Research Statement

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

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. I spend time 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), differential privacy, and fair principle component analysis (fair PCA).

Optimal design Mohit Singh 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 [BMI13], sensor placement [JB09], matrix sparsification [BSS12, SS11], column subset selection in numerical linear algebra [AB13], 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) [KT+12] 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 practice include local search heuristics, such as Federov exchange [F+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 most recent result with Mohit Singh and Aleksandar Nikolov [3] obtains the best approximation ratio known for two very commonly used criteria (A 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.

Future Work In many applications, the design the 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 [3] 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 [NS16], theory of stable polynomial [Gur08], expander graphs, graph sparsification, the Alon-Boppana bound [Alo86, Nil91] in spectral graph theory, and restricted invertibility principle (RIP) [BT87]. 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.

The second direction is the analysis of combinatorial approaches that are used in practice, such as a local search algorithm known as Fedorov exchange and implemented in SAS software. As mentioned earlier, no strong theoretical guarantee for these types of algorithms has been proven. The aim is to provide theoretical underpinning to these widely-used algorithms. We have partial result on one specific criterion, and plan to extend our analysis to more optimality criteria and other combinatorial algorithms. The partial result gives an explanation why statistic community has enjoyed the seeming success of such heuristics in most cases, and we aim that a complete picture of combinatorial algorithms will provide a fix to those cases where common heuristics fail.

Thirdly, in collaboration with statistics professor Chien-Fu Jeff Wu and his students, I aim to extend our results to optimal design for the generalized linear model. Such models are utilized more regularly in practice. There is, however, less consensus on what kind of theoretical statement is considered satisfying, so the goal is to propose first appropriate models, and then obtain algorithms with guarantees.

2 Differential Privacy

Many learning algorithms today deal with personal, sensitive data, including patience health records, GPS locations, and browsing history. First defined by [DMNS06], differential privacy (DP) gives a mathematically rigorous worst-case bound on the maximum amount of information that can be learned about any one 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 [JLE14] for a survey). Differentially private algorithms have also begun to be 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 [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 [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) [GPAM+14] 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 [ACG+16], followed by several optimization works [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 [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. The implementation and performance will be submitted to NIST, and be publicly available for academic use.

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 [KCZ+18] 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 [DR14] through utilizing other mathematical notions of privacy [DR16, BS16, ACG$^+$16, Mir17] and algorithmic techniques that provably increase privacy, as known as privacy amplifications. All privacy amplifications known so far are obtained by subsampling the data (see, for example, [KLN$^+$11, WFS15, ACG$^+$16, WBK18, BDRS18]), until recently [FMTT18] provides another type of privacy amplification. [FMTT18] shows an increase of privacy on an individual's sensitive information after the algorithm runs more noisy iterations over that sensitive data. The result holds even if those later noisy iterations are nontrivial and depend on other individuals' information. An example of this setting is stochastic gradient descent, where an individual's data is used for computing the gradient at some step, and not used for many steps later. The goal of this project is to explore other types of privacy amplification or improve existing ones, either to get tighter bounds over the same sets of assumptions or results over more general settings. One practical example of this line of work is our DP GAN framework, which utilizes the privacy notion and amplification via sampling from [ACG$^+$16]. However, privacy amplification by iterations [FMTT18] applies to convex optimization and hence does not apply to our DP GAN, as training GANs in general is a nonconvex optimization. Therefore, I want to find an extension of [FMTT18] or new privacy amplification technique that applies to nonconvex optimization. This project is with Janardhan (Jana) Kulkarni, a researcher 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 [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 [4]. 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 [4], to appear in NIPS 2018, 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 ($\sim 10 - 15$ 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.

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. Due to its wide applicability, we plan to release our implementation that is easy-to-use for researchers in any discipline other than computer science.

In addition to extending the applications of fair PCA, we will 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. Preliminary research gives a method to solving 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 also 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. Preliminary research shows that we obtain algorithms and analysis to solve some of those variants with similar guarantee as the original one.

My Publication

[1] Digvijay Boob, Rachel Cummings, Dhamma Kimpara, Uthaipon Tao Tantipongpipat, Chris Waites, and Kyle Zimmerman. Differentially private synthetic data generation via GANs. *To be presented at Theory and Practice of Differential Privacy (TPDP 2018) workshop*, 2018.

