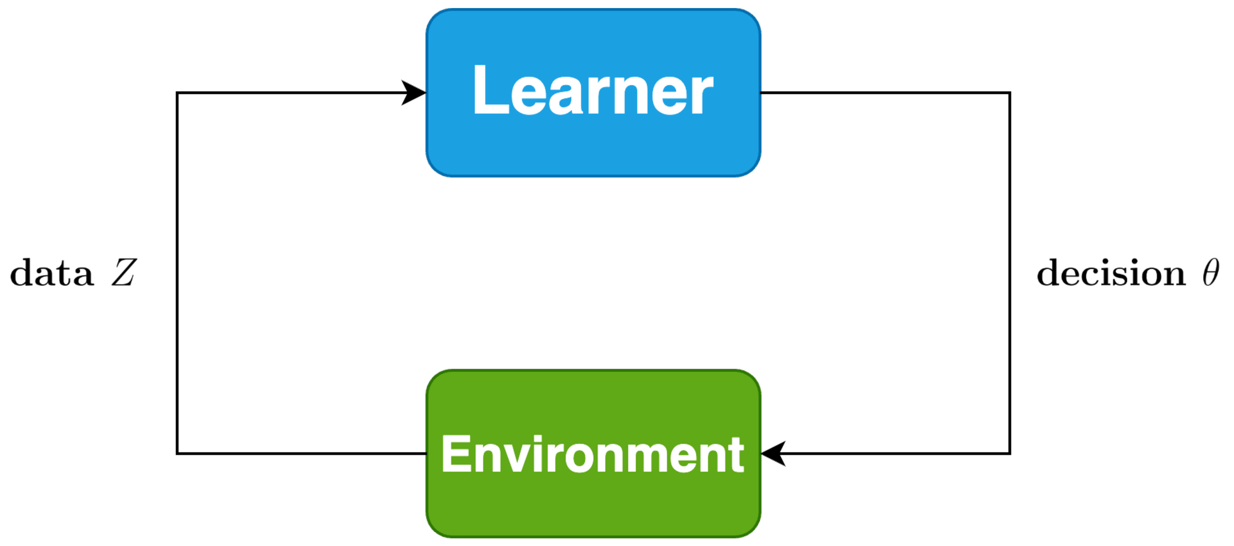
Two-timescale Derivative Free Optimization for Performative Prediction with Markovian Data

Haitong Liu, Dept. of CS, ETH Zurich, Qiang Li, Hoi-To Wai Dept. of SEEM, CUHK

Performative Prediction

- **Performative Prediction**: data distribution depends on decision variables.
- **Motivating Examples**: loan classification, pricing, ride sharing.



- **Goal**: minimize the performative risk,

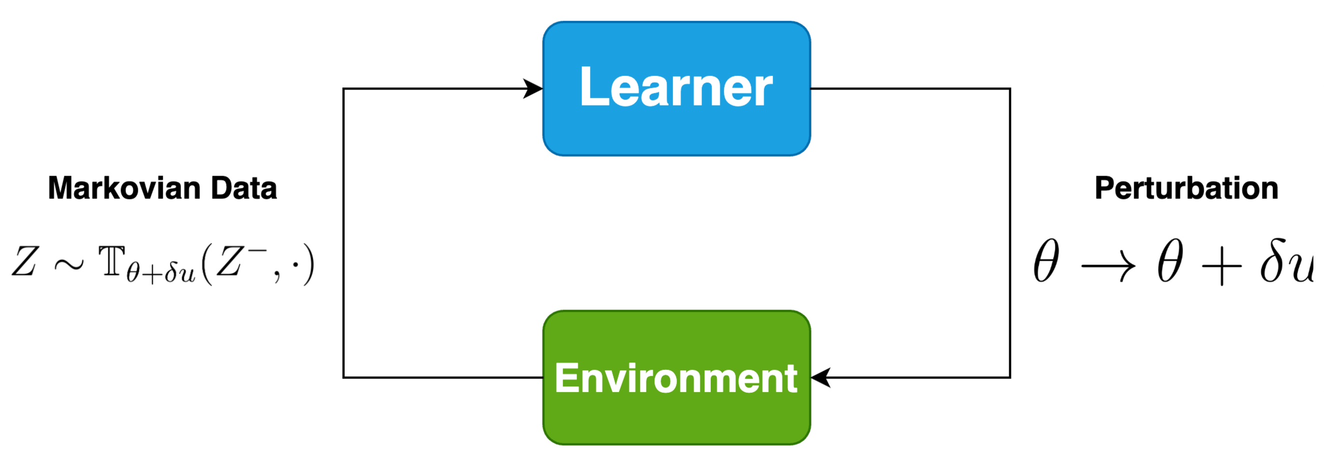
$$\min_{\theta} \mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{Z\sim\Pi(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta};Z)] \rightarrow \text{ncvx}$$

