Algorithmic Fairness

in ~~classification and~~ regression

Evgenii Chzhen

works with C. Denis, M. Hebiri, L. Oneto, M. Pontil, and N. Schreuder

Content

1. Fairness aware learning
2. Regression with Demographic Parity
3. Quantification of risk/fairness trade-off

Fairness aware learning

General setup

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P} \text{ on } \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$$

General setup

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$

Prediction: $f : \mathcal{Z} \to \mathcal{Y}$

- Fairness through awareness: $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$ and $\boldsymbol{Z} = (\boldsymbol{X}, S)$
- Fairness through unawareness: $\mathcal{Z} = \mathcal{X}$ and $\boldsymbol{Z} = \boldsymbol{X}$

General setup

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$

Prediction: $f : \mathcal{Z} \to \mathcal{Y}$

- Fairness through awareness: $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$ and $\boldsymbol{Z} = (\boldsymbol{X}, S)$
- Fairness through unawareness: $\mathcal{Z} = \mathcal{X}$ and $\boldsymbol{Z} = \boldsymbol{X}$

Risk: $f \mapsto \mathcal{R}(f)$

- regression: $\mathcal{R}(f) = \mathbb{E}(Y - f(\boldsymbol{Z}))^2$
- classification: $\mathcal{R}(f) = \mathbb{P}(Y \neq f(\boldsymbol{Z}))$

General setup

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$

Prediction: $f : \mathcal{Z} \to \mathcal{Y}$

- Fairness through awareness: $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$ and $\boldsymbol{Z} = (\boldsymbol{X}, S)$
- Fairness through unawareness: $\mathcal{Z} = \mathcal{X}$ and $\boldsymbol{Z} = \boldsymbol{X}$

Risk: $f \mapsto \mathcal{R}(f)$

- regression: $\mathcal{R}(f) = \mathbb{E}(Y - f(\boldsymbol{Z}))^2$
- classification: $\mathcal{R}(f) = \mathbb{P}(Y \neq f(\boldsymbol{Z}))$

Fairness constraint:

- Demographic Parity (DP): $f(\boldsymbol{Z}) \perp\!\!\!\perp S$
- Equalized Odds: $(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y$
- Group-risk equality: $\mathbb{E}[(Y - f(\boldsymbol{Z}))^2 | S = s] = \mathbb{E}(Y - f(\boldsymbol{Z}))^2$
- many more ... (Barocas, Hardt, and Narayanan, 2018)

Main approaches

Observations: $(\boldsymbol{X}_1, S_1, Y_1), \dots, (\boldsymbol{X}_n, S_n, Y_n) \in \mathcal{X} \times \mathcal{S} \times \mathbb{R}$

Goal: build $\hat{f} : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$ which has low risk and low "unfairness"

Unfairness: $f \mapsto \mathcal{U}(f)$ quantifies violations of fairness constraint

Main approaches

Observations: $(\boldsymbol{X}_1, S_1, Y_1), \ldots, (\boldsymbol{X}_n, S_n, Y_n) \in \mathcal{X} \times \mathcal{S} \times \mathbb{R}$

Goal: build $\hat{f} : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$ which has low risk and low "unfairness"

Unfairness: $f \mapsto \mathcal{U}(f)$ quantifies violations of fairness constraint

- Fairness at training: (Agarwal et al., 2018; Donini et al., 2018; Oneto, Donini, and Pontil, 2019; Agarwal, Dudik, and Wu, 2019) …
- Data transformation: (Donini et al., 2018; Adebayo and Kagal, 2016; Calmon et al., 2017; Zemel et al., 2013) …
- Post-processing: (Hardt, Price, and Srebro, 2016; Chiappa et al., 2020; Chzhen et al., 2020a; Chzhen et al., 2020b; Le Gouic, Loubes, and Rigollet, 2020) …

Regression with Demographic Parity

based on joint work with C. Denis, M. Hebiri, L. Oneto, M. Pontil

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathbb{R}^d \times \mathcal{S} \times \mathbb{R}$

Prediction: $f : \mathcal{Z} \to \mathbb{R}$

- Fairness through awareness: $\mathcal{Z} = \mathbb{R}^d \times \mathcal{S}$ and $\boldsymbol{Z} = (\boldsymbol{X}, S)$

Risk: $f \mapsto \mathcal{R}(f)$

- regression: $\mathcal{R}(f) = \mathbb{E}(Y - f(\boldsymbol{X}, S))^2$

Fairness constraint:

- Demographic Parity (DP): $f(\boldsymbol{X}, S) \perp\!\!\!\perp S$

Bayes rule $f^*(\boldsymbol{X}, S) = \mathbb{E}[Y|\boldsymbol{X}, S]$ is unfair

$$f(\boldsymbol{X}, S) \perp\!\!\!\perp S \iff (f(\boldsymbol{X}, S) \mid S{=}s) \overset{d}{=} (f(\boldsymbol{X}, S) \mid S{=}s') \ \forall s, s' \in \mathcal{S}$$

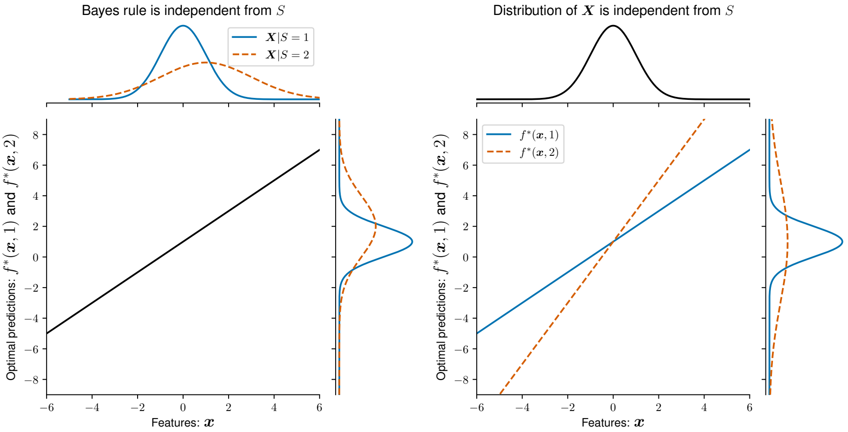


Figure: Two sources of DP unfairness of $f^*(\boldsymbol{X}, S) = \mathbb{E}[Y|\boldsymbol{X}, S]$.

Optimal prediction under DP

Optimal fair: $f_0^* \in \operatorname*{arg\,min}_{f:\mathcal{Z}\to\mathbb{R}} \left\{\mathbb{E}(Y - f(\boldsymbol{X}, S))^2 \,:\, f(\boldsymbol{X}, S) \perp\!\!\!\perp S\right\}$

Bayes optimal: $f^* \in \operatorname*{arg\,min}_{f:\mathcal{Z}\to\mathbb{R}} \mathbb{E}(Y - f(\boldsymbol{X}, S))^2$

Question: is there a link between f_0^* and f^*?

Optimal prediction under DP

Optimal fair: $f_0^* \in \arg\min_{f:\mathcal{Z}\to\mathbb{R}} \left\{\mathbb{E}(Y - f(\boldsymbol{X}, S))^2 \;:\; f(\boldsymbol{X}, S) \perp\!\!\!\perp S\right\}$

Bayes optimal: $f^* \in \arg\min_{f:\mathcal{Z}\to\mathbb{R}} \mathbb{E}(Y - f(\boldsymbol{X}, S))^2$

Question: is there a link between f_0^* and f^*?

Theorem

Assume $(f^*(\boldsymbol{X}, S) \mid S = s)$ are non-atomic with finite second moment. Set $w_s = \mathbb{P}(S{=}s)$, $F_{f^*|S=s}(t) = \mathbb{P}(f^*(\boldsymbol{X}, S){\le}t|S{=}s)$, then

$$\mathrm{Law}(f_0^*(\boldsymbol{X}, S)) = \arg\min_{\nu} \sum_{s\in\mathcal{S}} w_s W_2^2\left(\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S = s), \nu\right) \quad ,$$

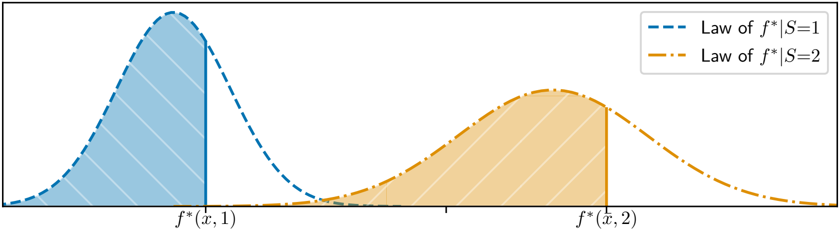
$$f_0^*(\boldsymbol{x}, s) = \left(\sum_{s'\in\mathcal{S}} w_{s'} F^{-1}_{f^*|S=s'}\right) \circ F_{f^*|S=s} \circ f^*(\boldsymbol{x}, s) \quad .$$

