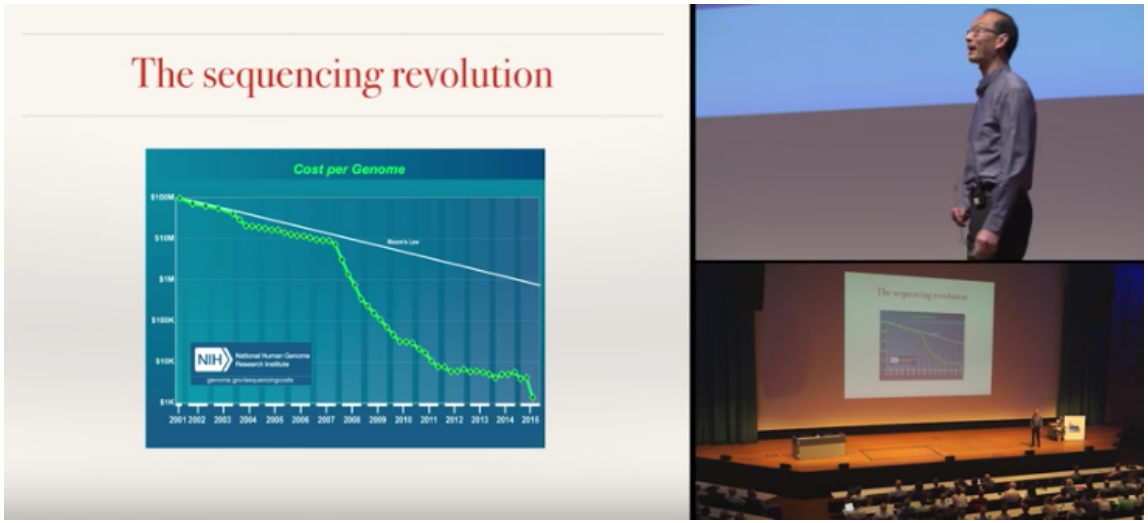


Figure 1: ISIT 2017, David Tse, The Spirit of Information Theory, Available in Youtube

  - Multi-armed bandits
  - Statistical estimation
  - Supervised learning
  - etc.
- Note: More than just using entropy / mutual information

Concept Analogies

Same concepts but different terminology:

Communication Problems	Data Problems
Channons with feedback	Adaptive Learning
Rate Distortion Theory	Approximate Recovery
Joint Source-Channel Coding	Non-uniform prior
Random Coding	Random Sampling
Channels with memory	Statistically dependent measurements
...	...

Cautionary Notes for using Information Theory

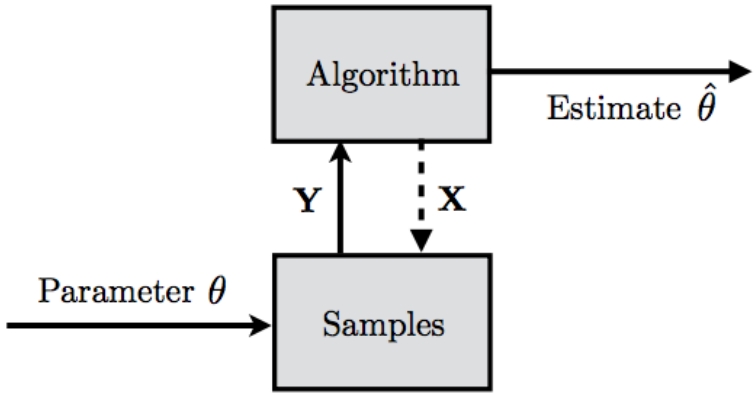
- The simple models we can analyze now may be over-simplified
- Information-theoretic Limits not yet considered in practice, but they do guide algorithm design
- Often encounter gaps between information-theoretic limits and computation limits

Terminologies: Achievability and Converse

- Achievability: Given $\bar{n}(\epsilon)$ data samples, there exists an algorithm achieving an "error" of at most ϵ.
  - Estimation Error: $\|\overline{\theta} - \theta_{\text{true}}\| \leq \epsilon$
  - Optimization Error: $f(x_{\text{selected}}) \leq \min_x f(x) + \epsilon$
- Converse: In order to achieve an "error" of at most ϵ, any algorithm requires at least $\underline{n}(\epsilon)$ data samples.

