Fairness in Multi-Task Learning

via Wasserstein Barycenters

François HU
Université de Montréal, Department of Mathematics and Statistics

Joint work with:

Philipp Ratz and **Arthur Charpentier**
Université du Québec à Montréal

ECML-PKDD 2023

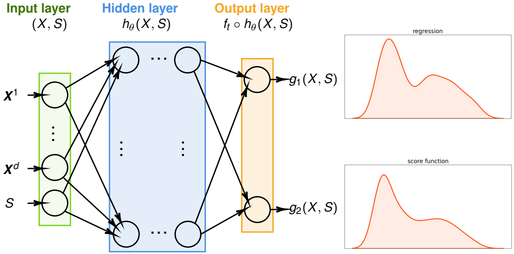
September 21, 2023

Outline

Multi-Task Learning

- Suppose that you have a small data set, but for a **related problem** you have much more data at hand. **Multi-task learning (MTL)** can help exploit these similarities.
- Examples: self-driving car, facial recognition, ...

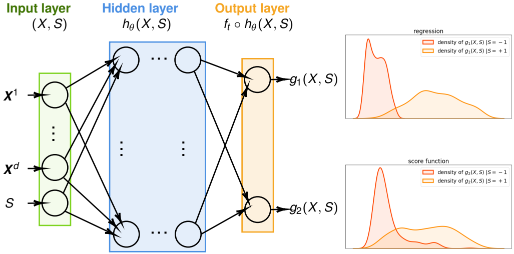
Data: $(\underbrace{\textit{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\textit{tasks}}_{\boldsymbol{Y}}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y} \subset \mathbb{R}^d \times \{-1, 1\} \times \mathbb{R}^2$



Multi-Task Learning

- Suppose that you have a small data set, but for a **related problem** you have much more data at hand. **Multi-task learning (MTL)** can help exploit these similarities.
- Examples: self-driving car, facial recognition, ...

Data: $(\underbrace{\textit{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\textit{tasks}}_{\boldsymbol{Y}}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y} \subset \mathbb{R}^d \times \{-1, 1\} \times \mathbb{R}^2$

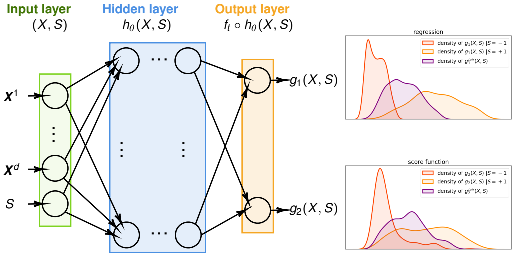


- **Fairness challenge:** Achieving fairness in one task does not necessarily extend fairness to others, despite equal representation $\longrightarrow$ Surprisingly, limited research on fairness in MTL.

Multi-Task Learning

- Suppose that you have a small data set, but for a **related problem** you have much more data at hand. **Multi-task learning (MTL)** can help exploit these similarities.
- Examples: self-driving car, facial recognition, ...

Data: $(\underbrace{\textit{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\textit{tasks}}_{\boldsymbol{Y}}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y} \subset \mathbb{R}^d \times \{-1, 1\} \times \mathbb{R}^2$



- **Fairness challenge:** Achieving fairness in one task does not necessarily extend fairness to others, despite equal representation $\longrightarrow$ Surprisingly, limited research on fairness in MTL.
- **Objective:** Leverage optimal transport theory to ensure fairness in MTL settings while minimizing the impact on predictive performance.

Risk and unfairness measure

- A predictor $g_t : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}_t \subset \mathbb{R}$ is called fair under **Demographic Parity** if

$$\sup_{u \in \mathcal{Y}_t} \left| \underbrace{\mathbb{P}(g_t(\boldsymbol{X}, S) \leq u \mid S = 1)}_{F_{g_t|1}(u)} - \underbrace{\mathbb{P}(g_t(\boldsymbol{X}, S) \leq u \mid S = -1)}_{F_{g_t|-1}(u)} \right| = 0 \ .$$

- The **unfairness** and the (squared) **risk** of g_t are resp. quantified by

$$\mathcal{U}(g_t) := \sup_{u \in \mathcal{Y}_t} \left| F_{g_t|1}(u) - F_{g_t|-1}(u) \right| \quad \text{and} \quad \mathcal{R}(g_t) := \mathbb{E}\left[(Y_t - g_t(\boldsymbol{X}, S))^2\right] \ .$$

