On the Privacy Risks of Algorithmic Recourse

Martin Pawelczyk
University of Tübingen

Himabindu Lakkaraju*
Harvard University

Seth Neel*
Harvard University

Abstract

As predictive models are increasingly being employed to make consequential decisions, there is a growing emphasis on developing techniques that can provide algorithmic recourse to affected individuals. While such recourses can be immensely beneficial to affected individuals, potential adversaries could also exploit these recourses to compromise privacy. In this work, we make the first attempt at investigating if and how an adversary can leverage recourses to infer private information about the underlying model's training data. To this end, we propose a series of novel membership inference attacks which leverage algorithmic recourse. More specifically, we extend the prior literature on membership inference attacks to the recourse setting by leveraging the distances between data instances and their corresponding counterfactuals output by state-of-the-art recourse methods. Extensive experimentation with real world and synthetic datasets demonstrates significant privacy leakage through recourses. Our work establishes unintended privacy leakage as an important risk in the widespread adoption of recourse methods.

1 INTRODUCTION

Machine learning (ML) models are increasingly being deployed in domains such as finance, healthcare, and policy to make a variety of consequential decisions. As a result, there is a growing emphasis on providing *recourse* to individuals who have been adversely impacted by the predictions of these models [43]. For example, an individual who was denied a loan by a predictive model employed by a bank should be informed about what can be done to reverse this decision. Several approaches in the recent literature tackled the problem of providing recourse by generating counterfactual explanations [12, 26, 40, 41, 45] which highlight what features need to be changed and by how much to flip a model's prediction. For instance, Wachter et al. [45] proposed a gradient based approach to find the nearest counterfactual resulting in the desired prediction. More recently, Karimi et al. [13, 14] shed light on the spuriousness of the recourses generated by counterfactual/contrastive explanation techniques [40, 45], and advocated for leveraging the causal structure of the underlying data when generating recourses [3, 17, 21].

As algorithmic recourses seep into real-world applications, adversaries could potentially exploit these recourses to extract information about the underlying models and their training data, thereby leaking sensitive information (e.g., a bank's customer data) and enabling fraudulent activities. Therefore, there is a clear and urgent need to investigate the privacy risks associated with algorithmic recourse. While there is extensive literature on privacy attacks and defenses for machine learning models [1, 19, 30, 32], there is very little research [2, 33] that focuses on the privacy risks that arise when adversaries have access to *explanations* which highlight the rationale behind one or more model predictions. Recently, Shokri et al. [33] studied if and how feature attribution based explanations (which capture the feature importances associated with individual model predictions) leak sensitive information, and Aïvodji et al. [2] developed model extraction (i.e., reconstructing the underlying model) attacks against counterfactual explanations. However, neither of these works explore if and how adversaries may leverage recourses to infer sensitive information about the underlying model's training data.

In this work, we address the aforementioned gaps by initiating a study of *if and how an adversary can leverage algorithmic recourses to leak sensitive information about the training data of the underlying model*. To this end, we introduce a general class of membership inference attacks called *counterfactual distance-based attacks* which leverage algorithmic recourse to determine if an instance belongs to the training data of the underlying model or not. In formulating this new class of attacks, we exploit the intuition that the distance between an instance and its corresponding recourse may capture information about whether that instance was

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. *Equal contribution.

used to train the model. We instantiate the aforementioned class of attacks to propose two novel membership inference attacks. Our first attack infers membership by thresholding on the distance between a given instance and its corresponding algorithmic recourse. Our second attack draws inspiration from state-of-the-art loss-based membership inference attacks [5, 46] and proposes a likelihood ratio test (LRT) that accounts for algorithmic recourse. Our attacks operate under the assumption that the adversary can only query the recourse algorithm once. This assumption is a lot more practical than those considered in related works [2], and is inline with real-world settings where an end user would typically be provided with a single recourse and will not be able to query the underlying model or recourse algorithm multiple times [40, 45]. *To the best of our knowledge, our work is the first to introduce membership inference attacks which leverage algorithmic recourse.*

We experiment with multiple real world datasets spanning diverse domains such as lending, healthcare, and law to evaluate the effectiveness of the proposed attacks. Our experimental results clearly demonstrate the efficacy of the proposed attacks, and highlight significant privacy leakage through recourses generated by a wide range of recourse algorithms. In addition, the proposed attacks also outperform the state-of-the-art loss-based membership inference attacks (which do not leverage recourses) on data with sufficiently high dimensionality, thus highlighting the promise of our recourse-based attacks as generic membership inference attacks. We also empirically analyze the factors contributing to the success of our attacks, and find that our attacks are highly successful when the underlying model overfits to the training data [5] and the dimensionality of the data is high. Overall, our results establish unintended privacy leakage as an important risk in the widespread adoption of recourse algorithms.

2 RELATED WORK

Algorithmic Recourse. Several approaches have been proposed in literature to provide recourse to individuals who have been negatively impacted by model predictions [12, 15, 17, 18, 21, 27, 38, 40, 45]. These approaches can be broadly categorized based on [42]: *type of the underlying predictive model* (e.g., tree vs. differentiable classifier), *type of access* they require to the underlying predictive model (e.g., black box vs. gradient access), 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*, whether the underlying *causal relationships* should be accounted for when generating counterfactuals, and whether the output produced by the method should be *multiple diverse counterfactuals* or a single counterfactual. Recent research also highlighted and addressed various challenges pertaining to the robustness [4, 8, 23–25, 28, 34, 39] and fairness [11, 44] of algorithmic recourse. However, none of the aforementioned works explore the privacy risks associated with algorithmic recourse which is the focus of our work.

Privacy Attacks for ML Models. There is a long line of prior work developing privacy attacks on machine learning models [5, 30–32, 46]. One class of attacks called *membership inference attacks* focus on determining if a given instance is present in the training data of a particular model [5, 31, 32, 46]. These attacks typically exploit the differences in the distribution of model confidence on the true label (or the loss) between the instances that are in the training set and those that are not. For example, Shokri et al. [32] proposed a loss-based membership inference attack which determines if an instance is in the training set by testing if the loss of the model for that instance is less than a specific threshold. Other membership inference attacks are also predominantly loss-based attacks where the calibration of the threshold varies from one proposed attack to the other [5, 31, 46].

Some works leverage additional information beyond loss functions to do membership inference attacks. For instance, [7] leverages adversarial examples to orchestrate membership inference attacks. While there exist similarities between adversarial examples and recourses output by `SCFE` [45], the recourses output by other SOTA methods such as `CCHVAE` [21] and `GS` [15] are quite different from adversarial examples and their framework does not apply to algorithmic recourse broadly.

Intersections between Privacy and Explainability. The intersection between privacy and explainability, which is the focus of our work, is relatively under explored. Recently, Shokri et al. [33] developed a membership inference attack against a variety of feature attribution based explanation methods (e.g., LIME [29], SHAP [16], Gradient-based methods [35, 37]). Their attack exploits the intuition that the higher the variance of the feature attribution corresponding to an instance, the more uncertain the corresponding prediction (i.e., higher model loss), and therefore the less likely it is that the instance belongs to the training set. Furthermore, Aïvodji et al. [2] developed a model extraction attack against counterfactual explanation methods. Their attack leverages model predictions and counterfactual explanations (output in case of unfavorable predictions) corresponding to a set of instances, and learns a proxy model that mimics the behavior of the model under attack as closely as possible. None of these works develop membership inference attacks against counterfactual explanation (algorithmic recourse) methods which is the main theme of our work.

3 PRELIMINARIES

Let us consider a predictive model $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$ is the feature space, $\mathcal{Y}$ is the space of outcomes, and $\theta \in \Theta$ denotes the parameters of the model $f_{\boldsymbol{\theta}}$. Let $\mathcal{Y} =$

$\{0, 1\}$ where 0 and 1 denote an unfavorable outcome (e.g., loan denied) and a favorable outcome (e.g., loan approved) respectively. In practice $f_{\boldsymbol{\theta}}(x)$ will output a probability in $[0, 1]$ of a positive classification, which is then thresholded to obtain a binary classification, and we will denote by $(f_{\boldsymbol{\theta}}(x))_y$ the probability $f_{\boldsymbol{\theta}}$ assigns to a binary label y. Let us assume that the model $f_{\boldsymbol{\theta}}$ was trained using some data set $D_t = (X_t, Y_t)$ where each $(x, y) \in D_t$ is sampled from an underlying data distribution $\mathcal{D}$. A training algorithm $\mathcal{T} : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$ is a potentially randomized algorithm that takes in a dataset D_t and outputs a model $f_{\boldsymbol{\theta}}$. With this notation in place, we provide an overview of the standard formulations for algorithmic recourse as well as membership inference attacks.

