Research Statement

An Information-Theoretic Approach to Data Science

Today's data deluge is driving a transformation of the world-economy and even the modern way of life. Cutting edge data science challenges are pushing the boundaries of what can be communicated, stored, and learned. While data science is a rapidly developing field with enormous industrial impact, academia is just now formalizing and developing data science as a subject of study and research activity. It is a privilege to have been part of the departments of electrical engineering and statistics at Stanford, where several major advances underlying the data revolution first saw the light of day. My recent research at Stanford encompasses both statistical approaches for discovering subtle patterns within large datasets, with information-theoretic tools explaining these patterns. More generally, I see myself developing foundations of data science by considering explicit restrictions on data acquisition, representation, and inference.

Consider the following diagram of a generic data processing system:

![Diagram of a generic data processing system](image.png)

Figure 1: Stages in extracting information from raw data that is needed for a decision or action

The components in Fig. 1 nicely organize some fundamental processes in data science where resource limitations intrude. Typical limitations include:

- Limited sampling rates and other resources in the acquisition stage.
- Limited bits for the compressed representation.
- Limited processing power for the inference procedure.

For example, raw data collected by sensors of an autonomous vehicle is disproportionately large compared to the time available to process it using current artificial intelligence techniques before the vehicle must reach a steering decision. As a result, the performance of such a system is dictated by sampling and data compression constraints in the acquisition stage, rather than the amount of data available to the sensors. In another example, classifying large text corpora using an end-to-end learning approach may require amounts of training data and computation resources that far exceed system specifications. Consequently, deriving a compressed representation of the text is a key enabler in large-scale text classification systems.

Both examples face the same principal challenge: transforming raw data most efficiently into a compressed representation that satisfies system's restrictions in terms of storage, bandwidth, or processing power. My expertise in statistics and data science combined with information theory and data compression provides me with a dual and powerful attitude to this challenge: data-driven approaches for modeling, with an information-theoretic approach for analyzing fundamental limits of information retrieval under these models.

My doctorate research was focused on the effects of bit-level compression on techniques in signal processing and machine learning, i.e., the case in which the compressed representation in Fig. 1 is subject to an explicit memory restriction. Specifically, I considered the way raw, high dimensional, and potentially analog signals are sampled or transformed before they are compressed into their digital representation. The ability to consider data in its analog form distinguishes my work from other compression and dimensionality reduction techniques that operate only after the analog-to-digital converter stage, namely, the data in its digital form. My work was able to characterize the minimal distortion that can be attained in this procedure as a function of the sampling rate and the bitrate of the digital representation, as well as the optimal sampling and compression schemes attaining it. These results provide a unified treatment of signal acquisition under the following three...
information inhibiting processes: sampling, compression, and additive noise. While the effect of each of these processes alone had been well-understood before, my work shows that the combination of them undermines standard conceptions. For example, I have shown that sampling at a rate smaller than Nyquist is optimal for most signals (even non-sparse ones) when the bitrate of the digital representation is restricted.

My postdoctoral research addresses challenges in processing data past the acquisition stage by exploiting structures of the data for efficient inference with limited computing power compared to available data. The goal of this research is to characterize the limits of discriminating between discrete distributions over large alphabets associated with two different frequency tables (counts of items from these distributions). For this problem, I conjecture that the Higher-Criticism (HC) test is optimal in most problem regimes. Since this problem was studied extensively in statistics, computer science, and information theory, I expect that proving it to be correct would have a significant impact in these fields. A parallel line of this work explores applications of the HC test of two discrete distribution in real-world data. Based on this test, I developed a procedure for authorship attribution and other text classification challenges that is extremely fast and scalable; it achieves results comparable to other techniques but without handcrafting or tuning. I used a vast number of cluster computations to evaluate the performance of this test and explain its success.

While no theory of data science exists at the moment, the diagram of Fig. 1 has proven to provide a fruitful viewpoint across many data science challenges. My future research goals are directed toward developing a unified framework that incorporates restrictions arising in this diagram. Such a framework is necessary to incorporate data acquisition and processing constraints into modern techniques in machine learning, signal processing, and statistics.

Shannon’s abstraction of communication systems enabled the unprecedented success of information theory – from the definition of the bit to smart-phones and cloud services in less than 70 years. The diagram of Fig. 1 can play an analogous role in the development of data science; it would transform this field into a successful academic discipline and inspire technological development. Indeed, the ongoing academic effort to formalize data science is driven by many of the unsolved problems on the interactions of the components in Fig. 1. I have identified various important and timely challenges in data science that are reflections of this figure and that I plan to pursue; see the section on future research below.

The sections below describe in more detail my past work and future research goals.

