Yu Sun

✉ ysun258@wisc.edu · ☎ (+1) 608-422-2239 · in yusunwisc · YuSunwisc

Education

Math PhD Candidate

- University of Wisconsin Madison
- Sep 2017– expected Dec 2022 or May 2023
- Advanced Stochastic Analysis, ML, DL, Bayesian Inference, Graph Theory, SPDE, PDE, Phylogeny

CS/Math Master

- University of Wisconsin Madison
- Sep 2020– May 2022/ Sep 2016– May 2017
- CV, NLP, Ranking, Convex Analysis, Algorithm Design. Advanced Stochastic Analysis, Computational Math.

Math BS

- Shanghai Jiao Tong University(上海交大)
- Sep 2012– June 2016
- Probability Theory, Computational Math, Statistics, Algorithm Design

Internship Experience

Meta Instagram Reels - MLE intern (Managers: Martin Wang, Boxuan Zhong) May 2022 – Aug 2022

- Set an end-to-end media relevance flow to detect anomaly, and improve final delivery quality.
- Set an end-to-end flow to increase the media sourcing pool based on clustering.
- Build interactive data visualization and flow performance evaluation tool.

Technical Skills

- Programming Languages: Python, SQL, Matlab, R, C++.
- Tools: Pytorch, Tenserflow, Jupyter, Pandas, Sklearn, Presto, Hive, Bash, Mercurial, Git.
- Ranking, Bayesian Inference, CV, NLP, Predictive Modeling, Clustering, Complex Stochastic Analysis.

Research Experience

Machine Learning & Deep Learning

- **Community Detection** and **Semi-Supervised Learning** : Adapt recent developed tensor decomposition technology for Degree Correlated Block Model(DCBM) to semi-supervised learning, i.e. with extra matrix info about partial labeling. Provide a poly-$O(\log n)$ algorithm with consistency results. Quantify the impact of semi-supervised learning than supervised learning in community detection. Aug 2021 – Present
- **Few Shots Learning**, **Differential Privacy** and **Fairness**: Using few shots learning technique to reduce the model accuracy lose due to differential privacy, with theoretical guarantees on mild regularized condition, and empirical results with Reddit posts dataset. We also find a way to control the loss of differential privacy and fairness loss concurrently, under some mild regularity condition. Sep 2021 – Present

Bioinformatic Algorithm Design

- **Distance Based Method**: Using internode distance to recover the network structure of phylogeny under network multispecies coalescent model. Also providing a poly-$O(\log n)$ algorithm with high probability to successfully reconstruct. This is the first distance based method proof in a phylogeny network. Nov 2020 – Aug 2021
- **Quartets Based Method**: Derive a new algorithm with high probability reconstruct the specie tree with incomplete lineage sorting, with rigorous upper bound on sample complexity depends poly-logarithmically on number gene tree samples, proved by stochastic domination technique. This result improves the state of the art algorithms when species tree branch lengths are more evenly distributed. March 2021 – Present

Integrable Probability

- **Right tail LDP rate function**: Calculate the right tail LDP rate function of free energy in Beta-Gamma directed polymer models. This is the first summarized right tail LDP result in discrete KPZ universality class. May 2018 – Dec 2019

ML & DL Project

- **NLP in phylogeny**: Use **Transformer** to solve the phylogeny sequence placement problem. Analyze how the attention mechanism performs better than traditional RNN in long sequence cases. Compare the effect of choosing different attention layers to reduce the computational cost. May 2021 – Present
- **Outlier Robust Learning**: Learn the distribution of Ising model with ϵ-adversarially corruption. Give a poly-$O(\log(N))$ algorithm to reconstruct the almost identical distribution with high probability. Feb 2021 – May 2021
- **Clustering**: Using state of the art clustering algorithm AncestralClust on COVID-19 RNA sequences datasets provided by NCBI. Compare with other unsupervised clustering techniques, measured by evenness. Jan 2021 – May 2021
- **Correlation detection**: Use alignment and slicing techniques to detect the correlation between gene sequences on large scale datasets. Oct 2020 – Dec 2020
- **Supervised Learning**: Use decision tree, random forest and logistic regression to predict whether an indigenous person that tested positive for COVID-19 would end up in the ICU. Sep 2020 – Dec 2020