留学報告書　　上原雅俊 2020/1/8

今、１年間の猶予期間をもらって UIUC がある Champaign(イリノイ州)にいます。Champaign は主に大学しかない街です。人が多すぎずちょうど勉強できるカフェがたくさんあるのがいいことです。それ以外はあまり街としていいことはない気がします。毎日、３個ぐらいカフェをはしごして研究しています。主に研究、学会、授業について書こうと思います。

研究

研究でよかったことは以前書いていた主著の論文たちが Neurips に１本、AISTATS に２本通ったことです。(どっちとも機械学習だといい会議です。) 特に AISTATS の２本の論文は一度、他の所でリジェクトを受けていたので、一安心です。去年は主著で 6 本論文、2 本共著者として書けて少しはヒットが打てるようになってきたかなと思うので、これからはインパクトがある研究をやっていくように心がけていきたいです。以下では最近、書き上げた論文について簡単に書きます。

Efficiently Breaking the Curse of Horizon in Off-Policy Evaluation with Double Reinforcement Learning
https://arxiv.org/abs/1909.05850

今までの書いてきた中で一番の自信作の論文です。Cornell のメンターの Nathan との共著です。貢献は強化学習の Off policy evaluation の誤差のあらゆる推定量の漸近下限を求め、実際に漸近下限を達成する推定量を構成したことです。因果推論勉強したことある人なら聞いたことある推定量の double robustness 性が鍵になってきます。以前も似た様な論文を出しているのですが、こちらの方では MDP 上(transition density や reward distribution が Markov かつ Time invariant な条件のもと)の Off policy evaluation を行っています。以前、書いた論文では TMDP 上(Markov だが time variant な条件のもと)の Off policy evaluation の論文を書きました。違いは time invariance 性のみですが、この性質によってエルゴード性が使えて、time invariance だと Horizon が長くなるにつれ Off policy evaluation が精確になるということが言えます。そのためにタイトルで breaking the curse of horizon と言っています。

Minimax Weight and Q-Function Learning for Off-Policy Evaluation
https://arxiv.org/abs/1910.12809

UIUC に移ってからメンターの Nan と書いた論文です。強化学習で大事な Q-function を Minimax estimation するというのがメインアイデアです。応用として Off policy evaluation のことも議論したり、最近、popular な quantity になりつつある behavior policy と target policy から生まれる state distribution の比に関する minimax estimation についても話して

います。Minimax estimation の idea 自身は他の文脈であるのですが、この状況においては Discriminator の double robustness という話が出てきてそれが面白いです。また Model based RL が tabular(離散 state, action)の状況で前述の漸近下限を達成するということも appendix で証明していて、もっと一般の空間でも証明できたらいいなと思っています。

Localized Debiased Machine Learning: Efficient Estimation of Quantile Treatment Effects, Conditional Value at Risk, and Beyond
https://arxiv.org/abs/1912.12945

Cornell のメンターの Nathan と友達の Xiojie との論文です。今、causal inference の Average treatment effect の推定だと debiased (double) machine learning という手法がメインストリームになっています。Quantile treatment effect を推定する場合はそのまま使いにくいので、そこをどう簡単にやるかを提案した論文です。元が one step の所を Two step estimator みたいな感じでやります。もともと、一般の causal DAG に拡張の話をやろうという話で始めたので、そこらへんでもっと貢献できたらと思います。

最近の研究

今、頑張って Off policy learning の論文を何本か書いています。Off policy learning は過去のデータに基づいて、どう薬を個々の人に投薬するか、どう広告を個々に配信するか、どう教育メソッドを個々に施すか、などという policy を学習するという領域です。 ML, Stat, Biostat, Econometrics などでそれぞれ強い人がやっていて、聴衆も多くてやっていて楽しいです。合間を縫って、RL の Online learning、causal DAG の identification の理論、IO (Industrial organization)の構造推定の勉強をしています。それらの研究も早くやれたらなと思います。

学会

通った論文を発表するため Neurips に参加しました。機械学習で多分一番大きい学会です。人多すぎてポスター発表で他の人の研究を聞くみたいなことは正直、難しかったのですが、色んな人と話せて楽しかったです。勿論、似た研究やっている人と色々、知り合いになれたのもよかったです。あとインターンの面接的な感じで Google research, Microsoft research の人と話したりしました。毎年、コンスタントに出せるよう頑張りたいです。

授業

聴講なのですが Online learning の授業をとっていました。Bandit から始まって、強化学習の online learning まで習いました。 例えば Linear UCB の証明を追ったことがなかったのですが、意外と難しくて面白かったです。早く強化学習の Online learning の研究はしたいです。残タスクが多くて本腰がなかなか入れられていないですが。