All-Purpose Mean Estimation over $\mathbb{R}$

Jasper C.H. Lee

University of California, Davis
jasperlee@ucdavis.edu

Abstract

Given society's increasing reliance on data, its collection and processing into useful information is a technical problem of growing focus, and perhaps paradoxically, a critical bottleneck in many data science and machine learning applications. Yet, even for the most basic statistical problems such as mean estimation, there is a theory-practice divide. Conventional methods like the sample mean, while supported by theoretical results under strong assumptions, are often brittle in the presence of extreme data. Practitioners thus often use ad-hoc and unprincipled "outlier removal" heuristics, but which can lead to wrong conclusions (e.g. Milikan's underestimation of the electron charge).

In this talk, I will describe my work that essentially resolves the fundamental 1-d mean estimation problem. I will show the construction of a statistically-optimal and computationally-efficient 1-dimensional mean estimator, whose estimation error is optimal even in the leading multiplicative constant, under bare minimum distributional assumptions (FOCS 2021). Furthermore, I will discuss its various robustness properties (ICML 2025 Oral), in particular highlighting robustness to adversarial sample corruption.

Setup

Mean estimation is one of the most fundamental statistical tasks and primitives, underlying many statistical methods as well as randomized algorithms in computer science (Mitzenmacher and Upfal 2017). Even in 1 dimension, mean estimation has plenty of natural and direct applications (e.g. drug trials (Bolton and Bon 2003), social science (Hanushek and Jackson 2013)). Concretely, the problem is as follows.

Problem 1. Assume there is a 1-d distribution D with mean μ and variance σ^2, both unknown. Given n i.i.d. samples from D, the task is to produce an estimate $\hat{\mu}$ such that $|\hat{\mu} - \mu| \leq \epsilon$ for some ϵ with $1 - \delta$ probability. The algorithmic problem then is to minimize the estimation error ϵ.

Optimal Sub-Gaussian 1-D Mean Estimation

My work with Paul Valiant in FOCS 2021 (Lee and Valiant 2022) resolved the above fundamental problem, achieving the smallest possible ϵ even in the leading multiplicative constant.

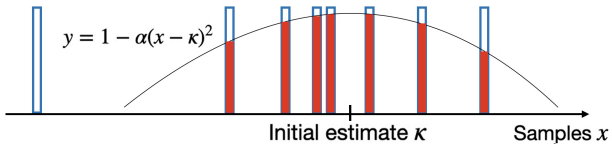

Figure 1: Diagram for the (Lee and Valiant 2022) estimator

Figure 1 is a diagrammatic representation of our estimator. In essence, it trims each sample in a fractional manner using a carefully-computed quadratic parabola centered at an initial estimate. The farther a sample is from the initial estimate, the more of it is trimmed, before the algorithm returns the sample mean of the remaining weighted samples.

In a recent ICML 2025 paper (Lee et al. 2025), we further showed that the same estimator, without any change to its structure or parameters, is robust in a variety of contexts, including when the input samples can be adversarially corrupted. The strong optimal guarantee is captured in Fact 1.

Fact 1 ((Lee et al. 2025))**.** Given n samples from a distribution with mean μ and variance σ^2, and parameter $\delta > 0$, if at most ηn samples are corrupted for $\eta \leq \frac{1}{24n} \log \frac{2}{\delta}$, then the Lee-Valiant estimator has error

$$|\hat{\mu} - \mu| \leq \sigma \cdot \left((1 + o(1)) \sqrt{\frac{2 \log \frac{2}{\delta}}{n}} + O(\sqrt{\eta}) \right)$$

Here, the $o(1)$ term tends to 0 as $\left(\frac{\log \frac{2}{\delta}}{n}, \delta \right) \to (0, 0)$ and, crucially, is independent of D.

The above result is "tight up to a $1 + o(1)$ factor" in the sense that no algorithm can achieve error less than $(1 - o(1)) \cdot \sigma \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}$ across all distributions, even in the absence of data corruption. Furthermore, the extra $O(\sqrt{\eta})$ term is also well-known in the literature to be optimal (Diakonikolas and Kane 2023).

Combined with further results we showed in the paper, including the robustness of the estimator to infinite-variance distributions, we argue that our mean estimator is "all-purpose", and should replace the conventional sample mean in standard statistical practice.

References

Bolton, S.; and Bon, C. 2003. *Pharmaceutical statistics: Practical and clinical applications, revised and expanded*. CRC press.

Diakonikolas, I.; and Kane, D. M. 2023. *Algorithmic high-dimensional robust statistics*. Cambridge university press.

Hanushek, E. A.; and Jackson, J. E. 2013. *Statistical methods for social scientists*. Academic Press.

Lee, J. C. H.; McKelvie, W.; Song, M.; and Valiant, P. 2025. All-Purpose Mean Estimation over R: Optimal Sub-Gaussianity with Outlier Robustness and Low Moments Performance. In *Proc. ICML'25*.

Lee, J. C. H.; and Valiant, P. 2022. Optimal Sub-Gaussian Mean Estimation in R. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 672–683. IEEE.

Mitzenmacher, M.; and Upfal, E. 2017. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press.