Multi-agent Performative Prediction with Greedy deployment and Consensus Seeking Agents

Qiang Li, Chung-Yiu Yau, Hoi-To Wai

Dept of System Engineering and Engineering Management,
The Chinese University of Hong Kong

November 5, 2024

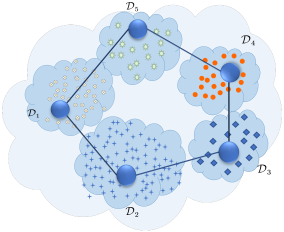
Overview

Decentralized Optimization

- Distributed optimization uses a set of networked computers, called **agents**, to solve optimization problems.
- Challenge: an algorithm running on one computer does not meet the expected performance.
- Approaches:
  - upgrade CPU, GPU, memory... ☹
  - use more computers, decompose the problem, run a decentralized optimization algorithm. More favorable (often the only)
- Examples: wireless sensor network, applications of real-time decisions made based on agents' local data.

Concrete Example - Clinical Data

- Each agent (hospital) wishes to learn about the treatment of a certain medical condition.
- But no previous experience (i.e., existing samples) in its local database.

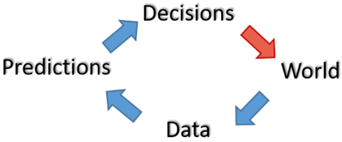


- Clinical data are privacy sensitive $\Longrightarrow$ shared directly ✗.
- Small amout of Data $\Longrightarrow$ consensus design [1]
- Candidate algorithm: Decentralized SGD (only requiring sharing the model among neighboring agents)
- More complex factors for local agents...

[1] also used in federated learning.

Local Performativity

- When predictions are used to support decisions, the distribution of future observations is altered.



- But *decision* (classifier) can cause distribution shift in the *world*.
- **Classical Supervised Learning**: *static world* with i.i.d. data.
- **Performative Prediction**: stochastic optimization problem whose data distribution depends on the decision variable.
- **Clinical Data Example**: After deploying a model, patients may overstates their symptoms to receive better treatment.

Performative Prediction for Single Agent

- **Data:** $z = (x, y) \sim \mathcal{D}(\theta)$.
- **Goal:** minimize *performative risk*

$$\min_{\theta} \ \mathcal{L}(\theta) := \mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)]$$

- Inspired by [Perdomo et al., 2020], use $\mathcal{D}(\theta)$ to capture distribution shift (agents' response) of z due to learner's state.
- Two different solutions to performative prediction:

$$\theta_{PO} \in \arg\min_{\theta\in\mathbb{R}^d} \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(\theta;z), \ \ \theta_{PS} = \arg\min_{\theta'\in\mathbb{R}^d} \mathbb{E}_{z\sim\mathcal{D}(\theta_{PS})}[\ell(\theta';z)].$$

- *Agnostic Setting:* No extra knowledge on local data, like distribution... $\Longrightarrow \theta_{PS}$ is the best to hope for.

How should the agent (local hospital) do?

- SGD/GD on $\ell(z;\theta)$ with $z \sim \mathcal{D}(\theta)$ [2]

---

[2] [Perdomo et al., 2020, Mendler-Dünner et al., 2020].

Finding θ_{PS}

- [Mendler-Dünner et al., 2020] considers an SGD-like recursion:

$$\begin{aligned} &\underline{\textit{Sampling}}: \quad z_{k+1} \sim \mathcal{D}(\theta_k) \\ &\underline{\textit{Update}}: \quad \theta_{k+1} = \theta_k - \gamma_{k+1} \nabla \ell(\theta_k; z_{k+1}), \end{aligned}$$

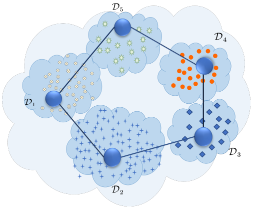
  i.e., a *greedy deployment* scheme. Assume that[3]:
  - A1: $\ell(\boldsymbol{\theta}; z)$ is μ-strongly convex.
  - A2: $\nabla \ell(\boldsymbol{\theta}; z)$ has L-Lipschitz gradient.
  - A3: $\epsilon-$sensitivity: $W_1(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq \epsilon \|\theta - \theta'\|, \forall\ \theta, \theta'$
- Convergence Region: $\epsilon < \mu/L$.
- *Issue* in multi-agent case: sensitive agent $\epsilon_i \geq \mu/L$.

