Algorithmic Statistics
Lecture 1: Introduction & Uniformity Testing

What can we learn about the world by observing data? How much data do we need? What should we do with it?

The field of *statistics* developed from the early 1900s to answer these questions when datasets were gathered by hand and could be written on a few pieces of paper. But that is no longer the world we live in – datasets are huge and high-dimensional, and they demand tremendous computational resources to process. (Witness: as these notes are being written, the hyperscalars are on track to spend one third of a **trillion** dollars in 2025 alone building out compute infrastructure to train and serve data-driven artificial intelligence.)

This class is about the intersection of statistics and computation. We will adopt a theoretical computer science approach to reason rigorously about the guarantees of algorithms which learn from statistical data. We will study simple models and ask basic questions: *which statistical learning tasks can be accomplished in polynomial time? what are the basic principles for designing algorithms for those tasks? what assumptions about the world must we make* a priori *to believe the outputs of our algorithms?*

Today we will give some very simple examples to describe why we need this course in the first place – there are very simple statistical problems in high dimensions which are simply unsolvable!

1 Example 1: Polling

We ask n people independently whether they approve of a policy/candidate. Our goal is to estimate what fraction of the population as a whole approves. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Ber}(p)$. The natural estimator for p is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \mathbb{E}[\hat{p}] = p, \qquad \mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n} \leq \frac{1}{4n}.$$

Hence $\mathrm{Std}(\hat{p}) \leq \frac{1}{2\sqrt{n}}$, and to estimate p within ε (with constant confidence) it suffices to take $n = \Theta(1/\varepsilon^2)$. Recall that $\mathrm{TV}(P, Q) = \frac{1}{2}\sum_{x \in \mathcal{X}} |P(x) - Q(x)|$ is the *total variation distance* between distributions P and Q. Since the total variation distance between $\mathrm{Ber}(p)$ and $\mathrm{Ber}(p + \varepsilon)$ is $O(\varepsilon)$, an alternative perspective is that this estimator learns the distribution of X up to total variation distance ε using $O(1/\varepsilon^2)$ samples.

Is there a better estimator? Perhaps we can get away with $n = 1/\varepsilon^{1.99}$ samples?

2 Le Cam's Two-Point Method

Let P, Q be distributions over a finite domain $\mathcal{X}$. A (deterministic) test is a function $T : \mathcal{X}^n \to \{P, Q\}$. The error probability of T against the pair (P, Q) is

$$\max\Big\{ \Pr_{X\sim P^n}[T(X) = Q], \ \Pr_{X\sim Q^n}[T(X) = P]\Big\}.$$

Lemma 2.1 (Le Cam)**.** *For all tests T,*

$$\textit{error} \geq \frac{1}{2} - \mathrm{TV}(P^n, Q^n)\,.$$

Proof. Write $A = \{x : T(x) = P\}$, so $A^c = \{x : T(x) = Q\}$. Then

$$\begin{aligned}
\Pr_P[T(X) = Q] + \Pr_Q[T(X) = P] &= P(A^c) + Q(A) \\
&= 1 - P(A) + Q(A) \\
&= 1 - \big(P(A) - Q(A)\big) \\
&\geq 1 - \sup_{B\subseteq\mathcal{X}} |P(B) - Q(B)| \\
&= 1 - 2\mathrm{TV}(P, Q) \quad (\text{since } \mathrm{TV}(P, Q) = \tfrac{1}{2}\sup_B |P(B) - Q(B)|)
\end{aligned}$$

Dividing by 2 gives the claim. □

3 Lower Bound for Bernoulli Mean Estimation

Consider distinguishing $\mathrm{Ber}(1/2)$ from $\mathrm{Ber}(1/2+\varepsilon)$ using n i.i.d. samples. By Lemma 2.1, it suffices to upper bound $\mathrm{TV}(P^n, Q^n)$ where $P = \mathrm{Ber}(1/2)$ and $Q = \mathrm{Ber}(1/2 + \varepsilon)$. Now we introduce one of the first real technical ideas of the course: *tensorization*. It turns out that relating $\mathrm{TV}(P, Q)$ directly to $\mathrm{TV}(P^n, Q^n)$ is not so easy. Instead, it's better to go via a different measure of distance between P and Q, one which behaves well under taking an n-fold product.

