An Introduction to Statistical Lower Bounds for Estimation and Learning

Part 2: Other Methods from Statistics, Information Theory, and Theoretical Computer Science

Jonathan Scarlett

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

Preliminaries: Measuring Distances Between Distributions

- **Divergences/distances we will use:**
  - KL divergence:
$$D(P\|Q) = \mathbb{E}_P\Big[\log\frac{P(X)}{Q(X)}\Big]$$
  - TV distance:
$$d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$$
    where $\sup(\cdot)$ is over all events. If discrete, $d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$; if continuous $d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\int |P(x) - Q(x)| dx$.
  - χ^2-divergence:
$$\chi^2(P,Q) = \mathbb{E}_Q\Big[\Big(\frac{P(X)}{Q(X)} - 1\Big)^2\Big]$$
    or expanding the square gives $\chi^2(P,Q) = \mathbb{E}_P\big[\frac{P(X)}{Q(X)}\big] - 1$.

Preliminaries: Measuring Distances Between Distributions

- **Divergences/distances we will use:**
  - KL divergence:
$$D(P\|Q) = \mathbb{E}_P\Big[\log\frac{P(X)}{Q(X)}\Big]$$
  - TV distance:
$$d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$$
    where $\sup(\cdot)$ is over all events. If discrete, $d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$; if continuous $d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\int |P(x) - Q(x)| dx$.
  - χ^2-divergence:
$$\chi^2(P,Q) = \mathbb{E}_Q\bigg[\Big(\frac{P(X)}{Q(X)} - 1\Big)^2\bigg]$$
    or expanding the square gives $\chi^2(P,Q) = \mathbb{E}_P\big[\frac{P(X)}{Q(X)}\big] - 1$.

- **Other useful ones (that we won't use here):**
  - Hellinger distance
  - Wasserstein distances
  - Generalizations of the above (e.g., f-divergences, Rényi divergences)

Properties (I)

Example uses/results:

- As mentioned earlier, $e^{-nD(P\|Q)}$ is roughly the probability of symbol proportions P when we draw n i.i.d. samples from Q
- TV norm naturally leads to additive change of measure:

$$\mathbb{P}_P[A] \le \mathbb{P}_Q[A] + d_{\mathrm{TV}}(P, Q),$$

  e.g., where A is some "success" event that has low probability under Q

Properties (I)

Example uses/results:

- As mentioned earlier, $e^{-nD(P\|Q)}$ is roughly the probability of symbol proportions P when we draw n i.i.d. samples from Q
- TV norm naturally leads to additive change of measure:

$$\mathbb{P}_P[A] \le \mathbb{P}_Q[A] + d_{\mathrm{TV}}(P, Q),$$

  e.g., where A is some "success" event that has low probability under Q

Example properties:

- Non-negativity: All are ≥ 0 with equality if and only if $P = Q$.
- Tensorization: $D(\prod_i P_i \| \prod_i Q_i) = \sum_i D(P_i \| Q_i)$ (more generally chain rule). For d_{TV} only $\le$ instead of $=$. For χ^2 we get an equality with $\prod_i (1 + \chi^2(P_i.Q_i)) - 1$.
- Triangle inequality: d_{TV} satisfies triangle inequality, KL and χ^2 don't
- Data processing inequality: $D(P_Y \| Q_Y) \le D(P_X \| Q_X)$ if $P_X \stackrel{P_{Y|X}}{\to} P_Y$ and $Q_X \stackrel{P_{Y|X}}{\to} Q_Y$. This also holds for TV, χ^2, and others.
- Variational forms: e.g., $D(P\|Q) = \sup_f \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[e^{f(X)}]$

See Yihong Wu's lecture notes for a lot more on the above concepts.

