船井情報科学振興財団留学報告書　上原雅俊 2020/6/16

最近、主著の強化学習の論文が３本、International Conference on Machine Learning (ICML)という機械学習の良い学会に通りました。やっぱり論文が通るととても嬉しいです。ジャーナルに投稿していた、他の論文たちも査読結果が返ってきてリバイズすればおそらく通りそうなので、その努力が報われればと思います。

また現時点で所属はハーバードの統計学科なのですが、このまま何もなければ９月からコーネルの CS の PHD に移ることになりました。キャンパスはイサカではなくニューヨーク市になります。移ることにした理由は（１）遠隔で仕事しているコーネルの先生と気が合う、（２）NYC に優秀な強化学習の理論系の研究者が多いから一緒に働けそう（コロンビア, MSR, Google brain とか）、（３）PHD に入って書いた１４本の論文のうちハーバードの先生とまだ論文書いていないことが理由です。自分がメインでやっている研究は因果推論と強化学習の中間ですが、因果推論についてはハーバードで様々な先生や生徒と Discussion して広範な視野を手に入れることができて、それが今の自分の武器になり、その視野で強化学習の研究をするのが独自性になっています。なので、特にハーバードの環境に不満があった訳ではないです。友達もできたし生活するにはいい場所なので、離れることは少し寂しいですが、新しい所に行っても頑張りたいと思います。

最後に前回報告書書いた時から何個か論文を書いたので簡単に概要を記します。

Statistically Efficient Off-Policy Policy Gradients
https://arxiv.org/abs/2002.04014
ICML に通った論文です。強化学習における Offline Policy Gradient の理論研究です。Policy Gradient とは最適な方策を直接的に学ぶアルゴリズムの総称をさします。Offline(観察研究) のもと、少ないサンプルで最適な方策を学べるというアルゴリズムを提案しました（専門的にいうと Horizon に対して Polynomial なリグレット保証)。主要なアイディアは統計的に最適な勾配推定法を考えてそれを学習に使うということです。ただ、残念ながら提案アルゴリズムは複雑な関数近似法を使うと機能しないです。強化学習はまだ理論と実験の隔たりが大きいです…

Off-Policy Evaluation and Learning for External Validity under a Covariate Shift
https://arxiv.org/pdf/2002.11642.pdf

観察研究の元、Historical data と Evaluation data の共変量の分布が違う時にどのように Off policy evaluation(Evaluation policy の Policy value の推定)を行うか研究した論文です。例えば、どのような方策(マスクつける、何か注射打つ)に対する COVID-19 の罹患率の因果効果を知りたいとして、仮にアメリカで得られたデータを使って因果効果を推定しても、その結論が日本にそのままあてはまるかは怪しいでしょう。一つの簡単に考えられる原因はそれぞれの Population の共変量 が違う（人種構成や年齢構成とか）ことです。この問題は推定された因果効果が他の集団においてそのまま適用はできないという外的妥当性の問題として知られていて、今回はその設定のもと推定下限やそれを達成する推定量を考えたというのが貢献です。外的妥当性の研究はモチベーションがわかりやすいので、研究としても盛んだと思いますが、統計的な意味で意外と理論的に深く掘られた研究が少なかったので、Cyber Agent の方としたという研究です。

Efficient Evaluation of Natural Stochastic Policies in Offline Reinforcement Learning
https://arxiv.org/abs/2006.03886

Natural Stochastic Policy と呼ばれる Policy たちを学ぶための強化学習の枠組みを考えたという研究です。例えば、観察実験において、１週間に A 分運動するという Policy が５年以内の致死率に与える影響を調べた高齢者のデータがあったとします。そこでよくあるタスクが、個々のデータ（年齢、性別、飲酒してるか…）が与えられたとき、その人にとって致死率が最も低くなるような最適な運動時間を返す（最適な Policy を求める）ことです。そのとき、A＋delta 分みたいな観察実験で観測された値に依存した Policy を Natural Stochastic Policy と言います。この値に依存させて（例えば Delta に制約をかけて）、実装可能な Policy にします。観察実験の値を無視していないというのが、普通の Policy との違いです。（普通の Policy では、３０分１週間に運動している人に６時間運動させるような Policy になってしまい、実装可能ではなくなってしまいます。）

本論文では Natural Stochastic Policy の Policy value を評価する上での、理論限界とそれを達成する手法を考えたというのが貢献です。

Doubly Robust Off-Policy Value and Gradient Estimation for Deterministic Policies
https://arxiv.org/abs/2006.03900

行動空間が連続の時に決定的な Policy を評価し、学ぶ理論保証つきのアルゴリズムを考えたという論文です。強化学習のベンチマークだとよく、車を上下左右に動かす（行動

空間は 4 個）、臨床試験のベンチマークだと薬を与える、与えない（行動空間は 2 個）だったりします。しかし、現実的には行動空間は連続であることも多いです。例えば車をただ上下左右ではなく A 度の向きに動かすなど。その場合、Stochastic な Policy の評価は簡単ですが、Deterministic な policy の評価は難しいことが知られています。今回は行動空間の滑らかさを仮定して、Deterministic な policy の評価手法と学習手法を良いレートの保証付きで考えたというのが貢献です。例えば Deterministic Policy Gradient という既存手法が有名ですが、この場合、良いレートを手に入れるためには条件が強くなります。提案手法だとその条件が緩いというのが売りです。