COMP 585 **Spring 2026**
Probabilistic Toolkit for Learning and Computing *Instructor: Maryam Aliakbarpour*
February 3, 2026 *Scriber: Maryam Aliakbarpour*

Lecture 7

Estimating the ℓ_2 Distance between Two Discrete Distributions

A fundamental task in distribution testing is estimating the distance between two distributions p and q given sample access. While ℓ_1 distance is the standard metric for closeness testing, it is analytically difficult to work with directly. The ℓ_2 distance, however, is much more tractable. It serves as a crucial algorithmic primitive: if we can accurately estimate $\|p-q\|_2^2$, we can use it to distinguish whether $p = q$ or $\|p-q\|_1 \geq \epsilon$.

In this lecture, we develop an unbiased estimator for the squared ℓ_2 distance. The high-level approach is to draw two sample sets from p and q and design a statistic that can capture the difference between the histograms of these samples. To make this analysis feasible, our first step is to use the Poissonization trick that we discussed in the previous lecture.

Poissonized sampling model. Instead of drawing a fixed number of samples m, which introduces dependencies between the counts of different elements, we adopt the Poissonized model we discussed in the previous lecture. For distributions p, q over $[n]$:

- Draw $\hat{m} \sim \text{Poi}(m)$ and $\hat{m}' \sim \text{Poi}(m)$.
- Take $\hat{m}$ samples from p and $\hat{m}'$ samples from q.
- Let X_i and Y_i be the number of occurrences of element i in each sample set.

By the properties of Poissonization, the variables $\{X_1, \ldots, X_n, Y_1, \ldots, Y_n\}$ are all mutually independent, with $X_i \sim \text{Poi}(mp_i)$ and $Y_i \sim \text{Poi}(mq_i)$. This independence is the technical engine that allows us to analyze the variance of our estimator coordinate-wise.

The Statistic for ℓ_2-Distance. We define our statistic Z as a function of the frequencies we observed in the sample sets:

$$Z := \sum_{i=1}^{n} \left((X_i - Y_i)^2 - X_i - Y_i \right) . \tag{1}$$

We will see that Z constitutes an unbiased estimator of the ℓ_2-distance between the two distributions (up to a normalization factor). Like the Pearson's χ^2 statistic, the quadratic term $(X_i - Y_i)^2$ prioritizes elements with larger discrepancies, amplifying the signal where p_i and q_i diverge.

Crucially, the correction term $-(X_i+Y_i)$ removes the "artificial" distance caused by sampling noise. In sub-linear regimes where we have fewer samples than the domain size of the distribution, even if $p = q$, one often observes an element in one sample set but not the other (e.g., $X_i = 1, Y_i = 0$). The correction term essentially allows the statistic to ignore such occurrences and reduces the variance by ensuring these random fluctuations do not accumulate. This leads to an estimator that is centered at zero whenever the distributions are identical.

Analyzing Mean. We first verify that Z behaves correctly in expectation.

Lemma 1. $\mathbf{E}[Z] = m^2 \|p - q\|_2^2$.

Proof. Recall that for a Poisson random variable, the variance and the mean are identical. Thus, for $W \sim \mathrm{Poi}(\lambda)$ we have:

$$\mathbf{Var}[W] = \mathbf{E}\big[W^2\big] - \mathbf{E}[W]^2 \quad \Longrightarrow \quad \mathbf{E}\big[W^2\big] = \lambda + \lambda^2 \,.$$

By independence of the X_i's and Y_i's, we obtain:

$$\begin{aligned}
\mathbf{E}\big[(X_i - Y_i)^2 - X_i - Y_i\big] &= \mathbf{E}\big[X_i^2\big] - 2\mathbf{E}[X_i]\mathbf{E}[Y_i] + \mathbf{E}\big[Y_i^2\big] - \mathbf{E}[X_i] - \mathbf{E}[Y_i] \\
&= (m^2 p_i^2 + m p_i) - 2m^2 p_i q_i + (m^2 q_i^2 + m q_i) - m p_i - m q_i \\
&= m^2(p_i^2 - 2p_i q_i + q_i^2) = m^2 (p_i - q_i)^2 .
\end{aligned}$$

Summing over i yields $\mathbf{E}[Z] = \sum_{i=1}^n m^2 (p_i - q_i)^2 = m^2 \|p - q\|_2^2$. □

Analyzing Variance. Next, we focus on bounding the variance of Z which will help us to show concentration later.

