Towards Bridging the Gaps between the Right to Explanation and the Right to be Forgotten

Satyapriya Krishna [*1] **Jiaqi Ma** [*1] **Himabindu Lakkaraju** [1]

Abstract

The Right to Explanation and *the Right to be Forgotten* are two important principles outlined to regulate algorithmic decision making and data usage in real-world applications. While the right to explanation allows individuals to request an actionable explanation for an algorithmic decision, the right to be forgotten grants them the right to ask for their data to be deleted from all the databases and models of an organization. Intuitively, enforcing the right to be forgotten may trigger model updates which in turn invalidate previously provided explanations, thus violating the right to explanation. In this work, we investigate the technical implications arising due to the interference between the two aforementioned regulatory principles, and propose *the first algorithmic framework* to resolve the tension between them. To this end, we formulate a novel optimization problem to generate explanations that are robust to model updates due to the removal of training data instances by data deletion requests. We then derive an efficient approximation algorithm to handle the combinatorial complexity of this optimization problem. We theoretically demonstrate that our method generates explanations that are provably robust to worst-case data deletion requests with bounded costs in case of linear models and certain classes of non-linear models. Extensive experimentation with real-world datasets demonstrates the efficacy of the proposed framework.

1. Introduction

Over the past decade, machine learning models have been increasingly deployed in various high-stakes decision making scenarios including hiring and loan approvals. Consequently, a number of regulatory policies and principles (GDPR, 2016; CCPA, 2018) were introduced to ensure that algorithmic decisions and data usage practices in real-world applications do not cause any undue harm to individuals. *The Right to Explanation* and *the Right to be Forgotten* are two such notable regulatory principles which were first introduced by the European Union's General Data Protection Regulation (GDPR) (GDPR, 2016). While the right to explanation ensures that individuals who are negatively impacted by adverse algorithmic outcomes are provided with an actionable explanation, the right to be forgotten ensures that individuals have the right to ask for their data to be removed from all the databases and models of an organization.

To operationalize the right to explanation in practice, several strategies have been considered in recent literature. A particular class of explanations commonly referred to as *counterfactual explanations* or *algorithmic recourse* are often considered very promising in this regard. For instance, when an individual is denied a loan by a predictive model employed by a bank, a counterfactual explanation (or an algorithmic recourse) provides them with inputs about what aspects (features) of their profile should be changed and by how much in order to obtain a positive outcome. Several approaches in recent literature tackled the problem of generating such counterfactual explanations (Wachter et al., 2018; Ustun et al., 2019; Pawelczyk et al., 2020; Karimi et al., 2020).

Prior research has also explored various strategies to operationalize the right to be forgotten (Cao & Yang, 2015; Ginart et al., 2019; Garg et al., 2020). Since the right to be forgotten requires organizations to delete pertinent user data from all their databases and models, it often involves retraining or updating their models. To this end, several methods were proposed to efficiently update machine learning models in the face of (training) data deletion requests without having to retrain them from scratch (Guo et al., 2020; Bourtoule et al., 2021; Izzo et al., 2021; Neel et al., 2021).

Despite the significance of the two aforementioned regulatory principles, there is very little research that explores potential

[*]Equal contribution [1]Harvard University. Correspondence to: Himabindu Lakkaraju <hlakkaraju@hbs.edu>.

interference between them. Intuitively, enforcing the right to be forgotten may trigger model updates which in turn invalidate previously provided actionable explanations that end users may act upon, thus violating the right to explanation. For instance, consider a scenario where a user was asked to increase their salary by 5K to get a loan and they start working towards it, but the underlying model gets updated in the meanwhile to accommodate (training) data deletion requests. Consequently, the user may no longer receive the desired outcome even if their salary increases by 5K as the previously prescribed recourse may no longer hold with respect to the new model. Pawelczyk et al. (2022b) highlighted this challenge and argued that the right to explanation and the right to be forgotten are in conflict with each other, and that existing methods are not capable of dealing with this tension.

In this work, we make one of the first attempts to resolve the aforementioned tension and bridge the operational gaps between the right to explanation and the right to be forgotten. More specifically, we propose the *first algorithmic framework*, RObust Counterfactual Explanations under the Right to be Forgotten (ROCERF), to address this problem. To this end, we formulate a novel optimization problem to generate counterfactual explanations that remain valid in the face of model updates (changes) arising due to (training) data deletion requests. This optimization problem turns out to be combinatorially complex as it considers n training instances and k data deletion requests resulting in $\binom{n}{k}$ possible ways of the model being updated. To mitigate this computational challenge, we propose a novel algorithm which can efficiently approximate model updates relative to the original model, and select those with most significant deviations, thus eliminating the need for retraining $\binom{n}{k}$ models. With this approximation, we are able to develop a practically efficient algorithm to learn effective counterfactual explanations that remain valid on model updates triggered by data deletion requests.

We theoretically and empirically analyze the validity and costs of the counterfactual explanations generated by our framework ROCERF. In case of linear models and non-linear models with certain regularity assumptions, we theoretically demonstrate that our method generates counterfactual explanations that are provably valid in the face of worst-case data deletion requests, while incurring additional costs upper bounded by $O(\frac{k}{n})$. Empirically, we evaluate the proposed ROCERF and state-of-the-art counterfactual explanation methods using logistic regression and neural network models on three real-world datasets. The proposed method outperforms baseline methods in most experimental settings. In comparison, baseline methods either fail dramatically in terms of validity, or achieve high validity with significantly higher cost. Our results establish that our framework ROCERF enables us to simultaneously enforce both the right to explanation as well as the right to be forgotten, thus bridging a critical operational gap between the two regulatory principles.

2. Related Work

Over the past few years, there has been a lot of exciting research on counterfactual explanations or algorithmic recourse (Tolomei et al., 2017; Laugel et al., 2017; Wachter et al., 2017; Ustun et al., 2019; Van Looveren & Klaise, 2019; Mahajan et al., 2019; Mothilal et al., 2020; Karimi et al., 2020; Rawal & Lakkaraju, 2020; Dandl et al., 2020). Several of the proposed approaches can be roughly categorized along the following dimensions (Verma et al., 2020b): *type of the underlying predictive model* (e.g., tree based vs. differentiable classifier), whether they encourage *sparsity* in counterfactuals (i.e., only a small number of features should be changed), whether counterfactuals should lie on the *data manifold* and whether the underlying *causal relationships* should be accounted for when generating counterfactuals. Most of these approches assume that the underlying predictive model remains unchanged before and after the end users implement the prescribed recourses.

More recently, few studies have investigated the impact of changes in the underlying predictive models on the the validity of recourses (Rawal et al., 2021; Upadhyay et al., 2021). To improve the robustness of the recourses in the face of such model changes, prior work has proposed adversarial training methods that generate counterfactual explanations robust to small (and often Gaussian) perturbations of the underlying model parameters (Upadhyay et al., 2021). While such methods could potentially be considered to mitigate the challenges brought about by the right to be forgotten, it is unclear how the removal of training data points will affect the model parameters. There is no guarantee that counterfactual explanations robust to Gaussian perturbations of model parameters will be valid under model updates (changes) occurring due to data deletion requests.

On the other hand, the right to be forgotten has also inspired considerable research in machine learning literature (Cao & Yang, 2015; Ginart et al., 2019; Garg et al., 2020; Guo et al., 2020; Bourtoule et al., 2021; Izzo et al., 2021; Neel et al., 2021). Majority of work along these lines focuses on developing methods to efficiently update models in the face of training data deletion requests, without having to retrain models from scratch. Such approaches are referred to as *Machine Unlearning* methods.

To the best of our knowledge, the only prior work at the intersection of the right to explanation and the right to be forgotten is by Pawelczyk et al. (2022b). Pawelczyk et al. (2022b) analyzed the impact of (training) data deletion requests on the validity of counterfactual explanations generated by existing methods, and concluded that the explanations generated by state-of-the-art methods become invalid in the face of model updates due to data deletion requests. While the above work highlighted the tension between the right to explanation and the right to be forgotten, they do not provide a solution to this critical problem. In contrast, our work proposes the first algorithmic framework to address this tension.

3. Our Framework ROCERF

In this section, we introduce our framework, RObust Counterfactual Explanations under the Right to be Forgotten (ROCERF). Specifically, we first formally define the problem of finding robust counterfactual explanations in the presence of training data removal required by the right to be forgotten. Then, we present an efficient approximation algorithm to solve this problem. We also discuss practical considerations including computation costs and further approximations in implementation.