(C., Denis, Hebiri, Oneto, Pontil, 2020b)(Le Gouic, Loubes, and Rigollet, 2020)

Interpretation for $\mathcal{S} = \{1, 2\}$

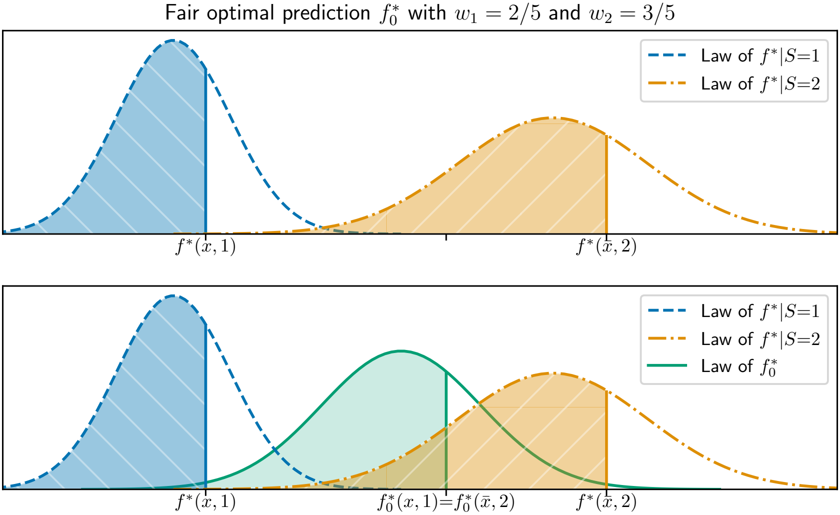
Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 F^{-1}_{f^*|S=2} \circ F_{f^*|S=1} \circ f^*(\boldsymbol{x}, 1)$

Fair optimal prediction f_0^* with $w_1 = 2/5$ and $w_2 = 3/5$



Interpretation for $\mathcal{S} = \{1, 2\}$

Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 F^{-1}_{f^*|S=2} \circ F_{f^*|S=1} \circ f^*(\boldsymbol{x}, 1)$



Generic post-processing estimator

Fair optimal: $f_0^*(\boldsymbol{x}, s) = \left(\sum_{s' \in \mathcal{S}} w_{s'} F^{-1}_{f^*|S=s'}\right) \circ F_{f^*|S=s} \circ f^*(\boldsymbol{x}, s)$

- Data: $\forall s \in \mathcal{S}$ we observe $\boldsymbol{X}_1^s, \ldots, \boldsymbol{X}_{2N_s}^s \overset{i.i.d.}{\sim} \mathbb{P}_{\boldsymbol{X}|S=s}$
- Base estimator: $\hat{f}$ independent from the above data.

Plug-in: $\hat{f}_0(\boldsymbol{x}, s) = \left(\sum_{s' \in \mathcal{S}} w_{s'} \hat{G}^{-1}_{\hat{f}|S=s'}\right) \circ \hat{F}_{\hat{f}|S=s} \circ \hat{f}(\boldsymbol{x}, s)$

$$\hat{G}_{\hat{f}|S=s}(t) = \frac{1}{N_s} \sum_{i \le N_s} \mathbb{I}\left\{\hat{f}(\boldsymbol{X}_i^s, s) + \xi_i^s \le t\right\}$$

$$\hat{F}_{\hat{f}|S=s}(t) = \frac{1}{N_s} \sum_{i > N_s} \mathbb{I}\left\{\hat{f}(\boldsymbol{X}_i^s, s) + \xi_i^s \le t\right\}$$

Theoretical guarantees: fairness

Theorem

For any joint distribution $\mathbb{P}$ of $(\boldsymbol{X}, S, Y)$, any base estimator $\hat{f}$ constructed on labeled data, and for all $s, s' \in \mathcal{S}$, the estimator $\hat{f}_0$ satisfies

$$\sup_{t\in\mathbb{R}} \left|\mathbf{P}(\hat{f}_0(\boldsymbol{X}, S) \leq t \mid S{=}s) - \mathbf{P}(\hat{f}_0(\boldsymbol{X}, S) \leq t \mid S{=}s')\right| \lesssim \frac{1}{N_s \wedge N_{s'}}$$

$$\mathbf{E}\sup_{t\in\mathbb{R}} \left|\mathbf{P}(\hat{f}_0(\boldsymbol{X}, S) \leq t | S{=}s, \mathcal{D}) - \mathbf{P}(\hat{f}_0(\boldsymbol{X}, S) \leq t | S{=}s', \mathcal{D})\right| \lesssim \frac{1}{\sqrt{N_s \wedge N_{s'}}}$$