Lemma 2. *Let* $b \coloneqq \max(\|p\|_2^2, \|q\|_2^2)$. *Then,*

$$\mathbf{Var}[Z] \le 8m^2 b + 8m^3 \|p - q\|_4^2 \sqrt{b} \,.$$

Proof sketch. The proof is largely computational and follows from a routine expansion using standard properties of Poisson moments. While the algebra is extensive, the underlying logic provides a standard template for analyzing Poissonized statistics. The key observation is that under Poissonization, the terms $Z_i \coloneqq (X_i - Y_i)^2 - X_i - Y_i$ are mutually independent, allowing us to decompose the total variance as $\mathbf{Var}[Z] = \sum_{i=1}^n \mathbf{Var}[Z_i]$.

To compute $\mathbf{Var}[Z_i]$, we first rewrite the term as:

$$Z_i = (X_i^2 - X_i) + (Y_i^2 - Y_i) - 2X_i Y_i \,.$$

Using the identity $\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$, we expand $\mathbf{Var}[Z_i]$ into a polynomial in X_i and Y_i of degree up to four. By applying the linearity of expectation and substituting the known moments of the Poisson distribution, the expression simplifies significantly.

To illustrate this mechanical process, consider the term $\mathbf{E}[(X_i^2 - X_i)^2]$. Expanding the square within the expectation yields:

$$\mathbf{E}\big[(X_i^2 - X_i)^2\big] = \mathbf{E}\big[X_i^4 - 2X_i^3 + X_i^2\big] = \mathbf{E}\big[X_i^4\big] - 2\mathbf{E}\big[X_i^3\big] + \mathbf{E}\big[X_i^2\big]$$

Substituting the raw moments of $X_i \sim \text{Poi}(mp_i)$:

$$\begin{aligned}\mathbf{E}\big[(X_i^2 - X_i)^2\big] &= \big((mp_i)^4 + 6(mp_i)^3 + 7(mp_i)^2 + mp_i\big) \\ &\quad - 2\big((mp_i)^3 + 3(mp_i)^2 + mp_i\big) \\ &\quad + \big((mp_i)^2 + mp_i\big) \\ &= (mp_i)^4 + 4(mp_i)^3 + 2(mp_i)^2\end{aligned}$$

We skip the calculation of the rest of the terms in $\mathbf{Var}[Z_i]$, but repeating this procedure for the cross-terms and summing the results, we find:

$$\mathbf{Var}[Z_i] = 4m^3(p_i - q_i)^2(p_i + q_i) + 2m^2(p_i + q_i)^2$$

To arrive at the final lemma, we bound the summation $\sum_{i=1}^n \mathbf{Var}[Z_i]$ in two parts. First, for the m^2 term, we use the inequality $(p_i + q_i)^2 \le 2(p_i^2 + q_i^2)$:

$$\begin{aligned}\sum_{i=1}^n 2m^2(p_i + q_i)^2 &\le 4m^2 \sum_{i=1}^n (p_i^2 + q_i^2) \\ &\le 4m^2(2\max(\|p\|_2^2, \|q\|_2^2)) = 8m^2 b\end{aligned}$$

Second, we bound the m^3 term by applying the Cauchy-Schwarz inequality:

$$\begin{aligned}\sum_{i=1}^n 4m^3(p_i - q_i)^2(p_i + q_i) &\le 4m^3 \sqrt{\sum_{i=1}^n (p_i - q_i)^4} \sqrt{\sum_{i=1}^n (p_i + q_i)^2} \\ &= 4m^3 \|p - q\|_4^2 \sqrt{\|p + q\|_2^2}\end{aligned}$$

Using the property $\|p+q\|_2^2 \le 2(\|p\|_2^2 + \|q\|_2^2) \le 4b$, the square root term becomes $\sqrt{4b} = 2\sqrt{b}$. Substituting this back gives:

$$4m^3 \|p - q\|_4^2 (2\sqrt{b}) = 8m^3 \|p - q\|_4^2 \sqrt{b}$$

Combining these two upper bounds completes the proof. □

Concentration via Chebyshev. To ensure $\frac{Z}{m^2}$ is an accurate empirical estimate, it must concentrate around its mean: the ℓ_2^2-distance between p and q. We apply *Chebyshev's*

inequality to bound the probability that our estimate deviates from the mean by more than ϵ^2:

$$\mathbf{Pr}\left[\left|\frac{Z}{m^2} - \|p-q\|_2^2\right| \geq \epsilon^2\right] = \mathbf{Pr}\big[|Z - \mathbf{E}[Z]| \geq m^2\epsilon^2\big] \leq \frac{\mathbf{Var}[Z]}{m^4\epsilon^4}.$$

