Multi-Distribution Learning

Simons Institute

Data-Driven Decision Processes Bootcamp

Nika Haghtalab

EECS, UC Berkeley

More Data … More Stakeholders

Learning guarantees that hold over agents, with individualized objectives, needs, and limitations.

Ownership: Data is spread across several sources.

Learning Guarantees: Solutions must be deployed across communities with different localities, populations, models, and resources.

Costly samples: Taking samples in physical domains is costly to individuals and data curators, e.g., medical tests, lead pipe testing, …

More Data … More Stakeholders … New challenges

1. Data is spread across several sources

2. Individualized and heterogenous learning objectives

3. Procurement of resources from multiple agents and sources

Learning Across Multiple Distributions

Enabling learning processes that benefit

multiple agents to learn from **collectively**

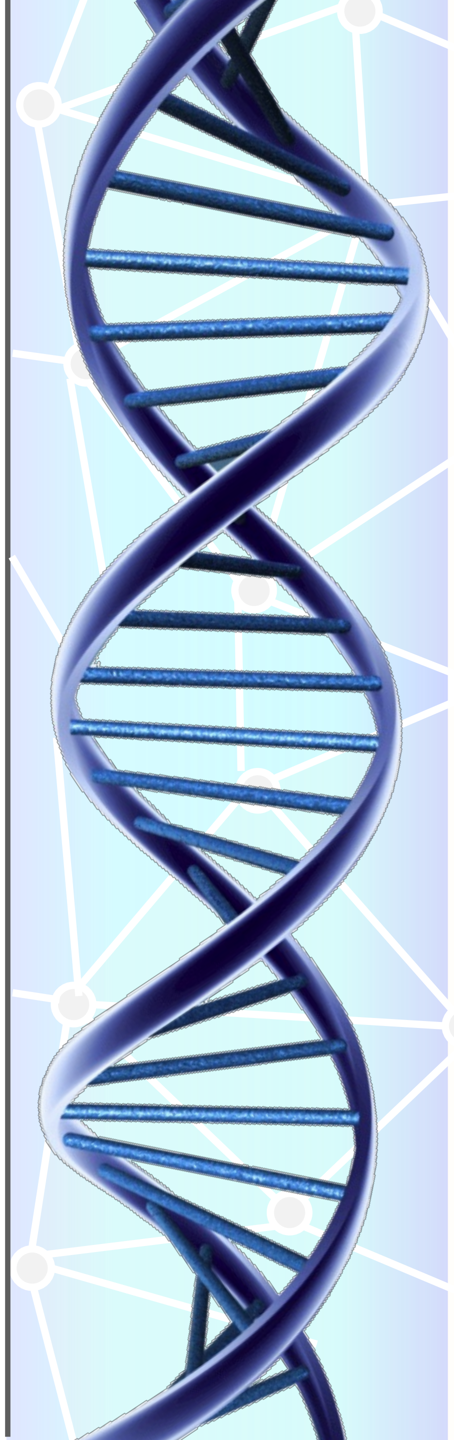
fewer resources.

Practical Applications: Data sharing and joint learning

- Starting to be used across network of devices, hospitals, etc.
- Behind recent major scientific discoveries, especially in genomic studies.

Theoretical Foundation:

- Multi-agent collaboration, welfare, and fairness,
- Fundamental to robust learning.



Large Scale Impact from **Mass Participation**

Recruit and Retain

Fundamental Questions We Need to Answer

Q1. How to measure learning performance across different tasks and distributions.

→What objective functions capture this performance?

→This tutorial: One unifying model (without too much detail).

Q2. How much resources are needed for learning and meeting these objectives?

→ Sample complexity and computational complexity.

→ Relying on decades of efforts for learning a single distribution.

→This Tutorial: Focus on a unifying view of multi-distribution learning problems and a powerful toolset.

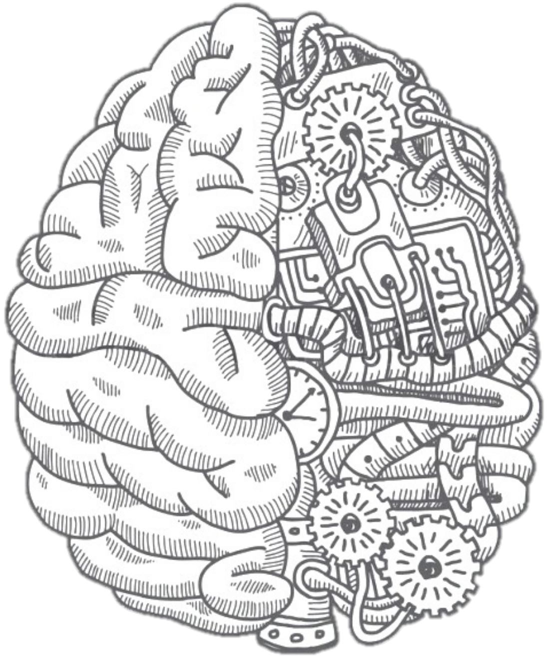
Q3. How should we procure these resources?

→Agents' incentives in providing resources in return for high quality solutions.

