NEURAL INFORMATION PROCESSING SYSTEMS

Responsible AI (RAI) Games and Ensembles

Yash Gupta, Runtian Zhai, Arun Suggala, Pradeep Ravikumar

Paper: https://arxiv.org/abs/2310.18832
Code: https://github.com/yashgupta-7/rai-games

CARNEGIE MELLON UNIVERSITY
PITTSBURGH PENNSYLVANIA 1900

ML
MACHINE LEARNING
DEPARTMENT

Responsible AI (RAI): Introduction and Motivation

- **Motivation:** AI is increasingly being used in high-stakes decision-making contexts such as hiring, criminal justice, and healthcare.

- **Setting:** Under the umbrella of "responsible AI", an emerging line of work has attempted to formalize desiderata ranging over ethics, fairness, robustness, and safety, many of which can be written as *min-max problems* involving optimizing some worst-case loss under a set of predefined distributions.

- **Problem:** majority of recent work around these problems is fragmented and usually focuses on optimizing one of these aspects at a time (DRO [Namkoong and Duchi, 2017, Duchi and Namkoong, 2018], GDRO [Sagawa et al., 2019], CVaR [Zhai et al., 2021a], Distribution Shift [Hashimoto et al., 2018, Zhai et al., 2021b]).

- **Proposal:** a general game-theoretic framework for solving these problems and learning responsible AI models. We propose practical algorithms to solve these games, as well as statistical analyses of solutions of these games.

Problem Setup

- standard supervised prediction setting: input random variable $X \in \mathcal{X} \subseteq \mathbb{R}^d$, output random variable $Y \in \mathcal{Y}$, and samples $S = \{(x_i, y_i)\}_{i=1}^n$ drawn from a distribution P_{data} over $\mathcal{X} \times \mathcal{Y}$

- The empirical distribution $\widehat{P}_{\text{data}}$ over the samples, set H of hypothesis functions $h : \mathcal{X} \mapsto \mathcal{Y}$

- Goodness of a predictor via a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, which yields the empirical risk: $\widehat{R}(h) = \mathbb{E}_{\widehat{P}_{\text{data}}} \ell(h(x), y) \quad \text{where} \quad \mathbb{E}_{\widehat{P}_{\text{data}}}(f(x, y)) = \frac{1}{n} \sum_{i=1}^n f(x_i, y_i)$.

- Apart from having low expected risk, h is required to have certain properties. e.g. robustness, fairness w.r.t subpopulations, superior tail performance, resistance to adversarial attacks, etc - cast all these subproblems into an umbrella term "Responsible AI".

RAI Risks - I

- Do not wish to compute an unweighted average over training samples - due to RAI considerations.

Definition 1 (RAI Risks) *Given a set of samples* $\{(x_i, y_i)\}_{i=1}^n$, *we define the class of empirical RAI risks (for Responsible AI risks) as:* $\widehat{R}_{W_n}(h) = \sup_{w \in W_n} \mathbb{E}_w(h(x), y)$, *where* $W_n \subseteq \Delta_n$, *is some set of sample weights (a.k.a uncertainty set), and* $\mathbb{E}_w(f(x, y)) = \sum_{i=1}^n w_i f(x_i, y_i)$.

- Given the empirical RAI risk of a hypothesis - naturally wish to obtain the hypothesis that minimizes the empirical RAI risk

- Can be seen as solving a zero-sum game

Definition 2 (RAI Games) *Given a set of hypothesis* H, *and a RAI sample weight set* W_n, *the class of RAI games is given as:* $\min_{h \in H} \max_{w \in W_n} \mathbb{E}_w(h(x), y)$.

RAI Risks - II

- Various choices of $\mathbf{W_n}$ give rise to various RAI risks.

Name	W_n	Description
Empirical Risk Minimization	$\{\widehat{P}_{\text{data}}\}$	object of focus in most of ML/AI
Worst Case Margin	Δ_n, entire probability simplex	used for designing margin-boosting algorithms [Warmuth et al., 2006, Bartlett et al., 1998]
Soft Margin	$\{w : KL(w\|\|\widehat{P}_{\text{data}}) \leq \rho_n\}$	used in the design of AdaBoost [Freund and Schapire, 1995]
α-Conditional Value at Risk (CVaR)	$\{w : w \in \Delta_n, w \preceq \frac{1}{\alpha n}\}$	used in fairness [Zhai et al., 2021a, Sagawa et al., 2019]
Distributionally Robust Optimization (DRO)	$\{w : D(w\|\|\widehat{P}_{\text{data}}) \leq \rho_n\}$	various choices for D have been studied f-divergence [Duchi and Namkoong, 2018]
Group DRO	$\{\widehat{P}_{\text{data}}(G_1), \widehat{P}_{\text{data}}(G_2), \ldots \widehat{P}_{\text{data}}(G_K)\}$ $\widehat{P}_{\text{data}}(G_i)$ is dist. of i^{th} group	used in group fairness, agnostic federated learning [Mohri et al., 2019]

Table 1: Various ML/AI problems that fall under the umbrella of RAI risks.