Substituting the variance bound, we obtain:

$$\mathbf{Pr}\left[\left|\frac{Z}{m^2} - \|p-q\|_2^2\right| \geq \epsilon^2\right] \leq \frac{8m^3\|p-q\|_4^2\sqrt{b} + 8m^2 b}{m^4\epsilon^4} = \frac{8\|p-q\|_4^2\sqrt{b}}{m\epsilon^4} + \frac{8b}{m^2\epsilon^4}$$

This bound indicates that the estimator's reliability is tied to the unknown distribution-dependent quantities $\|p-q\|_4^2$ and b. In a setting where we cannot assume prior knowledge of p and q, we must rely on a worst-case analysis. A naive bound, setting these quantities to their global maximum of 1, suggests a sample complexity of $m = \Omega(\epsilon^{-4})$ to achieve a constant failure probability. However, this expression highlights that more favorable sample complexities are possible if the distributions are restricted. Specifically, if we assume the distributions are relatively "flat" (e.g., $b = O(1/n)$), we can achieve a more desirable sample complexity. This motivates the *flattening* technique we will discuss in the next lecture.

Closeness Testing via ℓ_2-Distance Estimation

The goal of closeness testing is to determine, given sample access to two distributions p and q over a domain of size n, whether $p = q$ or $\|p-q\|_1 > \epsilon$. As previously derived, the expectation of this statistic is $\mathbf{E}[Z] = m^2\|p-q\|_2^2$, making $\frac{Z}{m^2}$ an unbiased estimator for the squared ℓ_2-distance. To distinguish between the two cases, we use the relationship between the ℓ_1 and ℓ_2 norms, specifically $\|p-q\|_1^2 \leq n\|p-q\|_2^2$, and show that our ℓ_2-distance estimator can be used for this task:

- **Case 1: Distributions are identical ($p = q$).** If the distributions are equal, then $\|p-q\|_1 = 0$, which implies $\|p-q\|_2^2 = 0$. In this scenario, the expected value of our estimator is $\mathbf{E}[Z/m^2] = 0$.
- **Case 2: Distributions are far ($\|p-q\|_1 > \epsilon$).** If the distributions are ϵ-far in L_1 distance, the Cauchy-Schwarz inequality provides a lower bound on the squared ℓ_2 distance:

  $$\left(\sum_{i=1}^{n}|p_i - q_i|^2\right)\cdot\left(\sum_{i=1}^{n}1\right) \geq \left(\sum_{i=1}^{n}|p_i - q_i|\right)^2$$

  Thus, we get:

  $$\|p-q\|_2^2 \geq \frac{\|p-q\|_1^2}{n} > \frac{\epsilon^2}{n}.$$

  Here, the expected value $\mathbf{E}[Z/m^2]$ is at least ϵ^2/n.

That is, there is a gap between the expected value of the Z/m^2 in the two cases. To distinguish these cases, we set a threshold $T := \frac{\epsilon^2}{2n}$, which lies exactly at the midpoint of the gap between the two possible expectations. We accept the hypothesis that $p = q$ if $Z/m^2 \leq T$ and reject otherwise.

As illustrated in Figure 1, the classification task is successful when the variance of our estimator is small enough that the densities of Z/m^2 under Case 1 (centered at 0) and Case 2 (centered at $\|p-q\|_2^2$) place negligible probability mass on the wrong side of the threshold T.

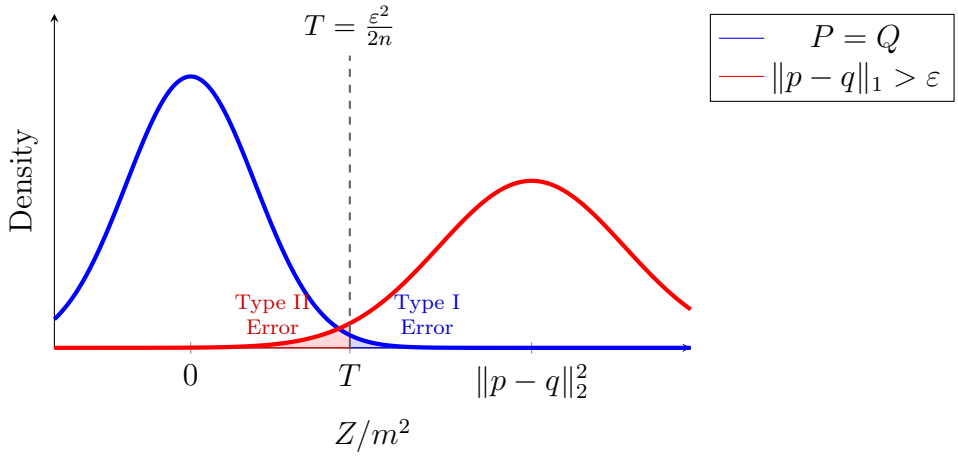

Figure 1: Concentration of the estimator Z/m^2. The blue distribution ($P = Q$) has a non-negligible tail exceeding the threshold T. The red distribution is significantly wider, illustrating the higher variance typically encountered when the distributions are far.