3.1. Problem Definition

Suppose we have a training dataset $D = \{(x_i, y_i)\}_{i=1}^n$ with n data points, where $x_i \in \mathcal{X}, y_i \in \{-1, 1\}$ are respectively features and labels. Given a family of classifiers $f_\theta : \mathcal{X} \to \mathbb{R}$ parameterized by $\theta \in \Theta$, a classifier f_θ predicts 1 on a data point x if $f_\theta(x) \geq 0$ and predicts -1 otherwise. Assume $B := \sup_{x \in \mathcal{X}} \|x\|_2$ is $O(1)$ in terms of n.

To characterize the data removal, we introduce a data weight vector $w \in \{0, 1\}^n$. For each data point i, let $w_i = 0$ if this data point is removed, and let $w_i = 1$ otherwise. Specially, when $w = \mathbf{1}$, where $\mathbf{1}$ is an all-one vector, there is no data point being removed.

Denote the loss of f_θ on each data point i as $l_i(\theta)$ and assume $l_i(\theta)$ has continuous second derivatives. We denote the classifier trained on the dataset D as $f_{\hat{\theta}_\mathbf{1}}$, where $\hat{\theta}_\mathbf{1} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n l_i(\theta)$. A classifier trained on the dataset with some removals indicated by w can then be denoted as $f_{\hat{\theta}_w}$, where $\hat{\theta}_w = \arg\min_{\theta \in \Theta} \frac{1}{\|w\|_1} \sum_{i=1}^n w_i l_i(\theta)$. To simplify notations, for $i = 1, 2, \ldots n$, define $g_i(\theta) := \frac{\partial l_i(\theta)}{\partial \theta}$ and $h_i(\theta) := \frac{\partial g_i(\theta)}{\partial \theta^T}$. Then $H := \frac{1}{n} \sum_{i=1}^n h_i(\hat{\theta}_\mathbf{1})$ is the Hessian matrix of the loss function on the whole dataset D.

In the literature (Wachter et al., 2017; Verma et al., 2020a), the problem of finding counterfactual explanations (CFEs) for the model $f_{\theta_\mathbf{1}}$ trained on the original full dataset is often defined as an optimization problem like the following Definition 3.1.

Definition 3.1 (Counterfactual Explanation (CFE))**.** For any data point $x_0 \in \mathcal{X}$, the *CFE* ($\tilde{x}_0 \in \mathcal{X}$) of x_0, is defined as the solution of the following optimization problem

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \|x - x_0\|_2 \\ \text{subject to} \quad & f_{\hat{\theta}_\mathbf{1}}(x) \geq 0. \end{aligned} \tag{1}$$

Intuitively, we hope to find a *valid* CFE (classified as 1) with the minimum *cost*, as measured by the L2 distance to x_0. In practice, the L2 distance could be replaced by other distance functions, such as L1 distance or any other metrics suitable for the application. In this paper, however, we stick to the L2 distance following the convention of recent literature (Pawelczyk et al., 2021).

In this paper, we aim to obtain CFEs that is robustly valid against potential data point removal required by right to be forgotten. To formalize this problem, we define the following *k-Removal-Robust CFE (kRR-CFE)* that is supposed to be robust with respect to any removal of k data points.

Definition 3.2 (k-Removal-Robust CFE (kRR-CFE))**.** Given an integer $k > 0$, denote the set of all possible weight vectors with k data removals as $\mathcal{W}^{(k)} = \{w \in \{0, 1\}^n : \|w\|_1 = n - k\}$. For any data point $x_0 \in \mathcal{X}$, the *k-RR CFE* ($\tilde{x}_0^{(k)} \in \mathcal{X}$) of x_0, is defined as the solution of the following optimization problem

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \|x - x_0\|_2 \\ \text{subject to} \quad & f_{\hat{\theta}_w}(x) \geq 0, \forall w \in \mathcal{W}^{(k)}. \end{aligned} \tag{2}$$

While the k-RR CFE defined in Definition 3.2 is robust to any removal of k data points by construction, a naive implementation to obtain the k-RR CFE requires one to retrain $|\mathcal{W}^{(k)}| = \binom{n}{k}$ classifiers that appear in the constraint of the optimization problem (2), which is computationally impractical.

3.2. Approximating k-RR CFE

To address the computational challenge, we propose an efficient algorithm to approximate k-RR CFE. The proposed method first efficiently approximates the classifier $f_{\hat{\theta}_w}$, for any $w \in \mathcal{W}^{(k)}$, without the need of retraining from scratch. Then we show that, we can reduce the constraint set with $\binom{n}{k}$ classifiers to an equivalent constraint that only requires a linear computation complexity with respect to n.

Approximating the Classifier. A key observation that makes it possible to efficiently approximate $f_{\hat{\theta}_w}$ is that these classifiers can be viewed as *leave-k-out (LKO)* estimators, which can be efficiently approximated by leveraging recent advances in LKO analysis (Giordano et al., 2019; Broderick et al., 2020).

For each of the classifier $f_{\hat{\theta}_w}(x)$ in the constraint set of the problem 2, note that $f_{\hat{\theta}_w}(x)$ is also a function of w[1]. For any fixed x, we can take a first-order Taylor approximation of $f_{\hat{\theta}_w}(x)$ with respect to w at $w = \mathbf{1}$, and denote this first-order approximation as $\tilde{f}_{\hat{\theta}_w}(x)$, i.e.,

$$\tilde{f}_{\hat{\theta}_w}(x) = f_{\hat{\theta}_\mathbf{1}}(x) + \frac{\partial f_{\hat{\theta}_w}(x)}{\partial w}\Big|_{w=\mathbf{1}}(w-\mathbf{1}) \tag{3}$$

$$= f_{\hat{\theta}_\mathbf{1}}(x) + \frac{\partial f_\theta(x)}{\partial \theta}\Big|_{\theta=\hat{\theta}_\mathbf{1}} \frac{\partial \hat{\theta}_w}{\partial w}\Big|_{w=\mathbf{1}}(w-\mathbf{1}), \tag{4}$$

where from Eq. (3) to Eq. (4), we have applied the chain rule. Using results in Giordano et al. (2019), we can show that[2]

$$\frac{\partial \hat{\theta}_w}{\partial w}\Big|_{w=\mathbf{1}}(w-\mathbf{1}) = \frac{1}{n}\sum_{i:w_i=0} H^{-1} g_i(\hat{\theta}_\mathbf{1}). \tag{5}$$

Therefore, the first-order Taylor approximation can be written as

$$\tilde{f}_{\hat{\theta}_w}(x) = f_{\hat{\theta}_\mathbf{1}}(x) + \frac{1}{n}\sum_{i:w_i=0} \beta(x)^T H^{-1} g_i(\hat{\theta}_\mathbf{1}), \tag{6}$$

where $\beta(x) := \left(\frac{\partial f_\theta(x)}{\partial \theta}\Big|_{\theta=\hat{\theta}_\mathbf{1}}\right)^T$ is the gradient of $f_\theta(x)$ with respect to the model parameters θ at $\theta = \hat{\theta}_\mathbf{1}$.

Replacing $f_{\hat{\theta}_w}$ with $\tilde{f}_{\hat{\theta}_w}$, we approximate the problem (2) with a new problem below.

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & \|x - x_0\|_2 \\ \text{subject to} \quad & \tilde{f}_{\hat{\theta}_w}(x) \geq \delta, \forall w \in \mathcal{W}^{(k)}, \end{aligned} \tag{7}$$

where $\delta > 0$ is a constant accounting for the approximation error of $\tilde{f}_{\hat{\theta}_w}$ for $f_{\hat{\theta}_w}$, which should be chosen in a way such that for any $w \in \mathcal{W}^{(k)}$ and $x \in \mathcal{X}$, $\tilde{f}_{\hat{\theta}_w}(x) \geq \delta$ implies $f_{\hat{\theta}_w}(x) \geq 0$. The proper choice of δ is model dependent and, in practice, can be treated as a hyperparameter selected using a validation set. We also provide some theoretical insights on δ in Section 4.

Reducing the Constraint Set. With the first-order Taylor approximate classifier $\tilde{f}_{\hat{\theta}_w}$, we can further reduce the constraint set with $\binom{n}{k}$ inequalities to a single inequality.