Properties (II)

Example relations:

- Pinsker's inequality:

$$d_{\mathrm{TV}}(P,Q) \le \sqrt{\frac{1}{2} D(P\|Q)}$$

  - If the PMF (or PDF) is uniformly lower bounded, a similar *lower* bound holds
- Bretagnolle-Huber inequality:

$$d_{\mathrm{TV}}(P,Q) \le \sqrt{1 - e^{-D(P\|Q)}}.$$

- χ^2 divergence upper bound:

$$D(P\|Q) \le \log(1 + \chi^2(P,Q)) \le \chi^2(P,Q).$$

Le Cam & Assouad Methods

Le Cam's Method

- Let $P_0(\mathrm{y})$ and $P_1(\mathrm{y})$ be two distributions on the observations

Le Cam's Method

- Let $P_0(\mathrm{y})$ and $P_1(\mathrm{y})$ be two distributions on the observations
- A very basic inequality (essentially by definition):

$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \leq d_{\mathrm{TV}}(P_0, P_1)$$

for any event A

  - This is a simple form of Le Cam's method (more general form later based on sets of distributions)
  - We can use this inequality to lower bound hypothesis testing error probability in terms of TV norm

Le Cam's Method

- Let $P_0(\mathrm{y})$ and $P_1(\mathrm{y})$ be two distributions on the observations
- A very basic inequality (essentially by definition):

$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \le d_{\mathrm{TV}}(P_0, P_1)$$

for any event A

  - This is a simple form of Le Cam's method (more general form later based on sets of distributions)
  - We can use this inequality to lower bound hypothesis testing error probability in terms of TV norm

- Weakened version (via Pinsker's inequality):

$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \le \sqrt{\frac{1}{2} D(P_1 \| P_0)}$$

(could also swap P_0 and P_1 on the right-hand side)

Le Cam's Method

- Let $P_0(\mathrm{y})$ and $P_1(\mathrm{y})$ be two distributions on the observations
- A very basic inequality (essentially by definition):

$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \le d_{\mathrm{TV}}(P_0, P_1)$$

for any event A

  - This is a simple form of Le Cam's method (more general form later based on sets of distributions)
  - We can use this inequality to lower bound hypothesis testing error probability in terms of TV norm

- Weakened version (via Pinsker's inequality):

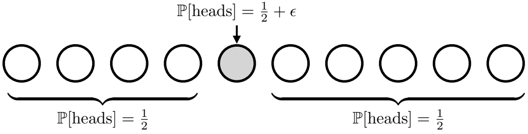
$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \le \sqrt{\frac{1}{2} D(P_1 \| P_0)}$$

(could also swap P_0 and P_1 on the right-hand side)

- **Applications:**
  - Statistical estimation [Le Cam, 1973]
  - Multi-armed bandits [Auer *et al.*, 1995]
  - Black-box optimization [Scarlett *et al.*, 2017]

Example 1: Finding a Biased Coin

- **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)

- **Note:** This is a simple example of the multi-armed bandit problem, for which similar analysis techniques have also given tight lower bounds

Example 1: Finding a Biased Coin

- **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)
- **Analysis:**
  - Apply weakened bound above to get

$$\mathbb{P}_v[\hat{V} = v] \leq \mathbb{P}_0[\hat{V} = v] + \sqrt{\frac{1}{2} D(P_0 \| P_v)}$$

where $P_v(\mathrm{y})$ corresponds to $V = v$, and $P_0(\mathrm{y})$ corresponds to all fair coins

Example 1: Finding a Biased Coin

- **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)
- **Analysis:**
  - Apply weakened bound above to get

$$\mathbb{P}_v[\hat{V} = v] \le \mathbb{P}_0[\hat{V} = v] + \sqrt{\frac{1}{2} D(P_0 \| P_v)}$$

    where $P_v(\mathrm{y})$ corresponds to $V = v$, and $P_0(\mathrm{y})$ corresponds to all fair coins
  - By chain rule for KL divergence and the fact that only coin v differs:

$$D(P_0 \| P_v) \lesssim \mathbb{E}_0[N_v] \epsilon^2$$

    where N_v is the number of flips of coin v (Note: $\mathrm{KL}(\frac{1}{2} \| \frac{1}{2} + \epsilon) \simeq \epsilon^2$)

