Accurate, Explainable, and Private Models: Providing Recourse While Minimizing Training Data Leakage

Catherine Huang [*1] **Chelse Swoopes** [*1] **Christina Xiao** [*1] **Jiaqi Ma** [1] **Hima Lakkaraju** [1]

Abstract

Machine learning models are increasingly utilized across impactful domains to predict individual outcomes. As such, many models provide algorithmic recourse to individuals who receive negative outcomes. However, recourse can be leveraged by adversaries to disclose private information. This work presents the first attempt at mitigating such attacks. We present two novel methods to generate differentially private recourse: Differentially Private Model (`DPM`) and Laplace Recourse (`LR`). Using logistic regression classifiers and real world and synthetic datasets, we find that `DPM` and `LR` perform well in reducing what an adversary can infer, especially at low `FPR`. When training dataset size is large enough, we find particular success in preventing privacy leakage while maintaining model and recourse accuracy with our novel `LR` method.

1. Introduction

Explainability and privacy are two important pillars of trustworthy machine learning (ML), but they are often viewed as conflicting. A right to privacy is often viewed as a limit on transparency (Weller, 2019). Still, users may want an explanation of how a ML system works or why it gave a particular outcome, even as they want their personal data to be kept private (Weller, 2019) (Shokri et al., 2021). As users are increasingly impacted by negative model predictions in domains such as healthcare, medicine, and criminal justice, there is a growing emphasis on providing algorithmic recourse to individuals, so that they can understand and contest decisions, or alter their behavior to achieve a preferred outcome (Wachter et al., 2017).

Recourse commonly takes the form of counterfactual explanations (CFEs), which highlight what feature changes a user would need to make for a model's predicted label to change (Pawelczyk et al., 2022). However, recent work by Pawelczyk et al. showed severe privacy risks with CFEs, as they developed two successful membership inference (MI) attacks utilizing CFEs to leak training data (2022).

In this work, we develop private recourse methods that protect against MI attacks while maintaining reasonable model accuracy. We hypothesize that the mathematical framework *differential privacy (DP)* can be used to create these private recourses (Dwork & Roth, 2014). Investigating the privacy, accuracy, and explainability trade-off is an under-explored area. Our work is the first attempt at mitigating the MI attacks presented by Pawelczyk et al. and the first work generating recourses in a DP manner.

2. Related Work

The research area of machine learning explainability and privacy is quite recent, and thus contains a smaller amount of prior work. Furthermore, most works have revolved around attacks leveraging explainability to the detriment of privacy, rather than creating explainable and still private models. Shokri et al. demonstrated that feature-based explanations may leak sensitive information about training data (2021). Milli et al. demonstrated that gradient-based explanations could be used to quickly reconstruct the underlying model (2019). Aïvodji et al. showed that CFEs could also be leveraged for highly accurate model extraction attacks (2020).

We are interested in CFEs that may expose information on training data. Pawelczyk et al. provided the first work in this area, with two novel counterfactual distance-based MI attacks: **1)** thresholding on counterfactual distance (`CFD`), and **2)** likelihood ratio test using counterfactual distance (`CFD LRT`) (2022). We discuss the details of these attacks, which are the basis of our work, in Section 3. With enough recourse queries, adversaries can use these attacks to reconstruct the training data of a non-private, recourse-supporting

[*]Equal contribution [1]Department of Computer Science, Harvard University School of Engineering and Applied Sciences, Cambridge, Massachusetts. Correspondence to: Catherine Huang <catherinehuang@college.harvard.edu>, Chelse Swoopes <cswoopes@g.harvard.edu>, Christina Xiao <christinaxiao@college.harvard.edu>, Jiaqi Ma <jima@hbs.edu >, Hima Lakkaraju <hlakkaraju@hbs.edu>.

2^{nd} *AdvML Frontiers workshop at* 40^{th} *International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023.

model (Pawelczyk et al., 2022).

3. Preliminaries

3.1. Algorithmic Recourse

We base our recourse definition off of Wachter et al.'s first work in this area (2017). Wachter et al.'s definition is also adopted by our motivating work (Pawelczyk et al., 2022) and other explainability-privacy papers.