Note that in Eq. (6), $f_{\hat{\theta}_\mathbf{1}}(x), \beta(x), H$, and $g_i(\hat{\theta}_\mathbf{1})$ are all calculated based on the model $f_{\hat{\theta}_\mathbf{1}}$ trained on the original full dataset D, and are independent of the data weight vector w. Define a set $\mathcal{A}(x) := \{\beta(x)^T H^{-1} g_i(\hat{\theta}_\mathbf{1})\}$. Then satisfying the constraints in the problem (7) is equivalent to having the following condition.

$$f_{\hat{\theta}_\mathbf{1}}(x) + \frac{1}{n}\min_{\mathcal{B}\subseteq\mathcal{A}(x), |\mathcal{B}|=k} \sum_{b\in\mathcal{B}} b \geq \delta, \tag{8}$$

[1] Strictly speaking, we need to assume uniqueness of $\hat{\theta}_w$ for any given w. Although in practice we can often make this assumption approximately hold locally.

[2] See Proposition 3 in Appendix A.3 of Giordano et al. (2019).

where one shall recall that $w \in \mathcal{W}^{(k)}$ always has k entries as 0 and the remaining as 1.

Defining

$$f_{\mathcal{A}}^{(k)}(x) := f_{\hat{\theta}_{\mathbf{1}}}(x) + \frac{1}{n} \min_{\mathcal{B} \subseteq \mathcal{A}(x), |\mathcal{B}|=k} \sum_{b \in \mathcal{B}} b,$$

we have shown that solving the problem (7) is equivalent to solving the problem below.

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \|x - x_0\|_2 \\ \text{subject to} \quad & f_{\mathcal{A}}^{(k)}(x) \geq \delta. \end{aligned} \tag{9}$$

Optimization. To solve the constrained optimization problem (9), we use the the penalty method (Freund, 2004). Define the penalty function as $\phi(z) := \max(z, 0)^2$. We solve a series of unconstrained relaxation of the original problem (9):

$$\min_{x \in \mathcal{X}} \quad J_t(x) = \lambda_t \phi(\delta - f_{\mathcal{A}}^{(k)}(x)) + \|x - x_0\|_2, \tag{10}$$

for $t = 1, 2, \ldots T$ as the iteration index. And $\lambda_t \geq 0$ is the penalty coefficient controlling the relative strength between the penalty and the original objective for each iteration t. Denote the solution of the t-th iteration as x_t^*. We start with a small λ_1 and double it until the first t_0 where $f_{\mathcal{A}}^{(k)}(x_{t_0}^*) \geq \delta$ while $f_{\mathcal{A}}^{(k)}(x_{t_0-1}^*) < \delta$. Then we have a binary search on the penalty coefficient between λ_{t_0-1} and λ_{t_0} to obtain a feasible solution with as small cost as possible. Please see Algorithm 1 in Appendix A for more details.

3.3. Practical Considerations

Computation Costs. Finally, we make a few remarks on the computation costs of the proposed method. Given the dataset D and the original model $f_{\hat{\theta}_{\mathbf{1}}}$ trained on D, we need to first calculate the gradients $g_i(\hat{\theta}_{\mathbf{1}}), i = 1, \ldots, n$ and the Hessian inverse H^{-1}. Calculating the exact Hessian inverse may be expensive for models with high-dimensional parameters, such as neural networks. However, we can leverage computational tricks calculating influence functions (Koh & Liang, 2017) to efficiently approximate H^{-1}. In addition, H^{-1} and $g_i(\hat{\theta}_{\mathbf{1}})$'s only need to be calculated once for the whole process and are shared for all the test samples.

The major computation cost comes from repeated evaluations of $f_{\mathcal{A}}^{(k)}(x)$ at different x during the optimization procedure. For each x, we can use automatic differentiation tools such as PyTorch (Paszke et al., 2017) to evaluate $f_{\hat{\theta}_{\mathbf{1}}}(x)$ and $\beta(x)$ by one forward pass and one backward pass. Suppose we have pre-computed and stored the values of $H^{-1}g_i(\hat{\theta}_{\mathbf{1}}), i = 1, \ldots, n$, we can obtain the set $\mathcal{A}(x)$ by n vector multiplications. Finally, evaluating $f_{\mathcal{A}}^{(k)}(x)$ requires partially sorting $\mathcal{A}(x)$ and obtaining the bottom-k values, which has a complexity of $O(n \log k)$. Overall, evaluating the constraint of problem (9) has a complexity that is linear in n, which is much smaller than evaluating the original constraint set with $\binom{n}{k}$ models.

Hyperparameters. The proposed method has two hyperparameters, k and δ. The hyperparameter k should be set from a rough estimate of the number of data removals, which relies on domain knowledge of the application. Empirically, however, we find the method is not very sensitive to the value of k so there is a good tolerance on the choice of k. The hyperparameter δ measures how good is the Taylor approximation of the function. In practice, we can choose δ on a validation set and simulating a few models trained after random removals. But in our experiments, we find fixing it as 0 also works well empirically.

A Special Case: Linear Models. When $f_\theta(x) = \theta^T x$ is a linear model, the first-order Taylor approximation in Eq. (6) simplifies to the following form:

$$\tilde{f}_{\hat{\theta}_w}(x) = \hat{\theta}_{\mathbf{1}}^T x + \frac{1}{n} \sum_{i: w_i = 0} x^T H^{-1} g_i(\hat{\theta}_{\mathbf{1}}), \tag{11}$$

since $\beta(x) = x$ for linear models. In this special case, we can avoid going through the backward pass when evaluating $\beta(x)$, which makes the optimization much more efficient.

Local Linear Approximation of Nonlinear Models. Owing to computational efficiency considerations, it is a common practice in recourse literature to first obtain a local linear approximation of the underlying model at each test sample, and then leverage this to compute counterfactual explanations (Upadhyay et al., 2021; Ustun et al., 2019; Rawal & Lakkaraju, 2020). Along similar lines, we propose to apply ROCERF on local linear approximations of nonlinear models to further improve the computational efficiency in practice. Specifically, we use LIME (Ribeiro et al., 2016) to obtain local linear approximations of the underlying models.

4. Theoretical Analysis of Validity and Cost

In this section, we provide theoretical guarantees on validity and cost of CFEs obtained by the proposed method, under a small fraction of data removal in the training set. In particular, we characterize the trade-off between validity and cost and provide upper bounds on the cost needed to guarantee that the CFE is robustly valid. We first present an analysis for linear models and then for nonlinear models with regularity assumptions.

4.1. Analysis on Linear Models

Assume the machine learning models are regularized logistic regression, i.e., $l_i(\theta) = \log(1 + \exp(-y_i\theta^T x_i)) + \gamma\|\theta\|_2^2$, and the model parameters have bounded norm. In this case, the following Theorem 4.1 provides theoretical guarantees on the validity and cost, and the detailed proof of which can be found in Appendix B.

Theorem 4.1 (Validity and Cost on Logistic Regression)**.** *For any data point $x_0 \in \mathcal{X}$, let $\tilde{x}_0$ be the CFE of x_0, and let $\tilde{x}_0^{(k)}$ be the solution of the optimization problem (9) when the classifiers are regularized logistic regression. Then we can properly choose δ such that, if $\tilde{x}_0^{(k)}$ exists, $\tilde{x}_0^{(k)}$ remains a valid CFE for all possible removal of k data points, i.e.,*

$$f_{\hat{\theta}_w}(\tilde{x}_0^{(k)}) \geq 0, \forall w \in \mathcal{W}^{(k)}.$$

Furthermore, the cost of implementing $\tilde{x}_0^{(k)}$ is upper bounded as following,

$$\|\tilde{x}_0^{(k)} - x_0\|_2 \leq \|\tilde{x}_0 - x_0\|_2 + \frac{kC}{n\|\hat{\theta}_{\mathbf{1}}\|_2}, \tag{12}$$

where C is a constant independent of n.

Proof Sketch. The proof of Theorem 4.1 involves two key steps. The first step is to derive a bound on the difference between the actual retrained model $f_{\hat{\theta}_w}$ and its Taylor approximation model $\tilde{f}_{\hat{\theta}_w}$. This bound gives us an estimate on how large δ is needed in the optimization problem (9) in order to ensure validity. The second step is to derive a bound on the difference between the Taylor approximation model $\tilde{f}_{\hat{\theta}_w}$ and the original model $f_{\hat{\theta}_{\mathbf{1}}}$. For regularized logistic regression, both differences can be well bounded without further assumptions. Note that the difference between $\tilde{x}_0^{(k)}$ and $\tilde{x}_0$ is that the former is constrained by $\tilde{f}_{\hat{\theta}_w}(\tilde{x}_0^{(k)}) \geq \delta$ while the latter is constrained by $f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) \geq 0$. So together with the estimate on δ, the bound on the difference between $\tilde{f}_{\hat{\theta}_w}$ and $f_{\hat{\theta}_{\mathbf{1}}}$ allows us to bound the additional cost of $\tilde{x}_0^{(k)}$ in comparison to $\tilde{x}_0$. □

This result states that the additional cost needed to achieve robust validity has an upper bound of $O(\frac{k}{n})$, and this additional cost vanishes when the training set size n is very large and the number of removals k is relatively small. As a sanity check, in the degenerate case where there is no data removed, i.e., $k = 0$, there is also no additional cost.