Example 1: Finding a Biased Coin

- **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)
- **Analysis:**
  - Apply weakened bound above to get

$$\mathbb{P}_v[\hat{V} = v] \le \mathbb{P}_0[\hat{V} = v] + \sqrt{\frac{1}{2} D(P_0 \| P_v)}$$

where $P_v(\mathrm{y})$ corresponds to $V = v$, and $P_0(\mathrm{y})$ corresponds to all fair coins

  - By chain rule for KL divergence and the fact that only coin v differs:

$$D(P_0 \| P_v) \lesssim \mathbb{E}_0[N_v]\epsilon^2$$

where N_v is the number of flips of coin v (Note: $\mathrm{KL}(\frac{1}{2} \| \frac{1}{2} + \epsilon) \simeq \epsilon^2$)

  - Apply $\frac{1}{M} \sum_{v=1}^{M}$ on both sides of first step, then Jensen's inequality:

$$\mathbb{P}[\hat{V} = V] \lesssim \frac{1}{M} + \sqrt{\frac{n\epsilon^2}{M}}$$

since $\sum_{v=1}^{M} \mathbb{P}_0[\hat{V} = v] = 1$ and $\sum_{v=1}^{M} \mathbb{E}[N_v] = n$

Example 1: Finding a Biased Coin

- **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)
- **Analysis:**
  - Apply weakened bound above to get

$$\mathbb{P}_v[\hat{V} = v] \le \mathbb{P}_0[\hat{V} = v] + \sqrt{\frac{1}{2} D(P_0 \| P_v)}$$

    where $P_v(\mathrm{y})$ corresponds to $V = v$, and $P_0(\mathrm{y})$ corresponds to all fair coins
  - By chain rule for KL divergence and the fact that only coin v differs:

$$D(P_0 \| P_v) \lesssim \mathbb{E}_0[N_v]\epsilon^2$$

    where N_v is the number of flips of coin v (Note: $\mathrm{KL}(\frac{1}{2} \| \frac{1}{2} + \epsilon) \simeq \epsilon^2$)
  - Apply $\frac{1}{M} \sum_{v=1}^{M}$ on both sides of first step, then Jensen's inequality:

$$\mathbb{P}[\hat{V} = V] \lesssim \frac{1}{M} + \sqrt{\frac{n\epsilon^2}{M}}$$

    since $\sum_{v=1}^{M} \mathbb{P}_0[\hat{V} = v] = 1$ and $\sum_{v=1}^{M} \mathbb{E}[N_v] = n$
  - Hence, achieving $\mathbb{P}[\hat{V} = V] \ge \frac{1}{2}$ requires $n \gtrsim \frac{M}{\epsilon^2}$

Example 2: Gaussian Mean Estimation

• **Simple example:** Suppose that we have i.i.d. samples $\mathsf{Y} = (Y_1, \ldots, Y_n)$ drawn from either $P_+ : N(\epsilon, \sigma^2)$ or $P_- : N(-\epsilon, \sigma^2)$. When can we distinguish these two cases?

Example 2: Gaussian Mean Estimation

- **Simple example:** Suppose that we have i.i.d. samples $\mathrm{Y} = (Y_1, \dots, Y_n)$ drawn from either $P_+ \,:\, N(\epsilon, \sigma^2)$ or $P_- \,:\, N(-\epsilon, \sigma^2)$. When can we distinguish these two cases?

- Let A_v be the event that P_v is chosen ($v \in \{+,-\}$). By the weakened bound above,

$$|\mathbb{P}_+[A_v] - \mathbb{P}_-[A_v]| \le \sqrt{\frac{1}{2} D(P_+^n \| P_-^n)} = \sqrt{\frac{n}{2} D(P_+ \| P_-)} = \sqrt{\frac{n\epsilon^2}{\sigma^2}}.$$

For instance, this is less than $\frac{1}{2}$ if $n \le \frac{\sigma^2}{4\epsilon^2}$, and in this case, if we have $\mathbb{P}_+[A_+] \ge 1 - \delta$ (a good event), then we must have $\mathbb{P}_-[A_-] \le \frac{1}{2} + \delta$ (a bad event)

Example 2: Gaussian Mean Estimation

- **Simple example:** Suppose that we have i.i.d. samples $\mathrm{Y} = (Y_1, \dots, Y_n)$ drawn from either $P_+ : N(\epsilon, \sigma^2)$ or $P_- : N(-\epsilon, \sigma^2)$. When can we distinguish these two cases?

