A Framework for Testing Properties of Discrete Distributions: Monotonicity, Independence, and More

Jayadev Acharya
EECS, MIT
jayadev@csail.mit.edu

Constantinos Daskalakis
EECS, MIT
costis@mit.edu

Gautam Kamath
EECS, MIT
g@csail.mit.edu

ABSTRACT

Given data sampled from an unknown discrete probability distribution p, does the underlying distribution possess some property of interest? For instance, is the distribution uniform? Monotone? Are its marginals independent? This class of problems is one of the most fundamental questions in statistics, where it is known as hypothesis testing.

Classical work on this problem has focused on distributions over a domain of a fixed size, as the number of samples goes to infinity. However, in modern scenarios, distributions may be over massive domains, and we are often limited by the number of samples or computational power. As such, over the past two decades, there has been intense study with these goals in mind (see [2] for a survey). Nevertheless, even for many basic properties of distributions, the optimal sample complexity was unknown.

We provide a testing framework which achieves the optimal sample complexity for the properties mentioned above, and more. Notably, all properties we study have strongly sublinear complexities, requiring only a number of samples proportional to the square root of the domain size. The framework follows a conceptually simple learn-then-test approach. Naively, such methods seemed to be intrinsically statistically inefficient, due to an information-theoretic lower bound for robust ℓ_1 identity testing [3]. We bypass this lower bound by using χ^2 distance as an intermediary metric. χ^2 is a non-uniform rescaling of ℓ_2, and is more "punishing" than ℓ_1. This makes learning in χ^2 slightly harder, but testing in χ^2 drastically easier.

This work appeared at NIPS'15 as [1].

BODY

Wanna optimally test if your distribution has a property? Assume it does, learn the distribution, and test the hypothesis. Use χ^2 distance!

REFERENCES

[1] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

[2] C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22(63), 2015.

[3] G. Valiant and P. Valiant. Estimating the unseen: An $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 685–694, New York, NY, USA, 2011. ACM.

Volume 4 of Tiny Transactions on Computer Science