Past Research

In this section, I describe the main contributions of my doctorate and postdoctorate research.

Higher Criticism for Discriminating Word Frequency Tables and Testing Authorship

Suppose we have several corpora of documents, each of homogeneous and known authorship. We encounter a new document of unknown authorship we wish to determine its author. The abundance of textual data in our age has spurred many studies of authorship attribution problems of this form. Existing approaches to this problem rely on handcrafted features and test statistics constructed for each specific problem; it is unclear whether these features or tuned parameters can be reused in other problem domains. My recent work [Kip19] proposes a technique to authorship attribution that can be used “out-of-the-box”. It relies on a relatively simple statistical tool: The Higher Criticism (HC) statistic as a measure of closeness between word-frequency tables. Based on this closeness, we determine the most likely author according to a standard statistical test.

I developed software packages providing efficient implementations of this technique in different settings, and conducted a massive number of cluster computations to evaluate its performance in authorship challenges. This technique performs about as well as state-of-the-art techniques but without handcrafting or tuning. In trying to understand the reasons for the success of our method, we discovered that HC is mostly affected by words characteristic of the author and is relatively unaffected by the topic structure of the text.
Sub-Nyquist Sampling is Optimal for Analog-to-Digital Conversion

Sampling is a fundamental data processing technique. It enables the effective digital manipulation of continuous-time information sources of the physical world. In the big-data era, sampling is also viewed as an effective dimensionality reduction technique widely used in machine learning. The effort of my doctorate was focused on characterizing the minimal distortion that can be attained in recovering signals from a digital representation of their samples. This work appears in [KGEW16, KEG18b, KGE19, MKG18] and summarized in the magazine article [KEG18a] and as a chapter in the book [RE20]. It provides a full characterization of the minimal distortion that can be attained in recovering the original signal as a function of the sampling rate and the bitrate of the digital representation. Note that without the sampling constraint, this minimal distortion is described by Shannon’s distortion-rate function (DRF) of the original signal. As illustrated in Fig. 2 below, an important corollary from my characterization is the fact that, for most signal models, this DRF is attained by sampling at a rate smaller than the Nyquist rate. That is, Nyquist rate sampling is necessary to attain zero distortion without any constraint on the bitrate of the digital representation. However, when such a constraint is imposed, the minimal distortion can be attained by sampling at a new critical sampling rate which is typically lower than the Nyquist rate. This new critical sampling rate depends on the bitrate of the digital representation and can be precisely derived in many important cases. Furthermore, the theory developed in this work applies regardless if the signal is bandlimited or not. In this sense, my doctorate work provided a comprehensive signal acquisition theory that includes classical sampling theory in signal processing and lossy compression in information theory as special cases.

Estimation from Compressed Data

Due to the disproportionate size of modern datasets compared to available computing and communication resources, many inference techniques are applied to a compressed representation of the data rather than the data itself. In these situations, the overall performance is determined both by the quality of the data (sample size and noise intensity) as well as the number of bits available to the estimator. A major theme in my research is the study of classical inference an learning techniques under bit-level compression constraints on the data; my works [KRG15, KRG16, RKSG19, KRG19, KD17, KREG17, KRE18, KR19a, KD19, KR19b] provided a comprehensive understanding of this research area.

One line of these works considered a compressed-sensing (CS) framework where the data represents random linear projections of a high-dimensional sparse signal. Empirical results and successful applications in MRI and ultrasound have shown that CS can perform well if the signal is very sparse, the sampling noise is very low, and the number of bits available for representing the measurements is sufficiently large. However, despite a myriad of theoretical works in this field, a precise characterization of the fundamental trade-offs between these quantities has remained elusive. My works [KREG17, KRE18, KR19a] leveraged ideas from statistical physics and random matrix theory to provide such characterization. Specifically, we show that the aforementioned trade-off can be analyzed precisely by reducing the original high-dimensional-problem into an equivalent scalar measurement model which can be analyzed precisely.

Another major contribution is described in my recent work [KR19b]. This work provides a strong and relatively simple characterization of the distribution of the lossy compression error resulting from

![Figure 2: Sub-Nyquist sampling is optimal under bit-level compression](figure.png)
spherical coding. Specifically, we show that this error can be approximated by Gaussian noise whose variance is a simple function of the coding bitrate. This approximation implies that the mean-square error of an estimator applied to the compressed representation of the data is asymptotically equivalent to the mean-square error of the same estimator applied to a Gaussian-noise corrupted version of the data. The benefit from such connection is twofold: (1) inference techniques from Gaussian-noise corrupted observations can now be applied directly on the compressed representation; and (2) it provides a mechanism to characterize the performance of inference using such techniques.