(C., Denis, Hebiri, Oneto, Pontil, 2020b)

Theoretical guarantees: risk

Assumption

$\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S = s)$ admits a density q_s, which is lower bounded by $\underline{\lambda}_s > 0$ and upper-bounded by $\overline{\lambda}_s \geq \underline{\lambda}_s$ for all $s \in \mathcal{S}$.

Theoretical guarantees: risk

Assumption

$\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S = s)$ admits a density q_s, which is lower bounded by $\underline{\lambda}_s > 0$ and upper-bounded by $\overline{\lambda}_s \geq \underline{\lambda}_s$ for all $s \in \mathcal{S}$.

Theorem

Set $\xi_i^s \overset{i.i.d.}{\sim} \mathrm{Unif}[0, \sum_{s \in \mathcal{S}} w_s N_s^{-1/2}]$, then under the above assumption it holds that

$$\mathbf{E}\|\hat{f}_0 - f_0^*\|_1 \lesssim \underbrace{\mathbf{E}\|\hat{f} - f^*\|_1}_{\text{quality of base estimator}} \bigvee \underbrace{\sum_{s \in \mathcal{S}} w_s N_s^{-1/2}}_{\text{CDF + quantile estimation}},$$

where the leading constant depends only on $\underline{\lambda}_s, \overline{\lambda}_s$

(C., Denis, Hebiri, Oneto, Pontil, 2020b)

Quantification of risk/fairness trade-off

based on joint work with N. Schreuder

Demographic Parity: $f(\boldsymbol{X}, S) \perp\!\!\!\perp S$

Quantification of risk/fairness trade-off

based on joint work with N. Schreuder

Demographic Parity: $f(\boldsymbol{X}, S) \perp\!\!\!\perp S$

- **Problem:** too stiff — either fair or unfair
- **Question:** how to quantify unfairness *i.e.*, violation of DP?
- **Question:** which risks are achievable for a fixed unfairness level?

What was used?

Demographic Parity: $f(\boldsymbol{X}, S) \perp\!\!\!\perp S$

Unfairness:

$$\mathrm{KS}(\mathrm{Law}(f | S = s), \mathrm{Law}(f))$$

$$\mathrm{TV}(\mathrm{Law}(f | S = s), \mathrm{Law}(f))$$

$$\mathrm{KL}(\mathrm{Law}(f | S = s), \mathrm{Law}(f))$$

What was used?

Demographic Parity: $f(\boldsymbol{X}, S) \perp\!\!\!\perp S$

Unfairness:

$$\mathrm{KS}(\mathrm{Law}(f|S=s), \mathrm{Law}(f))$$

$$\mathrm{TV}(\mathrm{Law}(f|S=s), \mathrm{Law}(f))$$

$$\mathrm{KL}(\mathrm{Law}(f|S=s), \mathrm{Law}(f))$$

We consider:

$$\mathcal{U}(f) = \min_{\nu} \sum_{s \in \mathcal{S}} w_s W_2^2(\mathrm{Law}(f|S=s), \nu)$$

Previous result

$$\min_{f:\mathcal{Z}\to\mathbb{R}} \left\{\mathbb{E}(f(\boldsymbol{X}, S) - f^*(\boldsymbol{X}, S))^2 \;:\; f(\boldsymbol{X}, S) \perp\!\!\!\perp S\right\} = \mathcal{U}(f^*)$$

(C., Denis, Hebiri, Oneto, Pontil, 2020b)(Le Gouic, Loubes, and Rigollet, 2020)

Fairer predictions

α-Relative Improvement $\quad f^*_\alpha \in \arg\min \{\mathcal{R}(f) \,:\, \mathcal{U}(f) \le \alpha\,\mathcal{U}(f^*)\}$

- f^*_α – $1/\alpha$ times fairer than f^*
- f^*_0 – optimal DP fair prediction
- $f^*_1 \equiv f^*$ – Bayes optimal prediction

Fairer predictions

α-Relative Improvement $\quad f^*_\alpha \in \arg\min\{\mathcal{R}(f) \,:\, \mathcal{U}(f) \leq \alpha\,\mathcal{U}(f^*)\}$