Statistical Estimation Problem Setting

- Unknown parameter $\theta \in \Theta$
- Given samples $Y = (Y_1, \ldots, Y_n)$ drawn from the distribution $P_\theta(y)$
  - Or more generally, generate input $X = (X_1, \ldots, X_n)$, we get $Y = (Y_1, \ldots, Y_n)$ with $Y = P_\theta(X)$.
- Given Y (and possibly X), construct estimate $\hat{\theta}$.



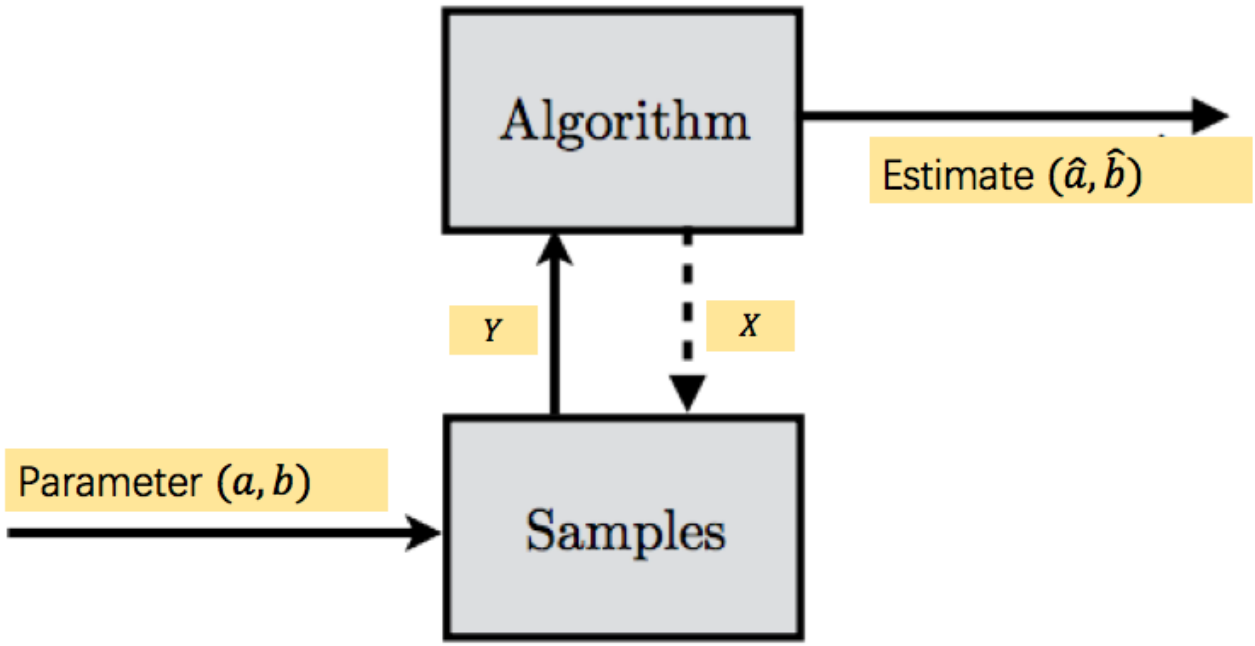
Question: How to obain a minimmax lower bound, i.e., in order to achieve (below) a certain loss for estimation, how many samples are needed, no matter which algorithm is chosen?

Highlights

- Discrete or continuous: The parameter set Θ may be discrete or continuous.
- Minimax: We seek a lower bound for the number of samples such that the loss is small for any given $\theta \in \Theta$, i.e., a worse case performance is considered.
- Choice of loss function:
  - 0-1 loss: $\ell(\theta, \hat{\theta}) = 1_{\theta \neq \hat{\theta}}$
  - quadratic loss: $||\theta - \hat{\theta}||^2$

Typical Example: Linear Regression

- Given data points on two dimension plane, $(x_i, y_i)_{i=1}^n$.
- Assume that $y = a + bx + \mathcal{Z}$.
- Estimate (a, b) from data points.