RAI Games - Moving to ensembles

- Good worst-case performance over the sample weight set W_n is generally harder, especially for a simpler set of hypotheses

- Natural to consider deterministic ensemble models
  - Gives us more powerful classes

Definition 3 (Deterministic Ensemble) *Consider the problem of classification, where $\mathcal{Y}$ is a discrete set. Given a hypothesis class H, a deterministic ensemble is specified by some distribution $Q \in \Delta_H$, and is given by: $h_{det;Q}(x) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{h \sim Q} \mathbb{I}[h(x) = y]$. Correspondingly, we can write the deterministic ensemble RAI risk as $\widehat{R}_{W_n}(h_{det;Q}(x)) = \max_{w \in W_n} \mathbb{E}_w \ell(h_{det;Q}(x), y)$.*

- This admits a class of deterministic RAI games

Definition 4 (Deterministic Ensemble RAI Games) *Given a set of hypothesis H, a RAI sample weight set W_n, the class of RAI games for deterministic ensembles over H is given as:*

$$\min_{Q \in \Delta_H} \max_{w \in W_n} \mathbb{E}_w \ell(h_{det;Q}(x), y).$$

RAI Games - Moving to *random* ensembles

- Aforementioned game is computationally less amenable because of the **non-smooth** nature of de-randomized predictions.

- To this end, we consider the following randomized ensembles:

Definition 5 (Randomized Ensemble) *Given a hypothesis class H, a randomized ensemble is specified by some distribution $Q \in \Delta_H$, and is given by:* $\mathbb{P}[h_{rand;Q}(x) = y] = \mathbb{E}_{h\sim Q}\mathbb{I}[h(x) = y]$. *Similarly, we can define its corresponding randomized ensemble RAI risk:* $\widehat{R}_{rand;W_n}(Q) = \max_{w \in W_n} \mathbb{E}_{h\sim Q}\mathbb{E}_w \ell(h(x), y)$.

Definition 6 (Randomized Ensemble RAI Games) *Given a set of hypothesis H, a RAI sample weight set W_n, the class of mixed RAI games is given as:*

$$\min_{Q \in \Delta_H} \max_{w \in W_n} \mathbb{E}_{h\sim Q}\mathbb{E}_w \ell(h(x), y). \tag{1}$$

- Much better class of zero-sum games
  - **linear** in both the hypothesis distribution P well as the sample weights
  - if the sample weight set is convex, is a **convex-concave** game.
  - under some mild conditions, this game has a **Nash** equilibrium

RAI Algorithms - I

- **Game Play -** Both players rely on no-regret algorithms to decide their next action
  - Follow-The-Regularized-Leader (FTRL) update for weights
  - Best Response (BR) update for hypotheses

Algorithm 1 Game play algorithm for solving Equation (1)

Input: Training data $\{(x_i, y_i)\}_{i=1}^n$, loss function ℓ, constraint set W_n, hypothesis set H, strongly concave regularizer R over W_n, learning rates $\{\eta^t\}_{t=1}^T$

for $t \leftarrow 1$ to T **do**

    FTRL: $w^t \leftarrow \text{argmax}_{w \in W_n} \sum_{s=1}^{t-1} \mathbb{E}_w \ell(h^s(x), y) + \eta^{t-1}\text{Reg}(w)$

    BR: $h^t \leftarrow \text{argmin}_{h \in H} \mathbb{E}_{w^t} \ell(h(x), y)$

end for

return $P^T = \frac{1}{T} \sum_{t=1}^T w^t$, $Q^T = \text{Unif}\{h^1, \ldots h^T\}$

RAI Algorithms - II

- **Greedy -** use Frank Wolfe (FW) for the inner maximization problem
  - when it is smooth, updates given by:

$$Q^t \leftarrow (1-\alpha^t)Q^{t-1} + \alpha^t G, \quad \text{where } G = \underset{Q}{\text{argmin}} \left\langle Q, \nabla_Q L(Q^{t-1}) \right\rangle .$$

  - when non-smooth, perform Moreau smoothing

$$L_\eta(Q) = \max_{w \in W_n} \mathbb{E}_{h\sim Q}\mathbb{E}_w \ell(h(x), y) + \eta \text{Reg}(w).$$

  - a slightly different AdaBoost-like algorithm by relaxing the simplex constraint on Q

Algorithm 2 Greedy algorithms for solving Equation (1)

Input: Training data $\{(x_i, y_i)\}_{i=1}^n$, loss function ℓ, constraint set W_n, hypothesis set H, strongly concave regularizer R over W_n, regularization strength η, step sizes $\{\alpha^t\}_{t=1}^T$

1: **for** $t \leftarrow 1$ to T **do**
2: $G^t = \text{argmin}_Q \left\langle Q, \nabla_Q L_\eta(Q^{t-1}) \right\rangle$
3: **FW:** $Q^t \leftarrow (1-\alpha^t)Q^{t-1} + \alpha^t G^t$ / **Gen-AdaBoost:** $Q^t \leftarrow Q^{t-1} + \alpha^t G^t$
4: **end for**
5: **return** Q^T