Definition 3.1. Let $x \in \mathcal{X}$ be a data observation that received a negative label when fed through f_θ, where $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ is a classifier model parameterized by θ, $\mathcal{X} \in \mathbb{R}^d$, and $\mathcal{Y} \in \{0, 1\}$.

Finding an **algorithmic recourse** for x means finding a counterfactual

$$x' = x + \delta : f_\theta(x') = f(x + \delta) = 1.$$

We aim to minimize the cost $c(x, x')$ to change x to x' so that the recourse is easily implementable; in practice, ℓ_2 distance is commonly used as a cost function.

3.2. Counterfactual Distance-Based Membership Inference Attacks for ML Models

MI attacks infer whether an instance x belongs to the training data for f_θ. Across the literature, MI attacks are commonly *loss-based*, following the intuition that models have lower loss on instances observed during training (Yeom et al., 2017) (Carlini et al., 2021). Pawelczyk et al.'s novel counterfactual distance-based MI attacks show that counterfactual distance can also be used to leak training data (2022).

Thresholding on counterfactual distance (`CFD`): Intuitively, during training, the decision boundary is pushed away from training points (as in margin maximization), resulting in test set points being closer to the decision boundary.

$$M_{\texttt{CFD}}(x) = \begin{cases} \texttt{MEMBER}, & \text{if } c(x, x') \geq \tau_{\mathcal{D}} \\ \texttt{NON-MEMBER}, & \text{if } c(x, x') < \tau_{\mathcal{D}} \end{cases}.$$

Counterfactual distance likelihood ratio test (`CFD LRT`): In this attack, the adversary trains shadow models to estimate the likelihood ratio

$$\Lambda = \frac{\Pr[c(x, x')|x \in \mathcal{D}_t]}{\Pr[c(x, x')|x \notin \mathcal{D}_t]}.$$

The attack then thresholds on this Λ.

See Appendix .1 for the formulation and implementation details of this attack.

3.3. Differential Privacy

Our solution involves using *differential privacy* (DP) to counteract the success of the counterfactual distance-based MI attacks in Section 3.2. DP is a mathematically provable definition of privacy that provides a quantifiable metric of privacy loss, providing a computational method whose output is random enough to obscure any single participant's presence in the training data (Dwork & Roth, 2014).

Definition 3.2. A randomized mechanism $\mathcal{M}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies ε**-differential privacy** (ε-DP) if for any two adjacent input datasets $d, d' \in \mathcal{D}$ differing by one row, and any subset of outputs $S \subseteq \mathcal{R}$

$$Pr[\mathcal{M}(d) \in S] \leq e^\varepsilon Pr[\mathcal{M}(d') \in S].$$

The ε parameter represents *privacy loss*: the lower the ε, the stronger the privacy protection.

3.3.1. Laplace Mechanism of Differential Privacy

The Laplace Mechanism, a widely used DP mechanism, is useful on numerical queries. It involves adding Laplacian distributed random noise on the output of a sensitive query (Dwork & Roth, 2014).

Definition 3.3. For sensitive query $f(d)$ on input dataset d, the ε**-DP Laplace Mechanism** $\mathcal{M}_{Lap}$ is defined as

$$\mathcal{M}_{Lap}(d) = f(d) + \text{Laplace}(GS_f/\varepsilon),$$

where $\text{Laplace}(GS_f/\varepsilon)$ is a Laplace random variable with scale parameter GS_f/ε.

GS_f is the *global sensitivity* of query f, bounding how much the sensitive query outcome can change across any two possible neighboring datasets $d, d' \in \mathcal{D}$:

$$GS_f = \max_{d,d'} ||f(d) - f(d')||_1.$$

3.3.2. Differential Privacy Under Post-Processing

Dwork et al. prove that once a quantity is "made private" through DP, it cannot be subsequently "made un-private" (2014). This is formalized in the following theorem.

