Multi-objective Learning:

An Algorithmic Toolbox for

Optimal Predictions on any Task and Loss

Nika Haghtalab

EECS, UC Berkeley

Brian Lee
UC Berkeley

Eric Zhao
UC Berkeley -> OpenAI

Siva Balakrishnan
CMU

Daniel Hsu
Columbia

Mike Jordan
UC Berkeley, Inria

On-demand Sampling: Learning Optimally from Multiple Distributions
*NeurIPS 2022: **H.**, Jordan, Zhao*

A Unifying Perspective on Multi-Calibration
*NeurIPS 2023: **H.**, Jordan, Zhao*

Panprediction: Optimal Predictions for Any Downstream Task and Loss
*AISTATS 2026: Balakrishnan, **H.**, Hsu, Lee, Zhao*

Multi-objective Learning:

An Algorithmic Toolbox for

Optimal Predictions on any Task and Loss

AISTATS's Oral Yesterday

Brian Lee

Optimal Predictions for Any Downstream Task and Loss

What does this mean?

What does this mean?

AISTATS's Oral Yesterday

Task (and Loss)

Brian Lee

- **Patient health records** $X \in \mathscr{X}$
- **Cardiac event in 24 hours** $Y \in \{0,1\}$
- Joint distribution $\mathscr{D}$

Predict **probability** p
of **cardiac event**
using **health records**

Loss $\ell \in \mathscr{L}$ compares p vs Y

AISTATS's Oral Yesterday

Task (and Loss)

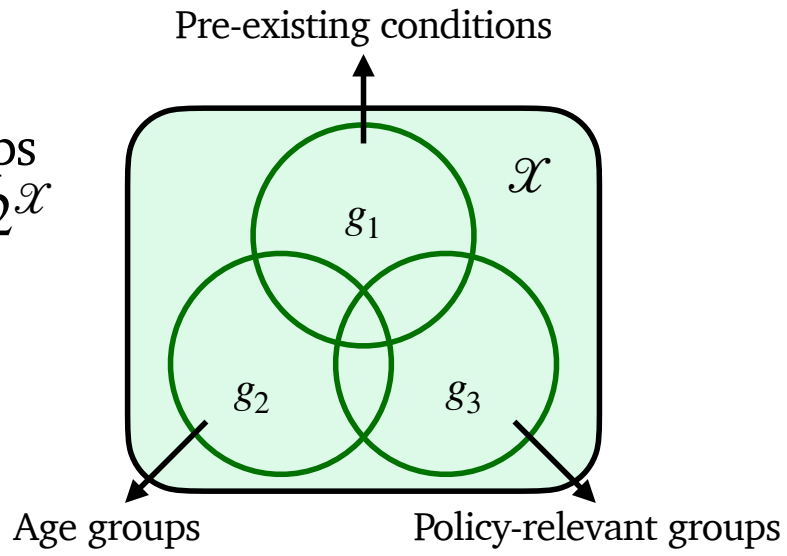
Brian Lee

- **Patient health records** $X \in \mathcal{X}$
- **Cardiac event in 24 hours** $Y \in \{0,1\}$
- Joint distribution $\mathcal{D}$

Predict **probability** p of **cardiac event** using **health records**

Loss $\ell \in \mathscr{L}$ compares p vs Y

Groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$



A **task** induced by g is the group-conditional distribution $D_g = D\,|_{g(X)=1}$

Brian Lee

AISTATS's Oral Yesterday

Optimality

Competitor hypothesis class $\mathscr{H} \subseteq \{h : \mathscr{X} \to [0,1]\}$

Physician (ℓ_1, g_1) → $h_1 = \arg\min_{h \in H} \mathbb{E}_{(X,Y)\sim D_{g_1}} \left[\ell_1(h(X), Y)\right]$

Discharge coordinator (ℓ_2, g_2) → $h_2 = \arg\min_{h \in H} \mathbb{E}_{(X,Y)\sim D_{g_2}} \left[\ell_2(h(X), Y)\right]$

Actuary (ℓ_3, g_3) → $h_3 = \arg\min_{h \in H} \mathbb{E}_{(X,Y)\sim D_{g_3}} \left[\ell_3(h(X), Y)\right]$

Want our model to compete with h_1, h_2, h_3 **simultaneously**

Brian Lee

What notion of calibration is sufficient, and statistically optimal? How this leads to improved and optimal sample complexity for a range of well-studied problems.