→This tutorial: Quantifying tradeoffs, highlighting technical challenges, and a call to action!

Question 1

How should we measure the learning performance across different tasks?



The Issue with Average Guarantees

Typical learning algorithms work well **on average** over the data sources

- Good for learning across data centers.
- Good for when the data is homogenous across sources.

Human and organization data:

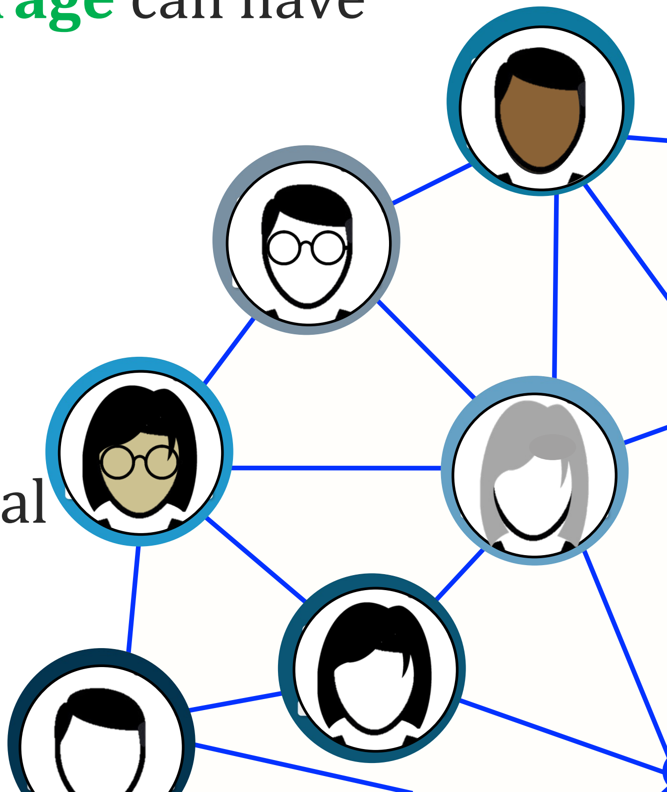
- For non-homogenous tasks, a model that has **5% error on average** can have **50% error for $^1/_{10}$ of the agents**.

Task difficulty varies significantly

→ Some populations are easier to learn than others.

→ Also depends on similarity across different populations.

→ Bad idea to have pre-fixed allocation of statistical/computational resources.



The Issue with Average Guarantees

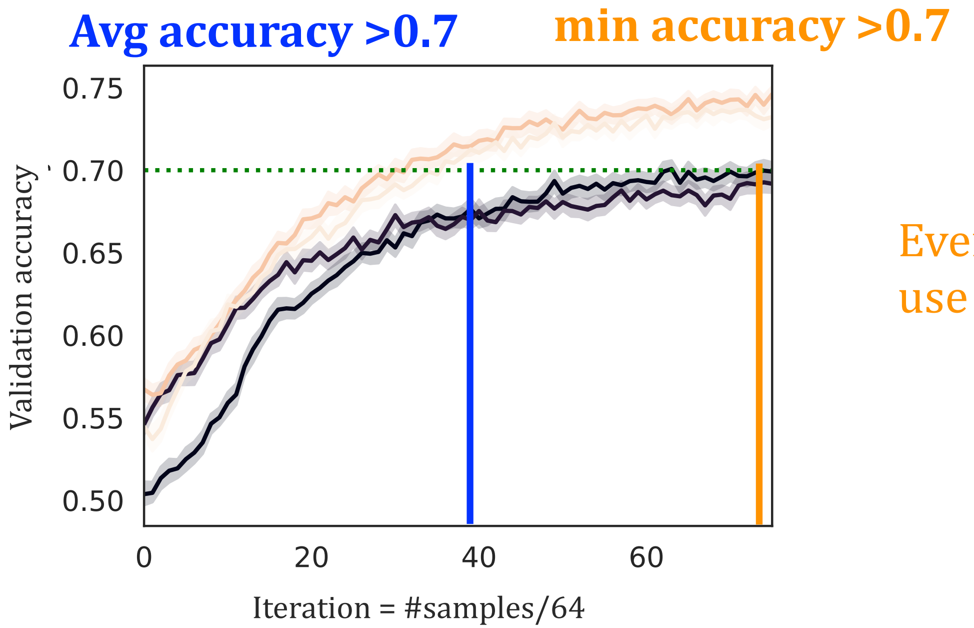
Typical learning algorithms work well **on average** over the data sources

- Good for learning across data centers.
- Good for when the data is homogenous across sources.

Human and organization data:

- For non-homogenous tasks, a model that has **5% error on average** can have **50% error for $^1/_{10}$ of the agents**.

Every agent uses 40 iterations.

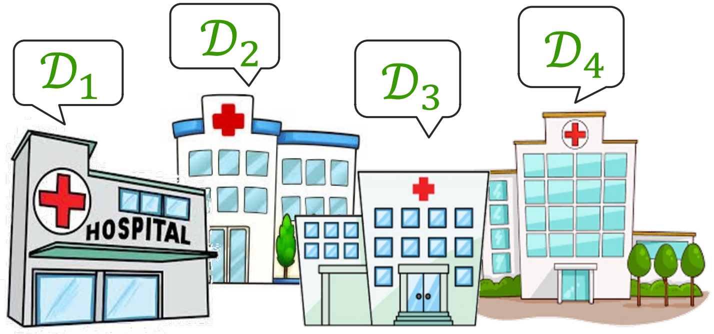


Every agent has to use 75 iterations.