- f^*_α – $^1/_\alpha$ times fairer than f^*
- f^*_0 – optimal DP fair prediction
- $f^*_1 \equiv f^*$ – Bayes optimal prediction

Theorem

Under the same assumptions as before, for all $\alpha \in [0, 1]$ it holds that

$$f^*_\alpha \equiv \sqrt{\alpha} f^*_1 + (1 - \sqrt{\alpha}) f^*_0 \ .$$

(C. and Schreuder, 2020)

Risk/fairness trade-off

α-Relative Improvement $\quad f^*_\alpha \in \arg\min \{\mathcal{R}(f) : \mathcal{U}(f) \le \alpha \mathcal{U}(f^*)\}$

Lemma

Under the same assumptions as before, for all $\alpha \in [0,1]$ it holds that

$$\mathcal{R}(f^*_\alpha) = (1-\sqrt{\alpha})^2 \mathcal{U}(f^*) \quad \text{and} \quad \mathcal{U}(f^*_\alpha) = \alpha \mathcal{U}(f^*) \ .$$

(C. and Schreuder, 2020)

Risk/fairness trade-off

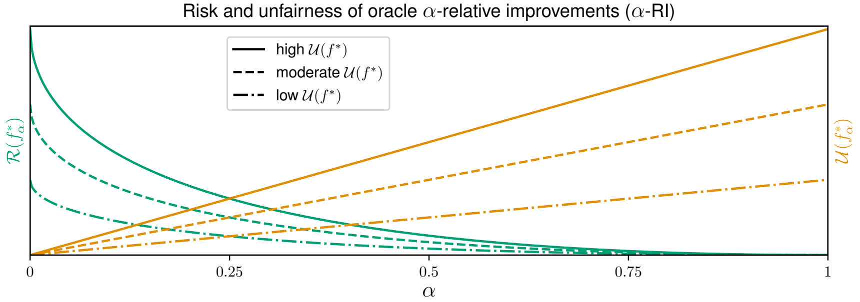
α-Relative Improvement $\quad f^*_\alpha \in \arg\min\{\mathcal{R}(f) \,:\, \mathcal{U}(f) \le \alpha\,\mathcal{U}(f^*)\}$

Lemma

Under the same assumptions as before, for all $\alpha \in [0,1]$ it holds that

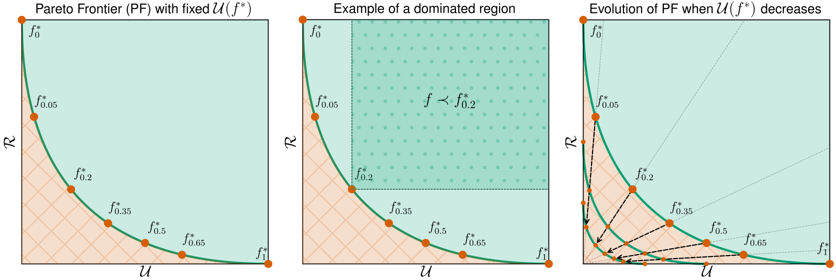
$$\mathcal{R}(f^*_\alpha) = (1-\sqrt{\alpha})^2\,\mathcal{U}(f^*) \quad \text{and} \quad \mathcal{U}(f^*_\alpha) = \alpha\,\mathcal{U}(f^*)\ .$$

(C. and Schreuder, 2020)



Pareto interpretation

- Multi-objective optimization: $\min_{f:\mathcal{Z}\to\mathbb{R}} \Big(\mathcal{U}(f), \mathcal{R}(f)\Big)$.
- Each prediction f defines a point $(\mathcal{U}(f), \mathcal{R}(f))$
- f is dominated by f' iff $\mathcal{R}(f') \leq \mathcal{R}(f)$ and $\mathcal{U}(f') \leq \mathcal{U}(f)$



Summary

- Regression with Demographic Parity is connected to the problem of Wasserstein barycenters
- A generic post-processing estimator is proposed, which requires only unlabeled data and enjoys plug-n-play guarantees
- Introduced notion of unfairness allows to provide precise quantification of the risk/fairness trade-off

Summary

- Regression with Demographic Parity is connected to the problem of Wasserstein barycenters
- A generic post-processing estimator is proposed, which requires only unlabeled data and enjoys plug-n-play guarantees
- Introduced notion of unfairness allows to provide precise quantification of the risk/fairness trade-off

Further details / Thank you!

