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 hyperscalers 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

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

Hence $\mathrm{Std}(\hat{p}) \le \frac{1}{2\sqrt{n}}$, 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 $\mathcal{X}$,

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

The KL divergence is the expected log likelihood ratio between P and Q, with the expectation taken under P. We could spend several lectures discussing the meaning of KL divergence, but we don't have time in this course – take an information theory course!

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 $|\varepsilon| \leq 1/4$,*

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

Proof.

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

using $\log(1-x) = -x - O(x^2)$ 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,

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

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

4 Uniformity Testing and Learning on the Hypercube

In the polling example, every member of the population we drew samples from had just one feature – supporting vs not supporting the candidate/policy in question. In this course we are primarily concerned with high-dimensional populations/distribution. For example – images, documents, videos, cryptographic keys, ... The canonical high-dimensional "universe" is the d-dimensional hypercube $\{0,1\}^d$. Mathematically, a population of d-bit individuals will be represented as a distribution P on $\{0,1\}^d$.

Gold Standard: Learning in Total Variation Distance The most ambitious goal we could have is to learn such a distribution P in total variation distance – meaning that after looking at some samples from P, we find a distribution $\hat{P}$ on $\{0,1\}^d$ such that $\mathrm{TV}(P,\hat{P}) \le \varepsilon$. Such a model $\hat{P}$ will let us answer any question about the population P which we choose to pose, with high accuracy, without observing any more samples.

More formally, for any 0/1-valued question we can ask about the population (what fraction have attribute A? what fraction have feature 1 correctly predicted by the best linear predictor using features $2-d$? ...), we can estimate the true answer to the question using $\hat{P}$, since $\mathrm{TV}(P,\hat{P}) = \sup_{f:\{0,1\}^d\to[0,1]} |\mathbb{E}_P f - \mathbb{E}_{\hat{P}} f|$.

Impossibility of Learning in Total Variation Unfortunately this is an impossible goal, unless we get to see $\Omega(2^d)$ samples, for any nontrivial value of ε. We will argue why only very informally, since we are about to prove an even stronger result formally. The hypercube is *big* – there are 2^d strings of length d, so to specify P requires 2^d numbers. So we need to observe at least 2^d numbers – each sample gives us d numbers, at most.

5 Uniformity Testing

Learning in total variation distance is too ambitious. Perhaps there are simpler things we can learn about a high dimensional distribution using only $d^{O(1)}$ samples? There are, but it is not so trivial to see which ones – that is part of the purpose of this class. Let's an example of a seemingly simpler problem which still cannot be solved with fewer than exponentially-many samples.

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

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

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

For example, if someone claims to you that they have a source of true randomness generating uniform samples from $\{0,1\}^d$[1], and you want to see if they are lying, this is the hypothesis test you want to perform.

Theorem 5.1 (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 $\{0,1\}^d$, 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.

5.1 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 $q_{2i} = (1 + Z_i \varepsilon)/N$ and $q_{2i-1} = (1 - Z_i \varepsilon)/N$.
- 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 5.2. $\mathrm{TV}(P, Q) \leq O(\sqrt{\exp(O(n^2 \varepsilon^4/N)) - 1})$

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.

[1] https://scitechdaily.com/better-cybersecurity-with-a-new-quantum-random-number-generator/

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

Definition 5.3 (Poisson Distribution)**.** A random variable X is said to follow a *Poisson distribution* with parameter $\lambda > 0$, denoted $X \sim \text{Poi}(\lambda)$, 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 $X_i \sim \text{Poi}(\lambda)$ where $\lambda = n/N$.
- Under Q: the pairs (X_{2i-1}, X_{2i}) are independent across i, and

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

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 5.4 (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

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

Fact 5.5. $\text{TV}(P, Q) \leq \frac{1}{2}\sqrt{\chi^2(P\|Q)}$

Exercise (on problem set 1) Using the above facts about Poissonization and χ^2 divergence freely, finish the proof of Paninski's sample complexity lower bound (poissonized variant) by proving the poissonized version of Lemma 5.2.

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

6 Acknowledgements

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