Definition 3.1 (Kullback–Leibler divergence)**.** For distributions P, Q on $\mathcal{X}$,

$$\mathrm{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}.$$

Lemma 3.2 (Tensorization and Pinsker)**.** *For product distributions P^n, Q^n, $\mathrm{KL}(P^n\|Q^n) = n\,\mathrm{KL}(P\|Q)$; moreover $\mathrm{TV}(P, Q) \leq \sqrt{\tfrac{1}{2}\mathrm{KL}(P\|Q)}$.*

Lemma 3.3. *For $P = \mathrm{Ber}(1/2)$ and $Q = \mathrm{Ber}(1/2 + \varepsilon)$ with $|\varepsilon| \leq 1/4$,*

$$\mathrm{KL}(P\|Q) = 2\varepsilon^2 + O(\varepsilon^4).$$

Proof.

$$\begin{aligned}
\mathrm{KL}\big(\mathrm{Ber}(\tfrac{1}{2})\big\|\mathrm{Ber}(\tfrac{1}{2}+\varepsilon)\big) &= \tfrac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}+\varepsilon} + \tfrac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}-\varepsilon} \\
&= -\tfrac{1}{2}\log(1+2\varepsilon) - \tfrac{1}{2}\log(1-2\varepsilon) \\
&= -\tfrac{1}{2}\log(1-4\varepsilon^2) \\
&= 2\varepsilon^2 + O(\varepsilon^4),
\end{aligned}$$

using $\log(1-x) = -x - x^2/2 - \cdots$ for small x. □

Proposition 3.4 (Necessity of $n = \Omega(1/\varepsilon^2)$). *Any estimator that distinguishes* $\mathrm{Ber}(1/2)$ *from* $\mathrm{Ber}(1/2+\varepsilon)$ *with constant advantage requires* $n = \Omega(1/\varepsilon^2)$ *samples.*

Proof. By Lemmas 3.2 and 3.3,

$$\mathrm{TV}(P^n, Q^n) \le \sqrt{\tfrac{1}{2}\,\mathrm{KL}(P^n \| Q^n)} = \sqrt{\tfrac{1}{2}\,n\,\mathrm{KL}(P\|Q)} = \Theta(\sqrt{n}\,\varepsilon).$$

By Lemma 2.1, the error is at least $\frac{1}{2}(1 - \Theta(\sqrt{n}\,\varepsilon))$, which is $\ge 1/4$ unless $n = \Omega(1/\varepsilon^2)$. □

4 Uniformity Testing

For a simple low-dimensional distribution $\mathrm{Ber}(p)$, we could learn the distribution in total variation using a reasonable number of samples. What happens in high dimensions? It turns out that not only can we not learn a high dimensional distribution in total variation distance with a "reasonable" number of samples – we can't even tell if it is equal to one specific distribution: the uniform distribution.

Let $\mathcal{X}$ be a domain of size N. Given sample access to an unknown distribution P over $\mathcal{X}$, decide

$$H_0 : \; P = U(\mathcal{X}) \qquad \text{vs.} \qquad H_1 : \; \mathrm{TV}(P, U(\mathcal{X})) \ge \varepsilon.$$

Here $U(\mathcal{X})$ is the uniform distribution on $\mathcal{X}$.

Theorem 4.1 (Paninski). $\Theta\left(\frac{\sqrt{N}}{\varepsilon^2}\right)$ *samples are necessary and sufficient for uniformity testing.*

In these notes we give the *lower bound* proof. What does this theorem have to do with high-dimensional learning? Note that if we have an unknown distribution P on $\{0,1\}^d$, the domain size for this distribution is 2^d. Paninski's theorem tells us that we need $\Omega(2^{d/2})$ samples even to test if P is the uniform distribution. The intuition behind Paninski's theorem is that with $\ll \sqrt{N}$ samples we cannot tell the difference between $U([N])$ and the uniform distribution on a randomly chosen subset of half the support, since in either case with good probability no element is repeated in the list of samples.

