Machine Learning via Statistical Discrepancies

Song Liu (song.liu@bristol.ac.uk)

Abstract

In this document, we highlight several exciting research opportunities on statistical discrepancies within the context of machine learning and data science. Note that this document is a public facing research profile thus I encourage interested students to talk to me directly for more specific research ideas and projects.

1 Background

Machine Learning is about identifying patterns from data [1]. *Statistical discrepancies* measure differences between two probability distributions and play essential roles in searching for patterns.

For example, when classifying images of cats and dogs, we want to find imaging patterns separating dogs and cats. Thus, an ideal searching criterion is looking for features that *maximizes* a particular statistical discrepancy between cats and dogs images.

When generating artificial images, we want synthesized images to have patterns similar to real ones. To do so, we can "train" a neural network by *minimizing* a statistical discrepancy between the synthesized and original images.

When detecting cyberattacks, we can identify data points that contribute the most to the discrepancy between a contaminated and a reference data set. There are many other machine learning applications whose learning criteria depend on measuring statistical discrepancies between distributions.

As a result, the study of statistical discrepancy is one of the most fundamental problems in statistics and has attracted much attention in our community for a long time. It was thought that many classic discrepancy measures (such as Kullback-Leibler divergence) are not tractable for large datasets and complicated models. However, the research on computing discrepancy has made massive progress in recent years, enabling many exciting applications and providing plentiful research opportunities.

2 Research Opportunities

I am interested in supervising research projects along the following lines.

- **Machine Learning Applications for Discrepancy Measures** Many discrepancies become computationally tractable due to approximation algorithms discovered very recently. Their potential has not been fully realized yet and new applications are being identified now and then. Approximate Bayesian Inference by minimizing f-divergence [5] and Kullback-Leibler variational gradient descent [11] are examples of research along this line. Both are applications of recently discovered computational methods [13, 4, 10]. Moreover, in some domains of statistics, the problems are often solved by non-probabilistic approaches. Statistical discrepancies usually provide a more general and principled framework to tackle these problems.

- **Approximating Discrepancy Measures from Data**: Proposing novel algorithms for useful discrepancy measures will significantly contribute to the foundation of machine learning. Many modern machine learning tasks, such as Generative Adversarial Net (GAN) [6] and Variational Inference (VI) [2], rely heavily on approximating certain statistical divergences. However, we still do not have efficient approximation algorithms for some popular discrepancies (such as Wasserstein distance) and some existing approximation algorithms suffer from stability issues [14]. Theoretical works can also provide insights and unify existing approximating algorithms.

- **New Types of Discrepancy Measure** Some machine learning applications call for specialized differences different from any known family of discrepancies. Fisher-Hyvarian Divergence [8, 12] and Kernalized Stein Discrepancy [10, 4] are great examples of this line of research. They were motivated by calculating the difference between an unnormalized density model and a dataset. Other discrepancy measures have been motivated by the need for robustness [17] and computational efficiencies recently [9, 15].

- **By-products of Approximating Discrepancy** While a discrepancy is only a single numeric measure of differences, the by-product of approximating a discrepancy can paint a much richer picture of two distributions. For example, the density ratio function in f-divergence estimation [13, 16] and witness function in Kernel Mean Discrepancy [7] are useful functionals in many applications in their own rights.

- **Identifying Limits and Challenges of Approximating Discrepancy** Discrepancy-based machine learning algorithms enjoy success in many applications, but it is also important to realize the limits of such methods and propose potential fixes. For example, it has been recognized that approximating information divergences suffers from a "density chasm" problem [14]: the density ratio function, on which the divergence is based, cannot be accurately estimated when two distributions are far away. Potential fixes have been proposed by constructing a "bridge" between two datasets [14, 3]. Research along this line will provide insights into developing more robust and versatile discrepancy measures for a much more comprehensive range of machine learning applications.

3 Supervision

Although I am interested in supervising PhD students working on above topics, I am also open to co-supervise students working on topics that are partially related to the concepts above. Students who are interested in co-supervision are advised to reach out to me early in their PhD study.

4 PhD Thesis

Students are free to work on *any* of the research opportunities above as long as their works can be coherently combined in one PhD thesis. Below are "example" thesis titles:

- Generating Realistic Images via Statistical Discrepancy Minimization (application based)
- Robust KL Divergence Estimation via Density Ratio Estimation (methodology based)
- On the Convergence Rate of f-divergence Approximation (theory based)

Students who want to study a specialized application (such as geological science) are advised to reach out early in order to find an appropriate co-supervisor.

References

[1] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[2] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[3] K. Choi, C. Meng, Y. Song, and S. Ermon. Density ratio estimation via infinitesimal classification. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2552–2573. PMLR, 28–30 Mar 2022.

[4] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 2606–2615, 2016.

[5] M. Glöckler, M. Deistler, and J. H. Macke. Variational methods for simulation-based inference. In *The Tenth International Conference on Learning Representations*, 2022.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NeurIPS 2014)*, pages 2672–2680, 2014.

[7] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[8] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[9] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[10] Q Liu, J. D. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 276–284, 2016.

[11] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems 29*, pages 2378–2386, 2016.

[12] S. Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2009.

[13] S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*, pages 271–279, 2016.

[14] B. Rhodes, K. Xu, and M. U. Gutmann. Telescoping density-ratio estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 4905–4916. Curran Associates, Inc., 2020.

[15] Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 574–584. PMLR, 22–25 Jul 2020.

[16] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

[17] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.