Optimality

Competitor hypothesis class $\mathscr{H} \subseteq \{h : \mathscr{X} \to [0,1]\}$

Physician (ℓ_1, g_1) → $h_1 = \arg\min_{h \in H} \mathbb{E}_{(X,Y)\sim D_{g_1}} \left[\ell_1(h(X), Y)\right]$

Discharge coordinator (ℓ_2, g_2) → $h_2 = \arg\min_{h \in H} \mathbb{E}_{(X,Y)\sim D_{g_2}} \left[\ell_2(h(X), Y)\right]$

Actuary (ℓ_3, g_3) → $h_3 = \arg\min_{h \in H} \mathbb{E}_{(X,Y)\sim D_{g_3}} \left[\ell_3(h(X), Y)\right]$

Want our model to compete with h_1, h_2, h_3 **simultaneously**

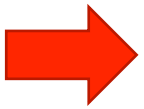
Multi-objective Learning:

An Algorithmic Toolbox for

Optimal Predictions on any Task and Loss

Foundations of Multi-objective Learning

Rest of this talk:



Multi-objective Learning, one unifying framework

- A toolkit for designing algorithms with multi-objective guarantees
  - → On-demand Sampling and Optimization
  - → Improved sample complexity of multi-objective learning
- How this applies to optimal predictions, for any downstream task and loss

Multi-objective Learning: Fine-grained guarantees

There are k distributions (tasks), represented by unknown $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ from which we can sample.

There are r possible loss functions $\ell^1, \ldots, \ell^r$, e.g., $\ell^j(x, y, f) = 1(y \neq f(x))$ or $(y - f(x))1(f(x) = v)$, and $\mathrm{L}^j_{\mathcal{D}_i}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\ell^j(x, y, f)]$

Multi-Objective Learning

Learn a function f that is simultaneously good for every distribution and every loss function in consideration.

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

$$\max_{j\in[r]} \max_{i\in[k]} \mathrm{L}^j_{\mathcal{D}_i}(f) - \min_{h^*\in H} \max_{j\in[r]} \max_{i\in[k]} \mathrm{L}^j_{\mathcal{D}_i}(h^*) \leq \epsilon \qquad \text{(best in class } H\text{)}$$

Example 1: Per-Distribution Guarantees

A function $h: X \to [0,1]$ and a loss function $\ell(x, y, h)$ measuring binary classification or regression error.

Classic Statistical Learning

Learn h with $L_D(h) \leq \epsilon$.
Even if $L_D(h) \leq 0.05$, h could have
50% error for $1/10$ of the population.

Problematic when that $1/10$ of the population correlates with a type.

[e.g. Valiant '84]

Per-Group Guarantee

Holding for each one of given distributions $\mathcal{D} = \{D_1, \dots, D_k\}$

all $D_i \in \mathcal{D}$, $L_{D_i}(h) \leq \epsilon$.

[formalized by Blum **H.** Procaccia Qiao'17]

Example 2: Per-Prediction Guarantees

Given a predictor $p: X \to [0,1]$ and measuring its biases

Unbiasedness

Rather a weak guarantee to ask for a predictor that's unbiased on average.

$$\mathbb{E}[y - p(x)] = 0.$$

May be biased at specific level sets.

Calibration Error

Seeking unbiased prediction, at every prediction level:

for all v, $\mathbb{E}_{x,y}[y|p(x) = v] = v$

Also for every downstream population:

for all v, $S \in 2^X$,

$\mathbb{E}_{x,y}[y|p(x) = v, x \in S] = v$

Relevant losses for calibration:

- $\ell^v(x, y, p) = \big(y - p(x)\big)1(p(x) = v)$ or $\ell^{v,S}(x, y, p) = \big(y - p(x)\big)1(p(x) = v, x \in S)$

A Unifying Perspective

In recent years, similar models were introduced by several different communities. Mostly inspired by ideas of fairness, robustness, and collaborations.

→ Collaborative Learning [Blum, H, Procaccia, Qiao '17; H, Jordan, Zhao 22-23]

→ $\mathcal{D}_i$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.

→ Multi-group Agnostic PAC [Rothblum, Yona'21]

→ $\mathcal{D}_i$s represent subpopulations and loss functions capture regret to optimal loss

→ Muti-Calibration [Herbert-Johnson, Kim, Reingold, Rothblum '18]

→ $\mathcal{D}$ a single distribution, with calibration loss functions $\big(y - f(x)\big)1(f(x) = v, x \in S)$ taking membership and predicted value.