Algorithmic Recourse. Let $x \in \mathcal{X}$ be an instance which received a negative outcome i.e., $f_{\boldsymbol{\theta}}(x) = 0$. The goal here is to find a recourse for this instance x i.e., to determine a set of changes δ that can be made to x in order to reverse the negative outcome. The problem of finding a recourse for x involves finding a counterfactual $x' = x + \delta$ for which the predictive model outputs a positive outcome i.e., $f_{\boldsymbol{\theta}}(x') = f_{\boldsymbol{\theta}}(x + \delta) = 1$. Note that it is desirable to minimize the cost $c(x, x')$ required to change x to x' so that the recourse is easily implementable. In practice, ℓ_1 or ℓ_2 distance are commonly used as cost functions [45]. Furthermore, since recommendations to change features such as gender or race would be unactionable, it is important to restrict the search for counterfactuals so that only actionable changes are allowed. Let $\mathcal{A}^p$ denote the set of plausible or actionable counterfactuals.

Putting it all together, the problem of finding a recourse for instance x for which $f_{\boldsymbol{\theta}}(x) = 0$ can be formalized as [45]:

$$x' = \underset{x' \in \mathcal{A}^p}{\arg\min}\, \ell(f_{\boldsymbol{\theta}}(x'), 1) + \lambda \cdot c(x, x'), \tag{1}$$

where $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ denotes a differentiable loss function (e.g., binary cross entropy loss) which ensures that gap between $f_{\boldsymbol{\theta}}(x')$ and the favorable outcome 1 is minimized, and $\lambda > 0$ is a trade-off parameter. Eqn. (1) captures the generic formulation leveraged by several of the state-of-the-art recourse finding algorithms [45]. In general, we denote (potentially randomized) recourse algorithms as $\mathcal{R} : (\Theta, \mathcal{X}) \to \mathcal{S}$. In the standard recourse setting $\mathcal{R}$ returns a recourse x' and so $\mathcal{S} = \mathcal{X}$.

Membership Inference Attacks for ML Models. The goal of a membership attack is to create a function that accurately determines if an instance $z = (x, y)$ belongs to the training set of the model $f_{\boldsymbol{\theta}}$. Several membership inference attacks proposed in literature exploit the intuition that models have lower loss on instances that were observed during their training. Such approaches are commonly referred to as *loss-based attacks*. Below, we discuss two of the most popular loss-based attacks developed by Yeom et al. [47] and Carlini et al. [5].

Info	`Loss`	`CFD`	`Loss LRT`	`CFD LRT`
Query access to $f_{\boldsymbol{\theta}}$	✓	×	✓	×
Query access to $\mathcal{R}$	×	✓	×	✓
Known loss function	✓	×	✓	×
Access to $\mathcal{D}^N$	×	×	✓	✓
Access to true labels	✓	×	✓	×

Table 1: Summarizing the assumptions underlying the different MI attacks. The recourse based attacks do not require access to the true labels nor do they need to know the correct loss functions.

Thresholding on Model Loss [47]. This attack takes an instance x and determines whether it is a member of the training set (`MEMBER`) by checking if the `Loss` of the model $f_{\boldsymbol{\theta}}$ is lower than or equal to a threshold τ_L:

$$M_{\text{Loss}}(x) = \begin{cases} \textit{MEMBER} & \text{if } \ell(\boldsymbol{\theta}, z) \leq \tau_L \\ \textit{NON-MEMBER} & \text{if } \ell(\boldsymbol{\theta}, z) > \tau_L. \end{cases} \tag{2}$$

Sablayrolles et al. [31] demonstrated that this attack is nearly optimal in the sense that it is approximately equivalent to the likelihood ratio test under certain conditions, which is the uniformly most powerful test for a given significance level (by the Neyman-Pearson Lemma). Note that this attack is only feasible when the adversary has access to the true label y of x, and the model's loss function ℓ and its parameters θ.

Likelihood Ratio Attack [5]. Recent work has attempted to further approximate a test based on the full likelihood ratio of the model θ by computing the likelihood ratio of the model loss, or equivalently confidence, $\text{conf}(f_\theta, z)$. Given sample access to $\mathcal{D}$ the adversary trains shadow models, and computes $\text{conf}(f_\theta, z)$ in the case when z is included in the training set and when it is a test point. Under the assumption that the logit scaled confidence is normally distributed, normal distributions are fit to the "in" and "out" scaled confidences, and an approximate likelihood ratio $\Lambda = \frac{\Pr[\text{conf}(f_{\boldsymbol{\theta}}, z) \mid \mathcal{N}(\mu_{in}, \sigma^2_{in})]}{\Pr[\text{conf}(f_{\boldsymbol{\theta}}, z) \mid \mathcal{N}(\mu_{out}, \sigma^2_{out})]}$ is computed. Finally, the adversary predicts `MEMBER` when $\Lambda > \tau$. We call this attack `Loss LRT`. In Subsection 4.2 we develop an LRT attack based not on $\text{conf}(f_\theta, z)$, but on the *counterfactual distance*, which does not require direct access to f_θ or y and can be implemented with algorithmic recourses.

4 OUR FRAMEWORK

In this section, we introduce a general class of novel membership inference attacks that leverage algorithmic recourse. First, we introduce the recourse-based membership inference game which generalizes the previously proposed membership inference attacks [47] on ML models to account for information captured in algorithmic recourse (see Table 1 for an overview). Then we introduce an attack that uses the distance between the recourse and the input (Subsection 4.1), and show how the attack can be improved with a

likelihood-ratio test (LRT) based approach in the style of [5] which we present in Subsection 4.2.

Definition 1 (Recourse-based MI Game)**.** *The game has two players: a model owner ($\mathcal{O}$) and an adversary ($\mathcal{A}$). The players take the following actions:*

(1) $\mathcal{O}$ draws a training set from the population $D_t \sim \mathcal{D}^N$, and using training algorithm $\mathcal{T}$, equipped with loss function ℓ trains a model $f_{\boldsymbol{\theta}} \sim \mathcal{T}(D_t)$. $\mathcal{O}$ then labels every point $z \in D_t$ with a binary label $f_{\boldsymbol{\theta}}(z)$. Let D_t^0 denote the subset of the training data with $f_{\boldsymbol{\theta}}(x) = 0$, and let $\mathcal{D}^{\theta,0}$ denote the conditional distribution $p(z) \sim D|f_{\boldsymbol{\theta}}(z) = 0$. $\mathcal{O}$ flips a coin: if "heads", $\mathcal{O}$ samples $x \sim \mathcal{D}^{0,\theta}$, else $x \sim D_t^0$. Using recourse generation algorithm $\mathcal{R}$, $\mathcal{O}$ generates a recourse $x' \sim \mathcal{R}(f_{\boldsymbol{\theta}}, x, D_t)$ for x. Then $\mathcal{O}$ sends $s = (x', x)$ to $\mathcal{A}$.

(2) In addition to s, $\mathcal{A}$ obtains query access to $\mathcal{D}$. We assume that $\mathcal{A}$ has full knowledge of all the implementation details of $\mathcal{O}$, including the specifics of $\mathcal{T}$ and $\mathcal{R}$. Finally $\mathcal{A}$ produces a binary guess G indicating whether $x \in D_t$ (MEMBER) or $x \notin D_t$ (NON-MEMBER).

We now present two attacks based on a statistic we call the *counterfactual distance*.

