船井情報科学振興財団留学報告書 上原雅俊 2021/7/15

今住んでいる NYC はだいぶ人が戻ってきて買い物も色々できるし、レストランも美味しいし、色んなアトラクションがあってとっても楽しいです。ボストンも住みやすかったけど、ニューヨークも違った良さがあって、ずっと住むのもいいなと思っています。

最近、 [1]が COLT という学習理論で一番いい会議に[2]が ICML という機械学習一般のトップカンファレンスに通りました。[2]は船井財団生の斎藤くんとの論文です。

[1] Fast Rates for the Regret of Offline Reinforcement Learning
[2] Optimal Off-Policy Evaluation from Multiple Logging Policies

論文の内容は前回の報告書に書いてあります。COLT はコロラドのデンバーで今のところ、In Person で会議やるらしいので、久しぶりの会議が楽しみです。

前回の報告書から三個論文書いて、一つ授業とったので、そのことについて簡単にかきます。特に私がフォーカスしているうちの一つの（理論ベースの）強化学習は最近、深層学習、オンライン学習の強い人達が入ってきたおかげで(^^)盛り上がりつつも競争が激しくなってきてなかなか大変です。実際に 3 個目の論文は他のグループがやっていると聞いて、とりあえず完成したバージョンを今週 Arxiv にあげました。

研究

Mitigating Covariate Shift in Imitation Learning via Offline Data Without Great Coverage
https://arxiv.org/abs/2106.03207

初めて書いた Imitation learning の論文です。（Cornell の PHD 生とアマゾンの研究者と最近 Microsoft Research からきたアドバイザーとの共著論文）Imitation learning は将棋のプロの指し手や、ゲームのプロの動き、テニスのプロの動きなど事前にあるエキスパートの情報を使って、アルゴリズムやロボットにエキスパートの知見をどうやって効率的に教えるという分野です。従来の強化学習が一からロボットに教える方針なのに対し、Imitation learning だとどうやってプロを模倣するかという話になってきます。今回の研究だと Expert data と Expert 以外からの Offline data があるときに、offline data をどうやって効率的に使うかというアルゴリズムを提案しています。特に実験的にアルゴリズムが優れていることを主張しているだけでなく、Distribution shift (Covariate shift) という従来から知られている imitation learning でよく出てくる問題を避けることができることを学習理論の観点から示しています。

Causal Inference Under Unmeasured Confounding With Negative Controls: A Minimax Learning Approach https://arxiv.org/abs/2103.14029

未観測の交絡因子がある時の因果効果の推定手法の研究です。（Cornell のメイン の Advisor と最近、卒業して tsinguha の助教になった友達との共著論文）社会科学や医学だと、観察データからの因果関係（マスクのコロナに対する因果関係とか）を、交絡因子が全てコントロールされている仮定のもとでよく解析します。ただ、そんな仮定が正しいとは誰もわからないし、常に未観測の交絡因子はある程度存在します。そのような場合、仮定から少しずらした上で色々解析する感度分析という手法、Point estimate を諦めて、区間を推定する Partial Identification の２種類がよく使われます。最近、それら以外の Negative control という Proxy variable を使う賢い手法が出てきて、今回はその手法論を発展させたという論文です。
従来の論文だと、手法がパラメトリックモデルに限定されていたのですが、RKHS とかニューラルネットをどうやって使い、その時どんな仮定のもとで推定量の収束レートや、漸近正規性が保証されるかを解析した論文です。単純な回帰や分類問題だと、ただパラメトリックモデルをニューラルネットに置き換えれば済む話ですが、Negative Control の話だとそういうことはできずに、本質的に推定量を変える（ミニマックス推定に置き換える）必要があります。実は前書いたオフラインの強化学習の話と繋がっていて、どちらもノンパラメトリック IV の問題として定式化できます。ただ、統計学者や計量経済学者がだいたい知っている単純なノンパラメトリック IV と比べて、色々な少し複雑に見えるけど綺麗な構造が入っていて、そこが面白いところです。

Pessimistic Model-based Offline RL: PAC Bounds and Posterior Sampling under Partial Coverage. https://arxiv.org/abs/2107.06226

オフライン強化学習の論文です。（Microsoft research からきた新しいアドバイザーとの共著論文）強化学習でもっとも一般的なオンラインの設定のもとだと Optimistic に学ぶ（どんどん探索していくように）のが実験的にも理論的にも何十年も当たり前になっています。オフラインの設定だと逆に Pessimistic にしたほうがいいというのがここ１年で通説?になってきています。実際にそうすると、昔のオフライン強化学習でよく使われていた Global Coverage という仮定を緩めることができます。今回はそのような仮定を外せるという理論保証のあるアルゴリズムで最も一般的な Model based の（ほとんど全てのモデルを学べる）手法を提案しました。具体的に Tabular MDPs, Linear MDPs, Linear Mixture MDPs, Kernelized nonlinear regulators (船井財団生の大西さんの論文のモデル)、Gaussian Processes での収束レートを求めています。ただ、そのアルゴリズムは Computationally Inefficient なので、Thompson sampling based の Natural

Policy Gradient を使うことを最終的に提案していて、それを使えば若干弱いベイジアンリグレットの意味で同じよう保証ができることを主張しました。Thompson sampling が Optimism も Pessimism も包含するのは不思議で面白いなと思います。

授業

Networks and Markets メカニズムデザイン・オークション・ゲーム理論の CS 的な観点からの授業です。オークションのプライシングの基礎になっている VCG メカニズムとか学んで感動しました。研究レベルまで知識をもっていくのはなかなか大変そうですが、将来的に自分がいま持っているものと繋げられたらいいなと思います。

最後に

最近は卒業後のこと考えることも増えて、時々大学や企業の人とキャリアについて話したりしています。あと今年は目の前の研究と色んな業務に忙殺されていて、前みたいにゆっくり色んな論文を読んだり新しいことを勉強する時間がかなり減っているので、それも確保できたらなと思います。あと特に Causal Graph と Online learning はかなり知識が増えたと思うので、なんかいい論文が書けたらなと思います。