Blum, **H**, Phillips, Shao ‘21

Multi-distribution Learning: Per-Group Guarantees

There are k populations/distributions. Represented by unknown $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$.



We want the to learn a function f that is good **for every population.**

A known loss function

$$\max_{i \in [k]} \mathrm{L}(\mathcal{D}_i, f) \quad \leq \epsilon$$

(uncovering a universally good model)

$$\max_{i \in [k]} \mathrm{L}\,(\mathcal{D}_i, f) \leq \min_{h^* \in H} \max_{i \in [k]} \mathrm{L}(\mathcal{D}_i, h^*) + \ \epsilon$$

(More general)

From a One to Multiple Distributions

Well-developed theory for how much resources are needed to learn a single distribution.

Import insights, algorithms, techniques, etc., from the single distribution setting to multi-distribution.

One Distribution

Given sample access to an unknown $\mathcal{D}$,

find f, s.t. with high probability,

$$\mathrm{L}(\mathcal{D}, f) \leq \min_{h^* \in H} \mathrm{L}(\mathcal{D}, h^*) + \epsilon$$

Multiple Distributions

Given sample access to unknown $\mathcal{D}_1, \dots, \mathcal{D}_k$,

find f, s.t. with high probability,

$$\max_{i \in [k]} \mathrm{L}(\mathcal{D}_i, f) \leq \min_{h^* \in H} \max_{i \in [k]} \mathrm{L}(\mathcal{D}_i, h^*) + \epsilon$$

Multi-distribution Learning: A Unifying Perspective

We want to learn a function f that is good **for every population.**

$$\max_{i\in[k]} \mathrm{L}(\mathcal{D}_i, f) \quad \leq \epsilon \qquad \text{(uncovering a universally good model)}$$

$$\max_{i\in[k]} \mathrm{L}\,(\mathcal{D}_i, f) \leq \min_{h^*\in H} \max_{i\in[k]} \mathrm{L}(\mathcal{D}_i, h^*) + \epsilon \qquad \text{(More general)}$$

Losses in this talk:

→ $\mathrm{L}(\mathcal{D}, f) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(z, f)]$ and $\ell(z, f)$ in finite range.

→ Take binary classification loss for convenience, i.e., $\mathrm{L}(\mathcal{D}, f)$ is expected error.

→ To emphasize, use notation $\mathrm{Loss}_{\mathcal{D}}(f)$.

→ Not every loss falls in this category (e.g. multi-calibration loss can be addressed with the same toolset, but does not follow this formulation)

Multi-distribution Learning: A Unifying Perspective

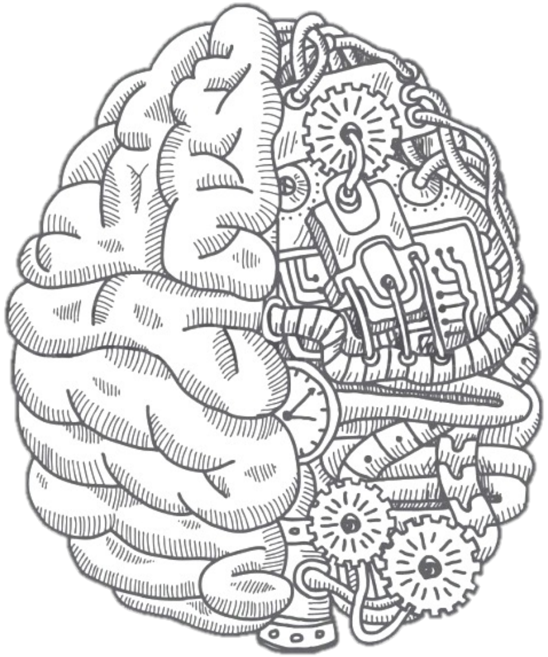
Within a span of 2-3 years, same study was initiated by 3 different communities. Mostly inspired by ideas of fairness, robustness, and collaborations.

→ Collaborative Learning [Blum, **H**, Procaccia, Qiao '17]
  → $\mathcal{D}_i$s represent agent distributions. Agents are willing to collaborate.
→ Agnostic (Fair) Federated Learning [Mohri, Sivek, Suresh'19]
  → $\mathcal{D}_i$s represent client distributions. Fairness goals and implications.
→ (Group) Distributionally Robust optimization [Sagawa, Koh, Hashimoto, Liang '19]:
  → $\mathcal{D}_i$s represent possible distribution shifts. Robustness and fairness goals.
→ And many more …

Beyond this talk: Economic and welfare perspective on minmax objectives
→Axiomatic and non-axiomatic approaches in cardinal welfare theory.
→Accuracy-fairness tradeoffs [Liang, Lu, Mu'21]

Question 2

How much resources do we need to meet the multi-distribution learning objective?



