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 26, 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.,

$$\begin{aligned} \mathbf{x}^f = &\operatorname*{argmin}_{\mathbf{x}' \in \mathcal{X}} c(\mathbf{x}, \mathbf{x}') \\ \text{s.t.}\ \ & h(\mathbf{x}') = 1, \\ & c(\mathbf{x}, \mathbf{x}') \leq B \end{aligned} \tag{2.1.1}$$

- 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

As discussed in section 2, the cost of changing features from $\mathbf{x}$ to $\mathbf{x}'$ in the strategic classification literature is typically a quadratic cost function of the form $c(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^2$ [**ADD CITATIONS**], or occasionally a quadratic form cost function $c(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T \mathbf{M}(\mathbf{x} - \mathbf{x}')$ where $\mathbf{M}$ is a fixed, known, square matrix (Bechavod et al., 2022).

However, these do not necessarily represent the true complexities of the cost of moving from $\mathbf{x}$ to $\mathbf{x}'$, for a number of (non-exhaustive) reasons. Consider the case of a individuals applying for a line of credit.

1. **Changing one feature can change the cost of changing another feature**. If an individual decides not to inquire about a loan for a number of months (which will change the feature “number of inquiries in the last 6 months”, the cost of decreasing the feature “number of inquiries in the last 6 months, excluding the last 7 days” will be very low or zero. However, if a quadratic cost function (or any L_p norm cost function) is used, this will be interpreted as two separate feature changes and the costs will be summed. Whilst this simple case can likely be handled by domain expertise, more complex causal relations will exist. Consider an individual obtaining two more credit cards. Whilst this may reduce the cost of increasing “number of credit cards”, this may also increase the cost of “monthly credit card payments” and may have less clear effects (which need not be linear) on other features.

2. **Changing feature costs can be different for different individuals**. For example, increasing the number of credit cards from 1 to 5 may be much easier for someone with a higher income or increasing income from £25,000 to £30,000 may be much easier for someone with a higher level of education. These are all modelled as the same in typical cost functions used in the literature.

[**ADD IN SECTION ON SPECIFYING CAUSAL GRAPHS**]

We do not observe the cost function itself, but one way to approximate it is to *learn from revealed preferences* (see section 2.3). We propose that each individual who is negatively classified is presented with N pairs of recourse options $(\mathbf{x}_n^a, \mathbf{x}_n^b)$. If option a is selected, then we assume $c(\mathbf{x}, \mathbf{x}_n^a) \leq c(\mathbf{x}, \mathbf{x}_n^b)$ and if option b is selected, then we assume $c(\mathbf{x}, \mathbf{x}_n^a) \geq c(\mathbf{x}, \mathbf{x}_n^b)$. The responses to the pairs of recourse options presented (the pairwise comparisons) reveal information about the individuals' preferences over recourse options, i.e., their cost functions.

Once a cost function is learned, we need to solve the constrained optimisation problem mentioned in equation 2.1.1 to generate the recourse package $\mathbf{x}^f$, where $\mathbf{x}$ are the individual's original features, h 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.

$$
\begin{aligned}
\mathbf{x}^f = & \operatorname*{argmin}_{\mathbf{x}' \in \mathcal{X}} c(\mathbf{x}, \mathbf{x}') \\
\text{s.t. } & h(\mathbf{x}') = 1, \\
& c(\mathbf{x}, \mathbf{x}') \leq B
\end{aligned} \tag{3.0.1}
$$

This chapter goes on to detail learning different functional forms for the cost function c. For a linear classifier and convex cost function (which the literature focuses on), this is typically handled as a convex optimisation problem. Therefore, in the following sections, where convex costs functions are learned, convex optimisation is used to find the recourse package $\mathbf{x}^f$.

