To Pool or Not To Pool: Analyzing the Regularizing Effects of Group-Fair Training on Shared Models

Cyrus Cousins cbcousins@umass.edu

I. Elizabeth Kumar iekumar@brown.edu

Suresh Venkatasubramanian suresh@brown.edu

Fairness and Overfitting

Given per-group samples $(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{z}_i \in (\mathcal{X} \times \mathcal{Y})^{\boldsymbol{m}_i}$ from $\mathcal{D}_i^{\boldsymbol{m}_i}$, assume hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$, loss function $\ell : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}$

- Subgroups have potentially different distributions $\mathcal{D}_{1:g}$
- Worst-case generalization error can be analyzed on a per-group basis, but:
  - There may **too few samples** to compute bounds for minority groups
  - Data from large groups may **regularize overfitting** to small groups
- Our goal: Use **majority data** to bound **minority overfitting**

Setting: Malfare-based Fair Machine Learning

Empirical per-group risk:

$$\hat{\mathrm{R}}_i(h, \boldsymbol{z}_i) \doteq \frac{1}{\boldsymbol{m}_i} \sum_{j=1}^{\boldsymbol{m}_i} \ell(h(\boldsymbol{x}_j), \boldsymbol{y}_j)$$

Choose a malfare function such as a *power mean:*

$$\mathrm{M}_p(i \to \mathcal{S}_i; \boldsymbol{w}) \doteq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i^p}$$

$p = 1$: $\boldsymbol{w}$-weighted risk minimization
$p \to \infty$: *minimax fair learning*

Our problem is *empirical malfare minimization:*

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\left(i \to \hat{\mathrm{R}}_i(h); \boldsymbol{w}\right)$$

Rademacher Averages

- **Rademacher averages** $\mathfrak{R}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \mathcal{D}_i)$ bound risk *generalization gap*
  - Suppose range r loss
  - **Supremum Deviation (SD) Bound:** With probability at least $1 - \delta$:

$$\forall i \in 1, \ldots g : \sup_{h \in \mathcal{H}} \left| \mathrm{R}_i(h) - \hat{\mathrm{R}}_i(h) \right| \leq \boldsymbol{\varepsilon}_i = 2\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + r\sqrt{\frac{\ln \frac{g}{\delta}}{2\boldsymbol{m}_i}}$$

- Can generalize this result to power-mean malfare

$$\sup_{h \in \mathcal{H}} \left| \mathrm{M}_p\left(i \mapsto \mathrm{R}_i(h); \boldsymbol{w}\right) \leq \mathrm{M}_p\left(i \mapsto \hat{\mathrm{R}}_i(h); \boldsymbol{w}\right) \right| \leq \max_{i \in 1, \ldots, g} \boldsymbol{\varepsilon}_i$$

Theoretical Restricted Hypothesis Classes

Let $\boldsymbol{\varepsilon}_i \doteq r\sqrt{\frac{\ln \frac{1}{\delta}}{2\boldsymbol{m}_i}}$ and $\boldsymbol{\eta}_i \doteq 2\mathfrak{R}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \mathcal{D}_i) + \boldsymbol{\varepsilon}_i$
We (pessimistically) upper-bound the objective value (w.h.p.)

$$\inf_{h' \in \mathcal{H}} \mathrm{M}\big(j \mapsto \hat{\mathrm{R}}(h', \boldsymbol{z}_j); \boldsymbol{w}\big) \leq \inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h', \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \end{cases}\right)$$

and (optimistically) lower-bound the empirical malfare of all $h \in \mathcal{H}$ (w.h.p.)

$$\mathrm{M}\big(j \mapsto \hat{\mathrm{R}}(h, \boldsymbol{z}_j); \boldsymbol{w}\big) \geq \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h, \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h, \mathcal{D}_i) - \boldsymbol{\eta}_i \end{cases}\right)$$

Set $\mathcal{H}_i^* \doteq \{h \in \mathcal{H}\}$, where

$$\mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h, \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h, \mathcal{D}_i) - \boldsymbol{\eta}_i \end{cases}\right) \leq \inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h', \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \end{cases}\right)$$

Theorem 1. *Assume as above; the following then hold:*
1. With probability at least $1 - 2\delta$ over choice of $\boldsymbol{z}_i$, it holds that $\hat{h} \in \mathcal{H}_i^*$.
2. With probability at least $1 - 4\delta$ over choice of $\boldsymbol{z}_i$,

$$\left| \mathrm{R}(\hat{h}, \mathcal{D}_i) - \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_i) \right| \leq 2\mathfrak{R}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}_i^*, \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \ .$$