Information: From a Single to Multiple Distributions

We will focus on the “number of samples” as a resource.
→ What we discuss also has implications on “computational power” as a resource.

For comparison

Single Distribution

Given sample access to an unknown $\mathcal{D}$,

how many samples sufficient to learn f, s.t.

with high probability,

$$\mathrm{Loss}_{\mathcal{D}}(f) \leq \min_{h^* \in H} \mathrm{Loss}_{\mathcal{D}}(h^*) + \epsilon$$

Multiple Distributions

Given sample access to unknown $\mathcal{D}_1, \ldots, \mathcal{D}_k$,

how many samples sufficient to learn f, s.t.

with high probability,

$$\max_{i \in [k]} \mathrm{Loss}_{\mathcal{D}_i}(f) \leq \min_{h^* \in H} \max_{i \in [k]} \mathrm{Loss}_{\mathcal{D}}(h^*) + \epsilon$$

Basics: Learning a Single Distribution

Recall goal: Using samples from $\mathcal{D}$ learn a hypothesis with near-optimal error.

ERM: Given a sample set S, choose $h \in H$ that has the smallest error on the sample set.

How many sample to make this work?

→ Sufficient to have: For all $h \in H$, estimated error of h is within $\frac{\epsilon}{2}$ of its true error.

- H finite: concentration and union bound gives

$$\Pr\left[\begin{array}{c}\text{For at least one } h \in H \\ |\mathrm{Loss}_S(h) - \mathrm{Loss}_{\mathcal{D}}(h)| > \epsilon\end{array}\right] \leq \overbrace{|H|}^{\text{Union bound}} \times \overbrace{2\exp(-2m\epsilon^2)}^{\text{Hoeffding}}$$

- H infinite: "VC dimension" controls the $\boxed{\text{effective size}}$ of the hypothesis class

Sample Complexity (Single Task)

For any H, optimal sample complexity (worst-case over all D) is

Sample complexity = $\widetilde{\Theta}\left(\frac{VCD(H)}{\epsilon^2}\right) \leq O\left(\frac{\log(|H|)}{\epsilon^2}\right)$ Avg. Regret = $\widetilde{\Theta}\left(\sqrt{VCD(H)/T}\right)$

What Can We Hope for?

How does the sample complexity of multi-distribution should compare to the sample complexity or learning 1 or k distributions in isolation?

$$\max_{i\in[k]} \mathrm{Loss}_{\mathcal{D}_i}(f) \le \min_{h^*\in H} \max_{i\in[k]} \mathrm{Loss}_{\mathcal{D}}(h^*) + \epsilon$$

Two forces at play:

1. Distributions could be related to each other, so we can cross-learn.

→As k grows, impossible to have $\mathcal{D}_1, \dots, \mathcal{D}_k$ that are all **independent** and **hard**.

2. Needs some coordination in addition to learning.

→ Finding same function f to perform well on all $\mathcal{D}_1, \dots, \mathcal{D}_k$.

→ Could potentially result in worst dependence on learning parameters, like ϵ, δ, d,...

→ Thought exercise: Same target function h^* labeled all distributions (realizability)

→ Identifying which distribution is the hard one and only learning that distribution.

$$O\left(\log(k)\times\begin{matrix}\text{sample complexty of}\\ \text{learning 1 distribution}\end{matrix}\right)$$ [BHPQ17, HJZ22]

$$O\begin{pmatrix}\text{sample complexty of}\\ \text{learning 1 distribution}\end{pmatrix}$$

$$O\left(k\times\begin{matrix}\text{sample complexty of}\\ \text{learning 1 distribution}\end{matrix}\right)$$ [MSS19,SKHL19]

Coordination, Interactions, Adaptivity

Lack of interactions:

- # of samples, learning rates, and update frequencies decided non-interactively.
- Ignores varying distribution difficulty and relevance.

To benefit from cross-learning, the distributions need to interact adaptively.
→ Decisions about $\mathcal{D}_i$ must depend on how well $\mathcal{D}_i$ has done so far, compared to $\mathcal{D}_j$.

Sample complexity of existing algorithms, for k agents $= \Theta(k) \times$ Learning for 1 agent separately
1 agent # samples

[Blum, **H**, Procaccia, Qiao ’17]

Without an “interactive” protocol,
Collaborative learning (almost) as ineffective as not collaborating at all.

Optimal Sample Complexity

Interactions/Coordination/Adaptivity: All enabled by online algorithms.

Back to the Basics: The MinMax Formulation of these problems

$$\max_{i\in[k]} \mathrm{L}(\mathcal{D}_i, f) \leq \min_{h^*\in H} \max_{i\in[k]} \mathrm{L}(\mathcal{D}_i, h^*) + \epsilon$$

Minimizing Agent 1

2 Maximizing Agent

We want to find f that's an approximate MinMax strategy for the minimizing agent.