- Let A_v be the event that P_v is chosen ($v \in \{+,-\}$). By the weakened bound above,

$$|\mathbb{P}_+[A_v] - \mathbb{P}_-[A_v]| \le \sqrt{\frac{1}{2} D(P_+^n \| P_-^n)} = \sqrt{\frac{n}{2} D(P_+ \| P_-)} = \sqrt{\frac{n\epsilon^2}{\sigma^2}}.$$

For instance, this is less than $\frac{1}{2}$ if $n \le \frac{\sigma^2}{4\epsilon^2}$, and in this case, if we have $\mathbb{P}_+[A_+] \ge 1 - \delta$ (a good event), then we must have $\mathbb{P}_-[A_-] \le \frac{1}{2} + \delta$ (a bad event)

- **Implication.** The minimax risk for 1D Gaussian mean estimation satisfies

$$\inf_{\hat{\mu}} \sup_{\mu} \mathbb{E}[(\mu - \hat{\mu}(\mathrm{Y}))^2] \ge \epsilon^2 \mathbb{P}[(\mu - \hat{\mu}(\mathrm{Y}))^2 \ge \epsilon^2] \ge \frac{\sigma^2}{16n}.$$

by setting $\epsilon^2 = \frac{\sigma^2}{4n}$ and considering $v \in \{+,-\}$ as occurring with probability $\frac{1}{2}$ each. (The above analysis leads to $\mathbb{P}[(\mu - \hat{\mu}(\mathrm{Y}))^2 \ge \epsilon^2] \ge \frac{1}{2}\delta + \frac{1}{2}\big(\frac{1}{2} + \delta\big) \ge \frac{1}{4}$ via similar steps to those used via Fano's inequality.)

Generalization 1: Using a Mixture Distribution

- **A useful generalization:**
  - Suppose that we are required to "distinguish" P_0 from not only P_1, but from *all of* $P_1, \dots, P_K$ (or more generally a continuum of distributions)
  - Obviously, if any $d_{\mathrm{TV}}(P_0, P_i)$ is small, this is a hard problem. Can we say more?
  - Le Cam's method using a mixture of distributions: For any non-negative $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ with $\sum_k \mu_k = 1$, if we define $P_{\boldsymbol{\mu}}(\cdot) = \sum_k \mu_k P_k$, then

$$|\mathbb{P}_0[A] - \mathbb{P}_\mu[A]| \le d_{\mathrm{TV}}(P_0, P_{\boldsymbol{\mu}}).$$

  Using, we can get a hardness result not just from individual $d_{\mathrm{TV}}(P_0, P_i)$ being small, but from $d_{\mathrm{TV}}(P_0, P_{\boldsymbol{\mu}})$ being small for any $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$

Example: Detecting a Hidden Clique

- **Example:**
  - **Goal:** Reliably distinguish between the following two scenarios:
    (i) G is an $\mathrm{ER}(\frac{1}{2})$ random graph (i.e., every edge included w.p. $\frac{1}{2}$ independently)
    (ii) An unknown set of k nodes is fully-connected (a clique) and the rest follow the $\mathrm{ER}(\frac{1}{2})$ model.
  - Let Q be the distribution in (i), and let P_S be the distribution in (ii) when S is the size-k subset of fully connected nodes.
  - It is not hard to show that $d_{\mathrm{TV}}(P_S, Q) = 1 - 2^{\binom{k}{2}}$, which isn't useful (for a hardness result we want to show that d_{TV} is small). The problem is that $d_{\mathrm{TV}}(P_S, Q)$ doesn't capture the fact that S is unknown.

