Maximum Weighted Loss Discrepancy

Fereshte Khani, Aditi Raghunathan, Percy Liang

Stanford University

Motivation

Focusing only on minimizing the average loss over a population might lead to large discrepancies between the losses across groups within the population. In this work, we study this inequality.

Setup

- individuals: $z = (x, y)$, underlying distribution: $p^\star$
- group: a measurable function $g : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$, all groups: $\mathcal{G}$
- predictor: $h : \mathcal{X} \to \mathcal{Y}$
- bounded measurable loss function : $\ell(h, z)$
- population loss: $\mathbb{E}[\ell]$, group loss: $\mathbb{E}[\ell \mid g = 1]$

Maximum Weighted Loss Discrepancy (MWLD)

$$\mathrm{MWLD}\,(w, \ell, h) \stackrel{\text{def}}{=} \sup_{g \in \mathcal{G}} w(g)\,|\mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell]|$$

Connection to Group Fairness

MWLD can be viewed as an upper bound for the loss of any group:

$$\mathbb{E}[\ell \mid g = 1] \le \mathbb{E}[\ell] + \frac{\mathrm{MWLD}(w, \ell, h)}{w(g)}.$$

- Existing statistical notions of fairness can be viewed as enforcing small $\mathrm{MWLD}(w)$ for different weighting functions.
- **Equalized opportunity:** weighting function is 1 on sensitive groups (e.g., different races) and 0 on all other groups.

Connection to AI safety (Robustness)

MWLD can be viewed as an upper bound for the loss on a population with shifted demographics:

$$\mathbb{E}_{z \sim q}[\ell] \le \mathbb{E}_{z \sim p^\star}[\ell] + \mathrm{MWLD}(w, \ell, h)$$

where $q(\cdot) \stackrel{\text{def}}{=} w(g)p^\star(\cdot \mid g = 1) + (1 - w(g))p^\star(\cdot \mid g = 0)$.

- This is similar in spirit to distributionally robust optimization (DRO) using a max-norm metric.
- The difference is that the mixture coefficient is group-dependent.

Questions?

- For what weighting functions we can estimate $\mathrm{MWLD}(w)$?

$$\forall \epsilon, \delta \in (0, \tfrac{1}{2}) : \quad \mathbb{P}\left[\left|\mathrm{MWLD}(w) - \widehat{\mathrm{MWLD}}_n(w)\right| \ge \epsilon\right] \le \delta$$

- When can we compute $\widehat{\mathrm{MWLD}}_n(w)$ efficiently?
- Is there any connection between $\mathrm{MWLD}(w)$ and other notions?