MinMax Games
Equilibria
and Regret

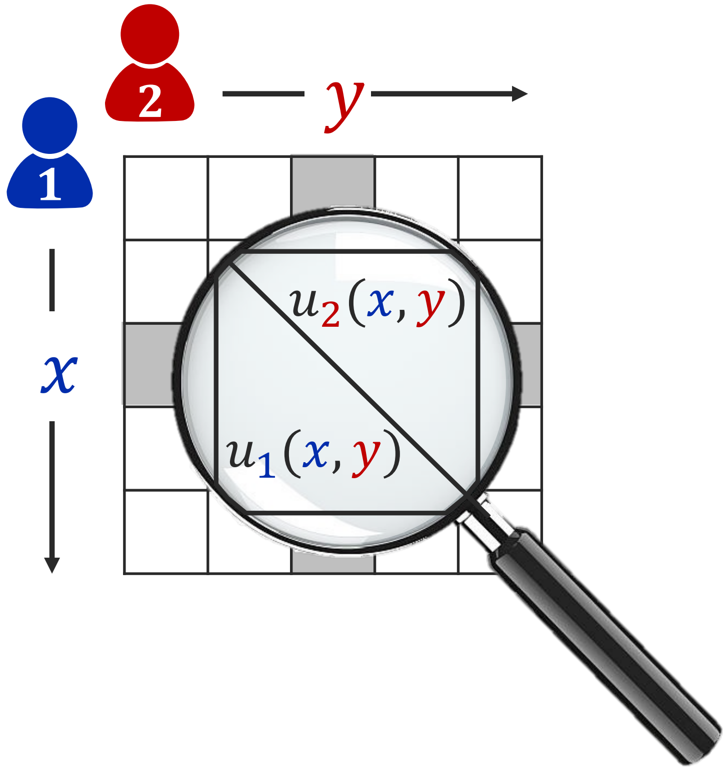
Basics: Two player Games

Players: Player 1 and 2

Strategies: Sets of actions X, Y

Payoffs: When 1 plays x and 2 plays y.

1's payoff : $u_1(x, y)$ 2's payoff : $u_2(x, y)$



Zero-sum games: focus of this section

$$-u_1(x, y) = u_2(x, y)$$

We'll call one of the loss and one gain/utility

$$\ell(x, y) = -u_1(x, y) \quad \text{(in this section)}$$

Basics: Solution Concepts

<u>Mixed Strategies:</u> 1 picks $P \in \Delta(X)$ and 2 picks $Q \in \Delta(Y)$. L(P, Q) is expected loss.

MinMax value

$$\min_{P} \max_{Q} \mathrm{L}(\mathrm{P}, \mathrm{Q})$$

(player 1 goes first)

MaxMin value

$$\max_{Q} \min_{P} \mathrm{L}(\mathrm{P}, \mathrm{Q})$$

(player 2 goes first)

(P, Q) is a **Nash equilibrium** if 1 can't improve their utility by unilaterally changing P, and 2 can't improve their utility by changing Q.

Von Neumann's MinMax Theorem

MinMax value = MaxMin value
Under some conditions, e.g., X and Y finite.

Basics: Why does MinMax Theorem hold?

1. Easy to see: Whoever goes second does a better job (minimizing or maximizing)

$$\min_{P} \max_{Q} \ \mathrm{L}(P, Q) \geq \max_{Q} \min_{P} \ \mathrm{L}(P, Q)$$

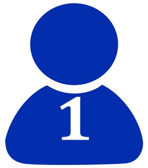
MinMax through online learning

[Freund-Schapire'96]

Online learnability and **MinMax** are about interactions with an adversary.

2. Interesting: One player plays **no-regret**, the other **best responds** (or also no-regret)

$$\min_{P} \max_{Q} \ \mathrm{L}(P, Q) \leq \max_{Q} \min_{P} \ \mathrm{L}(P, Q) + Avg.Regret$$



$$\bar{P} = \frac{1}{T} \sum P_t \qquad \bar{Q} = \frac{1}{T} \sum Q_t$$

$$\frac{1}{T} \sum L(P_t, Q_t) \ - \min_{P} \frac{1}{T} \sum L(P, Q_t) \leq Avg\ Regret$$

$$Q_t = \max_{Q} L(P_t, Q)$$

MinMax Games
Equilibria
and Regret

Multi-Distribution Learning as Game Solving

Re-imagining the multi-distribution learning objective as a zero–sum game.

$$\underbrace{\max_{i\in[k]} \mathrm{Loss}_{\mathcal{D}_i}(f) \leq \min_{h^*\in H} \max_{i\in[k]} \mathrm{Loss}_{\mathcal{D}_i}(h^*) + \epsilon}_{\textbf{Approximate MinMax equilibrium}}$$

Imagine two players:

- **Min Player:** Minimizing the loss over function class H.
- **Max Player:** Maximizing the loss over the class of distributions $\mathcal{D}_1, \ldots, \mathcal{D}_k$.
- **No-regret** algorithms to learn an approximate minmax equilibrium

Game loss: Unknown $\mathrm{Loss}_{\mathcal{D}_i}(f)$, we must estimate the game to solve it.

- **Sample complexity:** $\mathrm{Loss}_{\mathcal{D}_i}(f)$ is estimated through sampling from $\mathcal{D}_i$.

A Good (but not optimal) approach

An approach: Solve with a **no-regret algorithm** against a **best-responding agent**.