Theorem 3.4. *If mechanism M is ε-DP, and G is an arbitrary deterministic mapping, then $G \circ M$ is also ε-DP.*

4. Problem Statement & Methodology

4.1. Problem Statement

We hypothesize that DP can be used as a privacy preservation mechanism to protect algorithmic recourse models from MI attacks. To thoroughly evaluate the extent and nature of the impact of DP on CFD based attack success, we evaluate the influence of the following changes to our *ML classifier* $\rightarrow$ *algorithmic recourse* $\rightarrow$ *MI attack* pipeline:

- The presence versus absence of differential privacy.
- The particular DP mechanism being used: `DPM` versus `LR` (see Section 4.2.2).
- The privacy protection strength ε of the DP mechanism.
- The type of MI attack: `CFD` versus `CFD LRT`.
- The dataset, whether synthetic or real-world.
- The dimensionality of the data, for synthetic data.

4.2. Methodology

4.2.1. Recourse for Logistic Regression Classifiers

Logistic regression, our classifier of choice, has weights $\mathbf{w}$ after training which it uses to output a probability score: $f(x) = \mathbf{w}^T x = \log \frac{\Pr(y=1|x)}{1-\Pr(y=1|x)}$. In a linear model such as this one, it is standard for the counterfactual distance of instance x to be calculated using the ℓ_2 norm from $f(x)$ to the target score s in logistic regression space:

$$c(x, x') = \frac{s - f(x)}{||\mathbf{w}||_2^2}\mathbf{w}.$$

This is the counterfactual distance calculation method we use in our methodology. In our experiments, we set our decision boundary threshold to $s = 0$, which corresponds to the fitted $\Pr(y = 1|x)$ being equal to $\frac{1}{2}$ at the threshold.

4.2.2. Differentially Private Recourse Generation Methods

Differentially Private Model (`DPM`)

Our first DP method trains the *underlying* logistic regression classifier with DP. By post-processing of DP (see Section 3.3.2), an ε-DP logistic regression model gives rise to ε-DP counterfactual recourse. We use IBM's `diffprivlib` library (Holohan et al., 2019), which offers an implementation of DP logistic regression based on Chaudhuri et al.'s formulation of a DP empirical risk minimization mechanism (2009).

Differentially Private Laplace Recourse (`LR`)

We propose a novel method for DP post-hoc computation of counterfactual recourse that does not touch the underlying logistic regression model training process. The method is as follows:

1. Apply the Laplace mechanism onto the *predicted probability score* $\Pr'(y = 1|x) = \Pr(y = 1|x) + \text{Laplace}(1/\varepsilon)$.
2. Clamp $\Pr'(y = 1|x)$ to $[0, 1]$ so that $\Pr'(y = 1|x)$ can still be interpreted as a probability.
3. Calculate noisy logistic regression score $f'(x) = \log \frac{\Pr'(y=1|x)}{1-\Pr'(y=1|x)}$.
4. Calculate noisy CFD: $\mathcal{M}_{\texttt{CFD},Lap}(x) = \frac{s-f'(x)}{||\mathbf{w}||_2^2}\mathbf{w}$.

Claim: The above method is ε-DP.

Explanation: Our method begins with applying the Laplace mechanism (see Section 3.3.1) on the predicted probability $\Pr(y = 1|x)$, for data instance x. We claim that $GS_{P(y=1|x)} = 1$, i.e. the vector of all ones. First, note that $\Pr(y = 1|x)$ is a probability vector $\in [0, 1]^d$. For any two possible datasets $d, d' \in \mathcal{D}$, if we calculate $\Pr(y = 1|x)_d$ using the classifier trained on d, and then calculate $\Pr(y = 1|x)_{d'}$ using the classifier trained on d', the two probabilities can differ in ℓ_1 distance by at most 1. Steps 2-4 in the method are post-processing functions applied to $\Pr'(y = 1|x)$. By post-processing of DP (see Section 3.3.2), we retain ε-DP.

5. Experimental Results

5.1. Setup

Datasets: To stay relevant and methodologically consistent with the motivating work (Pawelczyk et al., 2022), we use similar datasets. For real world data, we use the datasets **1)** Heloc (Home Equity Line of Credit) (FICO Community) ($d = 23$), which scores whether individuals will repay their Heloc accounts within a fixed time window, **2)** MNIST (Khodabakhsh et al., 2019), which contains 28×28 pixel gray-scale images of handwritten digits between 0 and 9, and **3)** Adult (Dua & Graff, 2017) ($d = 14$), a variant of the 1994 Census database that labels whether an individual has annual income greater than \$50,000. However, based on the minimal success of baseline `CFD` attacks on Adult, we omit this analysis from this writeup, although our results are presented in the Appendix .2.