High-Level Steps for the converse

- Reduce the estimation problem to **multiple hypothesis testing setting**
- Apply a form of **Fano's inequality**
- Bound the resulting **mutual information** term

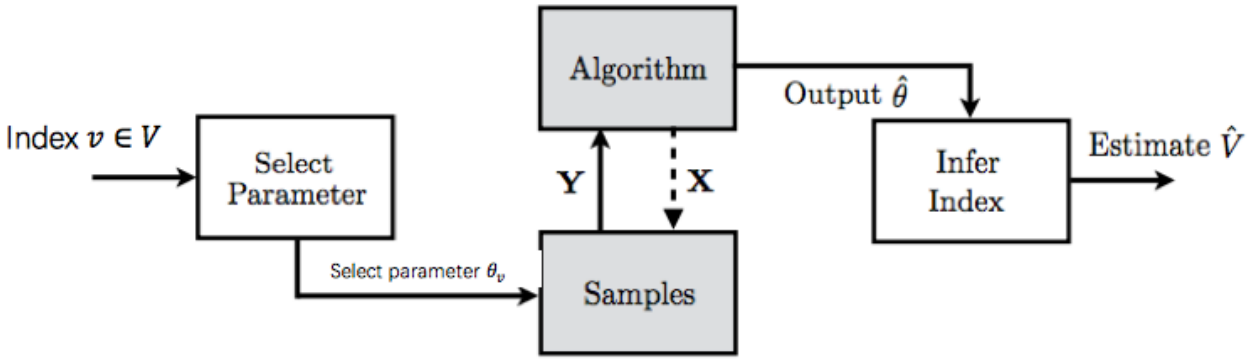
Introduction to Multiple hypothesis testing:

> *Given samples $Y_1, \ldots, Y_n$, which is drawn from an (unknown) underlying distribution, determine which distribution $P_1(Y), \ldots, P_M(Y)$ generated them.*

Step I: : Reduction to Multiple Hypothesis Testing

Lower bound worst-case error by average over the hard subset $\{\theta_1, \ldots, \theta_M\}$.

- Construct the discrete (finite) subset of Θ, say $\Theta_V = \{\theta_1, \ldots, \theta_M\}$.
- Define the index set $V = \{1, \ldots, M\}$.
- Perform the multiple hyphothesis testing based on samples $(X_i, Y_i)_{i=1}^n$. The estimate index is $v \in V$,



- The construction of set Θ_V should satisfy the condition that

  The successful estimation of $\hat{\theta}$ implies the correct estimation of v.

Step I: Some guidance for construction of Θ_V

- $\{\theta_1, \ldots, \theta_M\}$ cannot be too close.
- For discrete Θ, we can use the trivial reduction $\Theta_V = \Theta$, with a possibly non-uniform prior.
- The construction process could be existence, no need to write Θ_V explicitly.

Step I: Example of 1-sparse regression

Problem Setting:

- Estimate a parameter $\theta \in \mathbb{R}^p$ (have at most one non-zero entry)
- Given n samples $(X_i, Y_i)_{i=1}^n$, where the i-th sample Y_i is a nosiy sample of $\langle X_i, \theta \rangle$.
- The goal is to construct $\hat{\theta}$ such that $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$ is small.

Construction of Θ_V:

- Θ_V is the collection of vectors θ of the form

$$\theta = (0, \ldots, \pm\epsilon, 0, \ldots, 0)$$

  where $\epsilon > 0$ is a constant. Therefore, $M = 2p$.
- It follows that

$$\|\hat{\theta} - \theta_v\|_2 < \frac{\sqrt{2}}{2}\epsilon \implies \operatorname{argmin}_{v'=1,\ldots,M} \|\hat{\theta} - \theta_{v'}\| = v.$$

  Therefore, sufficient estimation of $\hat{\theta}$ implies success in identifying the index v.