This result also indicates that for simple models trained on abundant data, it is possible to provide robustly valid recourses to users with little additional costs, thus paving the way for *bridging critical operational gaps between the right to explanation and the right to be forgotten*. The technical insight behind this strong guarantee is that, when the number of data removals is not too large compared to the training set, the retrained model will not change too much (difference between $f_{\hat{\theta}_w}$ and $f_{\hat{\theta}_{\mathbf{1}}}$), and the change can be efficiently estimated (through $\tilde{f}_{\hat{\theta}_w}$).

Remark 4.2. Technically, neither the problem (2) nor its approximation (9) is guaranteed to be feasible. However, especially for linear models, we find that they are always feasible on the datasets we empirically tested. This is possibly because the difference among $f_{\hat{\theta}_w}$ for all $w \in \mathcal{W}^{(k)}$ is not dramatically large.

4.2. Analysis on Nonlinear Models

Next, we generalize Theorem 4.1 to nonlinear models with the following assumptions.

Assumption 4.3. Assume that there exist universal finite constants C_1, C_2, C_3, C_4, C_5 independent of n such that

1. $\sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \|h_i(\theta)\|_F \leq C_1$;
2. $\sup_{\theta\in\Theta} \|g_i(\theta)\|_2 \leq C_2, i = 1, \ldots, n$;
3. $\sup_{x\in\mathcal{X}} \|\beta(x)\|_2 \leq C_3$;
4. $H(\theta) := \frac{1}{n}\sum_{i=1}^{n} h_i(\theta)$ is nonsingular and
$$\sup_{\theta\in\Theta} \|H(\theta)^{-1}\|_{\text{op}} \leq C_4;$$
5. there exists suitable $\Delta > 0$, such that
$$\sup_{\|\theta-\hat{\theta}_{\mathbf{1}}\|_2<\Delta} \frac{1}{n}\sum_{i=1}^{n} \|h_i(\theta) - h_i(\hat{\theta}_{\mathbf{1}})\|_F \leq C_5\|\theta - \hat{\theta}_{\mathbf{1}}\|_2.$$

Theorem 4.4 (Validity and Cost on Nonlinear Models)**.** *For any data point $x_0 \in \mathcal{X}$, let $\tilde{x}_0$ be the CFE of x_0, and let $\tilde{x}_0^{(k)}$ be the solution of the optimization problem (9). Assume the classifiers satisfy Assumption 4.3. Then we can properly choose δ such that, if $\tilde{x}_0^{(k)}$ exists, $\tilde{x}_0^{(k)}$ remains a valid CFE for all possible removal of k data points, i.e.,*

$$f_{\hat{\theta}_w}(\tilde{x}_0^{(k)}) \geq 0, \forall w \in \mathcal{W}^{(k)}.$$

Furthermore, the cost of implementing $\tilde{x}_0^{(k)}$ is upper bounded as following,

$$\begin{aligned} &\|\tilde{x}_0^{(k)} - x_0\|_2 \\ \leq &\|\tilde{x}_0 - x_0\|_2 + \min_{\substack{x\in\mathcal{X}, \\ f_{\hat{\theta}_{\mathbf{1}}}(x) - f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) \geq \frac{kC}{n}}} \|x - \tilde{x}_0\|_2, \end{aligned} \tag{13}$$

where C is a constant independent of n.

Specially, if $f_{\hat{\theta}_{\mathbf{1}}}$ is μ-strongly convex, then

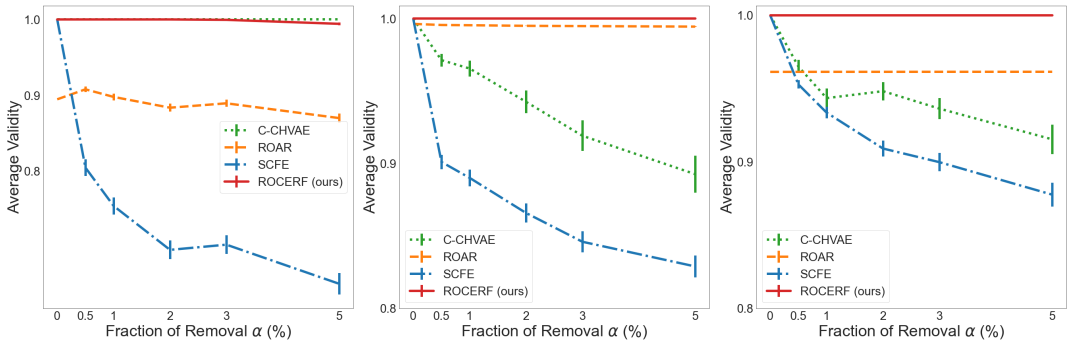
$$\|\tilde{x}_0^{(k)} - x_0\|_2 \leq \|\tilde{x}_0 - x_0\|_2 + \frac{2kC}{n\mu}. \tag{14}$$

The proof of Theorem 4.4 follows similar steps as Theorem 4.1. Assumption 4.3 is specifically baked to bound the difference $|\tilde{f}_{\hat{\theta}_w}(x) - f_{\hat{\theta}_w}(x)|$ and the difference $|\tilde{f}_{\hat{\theta}_w}(x) - f_{\hat{\theta}_{\mathbf{1}}}(x)|$. The detailed proof can be found in Appendix C.

Theorem 4.4 shares similar insights as the linear case while generalizing the results to a broader family of models beyond linear models. Admittedly, the assumptions are relatively strong for them to be held on very complex models such as neural networks. However, in applications where explainability is of major interest, simpler models are often preferred (Srinivas et al., 2022). So this result still provides valuable insights in practice.

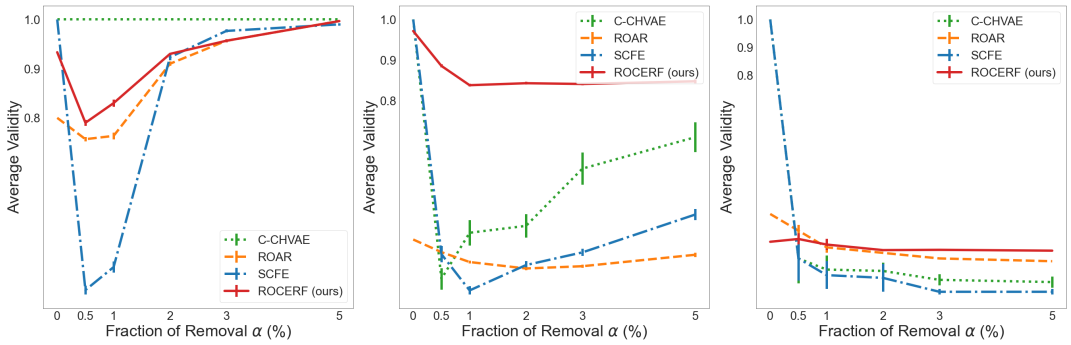
5. Experimental Evaluation

In this section, we empirically evaluate the validity and cost of the counterfactual explanations output by our framework, and compare them with other state-of-the-art counterfactual explanation methods. We first introduce the general experimental setup and then present experimental results on three real-world datasets with logistic regression and neural network models.



(a) German Credit. (b) COMPAS. (c) Adult.

Figure 1. Average validity of different counterfactual explanation methods applied to logistic regression models on three datasets. In each figure, the x-axis corresponds to the fraction of data removal α and the y-axis corresponds to the average validity. The error bars indicate the standard errors across $M = 100$ trials with each trial having an α fraction of training data points randomly removed.



(a) German Credit. (b) COMPAS. (c) Adult.

Figure 2. Average validity of different counterfactual explanation methods applied to neural network models on three datasets. See Figure 1 for more details about the plot setting.

5.1. Experimental Setup

We conduct experiments on three real-world datasets that are commonly used to benchmark counterfactual explanation methods. For each dataset, we split the dataset into training, validation, and test sets. We train a machine learning model (which we will refer as the *original model*) using the training set, and select the negative samples (data points classified as -1) in the test set. Then we apply different counterfactual explanation methods on these negative test samples to obtain a CFE for each of the sample. We calculate and report the average cost of the CFEs over all the negative test samples. To evaluate the validity under the right to be forgotten, we randomly remove a small fraction, α, of the training data points and retrain a new model (which we will refer as the *retrained model*), and then we evaluate the validity over all the negative test samples under the retrained model. We repeat this process of random removal, retrain, and validity evaluation for M times and report the average validity. We fix $M = 100$ on all experiments and vary $\alpha \in \{0.5\%, 1\%, 2\%, 3\%, 5\%\}$.

