Research Statement

Adway Girish, EPFL

I am broadly interested in **information theory** and its applications to **machine learning**, **compression**, and **communication**. In terms of applications, I look to study *fundamental limits* of what is and is not possible in problems that are of practical interest via probabilistic models. Knowledge of these baselines is crucial to fairly evaluate current methods, obtain insights into the inherent structure of the problem and its solution, and design improved methods. At a more abstract level, I look to better understand properties of *information measures*, which often characterize operational quantities in the aforementioned applications. While the engineering motivation is certainly reason enough to pursue these problems, I am drawn to them primarily for the elegant and curious mathematical problems they reduce to. Along these lines, specifically, I have been working on the following projects (and themes), which are further described below:

(1) prompt compression for language models (compression for machine learning),

(2) distributed hypothesis testing (compression and communication for decision making), and

(3) finite-entropy channel capacity (communication with practical constraints, contraction of information measures).

Prompt compression for language models. The recent success of transformer-based large language models (LLMs, e.g. ChatGPT) is due, in part, to its attention mechanism. A drawback is that its computational complexity scales quadratically with the length of the input — long inputs have been known to slow performance, decrease efficiency, and even result in incorrect responses. A natural solution for practitioners is to try to shorten these inputs before they are passed to the LLM, while ensuring that the meaning remains unchanged. Though several works proposed ad hoc methods that work well, the problem lacked a formal description. Looking at the problem as information theorists, in collaboration with applied researchers at UT Austin, we attempted to obtain a characterization of the fundamental limits involved. The quantities of interest are the rate of compression and the distortion incurred due to the compression (defined appropriately); the fundamental limit is simply the optimal trade-off between these quantities. We observed that the evaluation of this optimal trade-off takes the form of the solution to a linear program of an impossibly large dimension. Nonetheless, the structure of the Lagrange dual is such that it allows for a simple geometric algorithm, efficiently providing an exact solution for sufficiently small toy datasets. For large-scale practical datasets, we approximated the optimal curve by combining the geometric algorithm with a beam search.

A key insight that we obtained from studying the optimal trade-offs (in line with results from the classical information theory literature) is that variable-rate compression is essential to match the optimal performance. With this intuition, we observed that by simply tweaking a state-of-the-art scheme to have an adaptive parameter, making it variable-rate, there is an immediate improvement. Thus, our results show that there is room for active research in the area to bridge the gap to optimality, and provide structural insights into the problem that help in practice. A preliminary version of this work was selected to be an oral presentation (top 4 of 58) at the ICML 2024 workshop on Theoretical Foundations of Foundation Models and the complete version will appear in the proceedings of NeurIPS 2024 [1].

Distributed hypothesis testing. Most modern communication networks are comprised of a large number of independently operating agents that observe some partial information, such as in sensor networks, cloud computing, and Internet-of-Things. In many of these setups, the goal is to make a decision at a central server using information gathered from these agents. The communication between the agents and the server is rate-constrained, i.e., the agents must appropriately compress information that is relevant to the decision task. The fundamental limits in this problem are the trade-offs between the rate of compression and the decision error probabilities. Though well-studied in the information theory literature, exact solutions are known only in very specific agent configurations and choices of decision tasks. Seemingly simple settings such as testing from which of two possible joint distributions the samples obtained at two remote agents are drawn, remain open to this day. In ongoing work, we tackle this specific problem by assuming that the samples are binary or real-valued

Gaussian random variables; the decision task is then simply to test whether they have a low or high correlation. By considering simplified data models that make sense to us intuitively, we are able to focus on and identify key features of the underlying (general) decision task, without being distracted by the generality of the choice of data. As before, our aim is twofold: to characterize the fundamental trade-offs, and to design practical schemes that are close to optimality. We have a number of candidate schemes that seem to perform well in certain data parameter regimes, but without knowing the fundamental limits, an objective evaluation of these schemes is difficult. We hope that progress in these problems provides insights that take us closer to understanding the more general open problem.

Entropy-constrained channel capacity. A classical result in communication theory is that for a power-(variance-)constrained source to communicate over a noisy channel that adds Gaussian noise, the best input (i.e. achieving 'capacity', the best possible communication rate) is one that is picked from a Gaussian distribution. In practice, there are several constraints that make this choice impossible, e.g., the source can only produce a finite amount of randomness, or be drawn from a finite-support or bounded distribution, and so on. One example is in modern communication networks, where sources may be low-power devices that simply relay rate-limited information. Though such considerations are of immense practical utility, they remain underexplored, mainly due to the difficulty in formulating problems that can be tackled analytically while still offering operational insights. We are currently studying the effect of a finite-entropy constraint, which models the discrete nature inherent to all practical systems. Our goal is to characterize the reduction in the capacity incurred by such a constraint. Though this leads to a non-convex optimization problem, we find that the asymptotic regimes of low and high power and entropy are promising for analytical progress.

Independently of its engineering motivation, this problem can also be reformulated to fundamental questions relating to information measures and randomness. The classical result above says that the choice of power-constrained distribution that best preserves information when subjected to Gaussian noise is the Gaussian distribution. Hence, the finite-entropy constraint poses the following general questions: Given a finite information budget, how much of it can be preserved under a random transformation, such as addition of Gaussian noise? How does information dissipate in the presence of noise? These can be viewed as strengthenings of the classical data processing inequality, which says that further processing of data cannot increase information, or equivalently, any processing of data may only reduce information. Such questions relate to the inherent nature of randomness and find application in a variety of scenarios, far beyond the specific communication problem mentioned.

It is clear that information-theoretic techniques and analyses have the potential to be of use in several applications, particularly in the modern data-driven technological age. Moving forward, in collaboration with and learning from applied researchers, I hope to utilize my training in information theory to contribute further in various applications. Additionally, I aim to continue working on information-theoretic problems at the intersection of probability, statistics and computer science, tackling challenging mathematical problems that have the added benefit of offering fundamental insights into practical problems.

References

*, † denote equal contribution

[1] Alliot Nagle*, **Adway Girish***, Marco Bondaschi, Michael Gastpar, Ashok Vardhan Makkuva†, and Hyeji Kim†. "Fundamental Limits of Prompt Compression: A Rate-Distortion Framework for Black-Box Language Models". In: *The Thirty-Eighth Annual Conference on Neural Information Processing Systems (to appear)*. 2024. URL: https://arxiv.org/abs/2407.15504.