RESEARCH STATEMENT

Jennifer Tang (jstang@mit.edu)

My research is broadly in the area of **information science**, primarily involving a combination of **information theory**, **applied probability**, **networks**, and **collective social phenomena**. I want to apply information-focused mathematical techniques to large data-driven problems across and beyond electrical engineering, computer science, and statistics. Currently, the scale of data needed to sustain progress is quickly outpacing the hardware and energy resources available to support it. It is therefore crucial to develop new methods for reducing the memory, computational, and communication burden of large scale data and networks while minimally affecting the performance of the algorithms using it. For instance, one objective is to find better encodings for *quantization*, allowing devices to store and communicate data much more efficiently and greatly saving on resources like memory usage.

My plan is to find mathematically rigorous solutions and to determine *fundamental limits* on learning or communication for data science problems. Finding these fundamental limits is often crucial for understanding the underlying structure of these problems and can help guide the design of algorithms or lead to better understanding of underlying workings of data systems. I have embarked on this path so far by mainly focusing on two directions: 1) **information processing of high-dimensional data**, and 2) **analysis of and estimation on social networks**.

High-dimensional Data. One format of data relevant in many applications is probability or frequency data. However, this poses a computational challenge when, as is often the case, the relevant probability space is high-dimensional (over many objects or features). Current GPT-like models convert language into a sequence of tokens, where each token is one in 50000. Theory of probability over 50000 items is very different from theory of probability over a small number of items. My research accounts for this by using core techniques in information theory to find methods that scale efficiently with dimensionality to make problems such as the **encoding**, **transmission**, and **estimation** of large-scale data tractable. In my PhD thesis, I studied coverings of probability spaces under Kullback-Leibler (KL) divergence (a fundamental and natural way to compare probability distributions), in particular looking at how KL divergence coverings scale with the dimension of the probability space. This scaling is crucial, as KL divergence coverings can be used to help determine fundamental limits in a variety of problems in data science involving high-dimensional probability distributions, which I demonstrated in the following (three) projects:

- Developing a practical way of efficiently storing high-dimensional probability vectors [1].
- Better understanding how well we can estimate the next symbol from a sequence using only the pattern of previous data (in terms of minimax regret) [2].
- Finding the fundamental limits of sending information through a noisy channel where message symbols are randomly shuffled [3] (**Best Student Paper Award at ISIT 2022**).

Social Networks. My second area focuses on opinion dynamics under the broader context of understanding individual and group behavior in social networks under computational and information constraints. While there have existed models for social behavior in economic theory, recent evidence has brought into question whether the assumptions of these models are realistic. For instance, people commonly misrepresent their true beliefs in order to fit in socially, thus presenting only limited and distorted information about their true beliefs to their peers. This has implications to help better understand human behavior which can be beneficial for applications like political campaigns and marketing. Specifically, I study agent-based network models by:

- Modeling social phenomena and analyzing the model for convergence properties [4].
- Estimating unknown properties of agents from their outward behavior (inverse problems) [5].

Projects

Companders for Storing Probability Vectors [1]

In many applications, the data being studied comes in the form of frequency counts or probabilities.[1] It is often desirable to compress, or quantize, this data to make it smaller, while minimizing distortion. KL divergence is a common and natural measure of distortion (especially in estimation and machine learning), but it has not been examined in this context.

For this project, we employed *companders*, a classical technique originally used to compress radio signals, and derived optimal quantization schemes for any specific prior distribution on the data, as well as a *minimax* quantizer, which performs well over all possible prior distributions. A key advantage of companders is their simplicity and ease of implementation. Thus, our work yields a quantization technique that greatly reduces distortion while being easy to implement.

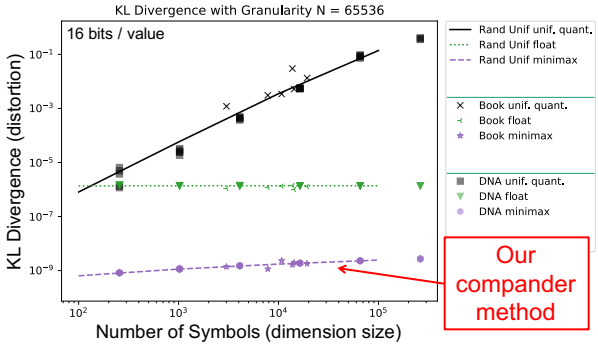

Figure 1: Results for storing probability distribution entries with our minimax compander technique, which achieves the lowest KL divergence distortion when compared to uniform quantization and floating points representations.

