An Introduction to Statistical Lower Bounds for Estimation and Learning

Part 1: Information Theory and Fano's Inequality

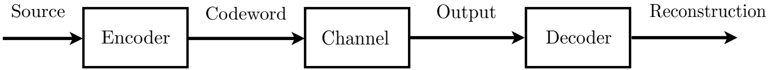
Jonathan Scarlett

National University of Singapore

Annual School on Mathematics of Data Science
[Darwin, 2024]

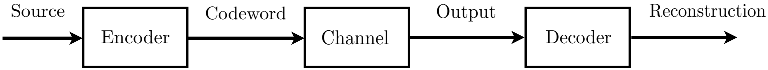
Information Theory

- 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

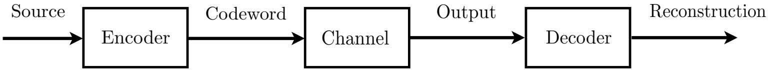


  - Information of source: Entropy
  - Information learned at channel output: Mutual information

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

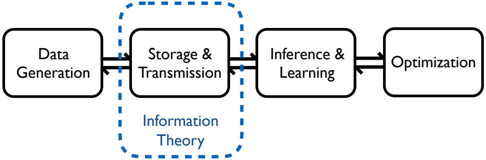
Principles:

- 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:

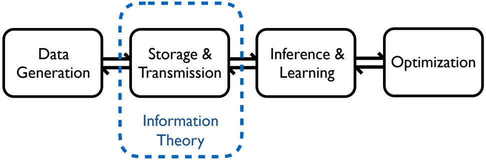
Information theory is a theory of communication



Information Theory and Data

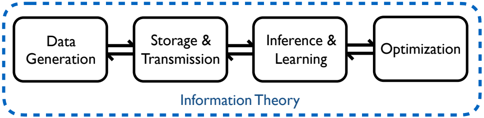
- Conventional view:

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
Feedback	Active learning / adaptivity
Rate-distortion theory	Approximate recovery
Joint source-channel coding	Non-uniform prior
...	...

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

Terminology: Achievability and Converse

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

- Estimation error: $\|\hat{\theta} - \theta_{\mathrm{true}}\| \le \epsilon$
- Optimization error: $f(x_{\mathrm{selected}}) \le \min_x f(x) + \epsilon$

Terminology: Achievability and Converse

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

- Estimation error: $\|\hat{\theta} - \theta_{\rm true}\| \leq \epsilon$
- Optimization error: $f(x_{\rm selected}) \leq \min_x f(x) + \epsilon$

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

Information Measures

Entropy

- **Definition:** The entropy of a discrete random variable X is defined as

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} = \mathbb{E}\Big[\log \frac{1}{P_X(X)}\Big].$$

This is measured in bits for $\log_2(\cdot)$, or nats for $\log_e(\cdot)$.

  - Interpetation: If we observe that $X = x$ then the amount of information learned is $\log \frac{1}{P_X(x)}$ ($\log \frac{1}{p}$ satisfies natural axioms). Entropy is the average information learned by observing X, or equivalently, the average uncertainty in X before observing it.
  - Examples: (i) If X is deterministic then $H(X) = 0$;
    (ii) If $X \sim \mathrm{Uniform}(\mathcal{X})$ then $H(X) = \log|\mathcal{X}|$
  - Source coding theorem: $H(X)$ is the fundamental limit of compression when a source emits i.i.d. symbols from P_X

Entropy

- **Definition:** The entropy of a discrete random variable X is defined as

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} = \mathbb{E}\Big[\log \frac{1}{P_X(X)} \Big].$$

This is measured in bits for $\log_2(\cdot)$, or nats for $\log_e(\cdot)$.

  - Interpetation: If we observe that $X = x$ then the amount of information learned is $\log \frac{1}{P_X(x)}$ ($\log \frac{1}{p}$ satisfies natural axioms). Entropy is the average information learned by observing X, or equivalently, the average uncertainty in X before observing it.
  - Examples: (i) If X is deterministic then $H(X) = 0$;
    (ii) If $X \sim \mathrm{Uniform}(\mathcal{X})$ then $H(X) = \log |\mathcal{X}|$
  - Source coding theorem: $H(X)$ is the fundamental limit of compression when a source emits i.i.d. symbols from P_X
- **Joint version:** $H(X, Y) = \mathbb{E}\big[\log \frac{1}{P_{XY}(X,Y)} \big]$; generally $H(\mathbf{X}) = \mathbb{E}\big[\log \frac{1}{P_{\mathbf{X}}(\mathbf{X})} \big]$.
  - Interpetation: Overall information/uncertainty in multiple variables
- **Conditional version:** $H(Y|X) = \sum_{x \in \mathcal{X}} P_X(x) H(Y|X = x)$
  - Interpetation: Remaining uncertainty in Y after observing X (on average)
- **Continuous RVs:** A counterpart exists for continuous RVs, but not as "well-behaved" (can be negative, no longer invariant under 1-to-1 maps)

Properties of Entropy

- **Non-negativity:**

$$H(X) \geq 0$$

  with equality iff X is deterministic

- **Uniform distribution has highest entropy:**

$$H(X) \leq \log |\mathcal{X}|$$

  with equality iff X is uniform

- **Conditioning can't increase entropy:** *(on average)*

$$H(X|Y) \leq H(X)$$

  with equality iff X and Y are independent

- **Chain rule:**

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$H(X_1, \dots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1})$$

- **Tensorization / sub-additivity:**

$$H(X_1, \dots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Relative Entropy (KL Divergence)

- **Definition:** For two distributions P and Q, the relative entropy (KL divergence) is defined as
$$D(P\|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)} = \mathbb{E}_P\Big[\log\frac{P(X)}{Q(X)}\Big]$$
  - Example usage: If we draw n i.i.d. samples from Q, the probability of getting symbol proportions P is roughly $e^{-nD(P\|Q)}$ (a more general statement: Sanov's theorem)
- **Key property:**
$$D(P\|Q) \ge 0$$
with equality iff $P = Q$
- **Conditional version:** $D(P_{Y|X}\|Q_{Y|X}|P_X) = \sum_x P_X(x) D(P_{Y|X=x}\|Q_{Y|X=x})$
  - This also leads to a chain rule: $D(P_{XY}\|Q_{XY}) = D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X)$
- Extends readily to continuous RVs (and beyond) while saying "well-behaved"

Mutual Information

- **Definition:** The mutual information between X and Y is defined as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= D(P_{XY} \| P_X \times P_Y) \end{aligned}$$

  - Interpretation 1: X has uncertainty $H(X)$, but after observing Y it has remaining uncertainty $H(X|Y)$, so $I(X;Y)$ is how much information Y revealed about X.
  - Interpretation 2: By the $D(P_{XY} \| P_X \times P_Y)$ form, this measures how far X and Y are from being independent
  - Channel coding theorem: $\max_{P_X} I(X;Y)$ is the fundamental limit of communication when the communication channel is probabilistic with transition law $P_{Y|X}$

Mutual Information

- **Definition:** The mutual information between X and Y is defined as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= D(P_{XY} \| P_X \times P_Y) \end{aligned}$$

  - Interpretation 1: X has uncertainty $H(X)$, but after observing Y it has remaining uncertainty $H(X|Y)$, so $I(X;Y)$ is how much information Y revealed about X.
  - Interpretation 2: By the $D(P_{XY} \| P_X \times P_Y)$ form, this measures how far X and Y are from being independent
  - Channel coding theorem: $\max_{P_X} I(X;Y)$ is the fundamental limit of communication when the communication channel is probabilistic with transition law $P_{Y|X}$
- Can again have joint version, e.g., $I(X_1, X_2; Y_1, Y_2)$, and conditional version, e.g., $I(X;Y|Z) = \sum_z P_Z(z) I(X;Y|Z=z)$
- Again well-behaved even for continuous variables

Properties of Mutual Information

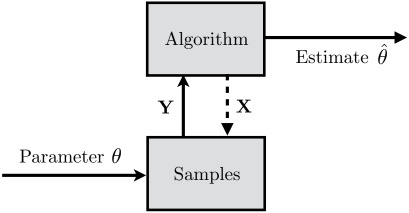
- **Non-negativity:**
$$I(X;Y) \geq 0$$
with equality iff X and Y are independent
- **Chain rule:**
$$I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y | X_1)$$
and similarly with n variables
- **Tensorization:** If $P_{\mathbf{Y}|\mathbf{X}} = \prod_{i=1}^n P_{Y_i|X_i}$, then
$$I(\mathbf{X}; \mathbf{Y}) \leq \sum_{i=1}^n I(X_i; Y_i)$$
(not true in general if the assumption on $P_{\mathbf{Y}|\mathbf{X}}$ is dropped)
- **Data processing inequality:** If $X \to Y \to Z$ forms a Markov chain, then
$$I(X;Z) \leq I(X;Y).$$
Similarly with more variables (e.g., $W \to X \to Y \to Z$ gives $I(W;Z) \leq I(X;Y)$)

Converse Bounds for Statistical Estimation via Fano's Inequality

(Based on survey chapter https://arxiv.org/abs/1901.00555)

