Research Summary

Yulai Zhao
Department of E.E.
Tsinghua University

Advisors:
Simon S. Du (UW)
Jason D. Lee (Princeton)
Aurelien Lucchi (ETH Zürich)

Outline

- Work Review
  1. Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games (accepted by AISTATS 2022)
  2. Blessing of Class Diversity in Pre-training (submitted to ICLR 2022)
  3. On the Theory of Optimizing Performative Risk (a technical report)

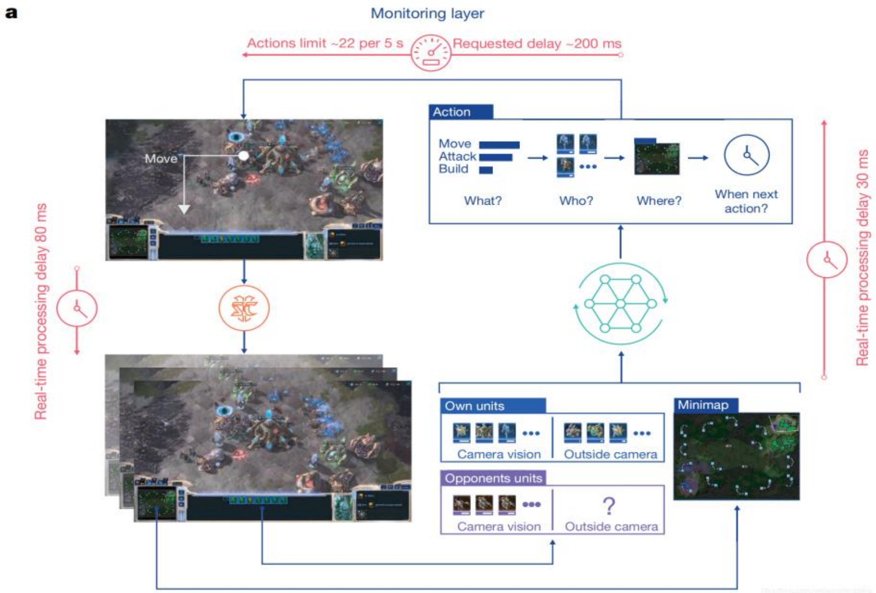
- Questions

Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games

Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Backgrounds

- Two-player zero-sum game is a widely used setting with applications (Go, StarCraft Ⅱ…)
- Policy optimization methods are widely used in solving zero-sum games (AlphaGo, LOLA…)

nature
At last — a computer program that can beat a champion Go player
ALL SYSTEMS GO



Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Problem

- Despite the large body of empirical work on using policy optimization methods for two-player zero-sum Markov games, theoretical studies are very limited.

Can we design a provably efficient policy optimization algorithm with function approximation for two-player zero-sum Markov games with a large state-action space?

Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Setup

- Two-Player zero-sum Markov Games
  - ➢ a tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathrm{r}, \gamma)$: A set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition probability $\mathcal{P}: \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to \Delta(\mathcal{S})$, a reward function $\mathrm{r}: \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to [0, 1]$, a discounted factor $\gamma \in [0, 1)$.
  - ➢ define policies as probability distributions over action space: $x, f \in \mathcal{S} \ \to \Delta(\mathcal{A})$, max player x seeks to maximize the reward while min player f seeks to minimize.
- value function
  - $V^{x,f}(s) = \ E_{\substack{a_t \sim x \\ b_t \sim f}} \left[\sum_{t=0}^{\infty} \gamma^t \, r(s_t, a_t, b_t) \middle| \, s_0 = s \ \right]$
  - $V^{x,f}(\rho) = E_{s \sim \rho} V^{x,f}(s)$

Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Setup

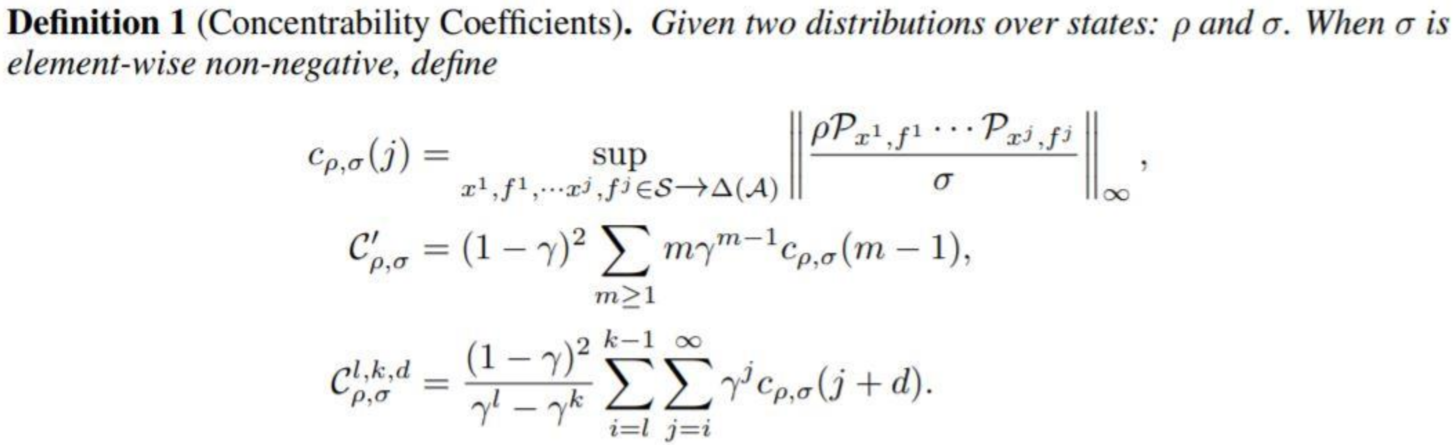
- (x^*, f^*) is a pair of **Nash equilibrium (NE)** if the following inequalities hold for any distribution ρ and policy pair (x, f):

$$V^{x,f^*}(\rho) \leq V^{x^*,f^*}(\rho) = V^*(\rho) \leq V^{x^*,f}(\rho)$$

- Our goal: find an approximate pair of Nash equilibrium, which means output x should make the following metric small

$$V^*(\rho) - \inf_f V^{x,f}(\rho)$$

- We use **concentrability coefficients** as in the previous work [Perolat et al., 2015].



Definition 1 (Concentrability Coefficients). *Given two distributions over states: ρ and σ. When σ is element-wise non-negative, define*

$$c_{\rho,\sigma}(j) = \sup_{x^1, f^1, \cdots x^j, f^j \in \mathcal{S} \to \Delta(\mathcal{A})} \left\| \frac{\rho \mathcal{P}_{x^1,f^1} \cdots \mathcal{P}_{x^j,f^j}}{\sigma} \right\|_\infty,$$

$$\mathcal{C}'_{\rho,\sigma} = (1-\gamma)^2 \sum_{m \geq 1} m \gamma^{m-1} c_{\rho,\sigma}(m-1),$$

$$\mathcal{C}^{l,k,d}_{\rho,\sigma} = \frac{(1-\gamma)^2}{\gamma^l - \gamma^k} \sum_{i=l}^{k-1} \sum_{j=i}^{\infty} \gamma^j c_{\rho,\sigma}(j+d).$$

- ➢ **σ** is the optimization measure we use to train the policy.
- ➢ **ρ** is the performance measure of our interest.

Population Algorithm for Tabular case

- We divide each outer loop into two steps.
  - I. In Greedy Step, we intend to find approximate solution (x, f) for Bellman operator $\mathcal{T}$ onto current value function V_{k-1} with T' updates. (towards V^*)
  - II. In Iteration Step, we run T NPG updates to solve $\arg\min_f V^{x,f}$ which is known as finding the best response of min player when fixing $x = x^k$.

