Paid with Models: Optimal Contract Design for Collaborative Machine Learning

Bingchen Wang[1*], Zhaoxuan Wu[1,2], Fusheng Liu[1], Bryan Kian Hsiang Low[3]

[1]Institute of Data Science, National University of Singapore, [2]Singapore-MIT Alliance for Research and Technology, [3]Department of Computer Science, National University of Singapore, *Corresponding Author (bingchen@nus.edu.sg)

NUS National University of Singapore

GLOW.AI

Paper:

Code:

Motivation

Training a model is no mean feat

Training a state-of-the-art model requires an enormous amount of data and compute.

Collaborative Machine Learning (CML)

Small parties can join their resources and train a good-performing model collectively.

The incentive problem of CML

Despite the promise, CML schemes may not work with the wrong incentives.

Conflict of interests

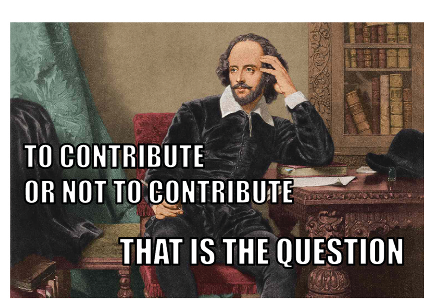
The common goal is to max model performance.

The private goal is to max net profit from joining CML.

More contribution means better model performance but also increased cost.

Private Information

Parties incur different contribution costs, which may only be privately observable.



Collaboration failure is a real concern.

Karimireddy, Guo and Jordan (2022)
Our paper (Appendix B.3)

Contract Design with Models as the Rewards

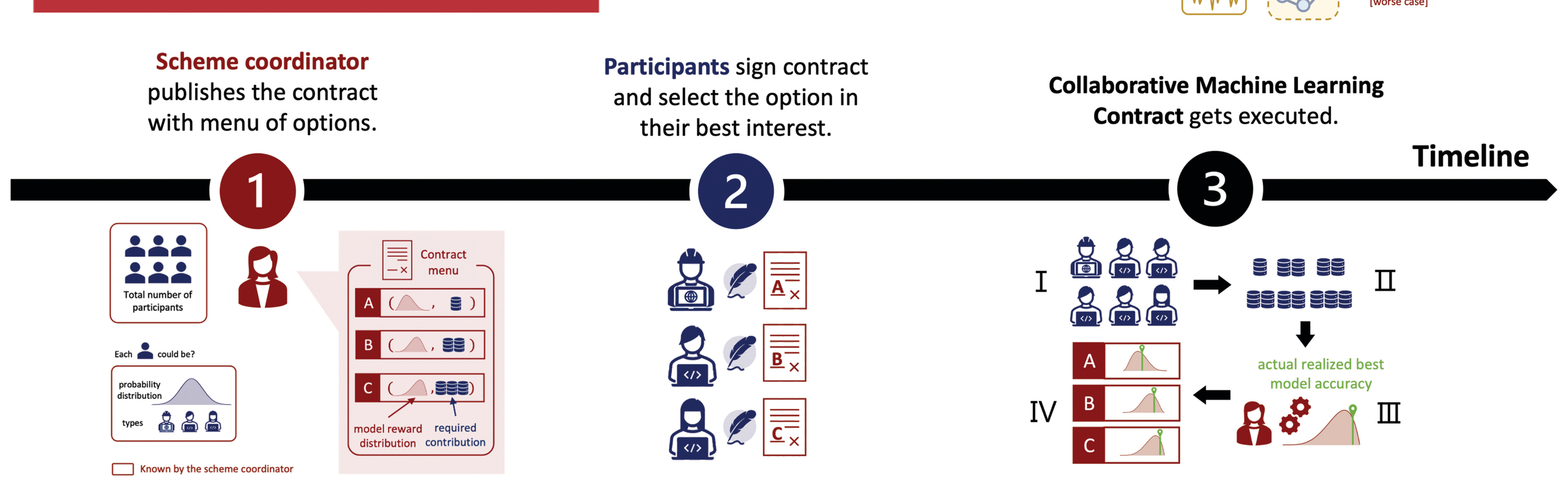
Optimal contract design for CML

We can design **contracts** to address the incentive problem, using models with different accuracy levels as the rewards.