For synthetic data, we follow Pawelczyk et al. (2022) and Shokri et al (2021): For $d \in \{100, 1000, 5000, 7000\}$, we randomly choose a vertex from a d-dimensional hypercube and sample $n = 5000$ random variables from a Gaussian distribution centered at the vertex with unit variance.

Pre-processing: We use 5000 data entries for each model training set, and give the adversary 5000 entries to train their own shadow models. Before model fitting, we pre-process data by removing multicollinear features (with correlation over 0.95), standardizing, and normalizing so that each feature's ℓ_2 norm is 1.

Attack specifications: For both attacks, we calculate counterfactual distance based on ℓ_2 norm to the decision boundary. For `CFD LRT` attacks, we consider two versions with global and local variance estimators, respectively, following the paper that introduced the `LRT` MI attacks (Carlini et al., 2021) that the motivating work (Pawelczyk et al., 2022) is based upon. In each `CFD LRT` attack, we train 5 shadow models and 20 ensemble models.

Settings: We evaluate the baseline setting as proposed by Pawelczyk et al. (2022), alongside our novel DP model (`DPM`) and Laplace recourse (`LR`) methods (refer to Section 4.2.2), each with $\varepsilon = 0.5, 1.0$.

5.2. Metrics

Following the motivating work (Pawelczyk et al., 2022), we use log-scaled `ROC` curves (receiver operating characteristic), `AUC` (area under the curve), and `BA` (balanced accuracy) to determine the efficacy of the MI attacks. Figs 1, 2, and 3 show these metrics. Based on motivating literature (Carlini et al., 2021) (Pawelczyk et al., 2022), we pay particular attention to successes at low `FPR`, as a MI attack is still successful if it can identify even a very small subset of the training data with high confidence.

The DP literature acknowledges the tradeoff of privacy and accuracy (Dwork & Roth, 2014). Because we aim to create models that are at once accurate, private, and explainable, we also consider the train accuracy on the last ensemble model, and the test accuracy across all 20 ensemble models. To examine recourse accuracy, we also compare the train and test distributions of CFDs and CFD LRTs on synthetic data, under all three settings (baseline, `LR`, `DPM`); these results are in Fig 4.

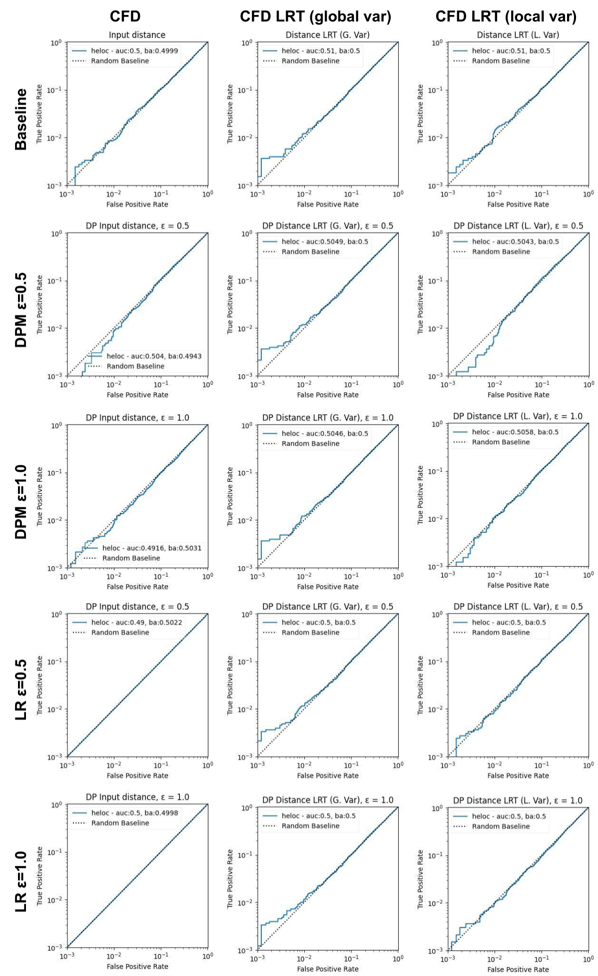

Figure 1. Log-scaled `ROC` curves (`TPR` v. `FPR`), `AUC`, and `BA` for all attacks on all settings using Heloc dataset. See **Privacy:** for analysis.