✻ Evaluate $\nabla\mathcal{L}(\boldsymbol{\theta})$ needs known $\Pi_{\boldsymbol{\theta}}(\cdot)$: $\mathbb{E}_{Z\sim\Pi_{\boldsymbol{\theta}}}[\nabla\ell(\boldsymbol{\theta};Z)+\ell(\boldsymbol{\theta};Z)\nabla\log\Pi_{\boldsymbol{\theta}}(Z)]$.

Zeroth Order Oracle & Markovian Data

- **One-point gradient estimator** $g_\delta(\cdot)$. Absence of prior knownledge on $\Pi_{\boldsymbol{\theta}} \rightarrow$ deploy and observe $\ell(\boldsymbol{\theta};\cdot)$ at perturbed decision points $\boldsymbol{\theta}+\delta u$ to estimate $\nabla\mathcal{L}$.
- **Markovian Sample**: $Z_t \sim \Pi_{\boldsymbol{\theta}}$ ✗.

✻ cannot draw samples directly from $\Pi_{\boldsymbol{\theta}} \rightarrow$ sample reweighting using the forgetting factor λ.



DFO(λ) Algorithm

- **Idea**: Construct zero-th order $\mathcal{O}(\delta)$-biased gradient estimator for $\mathcal{L}(\boldsymbol{\theta})$ as $\boldsymbol{g}_\delta$, to avoid evaluating a priori unknown $\Pi_{\boldsymbol{\theta}}(\cdot)$

$$g_\delta(\boldsymbol{\theta};\boldsymbol{u},Z) := \frac{d}{\delta}\ell(\check{\boldsymbol{\theta}};Z)\,\boldsymbol{u},\ \text{with } \check{\boldsymbol{\theta}} := \boldsymbol{\theta}+\delta\boldsymbol{u},\ Z\sim\Pi_{\check{\boldsymbol{\theta}}}(\cdot),\ \boldsymbol{u}\sim\text{Unif}(\mathbb{S}^{d-1})$$

- unbiased estimator for $\nabla\mathcal{L}_\delta(\boldsymbol{\theta})$, while $\mathcal{L}_\delta(\boldsymbol{\theta})$ is a smooth approx of $\mathcal{L}(\boldsymbol{\theta})$.

Two-timescale Derivative Free (DFO(λ)) Algorithm

Outer Loop ($k: 0\rightarrow T-1$) : Set stepsize δ_k and η_k, inner loop range τ_k

Inner loop ($m: 1\rightarrow\tau_k$) :

Deploy $\check{\boldsymbol{\theta}}_k^{(m)} = \boldsymbol{\theta}_k^{(m)}+\delta_k\boldsymbol{u}_k$, Sample $Z_k^{(m)}\sim\mathbb{T}_{\check{\boldsymbol{\theta}}_k^{(m)}}(Z_k^{(m-1)},\cdot)$

Update $\boldsymbol{\theta}_k^{(m+1)} = \boldsymbol{\theta}_k^{(m)} - \eta_k\lambda^{\tau_k-m}\boldsymbol{g}_{\delta_k}\left(\boldsymbol{\theta}_k^{(m)},\boldsymbol{u}_k,Z_k^{(m)}\right) \rightarrow$ forgetting factor

End inner loop : $Z_{k+1}\leftarrow Z_k^{(\tau_k)},\ \boldsymbol{\theta}_{k+1}\leftarrow\boldsymbol{\theta}_k^{(\tau_k+1)}$.