[2] Rachel Cummings, Sara Krehbiel, Kevin A Lai, and Uthaipon Tantipongpipat. Differential privacy for growing databases. *To appear in Thirty-second Conference on Neural Information Processing Systems (NIPS)*, 2018.

[3] Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat. Proportional volume sampling and approximation algorithms for A-optimal design. *To appear in ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2019.

[4] Samira Samadi, Uthaipon Tantipongpipat, Jamie Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair PCA: One extra dimension. *To appear in Thirty-second Conference on Neural Information Processing Systems (NIPS)*, 2018.

Other References

[AB13] Haim Avron and Christos Boutsidis. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[ALMK18] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias propublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2018. (Accessed on 05/16/2018).

[Alo86] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986. Theory of computing (Singer Island, Fla., 1984).

[BDRS18] Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86. ACM, 2018.

[BJWWG17] Brett K. Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, and Casey S. Greene. Privacy-preserving generative deep neural networks support clinical data sharing. bioRxiv preprint 159756, 2017.

[BMI13] Christos Boutsidis and Malik Magdon-Ismail. Deterministic feature selection for k-means clustering. *IEEE Transactions on Information Theory*, 59(9):6099–6110, 2013.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[BSS12] Joshua Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

[BT87] J. Bourgain and L. Tzafriri. Invertibility of large submatrices with applications to the geometry of banach spaces and harmonic analysis. *Israel journal of mathematics*, 57(2):137–224, 1987.

[CDKV16] L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, 2006.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(34):211–407, 2014.

[DR16] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[F+55] Walter Theodore Federer et al. Macmillan Co., New York and London, 1955.

[Fed72] Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 1972.

[FMTT18] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. *arXiv preprint arXiv:1808.06651*, 2018.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[Gur08] Leonid Gurvits. Van der waerden/schrijver-valiant like conjectures and stable (aka hyperbolic) homogeneous polynomials: one theorem for all. *the electronic journal of combinatorics*, 15(1):66, 2008.

[Hot33] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[HR10] Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proceedings of the 51st annual IEEE Symposium on Foundations of Computer Science*, FOCS '10, pages 61–70, 2010.

[JB09] Siddharth Joshi and Stephen Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009.

[JLE14] Zhanglong Ji, Zachary C. Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. arXiv pre-print 1412.7584, 2014.

[Jol86] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[KCZ+18] Sara Krehbiel, Rachel Cummings, Wanrong Zhang, Yajun Mei, and Rui Tuo. Differentially private change-point detection. *To appear in Thirty-second Conference on Neural Information Processing Systems (NIPS)*, 2018.

[KLN+11] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011.

[KPF01] LIII KPFRS. On lines and planes of closest fit to systems of points in space. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (SIGMOD)*, 1901.

[KT+12] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[Mir17] Ilya Mironov. Renyi differential privacy. In *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*, pages 263–275. IEEE, 2017.

[Nil91] A. Nilli. On the second eigenvalue of a graph. *Discrete Math.*, 91(2):207–210, 1991.

[NS16] Aleksandar Nikolov and Mohit Singh. Maximizing determinants under partition constraints. In *STOC*, pages 192–201, 2016.

[PSC18] NIST PSCR. The unlinkable data challenge: Advancing methods in differential privacy. https://www.herox.com/UnlinkableDataChallenge, 2018.

[Puk06] Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

[SS11] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

[WBK18] Yu-Xiang Wang, Borja Balle, and Shiva Kasiviswanathan. Subsampled r\'enyi differential privacy and analytical moments accountant. *arXiv preprint arXiv:1808.00087*, 2018.

[WFS15] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502, 2015.

[WUSK18] Yining Wang, Erva Ulu, A. Singh, and Levent Burak Kara. Detc 2018-85310 efficient load sampling for worst-case structural analysis under force location uncertainty. 2018.

[WYS17] Yining Wang, Adams Wei Yu, and Aarti Singh. On computationally tractable selection of experiments in measurement-constrained regression models. *The Journal of Machine Learning Research*, 18(1):5238–5278, 2017.

[ZJW18] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model. arXiv preprint 1801.01594, 2018.