When will the problem admit a stable and consensual solution? If so, how fast does it take for to converge to such solution?

[3] A1-A3 are mild - also in [Mendler-Dünner et al., 2020].

Multi-agent Performative Prediction (Multi-PfD)

1. n agents case: undirected and connected graph $G = (V, E)$.
2. Mixing matrix $\boldsymbol{W} \in \mathbb{R}_{+}^{n \times n}$ on G, doubly stochastic.
3. $\mathcal{D}_i(\boldsymbol{\theta}_i)$: Agent i draws samples from ith population of users.
4. Heterogeneous data: $\mathcal{D}_i(\boldsymbol{\theta}) \neq \mathcal{D}_j(\boldsymbol{\theta}')$, $i \neq j$, even if $\boldsymbol{\theta} = \boldsymbol{\theta}'$.



- **Goal**: find a *common decision vector* $\boldsymbol{\theta} \in \mathbb{R}^d$ in a collaborative fashion that minimizes the average of local losses.

$$\begin{aligned} \min_{\boldsymbol{\theta}_i \in \mathbb{R}^d,\, i=1,\ldots,n} \quad & \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i)} \big[\ell(\boldsymbol{\theta}_i; Z_i) \big] \\ \text{s.t.} \quad & \boldsymbol{\theta}_i = \boldsymbol{\theta}_j,\ \forall\ (i,j) \in E. \end{aligned} \tag{1}$$

- Define Multi-PS solution:

$$\boldsymbol{\theta}^{PS} = \mathcal{M}(\boldsymbol{\theta}^{PS}) := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z_i \sim \mathcal{D}_i(\boldsymbol{\theta}^{PS})} [\ell(\boldsymbol{\theta}; Z_i)]$$

DSGD-GD for Pref. Pred.

Decentralized SGD-Greedy Deployment (DSGD-GD)

(Phase 1) $Z_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t)$

(Phase 2) $\boldsymbol{\theta}_i^{t+1} = \sum_{j=1}^n W_{ij} \boldsymbol{\theta}_j^t - \gamma_{t+1} \nabla \ell(\boldsymbol{\theta}_i^t; Z_i^{t+1})$,

- $\nabla \ell(\boldsymbol{\theta}_i^t; Z_i^{t+1})$: the gradient taken w.r.t. $\boldsymbol{\theta}_i^t$, and the samples Z_i^{t+1} at each agent are iid.
- Extension of Greedy Deployment scheme over decentralized scenario.
- Contributions:
  - Provide sufficient and necessary condition for the existence and uniqueness of the Multi-PS solution.
  - Prove DSGD-GD converges to the Multi-PS solution ($\mathcal{O}(1/t)$).
  - Conduct numerical experiments on synthetic/real data.

Assumptions

A4. Doubly stochastic mixing matrix $\boldsymbol{W}$

Exist a constant $\rho \in (0, 1]$ such that $\left\|\boldsymbol{W} - (1/n)\mathbf{1}\mathbf{1}^\top\right\|_2 \leq 1 - \rho$.

A5. σ-perturbation with sampled gradient

$$\mathbb{E}_{Z_i \sim \mathcal{D}_i(\boldsymbol{\theta})}[\|\nabla \ell(\boldsymbol{\theta}; Z_i) - \nabla f_i(\boldsymbol{\theta}; \boldsymbol{\theta})\|^2] \leq \sigma^2(1 + \left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{PS}\right\|^2).$$

A6. Heterogeneity ς

$$\|\nabla f(\boldsymbol{\theta}; \boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta}; \boldsymbol{\theta})\|^2 \leq \varsigma^2(1 + \left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{PS}\right\|^2), \ \forall \ \boldsymbol{\theta} \in \mathbb{R}^d.$$

- A6 also implies $\max_{i=1,\ldots,n} \|\nabla f_i(\boldsymbol{\theta}^{PS}; \boldsymbol{\theta}^{PS})\|^2 \leq \varsigma^2$.