4.1 Thresholding on Counterfactual Distance

In the recourse-based MI Game, all the adversary has access to is the original instance x and its counterfactual x' generated by recourse algorithm. Loss-based attacks perform well at determining whether a point is a MEMBER of the training set or not, because the model typically over-fits to the training points, leading to lower losses on these points than on the test set. One explanation for this loss-disparity that has been given in prior work [33], is that during the training process, the decision boundary is forced away from training points. This suggests that points in the training set should be further from the boundary than points in the test set, motivating a distance-based attack that predicts a point is a MEMBER of the training set if its loss is below some threshold τ. The distance of a point x to the boundary can be computed as $c(x, x')$ where x' is the solution to Equation 1 with $\mathcal{A}^p = \mathbb{R}^d$. So if we exactly optimize the objective function in Equation 1 that underpins our recourse algorithms (with $\mathcal{A}^p = \mathbb{R}^d$), the counterfactual distance $c(x, x')$ is exactly the distance to the model boundary. While algorithms that focus on generating *realistic* recourses [14, 21] do not exactly optimize this objective in general, we can still view the distance to the recourse as a proxy for the distance of x to the model boundary. Based on this intuition we have the following counterfactual distance (CFD) based attack:

$$M_{\text{Distance}}(x) = \begin{cases} \textit{MEMBER} & \text{if } c(x, x') \geq \tau_D \\ \textit{NON-MEMBER} & \text{if } c(x, x') < \tau_D. \end{cases} \quad (3)$$

Following [5], we assume for the first two attacks below that $\mathcal{A}$ knows apriori an optimal threshold τ_α that maximizes a given TPR subject to a fixed FPR α, as the purpose of these simple attacks is to illustrate the potential privacy leakage through the recourse output. In practice, we will be plotting the TPR vs. FPR curves over all values of the threshold τ_α and so we will not need to pick a specific one.

4.2 LR Test using Counterfactual Distance

Algorithm 1 One-sided Distance-based Likelihood Ratio Test (CFD LRT)

1: **Inputs:** point (x, y), recourse output s = GetRecourse$(x, f_{\boldsymbol{\theta}}), \mathcal{D}$; FP-Rate: α, # Shadow Models: N, $\mathcal{T}$ = TrainClassifier$(\cdot)$
2: teststats = $[\,]$
3: **Compute**: $t_0 = T(s) = c(x, x')$
4: **for** $i = 1 : N$ **do**
5: Sample $\mathcal{D}_t^{(i)} \sim \mathcal{D}$
6: $f_{\boldsymbol{\theta}^{(i)}}$ = TrainClassifier$(\mathcal{D}^{(i)})$
7: $s^{(i)}$ = GetRecourse$(x, f_{\boldsymbol{\theta}^{(i)}})$
8: teststats $\leftarrow T(s^i) = c(x, x'^{(i)})$
9: **end for**
10: $\hat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\log c(x, \mathbf{x}'^{(i)})\right)$
11: $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{\text{MLE}} - \log\left(c(x, \mathbf{x}'^{(i)})\right)\right)^2$
12: **if** $t_0 > z_{1-\alpha}$ **then** ▷ $z_{1-\alpha}$ is the $1-\alpha$-quantile of $Z \sim \mathcal{LN}(\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}})$
13: **Output:** G = NON-MEMBER
14: **else**
15: **Output:** G = MEMBER
16: **end if**

Carlini et al. [5] showed that Loss LRT attacks perform better than simple Loss thresholding. In Algorithm 1 we present an LRT version of our counterfactual distance attack (CFD LRT). As in prior work [5, 31] since computing the LRT (Equation 4) exactly is intractable, we make several modifications to the attack that allow us to compute it efficiently. The full likelihood ratio given $c(x, x')$ is:

$$\Lambda = \frac{\Pr[c(x, x')|x \in D_t]}{\Pr[c(x, x')|x \notin D_t]}. \quad (4)$$

We model the distributions of our statistic $c(x, x')$ as log-Normal, and so in order to compute eqn. (4), we need to estimate $(\mu_{in}, \sigma_{in}), (\mu_{out}, \sigma_{out})$ where we assume that if $D_t \backslash \{x\} \sim \mathcal{D}, \theta \sim \mathcal{T}(\{x\} \cup D_t), x' \sim \mathcal{R}(x, \theta, D_t)$, then $\log c(x, x') \sim \mathcal{N}(\mu_{in}, \sigma_{in})$, and similarly when $x' \notin D_t$, $\log c(x, x') \sim \mathcal{N}(\mu_{out}, \sigma_{out})$. Given access to $\mathcal{D}$ we can estimate the parameters μ, σ by drawing fresh datasets, training models θ – with or without a given point x – and then computing the resulting counterfactual distances (Lines 5 – 8). However, as in [5] we note that to approximate the numerator, we have to perform this sampling and model training separately for each x that we perform the attack

on, which is computationally infeasible. Hence in Algorithm 1 we present a one-sided version of the LRT, where in Lines 10 – 11 we estimate μ_{out}, σ_{out}, and our attack predicts MEMBER if $c(x, x')$ has a sufficiently low likelihood under these parameters. Note that since μ_{out}, σ_{out} do not depend on x, we only need to perform the process of training Shadow models once, even if we are evaluating our attack on many different x's.

5 IS PRIVACY LEAKAGE THROUGH RECOURSES INEVITABLE?

The attacks developed above and empirical results in Section 6 suggest that recourses can be exploited to infer private information about the underlying training set. This raises a natural question: Is privacy leakage through recourses inevitable?

Over the last decade, differential privacy [10] has emerged as the canonical approach to provably preventing membership inference for a wide array of statistical tasks. Applying this to the recourse setting, results which have been folklore in the privacy community imply that if the recourse generation algorithm is DP in the training data, we can provably bound the success of *any* adversary $\mathcal{A}$ in the Recourse-based MI Game. In Theorem 2 (proof deferred to Supplement) we state a variant of the folklore result tailored to our setting, showing that not only can we bound the excess accuracy of the adversary over random guessing, we can also bound the balanced accuracy (BA). Since BA = $\frac{\text{TPR+TNR}}{2}$, this implies that for a small FPR α, the TPR of $\mathcal{A}$ is also close to α. Recent work advocates for evaluating the success of MI attacks at low FPR instead of just looking at the overall accuracy [5, 46].

Theorem 1. *Let $\mathcal{T} : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$ denote the training algorithm, draw $D_t \sim \mathcal{D}^n$ and and $\mathcal{A}$ be an arbitrary adversary that receives $z = (x, y), s \sim \mathcal{R}(f_\theta, x, D_t)$ from the recourse inference game, and produces a guess $G \in \{\textit{MEMBER}, \textit{NON-MEMBER}\}$. Then, if $\mathcal{R}$ is $(\epsilon, 0)$-differentially private, we have for all $\mathcal{A}$:*

$$BA_{\mathcal{A}} \leq \frac{1}{2} + \frac{1 - e^{-\epsilon}}{2}.$$

While Theorem 2 provides strong privacy guarantees, for several reasons both generic and specific to the recourse setting, *differential privacy is not a silver bullet to defend recourses against membership inference attacks*. It is known that training with DP causes a significant drop in accuracy on even relatively simple benchmarks [20], and so when accuracy is a concern this defense may not be feasible. Moreover, model accuracy aside, private training could alter the distance between training points and the model boundary, potentially leading to costlier and less actionable recourses for individuals.

6 EXPERIMENTAL EVALUATION

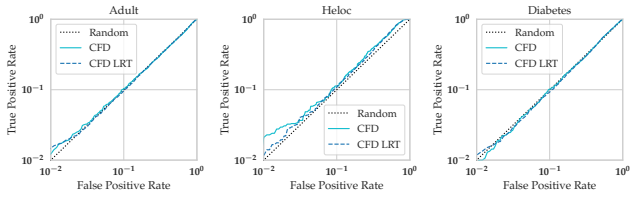
Here, we discuss the detailed experimental evaluation of our proposed attacks. Relative to related work [7], we are focusing on (i) a variety of recourse algorithms and (ii) evaluate our suggested LRT based attacks. We do so by first comparing the proposed recourse-based attacks to each other using log-scaled AUC curves, that emphasize the importance of the low false-positive rate regime. Recently, the latter metric has been advocated for [5, 46]. Second, we use average-case metrics such as AUC or balanced accuracy (BA) to understand which recourse algorithms are most vulnerable to membership inference attacks. We show these results in Section 6.2. Finally, in Section 6.3 we leverage synthetic data to analyze the determining factors of attack success. Below, we describe our experimental setup in more detail.