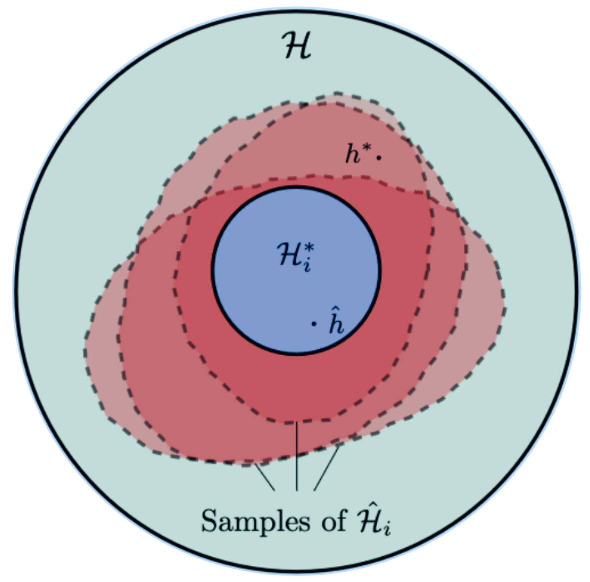
Empirical Restricted Hypothesis Classes

Take $\hat{\boldsymbol{\eta}}_i \doteq 2\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i$
Set $\hat{\mathcal{H}}_i \doteq \{h \in \mathcal{H}\}$, where

$$\mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h, \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h, \boldsymbol{z}_i) - \hat{\boldsymbol{\eta}}_i \end{cases}\right) \leq \inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h', \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \end{cases}\right)$$

Theorem 1. *Assume as above; the following then hold:*
1. With probability at least $1 - 4\delta$ over choice of $\boldsymbol{z}_i$, it holds that $\hat{h} \in \mathcal{H}_i^* \subseteq \hat{\mathcal{H}}_i$.
2. With probability at least $1 - 6\delta$, it holds that

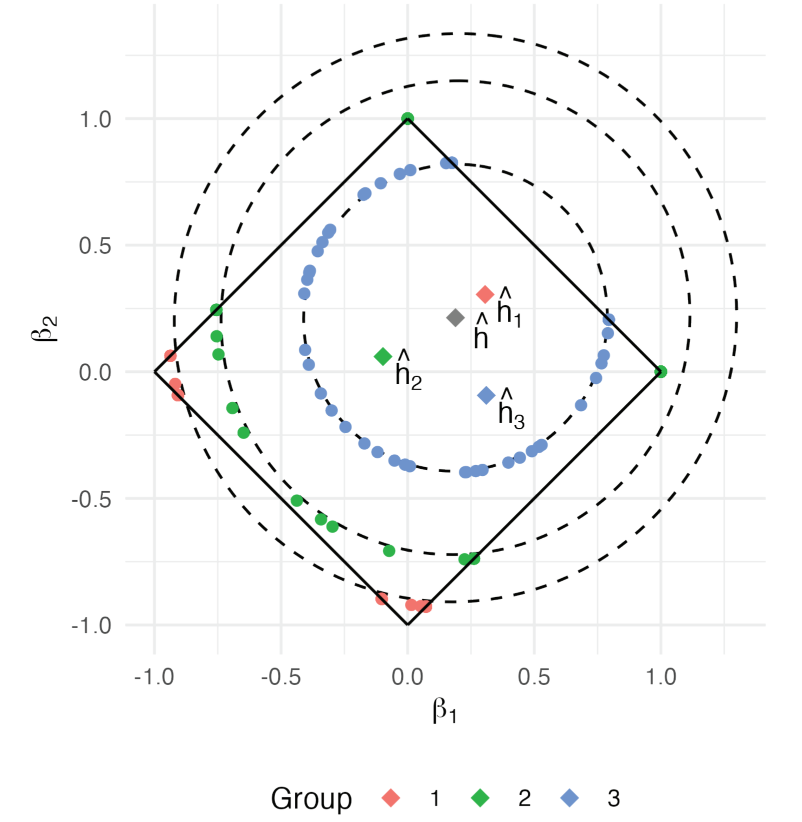
$$\left| \mathrm{R}(\hat{h}, \mathcal{D}_i) - \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_i) \right| \leq 2\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \hat{\mathcal{H}}_i, \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \ .$$



Visualization of unrestricted class $\mathcal{H}$,
theoretical restricted class $\mathcal{H}_i^*$,
and samples of empirical restricted class $\hat{\mathcal{H}}_i$ (varying $\boldsymbol{z}_i$).

Example with Linear Regression

- Take linear hypothesis class
  $\boldsymbol{B} \doteq \{\boldsymbol{\beta} \in \mathbb{R}^2 : \|\boldsymbol{\beta}\|_1 \leq 1\}$
- Sample $(\boldsymbol{x}_i, y_i)$ as
  $\boldsymbol{x}_i \sim \mathrm{Unif}([-1, 1]^2)$
  $y_i = \boldsymbol{x}_i \cdot \boldsymbol{\beta}_i + \mathrm{Unif}([-1, 1])$
- Plot values of $\boldsymbol{\beta}$ which realize each supremum in the empirical Rademacher average for some Rademacher sample $\boldsymbol{\sigma}_k$
- Points lie on either (the corner of) the ℓ_1 constraint boundary of $\boldsymbol{B}$ or the restricted hypothesis constraint boundary of $\hat{\mathcal{H}}_i$