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 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 $(\mathbf{x}^a_{kn}, \mathbf{x}^b_{kn})$. The response y_{kn} is defined in equation 3.1.2, where c^G_k represents the ground truth cost function of individual k.

$$y_{kn} = \begin{cases} -1 & \text{if } c^G_k(\mathbf{x}_{kn}, \mathbf{x}^a_{kn}) \leq c^G_k(\mathbf{x}_{kn}, \mathbf{x}^b_{kn}) \\ +1 & \text{if } c^G_k(\mathbf{x}_{kn}, \mathbf{x}^a_{kn}) > c^G_k(\mathbf{x}_{kn}, \mathbf{x}^b_{kn}) \end{cases} \tag{3.1.2}$$

To optimise for $\mathbf{M}$, we compare the squared Mahalanobis distances between $\mathbf{x}_k$ and $\mathbf{x}^a_{kn}$ and between $\mathbf{x}_k$ and $\mathbf{x}^b_{kn}$. The optimisation problem to learn $\mathbf{M}$ is shown in equation 3.1.3, where ℓ represents either the hinge or logistic loss function. The optimisation is an adaptation of the optimisation problem presented in Canal et al. (2022).

[TO ADD EXPLANATION ON WHY THIS IS A GOOD OPTIMISATION]

$$\begin{aligned} &\min_{\mathbf{M}} \frac{1}{KN}\sum_{k=1}^{K}\sum_{n=1}^{N} \ell\Big(y_{kn}(||\mathbf{x}_k-\mathbf{x}^a_{kn}||^2_{\mathbf{M}} - ||\mathbf{x}_k-\mathbf{x}^b_{kn}||^2_{\mathbf{M}})\Big) \\ &\text{s.t. } \mathbf{M} \succeq 0, \end{aligned} \tag{3.1.3}$$

3.2 Convex layers

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

4 Experiments

4.1 Synthetic data

The experiments that follow use synthetic data generated from a structural causal model, which is shown in Figure 4.1.

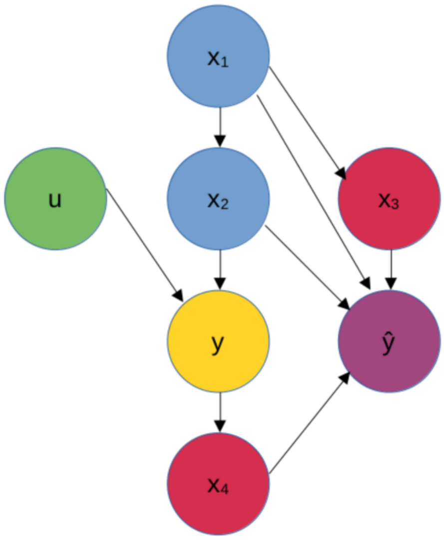

Figure 4.1: The Structural Causal Model used for synthetic data generation

The structural causal model contains 2 unobserved variables

- $\mathbf{u}$ - which is sampled from a normal distribution
- $\mathbf{y}$ - the true binary outcome which is a linear combination of $\mathbf{u}$ and $\mathbf{x}_2$

And 4 observed variables

- $\mathbf{x}_1$ - which is sampled from a normal distribution
- $\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ - which are linear combinations of other variables
- $\hat{\mathbf{y}}$ - The predicted value of $\mathbf{y}$

4.2 Simulation process

[TO CONVERT TO AN ALGORITHM/PSEUDOCODE]

The process for simulation is as follows. We assume that the individuals do *not* have any knowledge of the classifier.

1. Split the data into test and train sets.
2. Fit a classifier using just the train set and measure accuracy against the *true* labels.
3. Predict labels for *all* data points (both train and test)

4. Calculate recourse actions for all negatively classified data points, by minimising the current approximation of the cost function.

5. Perturb the recourse actions to create pairwise comparisons for the negatively classified individuals to evaluate.

6. Learn cost function from evaluated pairwise comparisons from *all* previous iterations.

7. Generate updated recourse with the current approximation of the cost function.

8. For individuals with a ground truth cost smaller than the cost threshold B, they modify their features to those suggested by recourse. If the ground truth cost of recourse is greater than C, we assume no action is taken.