Goal: Minimise risk measure under DP-fairness constraint

Risk and unfairness measure

- A predictor $g_t : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}_t \subset \mathbb{R}$ is called fair under **Demographic Parity** if

$$\sup_{u \in \mathcal{Y}_t} \mid \underbrace{\mathbb{P}(g_t(\boldsymbol{X}, S) \leq u \mid S = 1)}_{F_{g_t|1}(u)} - \underbrace{\mathbb{P}(g_t(\boldsymbol{X}, S) \leq u \mid S = -1)}_{F_{g_t|-1}(u)} \mid = 0 \ .$$

- The **unfairness** and the (squared) **risk** of g_t are resp. quantified by

$$\mathcal{U}(g_t) := \sup_{u \in \mathcal{Y}_t} \left| F_{g_t|1}(u) - F_{g_t|-1}(u) \right| \quad \text{and} \quad \mathcal{R}(g_t) := \mathbb{E}\left[(Y_t - g_t(\boldsymbol{X}, S))^2\right] \ .$$

- ⚠ In **MTL setting** $\boldsymbol{g} = (g_1, g_2)$,

$$\mathcal{R}\ (\boldsymbol{g}) := \mathbb{E}\left[\sum_{t=1,2} (Y_t - g_t(\boldsymbol{X}, S))^2\right] \quad \text{with} \quad g_t(\boldsymbol{X}, S) := f_t \circ \theta(\boldsymbol{X}, S) \ .$$

Goal: Minimise risk measure under DP-fairness constraint

Risk and unfairness measure

- A predictor $g_t : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}_t \subset \mathbb{R}$ is called fair under **Demographic Parity** if

$$\sup_{u \in \mathcal{Y}_t} \left| \underbrace{\mathbb{P}(g_t(\boldsymbol{X}, S) \leq u \mid S = 1)}_{F_{g_t|1}(u)} - \underbrace{\mathbb{P}(g_t(\boldsymbol{X}, S) \leq u \mid S = -1)}_{F_{g_t|-1}(u)} \right| = 0 \ .$$

- The **unfairness** and the (squared) **risk** of g_t are resp. quantified by

$$\mathcal{U}(g_t) := \sup_{u \in \mathcal{Y}_t} \left| F_{g_t|1}(u) - F_{g_t|-1}(u) \right| \quad \text{and} \quad \mathcal{R}(g_t) := \mathbb{E}\left[(Y_t - g_t(\boldsymbol{X}, S))^2\right] \ .$$

- ⚠ In **MTL setting** $\boldsymbol{g} = (g_1, g_2)$, with weight $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$,

$$\mathcal{R}_{\boldsymbol{\lambda}}(\boldsymbol{g}) := \mathbb{E}\left[\sum_{t=1,2} \lambda_t \cdot (Y_t - g_t(\boldsymbol{X}, S))^2\right] \quad \text{with} \quad g_t(\boldsymbol{X}, S) := f_t \circ \theta(\boldsymbol{X}, S) \ .$$

Goal: Minimise risk measure under DP-fairness constraint

$$\boxed{\min_{g_1 \text{ and } g_2 \text{ DP-fair}} \mathcal{R}_{\boldsymbol{\lambda}}(\boldsymbol{g})} \qquad (\textbf{objective})$$

Background on Wasserstein Barycenters

We consider two probability measures, ν_1 and ν_2. We define **distance function** between ν_1 and ν_2:

Definition (Wasserstein distance)

The squared Wasserstein distance between ν_1 and ν_2 is defined as

$$\mathcal{W}_2^2(\nu_1, \nu_2) = \inf_{\pi \in \Pi(\nu_1, \nu_2)} \mathbb{E}_{(Z_1, Z_2)\sim\pi} \left(Z_2 - Z_1\right)^2 \quad ,$$

where $\Pi(\nu_1, \nu_2)$ is the set of distributions on $\mathcal{Y} \times \mathcal{Y}$ having ν_1 and ν_2 as marginals.

The Wasserstein barycenter finds a **representative distribution** that lies between multiple given distributions in the Wasserstein space. It is defined for a family of K measures $(\nu_1, \ldots, \nu_K)$ in $\mathcal{V}$ and some positive weights $(w_1, \ldots, w_K) \in \mathbb{R}_+^K$.

Definition (Wasserstein Barycenters)

The Wasserstein barycenter, denoted as $\mathrm{Bar}\left\{(w_k, \nu_k)_{k=1}^K\right\}$ is the minimiser

$$\mathrm{Bar}(w_k, \nu_k)_{k=1}^K = \operatorname*{argmin}_{\nu} \sum_{k=1}^{K} w_k \cdot \mathcal{W}_2^2\left(\nu_k, \nu\right) \quad .$$

The barycenter exists and is unique if one of ν_k admits a density wrt the Lebesgue measure [AC11].