Main Result - Existence and Uniqueness

Define the map $\mathcal{M} : \mathbb{R}^d \to \mathbb{R}^d$

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}' \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{z_i \sim \mathcal{D}_i(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}'; z_i)] \qquad (2)$$

Proposition 1 Existence and Uniqueness of $\boldsymbol{\theta}^{PS}$

Under A1-A3,

- If $\epsilon_{\mathsf{avg}} < \mu/L$, then the map $\mathcal{M}(\boldsymbol{\theta})$ is a contraction with the unique fixed point $\boldsymbol{\theta}^{PS} = \mathcal{M}(\boldsymbol{\theta}^{PS})$.
- If $\epsilon_{\mathsf{avg}} \geq \mu/L$, then there exists an instance of (2) where $\lim_{T\to\infty} \left\|\mathcal{M}^T(\boldsymbol{\theta})\right\| = \infty$.

- Single agent case: $\epsilon < \mu/L$ vs Mult. agent case: $\epsilon_{avg} < \mu/L$.
- DSGD-GD converges even if ϵ_i exceed μ/L as long as $\epsilon_{avg} < \mu/L$.
- Benefit of consensus: improved robustness to node failure and local distribution shifts.

Main Result - Convergence of DSGD-GD

Theorem 1 [Li et al., 2022]

Under A1-A6. Let $\epsilon_{\text{avg}} < \frac{\mu}{(1+\delta)L}$ and with non-increasing and sufficient small step sizes, for any $k \geq 1$, there exists $\mathbb{C}$ where it holds

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|^2] \lesssim \underbrace{\prod_{i=1}^{t}\left(1 - \frac{\widetilde{\mu}\gamma_i}{2}\right) + \frac{L(\sigma^2+\varsigma^2)}{n\delta\widetilde{\mu}\rho^2\epsilon_{\text{avg}}}\gamma_t^2}_{\text{Transient}} + \underbrace{\frac{\sigma^2}{n\widetilde{\mu}}\gamma_t}_{\text{Fluctuation}},$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|^2\right] \lesssim \left(1 - \frac{\rho}{2}\right)^t + \frac{(\sigma^2+\varsigma^2)}{\rho^2}\gamma_t^2,$$

where δ is a parameter to be determined, $\widetilde{\mu} := \mu - (1+\delta)\epsilon_{\text{avg}}L$.

- Convergence needs $\epsilon_{avg} < \mu/L$ if $\delta = 0$.
- Consensus error: $\|\boldsymbol{\Theta}_{\text{o}}^t\|_F^2 := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|^2\right] \sim \mathcal{O}(\gamma_t^2)$
- Take $\gamma_t = \frac{a_0}{a_1+t}$ for some $a_0, a_1 > 0$, $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|^2] \to 0$ as $\mathcal{O}(1/t)$, while the consensus error $\to 0$ as $\mathcal{O}(1/t^2)$.
- Fluctuation term that only depends on the averaged noise variance $\mathcal{O}(\sigma^2/n)$. Decays at rate of $\mathcal{O}(\gamma_t)$.

Other Contributions

- B-connected graph: Extend our analysis of DSGD-GD on time-varying graph.
  - $G^{(t)} = (V, E^{(t)})$ be a simple, undirected graph, but possibly not connected. Weighted adjacency matrix $\boldsymbol{W}^{(t)}$.
  - Time-varying graph sequence $\{G^{(t)}\}_{t\geq 1} = \{(V, E^{(t)})\}_{t\geq 1}$ is B-connected.
  - Exists B such that undirected graph $(V, E^{(t)} \cup \cdots E^{(t+B-1)})$ is connected.

A4′-Time-varying doubly stochastic mixing matrix

For any $t \geq 1$, the mixing matrix $\boldsymbol{W}^{(t)} \in \mathbb{R}^{n\times n}$ satisfies:

1. (Topology) $\boldsymbol{W}_{ij}^{(t)} = 0$ if $(i, j) \notin E^{(t)}$.
2. (Doubly stochastic) $\boldsymbol{W}^{(t)}\mathbf{1} = (\boldsymbol{W}^{(t)})^\top \mathbf{1} = \mathbf{1}$.
3. (Fast mixing) Let $\boldsymbol{A}^{(t)} := \boldsymbol{W}^{(t)} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, there exists $\bar{\rho} \in (0, 1]$ such that $\|\boldsymbol{A}^{(t+B-1)} \cdots \boldsymbol{A}^{(t)}\|_2 \leq 1 - \bar{\rho}$.