Statistical Estimation

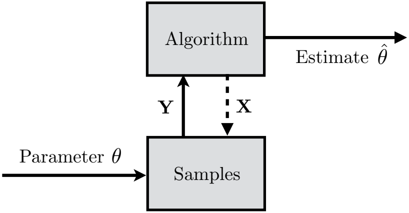
- **General statistical estimation setup:**
  - Unknown parameter $\theta \in \Theta$
  - Samples $\mathbf{Y} = (Y_1, \dots, Y_n)$ drawn from $P_\theta(\mathbf{y})$
    - More generally, from $P_{\theta,\mathbf{X}}$ with inputs $\mathbf{X} = (X_1, \dots, X_n)$
  - Given $\mathbf{Y}$ (and possibly $\mathbf{X}$), construct estimate $\hat{\theta}$



Statistical Estimation

- **General statistical estimation setup:**
  - Unknown parameter $\theta \in \Theta$
  - Samples $\mathbf{Y} = (Y_1, \dots, Y_n)$ drawn from $P_\theta(\mathbf{y})$
    - More generally, from $P_{\theta,\mathbf{X}}$ with inputs $\mathbf{X} = (X_1, \dots, X_n)$
  - Given $\mathbf{Y}$ (and possibly $\mathbf{X}$), construct estimate $\hat{\theta}$

- **Goal.** Minimize some loss $\ell(\theta, \hat{\theta})$
  - 0-1 loss: $\ell(\theta, \hat{\theta}) = \mathbf{1}\{\hat{\theta} \neq \theta\}$
  - Squared ℓ_2 loss: $\|\theta - \hat{\theta}\|^2$

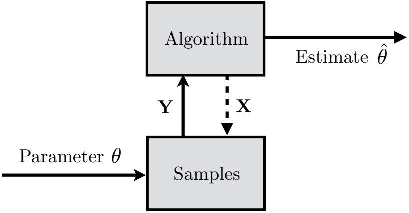


Statistical Estimation

- **General statistical estimation setup:**
  - Unknown parameter $\theta \in \Theta$
  - Samples $\mathbf{Y} = (Y_1, \dots, Y_n)$ drawn from $P_\theta(\mathbf{y})$
    - More generally, from $P_{\theta,\mathbf{X}}$ with inputs $\mathbf{X} = (X_1, \dots, X_n)$
  - Given $\mathbf{Y}$ (and possibly $\mathbf{X}$), construct estimate $\hat{\theta}$

- **Goal.** Minimize some loss $\ell(\theta, \hat{\theta})$
  - 0-1 loss: $\ell(\theta, \hat{\theta}) = \mathbf{1}\{\hat{\theta} \neq \theta\}$
  - Squared ℓ_2 loss: $\|\theta - \hat{\theta}\|^2$

- **Typical example.** Linear regression
  - Estimate $\theta \in \mathbb{R}^p$ from $\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}$



Defining Features

- There are many properties that impact the analysis:
  - Discrete θ (e.g., graph learning, sparsity pattern recovery)
  - Continuous θ (e.g., regression, density estimation)
  - Bayesian θ (average-case performance)
  - Minimax bounds over Θ (worst-case performance)
  - Non-adaptive inputs (all $X_1, \dots, X_n$ chosen in advance)
  - Adaptive inputs (X_i can be chosen based on $Y_1, \dots, Y_{i-1}$)

- **This talk.** Minimax bounds, mostly non-adaptive, first discrete and then continuous

High-Level Steps

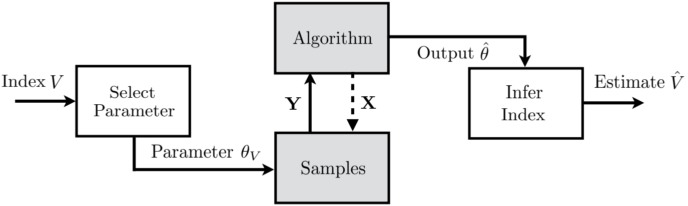
Steps in attaining a minimax lower bound (converse):

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.)

Step I: Reduction to Multiple Hypothesis Testing

- Lower bound worst-case error by average over hard subset $\theta_1, \dots, \theta_M$:

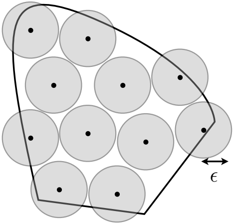


Idea:

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

Step I: Example

- **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.*

Step II: Application of Fano's Inequality

• Standard form of Fano's inequality from textbooks: For a random variable V and its estimate $\hat{V}$, defining $P_{\mathrm{e}} = \mathbb{P}[\hat{V} \neq V]$, we have

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

where $H_2(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha)\log\frac{1}{1-\alpha}$ is the entropy of $\mathrm{Bernoulli}(\alpha)$.

Intuition:

- Considering asking questions to resolve the uncertainty in V given $\hat{V}$?.
- First ask whether the two are equal; this has uncertainty $H_2(P_{\mathrm{e}})$
- When they differ, the remaining uncertainty is at most $\log\big(M-1\big)$.

Step II: Application of Fano's Inequality

• Standard form of Fano's inequality from textbooks: For a random variable V and its estimate $\hat{V}$, defining $P_{\rm e} = \mathbb{P}[\hat{V} \neq V]$, we have

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

where $H_2(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha)\log\frac{1}{1-\alpha}$ is the entropy of $\mathrm{Bernoulli}(\alpha)$.

Intuition:

- Considering asking questions to resolve the uncertainty in V given $\hat{V}$?.
- First ask whether the two are equal; this has uncertainty $H_2(P_{\rm e})$
- When they differ, the remaining uncertainty is at most $\log\big(M-1\big)$.

• Re-arranged and slightly weakened form for V uniform over M outcomes:

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

- Intuition: Need learned information $I(V;\hat{V})$ to be close to prior uncertainty $\log M$

Step II: Application of Fano's Inequality

- Standard form of Fano's inequality from textbooks: For a random variable V and its estimate $\hat{V}$, defining $P_{\mathrm{e}} = \mathbb{P}[\hat{V} \neq V]$, we have

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

where $H_2(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha)\log\frac{1}{1-\alpha}$ is the entropy of $\mathrm{Bernoulli}(\alpha)$.

Intuition:

  - Considering asking questions to resolve the uncertainty in V given $\hat{V}$?.
  - First ask whether the two are equal; this has uncertainty $H_2(P_{\mathrm{e}})$
  - When they differ, the remaining uncertainty is at most $\log\big(M-1\big)$.

- Re-arranged and slightly weakened form for V uniform over M outcomes:

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

  - Intuition: Need learned information $I(V;\hat{V})$ to be close to prior uncertainty $\log M$

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

Step III: Bounding the Mutual Information

- The key quantity remaining after applying Fano's inequality is $I(V; \hat{V})$
- **Data processing inequality:** (Based on $V \to \mathbf{Y} \to \hat{V}$ or similar)
  - No inputs: $I(V; \hat{V}) \le I(V; \mathbf{Y})$
  - Non-adaptive inputs: $I(V; \hat{V} | \mathbf{X}) \le I(V; \mathbf{Y} | \mathbf{X})$
  - Adaptive inputs: $I(V; \hat{V}) \le I(V; \mathbf{X}, \mathbf{Y})$

Step III: Bounding the Mutual Information

- The key quantity remaining after applying Fano's inequality is $I(V;\hat{V})$
- **Data processing inequality:** (Based on $V \to \mathbf{Y} \to \hat{V}$ or similar)
  - No inputs: $I(V;\hat{V}) \le I(V;\mathbf{Y})$
  - Non-adaptive inputs: $I(V;\hat{V}|\mathbf{X}) \le I(V;\mathbf{Y}|\mathbf{X})$
  - Adaptive inputs: $I(V;\hat{V}) \le I(V;\mathbf{X},\mathbf{Y})$
- **Tensorization:** (Based on conditional independence of the samples)
  - No inputs: $I(V;\mathbf{Y}) \le \sum_{i=1}^n I(V;Y_i)$
  - Non-adaptive inputs: $I(V;\mathbf{Y}|\mathbf{X}) \le \sum_{i=1}^n I(V;Y_i|X_i)$
  - Adaptive inputs: $I(V;\mathbf{X},\mathbf{Y}) \le \sum_{i=1}^n I(V;Y_i|X_i)$

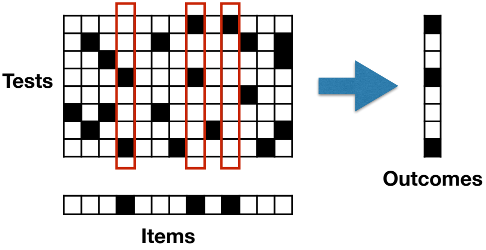
Step III: Bounding the Mutual Information