When tested on length 6 DNA data, compared to uniform quantization on 16-bits, our minimax compander gives smaller distortion as measured by KL divergence when using only 4-bits, a reduction of at least 75% in memory. Compared to 16-bits floating point (bfloat16), our minimax compander gives smaller KL divergence distortion using 12-bits, a reduction of at least 25%. This work was presented at ISIT 2022 and published in the IEEE Journal on Selected Areas in Information Theory issue on Modern Compression [1]. A precursor work, studying the information theoretic concept of rate distortion for a related loss function, was presented at ISIT 2021 [6]; these theoretical results then led to the development of the compander algorithm.

Minimax Regret for Learning Patterns [2]

Prediction with logarithmic loss is a classic problem in online learning which asks how well you can predict successive symbols in a sequence given past symbols. This is commonly done with the goal of minimizing *regret*, which compares the algorithm's performance to an oracle with hindsight knowledge of the arbitrarily generated string. When the data is high-dimensional (for instance, the dimension is around $300,000$ when modeling English words) it may not be possible to get vanishing per symbol regret, implying that we are unable to learn effectively. Thus, a variation is to try to predict *patterns*, which represent strings but are invariant under relabeling of the symbols. For certain classes of distributions, patterns essentially capture all of the entropy of the original distribution, meaning that the bulk of its information content lies in its pattern. Thus, studying online learning of patterns can bring immense insight to general online learning problems. Previous bounds such as those in [7] give the order on regret for patterns with a logarithmic factor gap. In my work, I improve the logarithmic term of the upper bound, giving the tightest currently-known bounds. This was also achieved with KL divergence covering bounds using a different approach that combines them with classical estimation techniques. The improved bounds help us better understand what can be achieved in online learning with high-dimensional data. This work was presented at COLT 2022 [2].

[1] For instance, DNA data can be stored as counts on k-mers, which are length k substrings. However, this data is a probability distribution over a set that has size exponential in k.

Capacity of Noisy Permutation Channels [3]

The noisy permutation channel is an abstract model of a communication channel where input symbols are not only subject to noise, but are randomly shuffled, removing information encoded in the order of the message symbols. This models practical problems such as out-of-order network routing and biological storage systems.[2]

Previously, bounds on the fundamental limits of the noisy permutation channel were given in [8]. However, there was a gap between the upper and lower bounds, which for some instances could be very large, leaving the exact characterization an open problem. My work resolves this open problem and determines exactly the correct capacity.

The fundamental difficulty of this problem is that no known central limit theorem scales as needed with large dimension. However, I overcame this by using KL divergence covering results from my thesis. Specifically, I determined that the KL divergence between noisy observations drawn from a distribution with replacement and noisy observations drawn from a distribution without replacement is, surprisingly, a constant. This is a difficult bound to obtain when the dimension is high and it can have consequences for other problems. The work was presented at ISIT, the flagship conference of information theory, in 2022 and received a **Best Student Paper Award** [3].

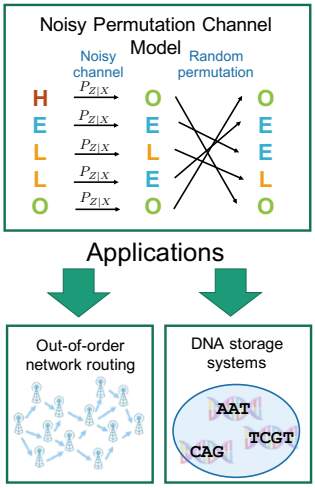

Figure 2: The noisy permutation channel and some applications; symbols undergo a noisy channel followed by a random permutation.