→ Multi-group Fairness [Kearns, Neel, Roth, Wu'18]

→ $\mathcal{D}$ a single distribution, loss functions capturing errors of various types on $1(x \in S)$.

Generalization of Classic Models of Learning From One to Multiple Distributions and Objectives

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

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

For comparison

One Distribution (Statistical Learning)

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

find f, s.t. whp,

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

Multiple Distributions Losses

Given sample access to unknown $\mathcal{D}_1, \ldots, \mathcal{D}_k$ and losses $\ell^1, \ldots, \ell^r$, find f, s.t whp

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

Generalization of Classic Models of Learning From One to Multiple Distributions and Objectives

Algorithm: Empirical Risk Minimization

Sample complexity: $\widetilde{\Theta}\left(\frac{VCD(H)}{\epsilon^2}\right)$

Algorithmic paradigm?

Sample complexity?

For comparison

One Distribution (Statistical Learning)

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 Losses

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

losses $\ell^1, \ldots, \ell^r$, find f, s.t whp

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

Foundations of Multi-objective Learning

Rest of this talk:

- Multi-objective Learning, one unifying framework
- A toolkit for designing algorithms with multi-objective guarantees
  - → On-demand Sampling and Optimization
  - → Improved sample complexity of multi-objective learning
- How this applies to optimal predictions, for any downstream task and loss

For conciseness, rest of the talk focuses on single loss and multiple distributions

The Need for Algorithmic Data Collection

Standard approach: Collect a priori fixed sized data sets
→ Ignores varying distribution difficulty and relevance.

Non-adaptive Data Collection

Sample complexity of existing algorithms, for k distributions $= \Theta(k) \times$ Learning for 1 distribution separately

[Blum, H, Procaccia, Qiao '17]
[H, Montasser, Qiao ongoing]

Without an adaptive protocol for data collection, multi-objective learning doesn't save data any more than learning many specialized models would do.

On-Demand Sampling

To effectively learn with multi-objective guarantees, the learning algorithm has to **actively curate and shape the data set**, **not just scale data sets.**

On-demand Sampling

Generate training data from where its needed, when its needed.

Multi-Objective Learning with On-demand Sampling

Standard approach: Collect a priori fixed sized data sets

→ Ignores varying distribution difficulty and relevance.

Non-adaptive Data Collection

Sample complexity of existing algorithms, for k distributions $= \Theta(k) \times$ Learning for 1 distribution separately

[Blum, H, Procaccia, Qiao '17]

[H, Montasser, Qiao ongoing]

Multi-Objective Learning with On-demand Sampling

Standard approach: Collect a priori fixed sized data sets

→ Ignores varying distribution difficulty and relevance.

Non-adaptive Data Collection

Sample complexity of existing algorithms, for k distributions $= O\left(k \ln(k) \cdot \frac{\log(|H|)}{\epsilon^2}\right)$

In this regime
Group DRO
Multi group agnostic
Agnostic Federated Learning
…

On-demand Sampling

On-demand approach: Collect samples adaptively, e.g., how many samples to take from $\mathcal{D}_i$ depends on intermediate performance on $\mathcal{D}_1, \ldots, \mathcal{D}_k$.

There is an algorithm
Overall # samples $= O\left(\frac{\log(|H|)}{\epsilon^2} + \frac{k \ln(k)}{\epsilon^2}\right)$

[Blum, H, Procaccia, Qiao '17]

[H, Jordan, Zhao '22]

Foundations of Multi-objective Learning

Rest of this talk:

- Multi-objective Learning, one unifying framework
- A toolkit for designing algorithms with multi-objective guarantees
  - → On-demand Sampling and Optimization
  - → Improved sample complexity of multi-objective learning
- How this applies to optimal predictions, for any downstream task and loss

Optimization with On-demand Sampling

Eric Zhao

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

Approximate MinMax equilibrium

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

Minimizing Agent: Minimize the loss over function class H

Maximizing Agent: Maximize the loss over the class of distributions $\mathcal{D}_1, \ldots, \mathcal{D}_k$.

Using no-regret algorithms to find an approximate minmax equilibrium.

- Sufficient for one player to play no-regret, and another to best respond or be no-regret.
  → Best-Response v. No-regret, No-regret v. No-regret, No-regret v. Best-Response.
- **Sample complexity:**
  → Number of samples and the quality of estimator for $\mathrm{L}_{\mathcal{D}_i}(f)$.
  → The rate at which game-playing dynamics converge.