Example: Detecting a Hidden Clique

- **Example:**
  - **Goal:** Reliably distinguish between the following two scenarios:
    - (i) G is an $\mathrm{ER}(\frac{1}{2})$ random graph (i.e., every edge included w.p. $\frac{1}{2}$ independently)
    - (ii) An unknown set of k nodes is fully-connected (a clique) and the rest follow the $\mathrm{ER}(\frac{1}{2})$ model.
  - Let Q be the distribution in (i), and let P_S be the distribution in (ii) when S is the size-k subset of fully connected nodes.
  - It is not hard to show that $d_{\mathrm{TV}}(P_S, Q) = 1 - 2^{\binom{k}{2}}$, which isn't useful (for a hardness result we want to show that d_{TV} is small). The problem is that $d_{\mathrm{TV}}(P_S, Q)$ doesn't capture the fact that S is unknown.
  - However, $P := \frac{1}{\binom{n}{k}} P_S$ and Q are much closer!
  - The χ^2-divergence turns out to be more convenient to work with, because it satisfies the following nice property:

$$\chi^2(\mathbb{E}_K[P_S], Q) = \mathbb{E}_{S,S'}\left[\mathbb{E}_{P_S}\left[\frac{P_{S'}(X)}{Q(X)}\right]\right] - 1$$

    with S, S' being independent draws from the $\binom{n}{k}$ possible k-cliques
  - Skipping details, $\chi^2(P, Q)$ is small unless $k_n \gtrsim 2\log_2 n - 2\log_2\log_2 n + \mathrm{constant}$
    - Small χ^2 implies small TV distance, which implies the problem can't be solved
    - The above bound is tight – if k_n is any larger, then w.h.p. the $\mathrm{ER}(\frac{1}{2})$ graph has no k_n-cliques, so the two distributions can be distinguished.

Generalization 2: Assouad's Method

- **Note:** Le Cam's method only concerns the difficulty of distinguishing two parameters/distributions (or mixtures thereof).

- **Widely-used generalization.** Assouad's method concerns the difficulty of distinguishing 2^d parameters/distributions, interpreted as vertices of an d-dimensional hypercube (i.e., representable as $\{\pm 1\}^d$)
  - Intuition: Each dimension acts as a sub-problem, and we characterize the difficulty of that sub-problem via Le Cam's method

- **Useful comparison of three methods:** "*Assouad, Fano, and Le Cam*" [Yu, 1997]
  - See also Chapter 15 of [Wainwright, 2019], lecture notes by John Duchi, or lecture notes by Yihong Wu

Example: *Multivariate* Gaussian Mean Estimation

- **General statement:** Consider a set of distributions P_{θ_v} with $v \in \{-1, 1\}^d$. If there exists some $\delta > 0$ such that the loss function satisfies

$$\ell(\theta_v, \theta_{v'}) \geq 2\delta d_{\mathrm{H}}(v, v')$$

with d_{H} denoting Hamming distance, then minimax risk is lower bound as follows:

$$\inf_{\hat{\theta}} \sup_{\theta} \ell(\theta, \hat{\theta}) \geq \delta \sum_{i=1}^{d} \left[1 - d_{\mathrm{TV}}(P_j^+, P_j^-)\right]$$

where $P_j^+ = \frac{1}{2^{d-1}} \sum_{v \,:\, v_j = 1} P_{\theta_v}$ and $P_j^- = \frac{1}{2^{d-1}} \sum_{v \,:\, v_j = -1} P_{\theta_v}$

Example: *Multivariate* Gaussian Mean Estimation

- **General statement:** Consider a set of distributions P_{θ_v} with $v \in \{-1,1\}^d$. If there exists some $\delta > 0$ such that the loss function satisfies

$$\ell(\theta_v, \theta_{v'}) \geq 2\delta d_{\mathrm{H}}(v, v')$$

with d_{H} denoting Hamming distance, then minimax risk is lower bound as follows:

$$\inf_{\hat{\theta}} \sup_{\theta} \ell(\theta, \hat{\theta}) \geq \delta \sum_{i=1}^{d} \big[1 - d_{\mathrm{TV}}(P_j^+, P_j^-)\big]$$

where $P_j^+ = \frac{1}{2^{d-1}} \sum_{v \,:\, v_j = 1} P_{\theta_v}$ and $P_j^- = \frac{1}{2^{d-1}} \sum_{v \,:\, v_j = -1} P_{\theta_v}$

