Advances in Nonparametric Hypothesis Testing

Aaditya Ramdas
Machine Learning Department
Carnegie Mellon University
aramdas@cs.cmu.edu

Executive Summary

The advent of "big data" is causing a merging of fields, due to the need of *simultaneously* solving a variety of problems which are usually in distinct domains. As someone with a Bachelors degree in computer science, my PhD involved a rapid change in mindset, as I quickly embraced the necessity of using statistics and optimization *together* for effectively solving modern data problems. Indeed, this spirit is captured by my wide range of successful research projects, exploring the depths of multiple related fields with collaborators from the Department of Statistics, Machine Learning Department and the Tepper School of Business, together informing my overall research agenda.

My research goal involves simultaneously addressing statistical & computational tradeoffs encountered in modern data analysis and high-dimensional machine learning (eg: hypothesis testing, regression, classification). My future interests include incorporating additional constraints like privacy or communication, and settings involving hidden utilities of multiple cooperative agents or competitive adversaries.

Along with the recent explosion in quantities of available data, our appetite for finding "patterns" in this data has grown exponentially faster, and we run a large risk of *false discoveries*. When one takes statistical precautions, being appropriately conservative in declaring findings, one is plagued with *false negatives (low power)*. Two basic questions that are not well understood are *for a given computational budget, what is the tradeoff between false positives and power?* and *what is the tradeoff involved in using general nonparametric hypothesis testing for simple or parametric alternatives?* These problems become critical when data is high-dimensional (the number of samples and features are comparable), especially when there are parameters to tune or a class of models to select from. I will later elaborate on free lunch theorems & computational-statistical tradeoffs for the practically important problems of nonparametric two-sample and independence testing.

On a related note, advances in numerical algorithms are *increasingly* critical for applications. Optimization is now a vital ingredient in ML and statistics, and the development of practical algorithms with theoretical convergence guarantees is a direction I have pursued with vigor, and intend to continue to pursue through my career. For maintaining brevity, the reader is directed to Ramdas and Tibshirani [2014] for trend filtering algorithms with potential to replace smoothing splines/kernels, Ramdas and Peña [2014a,b] for margin-based algorithms for classification, Ramdas *et al.* [2014] for randomized algorithms for regression, and Ramdas and Singh [2013b,a] for non-intuitive lower and upper bound connections between active learning and stochastic optimization.

Free Lunches and Comp./Stat. Tradeoffs in Two-Sample Testing

One of my current research aims is to characterize the behavior of nonparametric "two sample tests" in high dimensions, making *precise* non-asymptotic guarantees on their false positive and false negative rates, as a function of their computational cost. Given $X_1, ..., X_n \in \mathbb{R}^d$ drawn i.i.d. from a distribution P and $Y_1, ..., Y_n \in \mathbb{R}^d$ i.i.d. from Q, the two-sample or homogeneity testing problem involves testing the null hypothesis $P = Q$ against the alternative hypothesis $P \neq Q$. In the *high-dimensional setting*, the number of samples n and d can be comparable. It is easiest to motivate the problem and introduce the terminology with a simple *toy* example.

Example. To understand whether a particular brain region (say R) with 500 voxels (volume pixels) is involved in differentiating faces from non-faces, imagine placing patients inside an fMRI (functional MRI) machine, and showing 100 faces and 100 houses in some random order, while the machine records their "hemodynamic" response in region R. Given readings for faces $X_1, ..., X_{100}$ and houses $Y_1, ..., Y_{100}$, both in $\mathbb{R}^{500}$, we want to know if their underlying distributions are different.

If we choose to make no parametric assumptions like Gaussianity on P, Q, and the possible difference between P, Q could be arbitrary (in their means or variances or in higher moments), then we call this *nonparametric* two-sample testing against *general alternatives*. Neuroscientifically, suppose one expects a difference in the *mean* brain activity (if R differentiated faces from non-faces). Then we may instead choose to test $\mathbb{E}_P[X] = \mathbb{E}_Q[Y]$ against the simpler alternative $\mathbb{E}_P[X] \neq \mathbb{E}_Q[Y]$; the problem is still *nonparametric* (no assumptions on P, Q) but with *mean-difference alternatives*. Several high-dimensional nonparametric tests, based on kernels and distances, have been proposed for general alternatives over the last decade. However, there is almost no understanding of their *power* in high dimensions. After healthy

progress in [Ramdas *et al.*, 2015a,b], our most recent (unpublished) work successfully addresses some *very fundamental* open problems related to these tests, such as:

1. Can we precisely characterize the *false positive rate* & *power* in the high-dimensional setting?
2. How do tests that are designed to be consistent against general alternatives compare against "specialized" tests specifically designed for mean-difference alternatives?
3. Can one get the same power with cheaper computation, or can we prove there is a tradeoff?