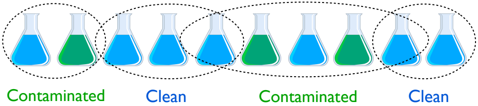
- The key quantity remaining after applying Fano's inequality is $I(V;\hat{V})$
- **Data processing inequality:** (Based on $V \to \mathbf{Y} \to \hat{V}$ or similar)
  - No inputs: $I(V;\hat{V}) \le I(V;\mathbf{Y})$
  - Non-adaptive inputs: $I(V;\hat{V}|\mathbf{X}) \le I(V;\mathbf{Y}|\mathbf{X})$
  - Adaptive inputs: $I(V;\hat{V}) \le I(V;\mathbf{X},\mathbf{Y})$
- **Tensorization:** (Based on conditional independence of the samples)
  - No inputs: $I(V;\mathbf{Y}) \le \sum_{i=1}^{n} I(V;Y_i)$
  - Non-adaptive inputs: $I(V;\mathbf{Y}|\mathbf{X}) \le \sum_{i=1}^{n} I(V;Y_i|X_i)$
  - Adaptive inputs: $I(V;\mathbf{X},\mathbf{Y}) \le \sum_{i=1}^{n} I(V;Y_i|X_i)$
- **KL Divergence Bounds:**
  - $I(V;Y) \le \max_{v,v'} D\big(P_{Y|V}(\cdot|v) \| P_{Y|V}(\cdot|v')\big)$
  - $I(V;Y) \le \max_{v} D\big(P_{Y|V}(\cdot|v) \| Q_Y\big)$ for any Q_Y
  - If each $P_{Y|V}(\cdot|v)$ is ϵ-close to the closest $Q_1(\mathbf{y}),\dots,Q_N(\mathbf{y})$ in KL divergence, then $I(V;Y) \le \log N + \epsilon$
  - (Similarly with conditioning on X)

Discrete Example 1

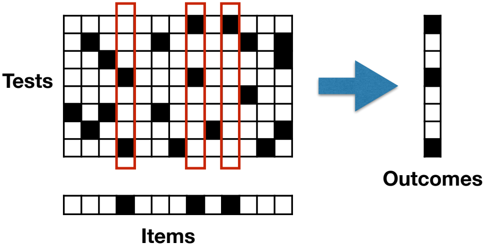
Group Testing

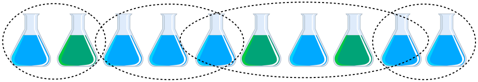
Group Testing





Group Testing



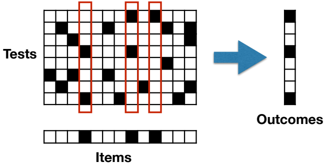


- **Goal:**

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

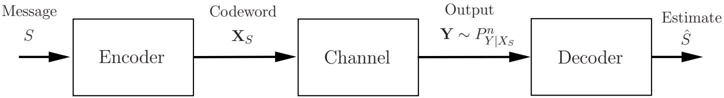
- **Sample complexity:** Required number of tests n

Information Theory and Group Testing



- **Information-theoretic viewpoint:**

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



- **Example formulation of general result:**

Entropy

(Model uncertainty)

Number of tests

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

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:**

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

where $p = (\#\text{items})$ and $k = (\#\text{defectives})$.

Converse via Fano's Inequality

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

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

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

where $p = (\#\text{items})$ and $k = (\#\text{defectives})$.

- **Bounding the mutual information:**
  - Data processing inequality: $I(S; \hat{S} | \mathbf{X}) \leq I(\mathbf{U}; \mathbf{Y})$ where $\mathbf{U}$ are pre-noise outputs
  - Tensorization: $I(\mathbf{U}; \mathbf{Y}) \leq \sum_{i=1}^{n} I(U_i; Y_i)$
  - Capacity bound: $I(U_i; Y_i) \leq C$ if outcome passed through channel of capacity C

Converse via Fano's Inequality

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

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

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

where $p = (\#\text{items})$ and $k = (\#\text{defectives})$.

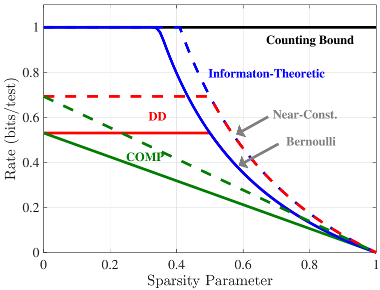
- **Bounding the mutual information:**
  - Data processing inequality: $I(S; \hat{S} | \mathbf{X}) \leq I(\mathbf{U}; \mathbf{Y})$ where $\mathbf{U}$ are pre-noise outputs
  - Tensorization: $I(\mathbf{U}; \mathbf{Y}) \leq \sum_{i=1}^{n} I(U_i; Y_i)$
  - Capacity bound: $I(U_i; Y_i) \leq C$ if outcome passed through channel of capacity C

- **Final result:**

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

Illustration of Bounds

Noiseless bounds:



Noisy bounds:

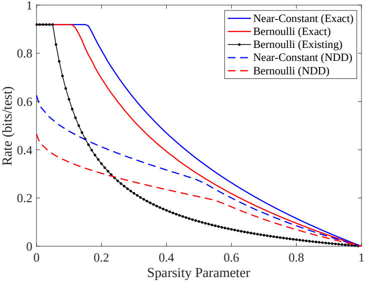
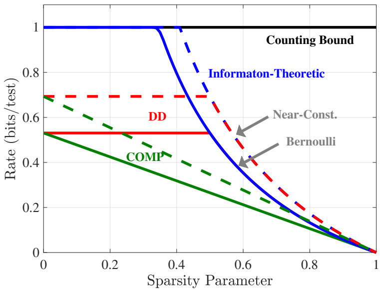
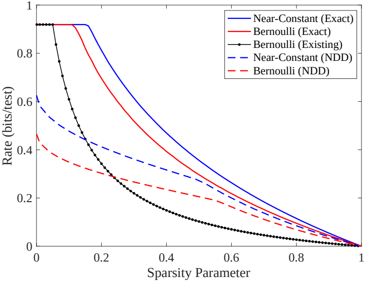


Illustration of Bounds

Noiseless bounds:



Noisy bounds:



- **Other Implications:**
  - Adaptivity and approximate recovery:
    - No gain at low sparsity levels
    - Significant gain at high sparsity levels
  - Information-theoretically optimal non-adaptive algorithms are now known

Discrete Example 2

Graphical Model Selection

Graphical Model Representations of Joint Distributions

Motivating example:

- In a population of p people, let

$$Y_i = \begin{cases} 1 & \text{person } i \text{ is infected} \\ -1 & \text{person } i \text{ is healthy,} \end{cases} \quad i = 1, \dots, p$$

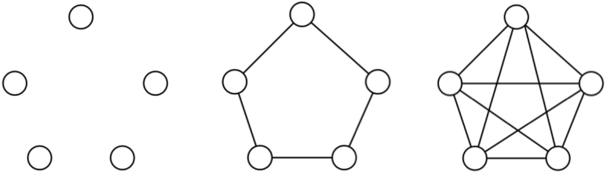
Graphical Model Representations of Joint Distributions

Motivating example:

- In a population of p people, let

$$Y_i = \begin{cases} 1 & \text{person } i \text{ is infected} \\ -1 & \text{person } i \text{ is healthy,} \end{cases} \qquad i = 1, \dots, p$$

- Example models:



[Abbe and Wainwright, ISIT Tutorial, 2015]

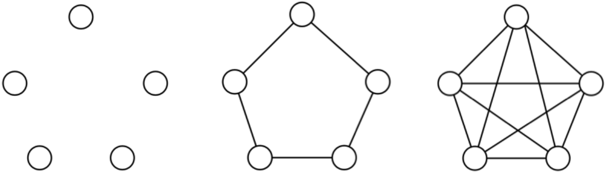
Graphical Model Representations of Joint Distributions

Motivating example:

- In a population of p people, let

$$Y_i = \begin{cases} 1 & \text{person } i \text{ is infected} \\ -1 & \text{person } i \text{ is healthy,} \end{cases} \qquad i = 1, \dots, p$$

- Example models:



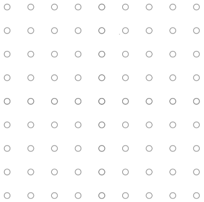
[Abbe and Wainwright, ISIT Tutorial, 2015]

- Joint distribution for a given graph $G = (V, E)$:

$$\mathbb{P}\big[(Y_1, \dots, Y_p) = (y_1, \dots, y_p)\big] = \frac{1}{Z} \exp\Bigg(\sum_{(i,j) \in E} \lambda_{ij} y_i y_j \Bigg)$$

Graphical Model Selection: Illustration

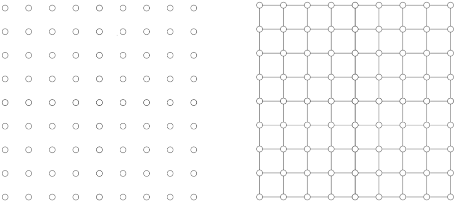
- A larger example from [Abbe and Wainwright, ISIT Tutorial 2015]:
  - Example graphs:



  - Sample images (Ising model):

Graphical Model Selection: Illustration

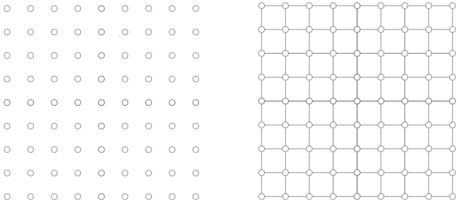
- A larger example from [Abbe and Wainwright, ISIT Tutorial 2015]:
  - Example graphs:



  - Sample images (Ising model):

Graphical Model Selection: Illustration

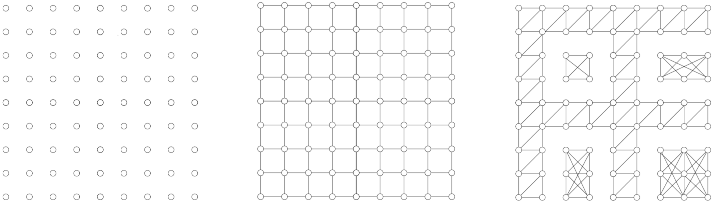
- A larger example from [Abbe and Wainwright, ISIT Tutorial 2015]:
  - Example graphs:



  - Sample images (Ising model):

Graphical Model Selection: Illustration

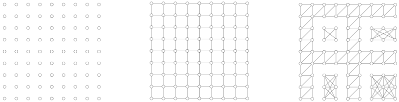
- A larger example from [Abbe and Wainwright, ISIT Tutorial 2015]:
  - Example graphs:



  - Sample images (Ising model):

Graphical Model Selection: Illustration

- A larger example from [Abbe and Wainwright, ISIT Tutorial, 2015]:
  - Example graphs:



  - Sample images (Ising model):

- **Goal:** Identify graph given n independent samples

Graphical Model Selection: Definition

General problem statement.