Output : $\boldsymbol{\theta}_s$, where $s\sim\text{Uniform}(\{0,1,\ldots,T\})$.

✿ **Highlights** : (i) two-timescales step sizes, (ii) make use of every sample, (iii) trade-off between sample accumulation and MC mixing time via λ.

Main Results

- A1. $\|\nabla\mathcal{L}(\boldsymbol{\theta})-\nabla\mathcal{L}(\boldsymbol{\theta}')\|\le L\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$ A2. (Bounded loss) $|\ell(\boldsymbol{\theta};z)|\le G$
- A3a. (Lipschitz distribution map) $\delta_{TV}(\boldsymbol{\theta},\boldsymbol{\theta}')\le L'\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$
- A3b. (L_1-sensitivity) $|\ell(\boldsymbol{\theta},z)-\ell(\boldsymbol{\theta},z')|\le L_0\|z-z'\|$, $W_1(\Pi_{\boldsymbol{\theta}},\Pi_{\boldsymbol{\theta}'})\le L_1\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$
- A4. (Geometric mixing) $\delta_{TV}(\mathbb{P}_{\boldsymbol{\theta}}(Z_k\in\cdot\mid Z_0=z),\Pi_{\boldsymbol{\theta}})\le M\rho^k$.
- A5. (Smooth Markov kernel) $\delta_{TV}(\mathbb{T}_{\boldsymbol{\theta}}(z,\cdot),\mathbb{T}_{\boldsymbol{\theta}'}(z,\cdot))\le L_2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$

Theorem 1: Using step sizes $\eta_k\propto k^{-2/3}$, $\delta_k\propto k^{-1/6}$, $\tau_k\propto\log k$, the iterate of DFO(λ) satisfies $\frac{1}{1+T}\sum_{k=0}^{T}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2\le\mathcal{O}\left(d^{2/3}/T^{1/3}\right)$

- **ϵ-stationary**: above metric achieves ϵ-target acc. after $\mathcal{O}\left(d^2/\epsilon^3\right)$ iter.
- **Sample complexity**: $\mathsf{S}_\epsilon=\mathcal{O}\left(d^2/\epsilon^3\right)\leftarrow$ worse than $\mathcal{O}(d/\epsilon^2)$.
- **Estimator (I)**: prior two point estimator $\boldsymbol{g}_{\text{2pt-I}}$ [Ghadimi & Lam, 2013] $\boldsymbol{g}_{\text{2pt-I}} := \frac{d}{\delta}[\ell(\boldsymbol{\theta}+\delta\boldsymbol{u};Z)-\ell(\boldsymbol{\theta};Z)]\,\boldsymbol{u} \rightarrow$ biased ✗ since $Z\sim\Pi_{\boldsymbol{\theta}+\delta\boldsymbol{u}}$, which is unique feature of decision-dependent sample distribution.
- **Estimator (II)**: $\boldsymbol{g}_{\text{2pt-II}} := \frac{d}{\delta}[\ell(\boldsymbol{\theta}+\delta\boldsymbol{u};Z_1)-\ell(\boldsymbol{\theta};Z_2)]\,\boldsymbol{u}\rightarrow$ unbiased Same variance $\mathbb{E}\|\boldsymbol{g}_{\text{2pt-II}}\|^2=\Omega(1/\delta^2)$, but higher sampling overhead ✗

Numerical Experiments

Markovian Regression —

- **Quadratic Loss**: $\ell(\boldsymbol{\theta};x,y)=(\langle x,\boldsymbol{\theta}\rangle-y)^2$
- **AR Model**: $(\tilde{X}_t,\tilde{Y}_t)=(1-\gamma)(\tilde{X}_{t-1},\tilde{Y}_{t-1})+\gamma(X_t,Y_t)$, where γ controls the mixing rate of the Markov chain.
- **Stationary samples**: (X_t,Y_t) is drawn according to