9. Update the predicted and true labels using the structural causal model (not all instances that involve recourse lead to a change in the true label).

10. Repeat steps 2-9 for a set number of iterations, keeping the same individuals in the train/test sets.

4.3 Mahalanobis distance

Figure 4.2 shows the cumulative proportion of individuals who were negatively classified and have been offered recourse packages with a ground truth cost $c^G(\mathbf{x},\ \mathbf{x'}) < B$. If the approximated cost function $c(\mathbf{x},\mathbf{x'})$ is close to the ground truth, the recourse suggestions will be closer to the optimal recourse. If the approximated cost is not similar to the ground truth cost function, the recourse suggestions will be further from the optimal recourse and it will be less likely that an individual is successfully offered recourse. The learned cost function outperforms the quadratic and other Mahalanobis cost functions. Figure 4.2 has an assumed ground truth cost function which takes a Mahalanobis form, with a randomly sampled $\mathbf{M}$. Additionally, it is assumed that this ground truth function is the same for all individuals.

The next steps to improve this are:

- Design a better method for generating the pairwise comparisons (perhaps look into techniques akin to Bayesian Optimisation).
- Introduce different ground truth cost functions.
- Consider *drift* - where features change over time independently from loan application.

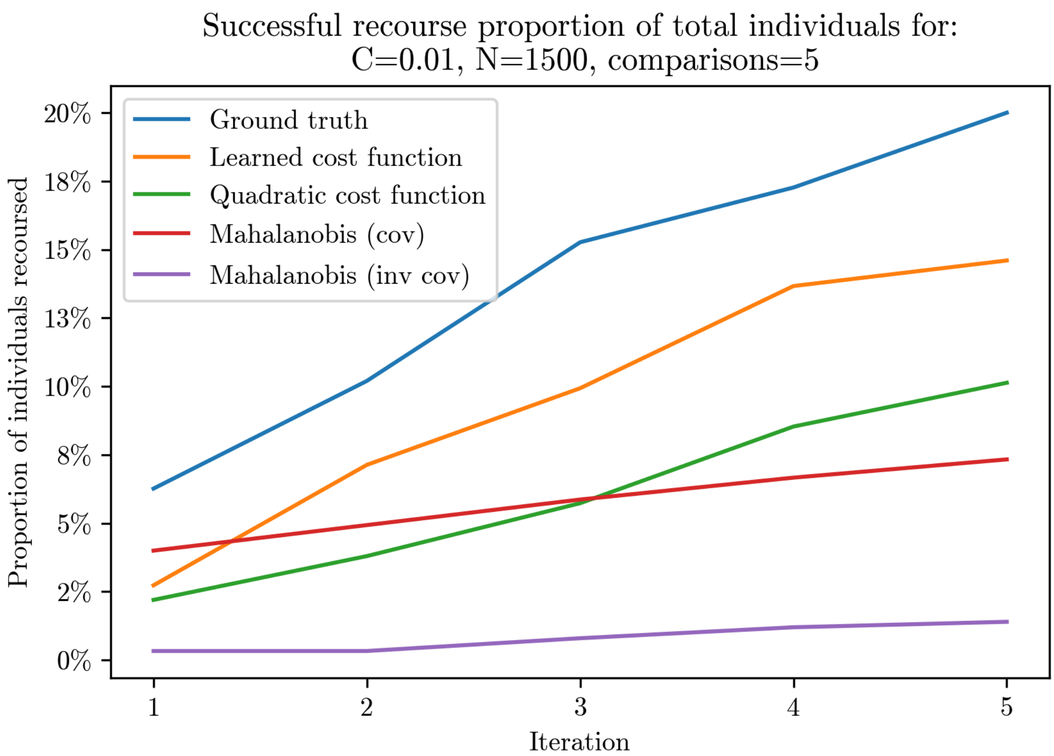

Figure 4.2: Cumulative proportion of individuals successfully offered recourse

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 pages 5, 7).

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, 8).

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 6).

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 6).

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).

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).