Data	Measures	CFD			CFD LRT		
		SCFE	GS	CCHVAE	SCFE	GS	CCHVAE
A	AUC	0.4971	0.5038	0.5008	0.4988	**0.5103**	0.5066
	BA	0.5115	0.5125	0.5056	**0.5132**	0.5098	0.5176
	TPR (0.1)	0.1039	0.1020	0.1058	0.1010	0.1043	**0.1298**
	TPR (0.01)	0.0121	0.0097	0.0157	**0.0158**	0.0095	0.0134
H	AUC	0.5887	0.5410	0.4874	0.5829	0.5027	**0.6789**
	BA	0.5904	0.5404	0.5473	0.5924	0.5326	**0.6389**
	TPR (0.1)	0.1130	0.1223	0.0863	0.1106	0.1142	**0.2635**
	TPR (0.01)	0.0155	0.0176	0.0016	0.0135	0.0372	**0.0513**
D	AUC	**0.5051**	0.5000	NA	0.5050	0.5047	NA
	BA	0.5100	0.5133	NA	**0.5145**	0.5136	NA
	TPR (0.1)	0.1020	0.0950	NA	0.0894	**0.1181**	NA
	TPR (0.01)	0.0093	0.0083	NA	0.0113	**0.0159**	NA

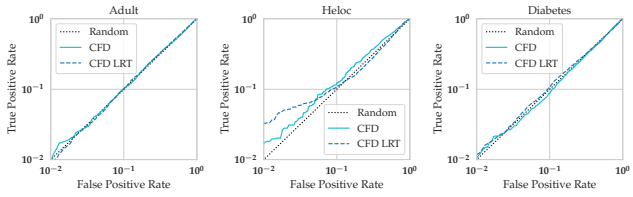
Table 2: Comparing the efficacy of the distance-based attacks for various recourse methods. The AUC denotes the area under the receiver operating characteristic curve, BA is the Balanced Accuracy, and TPR(x) measures the TPR when the FPR = x. For the Diabetes data set CCHVAE could not identify any recourses.

6.1 Setup

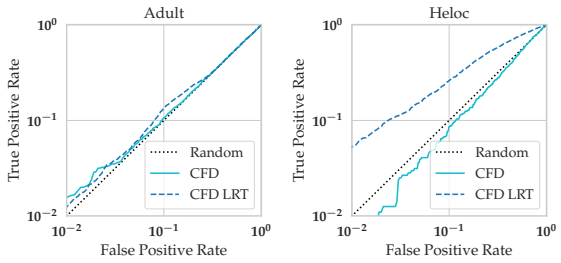
Datasets. Our first data set is the *Adult* (A) data set [9] that originates from the 1994 Census database, consisting of 14 attributes and 48,842 instances. The class label indicates whether an individual has an income greater than 50,000 USD/year. Second, we use the *Home Equity Line of Credit (Heloc)* (H) data set (d = 23). Here, the target variable records a score indicating whether individuals will repay the Heloc account within a fixed time window. Across both tasks we consider individuals in need of recourse if their scores lies below the median score across the entire data set, thresholding the scores based on the median to obtain binary target labels. Third, we use the *Diabetes* (D) data set which contains information on diabetic patients from 130 different US hospitals [36]. The patients are described using administrative (e.g., length of stay) and medical records (e.g., test results) (d = 42), and the prediction task is concerned with identifying whether a patient will be readmitted within the next 30 days. In line with [33], we sub sampled smaller data



(a) SCFE



(b) GS



(c) CCHVAE

Figure 1: Comparing the attack efficacy across different MI attacks for fully connected NN model trained on the real-world data sets when SCFE and GS are used for the attacks. Small upwards deviations from the diagonal at low FPRs (i.e. 0.01) indicate that a small fraction of points can accurately be identified as members of the training data set.

sets of 10000 points from each of these datasets. 5000 points are left to the model owner to train the their private model, while another 5000 data points are used by the adversary to train their shadow models when applicable.

Additionally, we follow Shokri et al. [33] and generate synthetic data sets from the sklearn library. For d features the method randomly chooses a vertex from the d-dimensional hypercube as a center for each of the classes, and samples Gaussian distributed random variables centered at the vertex with unit variance. For the linear model experiments, we use $d \in \{100, 1000, 5000, 7000\}$ with $n = 5000$ training and test samples. For the non-linear models we use $d \in \{50, 150\}$.

Recourse Algorithms and Predictive Models. We apply our techniques to three different methods which aim to generate low-cost recourses using different principles: SCFE is the method suggested by Wachter et al. [45] and uses a gradient-based objective to find recourses, GS conducts a random search for recourse in the input space [15], and CCHVAE [21] searches for recourse in a lower dimensional latent space using a generative model to encourage recourses to lie on the data manifold. All methods use a ℓ_1-regularizer to encourage sparse recourses. We use implementations from the CARLA library [22].

Baseline Attacks. We implement several baseline attacks: the baseline of random guessing that for a target FPR α predicts MEMBER with probability α, and baselines based on the loss ℓ. The loss-based baselines are simple thresholding on the loss [47] (i.e., Loss), and the offline loss-based LRT [5] (i.e., Loss LRT).

Evaluation Measures. We use several well established measures to validate the efficacy of our proposed membership inference attacks. Consistent with with previous works [32, 33, 47] we report balanced accuracy (BA) and receiver operating characteristic (ROC) area under the curve (AUC) scores. Additionally, we follow [5, 46] and also report log-scale ROC curves, and true positive rates of the attacks at low false positive rates. The authors argue that, for membership inference attacks, average case metrics such as BA and AUC are not well suited. The underlying idea is that if a membership inference attack can identify even a very small subset of the training data with very high confidence, then the attack should be considered successful. We follow this intuition and primarily report our findings using this metric.

6.2 Evaluating the Attack Efficacy

Inspecting Figure 1, we see that for nearly all datasets across all methods, at sufficiently low FPR the CFD LRT curve lies above the diagonal, outperforming the random baseline; the one exception being the Adult dataset with recourses generated from GS. CFD also often outperforms the random baseline in most cases, but not on the Diabetes dataset with recourse method SCFE or the Heloc dataset with method CCHVAE. These trends are reflected in the metrics captured in Table 2, where CFD LRT achieves TPR > .01 at FPR = .01, and AUC > .5 in 7 of the 8 dataset-recourse settings. CFD achieves TPR > .01 at FPR = .01 in 4 of the 8 settings. Focusing on the metric of TPR at FPR = .01, we see that in the 7 of 8 settings where either method outperforms the random baseline, CFD LRT achieves higher TPR than CFD 5 times. This difference is particularly evident on the Heloc dataset, where CFD LRT obtains TPR $35\%, 270\%$, and 413% above the random baseline for SCFE, GS, CCHVAE respectively. In summary, both methods often outperform the random baseline across all metrics, showing substantial privacy leakage from algorithmic recourses, with CFD LRT generally outperforming CFD. The results show that the Heloc dataset is particularly vulnerable to distance-based attacks across all recourse algorithms. On the Adult and Diabetes datasets the distance-based attacks usually outperform random guessing, however, the improvement over the random baseline is less pronounced.

Perhaps most interesting, is that only in the case of CCHVAE

do the CFD LRT attacks that outperform the random baseline and are plotted in Figure 1(c), actually reverse the direction of the threshold. This means that for the CFD LRT curves in Figure 1(c) and metrics in Table 2 in column CCHVAE, rather than predicting MEMBER in Line 12 of Algorithm 1 if $c(x, x') > z_{1-\alpha}$, we predict MEMBER iff $c(x, x') < z_\alpha$. While these results stand in contrast to our findings for SCFE and GS, there is an intuitive explanation. CCHVAE trains a VAE to model the data generating distribution, and performs sampling *in the latent space* to find a point in latent space z' that is close to the representation of x in the latent space encode$(x) = z_x$. Then it outputs the recourse $x' =$ decode(z') in the input space. Regardless of the specific z' found by the recourse algorithm, decode(z') is still in the range of the generative model. It is a known property of generative models like VAEs and GANs that their generated samples tend to be closer to training points than to test points, a fact which has been exploited for MI [6]. One explanation for the results in Figure 1(c) is that this property of generative models is in some cases outweighing the effect of the optimization during model training.

6.3 Towards Understanding Attack Success

To better understand the factors underlying these impressive results, in this section we use experiments on synthetic data with $\mathcal{R} =$ SCFE to examine the role of model type, model size, and number of features in attack success.