Fano's Inequality

Terminology:

- $H(V \mid \hat{V})$
- Error probability $P_e \triangleq \mathbb{P}[\hat{V} \neq V]$

Theorem (Fano's Inequality)

For any discrete random variable V and $\hat{V}$ on a common finite alphabet $\mathcal{V}$,

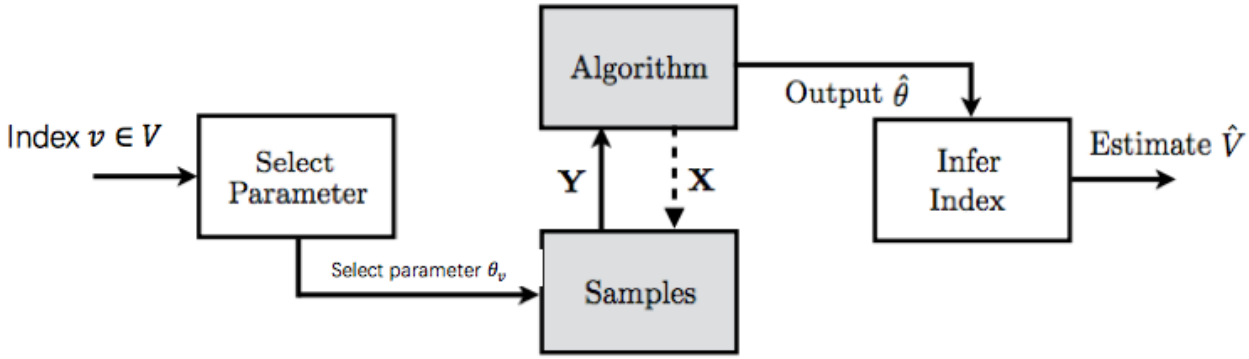
$$H(V \mid \hat{V}) \leq h_2(P_e) + P_e \log(|\mathcal{V}| - 1)$$

In particular, if V is uniform on $\mathcal{V}$, we have

$$I(V; \hat{V}) \geq (1 - P_e) \log |\mathcal{V}| - \log 2$$

Intuition

Step II: Application of Fanos Inequality



- Fano's inequality 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: The learned information is close to prior uncertainty.
- Variations of Fano's inequality

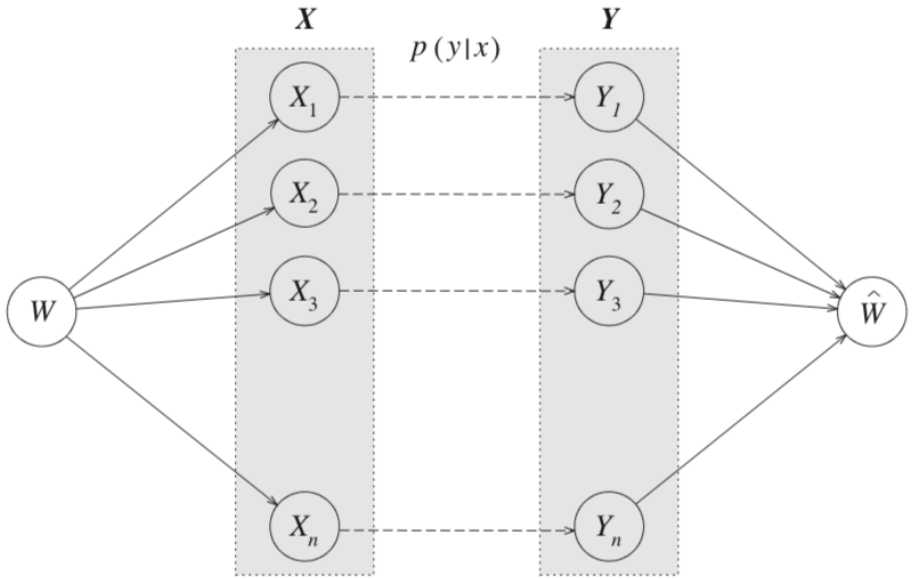
Step III: Upper Bounding the Mutual Information