Concentration Analysis. To show this tester succeeds with high probability, we bound the error probabilities in both cases using Chebyshev's inequality.

1. **Type I Error (falsely outputting reject):** When $p = q$, the expected value $\mathbf{E}[Z/m^2] = 0$. An error occurs if the estimate Z/m^2 exceeds the threshold $T = \frac{\epsilon^2}{2n}$. Since we need the deviation to be at least T:

$$\mathbf{Pr}\left[\frac{Z}{m^2} \geq \frac{\epsilon^2}{2n}\right] \leq \frac{\mathbf{Var}[Z]}{m^4(\frac{\epsilon^2}{2n})^2} = \frac{4n^2\mathbf{Var}[Z]}{m^4\epsilon^4}.$$

   Using our variance bound $\mathbf{Var}[Z] \leq 8m^2b$ (noting that $\|p-q\|_4 = 0$ in this case), we have:

$$\frac{4n^2(8m^2b)}{m^4\epsilon^4} = \frac{32n^2b}{m^2\epsilon^4} \leq \frac{1}{10}; \Leftarrow; m = \Omega\left(\frac{n\sqrt{b}}{\epsilon^2}\right).$$

2. **Type II Error (falsely outputting accept):** When $\|p-q\|_1 > \epsilon$, we have $\mathbf{E}[Z/m^2] = \|p-q\|_2^2 \geq \frac{\epsilon^2}{n}$. An error occurs if $Z/m^2 \leq T$. Since T is half the minimum

expected value, this implies a deviation of at least half the expectation:

$$\mathbf{Pr}\left[\frac{Z}{m^2} \leq T\right] \leq \mathbf{Pr}\left[\left|\frac{Z}{m^2} - \|p-q\|_2^2\right| \geq \frac{\|p-q\|_2^2}{2}\right] \leq \frac{4\mathbf{Var}[Z]}{m^4 \|p-q\|_2^4}.$$

Substituting the full variance $\mathbf{Var}[Z] \leq 8m^3\|p-q\|_4^2\sqrt{b} + 8m^2 b$, the error probability is bounded by:

$$\begin{aligned}\mathbf{Pr}\left[\frac{Z}{m^2} \leq T\right] &= \frac{32\|p-q\|_4^2\sqrt{b}}{m\|p-q\|_2^4} + \frac{32b}{m^2\|p-q\|_2^4} \\ &\leq \frac{32\sqrt{b}}{m\|p-q\|_2^2} + \frac{32b}{m^2\|p-q\|_2^4} \\ &\leq \frac{32n\sqrt{b}}{m\epsilon^2} + \frac{32bn^2}{m^2\epsilon^4}.\end{aligned}$$

In the first inequality above, we use the ℓ_p inequality. In particular, $\|p-q\|_4 \leq \|p-q\|_2$. In the second inequality, we use the lower bound ϵ^2/n for $\|p-q\|_2^2$. Setting this expression to be $\leq \frac{1}{10}$ yields the same sample complexity requirement of $m = \Theta\left(\frac{n\sqrt{b}}{\epsilon^2}\right)$.

Bibliographic Note: The study of distribution testing has its origins in classical statistics [Pea00, NP33], most notably in the work of Pearson, who introduced the χ^2 test for goodness-of-fit. However, classical theory typically focuses on the asymptotic regime. The modern *property testing* treatment of these problems—focusing on the sublinear regime where $m \ll n$—was initiated by [GR00, BFR+13]. [BFR+13] was the first to formalize closeness testing (the two-sample problem) and introduced the framework of using ℓ_2 distance as a proxy for ℓ_1 distance. The specific ℓ_2 estimator analyzed in this lecture is presented in [CDVV14], and further utilized in [DK16] for closeness testing.

References

[BFR+13] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *JACM*, 60(1):4:1–4:25, 2013.

[CDVV14] Siu-on Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[DK16] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 685–694, 2016.

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electron. Colloquium Comput. Complex.*, TR00-020, 2000.

[NP33] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[Pea00] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.