- Given n i.i.d. samples of $(Y_1, \dots, Y_p) \sim P_G$, recover the underlying graph G
  - Applications: Statistical physics, social and biological networks
- Error probability:

$$P_{\rm e} = \max_{G \in \mathcal{G}} \mathbb{P}[\hat{G} \neq G \mid G].$$

Graphical Model Selection: Definition

General problem statement.

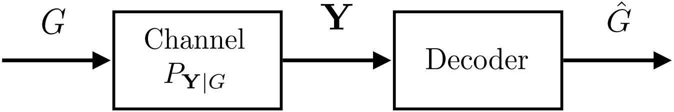
- Given n i.i.d. samples of $(Y_1, \dots, Y_p) \sim P_G$, recover the underlying graph G
  - Applications: Statistical physics, social and biological networks
- Error probability:
$$P_{\mathrm{e}} = \max_{G \in \mathcal{G}} \mathbb{P}[\hat{G} \neq G \mid G].$$

Assumptions.

- Distribution class:
  - Ising model
$$P_G(x_1, \dots, x_p) = \frac{1}{Z} \exp\bigg(\sum_{(i,j) \in E} \lambda_{ij} x_i x_j \bigg)$$
  - Gaussian model
$$(X_1, \dots, X_p) \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$
  where $(\mathbf{\Sigma}^{-1})_{ij} \neq 0 \iff (i,j) \in E$ [Hammersley-Clifford theorem]
- Graph class:
  - Bounded-edge (at most k edges total)
  - Bounded-degree (at most d edges out of each node)

Information-Theoretic Viewpoint

- **Information-theoretic viewpoint:**



Converse via Fano's Inequality

- **Reduction to multiple hypothesis testing:** Let G be uniform on hard subset $\mathcal{G}_0 \subseteq \mathcal{G}$
  - Ideally many graphs (lots of graphs to distinguish)
  - Ideally close together (harder to distinguish)

Converse via Fano's Inequality

- **Reduction to multiple hypothesis testing:** Let G be uniform on hard subset $\mathcal{G}_0 \subseteq \mathcal{G}$
  - Ideally many graphs (lots of graphs to distinguish)
  - Ideally close together (harder to distinguish)

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

$$\mathbb{P}[\hat{G} \neq G] \geq 1 - \frac{I(G; \hat{G}) + \log 2}{\log |\mathcal{G}_0|}$$

Converse via Fano's Inequality

- **Reduction to multiple hypothesis testing:** Let G be uniform on hard subset $\mathcal{G}_0 \subseteq \mathcal{G}$
  - Ideally many graphs (lots of graphs to distinguish)
  - Ideally close together (harder to distinguish)

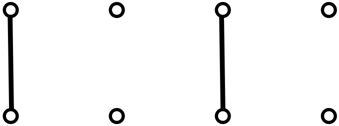
- **Application of Fano's Inequality:**

$$\mathbb{P}[\hat{G} \neq G] \geq 1 - \frac{I(G;\hat{G}) + \log 2}{\log |\mathcal{G}_0|}$$

- **Bounding the mutual information:**
  - Data processing inequality: $I(G;\hat{G}) \leq I(G;\mathbf{Y})$
  - Tensorization: $I(G;\mathbf{Y}) \leq \sum_{i=1}^{n} I(G;Y_i)$
  - KL divergence bound: Bound $I(G;Y_i) \leq \max_G D(P_{Y|G}(\cdot|G) \| Q_Y)$ case-by-case

Graph Ensembles

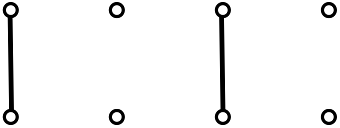
- Graphs that are difficult to distinguish from the empty graph:



  - Reveals $n = \Omega\big(\frac{1}{\lambda^2}\log p\big)$ necessary condition with "edge strength" λ and p nodes

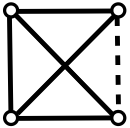
Graph Ensembles

- Graphs that are difficult to distinguish from the empty graph:



  - Reveals $n = \Omega\big(\frac{1}{\lambda^2}\log p\big)$ necessary condition with "edge strength" λ and p nodes

- Graphs that are difficult to distinguish from the complete (sub-)graph:



  - Reveals $n = \Omega\big(e^{\lambda d}\big)$ necessary condition with "edge strength" λ and degree d

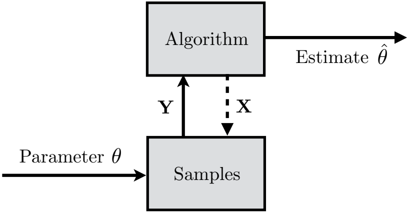
Upper vs. Lower Bounds

- Example results with maximal degree d, edge strength λ (slightly informal):
  - (Converse) $n = \Omega\big(\max\big\{ \frac{1}{\lambda^2}, e^{\lambda d} \big\} \log p \big)$ [Santhanam and Wainwright, 2012]
  - (Achievability) $n = O\big(\max\big\{ \frac{1}{\lambda^2}, e^{\lambda d} \big\} d \log p \big)$ [Santhanam and Wainwright, 2012]
  - (Early Practical) $n = O(d^2 \log p)$ but extra assumptions that are hard to cerify [Ravikumar/Wainwright/Lafferty, 2010]
  - (Further Practical) $n = O\big(\frac{d^2 e^{\lambda d}}{\lambda^2} \log p \big)$ [Klivans/Meka 2017] [Wu/Sanghavi/Dimakis 2018]
  - (Near-Optimality in Many Regimes)
    - Ising models [Lokhov/Vuffray/Misra/Chertkov, 2018]
    - Gaussian models [Misra/Vuffray/Lokhov, 2020]

What About Continuous-Valued Estimation?

Statistical Estimation

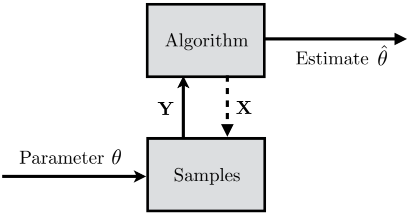
- **General statistical estimation setup:**
  - Unknown parameter $\theta \in \Theta$
  - Samples $\mathbf{Y} = (Y_1, \dots, Y_n)$ drawn from $P_\theta(\mathbf{y})$
    - More generally, from $P_{\theta,\mathbf{X}}$ with inputs $\mathbf{X} = (X_1, \dots, X_n)$
  - Given $\mathbf{Y}$ (and possibly $\mathbf{X}$), construct estimate $\hat{\theta}$



Statistical Estimation

- **General statistical estimation setup:**
  - Unknown parameter $\theta \in \Theta$
  - Samples $\mathbf{Y} = (Y_1, \dots, Y_n)$ drawn from $P_\theta(\mathbf{y})$
    - More generally, from $P_{\theta,\mathbf{X}}$ with inputs $\mathbf{X} = (X_1, \dots, X_n)$
  - Given $\mathbf{Y}$ (and possibly $\mathbf{X}$), construct estimate $\hat{\theta}$

- **Goal.** Minimize some loss $\ell(\theta, \hat{\theta})$
  - 0-1 loss: $\ell(\theta, \hat{\theta}) = \mathbf{1}\{\hat{\theta} \neq \theta\}$
  - Squared ℓ_2 loss: $\|\theta - \hat{\theta}\|^2$

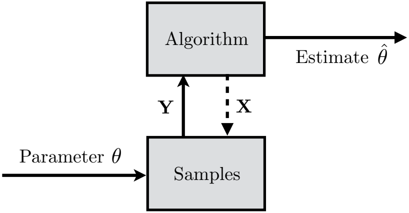


Statistical Estimation

- **General statistical estimation setup:**
  - Unknown parameter $\theta \in \Theta$
  - Samples $\mathbf{Y} = (Y_1, \dots, Y_n)$ drawn from $P_\theta(\mathbf{y})$
    - More generally, from $P_{\theta,\mathbf{X}}$ with inputs $\mathbf{X} = (X_1, \dots, X_n)$
  - Given $\mathbf{Y}$ (and possibly $\mathbf{X}$), construct estimate $\hat{\theta}$

- **Goal.** Minimize some loss $\ell(\theta, \hat{\theta})$
  - 0-1 loss: $\ell(\theta, \hat{\theta}) = \mathbf{1}\{\hat{\theta} \neq \theta\}$
  - Squared ℓ_2 loss: $\|\theta - \hat{\theta}\|^2$

- **Typical example.** Linear regression
  - Estimate $\theta \in \mathbb{R}^p$ from $\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}$



Minimax Risk

- Since the samples are random, so is $\hat{\theta}$ and hence $\ell(\theta, \hat{\theta})$
- So seek to minimize the average loss $\mathbb{E}_\theta[\ell(\theta, \hat{\theta})]$.
  - Note: $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ denote averages w.r.t. $\mathbf{Y}$ when the true parameter is θ.