Finally, my work [KD19] addresses the problem of learning the mean of a symmetric distribution subject to the constraint that each sample from this distribution is quantized into a single bit. Such a question has fundamental implications in the fields of privacy, analog-to-digital conversion, and optimization, to name a few. This work shows that any estimator in this setting is at most as efficient as the sample median as an estimator for the mean. Namely, the increase in the amount of data required to attain a prescribed error compared to an estimator that is not constrained by quantization is equivalent to the same increase in using the median instead of the sample mean (e.g., the sample size increases times $\pi/2$ for the normal distribution). This fundamental result demonstrates that even under severe bit restrictions, a clever quantizer design can lead to estimation performance that is close to the unrestricted estimator.

Estimation and Communication with Restricted Information

I now briefly discuss works falling outside the above categorization, but still within the domains of estimation and communication.

My work [RKJ+16] provided convergence guarantees for estimating the covariance matrix of vector autoregressive processes, under the assumption that observation of the process is only partially given. In [CKG15], I have shown that using multiple receiving antennas, reliable communication is possible even if the bit-resolution of an analog-to-digital converter at each antenna is very coarse compared to the received signal strength. Finally, in my Master’s work [Kip12], I proposed an extension of the Itô stochastic integral to a class of processes with possibly dependent differences. The Itô integral is widely used in financial mathematics and dynamical systems, and our extension allows the formulation of similar dynamical models driven by random perturbations correlated across long time lags. In continuation of this work, my Master’s adviser and I developed a technique for solving classical prediction problems concerning this class of processes [AK15].

Future Research

My past research has already shown that the diagram of Fig. 1 provides a fruitful abstraction for the fundamentals of several data science challenges. Indeed, a myriad of challenges in the field can be formalized as problems concerning resource limitations and interactions of the components of this diagram. Below are three exciting research problems that are reflections of this diagram and that I plan to pursue.

Feature Selection for Transfer Learning

Convolutional neural networks (CNN) have been successful in a host of important learning and estimation applications. Due to their architecture, they can be seen as automatic, data-driven, machines for extracting features from data. When data are scarce, however, the end-to-end training approach for classification and regression used by CNNs usually leads to poor results. Therefore, a by-now standard practice is to use the last layer of a CNN trained using training data for one problem as a set of features for the second problem, for which training data is scarce. The goal of this research project is to develop optimal procedures for such “transfer learning” scenario and understand its limitations. This project combines data science methodologies for uncovering patterns in the distribution of the features with insights from feature selection theories in statistics. The starting point for this project is the techniques and analysis I developed in my postdoctorate work on selecting features when the number of useful ones is small.
Data Compression over Function Classes

The fundamental limit of bit-level compression for stochastic processes is well-understood due to Shannon’s information theory. Much less is known about the fundamental limits of bit-level compression of functions belongs to some functional classes, a classical problem introduced by Kolmogorov. Recent works have shown an intriguing connection between the two problems provided by statistical decision theory. The goal of this research project is to exploit this connection to derive denoising procedures for function classes using insights from information theory. Characterizing fundamental limits of data compression over function classes would provide an important link between classical data compression theory in information theory and learning theory in machine learning, where such classes are elementary objects of study.

Approximate message passing (AMP) in deep architectures

AMP techniques can simplify a complicated inference problem into a sequence of computationally tractable “local” problems that incorporate prior information on the data. Furthermore, the per-iteration behavior in AMP can be rigorously characterized by a simple state-evolution procedure that leads, under some conditions, to Bayes optimal solutions. Such techniques have been applied successfully in analyzing inference in various high-dimensional settings, including compressed sensing, robust estimation based, and multi-class classification using logistic regression. In my recent work [KR19b], I have used AMP to characterize estimation error in high-dimensional linear models when the data undergo lossy data compression. The goal of this research project is to develop and analyze AMP techniques for inference in multi-layer (deep) settings. By utilizing the state-evolution property of AMP, it would be possible to study the effect of the number of layers in an artificial neural network on the quality of inference. Furthermore, such an analysis may provide an overarching explanation for the empirical success of techniques based on artificial neural networks and suggest ways to overcome their shortcomings.

Summary

It’s incredible to be alive in a time where data science is taking over the world, making data-driven scientific discovery possible with the right skills and insights. In this realm of data, powerful tools for the acquisition of data from the physical world and for rapidly extracting information from it are a necessity. While no theory of data science exists at the moment, the diagram of Fig. 1 exposes the principal bottlenecks in data science and has proven to be a fruitful viewpoint across many challenges.

My past research work studied the fundamental limitations of the components in this diagram. The rise of data science makes these insights valuable assets in addressing the most pressing issues in the field.

References