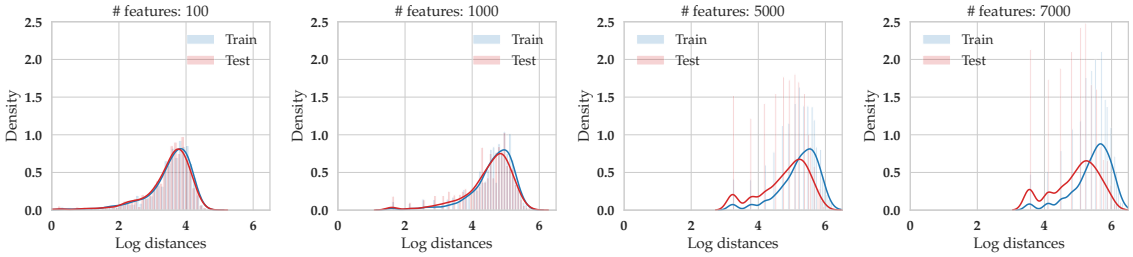
Linear Models. One potential confounding factor in assessing the results for SCFE, is that when $f_\theta(x)$ is non-convex so is the objective in Equation 1, and so there is no guarantee that the recourse output is close to globally optimal. In the case when f_θ is linear the objective is convex and has a closed form minimizer [7, 24]. In Figure 2 we train logistic regression models on the synthetic data, varying the number of features, and plotting the distribution of the counterfactual distances for train and test points, as well as the log-scaled ROC curves. We observe that the training distance distribution starts moving away from the test distance distribution as the number of feature dimensions increases (see Figure 2a), and as expected the distance-based attack starts performing better as the number of features increases (see Figure 2b). Strikingly, the CFD LRT in particular not only outperforms the CFD and random baseline, but also thresholding based on the Loss and the Loss LRT — attacks which have access to the full loss function and the label y. Figures 2a, 2b suggest two potential reasons why the results in Figure 1 for SCFE do not exhibit the same level of attack performance as on the synthetic data: (i) Since the model classes are non-convex it could be due to our optimization failing to find the recourse that minimizes Equation 1, and (ii) Distance-based membership inference attacks are more effective in high dimensions, and the tabular datasets we experiment on are of relatively small dimension $d < 50$.

Nonlinear models. So far we have studied the effect of the dimension on the attack success, which in the case of linear models is equal to the number of parameters. To disentangle the effect that the number of features has from the effect that model capacity has on attack success, we use fully-connected neural networks for which we can control both factors independently. We generate two synthetic data sets with $d = 50, 150$ respectively, and study to what extent an increase in model capacity, in this case the number of hidden nodes, impacts the performance of our attacks. For $d = 50$ the results in Figure 3a suggest that increasing model capacity does not yield performance increases for our distance-based attacks. On the other hand, when $d = 150$ the results in Figure 3b show that increasing model capacity drastically increases the attack success of the CFD LRT, while the simple CFD still does not work well. For the model with the largest capacity, three layers of 1000 hidden nodes per layer, CFD LRT performs considerably better than even the Loss LRT. Taken together these results suggest that a combination of high model capacity and high feature dimension increases the vulnerability of recourses to MI attacks, where feature dimension appears to play a more important role.

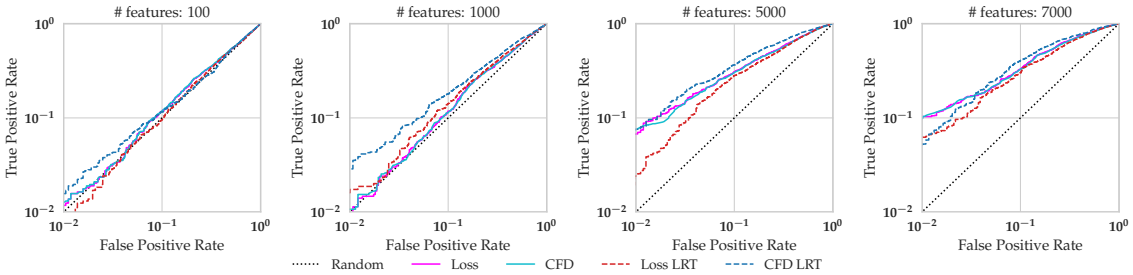
7 CONCLUSION

In this work, we investigated the privacy risks associated with algorithmic recourse. More specifically, we introduced a general class of membership inference attacks called *counterfactual distance-based attacks* which leverage algorithmic recourse to determine if an instance belongs to the training data of the underlying model. In formulating this new class of attacks, we exploit the intuition that the distance between an instance and its corresponding recourse may capture information about whether that instance was used to train the model. Empirical results on multiple synthetic and real-world datasets clearly demonstrate the efficacy of the proposed attacks, and highlight significant privacy leakage through recourses generated by a wide range of recourse methods. The proposed attacks also outperformed other state-of-the-art loss-based membership inference attacks on data with sufficiently high dimensionality. Overall, our results shed light on the critical risk of unintended privacy leakage through algorithmic recourse.

Our work paves the way for other important future research directions. For instance, exploring solutions to mitigate the privacy risks highlighted in our work, either through heuristic methods, or by developing novel ways to generate differentially private recourses which are provably robust to such attacks. Perhaps the most interesting future direction is to undertake a more thorough theoretical and empirical study of how the geometry of the decision boundary relates to privacy leakages.

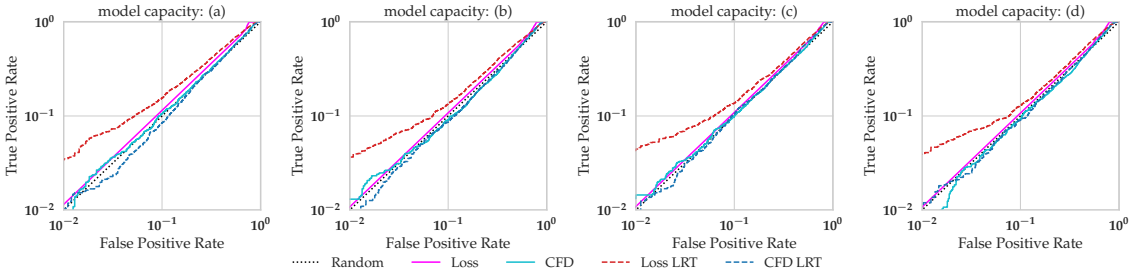


(a) Comparing log distances (ℓ_1) to the decision boundary across train and test points.

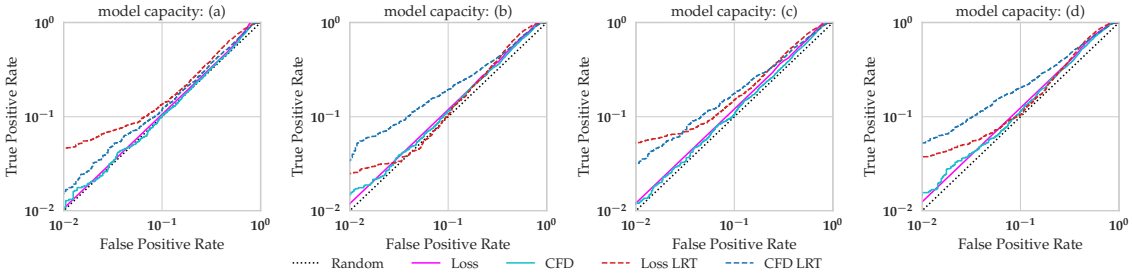


(b) Comparing the true positive rates against the false positive rates across different MI attacks using log-scaled ROC curves.

Figure 2: Demonstrating the efficacy of our proposed distance-based attack for logistic regression models trained on the synthetic data set when SCFE is used for the attack. At the interpolation threshold (i.e., when the number of training points equals the feature dimension: $d = n = 5000$) the baseline loss-based and distance-based attacks start outperforming the lrt-based attacks.



(a) # features = 50



(b) # features = 150

Figure 3: Demonstrating that both the network capacity and the number of features d matter for the efficacy of the distance-based attack. We trained neural network models on 10000 instances from the synthetic data set and used SCFE for the attack. From left to right the model capacity increases: (a): two-layer neural network with 1000 hidden nodes. (b): three-layer neural network with 100 hidden nodes in each hidden layer. (c): three-layer neural networks with 333 hidden nodes in each hidden layer. (d): three-layer neural networks with 1000 hidden nodes in each hidden layer.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful feedback. This work is supported in part by the NSF awards #IIS-2008461 and #IIS-2040989, and research awards from Google, JP Morgan, Amazon, Bayer, Harvard Data Science Initiative, and D^3 Institute at Harvard. HL would like to thank Sujatha and Mohan Lakkaraju for their continued support and encouragement. The views expressed here are those of the authors and do not reflect the official policy or position of the funding agencies.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884*, 2020.