We give exciting and surprising answers to all of these questions for the two most popular kernel and distance based test statistics - the Maximum Mean Discrepancy in RKHSs using the Gaussian kernel (GMMD) and the Energy Distance (ED) using the Euclidean distance, but these can be generalized to other kernels and distances. Below, I summarize some of our results, where the SNR or signal-to-noise-ratio (loosely speaking, the KL divergence) between P, Q is denoted by Ψ.

1. **Explicit characterization of power** as a function of n, d, Ψ in the high-dimensional setting as $(n, d) \to \infty$, for nonparametric P, Q differing in their means. This in itself is a big step forward - there has been *no* power analysis for kernel or distance based tests in high dimensions. Recent papers based on kernels and distances do not have any explicit rates for power or even asymptotic consistency when $(n, d) \to \infty$, and there currently exist many misconceptions that both kernel and distance based tests "work well" in high dimensional settings. These arise due to confusing low estimation error of the test statistic with high power of the test, or incorrect intuition from the normal means problem (see Ramdas *et al.* [2015a]).
2. **A clear and smooth computation-statistics tradeoff** in high dimensions, which includes linear, quadratic and sub-quadratic versions of GMMD and ED - ignoring small absolute constants, if computation scales as n^{2x} for $1 \leq x \leq 2$, then the power scales as $\Phi(n^x \Psi^2/\sqrt{d})$ when $\Psi \leqslant d/n$ (low SNR), for Gaussian CDF Φ. Specifically, linear-time and sub-quadratic block-based tests have been suggested as computationally cheaper alternatives to the full quadratic-time U-statistic, and we clearly characterize the tradeoffs involved. While there is existing analysis in the classical fixed d setting, the asymptotic power (ignoring constants) has been derived to be $\Phi(\sqrt{n})$, with computation time seemingly affecting only constants in the rate - our analysis shows that in high-dimensional settings we pay in exponents of n.
3. **ED & GMMD provably have *exactly the same power*** against mean-difference alternatives. While there has been recent work characterizing the similarity between distance and kernel based tests, there has been no formal power statement, especially about the performance of the "default" choice for kernel (Gaussian) and distance (Euclidean) in any setting.
4. **Free Lunch! ED & GMMD provably have the *same* power as specialized tests** that have been designed in the literature to test for differences in means, for example by Bai & Saranadasa (BS) or Chen & Qin (CQ). This is rather remarkable, since one does not lose *anything* for the extra generality! Indeed, it implies that unless one has any further information, one needn't *ever* use BS/CQ since GMMD/ED are strictly superior, being additionally consistent against *any* general alternatives.
5. **The power is provably independent of Gaussian kernel bandwidth**, as long as it is chosen to be $\Omega(\sqrt{d})$, which happens to be the choice made by the popular "median heuristic". This heuristic chooses the bandwidth as the median pairwise distance between all points, and has had *no* justification or power analysis in even the classical setting, and characterizing its behavior in any formal way is an important open problem.

Future work involves extending these results to the important problem of independence testing, as well as investigating the mathematical relationship between two sample testing and classification, independence testing and regression.

References

Aaditya Ramdas and Javier Peña. Margins, kernels and nonlinear smoothed perceptrons. In *Proceedings of The 31st International Conference on Machine Learning (ICML'14)*, pages 244–252, 2014.

Aaditya Ramdas and Javier Peña. Towards a deeper geometric, analytic and algorithmic understanding of margins. *preprint arXiv:1406.5311, submitted to Optimization Methods and Software*, 2014.

Aaditya Ramdas and Aarti Singh. Algorithmic connections between active learning and convex optimization. In *Algorithmic Learning Theory (ALT'13)*, pages 339–353. Springer, 2013.

Aaditya Ramdas and Aarti Singh. Optimal rates for stochastic convex optimization under tsybakov noise condition. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, pages 365–373, 2013.

Aaditya Ramdas and Ryan J. Tibshirani. Fast and flexible ADMM algorithms for trend filtering. *preprint arXiv:1406.2082, submitted to Journal of Computational and Graphical Statistics*, 2014.

Aaditya Ramdas, Deanna Needell, and Anna Ma. REGS: A randomized extended gauss-seidel algorithm for undercomplete systems. *in preparation for submission*, 2014.

Aaditya Ramdas, Sashank Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel- and distance-based nonparametric hypothesis tests in high dimensions. *Proceedings of the 29th AAAI Conf. on Aritificial Intelligence (AAAI'15)*, 2015.

Aaditya Ramdas, Sashank Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS'15)*, 2015.