Min Player: The best-responding agent. For any distribution over [k], $\alpha_1^t, \ldots, \alpha_k^t$, it uses an Empirical Risk Minimizer to learn $h^t \in H$ on the distribution $P^t = \sum \alpha_i^t D_i$

Sample

Max Player: The no-regret learning agent. Maintains a distribution over $[k]$, say weights $\alpha_1^t, \ldots, \alpha_k^t$ over the agents. Proxy of how poorly they've been doing so far.

Sample

The No-regret algorithm tells how to split our resources across distributions.

Every round, α_i^t fraction of the samples come from distribution D_i.

Analysis

Simplifying assumption:

$$\min_{h^* \in H} \max_{i \in [k]} \text{Loss}_{\mathcal{D}_i}(h^*) = 0 \text{ i.e., there is realizability with respect to } h^*$$

See the whiteboard!

Pointers to the Optimal Approach

Stochastic Mirror-Prox Algorithm: Use tools for **solving stochastic games optimally.**

- Two intertwined no-regret algorithm.
- Assumes stochastic gradients: noisy estimated of any $\mathrm{Loss}_{P^t}(h^t)$.
- (deterministic/stochastic) Faster convergence than a No-regret+Best-response algorithm.

What is missing from Stochastic Mirror-Prox?

- We can **control how** accurate the noisy estimates of $\mathrm{Loss}_{P^t}(h^t)$ should be.

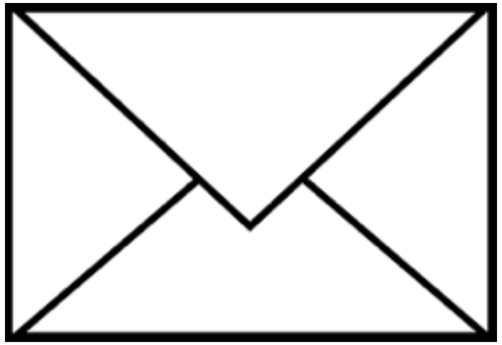
→We choose where, when, how much, to sample. Like adaptive sampling methods.

$$\underset{\text{Overall \# samples}}{\text{There is an Alg}} = \log(k)\begin{pmatrix}\text{sample complexity of}\\ \text{learning 1 distribution}\end{pmatrix}$$

[**H**, Jordan, Zhao '22] [Blum, **H.**, Procaccia, Qiao 17]

Important Message

Online Learning as a Powerful Medium
for Interactions in Learning
(beyond adversarial)



Beyond Accuracy Guarantees

Agents also incur cost for collecting information:

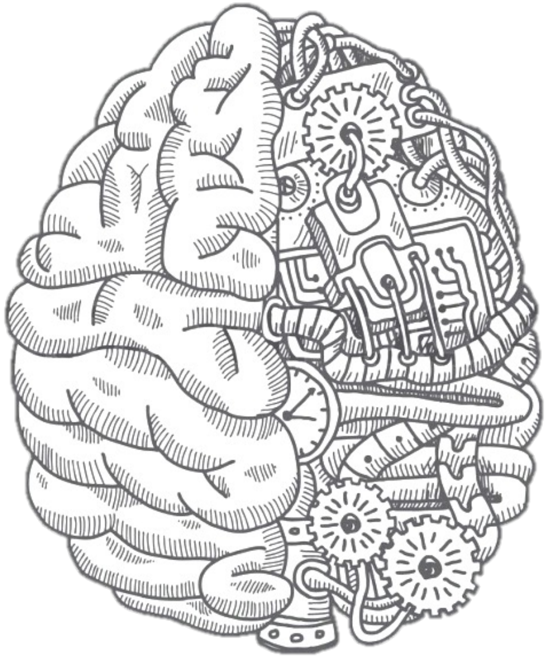
- E.g., cost for data set curation, privacy cost, etc.
- The protocol shouldn't ask for "unreasonable" amount of data.

→ Collaboration should be beneficial to all of its users.

Question 3

How should we procure resources needed for learning?

Theory for Multi-agent Sample Complexity!



Reasonable Share of Data

What we ask of agent i is <u>unreasonable</u> if:

- Ask i for more data than necessary, if he were to learn by himself.
- Part of i's contribution is exclusively used to meet the accuracy constraint of other agents and did not affect agent i.

[Blum, **H**, Phillips, Shao '21]

Individually Rational

1. Every agent's accuracy constraint is met, and
2. No agent collects more data than he needs, by himself.

If  's accuracy constraint is met, $m_i \leq m'_i$

$$\underbrace{(m_1, m_2, \ldots m_i \ \ldots m_k)}_{\text{\# of collected samples}}$$

$$(0,0, \ldots, m'_i, \ldots 0)$$

Reasonable Share of Data

What we ask of agent i is <u>unreasonable</u> if:

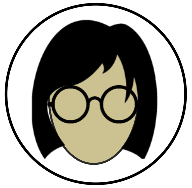
- Ask i for more data than necessary, if he were to learn by himself.
- Part of i's contribution is exclusively used to meet the accuracy constraint of other agents and did not affect agent i.