4.1 Lower Bound via a Random-Half Construction

We will use Le Cam's two-point method. Of course we choose $P = U([N])^n$. What should be other distribution Q be? If we take Q to be n draws from a specific subset of half the elements of $[N]$, say $[N/2]$, then P and Q will be easy to distinguish with a constant number of samples – just check if all the samples are from $[N/2]$. Instead, we have a be a bit more clever about how we choose Q – we will use a random subset of half of the domain. For analysis purposes, we will pick this random half in a slightly structured way.

To define Q:

- Sample $Z_1, \ldots, Z_{n/2} \sim \pm 1$
- Define a distribution q on $[N]$ by $q_{2i} = (1 + Z_i\varepsilon)/2$ and $q_{2i-1} = (1 - Z_i\varepsilon)/2$.
- Draw n samples independently from Q.

Note that any q which can be obtained in the above procedure satisfies $\mathrm{TV}(U[N], q) \geq \varepsilon$. So if we had a good test for H_0 vs H_1, we would be able to distinguish P from Q.

Lemma 4.2. $\mathrm{TV}(P, Q) \leq O(\sqrt{\exp(O(n^2\varepsilon^4/N)) - 1})$

So, if $n \ll \sqrt{N}/\varepsilon^2$, then the TV distance is close to 0, and by Le Cam's, the error probability of any test remains at least, say, $1/4$.

We will sketch the proof of this lemma in a slightly different setting, for technical convenience.

Technical slight-of-hand: Poissonization Rather than drawing exactly n samples, we consider the setting where we draw $\widetilde{n} \sim \mathrm{Poi}(n)$ i.i.d. samples.

Definition 4.3 (Poisson Distribution)**.** A random variable X is said to follow a *Poisson distribution* with parameter $\lambda > 0$, denoted $X \sim \mathrm{Poi}(\lambda)$, if

$$\Pr[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

This leads to some appealing technical simplifications.

Poissonization facts. Let X_i be the number of occurrences of element $i \in [N]$ in the (random-size) sample.

- Under P: $X_1, \ldots, X_N$ are independent with $X_i \sim \mathrm{Poi}(\lambda)$ where $\lambda = n/N$.
- Under Q: the pairs (X_{2i-1}, X_{2i}) are independent across i, and

$$X_{2i-1} \sim \mathrm{Poi}(\lambda(1 + Z_i\varepsilon)), \qquad X_{2i} \sim \mathrm{Poi}(\lambda(1 - Z_i\varepsilon)).$$

Second-moment (chi-squared) calculation. Instead of KL divergence, it will be simpler to use another quantity which also tensorizes nicely and similarly upper-bounds the total variation distance, called the χ^2 divergence.

Definition 4.4 (χ^2-divergence)**.** For two distributions P and Q on a finite domain $\mathcal{X}$ with $P(x) > 0$ whenever $Q(x) > 0$, the *χ^2-divergence* of Q from P is

$$\chi^2(Q\|P) \;=\; \sum_{x\in\mathcal{X}} \frac{(Q(x) - P(x))^2}{P(x)} \;=\; \mathbb{E}_{x\sim P}\left[\left(\frac{Q(x)}{P(x)} - 1\right)^2\right].$$

Let L_Z be the likelihood ratio dQ_Z/dP for the Poissonized model. For a single bin with mean $\lambda(1+\delta)$ versus λ,

$$\frac{d\mathrm{Poi}(\lambda(1+\delta))}{d\mathrm{Poi}(\lambda)}(x) = e^{-\lambda\delta}(1+\delta)^x.$$

Therefore for a *pair* $(2i-1, 2i)$,

$$L_{Z,i} = (1 + Z_i\varepsilon)^{X_{2i-1}}(1 - Z_i\varepsilon)^{X_{2i}},$$

as the exponential terms cancel. The full likelihood ratio factorizes: $L_Z = \prod_{i=1}^{N/2} L_{Z,i}$.