How does On-demand Sampling Help?

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 : proportional to α_i^t.

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.

Depending on $h^1, \ldots, h^{t-1}$.

Sample

Why does this work?

Simplifying assumption: $\min_{h^* \in H} \max_{i \in [k]} \mathrm{L}_{\mathcal{D}_i}(h^*) = 0$ i.e., realizable

Min Player: The best-responding player. 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 $P^t = \sum \alpha_i^t D_i$.

$$\mathrm{L}_{P^t}(h^t) \leq \epsilon' \qquad \text{Samples } O\left(\frac{\log(|H|)}{\epsilon'^2}\right)$$

Max Player: The no-regret learning player. 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.

$$|\mathrm{L}_{\mathcal{D}_i}(h^t) - \widehat{\mathrm{L}_{\mathcal{D}_i}}(h^t)| \leq \epsilon'.$$

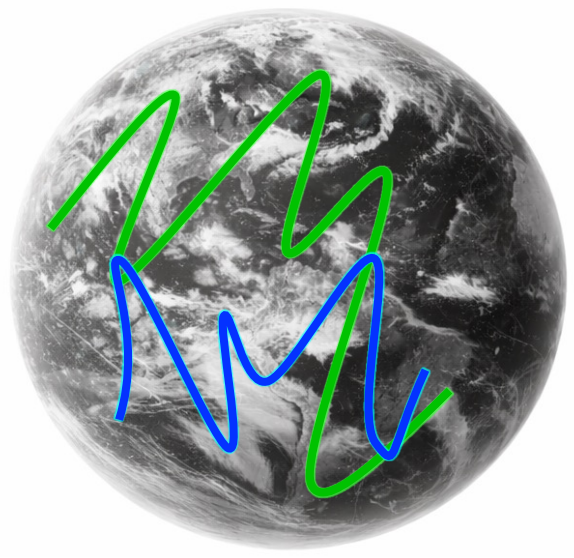
$$\epsilon' \geq \frac{1}{T}\sum \mathrm{L}_{P^t}(h^t) \geq \max_{i \in [k]} \frac{1}{T}\sum \mathrm{L}_{\mathcal{D}_i}(h^t) - \underbrace{\frac{\sqrt{T.\log(k)}}{T}}_{\leq \epsilon' \text{ for } \mathrm{T} = \frac{\log(k)}{\epsilon'^2}}.$$

$\max_{i \in [k]} \mathrm{L}_{\mathcal{D}_i}(\bar{h}_T)$ when $\bar{h}_T$ is a randomized classifier uniformly from $h^1, \ldots, h^T$

Improving the performance via a perspective on solving stochastic minmax games.

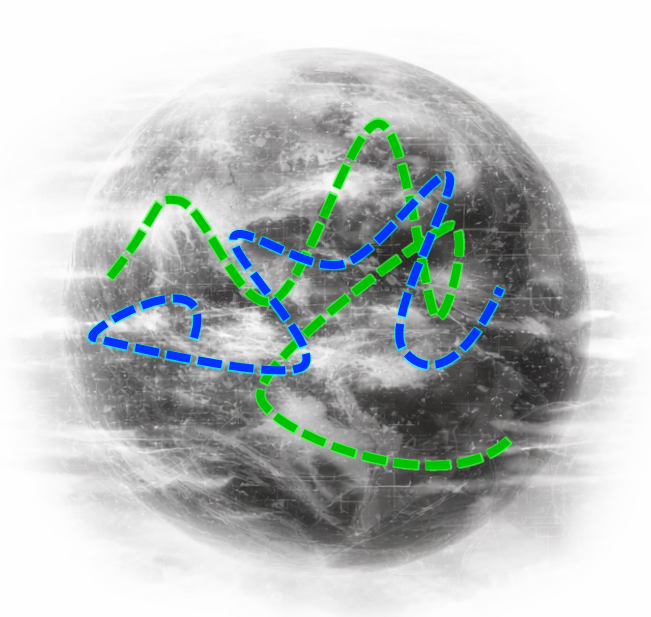
A perspective on solving stochastic min-max

Payoffs $\mathrm{L}_{\mathcal{D}_i}(f)$ are known



Trajectory taken by pairs of no-regret algorithms using losses $\mathrm{L}_{\mathcal{D}_i}(f)$