Datasets. We use three real-world binary classification datasets collected from high-stakes decision making scenarios. 1) *German Credit* (Dua & Graff, 2017) comprises of 1000 data points where each data point has 60 features including demographic (age, gender), personal (marital status), and financial (income, credit duration) information of a customer. These data points are labeled as "good" or "bad" in terms of credit risk. 2) *Adult* (Yeh & Lien, 2009) contains samples from 48,842 individuals, and each sample contains demographic (e.g., age, race, and gender), education (degree), employment (occupation, hours-per week), personal (marital status, relationship), and financial (capital gain/loss) features. 3) *COMPAS* (Jordan & Freiburger, 2015) comprises of criminal records and demographic features of 18,876 defendants who were released on bail at the US state courts during the period 1990-2009. The prediction target is "bail" or "no bail" given the defendant's data.

Methods	German Credit	COMPAS	Adult
SCFE	0.82 ± 0.12	0.78 ± 0.02	1.04 ± 0.006
C-CHVAE	8.51 ± 0.38	5.93 ± 0.11	3.79 ± 0.013
ROAR	1.45 ± 0.09	1.08 ± 0.01	1.07 ± 0.006
ROCERF (ours)	1.35 ± 0.14	0.87 ± 0.02	1.14 ± 0.006

Table 1. Average cost of different recourse methods applied to logistic regression models on three datasets. The cost is measured in terms of L2 norm.

Methods	German Credit	COMPAS	Adult
SCFE	1.18 ± 0.08	0.97 ± 0.11	1.00 ± 0.09
C-CHVAE	4.45 ± 0.18	5.98 ± 0.12	8.83 ± 0.31
ROAR	3.84 ± 0.33	1.09 ± 0.13	4.07 ± 0.55
ROCERF (ours)	2.76 ± 0.22	3.07 ± 0.08	4.06 ± 0.52

Table 2. Average cost of different recourse methods applied to neural network models on three datasets. The cost is measured in terms of L2 norm.

Predictive Models. We experiment with regularized logistic regression and deep neural networks. For regularized logistic regression, we use the default implementation from the Scikit-Learn package[3]. For neural networks, we use a 3-layer fully-connected feedforward neural network. Please see Appendix D.1 for more details about the implementation.

Evaluation Metrics. We evaluate the counterfactual explanation methods in terms of average validity and cost, which are the two most commonly used metrics in the counterfactual explanation literature (Verma et al., 2020a). Denote the set of negative samples under the original model $f_{\hat{\theta}_1}$ as $\mathcal{T}$ and the set of M random removals as $\mathcal{V} \subseteq \mathcal{W}^{(\lceil \alpha n \rceil)}$, $|\mathcal{V}| = M$. Suppose the CFE of a sample x is denoted as $c(x)$. Then the *average validity* is defined as

$$\frac{1}{M} \sum_{w \in \mathcal{V}} \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \mathbb{1}[f_{\hat{\theta}_w}(c(x)) = 1],$$

where $\mathbb{1}[\cdot]$ is the indicator function. And the *average cost* is defined as

$$\frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \|c(x) - x\|_2.$$

In Appendix D.2, we also report an alternative average cost with the L2 norm being replaced by L1 norm.

Baseline Methods. We compare the proposed method against three state-of-the-art counterfactual explanation methods, SCFE (Wachter et al., 2017), C-CHVAE (Pawelczyk et al., 2020), and ROAR (Upadhyay et al., 2021). SCFE uses gradient-based optimization to search for CFEs closest to the input sample, which can be viewed as the solution of the problem (1). C-CHVAE is a manifold-based method that searches for CFEs in a latent space. ROAR generates CFEs that are robust to small perturbations in model parameters, which is a strong baseline for the problem of interest in this paper.

Hyperparameters. For the proposed method, ROCERF, we set the hyperparameter k as 0.5% of the training set size and fix $\delta = 0$ in all experiments in this section. For SCFE, we use the hyper-parameter setting from Pawelczyk et al. (2022a). For C-CHVAE, we use the recommendations from Pawelczyk et al. (2020). We also use the same hyper-parameter setting for ROAR as suggested in Upadhyay et al. (2021). We refer the readers to Appendix D.1 for more details.

5.2. Experimental Results

The experimental results of average validity on logistic regression and neural network models are respectively shown in Figure 1 and Figure 2. The results of average cost on logistic regression and neural network models are respectively shown in Table 1 and Table 2.

We first look at the results on logistic regression models. As can be seen in Figure 1, the proposed method, ROCERF, achieves 100% average validity in almost all experimental settings for logistic regression. This result validates the strong

[3] `https://scikit-learn.org/stable/`.

theoretical guarantee on validity stated in Theorem 4.1. As a comparison, all the baseline methods suffer from significant drops in terms of average validity in some or all experimental settings.

In terms of the tradeoff between cost (Table 1) and validity (Figure 1), while SCFE always has the lowest cost, it has significantly worse validity than all other methods even for $\alpha = 0.5\%$; C-CHVAE is also inferior to the proposed method as it has both significantly higher costs on all datasets and worse validity on COMPAS and Adult; ROAR is closer to our method but our method consistently outperforms ROAR in terms of validity and has smaller or similar costs than ROAR. Overall, the empirical results both validate our theoretical analysis in Theorem 4.1 and verify that the proposed method outperforms baseline methods.

Next, we look at the results on neural network models. As the change of models after data removals becomes less predictable for these complex models, the performance of counterfactual explanation methods is more dataset dependent. However, we still see that the proposed method is consistently among the best performing methods.

On COMPAS dataset (Figure 2b), the proposed method clearly outperforms baseline methods in terms of validity. On Adult dataset (Figure 2c), SCFE and C-CHVAE are significantly worse in validity except for on the original model ($\alpha = 0\%$); the proposed method performs similarly as ROAR in terms of both validity and cost. On German Credit dataset (Figure 2a), the results of validity seem to be counter-intuitive: the average validity becomes 100% for all methods after removing a larger fraction of training data. This is possibly because the dataset is small and the decision boundary of the complex models changes dramatically after data removal. In addition, there are only 27 negative test samples under the original neural network model. The dramatical change in decision boundary may make all the test samples suddenly lie in a positive area. Nevertheless, on this dataset, C-CHVAE has the best validity but also with the highest cost. The proposed method has a similar validity as ROAR with a smaller cost.

6. Conclusions

In this work, we make one of the initial attempts at addressing the operational gaps between the right to explanation and the right to be forgotten. In particular, enforcing the right to be forgotten may invalidate actionable (counterfactual) explanations, which in turn violates the right to explanation. To resolve the tension between these two principles, we propose the first algorithmic framework, ROCERF, which generates counterfactual explanations that are provably robust to model updates triggered as a consequence of data deletion requests. The proposed framework not only enjoys theoretical guarantees on validity and cost, but also outperforms several other state-of-the-art counterfactual explanation methods. Our theoretical and empirical results establish that our framework ROCERF enables us to simultaneously enforce both the right to explanation as well as the right to be forgotten, thus bridging a critical operational gap between the two regulatory principles.

References

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, May 2021.

Broderick, T., Giordano, R., and Meager, R. An automatic Finite-Sample robustness metric: When can dropping a little data make a big difference? November 2020.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, May 2015.

CCPA. California consumer privacy act (ccpa), 2018. URL `https://oag.ca.gov/privacy/ccpa`.

Dandl, S., Molnar, C., Binder, M., and Bischl, B. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*, pp. 448–469. Springer, 2020.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Freund, R. M. Penalty and barrier methods for constrained optimization. *Lecture Notes, Massachusetts Institute of Technology*, 2004.

Garg, S., Goldwasser, S., and Vasudevan, P. N. Formalizing data deletion in the context of the right to be forgotten. In *Advances in Cryptology – EUROCRYPT 2020*, pp. 373–402. Springer International Publishing, 2020.

GDPR. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016.

Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. Making ai forget you: Data deletion in machine learning. *Adv. Neural Inf. Process. Syst.*, 2019.

Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. A swiss army infinitesimal jackknife. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1139–1147. PMLR, 2019.

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. In *International Conference on Machine Learning*, pp. 3832–3842, 2020.

Izzo, Z., Anne Smart, M., Chaudhuri, K., and Zou, J. Approximate data deletion from machine learning models. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2008–2016. PMLR, 2021.

