Learning from Revealed Algorithmic Recourse Preferences

Rory Creedon
University College London
Department of Computer Science

Submitted to University College London (UCL) in partial fulfilment of the requirements for the award of the degree of

Master of Science in Data Science and Machine Learning.

Industry supervisor: **Colin Rowat**

Academic supervisor: **Matthew Caldwell**

Submission date: **July 20, 2023**

Abstract

- Most works in algorithmic recourse/strategic classification assume a simple, pre-specified cost function for changing feature values.
- Understanding *individual* cost functions is important for generating recourse and understanding *global* cost functions is important for strategic classification.
- Whilst there has been research into generating individual recourse through preference elicitation, there has not been research into learning *global* cost functions.
- Learning algorithms are proposed to learn cost function from the users' revealed preferences - their responses to a series of pairwise comparisons of different recourse options.
- The algorithms are evaluated on synthetic and semi-synthetic data.
- Recourse costs are compared for users with different protected attributes, showing if learning costs functions aids or exacerbates fairness of recourse.

Contents

1 Introduction

Introduction chapter.

2 Literature Review

2.1 Algorithmic Recourse

2.1.1 Motivation

Description of what algorithmic recourse is and why it is important - use of automatic decision making, GDPR (Voigt and Bussche, 2017). Mention psychological factors causing humans to prefer recourse to explanations (**to find paper(s)**: was mentioned by Ruth Byrne in ICML panel session, from 25 minutes onwards).

2.1.2 Problem Set-up

- Description of the original set-up and problem - i.e.,

$$\mathbf{x}^f = \operatorname*{argmax}_{\mathbf{x}' \in \mathcal{X}} f(\mathbf{x}') - c(\mathbf{x}, \mathbf{x}') \tag{2.1.1}$$

- Description of the causal recourse set-up and problem (Karimi, Schölkopf, and Valera, 2021)
- Cost and distance functions, actionability of features

2.1.3 Recourse methods

Run through methods mentioned in survey paper (Karimi, Barthe, et al., 2022) and also those implemented in `CARLA`.

2.2 Strategic Classification

2.2.1 Standard Strategic Classification

- Begin with Hardt et al. (2016) and explain the set-up as a Stackelberg game with an example.
- Algorithms proposed for this task include Levanon and Rosenfeld (2021), Chen, Liu, and Podimata (2020) and Ahmadi et al. (2022).
- Mention extensions such as:
- Where the cost function is completely unknown to the lender (Dong et al., 2018)
- Where the response of lenders to the classifier is noisy (Jagadeesan, Mendler-Dünner, and Hardt, 2021).
- Where borrowers do not know the decision rule (Ghalme et al., 2021; Bechavod et al., 2022).
- Where the incentives of lender and borrower align (e.g., recommender systems) (Levanon and Rosenfeld, 2022).
- Where the cost functions are linked by graphs for the borrowers (Eilat et al., 2023).

- Where the borrowers act first (Nair et al., 2022).
- Where the borrowers and lenders update at different rates (Zrnic et al., 2021).

2.2.2 Causal Strategic Classification

A review of the *causal* strategic classification literature, which focuses more on causal identification of features which are strategically manipulated (without causing an improvement in underlying credit 'worthiness') and features which causally affect credit 'worthiness'.

2.3 Revealed Preferences

A brief primer on axioms of revealed preferences, and on the literature of *learning from revealed preferences*. To briefly discuss:

- Original paper by Beigman and Vohra (2006), where principal issues a list of prices and the agent purchases different quantities of each good. Over time, the principal learns from the different purchase amounts (which are the revealed preferences).
- When prices are of goods and budget of the agent are drawn from an unknown distribution (Zadimoghaddam and Roth, 2012; Balcan et al., 2014).
- Where the principal is maximising profit (Amin et al., 2015; Roth, Ullman, and Wu, 2016).
- Move onto a more detailed discussion of Dong et al. (2018).

2.4 Pairwise Metric Learning

- Start with an introduction of what pairwise metric learning is and key papers.
- Move onto specific proposed adaptation/simplification of the learning algorithm proposed in Canal et al. (2022).

2.5 Canonical Datasets

The canonical datasets used in the algorithmic recourse and strategic classification literature include:

- `Adult` - dataset to predict whether someone earns over $50,000 or more.
- `German Credit` - dataset identifies people as either good or bad credit risks.
- `FICO-HELOC` - dataset of HELOC applications, where applicants have applied for a credit line between $5,000 and $150,000. Outcome variable is whether they are a good or bad credit risk.
- `Finance` - dataset to predict financial distress for a number of companies. There are several over different time periods for each company.