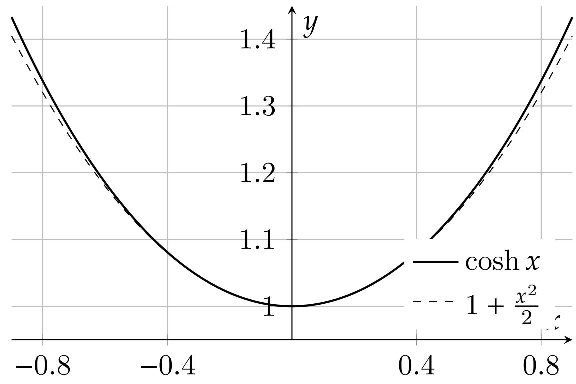

Figure 1: The cosh function and its quadratic approximation.

The chi-squared divergence of the mixture is

$$\chi^2(Q\|P) = \mathbb{E}_P\left[\left(\mathbb{E}_Z L_Z\right)^2\right] - 1 = \mathbb{E}_{Z,\tau}\left[\prod_{i=1}^{N/2} \mathbb{E}_P\left[L_{Z,i} L_{\tau,i}\right]\right] - 1.$$

For a fixed pair i and fixed $Z_i, \tau_i \in \{\pm 1\}$, using that if $X \sim \mathrm{Poi}(\lambda)$ then $\mathbb{E}[(1+\alpha)^X] = \exp(\lambda\alpha)$,

$$\begin{aligned}\mathbb{E}_P\left[L_{Z,i}L_{\tau,i}\right] &= \mathbb{E}\left[(1+Z_i\varepsilon)^{X_{2i-1}}(1+\tau_i\varepsilon)^{X_{2i-1}}\right] \cdot \mathbb{E}\left[(1-Z_i\varepsilon)^{X_{2i}}(1-\tau_i\varepsilon)^{X_{2i}}\right] \\ &= \exp\left(\lambda\big((1+Z_i\varepsilon)(1+\tau_i\varepsilon)-1\big)\right) \cdot \exp\left(\lambda\big((1-Z_i\varepsilon)(1-\tau_i\varepsilon)-1\big)\right) \\ &= \exp\left(2\lambda\, Z_i\tau_i\, \varepsilon^2\right).\end{aligned}$$

Averaging over Z_i, τ_i (independent uniform signs) gives

$$\mathbb{E}_{Z_i,\tau_i}\left[\mathbb{E}_P[L_{Z,i}L_{\tau,i}]\right] = \tfrac{1}{2}\left(e^{2\lambda\varepsilon^2} + e^{-2\lambda\varepsilon^2}\right) = \cosh(2\lambda\varepsilon^2).$$

By independence across pairs,

$$1 + \chi^2(Q\|P) = \left(\cosh(2\lambda\varepsilon^2)\right)^{N/2}.$$

For small x, $\cosh x = 1 + x^2/2 + O(x^4)$; with $x = 2\lambda\varepsilon^2$ this yields

$$\chi^2(Q\|P) = \left(1 + 2\lambda^2\varepsilon^4 + O(\lambda^4\varepsilon^8)\right)^{N/2} - 1 = \exp\left(\Theta\Big(\frac{n^2\varepsilon^4}{N}\Big)\right) - 1.$$

From χ^2 to total variation. Using $\mathrm{TV}(Q,P) \le \frac{1}{2}\sqrt{\chi^2(Q\|P)}$, we obtain

$$\mathrm{TV}(\overline{Q}, P) \;\le\; \tfrac{1}{2}\sqrt{\exp\Big(\Theta\big(\tfrac{n^2\varepsilon^4}{N}\big)\Big) - 1},$$

□

Remark. De-Poissonization changes constants only, so the same lower bound holds for a fixed sample size n.

5 Acknowledgements

These notes are based on handwritten notes by Costis Daskalakis from the 2023 version of the course.