Jordan, K. L. and Freiburger, T. L. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13(3):179–196, 2015. doi: 10.1080/15377938.2014.984045. URL https://doi.org/10.1080/15377938.2014.984045.

Karimi, A.-H., Barthe, G., Balle, B., and Valera, I. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017.

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

Mahajan, D., Tan, C., and Sharma, A. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

Mothilal, R. K., Sharma, A., and Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2020.

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-Delete: Gradient-Based methods for machine unlearning. In Feldman, V., Ligett, K., and Sabato, S. (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 931–962. PMLR, 2021.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. October 2017.

Pawelczyk, M., Broelemann, K., and Kasneci, G. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pp. 3126–3132, 2020.

Pawelczyk, M., Bielawski, S., Van den Heuvel, J., Richter, T., and Kasneci, G. Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. In *Advances in Neural Information Processing Systems (NeurIPS) (Benchmark and Datasets Track)*, volume 34, 2021.

Pawelczyk, M., Agarwal, C., Joshi, S., Upadhyay, S., and Lakkaraju, H. Exploring counterfactual explanations through the lens ofadversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022a.

Pawelczyk, M., Leemann, T., Biega, A., and Kasneci, G. On the Trade-Off between actionable explanations and the right to be forgotten. August 2022b.

Rawal, K. and Lakkaraju, H. Interpretable and interactive summaries ofactionable recourses. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Rawal, K., Kamar, E., and Lakkaraju, H. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv:2012.11788*, 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, pp. 1135–1144, 2016.

Srinivas, S., Matoba, K., Lakkaraju, H., and Fleuret, F. Efficient training of Low-Curvature neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Tolomei, G., Silvestri, F., Haines, A., and Lalmas, M. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. ACM, 2017.

Upadhyay, S., Joshi, S., and Lakkaraju, H. Towards robust and reliable algorithmic recourse. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16926–16937. Curran Associates, Inc., 2021.

Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

Van Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

Verma, S., Dickerson, J., and Hines, K. Counterfactual explanations for machine learning: A review. October 2020a.

Verma, S., Dickerson, J., and Hines, K. Counterfactual explanations for machine learning: A review. *arXiv:2010.10596*, 2020b.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electron. J.*, 2017.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2), 2018.

Yeh, I.-C. and Lien, C.-h. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. In *Expert Systems with Applications*, 2009.

A. The Optimization Algorithm

Algorithm 1 ROCERF.

Input: $x_0, f_{\mathcal{A}}^{(k)}, \delta, \mathcal{X}, T$.
Output: $\tilde{x}_0^{(k)}$.

Set $\lambda = 0.1$;
Set $x' = \arg\min_{x \in \mathcal{X}} \lambda\phi(\delta - f_{\mathcal{A}}^{(k)}(x)) + \|x - x_0\|_2$;
// Find initial left value λ
while $f_{\mathcal{A}}^{(k)}(x') \geq \delta$ **do**
  Set $\lambda = \lambda/2$;
  Set $x' = \arg\min_{x \in \mathcal{X}} \lambda\phi(\delta - f_{\mathcal{A}}^{(k)}(x)) + \|x - x_0\|_2$;
end while
Set $\lambda' = \lambda$;
// Find initial right value λ'
while $f_{\mathcal{A}}^{(k)}(x') < \delta$ **do**
  Set $\lambda' = \lambda' \times 2$;
  Set $x' = \arg\min_{x \in \mathcal{X}} \lambda\phi(\delta - f_{\mathcal{A}}^{(k)}(x)) + \|x - x_0\|_2$;
end while
// Binary search between λ and λ'
for $t = 1, 2, \ldots, T$ **do**
  Set $\lambda_t = (\lambda + \lambda')/2$;
  Set $x_t = \arg\min_{x \in \mathcal{X}} \lambda_t\phi(\delta - f_{\mathcal{A}}^{(k)}(x)) + \|x - x_0\|_2$;
  if $f_{\mathcal{A}}^{(k)}(x_t) < \delta$ **then**
    Set $\lambda = \lambda_t$;
  else
    Set $\lambda' = \lambda_t$;
  end if
end for
Set $\tilde{x}_0^{(k)} = \arg\min_{x \in \mathcal{X}} \lambda\phi(\delta - f_{\mathcal{A}}^{(k)}(x)) + \|x - x_0\|_2$;
Return $\tilde{x}_0^{(k)}$;

B. Proof of Theorem 4.1

B.1. Lemmas

We start by introducing a few useful lemmas.

For any $w \in \mathcal{W}^{(k)}$, define the following LKO estimator of $\hat{\theta}_w$:

$$\tilde{\theta}_w := \hat{\theta}_{\mathbf{1}} + H^{-1}\left(\frac{1}{n}\sum_{i:w_i=0} g_i(\hat{\theta}_{\mathbf{1}})\right). \tag{15}$$

Note that for linear models, Eq. (11) can be rewritten as $\tilde{f}_{\hat{\theta}_w}(x) = \tilde{\theta}_w^T x$.

The difference between $\tilde{\theta}_w$ and $\hat{\theta}_w$ can be bounded by the following lemma.

Lemma B.1 (Corollary 1 in Giordano et al. (2019)). *Let $H(\theta) = \frac{1}{n}\sum_{i=1}^n h_i(\theta)$. Assume the following quantities are bounded by constants independent of n: (1) $\sup_{\theta\in\Theta} \|H(\theta)^{-1}\|_{op}$; (2) $\frac{1}{n}\sum_{i=1}^n \|g_i(\theta)\|_2^2$; (3) $\frac{1}{n}\sum_{i=1}^n \|h_i(\theta)\|_F^2$. Also assume that there exists a suitable $\Delta > 0$, such that the following quantity is bounded by a constant independent of n: $\sup_{\|\theta-\hat{\theta}_{\mathbf{1}}\|_2<\Delta} \frac{1}{n}\sum_{i=1}^n \|h_i(\theta) - h_i(\hat{\theta}_{\mathbf{1}})\|_F / \|\theta - \hat{\theta}_{\mathbf{1}}\|_2$. Then for any small integer k, there exists a constant C_1 independent of n, such that*

$$\sup_{w\in\mathcal{W}^{(k)}} \|\tilde{\theta}_w - \hat{\theta}_w\|_2 \leq \frac{kC_1}{n}. \tag{16}$$

We also have the following results for strongly convex models.

Lemma B.2 (Lemma 8 in Neel et al. (2021)). *Suppose $l : \Theta \to \mathbb{R}$ is μ-strongly convex and let $\theta^* = \arg\min_{\theta\in\Theta} l(\theta)$. We have that for any $\theta \in \Theta$, $l(\theta) \geq l(\theta^*) + \frac{\mu}{2}\|\theta - \theta^*\|_2^2$.*

Lemma B.3. *Assume $l_i, i = 1, \ldots, n$ are L-Lipschitz and μ-strongly convex. For a fixed positive integer k, there exists a constant C_2 independent of n, such that for any $w \in \mathcal{W}^k$,*

$$\|\tilde{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2 \leq \frac{kC_2}{n}. \tag{17}$$

Proof of Lemma B.3. We bound $\|\tilde{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2$ by the summation of $\|\tilde{\theta}_w - \hat{\theta}_w\|_2$ and $\|\hat{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2$. From Lemma B.1, we already have $\|\tilde{\theta}_w - \hat{\theta}_w\|_2 \leq \frac{kC_1}{n}$. We now bound $\|\hat{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2$ largely following the proof of Lemma 8 (Sensitivity) in Neel et al. (2021).

WLOG, assume the the first k data points are removed in w, i.e., $w_1 = w_2 = \ldots = w_k = 0$ while $w_{k+1} = \ldots = w_n = 1$. Then we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} l_i(\hat{\theta}_w) &= \frac{n-k}{n}\frac{1}{n-k}\sum_{i=k+1}^{n} l_i(\hat{\theta}_w) + \frac{1}{n}\sum_{i=1}^{k} l_i(\hat{\theta}_w) \\
&\leq \frac{n-k}{n}\frac{1}{n-k}\sum_{i=k+1}^{n} l_i(\hat{\theta}_{\mathbf{1}}) + \frac{1}{n}\sum_{i=1}^{k} l_i(\hat{\theta}_w) \qquad (18) \\
&= \frac{1}{n}\sum_{i=k+1}^{n} l_i(\hat{\theta}_{\mathbf{1}}) + \frac{1}{n}\sum_{i=1}^{k}\Big(l_i(\hat{\theta}_w) - l_i(\hat{\theta}_{\mathbf{1}})\Big) \\
&\leq \frac{1}{n}\sum_{i=k+1}^{n} l_i(\hat{\theta}_{\mathbf{1}}) + \frac{kL}{n}\|\hat{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2, \qquad (19)
\end{aligned}$$

where (18) is because $\hat{\theta}_w$ is the minimizer of $\frac{1}{n-k}\sum_{i=k+1}^{n} l_i(\theta)$, while in (19) we have utilized the fact that each l_i is L-Lipschitz.

