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

Cyrus Cousins
Indra Elizabeth Kumar
Suresh Venkatasubramanian

AIStats 2024

cyruscousins.online/projects/fairlocalization

Regularization and Fair Learning

- So far, we have analyzed learning over *all of* Θ
  - Learning is a random process, but usually we learn $\hat{\theta} \approx \theta^*$
- Group fair learning: Data from other groups have a *regularizing effect*
  - Do small groups benefit from large group data?
  - Can we mathematically quantify the benefit of this regularization?
- For each group i: Analyze learning from $\boldsymbol{z}_i$, conditioned on $\boldsymbol{z}_{j \neq i}$
  - W.h.p. over $\boldsymbol{z}_i$: $\hat{\theta} \approx \underset{\theta \in \theta}{\operatorname{argmin}} \mathrm{M} \left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}_j(\theta) \\ j = i & \mathrm{R}_i(\theta) \end{cases} \right)$
  - Learning effectively occurs over a localized region
- Double-randomization technique [Cousins, Kumar, & Venkatasubramanian, AIStats 2024]
  - Construct theoretical class using $\hat{\mathrm{R}}_{j \neq i}(\theta)$ and $\mathrm{R}_i(\theta)$
  - Bound theoretical class with empirical class using $\hat{\mathrm{R}}_i(\theta)$

$(\boldsymbol{x}_1, \boldsymbol{y}_1) \sim \mathcal{D}_1^{m_1}$ $(\boldsymbol{x}_2, \boldsymbol{y}_2) \sim \mathcal{D}_2^{m_2}$ $(\boldsymbol{x}_3, \boldsymbol{y}_3) \sim \mathcal{D}_3^{m_3}$ $(\boldsymbol{x}_4, \boldsymbol{y}_4) \sim \mathcal{D}_4^{m_4}$

Improved Bounds with Localization

- Let's analyze fair learning from the perspective of group i
  - Training sample $\boldsymbol{z}_i$ is random, but we have $\boldsymbol{z}_j$ for $j \neq i$
  - Observed data $\boldsymbol{z}_j$ and distribution $\mathcal{D}_i$ determine $\hat{\theta}$ conditional distribution
- Define the *localized hypothesis class*:

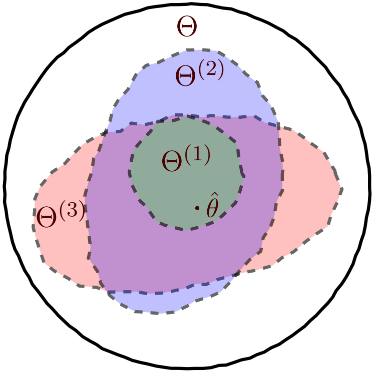
$$\Theta^{(i)} \doteq \left\{ \theta \in \Theta \;\middle|\; \underbrace{\mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}_j(\theta) \\ j = i & \hat{\mathrm{R}}_i(\theta) - 2\hat{\boldsymbol{\eta}}_i \end{cases} \right)}_{\text{Optimistic malfare estimate}} \leq \underbrace{\inf_{\theta' \in \Theta} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}_i(\theta') \\ j = i & \hat{\mathrm{R}}_j(\theta') + 2\boldsymbol{\varepsilon}_i \end{cases} \right)}_{\text{Pessimistic minimal malfare estimate}} \right\}$$

  - $\boldsymbol{\varepsilon}_i = \sqrt{\frac{\ln \frac{6}{\delta}}{2\boldsymbol{m}_i}}$
  - $\hat{\boldsymbol{\eta}}_i = 2\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \Theta, \boldsymbol{z}_i) + 2\sqrt{\frac{\ln \frac{6}{\delta}}{2\boldsymbol{m}_i}}$
- Learning *effectively occurs* over $\Theta^{(i)}$, not Θ