Only estimate $\widehat{\mathrm{L}_{\mathcal{D}_i}}(f)$ available

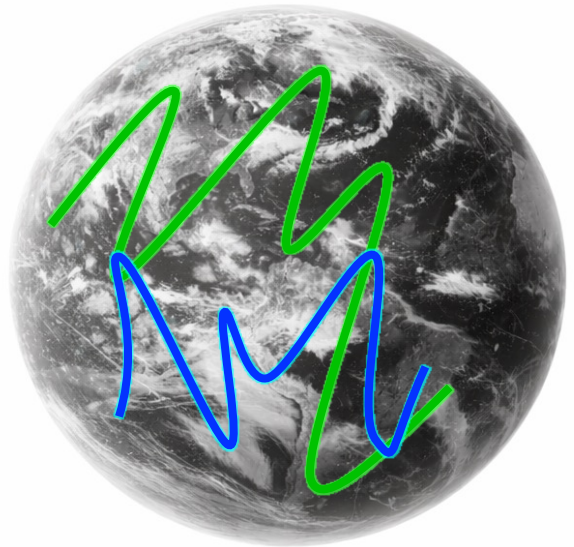


Trajectory of h^t and P^t taken by our algorithms.

We want our algorithm's trajectories (right) to be similar to hypothetical no-regret trajectories (left)

Up to now: Step-by-step similarity

Previous analysis: Ensure that **every step** of the adaptive process approximates the hypothetical no-regret dynamics we could have played if we knew expected losses.



Convenient:

- Every step has at most ϵ bias
- Regardless of correlations across time, total bias $\leq T \times \epsilon$

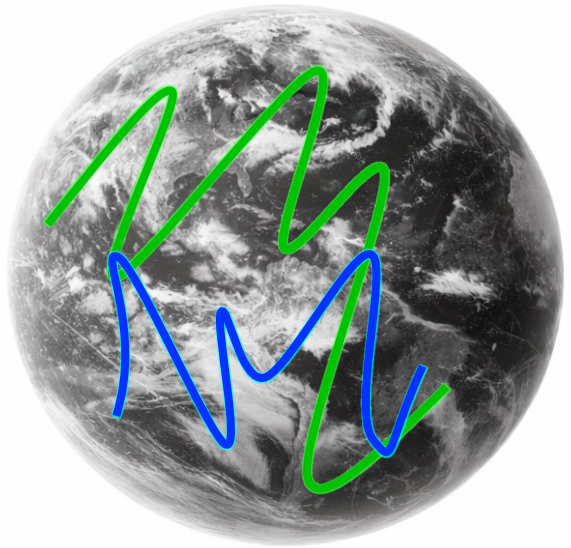
Bottleneck of this approach:

- We need many samples to train up to only $\leq \epsilon$ error at every round.
- Expensive for sample complexity

Optimal Multi-distribution learning

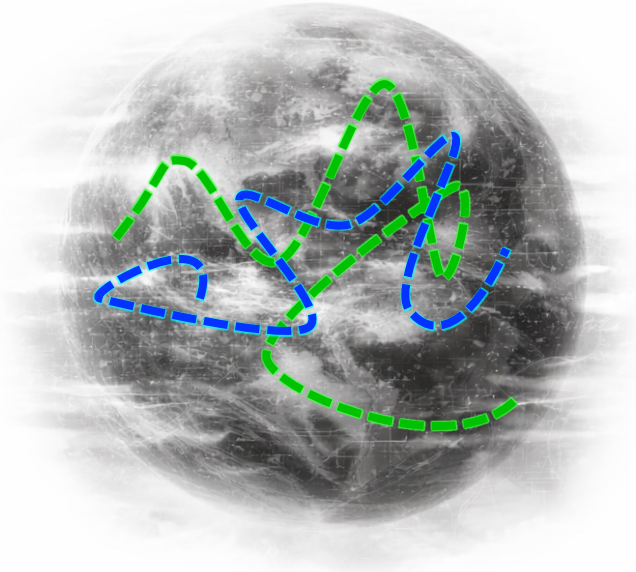
A different perspective:

- Don't worry about large per-step deviation, as long as they are **conditionally unbiased**.
- Control the **overall variance** of the stochastic process, rather than step-wise deviation.
- i.e., want true regret ≈ empirical regret for these choices.



No-regret vs No-regret perspective:

- Minimizer: Pick h^ts to get the weaker **no-regret guarantees.**
  → h^t can be picked before looking at $(x_t, y_t) \sim P^t$
  →Take a single sample (x_t, y_t)
- Maximizer: as before pick $P^t = \sum \alpha_i^t D_i$ to be no-regret.
  → P^t is picked before looking at h^t.