- Theorem 1 (informal): For this setting, after K outer loops

$$V^*(\rho) - \inf_f V^{x^K,f}(\rho) = \widetilde{O}\left(\frac{\mathcal{C}_{\rho,\sigma}^{1,K,0}}{(1-\gamma)^4 T} + \frac{\mathcal{C}_{\rho,\sigma}^{0,K,0}}{(1-\gamma)^4 T'} \log T' + \frac{\gamma^K}{1-\gamma} \mathcal{C}_{\rho,\sigma}^{K,K+1,0}\right).$$

$$\mathcal{T}_{x,f} v = r_{x,f} + \gamma\, \mathcal{P}_{x,f} v$$

$$\mathcal{T} v = \sup_x \inf_f \mathcal{T}_{x,f}\ v$$

Online Algorithm with Function Approximation

- We still divide each outer loop into two steps.
- Assume **Episodic Sampling Oracle** to provide unbiased estimates or a fixed state-action distribution ν_0, we can start from $s_0, a_0, b_0 \sim \nu_0$, then act according to any policy x, f, and terminate it when desired.
  - I. In Greedy Step, our goal is still to obtain a near-optimal x^k with respect to V_{k-1}. Different from tabular case, we use sample-based NPG updates.
  - II. After obtaining x^k from Greedy Step, we run T sample-based NPG updates (each with N samples) to find best response of min player.

- Theorem 2 (informal): For this setting, after K outer loops

$$E\left[V^*(\rho) - \inf_f V^{x,f}(\rho)\right] = \tilde{O}\left(\frac{1}{\sqrt{T}} + \frac{1}{N^{1/4}}\right)$$

Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Contributions

I. In Greedy Step, design a subroutine that found minmax solutions to a matrix game without prior knowledge of model parameters.

II. In Iteration Step, leverage NPG methods to update policies.

III. Finally, develop new perturbation analyses which may be of independent interest for provable multi-agent RL.

Blessing of Class Diversity in Pre-training

Blessing of Class Diversity in Pre-training -- Backgrounds

- Pre-training refers to training a model on a few or many tasks to help it learn parameters that can be used in other tasks

- Past works [Caruana, 1997, Baxter, 2000, Maurer et al., 2016, Du et al., 2021, Tripuraneni et al., 2020a,b, Thekumparampil et al., 2021] studied multi-task training. A notion ***diversity of tasks*** is shown to be crucial to allow pre-trained model to be useful for downstream tasks.

- Roughly speaking, **diverse tasks** ensure that, in the worst-case, representation difference for the downstream task is controlled when the pre-training representation difference is small.

Blessing of Class Diversity in Pre-training -- Problem

- Unfortunately, this line of theory cannot be used to explain the success of pre-training in NLP since they require **a large number** of diverse tasks, e.g., BERT generally pretrained on **few** tasks.

Can we go beyond multi-task training and develop a theory explaining the success of pre-training in NLP?

Besides, can we utilize our theory for better pretraining models in practice?

Blessing of Class Diversity in Pre-training -- Setup

- Follow transfer learning notations, divide the procedure into two phases
  1. Train the representation function and prediction function in pre-training

$$\hat{h} = arg\min_{h \in H} \min_{f^{pre} \in F^{pre}} \frac{1}{n} \sum_{i=1 \cdots n} l(f^{pre} \circ h(x_i^{pre}), y_i^{pre})$$

  2. Fix representation $\hat{h}$ and train the classifier for the downstream task

$$\hat{f}^{down} = arg \min_{f^{down} \in F^{down}} \frac{1}{m} \sum_{i=1 \cdots m} l(f^{down} \circ \hat{h}(x_i^{down}), y_i^{down})$$

- Goal: Let <u>Transfer Learning Risk</u> small

$$E\left[l\left(\hat{f}^{down} \circ \hat{h}\left(x^{down}\right), y^{down}\right)\right] - \min E\left[l\left(f^{down} \circ h\left(x^{down}\right), y^{down}\right)\right]$$