[3] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, New York, NY, USA, 2020. ACM.

[4] Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. Consistent counterfactuals for deep models. *arXiv:2110.03109*, 2021.

[5] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. *CoRR*, abs/2112.03570, 2021.

[6] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna, editors, *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, pages 343–362. ACM, 2020. doi: 10.1145/3372297.3417238. URL https://doi.org/10.1145/3372297.3417238.

[7] Christopher A. Choquette-Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. *CoRR*, abs/2007.14321, 2020.

[8] Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, and Bernhard Schölkopf. On the adversarial robustness of causal algorithmic recourse. *arXiv:2112.11313*, 2021.

[9] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[10] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. doi: 10.1561/0400000042.

[11] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

[12] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[13] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[14] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

[15] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

[16] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[17] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

[18] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2020.

[19] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[20] Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Ulfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy, 2020.

[21] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020 (WWW)*. ACM, 2020.

[22] Martin Pawelczyk, Sascha Bielawski, Johan Van den Heuvel, Tobias Richter, and Gjergji Kasneci. 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.

[23] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens ofadversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[24] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Algorithmic recourse in the face of noisy human responses. In *International Conference on Learning Representations (ICLR)*, 2023.

[25] Martin Pawelczyk, Tobias Leemann, Asia Biega, and Gjergji Kasneci. On the trade-off between actionable explanations and the right to be forgotten. In *International Conference on Learning Representations (ICLR)*, 2023.

[26] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 344–350, 2020.

[27] Kaivalya Rawal and Himabindu Lakkaraju. Interpretable and interactive summaries ofactionable recourses. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[28] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv:2012.11788*, 2021.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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)*, pages 1135–1144, 2016.

[30] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646*, 2020.

[31] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Herve Jegou. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[32] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[33] Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, page 231–241, 2021.

[34] Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[35] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: Removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[36] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[37] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017.

[38] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 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.

[39] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[40] Berk Ustun, Alexander Spangher, and Y. Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

[41] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

[42] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv:2010.10596*, 2020.

[43] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10: 3152676, 2017.

[44] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. *arXiv preprint arXiv:2010.06529*, 2020.

[45] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black

box: automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2), 2018.

[46] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, and Reza Shokri. Enhanced membership inference attacks against machine learning models. *CoRR*, abs/2111.09679, 2021.

[47] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. *arXiv preprint arXiv:1709.01604*, 2017.

A IS PRIVACY LEAKAGE THROUGH RECOURSES INEVITABLE?

The attacks developed above and empirical results in Section 6 suggest that recourses can be exploited to infer private information about the underlying training set. This raises a natural question: Is privacy leakage through recourses inevitable?

Over the last decade, differential privacy [10] has emerged as the canonical approach to provably preventing membership inference for a wide array of statistical tasks. Applying this to the recourse setting, results which have been folklore in the privacy community imply that if the recourse generation algorithm is DP in the training data, we can provably bound the success of *any* adversary $\mathcal{A}$ in the Recourse-based MI Game. In Theorem 2 (proof deferred to Supplement) we state a variant of the folklore result tailored to our setting, showing that not only can we bound the excess accuracy of the adversary over random guessing, we can also bound the balanced accuracy (BA). Since BA = $\frac{\text{TPR+TNR}}{2}$, this implies that for a small FPR α, the TPR of $\mathcal{A}$ is also close to α. Recent work advocates for evaluating the success of MI attacks at low FPR instead of just looking at the overall accuracy [5, 46].

Theorem 2. *Let $\mathcal{T} : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$ denote the training algorithm, draw $D_t \sim \mathcal{D}^n$ and and $\mathcal{A}$ be an arbitrary adversary that receives $z = (x, y), s \sim \mathcal{R}(f_\theta, x, D_t)$ from the recourse inference game, and produces a guess $G \in \{\textit{MEMBER}, \textit{NON-MEMBER}\}$. Then, if $\mathcal{R}$ is $(\epsilon, 0)$-differentially private, we have for all $\mathcal{A}$:*

$$BA_{\mathcal{A}} \leq \frac{1}{2} + \frac{1 - e^{-\epsilon}}{2}.$$

While Theorem 2 provides strong privacy guarantees, for several reasons both generic and specific to the recourse setting, *differential privacy is not a silver bullet to defend recourses against membership inference attacks*. It is known that training with DP causes a significant drop in accuracy on even relatively simple benchmarks [20], and so when accuracy is a concern this defense may not be feasible. Moreover, model accuracy aside, private training could alter the distance between training points and the model boundary, potentially leading to costlier and less actionable recourses for individuals.

B Proof of Theorem 2

Proof. Throughout the proof let the event that $\mathcal{A}$ recieves (z, s) and outputs G = MEMBER be denoted by $\mathcal{A}(z, s) = 1$. First we prove the following simple lemma in the Appendix, which says that since the recourse is generated privately, the probability it takes on any value can't be changed by more than e^ϵ depending on whether a given point z is in the training set.

Lemma 1. *If $\mathbf{o}$ is any event, for any z:*

$$\Pr_{s\sim\mathcal{R}, D_t\sim\mathcal{D}^n}[s \in \mathbf{o} | z \in D_t] \leq e^\epsilon \Pr_{s\sim\mathcal{R}, D_t\sim\mathcal{D}^n}[s \in \mathbf{o} | z \notin D_t]$$

Proof. Fix arbitrary $D_t \backslash \{z\} = (z_2, \dots z_n)$. Expanding

$$\Pr_{s\sim\mathcal{R}, D_t\sim\mathcal{D}^n}[s \in \mathbf{o} | z \in D_t] = \int_{(z_2,\dots z_n)\sim\mathcal{D}^{n-1}} \Pr_{s\sim\mathcal{R}}[s \in \mathbf{o} | z, D_t\backslash\{z\} = (z_2, \dots z_n)] \Pr_{\mathcal{D}}[(z_2, \dots z_n)] \tag{5}$$

By the definition of differential privacy, for arbitrary z':

$$\int_{(z_2,\dots z_n)\sim\mathcal{D}^{n-1}} \Pr_{s\sim\mathcal{R}}[s \in \mathbf{o} | z, D_t\backslash\{z\} = (z_2, \dots z_n)] \Pr_{\mathcal{D}}[(z_2, \dots z_n)] \leq$$
$$\int_{(z_2,\dots z_n)\sim\mathcal{D}^{n-1}} e^\epsilon \Pr_{s\sim\mathcal{R}}[s \in \mathbf{o} | z', D_t\backslash\{z\} = (z_2, \dots z_n)] \Pr_{\mathcal{D}}[(z_2, \dots z_n)] \tag{6}$$

Since this holds for arbitrary z', we have:

$$\Pr[s \in \mathbf{o} | z \in D_t] \leq \inf_{z'} \left(\int_{(z_2,\dots z_n)\sim\mathcal{D}^{n-1}} e^\epsilon \Pr_{s\sim\mathcal{R}}[s \in \mathbf{o} | z', D_t\backslash\{z\} = (z_2, \dots z_n)] \Pr_{\mathcal{D}}[(z_2, \dots z_n)] \right) \leq$$
$$\mathbb{E}_{z'\sim\mathcal{D}} \left[\int_{(z_2,\dots z_n)\sim\mathcal{D}^{n-1}} e^\epsilon \Pr_{s\sim\mathcal{R}}[s \in \mathbf{o} | z', D_t\backslash\{z\} = (z_2, \dots z_n)] \Pr_{\mathcal{D}}[(z_2, \dots z_n)] \right] = e^\epsilon Pr[s \in \mathbf{o} | z \notin D_t], \tag{7}$$

as desired.