The key quantity remaining after applying Fanos inequality is $I(V;\hat{V})$

- Data Processing Inequality: (Based on the Markov Chain $V \to Y \to \hat{V}$)
  - No inputs: $I(V;\hat{V}) \le I(V;Y)$
  - Non-adaptive inputs: $I(V;\hat{V} \mid X) \le I(V;Y \mid X)$
  - Adaptive inputs: $I(V;\hat{V}) \le I(V;X,Y)$
- Tensorization: (Based on conditional independence of the samples)
  - No inputs: $I(V;Y) \le \sum_{i=1}^{n} I(V;Y_i)$
  - Non-adaptive inputs: $I(V;Y \mid X) \le \sum_{i=1}^{n} I(V;Y_i \mid X_i)$
  - Adaptive inputs: $I(V;X,Y) \le \sum_{i=1}^{n} I(V;Y_i \mid X_i)$
- KL Divergence Bounds:
  - $I(V;Y) \le \max_{v,v'} D(P_{Y|V}(\cdot \mid v) \| P_{Y|V}(\cdot \mid v'))$
  - $I(V;Y) \le \max_{v} D(P_{Y|V}(\cdot \mid v) \| Q_Y)$ for any Q_Y
  - If each $P_{Y|V}(\cdot \mid v)$ is ϵ-close to the closest $Q_1(y), \ldots, Q_N(y)$ in KL-divergence measure, then $I(V;Y) \le \log N + \epsilon$
  - Similar for the case conditioning on X.

Toy Example I: Noisy-channel coding theorem

- Consider a channel code whose probability of error is arbitrarily small.



  - Apply the Fano's inequality;
  - Upper Bounding the Mutual Information

Therefore, $\frac{\log M}{n} \leq C \triangleq \max_{p(x)} I(X;Y)$.

Toy Example II: M-ary Hyphothesis Testing

- We wish to identify M hyphotheses;
- The v-th hyphothesis is $Y \sim P_v(y)$ for some distribution P_v on $\{0,1\}^n$;
- Apply the Fano's inequality:

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

- Upper Bound the Mutual Information:

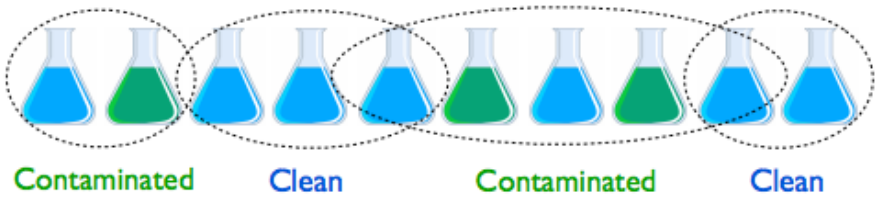
$$I(V;\hat{V}) \leq n \log 2$$

- As a result,

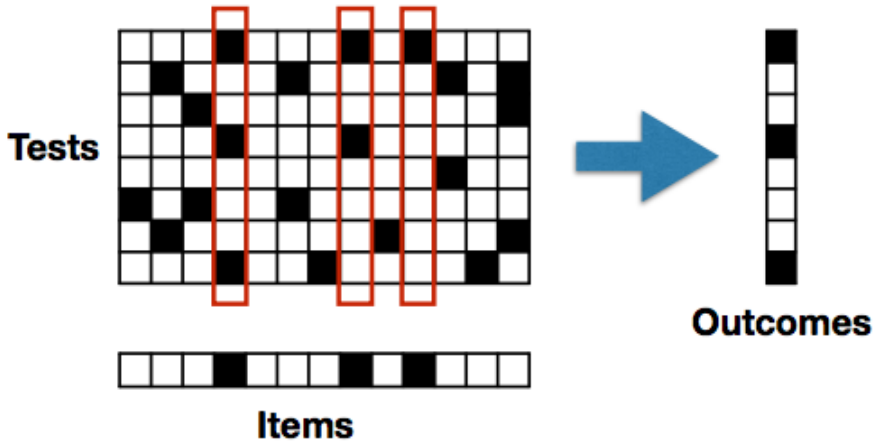
$$P_e \geq 1 - \frac{n+1}{\log_2 M} \implies P_e \leq \delta \text{ requires } n \geq (1-\delta)\log_2 M - 1.$$