Blessing of Class Diversity in Pre-training -- Results

- **Theorem 1 (Informal)** Under standard regularity conditions, we prove the upper bound of transfer learning risk is related with ν(diversity parameter), Lipschitz parameters, and model complexities.
- **Theorem 2, 3 (Informal)** Let H, f^{pre}, f^{down} be linear mappings. Assume pretraining and downstream training are k and k'-class classifications. Then

$$\nu = \Omega\left(\sigma_r\left(\alpha^{pre}(\alpha^{\mathrm{pre}})^{\mathrm{T}}\right)\right) > 0.$$

➡ In the benign case, transfer learning risk is bounded by

$$O\left(\sqrt{\frac{d\,r^2}{n}} + \sqrt{\frac{r}{m}}\right).$$

r and d are dimensions of representation and raw input, so $r < d$

This is significantly better than not using pre-training, where the risk scales $O\left(\sqrt{\frac{d}{m}}\right)$.

Blessing of Class Diversity in Pre-training -- Results

- We add a diversity regularizer to BERT pre-training(use determinant because least singular value is hard to optimize)

$$L'(\Theta) = L(\Theta) - \lambda \cdot \ln \det(\alpha^{\text{pre}}(\alpha^{\text{pre}})^T)$$

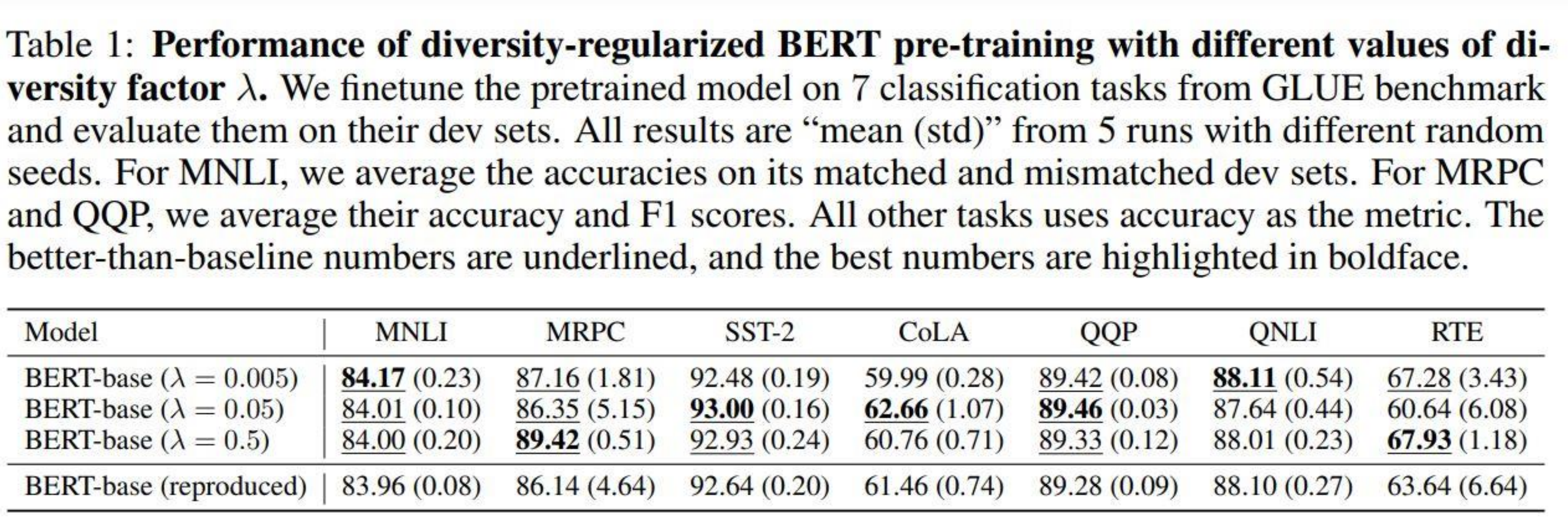


Table 1: **Performance of diversity-regularized BERT pre-training with different values of diversity factor λ.** We finetune the pretrained model on 7 classification tasks from GLUE benchmark and evaluate them on their dev sets. All results are "mean (std)" from 5 runs with different random seeds. For MNLI, we average the accuracies on its matched and mismatched dev sets. For MRPC and QQP, we average their accuracy and F1 scores. All other tasks uses accuracy as the metric. The better-than-baseline numbers are underlined, and the best numbers are highlighted in boldface.

