Research Statement

Yash Pote
Ph.D. candidate
National University of Singapore

A vast amount of data is collected daily, necessitating the need for algorithms that consume only a sublinear amount of resources in terms of the data size. My research is in the development and implementation of such algorithms. Specifically, my work has focused on distribution testing, where the input is a probability distribution over an exponentially large domain ($\{0,1\}^n$), and the goal is to test the properties of the distribution, such as identity or distance to some target. Although distribution testing has been studied extensively in the theoretical context, very little work has examined the algorithms in practice.

I have concentrated on adapting theoretical insights of distribution testing to practical scenarios in my research. This has involved not only fine-tuning the algorithms themselves for efficiency but also ensuring their compatibility and scalability in diverse, real-world environments. The core of my research lies in overcoming the barriers that prevent these theoretically efficient algorithms from being as effective in practice. The main idea behind our contribution is the use of more powerful but realistic models to access distributions.

Constrained samplers are randomized programs that take in a concise description of a distribution over a constrained set and output a sample from the set. Formally, a sampler $\mathcal{D}_{f,w}$ takes as input a function $f : \{0,1\}^n \to \{0,1\}$, a weight function $w : \{0,1\}^n \to \mathbb{R}^+$, and returns a sample from the set $\{x|f(x) = 1\}$, with probability proportional to $w(x)$. Given the widespread use of samplers in real-world applications where safety is essential, such as medicine and transportation, there is an urgent need for tests that can provide statistical guarantees on the output of these samplers. Unfortunately, even for the basic task of verifying the uniformity of a distribution over $\{0,1\}^n$ (i.e., for a constant $w(\cdot) = 1$), we require $\Omega(2^{n/2})$ samples. This raises the question: Can we prove rigorous statistical guarantees on the correctness of samplers in practice?

Since black-box sampling is clearly not enough, research has focused on richer access models, and one such model is conditioning. Conditional access allows one to specify a subset of the domain, and the sampler returns a sample from the distribution conditioned on the subset. Although conditioning has long been known to provide exponentially faster tests, its practical implementation has remained elusive. In [3, 5], we built on their approach and provided the first scalable test for samplers over *arbitrary* distributions, and our test required $\mathcal{O}(n)$ queries to the distribution. Our test has proved efficient and useful in practice and can be found on the following link (`https://github.com/meelgroup/barbarik`).

In our latest work, we tackled the more challenging problem of (additive) distance estimation, wherein the output is the total variation distance (TVD) between two input distributions. We were able to show the first polynomial query distance estimator in the conditional model in [2, 1], which requires $\mathcal{O}(n^3)$ queries to estimate the distance between two n-dimensional distributions. Moreover, we use a restricted form of conditioning known as Subcube Conditioning, which is particularly suitable for implementation.

The sampler formulation encompasses probabilistic circuits (PCs), a representation language for a large number of statistical models commonly used in AI. We were able to leverage our techniques to efficiently determine the total variation distance between large circuits [4]. To accompany the algorithmic contribution, we also presented the tool `Teq` (`https://github.com/meelgroup/teq`), which can estimate TVD between d-DNNF circuits, which form a large fragment of the PC family of circuits.

My current focus is on broadening the scope of testing and distance estimation algorithms to a larger class of objects, including large language and image models. The current state-of-the-art language and image models are built with the ability of conditional generation. In the near future, I aim to understand these conditional accesses deeply to enable $poly(n)$ algorithms on distributions over exponentially large (in n) domains.

References

[1] R. Bhattacharyya, S. Chakraborty, Y. Pote, U. Sarkar, and S. Sen. Testing self-reducible samplers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[2] G. Kumar, K. S. Meel, and Y. Pote. Tolerant testing of high-dimensional samplers with subcube conditioning, 2023.

[3] K. S. Meel𝔇, Y. Pote 𝔇, and S. Chakraborty. On testing of samplers. In *Proceedings of Advances in Neural Information Processing Systems*(*NeurIPS*), 12 2020.

[4] Y. Pote and K. S. Meel. Testing probabilistic circuits. In *Proceedings of Advances in Neural Information Processing Systems*(*NeurIPS*), 12 2021.

[5] Y. Pote and K. S. Meel. On scalable testing of samplers. *Advances in Neural Information Processing Systems*, 2022.