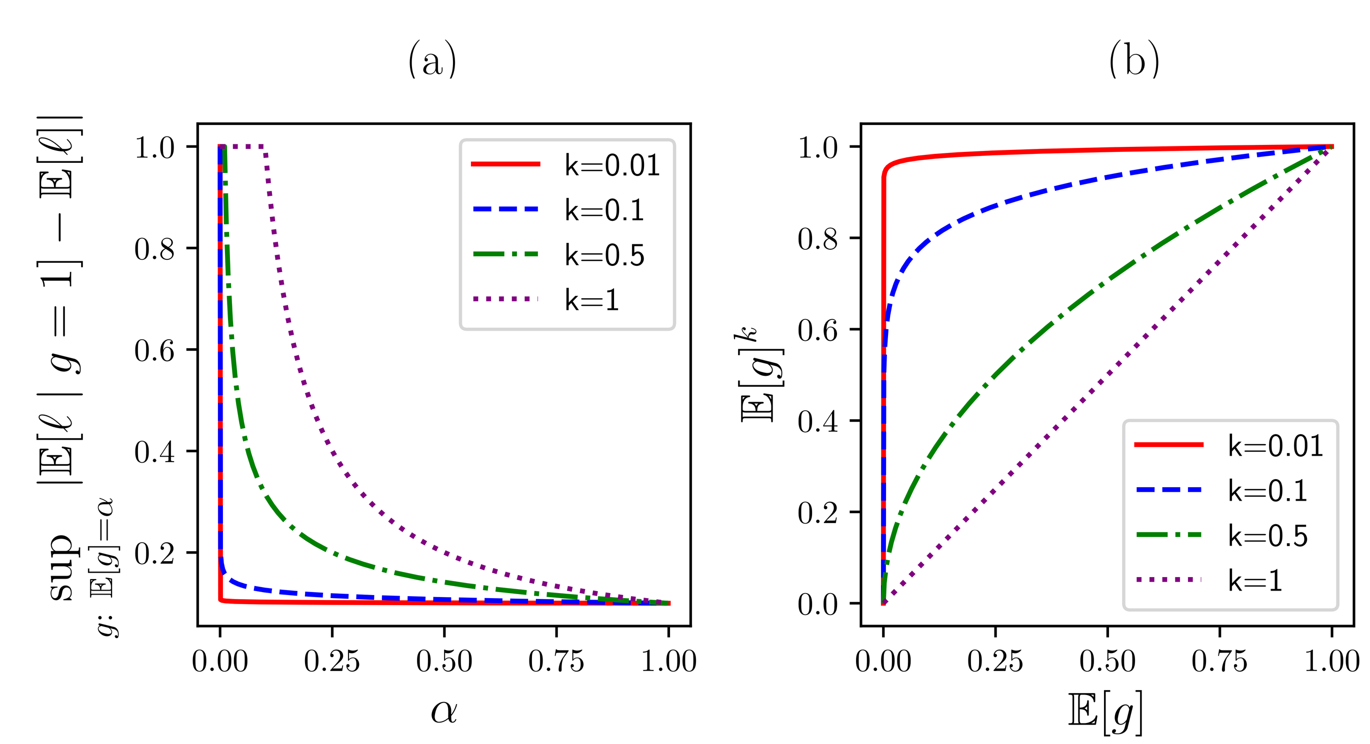
Proposition

Let $w^0(g) \stackrel{\text{def}}{=} \mathbb{I}[\mathbb{E}[g] > 0]$. For non-degenerate (ℓ, h), it is impossible to estimate $\mathrm{MWLD}(w^0, \ell, h)$.

Theorem

For $k \in (0, 1]$, let $w^k \stackrel{\text{def}}{=} \mathbb{E}[g]^k$. Given $n \ge \frac{C \log(1/\delta)}{\epsilon^{2+\frac{2}{k}}}$ i.i.d. samples from $p^\star$, we can estimate $\mathrm{MWLD}(w^k)$ efficiently.

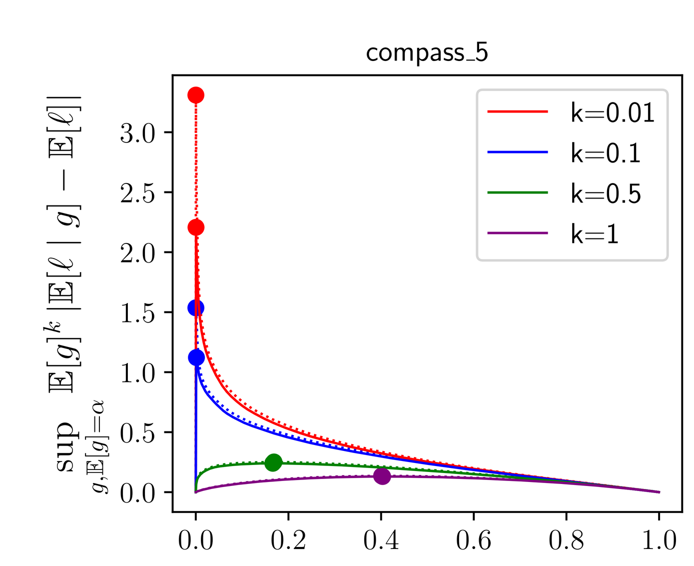
$w^k(g) = \mathbb{E}[g]^k$



The parameter k governs variation of (a) upper bound on loss discrepancy, (b) up-weighting factor across group sizes.