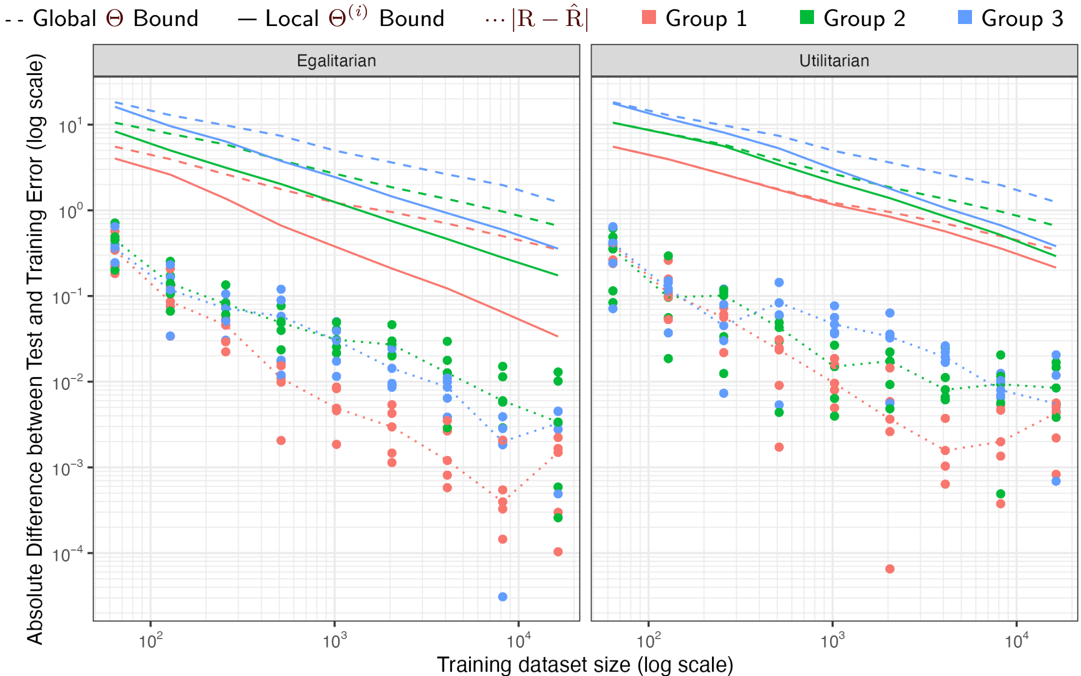
$$\mathbb{P}\left(\hat{\theta} \notin \Theta^{(i)} \right) < \tfrac{4}{6}\delta$$

- Get per-group generalization bounds

$$\mathbb{P}\left(\left| \mathrm{R}(\hat{\theta}, \mathcal{D}_i) - \hat{\mathrm{R}}(\hat{\theta}, \boldsymbol{z}_i) \right| > 2\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \Theta^{(i)}, \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \right) < \delta$$



Synthetic Localized Logistic Regression Experiment



Why Localize?

- **Goal:** Better understanding of overfitting and per-group risk
  - Make better decisions with the data we have
  - Decide where to sample more data
- Global bounds are loose for small groups
  - $\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \Theta, \boldsymbol{z}_i) \in \boldsymbol{\Theta} \frac{1}{\sqrt{\boldsymbol{m}_i}}$ ignores contributions of other groups
  - Usually $\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \Theta^{(i)}, \boldsymbol{z}_i) \ll \hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \Theta, \boldsymbol{z}_i)$
- Localization yields sharper generalization bounds
  - Use *majority data* to bound *minority overfitting*
  - Data from large groups *regularizes overfitting* to small groups
- Reveals an inherent tradeoff

$$\Theta^{(i)} \doteq \left\{ \theta \in \Theta \,\middle|\, \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}_j(\theta) \\ j = i & \hat{\mathrm{R}}_i(\theta) - 2\hat{\boldsymbol{\eta}}_i \end{cases} \right) \leq \inf_{\theta' \in \Theta} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}_i(\theta') \\ j = i & \hat{\mathrm{R}}_j(\theta') + 2\boldsymbol{\varepsilon}_i \end{cases} \right) \right\}$$

  - **Utilitarian:** Relatively insensitive to minority groups
  - **Egalitarian:** Highly sensitive to minority groups
  - Localized bounds depend on objective sensitivity to each group's risk!
  - Asymptotically measured by *malfare gradient* $\boldsymbol{\lambda} \doteq \nabla_{\mathrm{R}} \mathrm{M}\big(\mathrm{R}(\theta^*)\big)$