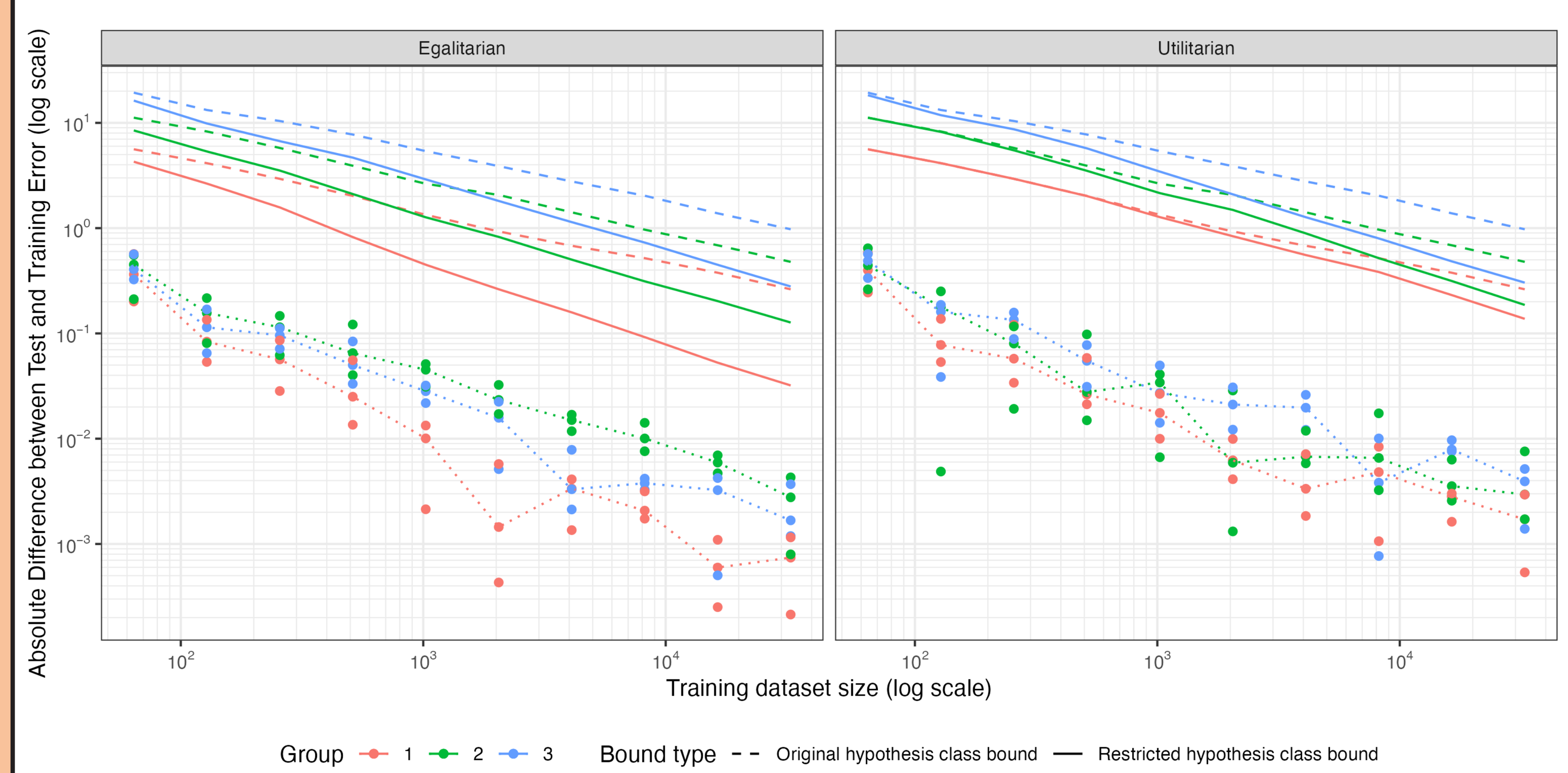
Experiments with Logistic Regression

$\mathcal{X} = [-1, 1]^{15}$
$\mathcal{Y} = \pm 1$
$\boldsymbol{B} \doteq \{\boldsymbol{\beta} \in \mathbb{R}^{15} \mid \|\boldsymbol{\beta}\|_1 \leq 15\}$
$\boldsymbol{x} \sim \mathrm{Unif}(\mathcal{X})$
$\mathbb{P}(y = 1) = \mathrm{logistic}(\boldsymbol{x} \cdot \boldsymbol{\beta}_i + \xi)$

$$\hat{h} \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(h, \boldsymbol{z}_i)\big)$$

$$\hat{\mathrm{R}}(h, \boldsymbol{z}_i) = \frac{1}{\boldsymbol{m}_i} \sum_{j=1}^{\boldsymbol{m}_i} \ln\big(1 + \exp(\boldsymbol{y}_{i,j} \cdot h(\boldsymbol{x}_{i,j}))\big)$$

	Data proportion	True parameters
Group 1	75%	$\beta_i = 0.3$
Group 2	20%	$\beta_i = 0.1$
Group 3	5%	$\beta_i = 0.2$



*This research was made possible in part by the generous support of the **Ford Foundation** and the **MacArthur Foundation**. Cyrus Cousins also wishes to acknowledge the **Center for Data Science at the University of Massachusetts Amherst**, where part of this work was conducted under a postdoctoral fellowship.*