To ensure the contract is amazing, we need to heed the unique features of model rewards:

(C1) Non-rivalrous. Model can be replicated free of charge.

(C2) Stochastic ex-ante. The exact accuracy of a model is unknown until the training completes.



Solving for the Optimal Contract

Notation

N	number of participants
I	number of possible cost types
n_i	number of type-i parties
m_i	contribution of type-i party
c_i	per-unit cost of type-i party
f_i	opportunity cost of type-i party
r_i	model reward for type-i party
$a(\cdot)$	accuracy function
$v(\cdot)$	valuation function

Information Assumption

The coordinator does not know the exact cost type of each party but knows the population distribution of cost types.

$$n \sim \text{Multinomial}(N, p)$$

Coordinator's Utility Function: $\mathbb{E}_{n\sim\text{Multi}(N,p)}\left[a\left(\sum_{i=1}^{I} n_i m_i\right)\right]$

Party's Utility Function: $\mathbb{E}_{n_i\geq 1}[v(r_i)] - c_i m_i$

Constrained optimization

$$\max_{(r_i,m_i)_{i=1}^{I}} \mathbb{E}_{n\sim\text{Multi}(N,p)}\left[a\left(\sum_{i=1}^{I} n_i m_i\right)\right]$$

$$\text{s.t.}\begin{cases} \mathbb{E}_{n_i\geq 1}[v(r_i)] - c_i m_i \geq f_i, \forall i \\ \mathbb{E}_{n_i\geq 1}[v(r_i)] - c_i m_i \geq \mathbb{E}_{n_j\geq 1}[v(r_j)] - c_i m_j, \forall i,j \\ \|r(n)\|_\infty \leq a\left(\sum_{i=1}^{I} n_i m_i\right), \forall n \in \text{Multi}(N,p) \end{cases}$$

Individual Rationality: Joining the CML should be better than opting out.

Incentive Compatibility: Telling the truth should be in the party's interest.

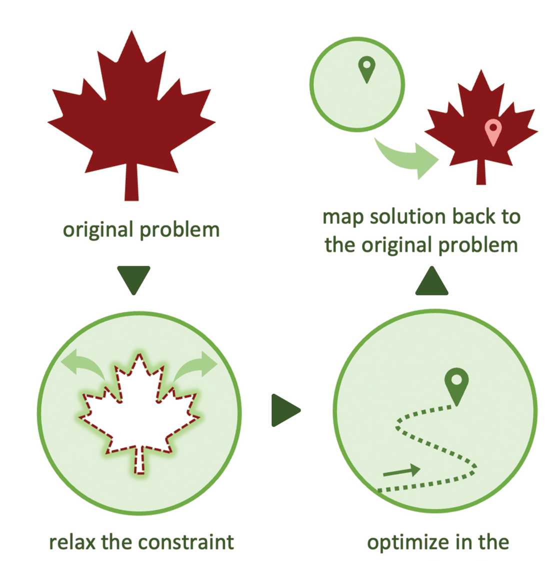
Budget Constraint: The coordinator cannot over-compensate parties.

The problem is hard to solve directly.

First-moment Problem

Zoom out to zoom in

The strategy to solve the above problem is to first solve an easier **convex problem** through constraint relaxation and then map its solution to that of the original problem.



Budget constraint transformation

The budget constraint in the original problem implies the budget constraint in first moments:

$$\|r(n)\|_\infty \leq a\left(\sum_{i=1}^{I} n_i m_i\right) \implies \mathbb{E}_{n_i\geq 1}[v(r_i)] \leq \mathbb{E}_{n_i\geq 1}\left[v\left(a\left(\sum_{i=1}^{I} n_i m_i\right)\right)\right]$$