Experiments

- **Goal:** demonstrate the generality of proposed RAI methods by studying a well studied problem i.e. **subpopulation shift** under various settings
  - Domain-oblivious (DO): we do not know the sub-populations [Hashimoto et al., 2018, Lahoti et al., 2020]
    - χ2-DRO constraint set to control
  - Domain-aware (DA): where we know the sub-populations [Sagawa et al., 2019]
    - Group DRO constraint set
  - Partially domain-aware (PDA): where only some might be known
    - intersection over Group DRO constraints over the known domains and χ2 constraints to control

- **Baselines -**
  - Deterministic classifiers trained on empirical risk (ERM) and DRO risks
    - the quasi-online algorithm for Group DRO [Sagawa et al., 2019] (Online GDRO)
    - ITLM-inspired SGD algorithm [Zhai et al., 2021b, Shen and Sanghavi, 2018] for χ^2 DRO (SGD (χ^2))
  - Ensemble models AdaBoost [Schapire, 1999].

Experiments - Results

- RAI-FW and RAI-GA methods significantly improve the worst-case performance with only 3-5 base learners across all datasets in all three settings, while maintaining average case performance.

- The plug-and-play framework allows for several different to enhance various responsible AI qualities at once. RAI is able to optimize effectively for both known and unknown subpopulations

Table 2: (Table 1 in the paper) Mean and worst-case expected loss for baselines, RAI-GA and RAI-FW. (Complex) indicates the use of larger models. Constraint sets W_n are indicated in (.). Each experiment is carried out over three random seeds and confidence intervals are reported.

Setting	Algorithm	COMPAS		CIFAR-10 (Imbalanced)		CIFAR10		CIFAR100	
		Average	Worst Group	Average	Worst Class	Average	Worst Class	Average	Worst Class
DO (Complex)	ERM	31.3 ±0.2	31.7 ±0.1	12.1 ±0.3	30.4 ±0.2	8.3 ±0.2	21.3 ±0.5	25.2 ±0.2	64.0 ±0.7
	RAI-GA (χ^2)	31.3 ±0.2	**31.2** ±0.2	11.7 ±0.4	**29.0** ±0.3	8.2 ±0.1	19.0 ±0.1	25.6 ±0.4	**56.8** ±0.8
	RAI-FW (χ^2)	31.2 ±0.1	31.4 ±0.3	11.9 ±0.1	29.1 ±0.2	8.0 ±0.3	**15.4** ±0.4	25.4 ±0.2	58.0 ±1.1
DO	ERM	32.1 ±0.3	34.6 ±0.4	14.2 ±0.1	33.6 ±0.3	11.4 ±0.4	27.0 ±0.1	27.1 ±0.3	66.0 ±1.1
	AdaBoost	31.8 ±0.4	32.6 ±0.3	15.2 ±0.2	40.6 ±0.2	12.0 ±0.1	28.7 ±0.3	28.1 ±0.2	72.2 ±1.2
	SGD (χ^2)	32.0 ±0.2	33.7 ±0.2	13.3 ±0.3	31.7 ±0.4	11.3 ±0.3	24.7 ±0.1	27.4 ±0.1	65.9 ±1.2
	RAI-GA (χ^2)	31.5 ±0.2	33.2 ±0.3	14.0 ±0.1	32.2 ±0.2	10.8 ±0.4	25.0 ±0.2	27.4 ±0.4	65.0 ±0.8
	RAI-FW (χ^2)	31.6 ±0.1	**32.5** ±0.5	13.9 ±0.1	32.6 ±0.3	10.9 ±0.4	**23.4** ±0.2	27.5 ±0.1	**63.8** ±0.6
DA	Online GDRO	31.7 ±0.2	32.2 ±0.3	13.1 ±0.2	26.6 ±0.2	11.2 ±0.1	21.7 ±0.3	27.3 ±0.1	57.0 ±0.5
	RAI-GA (Group)	32.0 ±0.1	32.7 ±0.1	13.0 ±0.3	27.3 ±0.4	11.5 ±0.1	22.4 ±0.2	27.4 ±0.2	56.6 ±1.1
	RAI-FW (Group)	32.1 ±0.2	32.3 ±0.2	13.0 ±0.2	**26.0** ±0.1	11.4 ±0.3	**20.3** ±0.1	27.9 ±0.2	**52.9** ±0.9
PDA	Online GDRO	31.5 ±0.1	32.7 ±0.2	13.4 ±0.1	32.2 ±0.2	11.3 ±0.2	25.2 ±0.1	27.7 ±0.2	64.0 ±0.8
	RAI-GA (Group $\cap$ χ^2)	31.4 ±0.4	32.9 ±0.2	13.0 ±0.3	30.1 ±0.1	10.8 ±0.2	**23.7** ±0.2	27.5 ±0.1	62.5 ±0.6
	RAI-FW (Group $\cap$ χ^2)	31.8 ±0.2	**32.3** ±0.1	13.5 ±0.3	**29.4** ±0.3	11.2 ±0.4	24.0 ±0.2	27.9 ±0.3	**58.9** ±0.7

Thank You!