Minimax Risk

- Since the samples are random, so is $\hat{\theta}$ and hence $\ell(\theta, \hat{\theta})$

- So seek to minimize the average loss $\mathbb{E}_\theta[\ell(\theta, \hat{\theta})]$.
  - Note: $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ denote averages w.r.t. $\mathbf{Y}$ when the true parameter is θ.

- **Minimax risk:**

$$\mathcal{M}_n(\Theta, \ell) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\ell(\theta, \hat{\theta})\right],$$

i.e., worst case average loss over all $\theta \in \Theta$

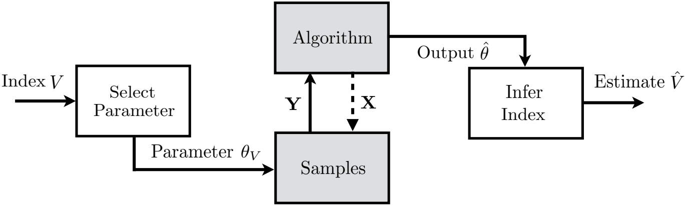
Minimax Risk

- Since the samples are random, so is $\hat{\theta}$ and hence $\ell(\theta, \hat{\theta})$
- So seek to minimize the average loss $\mathbb{E}_\theta[\ell(\theta, \hat{\theta})]$.
  - Note: $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ denote averages w.r.t. **Y** when the true parameter is θ.
- **Minimax risk:**

$$\mathcal{M}_n(\Theta, \ell) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \big[\ell(\theta, \hat{\theta})\big],$$

i.e., worst case average loss over all $\theta \in \Theta$

- **Approach:** Lower bound worst-case error by average over hard subset $\theta_1, \dots, \theta_M$:



General Lower Bound via Fano's Inequality

- To get a meaningful result, need a sufficiently "well-behaved" loss function. Subsequently, focus on loss functions of the form

$$\ell(\theta, \hat{\theta}) = \Phi\big(\rho(\theta, \hat{\theta})\big)$$

where $\rho(\theta, \theta')$ is some metric, and $\Phi(\cdot)$ is some non-negative and increasing function (e.g., $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$)

General Lower Bound via Fano's Inequality

- To get a meaningful result, need a sufficiently "well-behaved" loss function. Subsequently, focus on loss functions of the form

$$\ell(\theta, \hat{\theta}) = \Phi\big(\rho(\theta, \hat{\theta})\big)$$

where $\rho(\theta, \theta')$ is some metric, and $\Phi(\cdot)$ is some non-negative and increasing function (e.g., $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$)

- **Claim.** Fix $\epsilon > 0$, and let $\{\theta_1, \dots, \theta_M\}$ be a finite subset of Θ such that

$$\rho(\theta_v, \theta_{v'}) \geq \epsilon, \quad \forall v, v' \in \{1, \dots, M\}, v \neq v'.$$

Then, we have

$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\Big(\frac{\epsilon}{2}\Big)\bigg(1 - \frac{I(V; \mathbf{Y}) + \log 2}{\log M}\bigg),$$

where V is uniform on $\{1, \dots, M\}$, and $I(V; \mathbf{Y})$ is with respect to $V \to \theta_V \to \mathbf{Y}$.

Proof of General Lower Bound

- Using Markov's inequality:

$$\begin{aligned}\sup_{\theta \in \Theta} \mathbb{E}_\theta\big[\ell(\theta, \hat{\theta})\big] &\geq \sup_{\theta \in \Theta} \Phi(\epsilon_0) \mathbb{P}_\theta[\ell(\theta, \hat{\theta}) \geq \Phi(\epsilon_0)] \\ &= \Phi(\epsilon_0) \sup_{\theta \in \Theta} \mathbb{P}_\theta[\rho(\theta, \hat{\theta}) \geq \epsilon_0]\end{aligned}$$

Proof of General Lower Bound

- Using Markov's inequality:

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta\big[\ell(\theta,\hat{\theta})\big] \ge \sup_{\theta\in\Theta} \Phi(\epsilon_0)\mathbb{P}_\theta[\ell(\theta,\hat{\theta}) \ge \Phi(\epsilon_0)]$$
$$= \Phi(\epsilon_0) \sup_{\theta\in\Theta} \mathbb{P}_\theta[\rho(\theta,\hat{\theta}) \ge \epsilon_0]$$

- Suppose that $\hat{V} = \arg\min_{j=1,\dots,M} \rho(\theta_j,\hat{\theta})$. Then by the triangle inequality and $\rho(\theta_v,\theta_{v'}) \ge \epsilon$, if $\rho(\theta_v,\hat{\theta}) < \frac{\epsilon}{2}$ then we must have $\hat{V} = v$:

$$\mathbb{P}_{\theta_v}\left[\rho(\theta_v,\hat{\theta}) \ge \frac{\epsilon}{2}\right] \ge \mathbb{P}_{\theta_v}[\hat{V} \ne v].$$

Proof of General Lower Bound

- Using Markov's inequality:

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta\left[\ell(\theta,\hat{\theta})\right] \geq \sup_{\theta\in\Theta} \Phi(\epsilon_0)\mathbb{P}_\theta[\ell(\theta,\hat{\theta}) \geq \Phi(\epsilon_0)]$$
$$= \Phi(\epsilon_0) \sup_{\theta\in\Theta} \mathbb{P}_\theta[\rho(\theta,\hat{\theta}) \geq \epsilon_0]$$

- Suppose that $\hat{V} = \arg\min_{j=1,\dots,M} \rho(\theta_j,\hat{\theta})$. Then by the triangle inequality and $\rho(\theta_v,\theta_{v'}) \geq \epsilon$, if $\rho(\theta_v,\hat{\theta}) < \frac{\epsilon}{2}$ then we must have $\hat{V} = v$:

$$\mathbb{P}_{\theta_v}\left[\rho(\theta_v,\hat{\theta}) \geq \frac{\epsilon}{2}\right] \geq \mathbb{P}_{\theta_v}[\hat{V} \neq v].$$

- Hence,

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta\left[\rho(\theta,\hat{\theta}) \geq \frac{\epsilon}{2}\right] \geq \max_{v=1,\dots,M} \mathbb{P}_{\theta_v}\left[\rho(\theta_v,\hat{\theta}) \geq \frac{\epsilon}{2}\right]$$
$$\geq \max_{v=1,\dots,M} \mathbb{P}_{\theta_v}[\hat{V} \neq v]$$
$$\geq \frac{1}{M} \sum_{v=1,\dots,M} \mathbb{P}_{\theta_v}[\hat{V} \neq v]$$
$$\geq 1 - \frac{I(V;\mathbf{Y}) + \log 2}{\log M}$$

where the final step uses Fano's inequality.

Local Approach

- **General bound:** If $\rho(\theta_v, \theta_{v'}) \geq \epsilon$ for all v, v' then

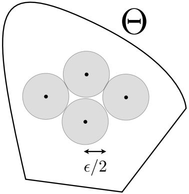
$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\Big(\frac{\epsilon}{2}\Big)\bigg(1 - \frac{I(V; \mathbf{Y}) + \log 2}{\log M}\bigg),$$

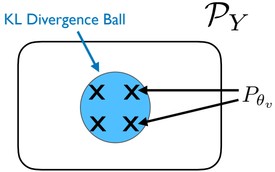
Local Approach

- **General bound:** If $\rho(\theta_v, \theta_{v'}) \geq \epsilon$ for all v, v' then

$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\left(\frac{\epsilon}{2}\right)\left(1 - \frac{I(V;\mathbf{Y}) + \log 2}{\log M}\right),$$

- **Local approach:** Carefully-chosen "local" hard subset:



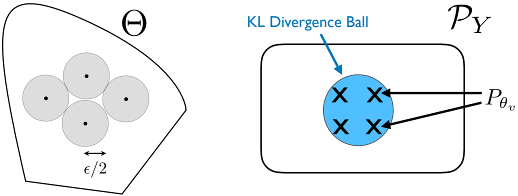


Local Approach

- **General bound:** If $\rho(\theta_v, \theta_{v'}) \ge \epsilon$ for all v, v' then

$$\mathcal{M}_n(\Theta, \ell) \ge \Phi\Big(\frac{\epsilon}{2}\Big)\bigg(1 - \frac{I(V;\mathbf{Y}) + \log 2}{\log M}\bigg),$$

- **Local approach:** Carefully-chosen "local" hard subset:



Resulting bound:

$$\mathcal{M}_n(\Theta, \ell) \ge \Phi\Big(\frac{\epsilon}{2}\Big)\bigg(1 - \frac{\min_{v=1,\dots,M} D(P_{\theta_v}^n \| Q_Y^n) + \log 2}{\log M}\bigg).$$

Global Approach

- **General bound:** If $\rho(\theta_v, \theta_{v'}) \geq \epsilon$ for all v, v' then

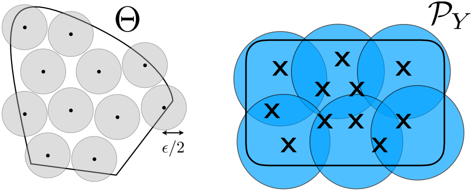
$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\left(\frac{\epsilon}{2}\right)\left(1 - \frac{I(V;\mathbf{Y}) + \log 2}{\log M}\right),$$

Global Approach

- **General bound:** If $\rho(\theta_v, \theta_{v'}) \geq \epsilon$ for all v, v' then

$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\left(\frac{\epsilon}{2}\right)\left(1 - \frac{I(V;\mathbf{Y}) + \log 2}{\log M}\right),$$

- **Global approach:** Pack as many ϵ-separated points as possible:



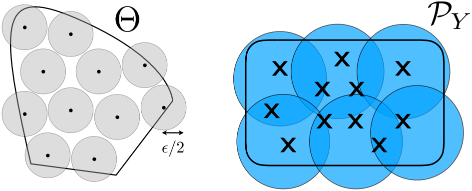
  - Typically suited to infinite-dimensional problems (e.g., non-parametric regression)

Global Approach

- **General bound:** If $\rho(\theta_v, \theta_{v'}) \ge \epsilon$ for all v, v' then

$$\mathcal{M}_n(\Theta, \ell) \ge \Phi\Big(\frac{\epsilon}{2}\Big)\bigg(1 - \frac{I(V;\mathbf{Y}) + \log 2}{\log M}\bigg),$$

- **Global approach:** Pack as many ϵ-separated points as possible:



  - Typically suited to infinite-dimensional problems (e.g., non-parametric regression)

- **Resulting bound:**

$$\mathcal{M}_n(\Theta, \ell) \ge \Phi\Big(\frac{\epsilon_{\rm p}}{2}\Big)\bigg(1 - \frac{\log N^*_{{\rm KL},n}(\Theta, \epsilon_{{\rm c},n}) + \epsilon_{{\rm c},n} + \log 2}{\log M^*_\rho(\Theta, \epsilon_{\rm p})}\bigg).$$

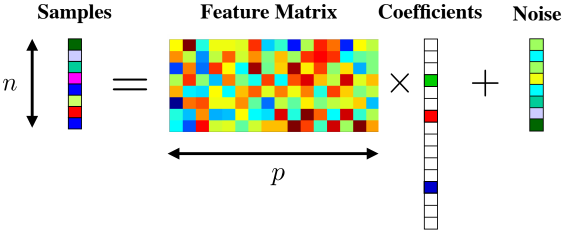
  - $M^*_\rho(\Theta, \epsilon_{\rm p})$: No. ϵ-separated θ we can pack into Θ (packing number)
  - $N^*_{{\rm KL},n}(\Theta, \epsilon_{{\rm c},n})$: No. $\epsilon_{{\rm c},n}$-size KL divergence balls to cover $\mathcal{P}_Y$ (covering number)

Continuous Example 1

Sparse Linear Regression

Sparse Linear Regression

- Linear regression model $\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}$:



  - Feature matrix $\mathbf{X}$ is given, noise is i.i.d. $N(0, \sigma^2)$
  - Coefficients are sparse – at most k non-zeros

Converse via Fano's Inequality

- **Reduction to hyp. testing:** Fix $\epsilon > 0$ and restrict to sparse vectors of the form

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

  - Total number of such sequences $= 2^k \binom{p}{k} \approx \exp\left(k \log \frac{p}{k}\right)$ (if $k \ll p$)
  - Choose a "well separated" subset of size $\exp\left(\frac{k}{4} \log \frac{p}{k}\right)$ (Gilbert-Varshamov)
  - Well-separated: Non-zero entries differ in at least $\frac{k}{8}$ indices

Converse via Fano's Inequality

- **Reduction to hyp. testing:** Fix $\epsilon > 0$ and restrict to sparse vectors of the form

$$\theta = (0, 0, 0, \pm\boldsymbol{\epsilon}, 0, \pm\boldsymbol{\epsilon}, 0, 0, 0, 0, 0, \pm\boldsymbol{\epsilon}, 0)$$

  - Total number of such sequences $= 2^k \binom{p}{k} \approx \exp\left(k \log \frac{p}{k}\right)$ (if $k \ll p$)
  - Choose a "well separated" subset of size $\exp\left(\frac{k}{4} \log \frac{p}{k}\right)$ (Gilbert-Varshamov)
  - Well-separated: Non-zero entries differ in at least $\frac{k}{8}$ indices

- **Application of Fano's inequality:**
  - Using the general bound given previously:

$$\mathcal{M}_n(\Theta, \ell) \geq \frac{k\epsilon^2}{32}\left(1 - \frac{I(V; \mathbf{Y}|\mathbf{X}) + \log 2}{\frac{k}{4} \log \frac{p}{k}}\right)$$

Converse via Fano's Inequality

- **Reduction to hyp. testing:** Fix $\epsilon > 0$ and restrict to sparse vectors of the form

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

  - Total number of such sequences $= 2^k \binom{p}{k} \approx \exp\left(k \log \frac{p}{k}\right)$ (if $k \ll p$)
  - Choose a "well separated" subset of size $\exp\left(\frac{k}{4}\log\frac{p}{k}\right)$ (Gilbert-Varshamov)
  - Well-separated: Non-zero entries differ in at least $\frac{k}{8}$ indices
- **Application of Fano's inequality:**
  - Using the general bound given previously:

$$\mathcal{M}_n(\Theta,\ell) \geq \frac{k\epsilon^2}{32}\left(1 - \frac{I(V;\mathbf{Y}|\mathbf{X}) + \log 2}{\frac{k}{4}\log\frac{p}{k}}\right)$$

- **Bounding the mutual information:**
  - By a direct calculation, $I(V;\mathbf{Y}|\mathbf{X}) \leq \frac{\epsilon^2}{2\sigma^2}\cdot\frac{k}{p}\|\mathbf{X}\|_F^2$ (Gaussian noise) [Actually extra steps (e.g., matrix Bernstein) needed when using Fano's inequality with exact recovery. But an "approximate recovery" version avoids it.]
  - Substitute and choose ϵ to optimize the bound: $\epsilon^2 = \frac{\sigma^2 p \log\frac{p}{k}}{2\|\mathbf{X}\|_F^2}$

Converse via Fano's Inequality

- **Reduction to hyp. testing:** Fix $\epsilon > 0$ and restrict to sparse vectors of the form

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

  - Total number of such sequences $= 2^k\binom{p}{k} \approx \exp\left(k \log \frac{p}{k}\right)$ (if $k \ll p$)
  - Choose a "well separated" subset of size $\exp\left(\frac{k}{4}\log\frac{p}{k}\right)$ (Gilbert-Varshamov)
  - Well-separated: Non-zero entries differ in at least $\frac{k}{8}$ indices
- **Application of Fano's inequality:**
  - Using the general bound given previously:

$$\mathcal{M}_n(\Theta,\ell) \ge \frac{k\epsilon^2}{32}\left(1 - \frac{I(V;\mathbf{Y}|\mathbf{X}) + \log 2}{\frac{k}{4}\log\frac{p}{k}}\right)$$

- **Bounding the mutual information:**
  - By a direct calculation, $I(V;\mathbf{Y}|\mathbf{X}) \le \frac{\epsilon^2}{2\sigma^2}\cdot\frac{k}{p}\|\mathbf{X}\|_F^2$ (Gaussian noise) [Actually extra steps (e.g., matrix Bernstein) needed when using Fano's inequality with exact recovery. But an "approximate recovery" version avoids it.]
  - Substitute and choose ϵ to optimize the bound: $\epsilon^2 = \frac{\sigma^2 p \log\frac{p}{k}}{2\|\mathbf{X}\|_F^2}$
- **Final result:** If $\|\mathbf{X}\|_F^2 \le np\Gamma$, then $\mathbb{E}[\|\theta - \hat{\theta}\|_2^2] \le \delta$ requires $n \ge \frac{c\sigma^2}{\Gamma\delta}\cdot k\log\frac{p}{k}$

Upper vs. Lower Bounds

- **Recap of model:** $\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}$, where θ is k-sparse
- **Lower bound:** If $\|\mathbf{X}\|_F^2 \le np\Gamma$, achieving $\mathbb{E}[\|\theta - \hat{\theta}\|_2^2] \le \delta$ requires $n \le \frac{c\sigma^2}{\Gamma\delta} \cdot k \log \frac{p}{k}$
- **Upper bound:** If $\mathbf{X}$ is a zero-mean random Gaussian matrix with power Γ per entry, then we can achieve $\mathbb{E}[\|\theta - \hat{\theta}\|_2^2] \le \delta$ using at most $n \ge \frac{c'\sigma^2}{\Gamma\delta} \cdot k \log \frac{p}{k}$ samples
  - Maximum-likelihood estimation suffices
- Tighter lower bounds could potentially be obtained under additional restrictions on $\mathbf{X}$