First-moment problem

Let $t_i \triangleq \mathbb{E}_{n_i\geq 1}[v(r_i)]$. The first-moment problem is:

$$\max_{(t_i,m_i)_{i=1}^{I}} \mathbb{E}_{n\sim\text{Multi}(N,p)}\left[a\left(\sum_{i=1}^{I} n_i m_i\right)\right]$$

$$\text{s.t.}\begin{cases} t_i - c_i m_i \geq f_i, \forall i \\ t_i - c_i m_i \geq t_j - c_i m_j, \forall i,j \\ t_i \leq \mathbb{E}_{n_i\geq 1}\left[v\left(a\left(\sum_{i=1}^{I} n_i m_i\right)\right)\right], \forall i \end{cases}$$

☑ Fewer variables ☑ Fewer constraints ☑ Convex

Proportional assignment

Denote the first-moment solution as:

$$(t_i^*, m_i^*)_{i=1}^{I}.$$

Additionally, let

$$\bar{t}_i^* \triangleq \mathbb{E}_{n_i\geq 1}\left[v\left(a\left(\sum_{i=1}^{I} n_i m_i^*\right)\right)\right].$$

The following mapping solves the original problem:

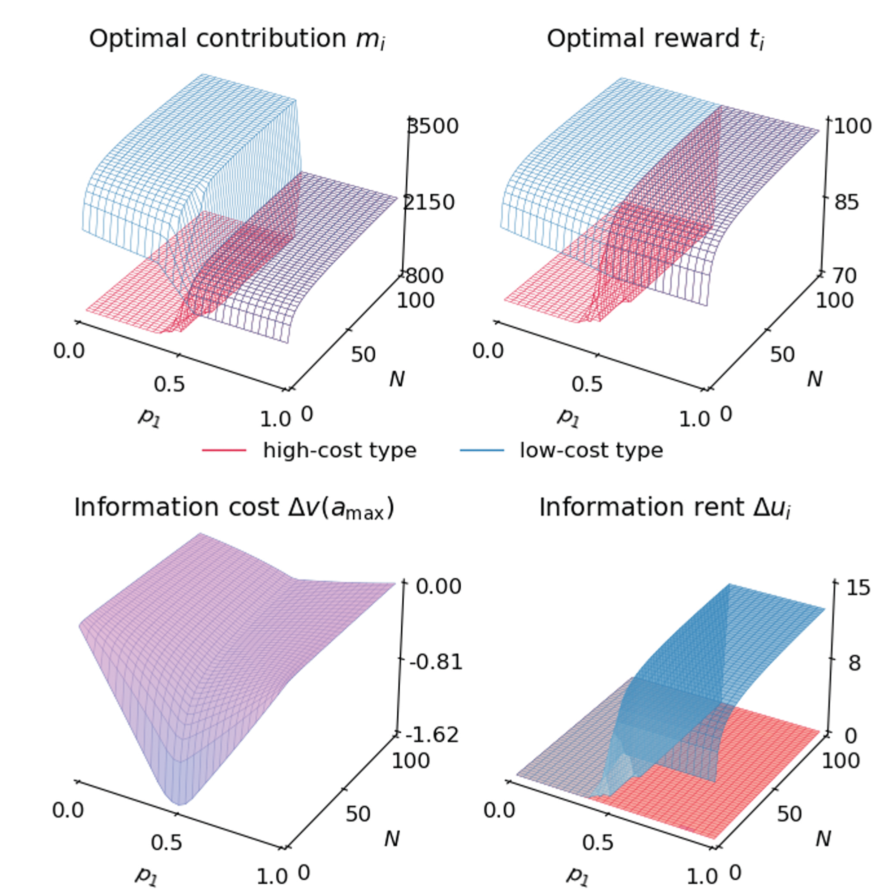
$$r_i^*(n) = v^{-1}\left(\frac{t_i^*}{\bar{t}_i^*} v\left(a\left(\sum_{i=1}^{I} n_i m_i^*\right)\right)\right)$$

Properties of Optimal Contracts

- Proportional Fairness. A party with lower cost contributes more and gets a better model.
- Weak Efficiency. The most cost-efficient party will be rewarded with the best model.
- Highest-cost Type Break Even. The least cost-efficient party (party who has the highest per-unit contribution cost) is indifferent between opting in and opting out.