□

Recall that BA = $\frac{\text{FPR+TNR}}{2} = \frac{\Pr[\mathcal{A}(z,s)=1|z\in D_t]+\Pr[\mathcal{A}(z,s)=0|z\notin D_t]}{2}$, and let $a^1 = \Pr[\mathcal{A}(z,s) = 1|z \in D_t]$, and $a^0 = \Pr[\mathcal{A}(z,s) = 0|z \notin D_t]$. Then:

$$a^1 = \Pr[\mathcal{A}(z,s) = 1|z \in D_t] = \int_{s\in\mathcal{S}} \int_{z\in\mathcal{X}\times\mathcal{Y}} \Pr[\mathcal{A}(z,s) = 1|z,s] \Pr[s|z \in D_t] \Pr_{\mathcal{D}}[z] \leq$$
$$\int_{s\in\mathcal{S}} \int_{z\in\mathcal{X}\times\mathcal{Y}} \Pr[\mathcal{A}(z,s) = 1|x,s] \Pr_{\mathcal{D}}[x](e^{\epsilon} \Pr[s|z \notin D_t]) =$$
$$\int_{s\in\mathcal{S}} \int_{z\in\mathcal{X}\times\mathcal{Y}} \Pr[\mathcal{A}(z,s) = 1|x,s] \Pr_{\mathcal{D}}[x](e^{\epsilon} \Pr[s]) =$$
$$e^{\epsilon} \Pr[A(x,s) = 1|x \notin D_t] = e^{\epsilon}(1 - a_0), \quad (8)$$

where the inequality follows from Lemma 1. Rearranging $a^1 \leq e^{\epsilon}(1 - a_0)$ we get that $a_0 \leq 1 - e^{-\epsilon}a_1$. Hence:

$$\text{BA} \leq \frac{a^1 + (1 - e^{-\epsilon}a^1)}{2} = \frac{1}{2} + \frac{a_1(1 - e^{-\epsilon})}{2}$$

Since $a_1 \leq 1$ this gives the result, and we note that for small ϵ this can be improved to BA $\leq \frac{1}{2} + \frac{(2-e^{-\epsilon})(1-e^{-\epsilon})}{4}$. □

C TRAINING DETAILS

C.1 Classification Models

Details on the trained models are provided in tables 3 and 4.

	Model			
	$d = 100$	$d = 1000$	$d = 5000$	$d = 7000$
Train	0.957	0.958	0.973	0.976
Test	0.951	0.936	0.899	0.872

Table 3: Model performances for the logistic regression classifiers trained on the synthetic data sets. We measured performance in terms of classification accuracy.

Data set	$d = 50$				$d = 150$			
Model size	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test	0.8632	0.8748	0.8820	0.8806	0.9060	0.9196	0.9130	0.9208

Table 4: Model performances for the neural network classifiers trained on the synthetic data sets for different numbers of features. We measured performance in terms of classification accuracy. (a): two-layer network with 1000 hidden nodes. (b): three-layer neural network with 100 hidden nodes in each hidden layer. (c): three-layer neural networks with 333 hidden nodes in each hidden layer. (d): three-layer neural networks with 1000 hidden nodes in each hidden layer.

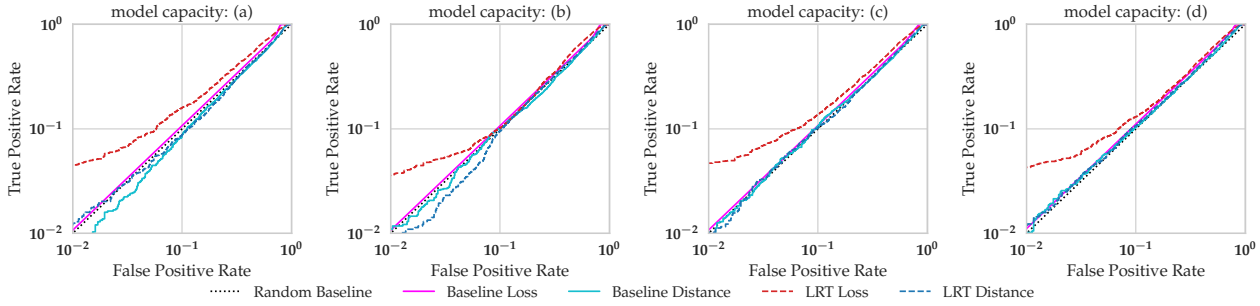
Data set	Adult ($d = 13$)	Heloc ($d = 23$)	Diabetes ($d = 42$)
Train	0.9755	1.00	0.9096
Test	0.8109	0.6701	0.5334

Table 5: Model performances for the neural network classifiers trained on the real-world data sets. We measured performance in terms of classification accuracy for two-layer network with 1000 hidden nodes.

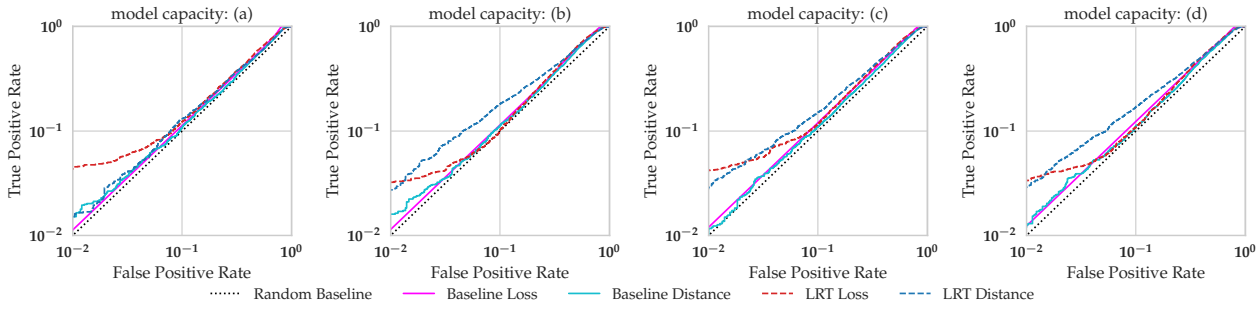
C.2 Recourse methods

- `SCFE`: As suggested in Wachter et al. [45], an Adam optimizer is used to optimize the recourse objective. We obtain recourses using an ℓ_1 distance function, and the binary cross entropy loss between the counterfactual label and the target.
- `GS`: The explanation model uses a counterfactual search algorithm in the input space. Particularly, instances are sampled within an ℓ_1-norm ball with search radius search radius r_i until recourse is successfully obtained. The search radius of the norm ball is increased until recourse is found.
- `C-CHVAE`: An autoencoder is additionally trained to model the data-manifold. The explanation model uses a counterfactual search algorithm in the latent space of the AE. Particularly, a latent sample within an ℓ_1-norm ball with search radius r_l is used until recourse is successfully obtained. The search radius of the norm ball is increased until recourse is found. All generative models use 8 latent dimensions, and 20 nodes in the first and third hidden layer.