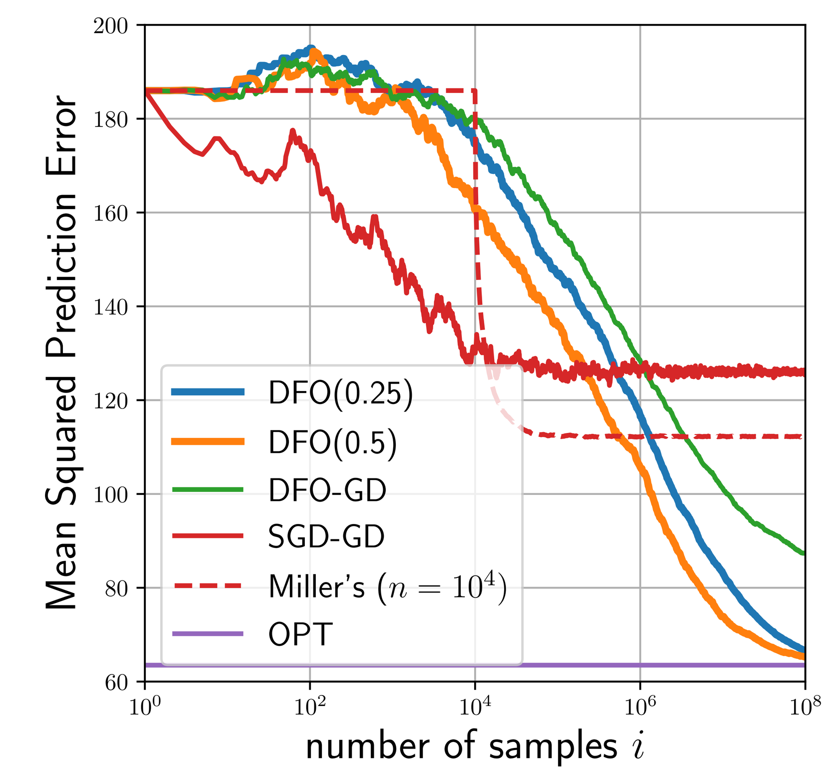
$$\begin{cases} X_t\sim\mathcal{N}(0,\frac{2-\gamma}{\gamma}\sigma_1^2 I),\\ Y_t|X_t\sim\mathcal{N}(\langle x_t+\kappa\boldsymbol{\theta}_{t-1},\boldsymbol{\theta}_0\rangle,\frac{2-\gamma}{\gamma}\sigma_2^2),\end{cases}$$

where $\gamma=0.25$, $\kappa=1/\|\boldsymbol{\theta}_0\|$, $\sigma_1=\sigma_2=1$.

- **Goal**: Comparison of 4 state-of-the-art algorithms:
  - SGD with greedy deployment from [Mendler et al., 2020],
  - Two-Phase algorithm from [Miller et al., 2021]:
    - (Phase I) Estimate distribution map $\Pi_{\boldsymbol{\theta}}$
    - (Phase II) Minimize finite-sample approx. of $\mathcal{L}(\boldsymbol{\theta})$,
  - DFO-GD (no burn-in phase),
  - Our DFO(λ) with $\lambda\in\{0.25,0.5\}$.

Observations

- ✗ DFO/SGD-GD fail to find a stationary solution to $\mathcal{L}(\boldsymbol{\theta})$.
- ✗ Two-Phase algorithm fail neither even with 10^4 (Markovian) samples gathered in the first phase.
- ✓Compared to above algorithms, DFO(λ) converges to a near-optimal solution after reasonable samples.



Reference

- Perdomo, Juan, et al. *Performative prediction*, ICML 2020.
- Mendler-Dünner, et al. *Stochastic optimization for performative prediction* NeurIPS 2020.
- Miller, et al. *Outside the echochamber: Optimizing the performative risk.* In ICML 2021.

Contact E-mail: liuhai@ethz.ch, {liqiang,htwai}@se.cuhk.edu.hk