- **Example:** Consider Gaussian mean estimation with $\theta_v = \epsilon v$ and $\ell(\theta, \theta') = \|\theta - \theta'\|^2$
  - This gives $\|\theta_v - \theta_{v'}\|^2 = \epsilon^2 d_{\mathrm{H}}(v, v')$, so $\delta = \epsilon^2/2$
  - If P_θ consists of n independent observations with $N(0, \sigma^2 \mathrm{I})$ noise, we can use Pinsker's inequality ($d_{\mathrm{TV}} \leq \sqrt{D_{\mathrm{KL}}/2}$) to get $d_{\mathrm{TV}}(P_j^+, P_j^-) \leq \sqrt{2n\epsilon^2/\sigma^2}$
  - Substituting into the above lower bound gives

$$\inf_{\hat{\theta}} \sup_{\theta} \ell(\theta, \hat{\theta}) \geq d\epsilon^2 \Big[1 - \sqrt{2n\epsilon^2/\sigma^2}\Big].$$

  Setting $\epsilon^2 = \frac{\sigma^2}{8n}$ gives a lower bound with dependence $\frac{\sigma^2 d}{n}$, which is tight (matched by the sample mean estimator)

Change-of-Measure Techniques

Multiplicative Change of Measure

- Le Cam's method can be viewed as an additive change of measure (e.g., $\mathbb{P}_P[A] \le \mathbb{P}_Q[A] + d_{\mathrm{TV}}(P, Q)$)

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

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

where γ is an arbitrary threshold

Multiplicative Change of Measure

- Le Cam's method can be viewed as an additive change of measure (e.g., $\mathbb{P}_P[A] \le \mathbb{P}_Q[A] + d_{\mathrm{TV}}(P,Q)$)

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

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

where γ is an arbitrary threshold

- **Applications:**
  - Channel coding [Wolfowitz, 1957] [Verdú and Han, 1994]
  - Multi-armed bandits [Lai and Robbins, 1985] [Kaufmann *et al.*, 2016]
  - Statistical estimation [Tsybakov, 2009] [Venkataramanan and Johnson, 2018]
  - Sparse recovery & group testing [Scarlett and Cevher, 2017]

Example: Binary Hypothesis Testing

- **Example: Binary hypothesis testing**
  - Goal: Given samples $\mathsf{X} = (X_1, \dots, X_n)$ i.i.d. from either P or Q, output 1 for P and 0 for Q. Let T denote the output.
  - Example question: If $\mathbb{P}_P[T = 1] \ge 0.99$, how does $\mathbb{P}_Q[T = 1]$ behave w.r.t n?

Example: Binary Hypothesis Testing

- **Example: Binary hypothesis testing**
  - Goal: Given samples $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. from either P or Q, output 1 for P and 0 for Q. Let T denote the output.
  - Example question: If $\mathbb{P}_P[T = 1] \ge 0.99$, how does $\mathbb{P}_Q[T = 1]$ behave w.r.t n?
  - Analysis:
    - Let $\mathcal{A}$ be the event that $T = 1$. The previous slide gives
$$\mathbb{P}_P\left[\frac{P^n(\mathbf{X})}{Q^n(\mathbf{X})} > \gamma\right] + \gamma \mathbb{P}_Q[T = 1] \ge 0.99$$
    - Write the condition $\frac{P^n(\mathbf{X})}{Q^n(\mathbf{X})} > \gamma$ as $\sum_{i=1}^n \log \frac{P(X_i)}{Q(X_i)} > \log \gamma$, and notice that $\mathbb{P}_P\big[\frac{P^n(\mathbf{X})}{Q^n(\mathbf{X})} > \gamma\big] \to 0$ if $\log \gamma$ is slightly above $nD(P\|Q)$ (law of large numbers)
    - This implies for n large enough that $\gamma \mathbb{P}_Q[T = 1] \ge 0.98$.
    - Since γ is roughly $e^{nD(P\|Q)}$, we conclude that $\mathbb{P}_Q[T = 1] \gtrsim e^{-nD(P\|Q)}$
  - This lower bound is tight; there exist testing strategies that get $\mathbb{P}_Q[T = 1] \ge 0.99$ and $\mathbb{P}_Q[T = 1] \lesssim e^{-nD(P\|Q)}$.