Model	MNLI	MRPC	SST-2	CoLA	QQP	QNLI	RTE
BERT-base ($\lambda = 0.005$)	**84.17** (0.23)	87.16 (1.81)	92.48 (0.19)	59.99 (0.28)	89.42 (0.08)	**88.11** (0.54)	67.28 (3.43)
BERT-base ($\lambda = 0.05$)	84.01 (0.10)	86.35 (5.15)	**93.00** (0.16)	**62.66** (1.07)	**89.46** (0.03)	87.64 (0.44)	60.64 (6.08)
BERT-base ($\lambda = 0.5$)	84.00 (0.20)	**89.42** (0.51)	92.93 (0.24)	60.76 (0.71)	89.33 (0.12)	88.01 (0.23)	**67.93** (1.18)
BERT-base (reproduced)	83.96 (0.08)	86.14 (4.64)	92.64 (0.20)	61.46 (0.74)	89.28 (0.09)	88.10 (0.27)	63.64 (6.64)

Blessing of Class Diversity in Pre-training -- Contributions

I. The first set of theoretical results that demonstrates the statistical gain of the standard practice of NLP pre-training

II. Technically, we introduce vector-form Radamacher complexity chain rule and modified self-concordance condition to refining previous multi-task theories.

III. Develop a new regularization technique for pre-training (e.g., BERT) which boosts the performance of real-world models.

On the Theory of Optimizing Performative Risk

On the Theory of Optimizing Performative Risk -- Setup

- **Performative Risk** is introduced when prediction causes a change in the distribution of the target variable, i.e.,

$$\boldsymbol{PR(\theta) = E_{z\sim D(\theta)} l(z;\theta)}$$

  - Our ultimate goal is to find $\boldsymbol{\theta_{po} = \ \ argmin_{\theta}\, PR(\theta)}$.
  - However, past work mainly focused on finding $\boldsymbol{\theta_{ps} = \mathrm{argmin}_{\theta}\, E_{z\sim D(\theta_{ps})} l(z;\theta)}$

- Example: predicting credit default risk. A bank might estimate that a loan applicant has an elevated risk of default if he applied for a loan, and will act on it by assigning a high interest rate.

On the Theory of Optimizing Performative Risk -- Problem

How and Under what conditions could we optimize performative risks?

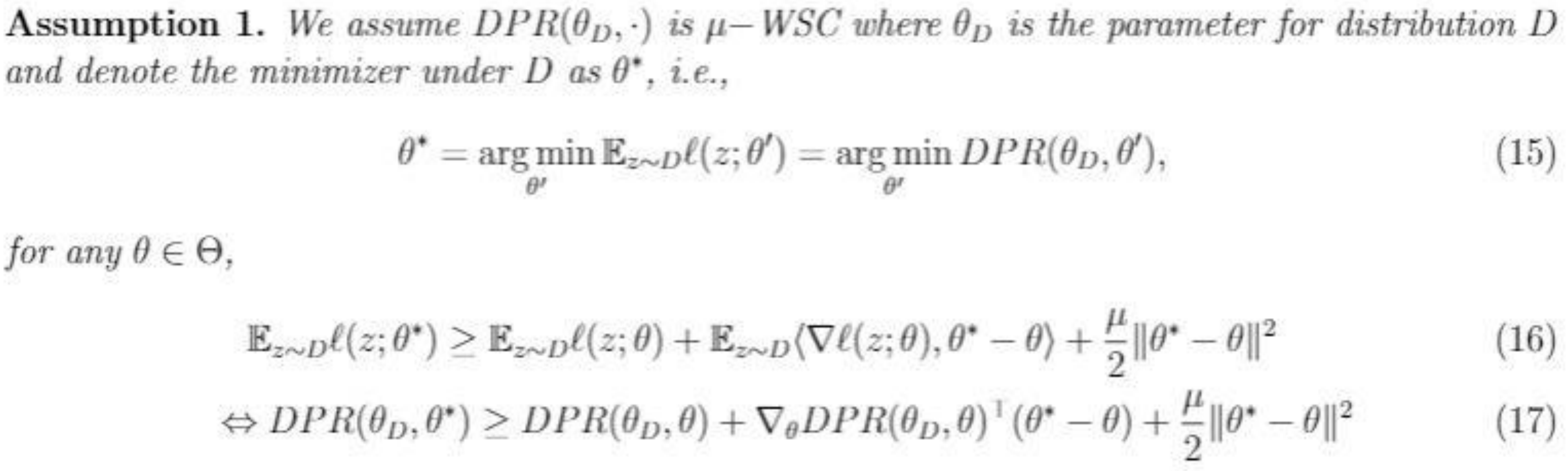
We aim to answer this question through two perspectives