Discrete Example I: Group Testing

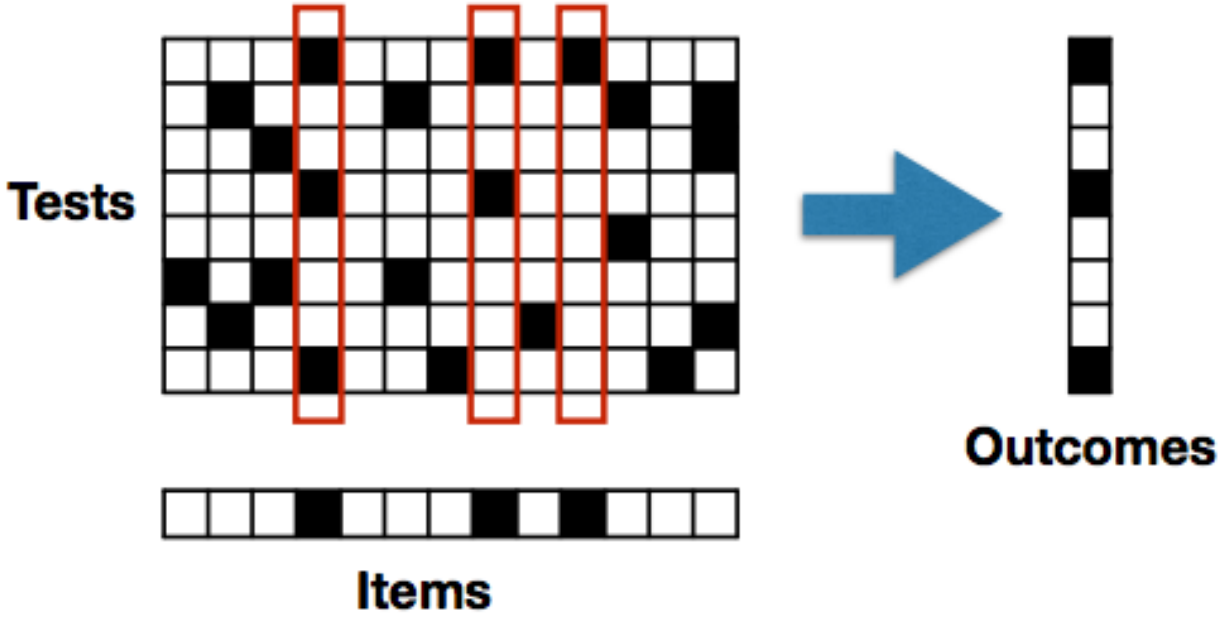
- In a population of p items, there are k unknown defective items;



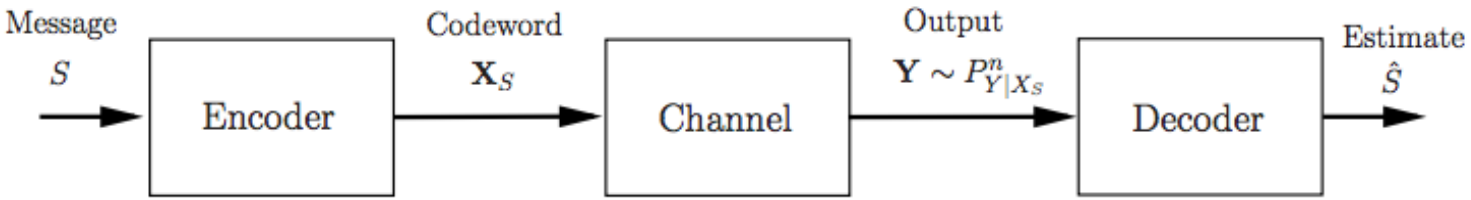
- Defective set $S \subseteq \{1, \ldots, p\}$, which is uniform over sets having cardinality k
- Test matrix $X \in \{0, 1\}^{n \times p}$.
- Given X, we obtain the observation $Y_i = \left(\bigvee_{j \in S} X_{ij} \right) \oplus Z_i$.