On the other hand, by Lemma B.2, we have

$$\frac{1}{n}\sum_{i=1}^{n} l_i(\hat{\theta}_w) \geq \frac{1}{n}\sum_{i=k+1}^{n} l_i(\hat{\theta}_{\mathbf{1}}) + \frac{\mu}{2}\|\hat{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2^2.$$

Combining the two inequalities above, we have $\|\hat{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2 \leq \frac{k(2L/\mu)}{n}$.

Therefore, letting $C_2 = C_1 + \frac{2L}{\mu}$, we have

$$\|\tilde{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2 \leq \|\tilde{\theta}_w - \hat{\theta}_w\|_2 + \|\hat{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2 \leq \frac{kC_2}{n}.$$

□

B.2. Useful Facts of Regularized Logistic Regression

Define $\sigma(x;\theta) = \frac{1}{1+\exp(-\theta^T x)}$. For regularized logistic regression with the loss defined as $l_i(\theta) = \log(1+\exp(-y_i\theta^T x_i)) + \gamma\|\theta\|_2^2$, we have

$$g_i(\theta) = -\frac{1}{1+\exp(y_i\theta^T x_i)}(y_i x_i) + \gamma\theta, \tag{20}$$

$$h_i(\theta) = \sigma(x_i;\theta)(1-\sigma(x_i;\theta))x_i x_i^T + \gamma I. \tag{21}$$

We can verify that regularized logistic regression satisfies all the assumptions in Lemma B.1. First, we know that the eigen value of the Hessian is lower-bounded by γ, so the eigen value of the inverse Hessian is upper bounded by $1/\gamma$. Hence $\sup_{\theta\in\Theta} \|H(\theta)^{-1}\|_{\text{op}}$ is bounded. Next, under the assumption that both the feature vector and model parameters have bounded norm, it is easy to show that $\|g_i(\theta)\|_2$ and $\|h_i(\theta)\|_F$ are bounded from Eq. (20) and Eq. (21). Hence both $\frac{1}{n}\sum_{i=1}^n \|g_i(\theta)\|_2^2$ and $\frac{1}{n}\sum_{i=1}^n \|h_i(\theta)\|_F^2$ are bounded. Finally, $h_i(\theta)$ is Lipschitz continuous so the last assumption is also verified.

We can also verify that l_i are Lipschitz and strongly convex so the regularized logistic regression satisfies the assumptions in Lemma B.3.

B.3. Proof of Theorem 4.1

Next, we are ready to prove Theorem 4.1.

Proof of Theorem 4.1. The validity of $\tilde{x}_0^{(k)}$ holds if for any $w \in \mathcal{W}^{(k)}$ and $x \in \mathcal{X}$, $\tilde{f}_{\hat{\theta}_w}(x) \geq \delta$ implies $f_{\hat{\theta}_w}(x) \geq 0$. Now we investigate the choice of δ that guarantees the above condition holds while not being too large.

Under the linear model assumption, for any $w \in \mathcal{W}^{(k)}$, we have

$$\begin{aligned} f_{\hat{\theta}_w}(x) &= \hat{\theta}_w^T x \\ &= \tilde{\theta}_w^T x + (\hat{\theta}_w - \tilde{\theta}_w)^T x \\ &\geq \tilde{f}_{\hat{\theta}_w}(x) - \|\hat{\theta}_w - \tilde{\theta}_w\|_2 \|x\|_2 \\ &\geq \tilde{f}_{\hat{\theta}_w}(x) - \frac{kC_1}{n} \cdot B, \end{aligned}$$

where recall that $B = \sup_{x\in\mathcal{X}} \|x\|_2$.

If we set $\delta = \frac{kC_1B}{n}$, then $\tilde{f}_{\hat{\theta}_w}(x) \geq \delta$ implies $f_{\hat{\theta}_w}(x) \geq 0$, in which case the validity of $\tilde{x}_0^{(k)}$ is guaranteed.

For the cost, as $\tilde{x}_0^{(k)}$ is the minimizer of the problem (7), we have $\|\tilde{x}_0^{(k)} - x_0\|_2 \leq \|x - x_0\|_2$ for any x in the feasible set of the the problem (7). Furthermore, for any x, $\|x - x_0\|_2 \leq \|\tilde{x}_0 - x_0\| + \|x - \tilde{x}_0\|$. So we only need to focus on the bound of $\|x - \tilde{x}_0\|_2$ for some x in the feasible set.

To begin with, we make the following transformation of $\tilde{f}_{\hat{\theta}_w}(x)$.

$$\begin{aligned} \tilde{f}_{\hat{\theta}_w}(x) &= \tilde{\theta}_w^T x \\ &= \hat{\theta}_{\mathbf{1}}^T \tilde{x}_0 + \hat{\theta}_{\mathbf{1}}^T (x - \tilde{x}_0) + (\tilde{\theta}_w - \hat{\theta}_{\mathbf{1}})^T x \\ &\geq \hat{\theta}_{\mathbf{1}}^T (x - \tilde{x}_0) + (\tilde{\theta}_w - \hat{\theta}_{\mathbf{1}})^T x && \hat{\theta}_{\mathbf{1}}^T \tilde{x}_0 \geq 0 \text{ by Definition 3.1} \\ &\geq \hat{\theta}_{\mathbf{1}}^T (x - \tilde{x}_0) - \|\tilde{\theta}_w - \hat{\theta}_{\mathbf{1}}\|_2 \|x\|_2 \\ &\geq \hat{\theta}_{\mathbf{1}}^T (x - \tilde{x}_0) - \frac{kC_2B}{n}. && \text{Lemma B.3} \end{aligned}$$

For x to be in a feasible set, it suffices to have $\tilde{f}_{\hat{\theta}_w}(x) \geq \delta$ for all $w \in \mathcal{W}^{(k)}$. Set $\delta = \frac{kC_1B}{n}$ and let $x' = \tilde{x}_0 + \frac{kC}{n\|\hat{\theta}_{\mathbf{1}}\|_2^2}\hat{\theta}_{\mathbf{1}}$, where $C := (C_1 + C_2)B$. Then for any w, we have

$$\tilde{f}_{\hat{\theta}_w}(x') - \delta \geq \hat{\theta}_{\mathbf{1}}^T (x' - \tilde{x}_0) - \frac{kC_2B}{n} - \frac{kC_1B}{n} = 0.$$

So x' is in the feasible set. Therefore,

$$\|\tilde{x}_0^{(k)} - x_0\|_2 \leq \|x' - x_0\|_2 \leq \|\tilde{x}_0 - x_0\|_2 + \frac{kC}{n\|\hat{\theta}_{\mathbf{1}}\|_2}.$$

□

C. Proof of Theorem 4.4

We first apply a result from Broderick et al. (2020) to bound the difference $|\tilde{f}_{\hat{\theta}_w}(x) - f_{\hat{\theta}_w}(x)|$.

Lemma C.1 (Direct Application of Theorem 1 in Broderick et al. (2020))**.** *Under Assumption 4.3, there exists a constant C_f independent of n, such that*

$$\sup_{x\in\mathcal{X}, w\in\mathcal{W}^{(k)}} |\tilde{f}_{\hat{\theta}_w}(x) - f_{\hat{\theta}_w}(x)| < \frac{kC_f}{n}.$$

Next, we use this result to prove Theorem 4.4.

Proof of Theorem 4.4. Similarly as Theorem 4.1, the validity of $\tilde{x}_0^{(k)}$ holds if for any $w \in \mathcal{W}^{(k)}$ and $x \in \mathcal{X}$, $\tilde{f}_{\hat{\theta}_w}(x) \geq \delta$ implies $f_{\hat{\theta}_w}(x) \geq 0$. By Lemma C.1, we know that setting $\delta = \frac{kC_f}{n}$ suffices to guarantee the validity.

For the cost, similarly as Theorem 4.1, we only need to bound $\|x - \tilde{x}_0\|_2$ for some feasible x in problem 7.