1. Validate several conditions in **first-order optimization that could guarantee a linear convergence rate**. Once recognizing such conditions, there are plenty implementations in literature.
2. Translate the problem into the language of gradient flow in terms of distributions.

On the Theory of Optimizing Performative Risk -- Results

- Define DPR(θ_1, θ_2) = $\mathrm{E}_{z\sim D(\theta_1)} l(z; \theta_2)$ for *decoupled performative risk.*
- We show: when DPR is WSC (weakly strong convex), PR is WC (weakly convex) to θ_{po}, namely

$$PR\left(\theta_{po}\right) \geq PR(\theta) + \langle \nabla PR(\theta), \theta_{po} - \theta \rangle$$



Assumption 1. *We assume $DPR(\theta_D, \cdot)$ is $\mu-WSC$ where θ_D is the parameter for distribution D and denote the minimizer under D as θ^*, i.e.,*

$$\theta^* = \arg\min_{\theta'} \mathbb{E}_{z\sim D}\ell(z;\theta') = \arg\min_{\theta'} DPR(\theta_D, \theta'), \tag{15}$$

for any $\theta \in \Theta$,

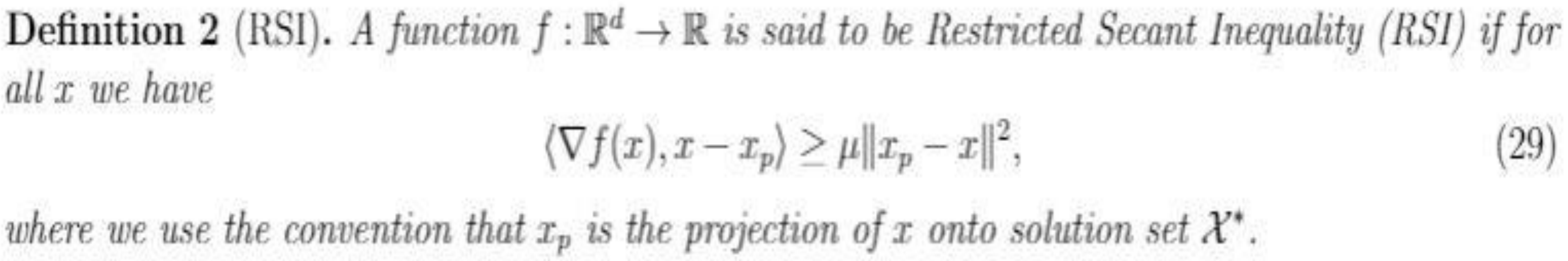
$$\mathbb{E}_{z\sim D}\ell(z;\theta^*) \geq \mathbb{E}_{z\sim D}\ell(z;\theta) + \mathbb{E}_{z\sim D}\langle \nabla\ell(z;\theta), \theta^* - \theta\rangle + \frac{\mu}{2}\|\theta^* - \theta\|^2 \tag{16}$$

$$\Leftrightarrow DPR(\theta_D, \theta^*) \geq DPR(\theta_D, \theta) + \nabla_\theta DPR(\theta_D, \theta)^\top(\theta^* - \theta) + \frac{\mu}{2}\|\theta^* - \theta\|^2 \tag{17}$$