Opinion Dynamics with Social Pressure [4, 5]

Opinion dynamics studies the interaction of agents in order to better understand certain social behavior. To explore how social networks evolve when their agents are affected by social pressure, I studied an interacting Pólya urn model (introduced in [9]) for opinion dynamics. Understanding the dynamics of this abstract model can help give insights on people's behavior in social situations involving strong differences of opinion such as political campaigns, marketing for commercial products, and behavior in online social networks. Two directions of this work are:

- We show precisely under what circumstances all agents in the network eventually converge to outwardly agreeing on the same opinion (a situation we call *consensus*), even if there are many agents who inwardly do not agree with that opinion [4]. We also show that on random networks, when agents are more connected with those with similar beliefs, consensus is less likely to happen. Key techniques used in this work are stochastic approximation and Lyapunov theory.
- We show that the agents' true beliefs can be estimated from their stated opinions using an estimator based on maximum likelihood, and find bounds on the rate of convergence of this estimator. The challenge of this estimation problem is that consensus causes the agents' stated opinions to become uninformative in the limit. Key techniques used in this work are martingale concentration bounds and Lyapunov theory [5].

Other Projects

Signal Separation. This project focuses on data-driven signal separation for removing unwanted interference from over-the-air signals. Techniques used involve signal processing and machine learning.

[2] In biological storage systems, a developing technology for storing data on snippets of DNA, strings of nucleotides are stored in a pool without regard to order.

This work was presented at Globecom 2022 [10] and won a **Best Student Paper Award** at MLSP 2022 [11].

Defect Tolerance. This project studied an information-theoretic model of how to add redundancy to computer hardware for error correction when physical elements fail. I characterized the trade-off between redundancy and wiring complexity for finite sized models and various asymptotic settings [12]. This work was the **Shannon Centennial Celebration Student Competition Winner**.

Future Projects

My vision is to investigate the **role of information** in emerging applications. I want to revolutionize how classical ideas in fields like information theory can be applied to problems where theoretical aspects of information were not previously considered in depth. Particular avenues I want to explore include **efficient storage for machine learning** and **information aggregation and spread in social networks**, where there are ample directions to find novel uses of classical information theory. A key feature of both areas is the presence of huge amounts of data and high dimensionality. Practical usage of information in these domains have quickly outpaced theoretical understanding.

Some fundamental questions in efficient storage for machine learning that I want to study are:

1. What can possibly be learned from the available data?
2. What information is redundant and can be removed?

Once there is a better understanding of what information is necessary for the task at hand, applications used on platforms like mobile devices or data centers that store or transmit data needed for machine learning can be optimized to save resource usage, such as memory or battery life.

For social networks, there are similar questions:

1. What can agents learn from data communicated between peers?
2. How does the available information affect the behavior of agents?

Such questions can help us understand the social phenomena we observe and how to leverage information for change, which are likely important for applications like public health campaigns.

My prior research equips me well to explore these topics. In particular, my Ph.D. thesis focused on studying the structure of information with respect to KL divergence, and gives me useful tools and insights which I have already applied to several problems. Additionally, my work on social networks, quantization, and learning has given me some experience in these domains, which I hope to expand as I pursue a deeper understanding of the role of information. *Specific project ideas are included below.*

Quantization for efficient machine learning. One exciting direction to expand my current companders project is to apply it to neural networks and other learning models in order to improve quantization. This approach in particular seems well-suited for analyzing large language models, which inherently involve high-dimensional probability distributions (over tokens), though other models also offer interesting and novel possibilities.

- **The problem**: Machine learning uses large language models with hundreds of billions of parameters[3], which overburdens the resources these models run on, particularly for storing data in memory. Current solutions to solve this problem are done mostly heuristically, with relatively little rigorous theoretical analysis or guarantees.
- **The goal**: I want to use theoretical foundations inspired by information theory and probability theory to determine the optimal trade-off between memory usage and accuracy in learning models.

[3]For instance, GPT-3 contains 175 billion parameters requiring 350 GB of memory in 16-bit floating points

- **Possible approach**: Techniques from my previous work on compander transformations for storing probability distributions could be applied to weights of large language models to find the set of provably optimal companders. This could reduce the quantization loss for learned models, via a very different approach compared to current methods.