- E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil (2020b). "Fair Regression with Wasserstein Barycenters". In: *to appear NeurIPS20*
- E. Chzhen and N. Schreuder (2020). "A minimax framework for quantifying risk-fairness trade-off in regression". In: *arXiv preprint arXiv:2007.14265*

Bibliography I

Adebayo, J. and L. Kagal (2016). “Iterative orthogonal feature projection for diagnosing bias in black-box models”. In: *Conference on Fairness, Accountability, and Transparency in Machine Learning*.

Agarwal, A., M. Dudik, and Z. S. Wu (2019). “Fair Regression: Quantitative Definitions and Reduction-Based Algorithms”. In: *International Conference on Machine Learning*.

Agarwal, A. et al. (2018). “A reductions approach to fair classification”. In: *arXiv preprint arXiv:1803.02453*.

Barocas, S., M. Hardt, and A. Narayanan (2018). *Fairness and Machine Learning*. `http://www.fairmlbook.org`. fairmlbook.org.

Calmon, F. et al. (2017). “Optimized Pre-Processing for Discrimination Prevention”. In: *Neural Information Processing Systems*.

Chiappa, S. et al. (2020). “A general approach to fairness with optimal transport”. In: *AAAI*.

Bibliography II

Chzhen, E. and N. Schreuder (2020). "A minimax framework for quantifying risk-fairness trade-off in regression". In: *arXiv preprint arXiv:2007.14265.*

Chzhen, E. et al. (2020a). "Fair Regression via Plug-In Estimator and Recalibration". NeurIPS20.

– (2020b). "Fair Regression with Wasserstein Barycenters". In: *to appear NeurIPS20.*

Donini, M. et al. (2018). "Empirical risk minimization under fairness constraints". In: *Neural Information Processing Systems.*

Hardt, M., E. Price, and N. Srebro (2016). "Equality of opportunity in supervised learning". In: *Neural Information Processing Systems.*

Le Gouic, T., J.-M. Loubes, and P. Rigollet (2020). "Projection to Fairness in Statistical Learning". In: *arXiv preprint arXiv:2005.11720.*

Oneto, L., M. Donini, and M. Pontil (2019). "General Fair Empirical Risk Minimization". In: *arXiv preprint arXiv:1901.10080.*

Bibliography III

Zemel, R. et al. (2013). “Learning fair representations”. In: *International Conference on Machine Learning.*

Adebayo, J. and L. Kagal (2016). “Iterative orthogonal feature projection for diagnosing bias in black-box models”. In: *Conference on Fairness, Accountability, and Transparency in Machine Learning.*

Agarwal, A., M. Dudik, and Z. S. Wu (2019). “Fair Regression: Quantitative Definitions and Reduction-Based Algorithms”. In: *International Conference on Machine Learning.*

Agarwal, A. et al. (2018). “A reductions approach to fair classification”. In: *arXiv preprint arXiv:1803.02453.*

Barocas, S., M. Hardt, and A. Narayanan (2018). *Fairness and Machine Learning.* `http://www.fairmlbook.org`. fairmlbook.org.

Calmon, F. et al. (2017). “Optimized Pre-Processing for Discrimination Prevention”. In: *Neural Information Processing Systems.*

Bibliography IV

Chiappa, S. et al. (2020). "A general approach to fairness with optimal transport". In: *AAAI*.

Chzhen, E. and N. Schreuder (2020). "A minimax framework for quantifying risk-fairness trade-off in regression". In: *arXiv preprint arXiv:2007.14265*.

Chzhen, E. et al. (2020a). "Fair Regression via Plug-In Estimator and Recalibration". NeurIPS20.

– (2020b). "Fair Regression with Wasserstein Barycenters". In: *to appear NeurIPS20*.

Donini, M. et al. (2018). "Empirical risk minimization under fairness constraints". In: *Neural Information Processing Systems*.

Hardt, M., E. Price, and N. Srebro (2016). "Equality of opportunity in supervised learning". In: *Neural Information Processing Systems*.

Le Gouic, T., J.-M. Loubes, and P. Rigollet (2020). "Projection to Fairness in Statistical Learning". In: *arXiv preprint arXiv:2005.11720*.

Bibliography V

Oneto, L., M. Donini, and M. Pontil (2019). "General Fair Empirical Risk Minimization". In: *arXiv preprint arXiv:1901.10080.*

Zemel, R. et al. (2013). "Learning fair representations". In: *International Conference on Machine Learning.*