- Extend our analysis to the scenario when the local distributions $\mathcal{D}_i(\cdot)$ are simultaneously influenced by other agents in the network.

Simulation-Synthetic Data

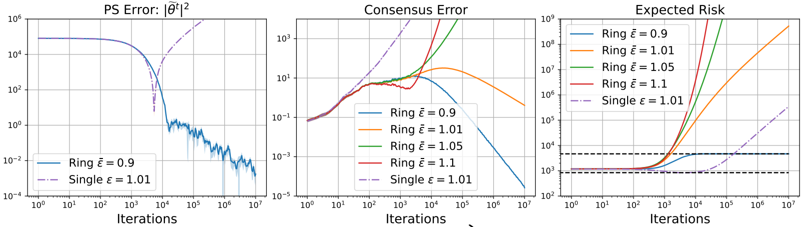
Multi-agent Gaussian Mean Estimation:
Consider $n = 25$-agent right graph and a quadratic loss

$$\ell(\boldsymbol{\theta}_i; Z_i) = (\boldsymbol{\theta}_i - Z_i)^2/2$$

Set the local distributions as $\mathcal{D}_i(\boldsymbol{\theta}_i) \equiv \mathcal{N}(\bar{z}_i + \epsilon_i \boldsymbol{\theta}_i, \sigma^2)$, where $\bar{z}_i$ is the mean value to be estimated.

- Parameters: $\mu = 1$, $L = 1$, $\gamma_t = \frac{a_0}{(a_1+t)}$.
- Multi-PS sol. $\boldsymbol{\theta}^{PS} = \sum_{i=1}^{n} \bar{z}_i / [n(1 - \epsilon_{\textsf{avg}})]$, if $0 < \bar{\epsilon} = \epsilon_{\textsf{avg}} < 1$.
- While $\boldsymbol{\theta}^{PS}$ does not exist if $\epsilon_{\textsf{avg}} \geq 1$.

Simulation-Synthetic Data (Cont'd)



$$\left.\begin{array}{r}\text{(left) when } \epsilon_{\text{avg}} < 1 \text{ converge,} \\ \text{(right) when } \epsilon_{\text{avg}} > 1, \text{ diverge.}\end{array}\right\} \implies \text{Prop. 1} ✓$$

$$\left.\begin{array}{r}\text{(left) } |\tilde{\boldsymbol{\theta}}^t|^2 \text{ decays at } \mathcal{O}(1/t), \\ \text{(middle) } \|\boldsymbol{\Theta}_{\circ}^t\|^2 \text{ decays at } \mathcal{O}(1/t^2)\end{array}\right\} \implies \text{Thm 1} ✓$$

- (dash-dotted) when $\epsilon_i = 1.01 > 1$, agent i disconnected and perform greedy deployment *individually*, its performative risk $f_i(\boldsymbol{\theta}_i^t; \boldsymbol{\theta}_i^t)$ diverges as $t \to \infty$.
- With consensus, performative risk of whole system $n^{-1} \sum_{i=1}^{n} f_i(\boldsymbol{\theta}_i^t; \boldsymbol{\theta}_i^t)$ can be stable.

Example-Binary Classification Problem

Recall that

$$\min_{\boldsymbol{\theta}_i \in \mathbb{R}^d,\, i=1,\ldots,n} \ \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i)}\big[\ell(\boldsymbol{\theta}_i; Z_i)\big] \ \text{ s.t. } \ \boldsymbol{\theta}_i = \boldsymbol{\theta}_j,\ \forall\ (i,j) \in E.$$

Take the logistic regression function as loss function, i.e.,

$$\ell(\boldsymbol{\theta}; Z_i) = \log\big(1 + \exp(\langle \boldsymbol{X}_i \,|\, \boldsymbol{\theta}\rangle)\big) - Y\langle \boldsymbol{X}_i \,|\, \boldsymbol{\theta}\rangle + \tfrac{\beta}{2}\|\boldsymbol{\theta}\|^2,$$

where $\beta > 0$ is a regularization parameter and $Z_i = (\boldsymbol{X}_i, Y_i)$ is the given data tuple.