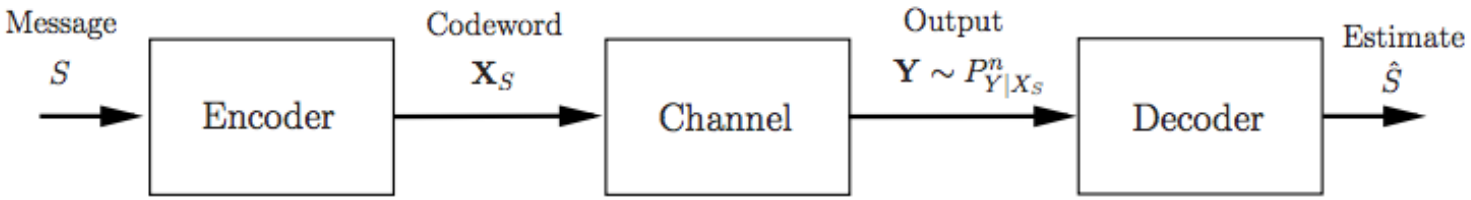
Information Theoretical Viewpoint of Group Testing



- S denotes the defective set
- X_S denotes the columns of X indexed by S.



Information Theoretical Viewpoint of Group Testing



General result:

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

where n^* is the sample complexity; $H(S)$ denotes the model uncertainty; $I(P_{Y|X_S})$ denotes the information learned from measurements.

Converse for Group Testing

- Reduction to multiple hypothesis testing: Select the index set $V = \{1, \ldots, \binom{p}{k}\}$ and $\Theta_v = S$.
- Application of Fanos Inequality:

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

- Bounding the mutual information:
  - Data processing inequality: $I(S; \hat{S} \mid X) \leq I(U; Y)$, where $U_i = \bigvee_{j \in S} X_{ij}$
  - Tensorization: $I(U; Y) \leq \sum_{i=1}^{n} I(U_i; Y_i)$
  - Capacity bound: $I(U_i; Y_i) \leq C$, where C denotes channel capacity

  Final Result:

$$n \leq \frac{\log \binom{p}{k}}{C} \implies \mathbb{P}[\hat{S} \neq S] \text{ cannot be arbitrary small}$$

Discussion

Limitations of Fanos Inequality:

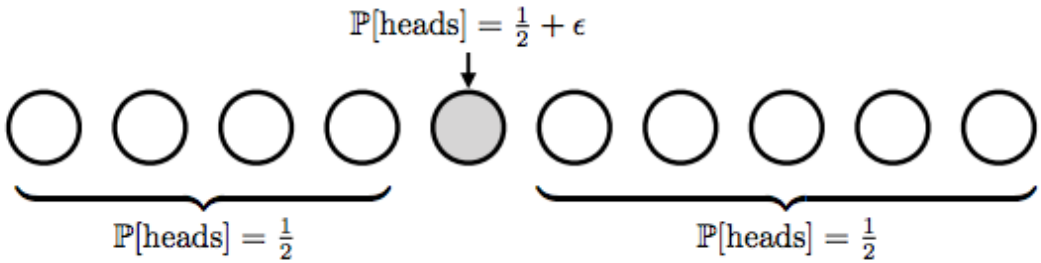
- Non-asymptotic weakness
- Typically hard to tightly bound mutual information in adaptive settings
- Restriction to KL divergence

Generalizations of Fanos Inequality:

- Non-uniform V
- More general f-divergence
- Continuous V

Information Theoretical Gap in Adaptive Settings

- Simple search problem: find the (only) biased coin using few flips

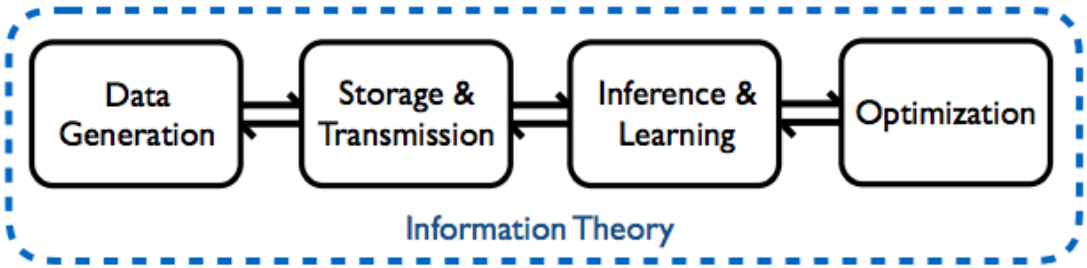


  - Heavy coin $V \in \{1, \ldots, M\}$ uniformly distributed
  - Selected coin at time $i = 1, \ldots, n$ is X_i, the observation is $Y_i \in \{0, 1\}$ (1 for heads)
- Non-adaptive setting:
  - Since X_i and V are independent, $I(V; Y_i \mid X_i) \leq \dfrac{\epsilon^2}{M}$
  - Substituting inot Fano's inequality gives $n \geq \dfrac{M \log M}{\epsilon^2}$

There is an adaptive algorithm that gives $n \geq \dfrac{M}{\epsilon^2}$

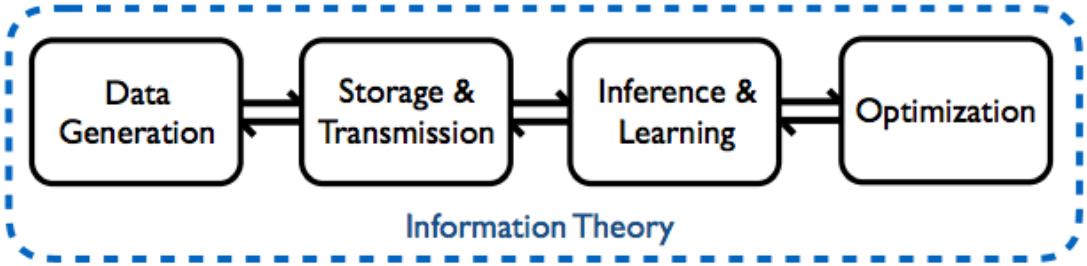
Conclusion

- Information Theory is a theory of data:



Conclusion

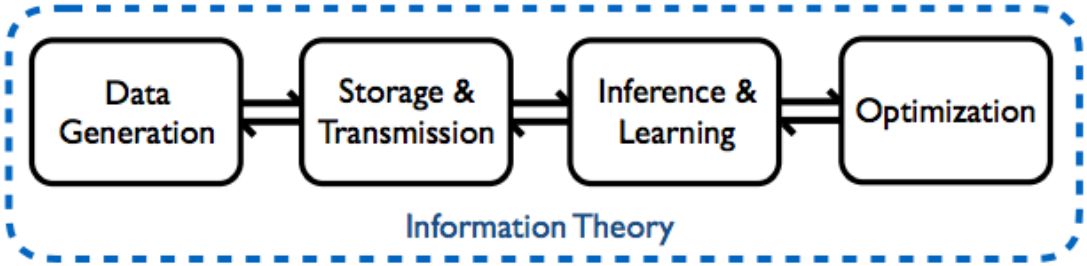
- Information Theory is a theory of data:



- Approach highlighted in this talk:
  - Reduction to multiple hypothesis testing
  - Application of Fanos inequality
  - Bounding the mutual information

Conclusion

- Information Theory is a theory of data:



- Approach highlighted in this talk:
  - Reduction to multiple hypothesis testing
  - Application of Fanos inequality
  - Bounding the mutual information
- Examples:
  - Group Testing
  - Graphical model selection
  - Sparse regression
  - Convex optimization
  - $\cdots$ hopefully many moe to come!