User-Friendly Simplifications

• **Note:** For all of the above methods, they are not necessarily applied "from scratch"; instead, "user-friendly" simplifications are often applied

• **Example 1:** Tsyabkov's textbook on non-parametric estimation gives "Fano-like" tools for minimax lower bounds that can be applied directly given:

(i) $\theta_1, \dots, \theta_M$ separated by distance 2δ

(ii) Bounds on quantities like $\frac{1}{M}\sum_{j=1}^M D(P_j \| P_0)$ or $\frac{1}{M}\sum_{j=1}^M P_j\big[\frac{P_0(\mathbf{Y})}{P_j(\mathbf{Y})} \ge \tau\big]$ for some "null distribution" P_0.

User-Friendly Simplifications

- **Note:** For all of the above methods, they are not necessarily applied "from scratch"; instead, "user-friendly" simplifications are often applied

- **Example 1:** Tsyabkov's textbook on non-parametric estimation gives "Fano-like" tools for minimax lower bounds that can be applied directly given:
  (i) $\theta_1, \dots, \theta_M$ separated by distance 2δ
  (ii) Bounds on quantities like $\frac{1}{M}\sum_{j=1}^M D(P_j \| P_0)$ or $\frac{1}{M}\sum_{j=1}^M P_j\big[\frac{P_0(\mathbf{Y})}{P_j(\mathbf{Y})} \ge \tau\big]$ for some "null distribution" P_0.

- **Example 2:** In sequential decision-making involving K distributions (e.g., arms) $\nu_1, \dots, \nu_K$, Kaufmann *et al.*'s "information complexity" paper gives a result of the form

$$\sum_{a=1}^K \mathbb{E}_\nu[N_a] D(\nu_a \| \nu'_a) \ge \log \frac{1}{2.4\delta},$$

where:
  - ν, ν' are any two instances for which the algorithm must output different results;
  - δ is the maximum allowed error probability;
  - N_a is the number of times a sample from ν_a is taken.

This result comes from a type of data processing inequality, and the proof also uses ideas from multiplicative change of measure

Lower Bounds Based on Communication Complexity

Very Brief Introduction to Communication Complexity

- **Communication complexity** is a major topic in theoretical computer science, and is not only of independent interest, but also has extensive uses in proving lower bounds

- **Setup:** *(2-agent 2-way case)*
  - Two agents Alice and Bob are given strings x and y respectively, and their goal is to compute some function $f(\mathrm{x},\mathrm{y})$
  - The communication complexity is the number of noiseless bits that need to be exchanged (summed over both directions) to achieve this
  - Allowing zero error probability can be too stringent, so it is common to allow randomization and to succeed with probability $1-\delta$
  - The randomness may be private to one agent, or common to both (public)

Very Brief Introduction to Communication Complexity

- **Communication complexity** is a major topic in theoretical computer science, and is not only of independent interest, but also has extensive uses in proving lower bounds

- **Setup:** *(2-agent 2-way case)*
  - Two agents Alice and Bob are given strings x and y respectively, and their goal is to compute some function $f(\mathrm{x}, \mathrm{y})$
  - The communication complexity is the number of noiseless bits that need to be exchanged (summed over both directions) to achieve this
  - Allowing zero error probability can be too stringent, so it is common to allow randomization and to succeed with probability $1 - \delta$
  - The randomness may be private to one agent, or common to both (public)

- **Example 1:** (EQUALS) If $f(\mathrm{x}, \mathrm{y}) = \mathbb{1}\{\mathrm{x} = \mathrm{y}\}$ with length-n strings, then:
  - With deterministic protocols, $\Omega(n)$ bits must be communicated
  - With common randomness, this drops to $O(\log n)$ or even $O(1)$, e.g., by sharing hash values and declaring 'YES' if they all match

- **Example 2:** (DISJOINT) If $f(\mathrm{x}, \mathrm{y}) = \mathbb{1}\{\{i : x_i = 1\}$ is disjoint from $\{i : y_i = 1\}\}$, then $\Omega(n)$ bits must be communicated even if randomization is allowed.

