RESEARCH STATEMENT

Uthaipon (Tao) Tantipongpipat (tao@gatech.edu)
Algorithms, Combinatorics, and Optimization (ACO) PhD, Georgia Institute of Technology
Advisor: Mohit Singh
September 5, 2019

My research lies broadly in applying algorithmic and optimization perspective to problems risen in machine learning and statistics. I propose and study theoretical questions well-motivated from practice, and bridge practice and theory by providing theoretical underpinning of commonly used algorithms. This includes finding challenges arise in practice and framing them into practically accurate yet mathematically precise questions; designing efficient algorithms to solve those questions; and rigorously proving theoretical properties (approximation ratio, running time, properties of the output distribution, etc.) of algorithms. My work can be broadly put into three topics: optimal design (in design of experiments in statistics), differential privacy, and fair principle component analysis (fair PCA). They fit in broader themes of big data subsampling or compression (for efficiency or diversity), privacy, and fairness in machine learning, respectively.

Optimal design  
My colleagues and I use combinatorial and continuous optimization tools to solve a classical design problem. Though classical in statistics, the problem found many applications in machine learning and connections to rich mathematical structure in theoretical computer science. The optimization and approximation algorithm approach provides a new perspective of the problem not explored until recently, leading to the best approximation guarantee known.

Differential privacy  
I propose and study private algorithms under the privacy models that are relevant to real-world application. My coauthors and I developed new algorithms that apply to new settings including when database is growing or when analysts want to access sensitive data with no restriction on the set of allowable queries. Both settings are primarily motivated from practice.

Fair PCA  
My coauthors and I propose a new notion of fairness to address unfair representation of minority groups after dimensional reduction. Well-motivated from the observation of biases seen in practice, the problem leads to a beautiful connection between geometry of a polyhedron and a semi-definite cone in optimization and the dimensional reduction method PCA in machine learning. Our algorithm to solve fair PCA is scalable and useful to a wide range of applications.

1 Optimal Design

Linear regression is arguably the most fundamental concept in supervised machine learning. In many settings, obtaining labels is costly in time and resources, but a researcher may choose from the pool of datapoints points to obtain labels from, as known as a active learning setting. The goal of optimal design is to choose the best smaller set of datapoints to obtain labels to maximize the accuracy and confidence the model an algorithm will learn.

Optimal design is a classical problem in statistics [Fed72, Puk06], and arises in many settings such as feature selection [BAM13], sensor placement [JH09], matrix sparsification [BSS12, SS11], column subset selection in numerical linear algebra [ABL12], efficient design of science experiments and CPU processors [WYS17], and material design [WUSK18].

Intuitively, the designer seeks to find a small set of datapoints that spreads over a wide region of space in order to maximize learning over the entire space. For example, in a sensor placement application, one ought not to put all sensors in a small region and hope to accurately predict measurements in other regions. Optimal design, then, gives rise to a notion of diversity sampling, where one seeks to maximize the diversity of a smaller set from a given pool of items. Diversity sampling has many connections with machine learning, such as determinantal point processes (DPPs) [KLW12] and fairness in machine learning [CDKV16].
**Current Work** Approaches to optimal design in statistics have no strong theoretical guarantees. Existing common approaches studied in theory and used in practice include local search heuristics, such as Federov exchange \([\text{F}\text{e}\text{d}\text{e}\text{r}\text{o}\text{v}+55]\), and approximate design which solves the continuous relaxation of the problem and uses heuristics rounding. Only until recently, a new perspective to optimal design problem through a more sophisticated randomized rounding algorithm gave a reasonable approximation ratio guarantee within a polynomial running time. My work is at the forefront of this rounding: to obtain new efficient rounding algorithms which leads to the strongest approximation ratio possible.

My recent result with Mohit Singh and Aleksandar Nikolov \([4]\) obtains the best approximation ratio known for two very commonly used criteria \((A\text{ and } D)\) in statistics, which is asymptotically optimal among any algorithms utilizing approximate design as the numbers of datapoints and dimension increase. Our work did not improve approximation guarantee on another popular criteria \((E)\), but we show a hardness result which implies that it is indeed impossible for any similar algorithm to do so.

Another recent result \([3]\) is the analysis of combinatorial approaches that are used in practice, such as a local search algorithm known as Federov exchange and implemented in SAS software. As mentioned earlier, though these algorithms are proposed and widely used for more than 50 years, no strong theoretical guarantee for these types of algorithms was proven. My coauthors Vivek Madan, Mohit Singh, Weijun Xie and I provided the theoretical underpinning of the performance by these algorithms. Moreover, through our analysis we discovered (not probable but existing) cases where these algorithms have bad performances, and provided a modification of the algorithm that has theoretical guarantee for all problem instances.