Estimating $\mathrm{MWLD}(w^k)$ on real world datasets

- Dots denote $\widehat{\mathrm{MWLD}}(w^k)$.
- For smaller k, there is a bigger gap between values of $\widehat{\mathrm{MWLD}}(w^k)$ corresponding to train (dashed lines) and test (solid lines) set.



Loss Variance

- Average individual discrepancy

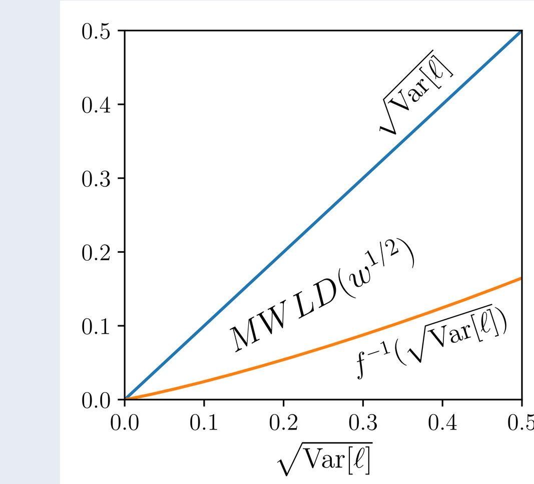
$$\mathrm{Var}[\ell] = \mathbb{E}\left[(\ell(z) - \mathbb{E}[\ell])^2\right]$$

- Generalization bounds

$$|\hat{\mathbb{E}}[\ell] - \mathbb{E}[\ell]| \le C_1\sqrt{\frac{\hat{\mathrm{Var}}[\ell]}{n}} + \frac{C_2}{n}$$

Proposition

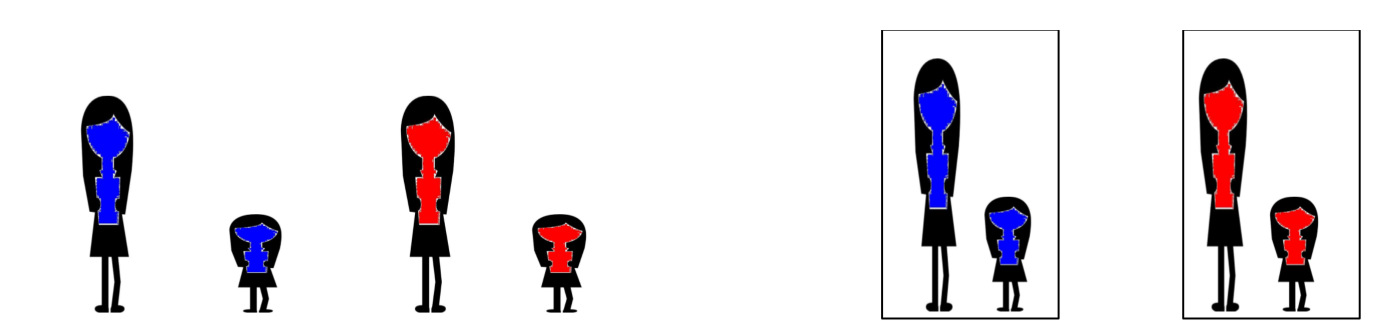
For $f(x) = x\sqrt{2 - 4\ln(x)}$:



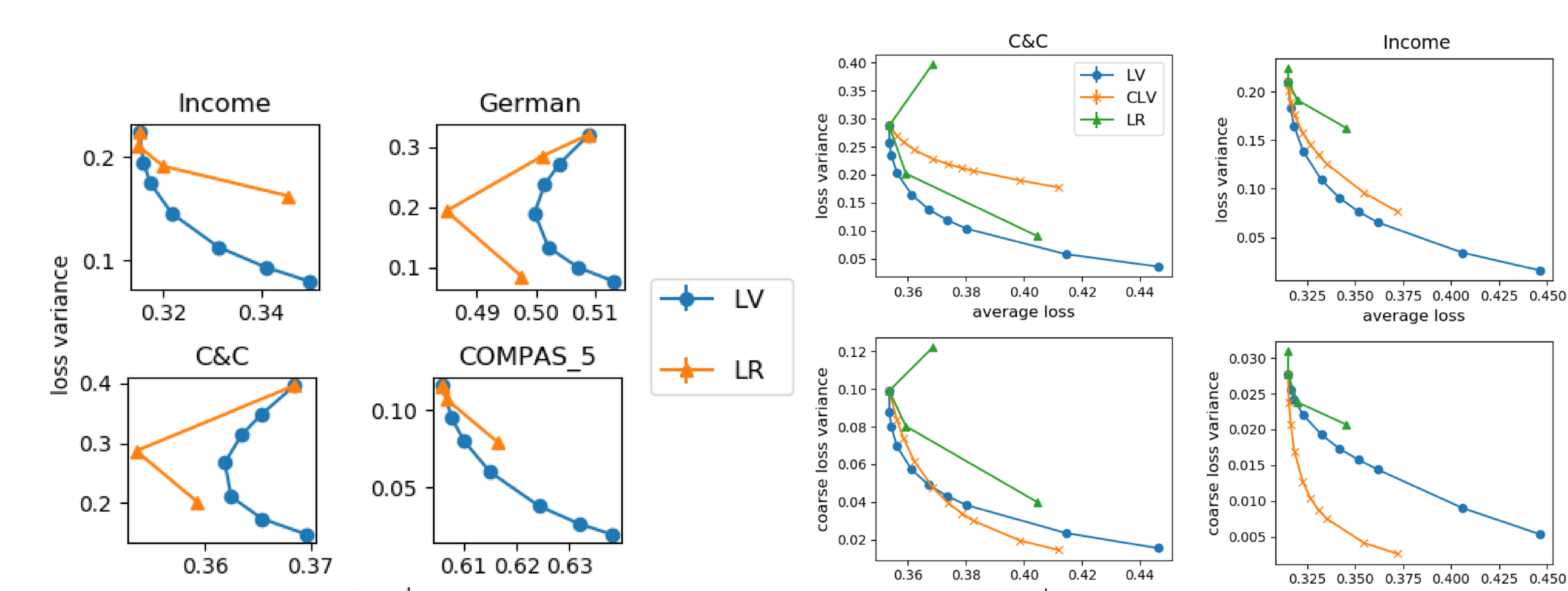
Coarse Loss Variance

Let A denote the sensitive attributes; e.g., $A = [\text{race}, \text{gender}, \ldots]$.

$$\mathrm{Var}\left[\mathbb{E}[\ell \mid A]\right] = \mathbb{E}\left[(\mathbb{E}[\ell \mid A] - \mathbb{E}[\ell])^2\right].$$



(Coarse) Loss Variance Regularization



$$\mathcal{O}_{LR} \stackrel{\text{def}}{=} \hat{\mathbb{E}}[\ell] + \eta\|\theta\|_2^2$$

$$\mathcal{O}_{LV} \stackrel{\text{def}}{=} \mathcal{O}_{LR} + \lambda\hat{\mathbb{E}}[\hat{\mathrm{Var}}[\ell \mid y]] \qquad \mathcal{O}_{CLV} \stackrel{\text{def}}{=} \mathcal{O}_{LR} + \lambda\hat{\mathbb{E}}[\hat{\mathrm{Var}}[\hat{\mathbb{E}}[\ell \mid A, y] \mid y]]$$

- We halve the loss variance with only a small drop in the average loss.
- In some cases, using loss variance as a regularizer simultaneously reduces the classification loss and loss variance.