We are now working with unbiased estimator of $\mathrm{L}_{P^t}(h^t)$ but constant variance.

Foundations of Multi-objective Machine Learning

Rest of this talk:

- Multi-objective Learning, one unifying framework
- A toolkit for designing algorithms with multi-objective guarantees
  - → On-demand Sampling and Optimization
  - → Improved sample complexity of multi-objective learning
- How this applies to optimal predictions, for any downstream task and loss

Back to Optimal Predictions for Any Task and Loss

All relevant information is captured in the Bayes predictor $p^\star(x) = \mathbb{E}[y|x]$ and we could just post-process $p^\star$ for any loss ℓ:

$$\text{Choose decision } h(x) = \mathrm{BR}\big(p^\star(x)\big) \quad \text{, where } \mathrm{BR}(\alpha) \text{ is the decision d}$$
$$\text{minimizing } \mathbb{E}_{\hat{y}\sim \mathrm{Ber}(\alpha)}[\ell(d, \hat{y})]$$

Problem with Bayes predictor: **Not learnable**.

Is there a **tractable and learnable** predictor working as well for any loss and any population?

A predictor $p: X \to [0,1]$ is a **panpredictor** for a hypothesis class H if for any loss ℓ, and any population $S \in G$ (for some $G \subseteq 2^X$),

$$\mathbb{E}[\ell(\mathrm{BR}(p(x), y)| \; x \in S] \leq \min_{h\in H} \mathbb{E}[\ell(h(x), y)| \; x \in S] + \epsilon$$

Treating panpredictor as the true Bayes predictor

Optimal hypothesis for population S according to ℓ

Balakrishnan, **H**, Hsu, Lee, Zhao '25

Multi-Objective Learning for Panprediction

Sufficient: a predictor calibrated, not on every level-set, but on thresholds of level sets.

→Intuition: For well-behaved losses, optimal decision switches when risk score passes a threshold. Only unbiasedness on thresholds matter.

This is **Multi-objective Learning with**

Biasedness Prediction range Decision range

Losses $\ell^{v,w}(x,y,p) = \overbrace{(y - p(x))}^{\text{Biasedness}} \cdot \overbrace{1(p(x) \le v)}^{\text{Prediction range}} \cdot \overbrace{1(h(x) \le w)}^{\text{Decision range}}$

Distributions $D_{|S}$ for any subgroup $S \in G$.

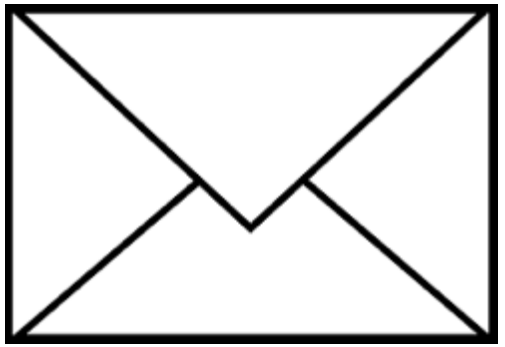
[Generalization of Step calibration [QZ25] aka proper calibration [OKK25]]

Optimal multi-objective algorithms using $\tilde{O}\left(\frac{\log(|H||G|)}{\varepsilon^2}\right)$ samples total, of the same order of sample complexity of single loss, single task learning up to log factors!

*Simultaneously minimizing **infinitely many losses** over **infinitely many tasks** can be as **statistically easy** as minimizing one loss over one task.*

Final Words

This approach has led to optimal sample complexity (and improved) for many problems in calibration, multi-calibration, omni-predictions, group fairness notions, …



A principled approach to multi-objective learning addresses the core design challenges of general-purpose machine learning models.

Brian Lee
UC Berkeley

Eric Zhao
UC Berkeley -> OpenAI

Siva Balakrishnan
CMU

Daniel Hsu
Columbia

Mike Jordan
UC Berkeley, Inria

On-demand Sampling: Learning Optimally from Multiple Distributions
*NeurIPS 2022: **H.**, Jordan, Zhao*

A Unifying Perspective on Multi-Calibration
*NeurIPS 2023: **H.**, Jordan, Zhao*

Panprediction: Optimal Predictions for Any Downstream Task and Loss
*AISTATS 2026: Balakrishnan, **H.**, Hsu, Lee, Zhao*