Outline

Optimal prediction under DP

$g^*_{t,\boldsymbol{\lambda}}$: optimal predictor **without** fairness constr. $g^{*(fair)}_{t,\boldsymbol{\lambda}}$: optimal predictor **with** fairness constr.

Continuity assumpt. For any $(s, t, \boldsymbol{\lambda}) \in \mathcal{S} \times \mathcal{T} \times \Lambda$, assume that $\nu_{g^*_{t,\boldsymbol{\lambda}}|s}$ has a density function. This is equivalent to assuming that the mapping $u \mapsto F_{g^*_{t,\boldsymbol{\lambda}}|s}(u)$ is continuous.

Theorem (Optimal fair predictions, adapted from [CDH+20, GLR20])

Assuming continuity. Let $\pi_s := \mathbb{P}(S = s)$. Then,

1. *A representation function $\theta^{*(fair)}_{\boldsymbol{\lambda}}$ satisfies (**objective**), iff, for each task t,*

$$\nu_{f_t \circ \theta^{*(fair)}_{\boldsymbol{\lambda}}} = \mathrm{Bar}(\pi_s, \nu_{g^*_{t,\boldsymbol{\lambda}}|s})_{s\in\mathcal{S}} = \operatorname*{argmin}_{\nu} \sum_{s\in\mathcal{S}} \pi_s \mathcal{W}_2^2(\nu_{g^*_{t,\boldsymbol{\lambda}}|s}, \nu) \ .$$

2. *Additionally, the optimal fair predictor $g^{*(fair)}_{t,\boldsymbol{\lambda}}(\cdot) = f_t \circ \theta^{*(fair)}_{\boldsymbol{\lambda}}(\cdot)$ can be rewritten as*

$$\boxed{g^{*(fair)}_{t,\boldsymbol{\lambda}}(\boldsymbol{x}, s) = \sum_{s'\in\mathcal{S}} \pi_{s'} Q_{g^*_{t,\boldsymbol{\lambda}}|s'} \circ F_{g^*_{t,\boldsymbol{\lambda}}|s}\left(g^*_{t,\boldsymbol{\lambda}}(\boldsymbol{x}, s)\right)}, \quad (\boldsymbol{x}, s) \in \mathcal{X} \times \mathcal{S} \ .$$

Post-processing estimator

Theoretical fair solution:

$$g^{*(\text{fair})}_{t,\boldsymbol{\lambda}}(\boldsymbol{x}, s) = \sum_{s' \in \mathcal{S}} \pi_{s'} Q_{g^*_{t,\boldsymbol{\lambda}}|s'} \circ F_{g^*_{t,\boldsymbol{\lambda}}|s} \left(g^*_{t,\boldsymbol{\lambda}}(\boldsymbol{x}, s) \right) , \quad (\boldsymbol{x}, s) \in \mathcal{X} \times \mathcal{S} \ .$$

Empirical fair solution

- **Plug-in**: Estimation based on independent samples of $(\boldsymbol{X}, S, Y_1, Y_2)$.
  - **Labeled data:** $\mathcal{D}^{\text{train}}_n = \{(\boldsymbol{X}_i, S_i, Y_{i,1}, Y_{i,2})\}_{i=1}^n$. $\zeta_{i,t}$ uniform perturbation on $[0, u]$ train estimators $\widehat{g}_{1,\boldsymbol{\lambda}}$ and $\widehat{g}_{2,\boldsymbol{\lambda}}$. Continuity assumption satisfied if

    $$\bar{g}_{t,\boldsymbol{\lambda}}(\boldsymbol{X}_i, S_i, \zeta_{i,t}) = \widehat{g}_{t,\boldsymbol{\lambda}}(\boldsymbol{X}_i, S_i) + \zeta_{i,t} \ .$$

  - **Unlabeled data:** $\mathcal{D}^{\text{pool}}_N = \{(\boldsymbol{X}_i, S_i)\}_{i=1}^N$, N *i.i.d.* copies of $(\boldsymbol{X}, S)$. emp. frequencies $(\widehat{\pi}_s)_{s \in \mathcal{S}}$, CDF $\widehat{F}_{\bar{g}_{t,\boldsymbol{\lambda}}|s}$ and quantile function $\widehat{Q}_{\bar{g}_{t,\boldsymbol{\lambda}}|s}$ via $\bar{g}_t$ and $\mathcal{D}^{\text{pool}}_N$.
- (Randomised) **fair estimator:**

$$\boxed{\widehat{g}^{(\text{fair})}_{t,\boldsymbol{\lambda}}(\boldsymbol{x}, s) = \sum_{s' \in \mathcal{S}} \widehat{\pi}_{s'} \widehat{Q}_{\bar{g}_{t,\boldsymbol{\lambda}}|s'} \circ \widehat{F}_{\bar{g}_{t,\boldsymbol{\lambda}}|s} \left(\bar{g}_{t,\boldsymbol{\lambda}}(\boldsymbol{x}, s, \zeta_t) \right)} , \quad (\boldsymbol{x}, s) \in \mathcal{X} \times \mathcal{S} \ .$$