For any $x \in \mathcal{X}$ and $w \in \mathcal{W}^{(k)}$, we have

$$\begin{aligned}
\tilde{f}_{\hat{\theta}_w}(x) &= f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) + f_{\hat{\theta}_{\mathbf{1}}}(x) - f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) + \tilde{f}_{\hat{\theta}_w}(x) - f_{\hat{\theta}_{\mathbf{1}}}(x) \\
&\geq f_{\hat{\theta}_{\mathbf{1}}}(x) - f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) + \tilde{f}_{\hat{\theta}_w}(x) - f_{\theta_{\mathbf{1}}}(x) && f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) \geq 0 \text{ by Definition 3.1} \\
&= f_{\hat{\theta}_{\mathbf{1}}}(x) - f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) + \frac{1}{n}\sum_{i:w_i=0} \beta(x)^T H^{-1} g_i(\hat{\theta}_{\mathbf{1}}) && \text{Eq. (6)} \\
&\geq f_{\hat{\theta}_{\mathbf{1}}}(x) - f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) - \frac{kC_2C_3C_4}{n}. && \text{Assumption 4.3}
\end{aligned}$$

For an x to be feasible, it needs to satisfy $\tilde{f}_{\hat{\theta}_w}(x) \geq \delta$ for all $w \in \mathcal{W}^{(k)}$. Set $\delta = \frac{kC_f}{n}$ and let

$$x' = \arg \min_{\substack{x\in\mathcal{X}, \\ f_{\hat{\theta}_{\mathbf{1}}}(x) - f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) \geq \frac{kC}{n}}} \|x - \tilde{x}_0\|_2,$$

where $C := C_f + C_2C_3C_4$. Then for any w, we have

$$\tilde{f}_{\hat{\theta}_w}(x') - \delta \geq f_{\hat{\theta}_{\mathbf{1}}}(x') - f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) - \frac{kC}{n} \geq 0.$$

So x' is in the feasible set. Therefore,

$$\|\tilde{x}_0^{(k)} - x_0\|_2 \leq \|x' - x_0\|_2 \leq \|\tilde{x}_0 - x_0\|_2 + \min_{\substack{x\in\mathcal{X}, \\ f_{\hat{\theta}_{\mathbf{1}}}(x) - f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) \geq \frac{kC}{n}}} \|x - \tilde{x}_0\|_2.$$

Finally, if $f_{\hat{\theta}_{\mathbf{1}}}$ is μ-strongly convex, then for any $z \in \mathcal{X}$,

$$f_{\hat{\theta}_{\mathbf{1}}}(z) \geq f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) + \frac{\partial f_{\hat{\theta}_{\mathbf{1}}}(x)}{\partial x}\Big|_{\tilde{x}_0}(z - \tilde{x}_0) + \frac{\mu}{2}\|z - \tilde{x}_0\|_2.$$

Denote $v = \left(\frac{\partial f_{\hat{\theta}_{\mathbf{1}}}(x)}{\partial x}\Big|_{\tilde{x}_0}\right)^T$. Let $z = \tilde{x}_0 + \frac{2kC}{n\mu}\frac{v}{\|v\|_2}$, then

$$f_{\hat{\theta}_{\mathbf{1}}}(z) - f_{\hat{\theta}_{\mathbf{1}}}(\tilde{x}_0) \geq \frac{kC}{n}.$$

So z is in the feasible set. Therefore,

$$\|\tilde{x}_0^{(k)} - x_0\|_2 \leq \|x' - x_0\|_2 \leq \|\tilde{x}_0 - x_0\|_2 + \frac{2kC}{n\mu}.$$

□

D. Experiment Details

D.1. More Detailed Experimental Setup

Model Training. We train two models for our experiments : (1) Logistic Regression (LR), and (2) Neural Network (NN). For NN, we have three intermediate layers with twice the number of input nodes for each intermediate layer. We apply centered-softplus activation (Srinivas et al., 2022) for each intermediate layer output. The training procedure involved minimizing the standard cross entropy loss using stochastic gradient descent with 0.01 as the learning rate. The accuracy achieved after training for all the datasets is shown in Table 3.

Table 3. The accuracy of LR and ANN models trained on the datasets.

Dataset	LR	NN
German Credit	72.2%	73.9%
COMPAS	85.8%	85.1%
Adult	84.0%	84.7%

Recourse Method Hyperparameters. We use default hyper-parameter setting for most baseline methods aligned with authors' guidelines. Specifically, we use step size = 0.05 with a sample size of 1000 per iteration for C-CHVAE, $\delta_{max} = 0.1$ for ROAR.

LIME Approximation of Neural Network Models. For neural network, we learn a local linear approximation of the model using the perturbation-based framework in LIME (Ribeiro et al., 2016). Specifically, we train a logistic regression model on 10,000 perturbations sampled from $\mathcal{N}(0,\ 0.1)$ around the input sample.

D.2. Additional Results

Average cost in terms of L1 norm. We provide the L1-norm based average cost of different recourse methods in Table 4 and Table 5, respectively for logistic regression and neural network models. The relative trend is almost the same as the results of L2-norm based average cost reported in the main paper.

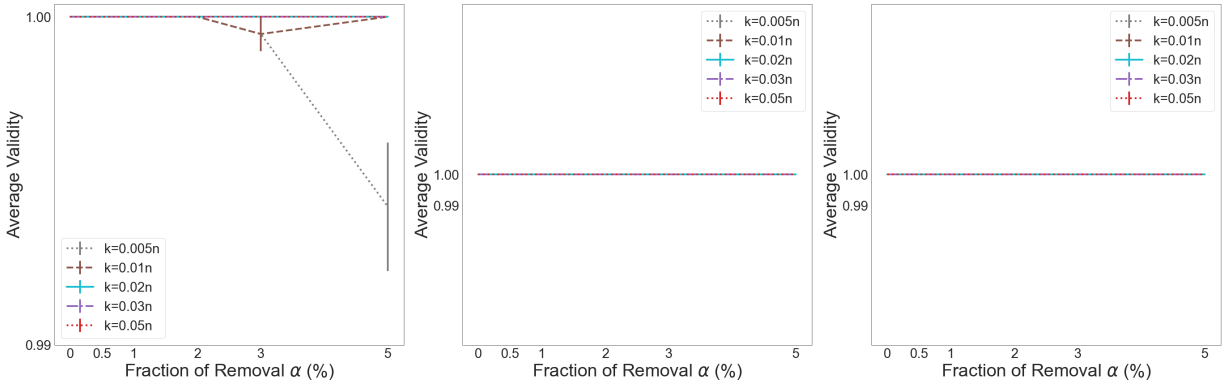
Methods	German Credit	COMPAS	Adult
SCFE	5.76 ± 0.92	1.91 ± 0.06	2.98 ± 0.01
C-CHVAE	48.04 ± 1.83	11.71 ± 0.23	10.37 ± 0.03
ROAR	9.61 ± 0.61	2.47 ± 0.04	2.61 ± 0.01
ROCERF (ours)	9.44 ± 1.08	2.13 ± 0.05	3.33 ± 0.02

Table 4. Average cost of different recourse methods applied to logistic regression models on three datasets. The cost is measured in terms of L1 norm.

Methods	German Credit	COMPAS	Adult
SCFE	2.97 ± 1.14	2.01 ± 0.23	7.03 ± 0.87
C-CHVAE	12.56 ± 0.53	11.31 ± 0.19	49.45 ± 1.88
ROAR	13.46 ± 0.25	2.25 ± 0.25	23.04 ± 3.68
ROCERF (ours)	7.85 ± 0.58	6.35 ± 0.14	19.61 ± 3.07

Table 5. Average cost of different recourse methods applied to neural network models on three datasets. The cost is measured in terms of L2 norm.

Sensitivity analysis with respect to the hyperparameter k. We also conduct a sensitivity analysis on different variants of our method with respect to the hyperparameter k. Figure 3 shows the results on logistic regression models. All variants of k achieves 100% validity COMPAS and Adult. On German Credit, the variant with the lowest k has a slight drop in validity for higher fraction of removal, which is fixed by for variants of higher k values. Notably, $k = 0.005n$ corresponds to $\alpha = 0.5\%$ and similarly for other values of k and α. The value of k refers to the hyperparameter of our method while the value of α refers to the actual fraction of data removal in the evaluation. In reality, our selection of hyperparameter k



(a) German Credit. (b) COMPAS. (c) Adult.

Figure 3. Sensitivity analysis with respect to the hyperparameter k.

may not exactly match the actual fraction of removals in the future. However, we note that, in Figure 3, the variants of our method always achieve 100% validity for αn below the hyperparamter k, e.g., $k = 0.01n$ achieves 100% validity for any $\alpha \leq 1\%$ and $k = 0.02n$ achieves 100% validity for any $\alpha \leq 2\%$.