[Blum, **H**, Phillips, Shao '21]

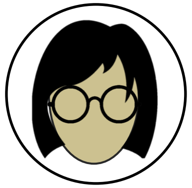
Stable **Equilibrium**

1. Every agent's accuracy constraint is met, and
2. No agent can reduce her contribution and still meet her accuracy constraint.

's accuracy constraint won't be met



$(m_1, m_2, \ldots m_i \ \ldots m_k)$



$(m_1, m_2, \ldots m'_i \ \ldots m_k)$

$m'_i < m_i$

Rationality and Equilibria Matter

Welfare of the agents:

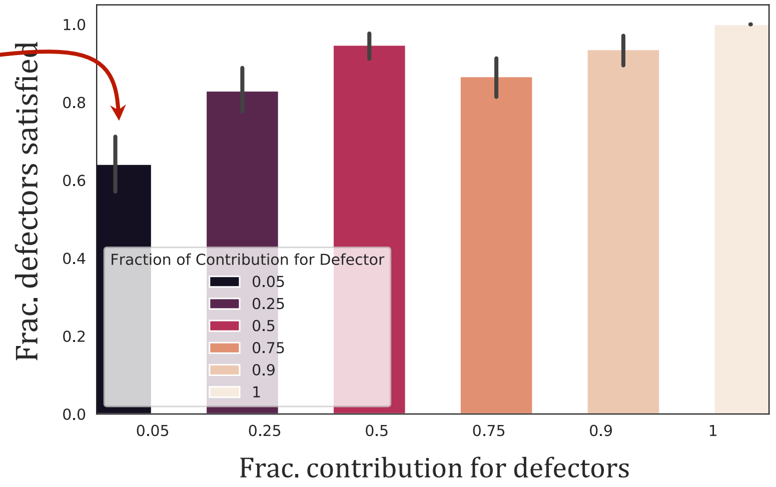
- Receiving a reasonable return in what resources you put in.

Usability and stability of systems over time:

- Even a small reduction in contribution across the agents impacts algorithmic performance.

State of the art learning algorithms are VERY far from equilibrium

60% of agents can unilaterally reduce their contributions to 5% of current levels.



Do these solution concepts exist?

Individually Rational allocations always exists, e.g., in a non-collaborative way.

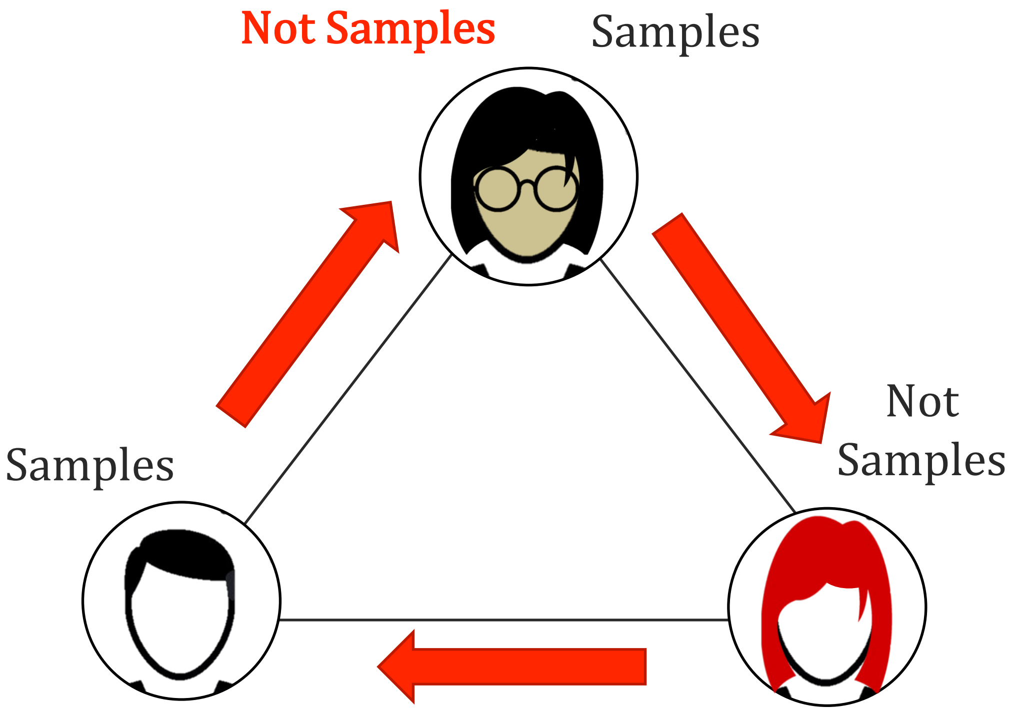
Unfortunately, some learning problems have no stable equilibrium!

But stable equilibria generally exist under mild assumptions.

Bad case for Equilibria

Each agent is much better at completing the next agent's task, then their own.

Let the feature of an instance in  's distribution encode the target function for the next agent, and its label reveal the target on their own data.



Cycling behavior:

- Non-continuous functions and actions
- More of a pure strategy equilibrium.

Good Case for Equilibria

Equilibria are guaranteed to exist, when the loss is monotone decreasing and Lipchitz in the **sampling effort**.

