Meta Optimality for Demographic Parity Constrained Regression via Post-Processing

Kazuto Fukuchi
University of Tsukuba / RIKEN AIP

July 13-19, 2025

ICML 2025

Unfairness in Machine Learning

- Real-world ML systems can be unfair:
  - Criminal risk assessment (Angwin et al. 2016)
  - Hiring (Dastin 2018)
  - Facial recognition (Crockford 2020; Najibi 2020)
  - Credit scoring (Vigdor 2019)

Unfairness in Machine Learning

- Real-world ML systems can be unfair:
  - Criminal risk assessment (Angwin et al. 2016)
  - Hiring (Dastin 2018)
  - Facial recognition (Crockford 2020; Najibi 2020)
  - Credit scoring (Vigdor 2019)
- These cases underscore the need for *fair* models.
- Many approaches exist to address different fairness criteria (Feldman et al. 2015; Chzhen et al. 2020; Chen et al. 2023; Jovanović et al. 2023; Khalili et al. 2023; Xian et al. 2023; Xu et al. 2023)

Unfairness in Machine Learning

- Real-world ML systems can be unfair:
  - Criminal risk assessment (Angwin et al. 2016)
  - Hiring (Dastin 2018)
  - Facial recognition (Crockford 2020; Najibi 2020)
  - Credit scoring (Vigdor 2019)
- These cases underscore the need for *fair* models.
- Many approaches exist to address different fairness criteria (Feldman et al. 2015; Chzhen et al. 2020; Chen et al. 2023; Jovanović et al. 2023; Khalili et al. 2023; Xian et al. 2023; Xu et al. 2023)

Research Question

What is the best algorithm for fair regression?

Unfairness in Machine Learning

- Real-world ML systems can be unfair:
  - Criminal risk assessment (Angwin et al. 2016)
  - Hiring (Dastin 2018)
  - Facial recognition (Crockford 2020; Najibi 2020)
  - Credit scoring (Vigdor 2019)
- These cases underscore the need for *fair* models.
- Many approaches exist to address different fairness criteria (Feldman et al. 2015; Chzhen et al. 2020; Chen et al. 2023; Jovanović et al. 2023; Khalili et al. 2023; Xian et al. 2023; Xu et al. 2023)

Research Question

What is the best algorithm for fair regression?

best = minimax optimal, fair = demographic parity

Fair Regression

- For each group $s \in [M]$:
  - $X^{(s)}$: non-sensitive features ($\mathcal{X}$)
  - $Y^{(s)}$: outcome on Ω ($\Omega \subset \mathbb{R}$ open, bounded)
- **Goal:** Given n_s i.i.d. copies of $(X^{(s)}, Y^{(s)})$ for each $s \in [M]$, construct an accurate and fair regressor $f_{:}$.

Fair Regression

- For each group $s \in [M]$:
  - $X^{(s)}$: non-sensitive features ($\mathcal{X}$)
  - $Y^{(s)}$: outcome on Ω ($\Omega \subset \mathbb{R}$ open, bounded)
- **Goal:** Given n_s i.i.d. copies of $(X^{(s)}, Y^{(s)})$ for each $s \in [M]$, construct an accurate and fair regressor $f_{:}$.
- **Fairness:**

$$\mathbb{P}_{\mu_s}\{f_s(X^{(s)}) \in E\} = \mathbb{P}_{\mu_{s'}}\{f_{s'}(X^{(s')}) \in E\}$$

Fair Regression

- For each group $s \in [M]$:
  - $X^{(s)}$: non-sensitive features ($\mathcal{X}$)
  - $Y^{(s)}$: outcome on Ω ($\Omega \subset \mathbb{R}$ open, bounded)
- **Goal:** Given n_s i.i.d. copies of $(X^{(s)}, Y^{(s)})$ for each $s \in [M]$, construct an accurate and fair regressor $f_:$.
- **Fairness:**

$$\mathbb{P}_{\mu_s}\{f_s(X^{(s)}) \in E\} = \mathbb{P}_{\mu_{s'}}\{f_{s'}(X^{(s')}) \in E\}$$

- **Accuracy:**

$$d^2_{\mu_{X,:}}(f_:, \bar{f}^*_{\mu,:}) = \sum_{s \in [M]} w_s \int \big(f_s(z) - \bar{f}^*_{\mu,s}(z)\big)^2 \mu_{X,s}(dz)$$

  - $\bar{f}^*_{\mu,:}$: Fair Bayes-optimal regressor (closest to Bayes-optimal, subject to demographic parity)

Minimax Optimal Fair Regression

- **Fair minimax optimal error:**

$$\bar{\mathcal{E}}_n(\mathcal{P}) = \inf_{\bar{f}_{n,:}:\text{fair}} \sup_{\mu_: \in \mathcal{P}} \mathbb{E}_{\mu_:^n}[d^2_{\mu_{X,:}}(\bar{f}_{n,:}, \bar{f}^*_{\mu,:})],$$

  - $\sup$: over all distributions $\mu_: \in \mathcal{P}$
  - $\inf$: over all fair regression algorithms $\bar{f}_{n,:}$
- **Fair minimax optimal regression algorithm:**
  - Achieves the minimax optimal error above
  - Guarantees the smallest possible error in the worst case

Existing Work

Fair minimax optimal algorithms have been developed for specific data generation models $\mathcal{P}$:

	Task	$\mathcal{P}$
Chzhen et al. (2022)	Regression	Linear w/ additive bias
Fukuchi et al. (2023)	Regression	Linear w/ group-dependent coefficients
Zeng et al. (2024)	Classification	Hölder class /w margin & density conditions