Continuous Example 2

Convex Optimization

Stochastic Convex Optimization

- **A basic optimization problem**

$$\mathbf{x}^\star = \mathrm{argmin}_{x \in D} f(x)$$

For simplicity, we focus on the 1D case $D \subseteq \mathbb{R}$ (extensions to $\mathbb{R}^d$ are possible)

Stochastic Convex Optimization

- **A basic optimization problem**

$$\mathbf{x}^\star = \mathrm{argmin}_{x \in D} f(x)$$

For simplicity, we focus on the 1D case $D \subseteq \mathbb{R}$ (extensions to $\mathbb{R}^d$ are possible)

- **Model:**
  - Noisy samples: When we query x, we get a noisy value and noisy gradient:

$$Y = f(x) + Z, \qquad Y' = f'(x) + Z'$$

  where $Z \sim N(0, \sigma^2)$ an $Z' \sim N(0, \sigma^2)$
  - Adaptive sampling: Chosen X_i may depend on $Y_1, \dots, Y_{i-1}$

Stochastic Convex Optimization

- **A basic optimization problem**

$$\mathbf{x}^\star = \operatorname{argmin}_{x \in D} f(x)$$

For simplicity, we focus on the 1D case $D \subseteq \mathbb{R}$ (extensions to $\mathbb{R}^d$ are possible)

- **Model:**
  - Noisy samples: When we query x, we get a noisy value and noisy gradient:

$$Y = f(x) + Z, \qquad Y' = f'(x) + Z'$$

  where $Z \sim N(0, \sigma^2)$ an $Z' \sim N(0, \sigma^2)$
  - Adaptive sampling: Chosen X_i may depend on $Y_1, \dots, Y_{i-1}$

- **Function classes:** Convex, strongly convex, Lipschitz, self-concordant, etc.
  - We will focus on the class of strongly convex functions
  - Strong convexity: $f(x) - \frac{c}{2}x^2$ is a convex function for some $c > 0$ (we set $c = 1$)

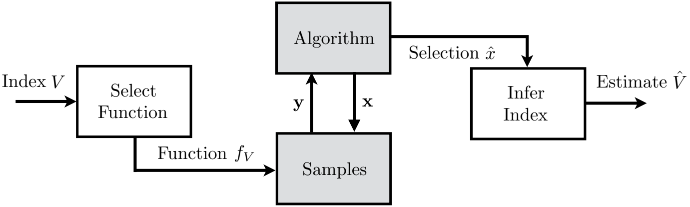
Performance Measure and Minimax Risk

- After sampling n points, the algorithm returns a final point $\hat{x}$
- The **loss** incurred is $\ell_f(\hat{x}) = f(\hat{x}) - \min_{x \in \mathcal{X}} f(x)$, i.e., the gap to the optimum
- For a given class of functions $\mathcal{F}$, the **minimax risk** is given by

$$\mathcal{M}_n(\mathcal{F}) = \inf_{\hat{X}} \sup_{f \in \mathcal{F}} \mathbb{E}_f[\ell_f(\hat{X})]$$

Reduction to Multiple Hypothesis Testing

- **The picture remains the same:**



  - Successful optimization $\implies$ Successful identification of V

General Minimax Lower Bound

- **Claim 1.** Fix $\epsilon > 0$, and let $\{f_1, \dots, f_M\} \subseteq \mathcal{F}$ be a subset of $\mathcal{F}$ such that for each $x \in \mathcal{X}$, we have $\ell_{f_v}(x) \le \epsilon$ for at most one value of $v \in \{1, \dots, M\}$. Then we have

$$\mathcal{M}_n(\mathcal{F}) \ge \epsilon \cdot \left(1 - \frac{I(V; \mathbf{X}, \mathbf{Y}) + \log 2}{\log M}\right), \tag{1}$$

where V is uniform on $\{1, \dots, M\}$, and $I(V; \mathbf{X}, \mathbf{Y})$ is w.r.t $V \to f_V \to (\mathbf{X}, \mathbf{Y})$.

General Minimax Lower Bound

- **Claim 1.** Fix $\epsilon > 0$, and let $\{f_1, \dots, f_M\} \subseteq \mathcal{F}$ be a subset of $\mathcal{F}$ such that for each $x \in \mathcal{X}$, we have $\ell_{f_v}(x) \le \epsilon$ for at most one value of $v \in \{1, \dots, M\}$. Then we have

$$\mathcal{M}_n(\mathcal{F}) \ge \epsilon \cdot \left(1 - \frac{I(V;\mathbf{X},\mathbf{Y}) + \log 2}{\log M}\right), \tag{1}$$

where V is uniform on $\{1, \dots, M\}$, and $I(V;\mathbf{X},\mathbf{Y})$ is w.r.t $V \to f_V \to (\mathbf{X},\mathbf{Y})$.

- **Claim 2.** In the special case $M = 2$, we have

$$\mathcal{M}_n(\mathcal{F}) \ge \epsilon \cdot H_2^{-1}\big(\log 2 - I(V;\mathbf{X},\mathbf{Y})\big), \tag{2}$$

where $H_2^{-1}(\cdot) \in [0, 0.5]$ is the inverse binary entropy function.

General Minimax Lower Bound

- **Claim 1.** Fix $\epsilon > 0$, and let $\{f_1, \dots, f_M\} \subseteq \mathcal{F}$ be a subset of $\mathcal{F}$ such that for each $x \in \mathcal{X}$, we have $\ell_{f_v}(x) \le \epsilon$ for at most one value of $v \in \{1, \dots, M\}$. Then we have

$$\mathcal{M}_n(\mathcal{F}) \ge \epsilon \cdot \left(1 - \frac{I(V; \mathbf{X}, \mathbf{Y}) + \log 2}{\log M}\right), \tag{1}$$

where V is uniform on $\{1, \dots, M\}$, and $I(V; \mathbf{X}, \mathbf{Y})$ is w.r.t $V \to f_V \to (\mathbf{X}, \mathbf{Y})$.

- **Claim 2.** In the special case $M = 2$, we have

$$\mathcal{M}_n(\mathcal{F}) \ge \epsilon \cdot H_2^{-1}\big(\log 2 - I(V; \mathbf{X}, \mathbf{Y})\big), \tag{2}$$

where $H_2^{-1}(\cdot) \in [0, 0.5]$ is the inverse binary entropy function.

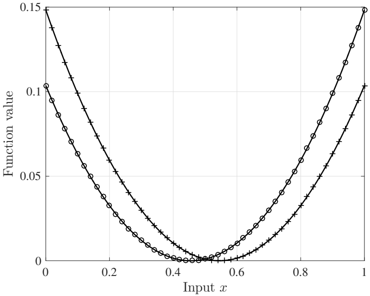
- Proof is like with estimation, starting with Markov's inequality:

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f[\ell_f(\hat{X})] \ge \sup_{f \in \mathcal{F}} \epsilon \cdot \mathbb{P}_f[\ell_f(\hat{X}) \ge \epsilon].$$

- Proof for $M = 2$ uses a (somewhat less well-known) form of Fano's inequality for *binary* hypothesis testing

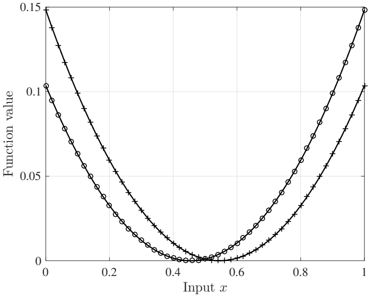
Strongly Convex Class: Choice of Hard Subset

- **Reduction to hyp. testing.** In 1D, it suffices to choose just two similar functions!
  - (Becomes $2^{\mathrm{constant} \times d}$ in d dimensions)



Strongly Convex Class: Choice of Hard Subset

- **Reduction to hyp. testing.** In 1D, it suffices to choose just two similar functions!
  - (Becomes $2^{\mathrm{constant} \times d}$ in d dimensions)



- The precise functions:

$$f_v(x) = \frac{1}{2}(x - x_v^*)^2, \quad v = 1, 2,$$

$$x_1^* = \frac{1}{2} - \sqrt{2\epsilon'} \qquad x_2^* = \frac{1}{2} + \sqrt{2\epsilon'}$$

Analysis

- **Application of Fano's Inequality.** As above,

$$\mathcal{M}_n(\mathcal{F}) \geq \epsilon \cdot H_2^{-1}(\log 2 - I(V; \mathbf{X}, \mathbf{Y}))$$

  - Approach: $H_2^{-1}(\alpha) \geq \frac{1}{10}$ if $\alpha \geq \frac{\log 2}{2}$
  - How few samples ensure $I(V; \mathbf{X}, \mathbf{Y}) \leq \frac{\log 2}{2}$?

Analysis

- **Application of Fano's Inequality.** As above,

$$\mathcal{M}_n(\mathcal{F}) \geq \epsilon \cdot H_2^{-1}(\log 2 - I(V; \mathbf{X}, \mathbf{Y}))$$

  - Approach: $H_2^{-1}(\alpha) \geq \frac{1}{10}$ if $\alpha \geq \frac{\log 2}{2}$
  - How few samples ensure $I(V; \mathbf{X}, \mathbf{Y}) \leq \frac{\log 2}{2}$?

