今年の夏はパラアルトにあるAmazonインターンでしていました。パラアルトは気候がいいし、スタンフォードや近くの企業にある色々な研究者に会えて刺激的でした。プロジェクト内容は具体的なことをどこまで書いていいのかわからないですが、Best Artm Idenficationのアルゴリズムを Amazon Sarchに応用するということをしていました。論文になるよう、今執筆しています。

また夏には3つの学会、Berkelyで行われた因果推論のワークショップとTTICで行われた強化学習のワークショップとICMLという機械学習全般でNeurips, ICLR?と並んで一番いいとされているに学会に梯子で参加しました（する予定でした）。ただ、最初のTTICで行われた学会の後にCOVIDにかかって、ICMLがその直後にあったためずっとホテルにいました。発表の準備もかなりしていてIn Personで参加できなたかったはとても残念でしたが、以下の論文がアクセプトされました。

Chengchun Shi (*), Masatoshi Uehara (*), Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in partially observable markov decision processes. ICML 2022 (Long presentation)

Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Wen Sun, and Alekh Agarwal. Efficient reinforcement learning in block mdps: A model-free representation learning approach. ICML 2022.

内容は前回の報告書に載っています。あとは前回の報告書以降書いた論文を簡単にまとめます。Google Brainで夏に講演したときの資料に詳しくはまとまっています。

Masatoshi Uehara, Ayush Sekhari, Jason D. Lee, Nathan Kallus, Wen Sun. Provably Efficient Reinforcement Learning in Partially Observable Dynamical Systems. arXiv preprint arXiv:2206.12020

POMDPのモデルを統一的に(オンラインで）統計的に学習するためのフレームワークを提示した論文です。今までの統計的学習に基づいた論文はほとんどMDP (Markovian Deiscion Process)というモデルに基づいていました。ただ、現実的にMDPが成立することはほとんどなくて、実際の学習環境はノイズが乗ったParitally Observable MDPs (POMDPs)に近いことが知られていますた。例えば自動運転では実際の位置は観測されてないですが、ノイズが乗った位置が観測されます。ただ、今までの統計的学習理論の枠組みの保証はMarkovianという仮定に完全に依存していて、POMDPを議論していた論文はここ数年で数本、あっても特定のモデルに合わせられたアルゴリズムしか知られていませんでした。今回の論文ではこのようなモデルや他の新しいモデルでも働くような統一的なアルゴリズムを提案しました。実験がなくて８０ページ近くある若干ゴツい論文ですが、基本の考え方やアルゴリズムはとてもシンプルで”未来のObservaitonsを使ったModel-free RL”という一言に語れます。長いのはフレームワークが色々

な具体的なモデル（最近の機械学習のモデルから古典的な制御論で学ぶLQGまでで) 働くことを示しているからです。今までの書いてきた論文の中でも最も好きな自信作です。

Masatoshi Uehara, Haruka Kiyohara, Andrew Bennett, Victor Chernozhukov, Nan Jiang, Nathan Kallus, Chengchun Shi, and Wen Sun. Future-Dependent Value-Based Off-Policy Evaluation in POMDPs. arXiv preprint arXiv:2207.13081

上のOnline RLの論文のOffline RL Versionです。この論文でも”未来のObservations”を使うことが鍵になるのですがオフラインの場合はさらに条件が必要で過去をどのように使うことに関しての議論が必要になってきます。統計の用語でいうとPOMDPsでは過去を操作変数として自然に見做せるという観察を使っています。

Masatoshi Uehara, Ayush Sekhari, Jason D. Lee, Nathan Kallus, Wen Sun. Computationally Efficient PAC RL in POMDPs with Latent Determinism and Conditional Embeddings.　arXiv preprint arXiv:2206.12081

上の二つは統計的な効率性にフォーカスを置いていましたが、この論文では計算論的な効率性についてもフォーカスを置いています。

Wenhao Zhan, Masatoshi Uehara, Wen Sun, Jason D. Lee.　PAC Reinforcement Learning for Predictive State Representations　arXiv preprint arXiv:2206.12081

PSRs (predictive state represetations) は強化学習を作ってきた巨人たち（Sutton,Littman, Satinder）によって２０年前に考案されたMDPやPOMDPsよりさらに強力な表現能力を持ったモデルです。その強力な利点にもかかわらず、提案時の原著の解読が難しいことや、スタンダードな教科書に載っていないことから、現代の新参の強化学習の研究者やユーザーにはほぼ知られていないです。その一人である自分も今年まで知らなかったです。この論文では初めて、PSRsでの統計的オンラインで効率的に学べることを示しました。アルゴリズムはモデルベースでとてもシンプルですが、証明はかなりの新しいテクニックを要します。