Existing Work

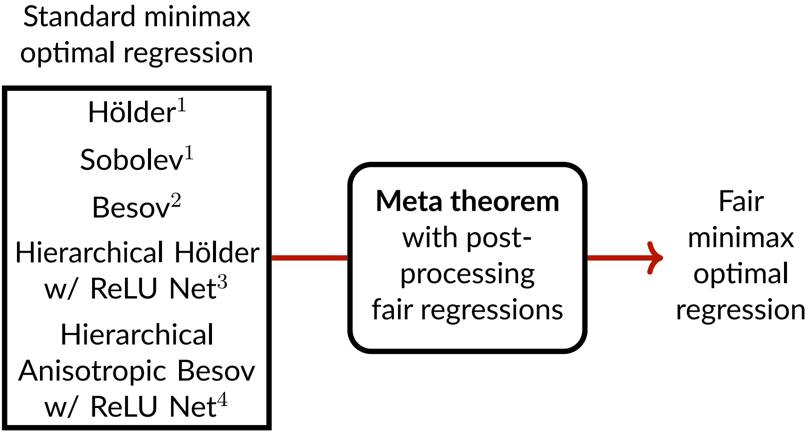
Fair minimax optimal algorithms have been developed for specific data generation models $\mathcal{P}$:

	Task	$\mathcal{P}$
Chzhen et al. (2022)	Regression	Linear w/ additive bias
Fukuchi et al. (2023)	Regression	Linear w/ group-dependent coefficients
Zeng et al. (2024)	Classification	Hölder class /w margin & density conditions

Key Limitation:

- Methods are tailored to their assumed $\mathcal{P}$.
- Generalizing to other models demands new theoretical analysis.

Contributions: Meta-Optimality



[1] (Giné et al. 2015), [2] (Donoho et al. 1998), [3] (Schmidt-Hieber 2020), [4] (Suzuki et al. 2021)

Developed a meta-theorem showing that post-processing standard minimax optimal regressors yields fair minimax optimality.

Summary

- Studied minimax optimal regression under demographic parity constraints.
- Proved a meta-optimality theorem for post-processing fair regression: this approach inherits minimax optimality from standard regression algorithms, enabling broad applicability across diverse settings.

Check out my poster for details!

References I

Donoho, David L. and Iain M. Johnstone (1998). Minimax estimation via wavelet shrinkage. In: *The Annals of Statistics* 26.3, pp. 879–921. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1024691081.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian (2015). Certifying and Removing Disparate Impact. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Ed. by Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams. Sydney, NSW, Australia: ACM, pp. 259–268. DOI: 10.1145/2783258.2783311.

References II

- Giné, Evarist and Richard Nickl (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. ISBN: 978-1-107-04316-9. DOI: 10.1017/CBO9781107337862.
- Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016). *Machine Bias*. ProPublica. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
- Dastin, Jeffrey (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

References III

- Vigdor, Neil (2019). Apple Card Investigated After Gender Discrimination Complaints. In: *The New York Times*. ISSN: 0362-4331.
- Chzhen, Evgenii, Christophe Denis, Mohamed Hebiri, and Massimiliano Pontil (2020). Fair Regression with Wasserstein Barycenters. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 7321–7331. arXiv: 2006.07286.
- Crockford, Kade (2020). *How is Face Recognition Surveillance Technology Racist?* American Civil Liberties Union. URL: https://www.aclu.org/news/privacy-technology/how-is-face-recognition-surveillance-technology-racist.

References IV

- Najibi, Alex (2020). *Racial Discrimination in Face Recognition Technology*. Science in the News. URL: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/.
- Schmidt-Hieber, Johannes (2020). Nonparametric regression using deep neural networks with ReLU activation function. In: *The Annals of Statistics* 48.4, pp. 1875–1897. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/19-AOS1875.
- Suzuki, Taiji and Atsushi Nitanda (2021). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 3609–3621.

References V

- Chzhen, Evgenii and Nicolas Schreuder (2022). A minimax framework for quantifying risk-fairness trade-off in regression. In: *The Annals of Statistics* 50.4, pp. 2416–2442. ISSN: 0090-5364. DOI: 10.1214/22-AOS2198.
- Chen, Wenlong, Yegor Klochkov, and Yang Liu (2023). Post-hoc bias scoring is optimal for fair classification. In: *The Twelfth International Conference on Learning Representations*.
- Fukuchi, Kazuto and Jun Sakuma (2023). Demographic parity constrained minimax optimal regression under linear model. In: *Advances in Neural Information Processing Systems*. Vol. 36, pp. 8653–8689. arXiv: 2206.11546.
- Jovanović, Nikola, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev (2023). FARE: Provably Fair Representation Learning with Practical Certificates. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, pp. 15401–15420.

References VI

- Khalili, Mohammad Mahdi, Xueru Zhang, and Mahed Abroshan (2023). Loss Balancing for Fair Supervised Learning. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, pp. 16271–16290.
- Xian, Ruicheng, Lang Yin, and Han Zhao (2023). Fair and Optimal Classification via Post-Processing. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, pp. 37977–38012.
- Xu, Shizhou and Thomas Strohmer (2023). Fair Data Representation for Machine Learning at the Pareto Frontier. In: *Journal of Machine Learning Research* 24.331, pp. 1–63. ISSN: 1533-7928.
- Zeng, Xianli, Guang Cheng, and Edgar Dobriban (2024). *Minimax Optimal Fair Classification with Bounded Demographic Disparity*. arXiv: 2403.18216[stat].