- **Bounding the Mutual Information.** Let $P_Y, P_{Y'}$ be the observation distributions (function and gradient), and $Q_Y, Q_{Y'}$ similar but with $f_0(x) = \frac{1}{2}x^2$. Then:

$$D(P_Y \times P_{Y'} \| Q_Y \times Q_{Y'}) = \frac{(f_1(x) - f_0(x))^2}{2\sigma^2} + \frac{(f_1'(x) - f_0'(x))^2}{2\sigma^2}$$

  - Simplifications: $(f_1(x) - f_0(x))^2 \leq \left(\epsilon + \sqrt{\frac{\epsilon}{2}}\right)^2 \leq 2\epsilon$ and $(f_1'(x) - f_0'(x))^2 = 2\epsilon$
  - With some manipulation, $I(V; \mathbf{X}, \mathbf{Y}) \leq \frac{\log 2}{2}$ when $\epsilon = \frac{\sigma^2 \log 2}{4n}$

Analysis

- **Application of Fano's Inequality.** As above,

$$\mathcal{M}_n(\mathcal{F}) \geq \epsilon \cdot H_2^{-1}(\log 2 - I(V; \mathbf{X}, \mathbf{Y}))$$

  - Approach: $H_2^{-1}(\alpha) \geq \frac{1}{10}$ if $\alpha \geq \frac{\log 2}{2}$
  - How few samples ensure $I(V; \mathbf{X}, \mathbf{Y}) \leq \frac{\log 2}{2}$?

- **Bounding the Mutual Information.** Let $P_Y, P_{Y'}$ be the observation distributions (function and gradient), and $Q_Y, Q_{Y'}$ similar but with $f_0(x) = \frac{1}{2}x^2$. Then:

$$D(P_Y \times P_{Y'} \| Q_Y \times Q_{Y'}) = \frac{(f_1(x) - f_0(x))^2}{2\sigma^2} + \frac{(f_1'(x) - f_0'(x))^2}{2\sigma^2}$$

  - Simplifications: $(f_1(x) - f_0(x))^2 \leq \big(\epsilon + \sqrt{\frac{\epsilon}{2}}\big)^2 \leq 2\epsilon$ and $(f_1'(x) - f_0'(x))^2 = 2\epsilon$
  - With some manipulation, $I(V; \mathbf{X}, \mathbf{Y}) \leq \frac{\log 2}{2}$ when $\epsilon = \frac{\sigma^2 \log 2}{4n}$

- **Final result:** $\mathcal{M}_n(\mathcal{F}) \geq \frac{\sigma^2 \log 2}{40n}$

Upper vs. Lower Bounds

- Lower bound for 1D strongly convex functions: $\mathcal{M}_n(\mathcal{F}) \geq c\,\frac{\sigma^2}{n}$

- Upper bound for 1D strongly convex functions: $\mathcal{M}_n(\mathcal{F}) \leq c'\,\frac{\sigma^2}{n}$
  - Achieved by stochastic gradient descent

- Analogous results (and proof techniques) known for d-dimensional functions, additional Lipschitz assumptions, etc. [Raginsky and Rakhlin, 2011]

Continuous Example 3

Density Estimation

Density Estimation Example

- **An example density estimation problem:**
  - Goal: Estimate the density f given n i.i.d. samples
  - Here we consider random variables defined on $[0,1]$, and consider the class $\mathcal{F}_{\eta,\Gamma}$ of density functions satisfying the following:

$$f(y) \geq \eta, \forall y \in [0,1], \qquad \|f\|_{\mathrm{TV}} \leq \Gamma,$$

    where $\|f\|_{\mathrm{TV}} = \sup_L \sup_{0 \leq x_1 \leq \ldots \leq x_L \leq 1} \sum_{l=2}^{L} \big(f(x_l) - f(x_{l-1})\big)$.
  - We measure performance via the ℓ_2^2-loss:

$$\ell(f, \hat{f}) = \|f - \hat{f}\|_2^2 = \int_0^1 (f(x) - \hat{f}(x))^2 dx$$

  - Minimax risk:

$$\mathcal{M}_n(\eta, \Gamma) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}_{\eta,\Gamma}} \mathbb{E}_f\big[\|f - \hat{f}\|_2^2\big],$$

Density Estimation Example

- **An example density estimation problem:**
  - Goal: Estimate the density f given n i.i.d. samples
  - Here we consider random variables defined on $[0,1]$, and consider the class $\mathcal{F}_{\eta,\Gamma}$ of density functions satisfying the following:

$$f(y) \geq \eta, \forall y \in [0,1], \qquad \|f\|_{\mathrm{TV}} \leq \Gamma,$$

    where $\|f\|_{\mathrm{TV}} = \sup_L \sup_{0 \leq x_1 \leq \ldots \leq x_L \leq 1} \sum_{l=2}^{L} \big(f(x_l) - f(x_{l-1})\big)$.
  - We measure performance via the ℓ_2^2-loss:

$$\ell(f, \hat{f}) = \|f - \hat{f}\|_2^2 = \int_0^1 (f(x) - \hat{f}(x))^2 dx$$

  - Minimax risk:

$$\mathcal{M}_n(\eta, \Gamma) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}_{\eta,\Gamma}} \mathbb{E}_f\big[\|f - \hat{f}\|_2^2\big],$$

- **Claim:** For constant η and Γ, attaining $\mathcal{M}_n(\eta,\Gamma) \leq \delta$ requires $n \geq c\big(\frac{1}{\delta}\big)^{3/2}$.
  - This scaling is tight; a matching upper bound is known
  - The proof uses the global packing/covering approach
  - See our survey introductory guide to Fano's inequality for this specific example, or Yang/Barron's original paper for many more classes

Limitations and Generalizations

Limitations and Generalizations

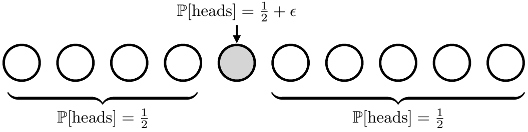
- **Limitations of Fano's Inequality.**
  - Non-asymptotic weakness
  - Often hard to tightly bound mutual information in adaptive settings
  - Restriction to KL divergence
    - Other useful measures: Total variation, Hellinger distance, χ^2-divergence, etc.

- **Generalizations of Fano's Inequality.**
  - Non-uniform V [Han/Verdú, 1994]
  - More general f-divergences [Guntuboyina, 2011]
  - Continuous V [Duchi/Wainwright, 2013]

(This list is certainly incomplete!)

Example: Difficulties in Adaptive Settings

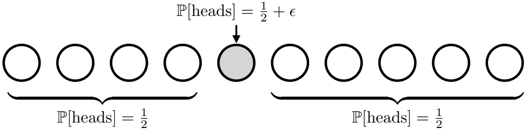
- **A simple search problem:** Find the (only) biased coin using few flips



  - 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

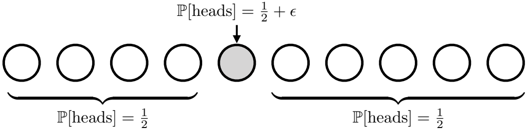


  - Heavy coin $V \in \{1, \dots, M\}$ uniformly at random
  - Selected coin at time $i = 1, \dots, n$ is X_i, observation is $Y_i \in \{0, 1\}$ (1 for heads)

- **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



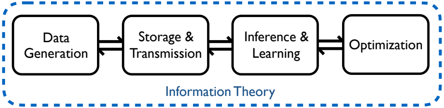
  - Heavy coin $V \in \{1, \dots, M\}$ uniformly at random
  - Selected coin at time $i = 1, \dots, n$ is X_i, observation is $Y_i \in \{0, 1\}$ (1 for heads)

- **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 $n \gtrsim \frac{\log M}{\epsilon^2}$
  - **Next lecture:** An alternative tool that gives $n \gtrsim \frac{M}{\epsilon^2}$

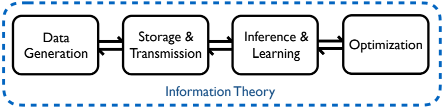
Conclusion

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



Conclusion

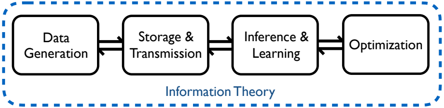
- **Information theory as a theory of data**:



- **Approach highlighted in this talk:**
  - Reduction to multiple hypothesis testing
  - Application of Fano's inequality
  - Bounding the mutual information

Conclusion

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



- **Approach highlighted in this talk:**
  - Reduction to multiple hypothesis testing
  - Application of Fano's inequality
  - Bounding the mutual information

- **Examples**:
  - Group testing
  - Graphical model selection
  - Sparse regression
  - Convex optimization
  - ...and many more!

- **Tutorial Chapter:** "An Introductory Guide to Fano's Inequality with Applications in Statistical Estimation" [S. and Cevher, 2019]

https://arxiv.org/abs/1901.00555

(Chapter in 2021 book *Information-Theoretic Methods in Data Science*, Cambridge University Press)