Outline

FOLKTABLES data

- **Data:** The FOLKTABLES dataset [DHMS21] comprises various binary prediction tasks derived from a substantial US Census data corpus encompassing income, employment, health, transportation, and housing (58,650 observations).
- **Tasks:** **Mobility** (Binary) and **Income** (Regression) using a feature set of 19 attributes, with **gender** as the binary sensitive variable.
- **Empirical MTL:** We use the **"You Only Train Once"** (YOTO) approach of [DD20]. The model is only trained once for different $\boldsymbol{\lambda}$ values by conditioning the parameters of the neural network directly on the task weights $\boldsymbol{\lambda}$. The idea is that different values for $\boldsymbol{\lambda}$ are sampled from a distribution and included directly in the estimation process.

Model / Data	MTL		MTL, Post-processed		STL	
	Performance	Unfairness	Performance	Unfairness	Performance	Unfairness
regression - all data	0.548 ± 0.02	0.109 ± 0.01	0.558 ± 0.02	0.018 ± 0.00	0.559 ± 0.02	0.107 ± 0.01
regression - 25% missing	0.558 ± 0.02	0.109 ± 0.02	0.572 ± 0.02	0.018 ± 0.00	0.570 ± 0.02	0.105 ± 0.02
regression - 50% missing	0.577 ± 0.02	0.109 ± 0.02	0.593 ± 0.03	0.018 ± 0.01	0.587 ± 0.02	0.099 ± 0.01
regression - 75% missing	0.612 ± 0.05	0.101 ± 0.02	0.627 ± 0.06	0.019 ± 0.01	0.632 ± 0.04	0.098 ± 0.01
regression - 95% missing	0.678 ± 0.05	0.105 ± 0.02	0.687 ± 0.05	0.018 ± 0.01	0.738 ± 0.06	0.108 ± 0.03
classification - all data	0.576 ± 0.01	0.080 ± 0.03	0.577 ± 0.01	0.018 ± 0.01	0.640 ± 0.03	0.042 ± 0.02

Table 1: Performance and unfairness for MTL and Single Task Learning (STL) models on the FOLKTABLES data. Each model was also post-processed and evaluated on performance and unfairness.

COMPAS data

- **Data:** The COMPAS dataset [LAKM16], used to assess the reoffending likelihood of criminal defendants, exhibits bias in favor of white defendants. The dataset includes two classification targets (**recidivism** and **violent recidivism**), employing 18 features. We study 6,172 observations with **race** as the sensitive attribute.
- **Empirical MTL:** **YOTO** approach [DD20].

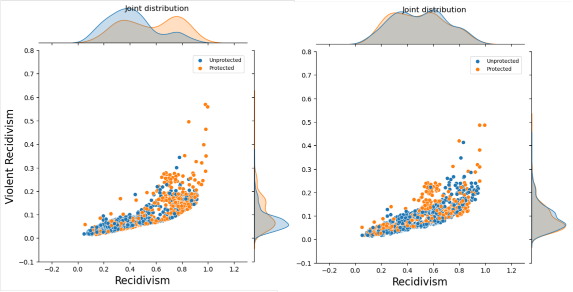


Figure 1: Joint distribution for scores under unconstrained and DP-fair regimes. Color indicates the presence of the sensitive feature. Note that the joint distribution appears more mixed and the marginal distributions overlap in the DP fair case.

Outline

Conclusion

MTL is gaining popularity but poses fairness challenges.

1. We propose an efficient **post-processing method** to incorporate fairness into MTL;
2. Extending this approach to domains like computer vision with pre-trained models (e.g., h_θ or θ) warrants further exploration as Transfer-Multitask-Fair learning;
3. Future work could address **fairness across multiple tasks simultaneously**, but it may require non-trivial solutions due to quantile estimation reliance.

Conclusion

Thank you !

References I

[AC11] M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[CDH+20] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with wasserstein barycenters. In *Advances in Neural Information Processing Systems*, 2020.

[DD20] Alexey Dosovitskiy and Josip Djolonga. You only train once: Loss-conditional training of deep networks. In *International conference on learning representations*, 2020.

[DHMS21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[GLR20] T. Gouic, J.M. Loubes, and P. Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

[LAKM16] Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya Mattu. How we analyzed the compas recidivism algorithm, May 2016.