5.3. Results and Analysis

Privacy: In Figs 1, 2, and 3, we hoped to see the `DPM` and `LR` methods flatten the `ROC` curves of the baseline, towards the random line with lower `AUC`, for all attacks—particularly at low `FPR`, as explained in Section 5.2. This would show the success of our methods in reducing the efficacy of the MI attacks. Overall, we see that both DP methods generally do flatten the `ROC` curves from the baseline, particularly at low `FPR`, across all datasets. For a particularly impressive example, compare `DPM` $\varepsilon = 1.0$ and baseline for `CFD LRT` (global variance) in Fig. 2. As expected, $\varepsilon = 0.5$, with more DP, protects against attacks more successfully than $\varepsilon = 1.0$.

At both values of ε, `LR` appears highly successful against `CFD`, with almost exactly random lines and `AUC` $= 0.5$, across all datasets. Similarly, `DPM` appears slightly more successful against both `CFD LRT` attacks, particularly in Fig. 2. This makes intuitive sense since both `LR` and `CFD` happen on the level of recourse, while `DPM` and `CFD LRT` happen on the level of model training.

The motivating work proposes this theorem: For a ε-DP recourse mechanism—such as our `LR`—the `BA` of all attacks is bounded by $\frac{1}{2} + \frac{1-e^{-\varepsilon}}{2}$ (Pawelczyk et al., 2022). For $\varepsilon = 0.5, 1.0$, this bound is $0.697, 0.816$, respectively. We are pleased to announce that our empirical `BA` far surpasses this theoretical bound, with `BA` across nearly all DP methods, attacks, and datasets at approximately 0.5.

Model Accuracy: Considering train versus test accuracies in Tables 1 and 2, we are not enthused about the tradeoff in accuracy and privacy under `DPM`. `LR` seems the more fruitful method.

Recourse Accuracy: Another way to assess accuracy is to compare the distributions of the DP CFD calculations with the baseline CFD distribution. These distributions are shown in Fig 4. For dimension $d \geq 1000$, both `DPM` and `LR` have inaccurate CFD distributions, as seen through large disparities in horizontal axis scaling. For $d = 100$, however,

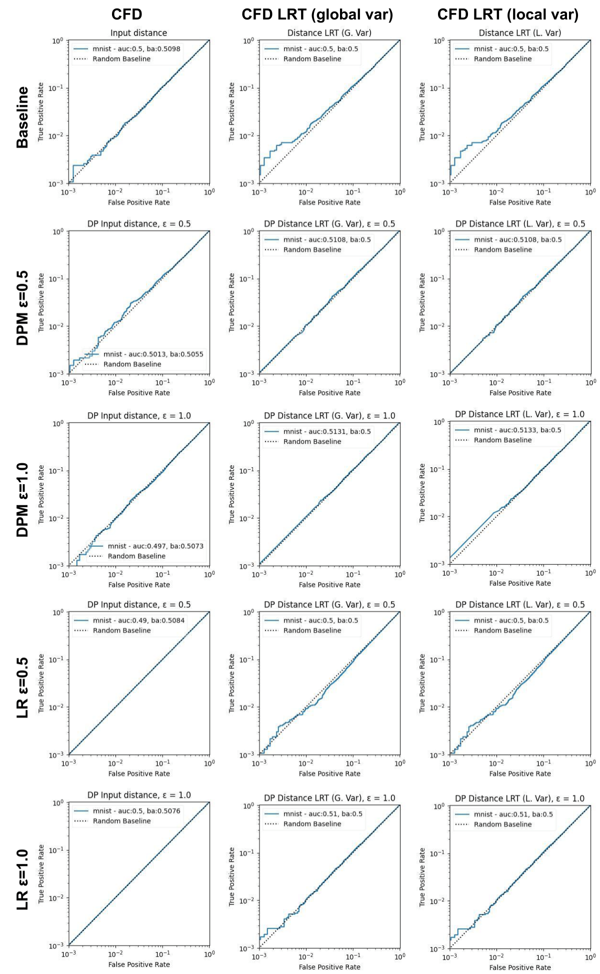

Figure 2. Same as Fig. 1, but using MNIST dataset.

