Statistical Collusion by Collectives on Learning Platforms

Etienne Gauthier[1] **Francis Bach**[1] **Michael I. Jordan**[1,2]

[1] Inria, ENS, PSL University [2] UC Berkeley

ICML International Conference On Machine Learning

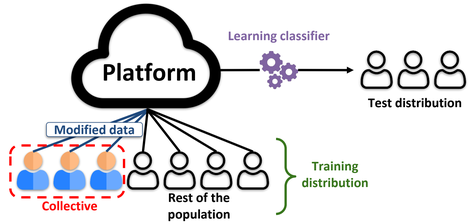
Framework Overview

Motivation

- Study how collectives can **influence learning platforms** by **strategically modifying their data** in a **coordinated** way.

Problem Setting

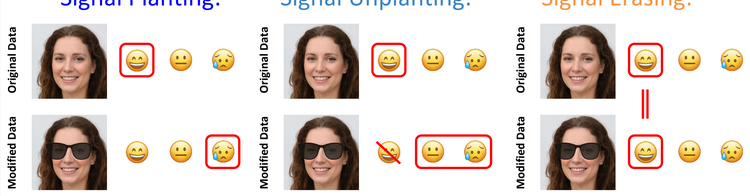
- Platform uses data from an i.i.d. population to train an algorithm.
- A subset (the **collective**) wants to steer the algorithm's behavior.
- Collective can **modify features/labels** via a shared strategy.



Challenges

- **Limited information**: The collective lacks access to platform internals and the rest of the population, requiring inference of key parameters and strategies from local data.
- **Goal**: Assess the collective's impact as a function of its size.

Three Different Objectives



Theoretical Analysis

Setup

- **Platform:** Trains a classifier f on a dataset of N consumers $D^{(n)} \uplus D^{(N-n)}$ initially drawn i.i.d. from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.
- **Collective:** The subset $D^{(n)}$ of $n < N$ consumers applies a shared strategy $h : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ to influence the platform, yielding a modified dataset $\tilde{D}^{(n)}$.
- **Data Model:**
  - Each data point: $(x, y) \in \mathcal{X} \times \mathcal{Y}$ (**finite universe**).
  - Collective creates a modified empirical distribution $\hat{\mathcal{P}}$ by applying h.

Assumptions

- The collective **knows the total number of users** N.
- It **does not know the data of non-collective users**.
- It **can pool its own data** to estimate distributions, parameters, and success of strategy h with concentration inequalities (e.g., Hoeffding).

Agent behaviors

- **Platform behavior:** selects $\hat{\mathcal{P}}$ that is Bayes-optimal for a distribution within total variation ε of $\hat{\mathcal{P}}$.
- **Collective goal:** influence test-time performance on $D_{\text{test}} \overset{i.i.d.}{\sim} \mathcal{D}$ by optimizing success metric $\hat{S}(n)$:

Objective	Signal planting	Signal unplanting	Signal erasing
Success $\hat{S}(n)$	$\hat{\mathbb{P}}_{x \sim D_{\text{test}}}(\hat{f}(g(x)) = y^*)$	$\hat{\mathbb{P}}_{x \sim D_{\text{test}}}(\hat{f}(g(x)) \neq y^*)$	$\hat{\mathbb{P}}_{x \sim D_{\text{test}}}(\hat{f}(g(x)) = \hat{f}(x))$

Results

- For each objective, we **analyze strategies** and **derive strategy-dependent high-probability lower bounds** on $\hat{S}(n)$.

Example: Signal Planting

Dataset

- **Synthetic tabular dataset:** Simulated vehicle data with features like *Model Type*, *Fuel Type*, and *Country of Manufacture*, labeled by evaluation (*Excellent*, *Good*, *Average*, *Poor*). Fixed transformation g.

Strategy

- **Natural strategy:** flood the platform with $h(x, y) = (g(x), y^*)$.

Theoretical Lower Bound

- We derive a high-probability lower bound on $\hat{S}(n)$, **fully computable by the collective**, which take the following form up to $1/\sqrt{n}$ error terms:

$$\hat{S}(n) \geq \hat{\mathbb{P}}_{\tilde{x} \sim \tilde{D}^{(n)}} [\text{Prevalence} - \text{Counteracting Influence} - \text{Robustness} > 0]$$

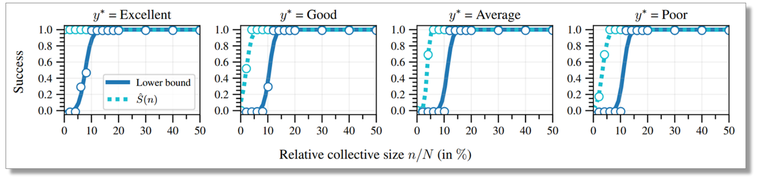
Indicates the prevalence of the modified feature in the modified dataset: the more frequently $\tilde{x}$ appears in the poisoned data, the greater the collective's ability to influence the associated label (proportional to n/N).

Captures how non-collective individuals hinder the collective's ability to plant the signal, reflecting how strongly the target label is tied to the features $\tilde{x}$; if other labels are far more likely than y^*, planting the signal becomes more difficult (scales with $1 - n/N$).

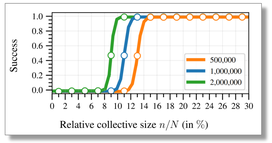
Platform robustness (increasing function of ε).

- **Interpretation:** As the collective size n/N grows, features $\tilde{x}$ are planted one by one, breaking in order of decreasing resistance.

Experimental Evaluation



Signal planting for different target labels. For example, the lower bound for *Poor* suggests 10% of agents are needed to plant the signal, but in practice only 5% suffice.



As N grows, collectives of the same proportion achieve better success bounds. **Larger platforms face higher risks from collective action.**

PAPER CODE