**Future Work** In many applications, the design of experiments is constrained not only by the number of experiments to perform, but also the number of experiments in each type of experiments. For example, a researcher may be allowed 10 sensors of the first type and 15 of the second type, for 25 in total. This is also known as partitioning constraint. Another type of constraint is when different experiments are present with different costs, as known as a knapsack constraint. More generally, we conjecture that our existing result \([4]\) can be obtained even for a general combinatorial structure that captures all these types of constraints, namely matroid. In preliminary research, we have found close connections of optimal design problem under general constraints to DPPs \([\text{N}\text{S}\text{1}\text{6}]\), theory of stable polynomial \([\text{Gur}\text{0}\text{8}]\), expander graphs, graph sparsification, the Alon-Boppana bound \([\text{Alo}\text{8}\text{6}, \text{Nil}\text{9}\text{1}]\) in spectral graph theory, and restricted invertibility principle (RIP) \([\text{B}\text{T}\text{8}\text{7}]\). The goal will be to gain deeper understanding that allows us to relate these seemingly disparate topics. The project is joint with Mohit Singh, Aleksandar Nikolov, and Vivek Madan.

## 2 Differential Privacy

Many learning algorithms today deal with personal, sensitive data, including patience health records, GPS locations, and browsing history. First defined by \([\text{D}\text{M}\text{N}\text{S}\text{0}\text{6}]\), differential privacy (DP) gives a mathematically rigorous worst-case bound on the maximum amount of information that can be learned about any individual’s data from the output of an algorithm. Differentially private algorithms that provide accuracy guarantees have been designed for a wide variety of machine learning problems (see \([\text{I}\text{L}\text{E}\text{1}\text{4}]\) for a survey). Differentially private algorithms have been implemented in practice by major organizations such as Apple, Google, Microsoft, Uber, and the United Status Census Bureau.

However, the models on which vast majority of differentially private algorithms are designed are not applicable to the real world. First, those private algorithms are designed for static databases, yet generally the data in practice are constantly changing and growing. Second, the models assume a trusted curator who has an access to all sensitive information and allows analysts to learn the data through a predefined restricted set of queries. In reality, practitioners usually choose analysis tasks to perform dynamically as they learn about the data. These are some of the examples where developed theory of differential privacy does not adequately address challenges faced in practice. My research in this area is to close the gap between theory and practice of DP by proposing and solving theoretical questions that are relevant in practice, and implement practical DP algorithms that is supported by theoretical guarantees.

**Current Work** Our first work \([2]\), with Rachel Cummings, Sara Krehbiel, and Kevin Lai, to appear in NIPS 2018, addresses the challenge of growing databases. We give the first private algorithm that can
handle arbitrary growth in database and answer exponentially large set of queries continuously as database size increases. The algorithm extends the state-of-the-art private multiplicative weight algorithm \cite{HR10} to the growing database setting with at most a constant loss in privacy leak, the best theoretical guarantee possible.

Our second line of work \cite{1} tackles the challenge of requiring a predefined restricted set of queries by providing a privately generated synthetic data that statistically resembles the original. The algorithm trains generative adversarial network (GAN) \cite{GPAM14} privately and output nothing from the original sensitive data but the generator of the trained GAN. An analyst may use the generator to synthesize as many data points as desired and perform any analysis tasks on them without losing any privacy. The use of GANs to privately generate synthetic data has been proposed \cite{ACG16}, followed by several optimization works \cite{ZJW18, BJWWG17}. Our work proposes further optimization improvement, combining recent advances in GANs and DP into one DP GAN framework. Our framework is scalable, applicable to wide range of data types, and adaptive to change or growth in database. This work won the first prize award and people’s choice award in the engineering privacy challenge by National Institute of Standards and Technology (NIST), hosted on Herox.com \cite{PSC18}.

Future Work  Because of the wide applicability of our DP GAN framework, our team plan to implement our proposed DP GAN and test its performance on many kinds of real world datasets, including the US census, geographical data, social networks, and patience health records.

I plan to obtain theoretical privacy guarantees to new DP models that are relevant to real world situations. The first research project is to develop an efficient DP algorithm that detects whenever the distribution of the new incoming data has changed. A work \cite{KCZ18} in this similar direction assumes that we know the current and the changed distribution of the data, but in practice, even if one may expect a change in distribution, one may not know to what the distribution will change. We propose a new task of privately detecting when distribution changes without knowing the new distribution, given that the new distribution differs statistically enough to be worth recognizing. The analysts then can estimate the new distribution using the portion of data after the detected change point. This project is with professors Rachel Cummings and Sara Krehbiel.

