CPSC 448 Proposal

Ayush Vora, Nathaniel Harms

December 2025

The primary objective of this directed study is to learn about the "moment-matching" technique, and a few important use-cases of it. One of which is to prove a lower bound on the number of random samples required to make certain decisions about probability distributions. I will show my understanding by creating a lecture note-style survey paper explaining the technique. If there is more time available, we will also attempt to prove a lower bound on the "support size" testing problem shown below.

1 The Moment Matching Technique

In [2], we see a lower bound on the support-size testing problem. This lower bound is proven using the moment matching technique. For this technique, we must first construct two random variables: X_1, X_2. Let X_1, X_2 take on positive integers as values, and let their expectations be very different. For the moment matching technique, we must also require the condition that for the first k moments of X_1, X_2,

$$\frac{E[X_1]}{E[X_2]} = \frac{E[X_1^2]}{E[X_2^2]} = ... = \frac{E[X_1^k]}{E[X_2^k]}$$

If we create two random variables that satisfy these conditions for a large enough value of k, we can obtain lower bounds for making decisions about probability distributions. The goal of this project is to understand why and how this can be achieved in different settings.

2 The Support Size Testing Problem

The support size testing problem focuses on determining the size of the support of a certain probability distribution. We assume that we do not know exactly what the probability distribution is, but that we can draw samples from it as needed. Because we do not know the size of the support with certainty, we want to get a good approximation for it.

Definition 2.1 (Total Variation Distance)**.** For probability distributions P and Q, the total variation distance between P and Q, is defined as

$$\delta(P,Q) = \max_{E \subseteq \mathbb{N}} |P(E) - Q(E)| = \frac{1}{2} \sum_{i \in \mathbb{N}} |P(i) - Q(i)|$$

Definition 2.2 (ε-far)**.** We say that a probability distribution P is ε-far from $|\operatorname{supp}(p)| \leq k$ if, $\forall$ probability distributions Q such that $|\operatorname{supp}(Q)| \leq k$, the total variation distance $\delta(P,Q) \geq \varepsilon$.

Let P be a discrete probability distribution taking values in $[n], n \in \mathbb{N}$. Given $k, \varepsilon \in \mathbb{N}$, and sample access to P, we want to create an algorithm $A(k, \varepsilon, S)$ such that

$$A(k,\varepsilon,S) = \begin{cases} \textsc{Yes} & |\operatorname{supp}(P)| \leq k \\ \textsc{No} & p \text{ is } \varepsilon\text{-far from } |\operatorname{supp}(P)| \leq k \end{cases}$$

and $\mathbb{P}(A \text{ is correct}) = 2/3$.

It has been shown that s, the number of samples required for A, has upper and lower bounds

$$\frac{k}{\varepsilon \log(k)} \leq s \leq O\left(\min\left(\frac{k\log(1/\varepsilon)}{\varepsilon\log(k)}, \frac{k}{\varepsilon}\right)\right)$$

Showing that $s \leq O(k/\varepsilon)$ can be done with some manipulation of P and Markov's Inequality. If my directed studies goes well, my next steps will be to find an optimal lower bound for the number of samples required. Alternatively, it is assumed, but not proven, that the $\log(1/\varepsilon)$ term is not necessary in the upper bound. We can investigate further and see if we can tighten the bound.

3 Deadlines

This directed studies will take place from September to December, 2025. By the end of the term (Dec 20, 2025), the goals listed above will be finished. I will be showing my work by presenting my findings to the professor, creating a lecture-note's style survey paper with condensed and easy-to-navigate notes, and creating example practice exercises to test the understanding of someone who would want to learn this topic as well.

4 Reading List

- Strong Lower Bounds for Approximating Distribution Support Size and the Distinct Elements Problem [2]
- Polynomial Methods in Statistical Inference: Theory and Practice [5]
- Estimating the unseen: an n/log(n)-sample estimator for entropy and support size [3]

- Chebyshev polynomials, moment matching, and optimal estimation of the unseen [4]
- A survey on distribution testing: Your data is big. But is it blue? [1]

References

[1] Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

[2] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

[3] Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 685–694, 2011.

[4] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019.

[5] Yihong Wu, Pengkun Yang, et al. Polynomial methods in statistical inference: Theory and practice. *Foundations and Trends® in Communications and Information Theory*, 17(4):402–586, 2020.