Example: Storage Requirements for Streaming Algorithms

- **Streaming distinct elements problem:** An algorithm processes a stream $a_1, \ldots, a_n$ of integers in $\{1, \ldots, n\}$ and seeks to output the number of distinct elements. The memory is limited and not all numbers can be stored.

- **Claim:** Any deterministic algorithm for this task requires $\Omega(n)$ memory.

Example: Storage Requirements for Streaming Algorithms

- **Streaming distinct elements problem:** An algorithm processes a stream $a_1, \dots, a_n$ of integers in $\{1, \dots, n\}$ and seeks to output the number of distinct elements. The memory is limited and not all numbers can be stored.
- **Claim:** Any deterministic algorithm for this task requires $\Omega(n)$ memory.
- **Proof outline:** Show that solving this problem implies solving EQUALS(x, y)
  - Let $L_x = \{i : x_i = 1\}$ and $L_y = \{i : y_i = 1\}$, so that "$x_i = 1$" means "integer i appears in the Alice's list" (similarly for y and Bob)
  - Bob computes the number L_y
  - Alice runs the streaming algorithm on L_x and passes the memory contents to Bob, who continues running it on L_y. Alice also sends Bob the number of distinct elements in L_x (using $O(\log n)$ bits).
  - Now Bob also knows the number of distinct elements in $L_x \cup L_y$
  - If the number of distinct elements in L_x, L_y, and $L_x \cup L_y$ are all the same, then he declares EQUALS(x, y)=1, otherwise 0.

Since EQUALS requires $\Omega(n)$ communication, it follows that distinct elements requires $\Omega(n)$ storage.

- Similar kinds of reductions are possible for randomized algorithms, but the reduction uses DISJOINT(x, y) instead of EQUALS(x, y).

Other Uses

Lower bounds based on communication complexity have appeared in many areas:

- Query complexity in property testing
- Number of measurements in compressive sensing problems
- Boolean circuit complexity
- Game theory (truthfulness vs. accuracy)
- ...

Summary

- **Summary:** Many useful lower bounding techniques from statistics, information theory, and theoretical computer science:
  - Fano's inequality
  - Le Cam's method
  - Assouad's method
  - Multiplicative change of measure
  - Communication complexity based

Summary

- **Summary:** Many useful lower bounding techniques from statistics, information theory, and theoretical computer science:
  - Fano's inequality
  - Le Cam's method
  - Assouad's method
  - Multiplicative change of measure
  - Communication complexity based

- **Many techniques/ideas not covered today:**
  - Direct analysis of the optimal estimator
  - Other tools from statistics (e.g., Cramér-Rao bound)
  - From high-dimensional probability (e.g., Sudakov's inequality)
  - Bounds for restricted algorithms (e.g., statistical query lower bounds)
  - Computational hardness (e.g., NP-hard, SETH-hard)

Summary

- **Summary:** Many useful lower bounding techniques from statistics, information theory, and theoretical computer science:
  - Fano's inequality
  - Le Cam's method
  - Assouad's method
  - Multiplicative change of measure
  - Communication complexity based
- **Many techniques/ideas not covered today:**
  - Direct analysis of the optimal estimator
  - Other tools from statistics (e.g., Cramér-Rao bound)
  - From high-dimensional probability (e.g., Sudakov's inequality)
  - Bounds for restricted algorithms (e.g., statistical query lower bounds)
  - Computational hardness (e.g., NP-hard, SETH-hard)
- **Further reading:**
  - Google theinformaticists lower bounds lecture X where $X \in \{1, \dots, 9\}$
  - Tsybakov's book Introduction to Nonparametric Estimation
  - John Duchi's lecture notes / Yihong Wu's lecture notes