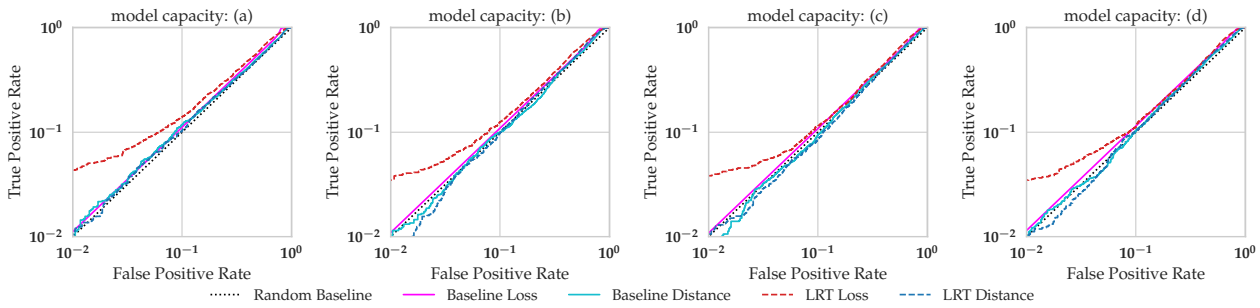
D ADDITIONAL EXPERIMENTAL RESULTS



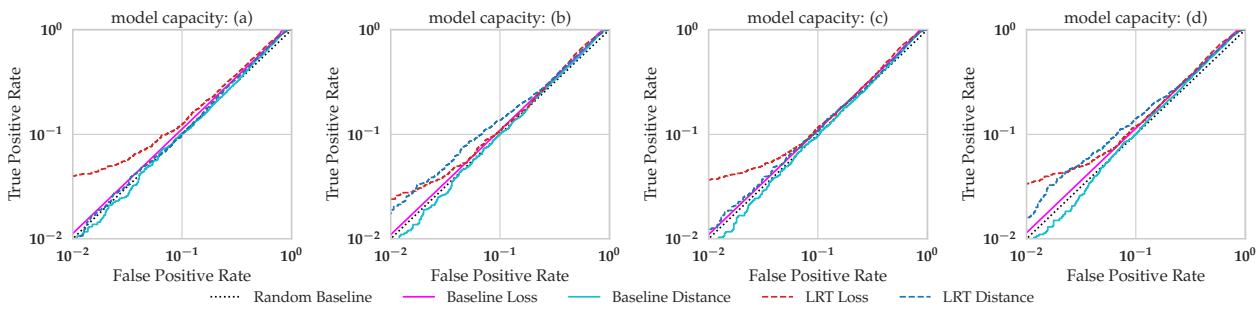
(a) $d = 50, n = 20000$



(b) $d = 150, n = 20000$

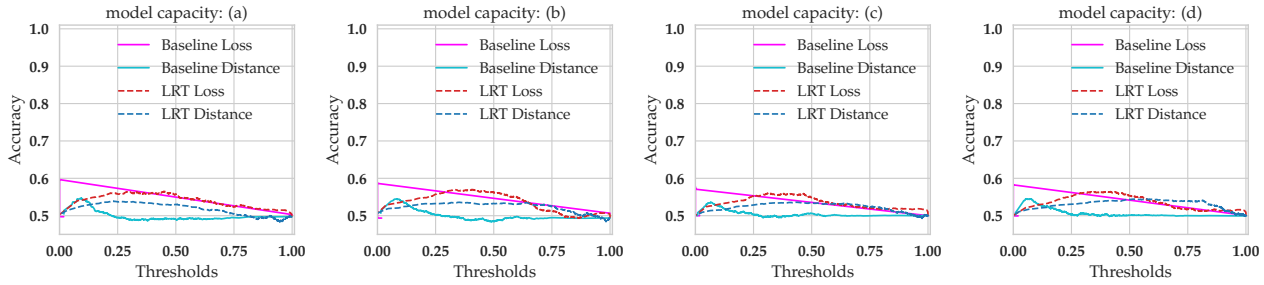


(c) $d = 50, n = 30000$

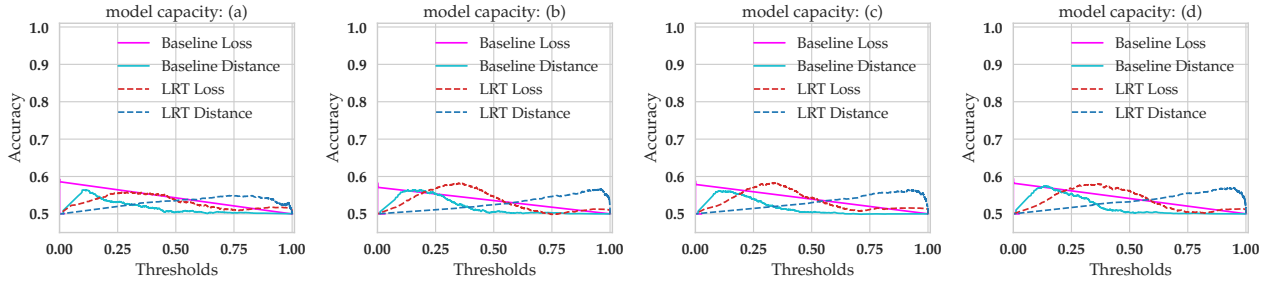


(d) $d = 150, n = 30000$

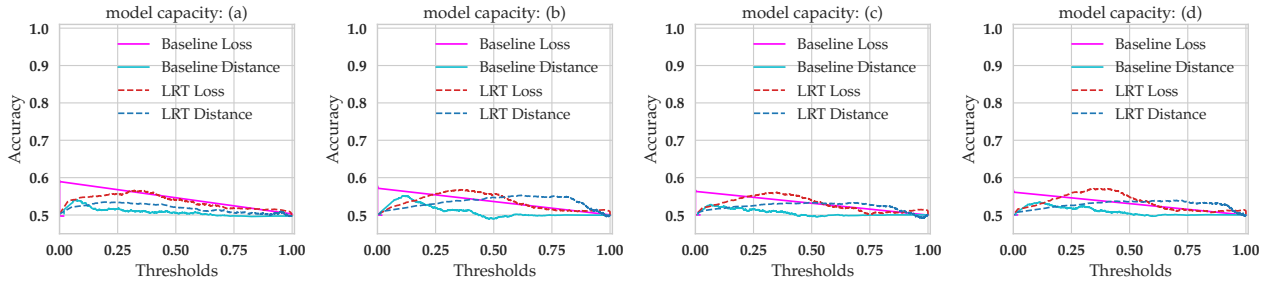
Figure 4: Demonstrating the efficacy of our proposed distance-based attack for neural network models trained on the synthetic data set when `SCFE` is used for the attack. (a): two-layer network with 1000 hidden nodes. (b): three-layer neural network with 100 hidden nodes in each hidden layer. (c): three-layer neural networks with 333 hidden nodes in each hidden layer. (d): three-layer neural networks with 1000 hidden nodes in each hidden layer.



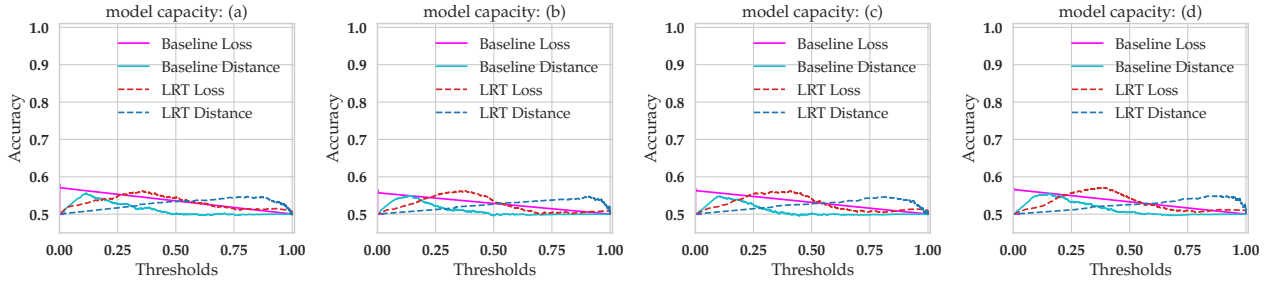
(a) $d = 50, n = 20000$



(b) $d = 150, n = 20000$



(c) $d = 50, n = 30000$



(d) $d = 150, n = 30000$

Figure 5: Demonstrating the efficacy of our proposed distance-based attack for neural network models trained on the synthetic data set when `SCFE` is used for the attack. (a): two-layer network with 1000 hidden nodes. (b): three-layer neural network with 100 hidden nodes in each hidden layer. (c): three-layer neural networks with 333 hidden nodes in each hidden layer. (d): three-layer neural networks with 1000 hidden nodes in each hidden layer. The thresholds have been normalized to the range $[0, 1]$ when necessary.

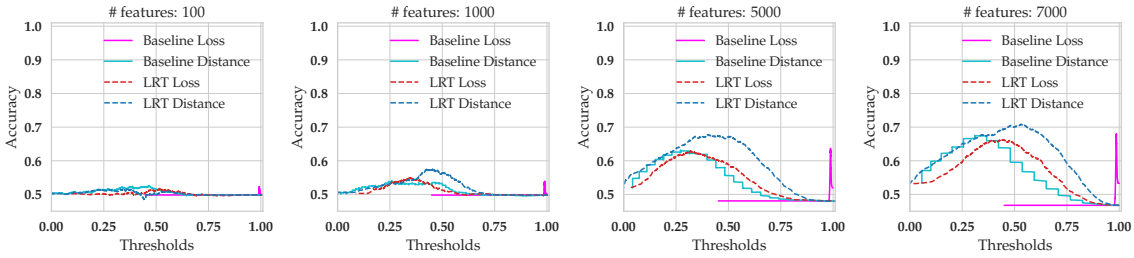


Figure 6: Demonstrating the efficacy of our proposed distance-based attack for logistic regression models trained on the synthetic data set by showing the attack accuracy as the threshold varies. Both the LRT-Distance as well as the baseline distance attacks are the most competetive attacks. The thresholds have been normalized to the range $[0, 1]$.