3 Cost Learning

In order to generate recourse selections, we need to solve the constrained optimisation problem mentioned in equation 2.1.1, where $\mathbf{x}$ are the individual's original features, f is the utility of being positively or negatively classified, c is the cost function and B is the individual's 'budget' for changing their features.

$$\begin{aligned} \mathbf{x}^f &= \underset{\mathbf{x}' \in \mathcal{X}}{\operatorname{argmax}}\, f(\mathbf{x}) - c(\mathbf{x}, \mathbf{x}') \\ \text{s.t. } & c(\mathbf{x}, \mathbf{x}') \leq B \end{aligned} \tag{3.0.1}$$

To solve for $\mathbf{x}^f$ effectively, this is typically handled as a convex optimisation problem. This requires the learned cost function c to be suitable to be convex/suitable for convex optimisation. Two different functional forms for the cost function are outlined below.

3.1 Mahalanobis distance

The Mahalanobis distance between the vector $\mathbf{x}$ and the vector $\mathbf{y}$ is defined in equation 3.1.1, where M is a positive semi-definite matrix.

$$||\mathbf{x} - \mathbf{y}||_{\mathbf{M}} = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{M}^{-1} (\mathbf{x} - \mathbf{y})} \tag{3.1.1}$$

The matrix $\mathbf{M}$ captures different distances relationships between the features within $\mathbf{x}$ and $\mathbf{y}$ in the off-diagonal elements of $\mathbf{M}$. If $\mathbf{M}$ is set to the identity matrix, then the Mahalanobis distance then becomes equal to the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$.

3.1.1 Learning the Mahalanobis distance

In order to use the Mahalanobis distance as a cost function, we must learn the matrix $\mathbf{M}$. In this set-up, each individual k with original features $\mathbf{x}_k$ is presented with N recourse options $(\mathbf{x}_{kn}^a, \mathbf{x}_{kn}^b)$ and responds with $y_{kn} = -1$ if offering a is preferred (preferences are defined by the ground truth cost function) and $y_{kn} = 1$ if offering b is preferred. The optimisation problem presented in Canal et al. (2022) is simplified (to only conduct metric learning, as opposed to metric and preference learning) in equation 3.1.2, where ℓ represents either the hinge or logistic loss function.

$$\begin{aligned} \min_{\mathbf{M}} & \frac{1}{KN} \sum_{k=1}^{K} \sum_{n=1}^{N} \ell\Big(y_{kn}(||\mathbf{x}_k - \mathbf{x}_{kn}^a||_{\mathbf{M}}^2 - ||\mathbf{x}_k - \mathbf{x}_{kn}^b||_{\mathbf{M}}^2)\Big) \\ \text{s.t. } & \mathbf{M} \succeq 0, \\ & ||\mathbf{M}||_F \leq \lambda_M \end{aligned} \tag{3.1.2}$$

The term λ_F is used to regularise the matrix $\mathbf{M}$. However, this is a non-convex problem due to the squared Mahalanobis terms. Therefore, to make this problem convex, we must use a substitution $\mathbf{v}_k = -2\mathbf{M}\mathbf{x}_k$. The optimisation problem we solve is presented in equation 3.1.3.

$$
\begin{aligned}
\min_{\mathbf{M},\{\mathbf{v}_k\}_{k=1}^{K}} & \sum_{k=1}^{K}\sum_{n=1}^{N} \ell\Big(y_{kn}(||\mathbf{x}_{kn}^{a}||_{\mathbf{M}}^{2} - ||\mathbf{x}_{kn}^{b}||_{\mathbf{M}}^{2} + \mathbf{v}_k^{T}(\mathbf{x}_{kn}^{a} - \mathbf{x}_{kn}^{b}))\Big) \\
\text{s.t. } & \mathbf{M} \succeq 0, \\
& ||\mathbf{M}||_F \leq \lambda_M
\end{aligned}
\tag{3.1.3}
$$

This is a convex problem that can be solved using an convex optimisation solver such as SCS (O'Donoghue, 2021).

3.2 Convex layers

To look into convex neural networks using `cvxpylayers`, which is based on Agrawal et al. (2019).

4 Experiments

Experiments section.

References

Agrawal, Akshay et al. (2019). "Differentiable Convex Optimization Layers". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9558–9570. URL: https://proceedings.neurips.cc/paper/2019/hash/9ce3c52fc54362e22053399d3181c638-Abstract.html (Cited on page 8).

Ahmadi, Saba et al. (2022). "On Classification of Strategic Agents Who Can Both Game and Improve". In: *3rd Symposium on Foundations of Responsible Computing (FORC 2022)*. Vol. 218. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 3:1–3:22. DOI: 10.4230/LIPIcs.FORC.2022.3 (Cited on page 5).

Amin, Kareem et al. (2015). "Online Learning and Profit Maximization from Revealed Preferences". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. AAAI Press, pp. 770–776. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9984 (Cited on page 6).

Balcan, Maria-Florina et al. (2014). "Learning Economic Parameters from Revealed Preferences". In: *Web and Internet Economics - 10th International Conference, WINE 2014, Beijing, China, December 14-17, 2014. Proceedings*. Vol. 8877. Springer, pp. 338–353. DOI: 10.1007/978-3-319-13129-0_28 (Cited on page 6).