Social learning with limited communication. Social learning is an active area which tries to model how agents combine their knowledge to learn collectively. This area seeks to understand how humans, who have computational limitations, might reason in a group setting. One direction is to understand what happens if there are limits on what information is transferred between neighbors.

- **The problem**: Agents connected on a network each collect private observations which can be used to infer a true parameter representing the world. However, their private observations are insufficient for any one agent to deduce the correct parameter. They must communicate with their neighbors to collect enough information in order to infer the true parameter.
- **The goal**: Understanding how limitations on peer-to-peer communication, such as limited bandwidth, affect the ability of the agents to learn the true parameter.
- **Possible approach**: First, we can attempt to use Lyapunov theory, similar to [4], to understand the dynamics. Second, nonlinear quantization techniques, like companders, can be explored to understand how quantized communication affects the network.

Optimal intervention in social networks. While [4] and [5] studied the dynamics of agents exchanging opinions under social pressure, as well as the problem of estimating true beliefs from observed behavior, a logical next step is to consider optimal interventions in such networks: using a limited 'budget' for influencing a small number of agents, what is the optimal strategy to shift the general opinion of the whole network in a desired direction?

Estimation with KL divergence covering. Many estimation problems (such as distribution or property estimation from sample data) still have gaps between the best known theoretical upper and lower bounds of their optimal estimation loss; by closing or narrowing such gaps, we can sharpen our understanding of these problems and potentially develop better estimation techniques. Since these problems center around probability distributions, it is very common for them to feature KL divergence or logarithmic loss functions, which connects them on a fundamental level to my thesis topic of KL divergence coverings. These gaps might therefore be narrowed or closed by applying or extending results similar to [3] and [2], including for more complicated distributions like those with a Markov chain structure.

References

[1] Aviv Adler, **Jennifer Tang**, and Yury Polyanskiy, "Efficient representation of large-alphabet probability distributions," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 4, pp. 651–663, 2022.

[2] **Jennifer Tang**, "Minimax regret on patterns using kullback-leibler divergence covering," in *Annual Conference on Learning Theory (COLT)*, 2022, vol. 178, pp. 1–18.

[3] **Jennifer Tang** and Yury Polyanskiy, "Capacity of noisy permutation channels," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 1987–1992.

[4] **Jennifer Tang**, Aviv Adler, Amir Ajorlou, and Ali Jadbabaie, "Convergence of opinion dynamics under social pressure for general networks," in *To appear at IEEE Conference on Decision and Control (CDC) 2023*.

[5] **Jennifer Tang**, Aviv Adler, Amir Ajorlou, and Ali Jadbabaie, "Estimating true beliefs from declared opinions," 2023.

[6] Aviv Adler, **Jennifer Tang**, and Yury Polyanskiy, "Quantization of random distributions under kl divergence," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 2762–2767.

[7] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh, "Tight bounds for universal compression of large alphabets," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 2875–2879.

[8] Anuran Makur, "Coding theorems for noisy permutation channels," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 672–6748, Nov 2020.

[9] Ali Jadbabaie, Anuran Makur, Elchanan Mossel, and Rabih Salhab, "Inference in opinion dynamics under social pressure," *IEEE Transactions on Automatic Control*, vol. 68, no. 6, pp. 3377–3392, 2023.

[10] Alejandro Lancho, Amir Weiss, Gary C.F. Lee, **Jennifer Tang**, Yuheng Bu, Yury Polyanskiy, and Gregory W. Wornell, "Data-driven blind synchronization and interference rejection for digital communication signals," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 2296–2302.

[11] Gary C.F. Lee, Amir Weiss, Alejandro Lancho, **Jennifer Tang**, Yuheng Bu, Yury Polyanskiy, and Gregory W. Wornell, "Exploiting temporal structures of cyclostationary signals for data-driven single-channel source separation," in *2022 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2022.

[12] **Jennifer Tang**, Da Wang, Yury Polyanskiy, and Gregory W. Wornell, "Defect tolerance: Fundamental limits and examples," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5240–5260, 2018.