These are similar in nature to “pure” Nash equilibria, since we need to identify a deterministic number of samples.

Lipschitzness assumption allows us to talk about a random number of samples, without losing the integrity of learning problems.

Types of randomness:
Fine: Take 500 samples or 501 samples with with probability ½ ½.
Not Ok: Take 1000 or 1 samples with probability ½, ½.

Are Equilibria Efficient?

They may require more collective resources than the optimal collaboration!

In some cases,

Best equilibrium → **Some agents don't contribute.**

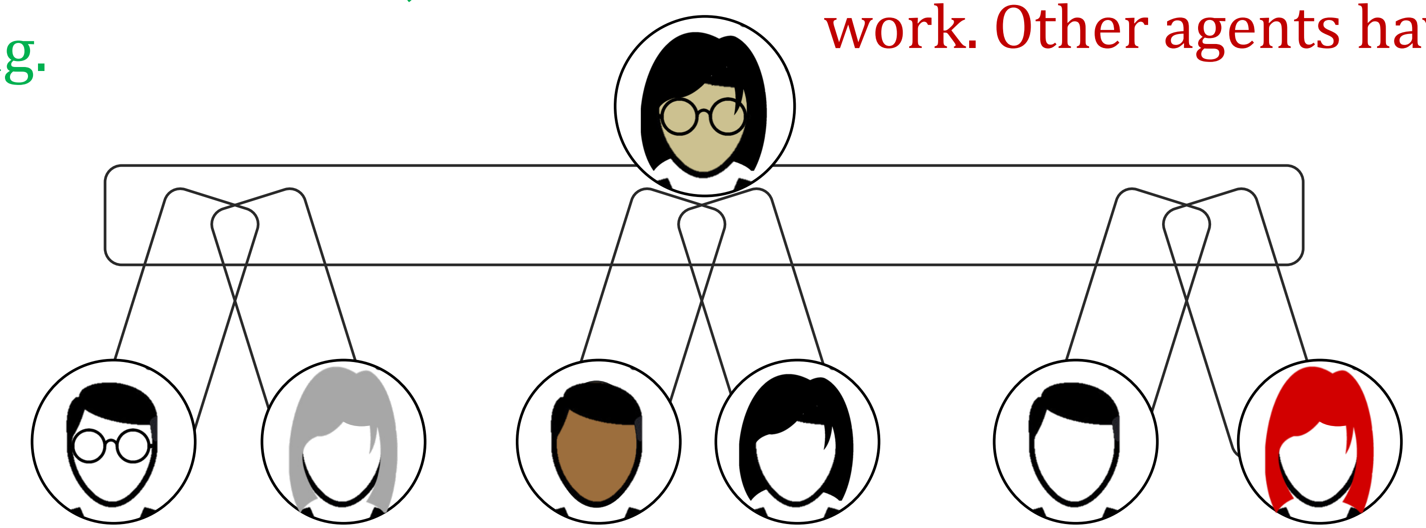
Judiciously introduce small inefficiencies, so everyone can continue benefitting from the system.

Price of Rationality and Stability

Individually rational or stable equilibria, require more collective resources than the optimal collaboration.

Optimal:  does all the work, others do nothing.

Stable/Rational:  does (almost) no work. Other agents have to do the work.



Equilibrium/Individual Rationality: Total work required to be done by other agents is large.

Overall # samples in the best IR/Stable allocation $= \Omega(\sqrt{k}) \times$ Overall # samples in the optimal collaboration

Optimality, Equilibria, and Free Riding

In some cases, equilibria are highly structured.

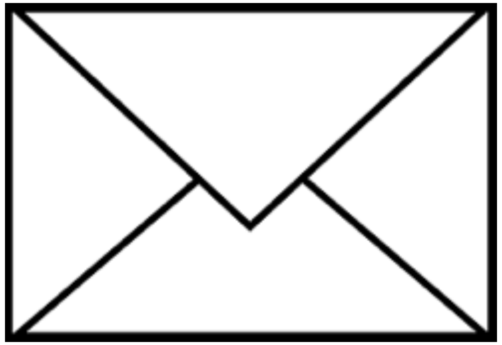
If the utility/loss of agents are linear functions of the contribution:

Difference between optimal:

- Any **equilibrium** is an optimal collaboration among **a subset of agents**.
- Free riding is part of equilibria.
- Free-riders don’t fundamentally change the optimal collaboration structure between participating agents.

Important Message

New mathematical foundation needed to
design learning algorithms that **act globally**,
and consider **per-agent incentives and objectives.**



Fundamental Questions We Discussed

Q1. How to measure learning performance across different tasks and distributions.

→What objective functions capture this performance?

→This tutorial: One unifying model (without too much detail).

Q2. How much resources are needed for learning and meeting these objectives?

→ Sample complexity and computational complexity.

→Relying on decades of efforts for learning a single distribution.

→This Tutorial: Focus on a unifying view of multi-distribution learning problems and a powerful toolset.

Q3. How should we procure these resources?

→Agents' incentives in providing resources in return for high quality solutions.

→This tutorial: Quantifying tradeoffs, highlighting technical challenges, and a call to action!