Bechavod, Yahav et al. (2022). "Information Discrepancy in Strategic Learning". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Vol. 162. PMLR, pp. 1691–1715. URL: https://proceedings.mlr.press/v162/bechavod22a.html (Cited on page 5).

Beigman, Eyal and Rakesh Vohra (2006). "Learning from Revealed Preference". In: *Proceedings of the 7th ACM Conference on Electronic Commerce*. New York, NY, USA: Association for Computing Machinery, pp. 36–42. DOI: 10.1145/1134707.1134712 (Cited on page 6).

Canal, Gregory et al. (2022). "One for All: Simultaneous Metric and Preference Learning over Multiple Users". In: *Advances in Neural Information Processing Systems* 35, pp. 4943–4956. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/1fd4367793bcd3ad38a0b820fcc1b815-Abstract-Conference.html (Cited on pages 6, 7).

Chen, Yiling, Yang Liu, and Chara Podimata (2020). "Learning Strategy-Aware Linear Classifiers". In: *Advances in Neural Information Processing Systems 33*. URL: https://proceedings.neurips.cc/paper/2020/hash/ae87a54e183c075c494c4d397d126a66-Abstract.html (Cited on page 5).

Dong, Jinshuo et al. (2018). "Strategic Classification from Revealed Preferences". In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. Ithaca, NY, USA, pp. 55–70. DOI: 10.1145/3219166.3219193 (Cited on pages 5, 6).

Eilat, Itay et al. (2023). “Strategic Classification with Graph Neural Networks”. In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. URL: https://openreview.net/pdf?id=TuHkVOjSAR (Cited on page 5).

Ghalme, Ganesh et al. (2021). “Strategic Classification in the Dark”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Vol. 139. PMLR, pp. 3672–3681. URL: http://proceedings.mlr.press/v139/ghalme21a.html (Cited on page 5).

Hardt, Moritz et al. (2016). “Strategic Classification”. In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science* (Cambridge, Massachusetts, USA). New York, NY, USA: Association for Computing Machinery, pp. 111–122. DOI: 10.1145/2840728.2840730 (Cited on page 5).

Jagadeesan, Meena, Celestine Mendler-Dünner, and Moritz Hardt (2021). “Alternative Microfoundations for Strategic Classification”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Vol. 139. PMLR, pp. 4687–4697. URL: http://proceedings.mlr.press/v139/jagadeesan21a.html (Cited on page 5).

Karimi, Amir-Hossein, Gilles Barthe, et al. (2022). “A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations”. In: *ACM Comput. Surv.* 55.5. DOI: 10.1145/3527848 (Cited on page 5).

Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera (2021). “Algorithmic Recourse: From Counterfactual Explanations to Interventions”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada). New York, NY, USA, pp. 353–362. DOI: 10.1145/3442188.3445899 (Cited on page 5).

Levanon, Sagi and Nir Rosenfeld (2021). “Strategic Classification Made Practical”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Vol. 139. PMLR, pp. 6243–6253. URL: http://proceedings.mlr.press/v139/levanon21a.html (Cited on page 5).

Levanon, Sagi and Nir Rosenfeld (2022). “Generalized Strategic Classification and the Case of Aligned Incentives”. In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Vol. 162. PMLR, pp. 12593–12618. URL: https://proceedings.mlr.press/v162/levanon22a.html (Cited on page 5).

Nair, Vineet et al. (2022). “Strategic Representation”. In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Vol. 162. PMLR, pp. 16331–16352. URL: https://proceedings.mlr.press/v162/nair22a.html (Cited on page 6).

O’Donoghue, Brendan (2021). “Operator Splitting for a Homogeneous Embedding of the Linear Complementarity Problem”. In: *SIAM J. Optim.* 31.3, pp. 1999–2023. DOI: 10.1137/20M1366307 (Cited on page 8).

Roth, Aaron, Jonathan R. Ullman, and Zhiwei Steven Wu (2016). “Watch and Learn: Optimizing from Revealed Preferences Feedback”. In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*. ACM, pp. 949–962. DOI: 10.1145/2897518.2897579 (Cited on page 6).

Voigt, Paul and Axel von dem Bussche (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide*. 1st ed. Springer Publishing Company, Incorporated. DOI: 10.1007/978-3-319-57959-7 (Cited on page 5).

Zadimoghaddam, Morteza and Aaron Roth (2012). "Efficiently Learning from Revealed Preference". In: *Internet and Network Economics - 8th International Workshop, WINE 2012, Liverpool, UK, December 10-12, 2012. Proceedings*. Vol. 7695. Springer, pp. 114–127. DOI: 10.1007/978-3-642-35311-6_9 (Cited on page 6).

Zrnic, Tijana et al. (2021). "Who Leads and Who Follows in Strategic Classification?" In: *Advances in Neural Information Processing Systems 34*, pp. 15257–15269. URL: https://proceedings.neurips.cc/paper/2021/hash/812214fb8e7066bfa6e32c626c2c688b-Abstract.html (Cited on page 6).