The other project is to improve privacy guarantees that are applicable to commonly used private machine learning algorithms, such as gradient-descent type algorithms. There have been several improvement of privacy bounds beyond standard composition theorems \cite{Dr14} through utilizing other mathematical notions of privacy \cite{BS16, ACC16, MTT17} and algorithmic techniques that provably increase privacy, as known as privacy amplifications. All privacy amplifications known so far are obtained by subsampling the data (see, for example, \cite{KLW11, MPS14, ACC16, WBKT18, BDRS18}), until recently \cite{MMTT18} provides another type of privacy amplification. As the state-of-the-art, it is possible to train large-scale machine learning model on large number of users with meaningful privacy guarantee without loss of accuracy \cite{MRLZ18}, but this requires significant runtime increases. In fact, in the setting of large ML model with large number of users, maintaining good accuracy of the model can be achieved but with a trade-off on runtime and privacy budget. The goal of this project is to explore other types of privacy amplification or improve existing techniques, either to get tighter bounds of privacy or reducing the runtime. One practical example of this line of work is our DP GAN framework, which utilizes the privacy notion and amplification via sampling from \cite{ACG16} and requires significant runtime increase for training for model differentially privately. This project is with Janardhan (Jana) Kulkarni and Sergey Yekhanin, researchers at Microsoft Research, Redmond.

3 Fair PCA

There are instances that machine learning algorithms produce “biased” outcomes in recent years. For example, recidivism prediction software has labeled low-risk African Americans as high-risk at higher rates than low-risk white people \cite{ALMK18}. Proposed solutions can be categorized mainly into two types: scaling or data sampling schemes of the training data, such as weighting labels in minority group heavier; and modifications in optimization objective and/or constraints in the machine learning algorithms. One missing piece of these attempts is that bias may be introduced during the intermediate steps of data processing, such as
in dimensional reduction. Our focus is on principle component analysis (PCA), probably the most fundamental dimensionality reduction technique in the sciences generally [Hot33, Jol86, KPF01]. For example, on a real-world faces data set, PCA incurs much higher reconstruction error for women than men, even if male and female faces are sampled with equal weight [5]. This motivates a search for a notion of “fair” PCA and an algorithm for fair PCA that performs well in theory and practice.

**Current Work** Our work [5] with Samira Samadi, Jamie Morgenstern, Mohit Singh, and Santosh Vempala, defines a new notion of fairness in PCA as equalizing the additional loss of structure in each group as a result of dimensional reduction. Though it is unclear whether solving for a fair PCA to optimality is possible, we obtain theoretically and practically intriguing results. Theoretically, we make a connection between the geometry of extreme points in polyhedron and the rank of a projection matrix, showing that a solution to fair PCA when two groups are present can be obtained with at most one extra dimension required to represent the data. Practically, we exploit the mathematical structure of the problem and hence propose a multiplicative weight update method to fasten the run of algorithms. Experiments on real-world data sets show that our algorithm takes at most constant factor (∼10−15 times) longer than a standard PCA, which is easily solvable (e.g. by Singular Value Decomposition), and does not require any extra dimension in order to obtain optimal fair PCA as suggested by theory. The code is publicly available for use.

The follow-up work [6] generalize the result of fair PCA to the setting when more than two groups are present and to other social welfare objectives of interests, which will widen the applicability and performance of our algorithm. For example, ethnicity and level of income can be divided into more than two groups. We developed an algorithm that solves fair PCA in arbitrary number of groups that adds significantly less number of extra dimensions to the solution than existing results, and an approximation algorithm with satisfying approximation guarantee. The new result follows from a connection between the rank of a matrix and the geometry of a semi-definite cone containing the space of all projection matrices, whereas the previous work make a connection to a geometry of a polyhedron. We connect fairness notion to social welfare literature, specifically bargaining solution in mathematical economics which provides multiple criteria of fairness. When more than two groups are present, each bargaining solution criterion leads to a different objective for fair PCA, generalizing our definition of fair PCA problem to multiple variants. The algorithms and analysis to solve many of those variants we obtained has similar guarantees as the standard fair PCA one.

**Future Work** Our fairness notion also applies beyond issues risen from ethical or legal obligation to variety of disciplines, such as in the sciences. Data in the sciences often contain unequal representation of different types of labels, e.g. due to different costs and ease of access to different types of data. PCA can conceal the structure of smaller group of labels due to the overwhelming presence of another group of labels. We plan to explore scientific datasets and show that our fair PCA algorithm maintains the structure of all groups better than existing technique.

In addition, we want to explore the empirical effect when standard PCA in the pipeline of machine learning process is replaced with our fair PCA. In particular, we want to explore to what extend the bias introduced in PCA propagates to the bias at the final stage of ML, and how much fair PCA are able to remedy the bias.

**My Publication**


**Other References**