On the Theory of Optimizing Performative Risk -- Results

- We show: when DPR is RSI (Restricted Secant Inequality), PR is RSI, namely

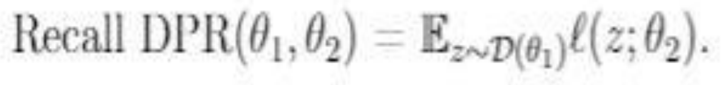
$$\langle \nabla PR(\theta), \theta - \ \theta_{po} \rangle \ \geq \mu' \left| \theta_{po} - \theta \right|^2$$



Definition 2 (RSI). *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be Restricted Secant Inequality (RSI) if for all x we have*

$$\langle \nabla f(x), x - x_p \rangle \geq \mu \| x_p - x \|^2, \tag{29}$$

where we use the convention that x_p is the projection of x onto solution set $\mathcal{X}^*$.



Recall $\mathrm{DPR}(\theta_1, \theta_2) = \mathbb{E}_{z \sim \mathcal{D}(\theta_1)} \ell(z; \theta_2)$.

On the Theory of Optimizing Performative Risk -- Results

- We also analyze this problem using gradient flow in a Wasserstein-2 space

$$\theta_{k+1} \in argmin_{\theta}\ PR(\theta) + \frac{W_2^2\big(D(\theta), D(\theta_k)\big)}{2\tau}$$

- However, updating functional in a probability distribution space is not realisitic in practice…

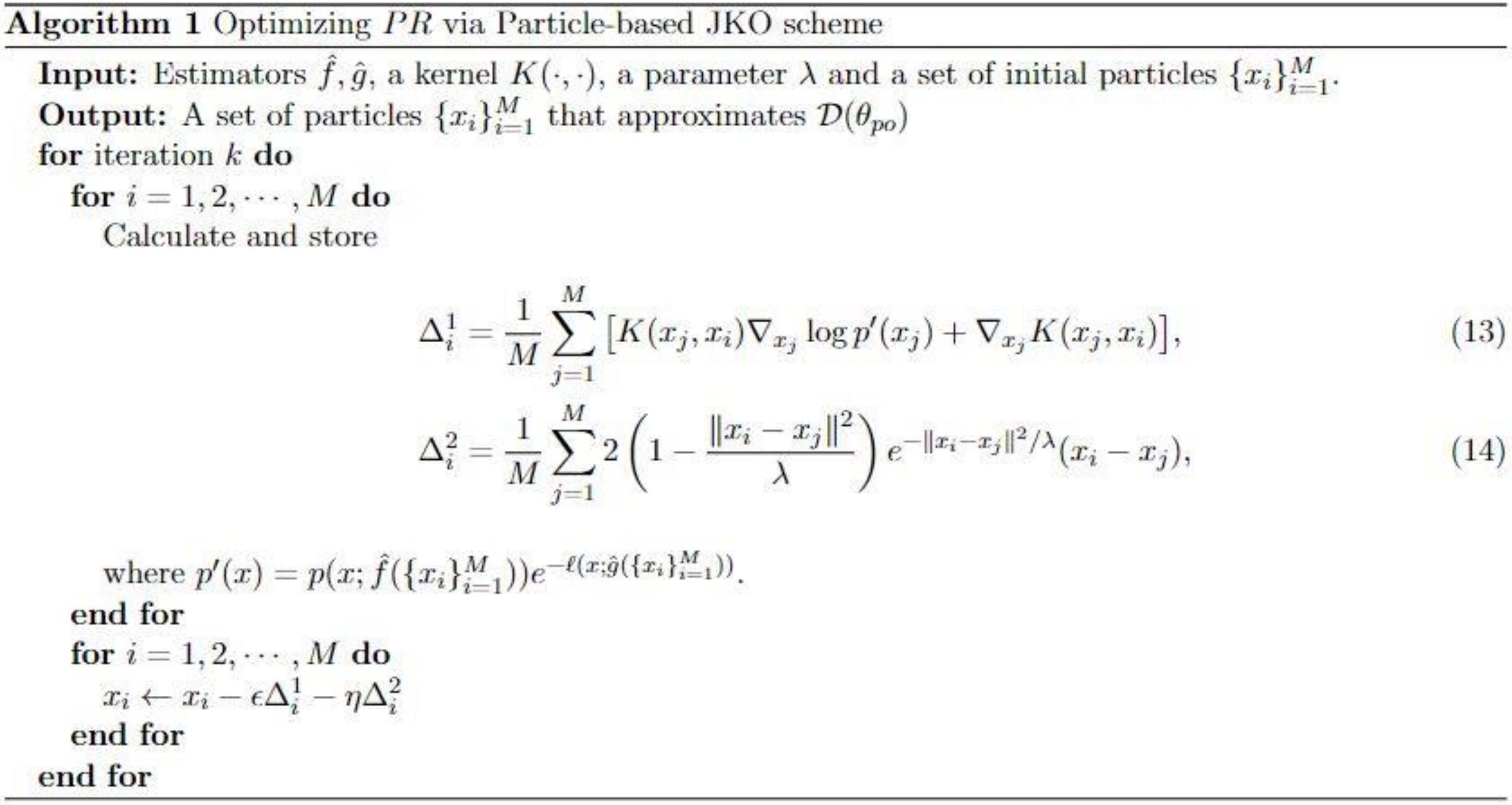
To deal with challenge, we need:

✓ Discretization of the flow: **particle-based approximations**

- ■ *Roughly speaking, we can iterate by only using M samples to approximate distributions.*

$$\mu_k^{(h)} \approx \frac{1}{M} \sum_{i=1}^{M} \delta(x_k^{(i)}).$$

On the Theory of Optimizing Performative Risk -- Results



Algorithm 1 Optimizing PR via Particle-based JKO scheme

```
Input: Estimators $\hat{f}, \hat{g}$, a kernel $K(\cdot,\cdot)$, a parameter $\lambda$ and a set of initial particles $\{x_i\}_{i=1}^M$.
Output: A set of particles $\{x_i\}_{i=1}^M$ that approximates $\mathcal{D}(\theta_{po})$
for iteration $k$ do
  for $i = 1, 2, \cdots, M$ do
    Calculate and store
        $\Delta_i^1 = \frac{1}{M}\sum_{j=1}^{M}\left[K(x_j, x_i)\nabla_{x_j}\log p'(x_j) + \nabla_{x_j}K(x_j, x_i)\right],$    (13)
        $\Delta_i^2 = \frac{1}{M}\sum_{j=1}^{M} 2\left(1 - \frac{\|x_i - x_j\|^2}{\lambda}\right)e^{-\|x_i - x_j\|^2/\lambda}(x_i - x_j),$    (14)
    where $p'(x) = p(x; \hat{f}(\{x_i\}_{i=1}^M))e^{-\ell(x;\hat{g}(\{x_i\}_{i=1}^M))}$.
  end for
  for $i = 1, 2, \cdots, M$ do
    $x_i \leftarrow x_i - \epsilon\Delta_i^1 - \eta\Delta_i^2$
  end for
end for
```

On the Theory of Optimizing Performative Risk -- Contributions

I. Studied several first-order conditions (Weak Strong Convexity, Restricted Secant Inequality, Polyak-Lojasiewicz inequality…) for performative prediction and what structural assumptions are needed.

II. Analyzed PR using gradient flow in Wasserstein space. We believe understanding performativity in terms of distributions could be beneficial.

III. Also developed a practical particle-based algorithm.

Questions

- What’s your expectation of a good Ph.D. student?
- What are some projects/directions you would like me to explore if I join the group?
- What is the research atmosphere like in MIT?

Thank you for your time
and consideration!