Thesis proposal
Handling unfairness in data: modelling, detecting, and debiasing

Evgenii Chzhen Luca Ganassali

CNRS, Université Paris-Saclay, Laboratoire de Mathématiques d'Orsay
first.last@universite-paris-saclay.fr

February 20, 2025

AI-driven processes increasingly influence critical domains such as healthcare, finance, criminal justice, and education. However, these systems often embed structural unfairness, either inherited from biased data or induced by model design. Understanding and addressing such biases requires a rigorous, model-driven and, consequently, statistical perspective.

A vast body of literature has tackled the issue of fairness in AI by formulating it as an optimization problem under constraints, aiming to enforce statistical fairness criteria without explicitly modeling unfairness itself. This approach has led to a rich and diverse set of methods, often focused on adjusting model outputs to satisfy predefined fairness conditions. In contrast, our goal is to take a more fine-grained, model-driven perspective by embedding unfairness within data-generating processes. We take a model-driven approach, planting unfairness in data—e.g., in online settings with feedback—to better understand, detect, and mitigate it.

This thesis focuses on three major themes:

1. Modeling biases present in the data, drawing inspiration from existing machine learning literature and extending it to the fields of economics and econometrics;
2. Conducting a rigorous statistical analysis for detecting/estimating the unfairness, once the model is established, including an investigation of the fundamental statistical limits of the problem;
3. Interactions of statistical modeling and sequential learning (such as linear bandits) is expected to be investigated;

1 Bias in linear bandits

To give an idea of the type of modeling that we envision and that we want to investigate deeply, we present a short overview of Gaucher et al. (2022). The are faced with the following problem: given a two finite pools of candidates $\mathcal{X}_1, \mathcal{X}_2 \subset \mathbb{R}^d$ (e.g., males and females) at each time step $t \geq 1$ the learner picks $x_t \in \mathcal{X}_1 \cup \mathcal{X}_2$ and receives a biased feedback:

$$y_t = \langle x_t, \theta^* \rangle + b_{s_t} + \eta_t \ ,$$

where $s_t = 1$ if $x_t \in \mathcal{X}_1$ and $s_t = 2$ otherwise. Here $b_s \in \mathbb{R}$ is the systematic group-dependent bias that is added to the feedback. The goal of the learner is to design a sequential strategy to minimize

the regret

$$R_T = \max_{x \in \mathcal{X}_1 \cup \mathcal{X}_2} \sum_{t=1}^{T} \langle x_t - x, \theta^* \rangle \ .$$

The work of Gaucher et al. (2022) studies the impact of the biased feedback, when the true regret only depends on the unbiased feedback and link the eventual rates to the geometry of the candidates.

Extending this model to a more complex bias in the feedback is one of the initial goals of the thesis, which would allow the candidate to get familiar with the literature on fairness and linear bandits. The first and rather natural extension is to consider

$$y_t = \langle x_t, \theta^* + \theta_{s_t} \rangle + \eta_t \ .$$

That is, the bias is in the alignment of feature vectors and the evaluator. Note that in general this problem is hopeless for example when $\theta_1 = \theta_2 \neq 0$, but it becomes more feasible if $\theta_1 = -\theta_2$. The first step is understanding when sub-linear regret is possible depending on the assumptions on θ_s. The second step is about designing algorithms that achieve sub-linear regret and eventually optimal regret, which would require understanding of fundamental limits of the problem.

2 In-feature unfairness

In contrast with the previous approach which considers a systematic bias in the decision-making process, we can also envision a model where the bias lies within the training data. Here, there is no explicit notion of groups, and the data are not necessarily isotropic, but the available observations are biased. A typical model for the features X_t and the outcome Y_t is:

$$X_t = AZ_t + E_t, \quad Y_t = BZ_t + W_t$$

where Z_t represents latent variables, which are thought to be fair, E_t and W_t are independent noises. This setting is related to the *errors-in-variables* problem. Matrix A represents the unfairness in the feature design, for instance by placing too much weight on a particular coordinate or making certain features invisible. The goal is to predict Y_t given X_t, first say, in a non online setting. Doing so, we also aim to propose an interpretable that somehow remove bias from feature, by getting back to the original latent variables Z_t. When A is invertible, we have

$$Y_t = BA^{-1}X_t - BA^{-1}E_t + W_t,$$

the noise may be structured under additional assumptions on A, B, which can help the learning procedure.

When these conditions are not met, designing an estimation procedure becomes an even more interesting problem, requiring a more refined modeling of biases.

3 Planted counterfactuals

Another approach which we envision to model unfairness in online learning is counterfactual reasoning.

Counterfactual reasoning consists in providing answers to the following question: 'all things being equal, had this individual been a man instead of a woman, how would the output of the algorithm have been modified?'. Counterfactual reasoning is notoriously difficult problem as it often requires very strong assumptions on the data-generating process.

