# On the Necessity of Collaboration for Online Model Selection with Decentralized Data

Junfan Li

School of Computer Science and Technology, Harbin Institute of Technology Shenzhen

Joint work with Zheshun Wu, Zenglin Xu*, Irwin King

lijunfan@hit.edu.cn

# Outline

# Problem Setting

## Online Model Selection with Decentralized Data (OMS-DecD)

**Protocol 1** OMS-DecD

1: **for** $t = 1, 2, \dots, T$ **do**
2: **for** $j = 1, \dots, M$ in parallel **do**
3: The adversary sends $\mathbf{x}_t^{(j)}$ to the $j$-th client
4: The learner selects a hypothesis space $\mathcal{F}_{I_t} \in \mathcal{F}$
5: The learner selects $f_t^{(j)} \in \mathcal{F}_{I_t}$ and outputs $f_t^{(j)}(\mathbf{x}_t^{(j)})$
6: The learner observes the true output $y_t^{(j)}$
7: **end for**
8: **end for**

The learner aims to design an algorithm without leaking the raw data and minimizing the regret,

$$\forall i \in [K], \quad \mathrm{Reg}_D(\mathcal{F}_i) = \sum_{t=1}^{T} \sum_{j=1}^{M} \ell\left(f_t^{(j)}(\mathbf{x}_t^{(j)}), y_t^{(j)}\right) - \min_{f \in \mathcal{F}_i} \sum_{t=1}^{T} \sum_{j=1}^{M} \ell\left(f(\mathbf{x}_t^{(j)}), y_t^{(j)}\right).$$

# Problem Setting

## A Trivial Approach

**Definition 1 (A non-cooperative algorithm)**

Let $\mathcal{A}_{OMS}$ be an algorithm for centralized online model selection. A non-cooperative algorithm for OMS-DecD is defined by independently running a copy of $\mathcal{A}_{OMS}$ on each client.

It is obvious that

- The regret bound is $O(M \cdot \mathrm{Reg}(\mathcal{A}_{OMS}))$.
- The non-cooperative algorithm will not leak the raw data.

There is a pessimistic result [1]:

- if $K = 1$, then the non-cooperative algorithm is optimal in full information. In other words, collaboration is unnecessary in full information setting.

# Problem Setting

## Question 1

*Whether collaboration is effective for OMS-DecD i.e.,* $K \geq 2$*. And if so, how?*

# Problem Setting

## Our results

| Algorithm | Regret | Computational constraint on clients |
|---|---|---|
| Any | $\Omega(M\sqrt{T\ln K})$ | No |
| non-cooperative | $\Omega(M\sqrt{TKJ^{-1}})$ | $O(J), 1 \leq J < K$ |
| FOMD-OMS | $\tilde{O}(M\sqrt{T\ln K} + \sqrt{MTKJ^{-1}})$ | $O(J), 2 \leq J < K$ |

Table 1: Summary of the main results.

We can conclude that

1. No computational constraints on clients.
   Collaboration is unnecessary for OMS-DecD.
2. The per-round time complexity on each client is limited to $o(K)$.
   Collaboration is necessary for OMS-DecD.
3. The collaboration in previous federated algorithms for distributed online multi-kernel learning [2, 3] is unnecessary.

# Outline

## Overview

Lower Bounds

- No computational constraint.
  Reducing OMS-DecD to prediction with expert advice. [4]
- The per-round time complexity on each client is limited to $o(K)$.
  Reducing OMS-DecD to prediction with limited advice. [5]

Algorithm

- A new federated online mirror descent framework, FOMD-No-LU.
- Decoupling model selection and prediction for efficient communications.

Theoretical analysis

- A new Bernstein's inequality for martingale
  high-probability regret bounds that adapt to the complexity of individual hypothesis space.

## Reducing to Prediction with Expert/Limited Advice

### Theorem 2 (Lower Bounds)

*Assuming that* $5 \leq K \leq \min\{d, T\}$. *For each* $i \in [K]$, *let* $\mathcal{F}_i = \{f_i(\mathbf{x}) = \mathbf{e}_i^\top \mathbf{x}\}$ *and* $\mathcal{D}_i = [\min_{\mathbf{x}\in\mathcal{X}} f_i(\mathbf{x}), \max_{\mathbf{x}\in\mathcal{X}} f_i(\mathbf{x})]$, *where* $\mathbf{e}_i$ *is the standard basis vector in* $\mathbb{R}^d$.
*Denote by* sup *the supremum over all examples.*
*(i) There are no computational constraints on clients. Let* $\ell(v, y) = |v - y|$. *The regret of any algorithm for OMS-DecD satisfies:*
$\lim_{T\to\infty} \sup \max_{i\in[K]} \mathrm{Reg}_D(\mathcal{F}_i) \geq 0.25M\sqrt{T \ln K}$;
*(ii) The per-round time complexity on each client is limited to* $O(J)$. *Let* $\ell(v, y) = 1 - v \cdot y$. *The regret of any, possibly randomized, noncooperative algorithm with outputs in* $\cup_{i\in[K]}\mathcal{D}_i$ *for OMS-DecD satisfies:*
$\sup \mathbb{E}[\max_{i\in[K]} \mathrm{Reg}_D(\mathcal{F}_i)] \geq 0.1M\sqrt{KTJ^{-1}}$, *where the expectation is taken over the randomization of algorithm.*

## A New Bernstein's Inequality for Martingale

### Lemma 3

*Let $X_1, \dots, X_n$ be a bounded martingale difference sequence w.r.t. the filtration $\mathcal{H} = (\mathcal{H}_k)_{1 \le k \le n}$ and with $|X_k| \le a$. Let $Z_t = \sum_{k=1}^{t} X_k$ be the associated martingale. Denote the sum of the conditional variances by $\Sigma_n^2 = \sum_{k=1}^{n} \mathbb{E}\left[X_k^2 | \mathcal{H}_{k-1}\right] \le v$, where $v \in [0, B]$ is a random variable and $B \ge 2$ is a constant. Then for any constant $a > 0$, with probability at least $1 - 2\lceil \log B \rceil \delta$,*

$$\max_{t=1,\dots,n} Z_t < \frac{2a}{3} \ln \frac{1}{\delta} + \sqrt{\frac{2}{B} \ln \frac{1}{\delta}} + 2\sqrt{v \ln \frac{1}{\delta}}.$$

The novelty is that the conditional variance $v$ is a random variable, not a constant.

# References

[1] Kumar Kshitij Patel, Lingxiao Wang, Aadirupa Saha, and Nathan Srebro. Federated online and bandit convex optimization. In ICML, pp. 27439–27460, 2023.

[2] Pouya M. Ghari and Yanning Shen. Personalized online federated learning with multiple kernels. In NeurIPS, pp. 33316–33329, 2022.

[3] Songnam Hong and Jeongmin Chae. Communication-efficient randomized algorithm for multi-kernel online federated learning. In IEEE TPAMI, pp. 9872–9886, 2022.

[4] Nicolò Cesa-Bianchi and Gábor Lugosi. Prediction, Learning, and Games. Cambridge University Press, New York, NY, 2006.

[5] Yevgeny Seldin, Peter L. Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In ICML, pp. 280–287, 2014.

Thank you!