Linear utility function for $Z_i = (\boldsymbol{X}_i, Y_i) \sim \mathcal{D}_i(\boldsymbol{\theta}_i)$ is given by

$$\boldsymbol{X}_i = \arg\max_{\hat{\boldsymbol{X}} \in \mathbb{R}^d} \big\{\langle \boldsymbol{\theta}_i \,|\, \hat{\boldsymbol{X}}\rangle - \tfrac{1}{2\epsilon_i}\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|^2\big\},\ Y_i = Y \ \text{ with } \ (\boldsymbol{X}, Y) \sim \mathcal{D}_i^{\circ},$$

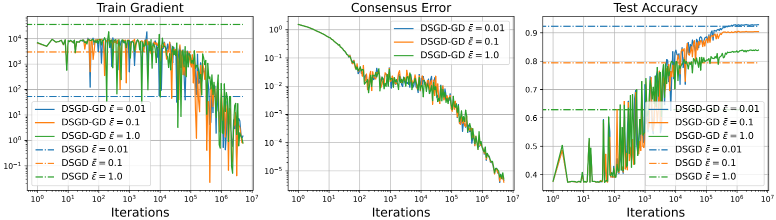
for some $\epsilon_i > 0$, where $\mathcal{D}_i^{\circ}$ is a base data distribution of the ith population.

- the closed form solution is $\boldsymbol{X}_i = \boldsymbol{X} + \epsilon_i \boldsymbol{\theta}_i$.

Simulation-Real Data

- Multi-agent spam classification task based on $\mathrm{spambase}$, a dataset [Hopkins, 1999]. Adopt Example 1 and simulate a scenario with 25 servers on a ring graph.
- Training: Test Data $= 3 : 1$. Each server has access to $1/25$ training data.
- Goal: find a common *spam filter classifier* via logistic loss.
- Strategic behavior of users: $\boldsymbol{X}_i$ are adapted to $\boldsymbol{\theta}_i$ through maximizing a linear utility function.
- Sensitivity parameters are set as $\epsilon_i \in \{0.4\epsilon_{\mathsf{avg}}, 0.45\epsilon_{\mathsf{avg}}, \ldots, 1.6\epsilon_{\mathsf{avg}}\}$ with $\bar{\epsilon} = \epsilon_{\mathsf{avg}} \in \{0.01, 0.1, 1\}$.

Simulation-Real Data (Cont'd)



†DSGD (dashed lines): non-performative opt. sol. on the shifted dataset.

- $\nabla f(\boldsymbol{\theta}^{PS}; \boldsymbol{\theta}^{PS}) = \mathbf{0}$, thus gradient norm measures the gap to $\boldsymbol{\theta}^{PS}$.
- (left) and (middle), DSGD-GD converges to the Multi-PS solution and reaches consensus at the rates $\mathcal{O}(1/t), \mathcal{O}(1/t^2)$, respectively.
- (right) Accur. $\downarrow$ as $\epsilon_{\text{avg}} \uparrow$, DSGD-GD achieves better accuracy than DSGD.

Conclusions

- *Multi-agent performative prediction problem* framework & extend analysis in [Perdomo et al., 2020], [Mendler-Dünner et al., 2020].
- Show that the MSE between DSGD-GD iterates and performative stable solution $\boldsymbol{\theta}^{PS}$ converges at $\mathcal{O}(1/t)$.
- *Necessary and sufficient* condition on the sensitivity of decision dependent data distributions for the existence and uniqueness of the Multi-PS solution.
- Numerical experiments validate our analysis.

Future Works:

- Multi-agent system based on non-iid data?

References I

Hopkins, Mark, R. (1999).
Spambase.
UCI Machine Learning Repository.

Li, Q., Yau, C.-Y., and Wai, H.-T. (2022).
Multi-agent performative prediction with greedy deployment and consensus seeking agents.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. (2020).
Stochastic optimization for performative prediction.
Advances in Neural Information Processing Systems, 33:4929–4939.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020).
Performative prediction.
In *ICML*.