Let us give a high-level idea of a possible modelization of the problem. Assume that we have K groups of candidates, $\mathcal{X}_1, \ldots, \mathcal{X}_k$, which represent the sensitive classes for which we want to guarantee

fairness in the decision process. At each time $t \geq 1$, the learner selects a group $k_t \in [K] := \{1, \dots, K\}$ and a candidate $x_t \in \mathcal{X}_{k_t}$, which yields an immediate reward of the form:

$$y_t = f_{k_t}(x_t) + \varepsilon_t \ ,$$

where ε_t represents the centered noise. Note that $f_1, \dots, f_k$ are the different reward functions across groups. In classical online learning, the focus is on maximizing cummulative reward $\sum_{t=1}^{T} f_{k_t}(x_t)$. Thus, the goal is purely reward-driven and is based on the assumption (or belief) that the obtained rewards are unbiased (or fair). Yet, in many real-world scenarios, this assumption does not hold.

Counterfactual regret Instead, we propose defining a notion of counterfactual reward and counterfactual feature. We assume the existence of counterfactual mappings $\Psi_k : \mathcal{X}_t \to \mathcal{X}$ which we call *planted counterfactuals*, and a counterfactual reward function $f : \mathcal{X} \to \mathbb{R}$. Each function Ψ_k takes a feature from group $\mathcal{X}_k$ and returns its counterpart in the counterfactual world $\mathcal{X}$. The counterfactual reward is then defined as $\sum_{t=1}^{T} f(\Psi_{k_t}(x_t))$.

The goal now is to maximize this reward instead. Morally, the mappings Ψ_k are precisely counterfactual features of individuals in group $\mathcal{X}_k$ (e.g., Ψ_k maps women to men) that are already planted. Meanwhile, the function f would correspond to the reward assuming that all the groups are counterfactually mapped into the same space, thus eliminating the effect of between-group bias that.

Naturally, this goal is infeasible in the most general setting, so we need to make reasonable assumptions about both Ψ_k and f. Developing these assumptions and building the corresponding algorithms lies at the core of this thesis.

Estimating the planted counterfactuals: a transport approach In order to address the problem of minimizing the counterfactual regret in previous section, a key step is to estimate the planted counterfactuals Ψ_k, as well as f.

A general approach to model Ψ_k is to design a transport map between the features' distribution in $\mathcal{X}_k$ and some reasonable distribution in $\mathcal{X}$ (either postulated, known or estimated depending on the problem). For instance, in the 1D case, when $\mathcal{X}_k, \mathcal{X} \subset \mathbb{R}$, a natural transport map Ψ_k is a non-decreasing mapping, thus preserving quantiles (or ranks) within $\mathcal{X}_k$. In higher dimensions, however, defining quantiles or ranks is more challenging, and different approaches have already been proposed in the recent literature (see e.g. Hallin et al. (2021) and Ghosal and Sen (2022)).

The goal of this part is to investigate theoretical properties of various notions of high-dimensional quantiles that, on one hand, have desirable statistical properties (e.g., enabling estimation without the curse of dimensionality) and, on the other hand, are computationally efficient (e.g., computable in polynomial time with respect to both sample size and dimension).

Regarding the estimation/computation of the transport map Ψ_k itself, we know that the cost of such computation rapidly increases with the dimension. However, note that in the above model, the only crucial map to estimate in order to minimize $\tilde{R}_T$ is $f \circ \Psi_k$, instead of Ψ_k. Hence making structural assumptions on f can help in estimating $f \circ \Psi_k$ accurately, escaping the curse of dimensionality.

To illustrate this, let us take the simple example where f is linear, that is, the reward y_t writes $\langle \beta^*, \Psi_k(x_t) \rangle + \varepsilon_t$. In this case, the goal is to learn transport Ψ_k only in the direction of β^*, without caring for the other orthogonal directions.

In line with this approach, some related concepts such as depths (Zuo and Serfling (2000)), Sliced Wasserstein (Kolouri et al. (2019)) can also be useful for our tasks. Studying these approaches more in depth and building new ones for our general problem will be one of the main directions of this thesis.

4 Required profile

Candidates with a background in pure or applied mathematics/physics are encouraged to apply. Our aim is to explore the intersection of three key areas: algorithmic fairness, online learning, and counterfactual and causal reasoning. Ideally, applicants will have prior experience (such as a Master's-level course) in at least one of these areas. Due to the theoretical nature of the project, a strong foundation in undergraduate mathematics and graduate-level statistics is a plus.

Acknowledgments The three-year thesis is funded by PEPR CAUSAL-IA[1]. The grant comes without any obligatory teaching load. Interactions with other members of the consortium are to be expected.

References

Gaucher, S., A. Carpentier, and C. Giraud (2022). The price of unfairness in linear bandits with biased feedback. *Advances in Neural Information Processing Systems 35*, 18363–18376.

Ghosal, P. and B. Sen (2022). Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *The Annals of Statistics 50*(2), 1012–1037.

Hallin, M., E. del Barrio, J. Cuesta-Albertos, and C. Matrán (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics 49*(2), 1139 – 1165.

Kolouri, S., K. Kolouri, U. Simsekli, R. Badeau, and G. Rohde (2019). Generalized sliced wasserstein distances. *Advances in neural information processing systems 32*.

Zuo, Y. and R. Serfling (2000). General notions of statistical depth function. *The Annals of Statistics 28*(2), 461–482.

[1] see https://www.pepr-ia.fr/projet/causali-t-ai/