Experiment Results

Two-type case



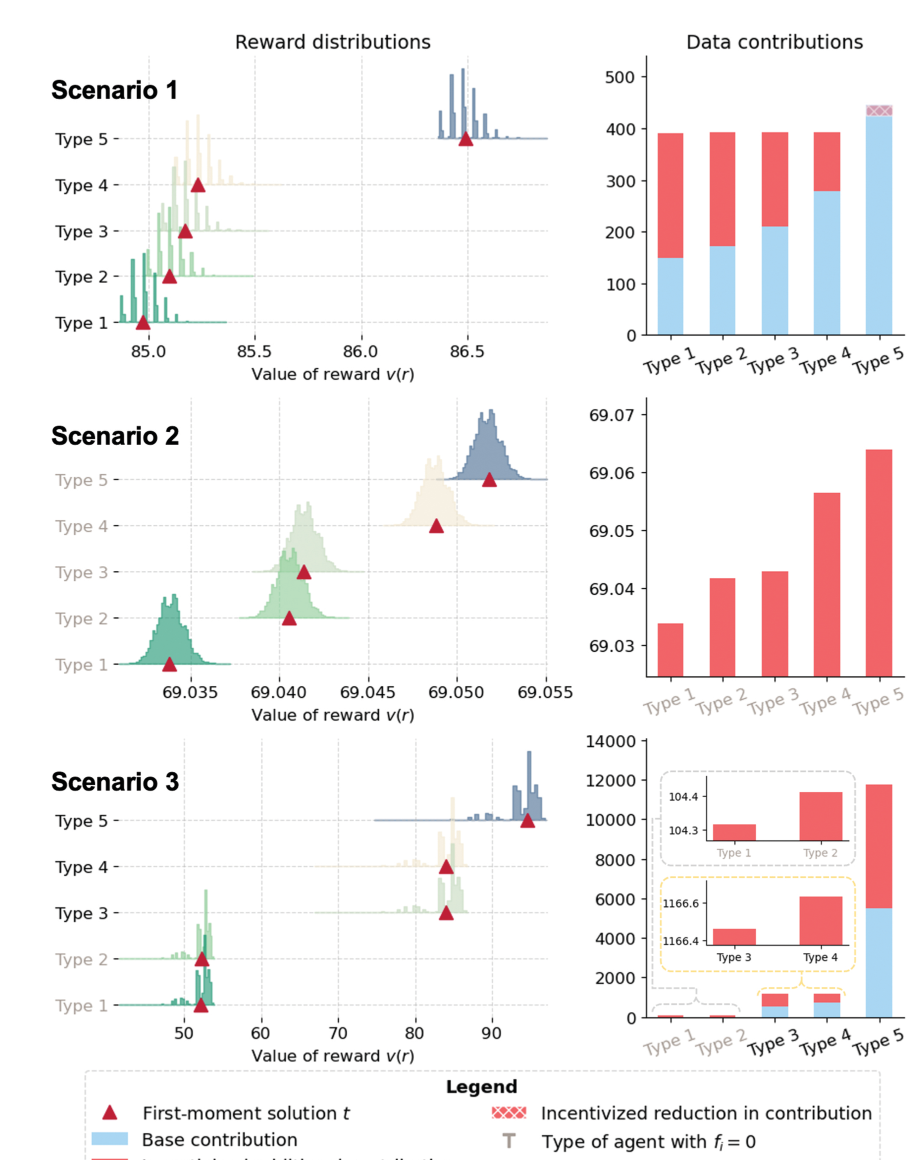
Optimal contract design depends crucially on the distribution of cost types and less so on the number of participants.

A pooling contract is more likely to be optimal when high-cost type is dominant in the population.

Low-cost type gains information rent when the coordinator cannot observe contribution costs.

Information rent is higher under pooling contracts.

Multi-type case



Scenario 1

All parties would be willing to train a model on their own.

A party may be incentivized to contribute less than their reservation level.

Scenario 2

All parties would not train a model if on their own.

Incentivized CML scheme can help small parties overcome the hurdle of model training.

Scenario 3

Some parties would train a model if on their own, and others not.

In the presence of dominant players, small parties can still gain from collaboration, demonstrating the trickle-down effect of the collaborative scheme.

Contract design is a viable tool for democratizing CML in an incentive-driven economy.

Acknowledgements

This research is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-018).