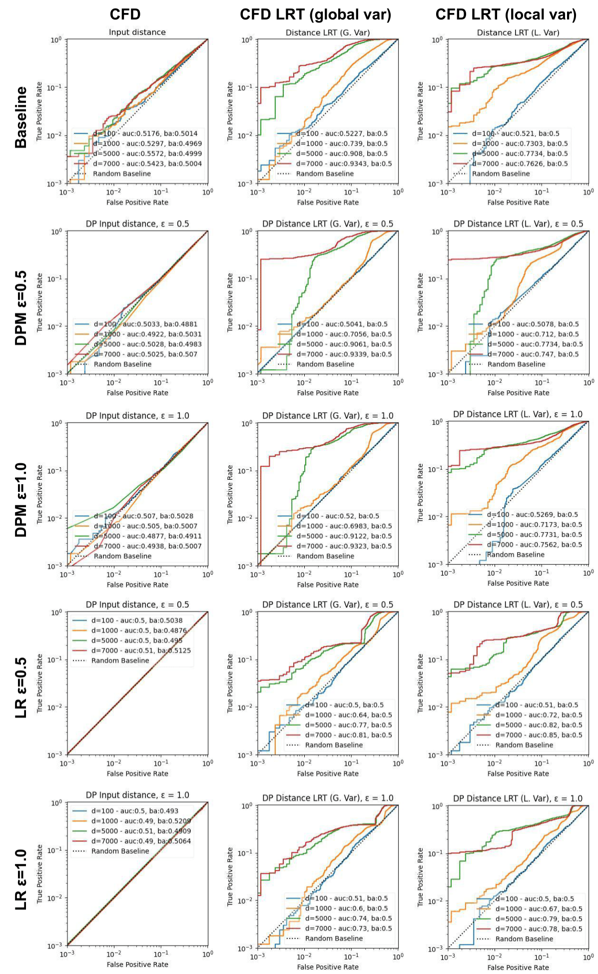

Figure 3. Same as Fig. 1, but using synthetic dataset.

Table 1. Train accuracy of last ensemble model and test accuracy over all 20 ensemble models, using real world datasets. (LR is excluded because it has no impact on model training; we observed very similar empirical results, with only minor variation based on the split of training data points between ensemble models.) We see that training with DP significantly lowers train and test accuracy for both datasets. As expected, $\varepsilon = 1.0$ is much more accurate than $\varepsilon = 0.5$ in training — but it is only slightly more accurate in testing.

	Heloc		Mnist	
	Train	**Test**	**Train**	**Test**
Baseline	0.9046	0.8506	1.0	0.9393
DPM, $\varepsilon = 0.5$	0.6079	0.5254	0.6239	0.4795
DPM, $\varepsilon = 1.0$	0.8011	0.5648	0.7317	0.4985

Table 2. Same as Tab. 1, but using synthetic datasets. Again, we see that training with DP lowers train and test accuracy across all datasets, and that $\varepsilon = 1.0$ provides more accuracy than $\varepsilon = 0.5$. However, given the low test accuracy on synthetic data under even the baseline condition, the discrepancy in test accuracy under DP is less noticeable.

	$d = 100$		$d = 1000$	
	Train	**Test**	**Train**	**Test**
Baseline	1.0	0.6799	1.0	0.5619
DPM, $\varepsilon = 0.5$	0.5323	0.4933	0.4913	0.5016
DPM, $\varepsilon = 1.0$	0.5386	0.5047	0.5165	0.5029

	$d = 5000$		$d = 7000$	
	Train	**Test**	**Train**	**Test**
Baseline	1.0	0.5261	1.0	0.5215
DPM, $\varepsilon = 0.5$	0.4551	0.4953	0.515	0.5035
DPM, $\varepsilon = 1.0$	0.4693	0.501	0.5244	0.5012

LR offers a CFD distribution close to that of the baseline, with the correct scaling. We hypothesize that the high bar to the right in the LR case—the main noticeable difference when compared with baseline—is a result of the clamping in the novel LR method, where we clamp the noisy predicted probability score to [0, 1], so that the score can still be interpreted as a probability. Namely, we believe the bar

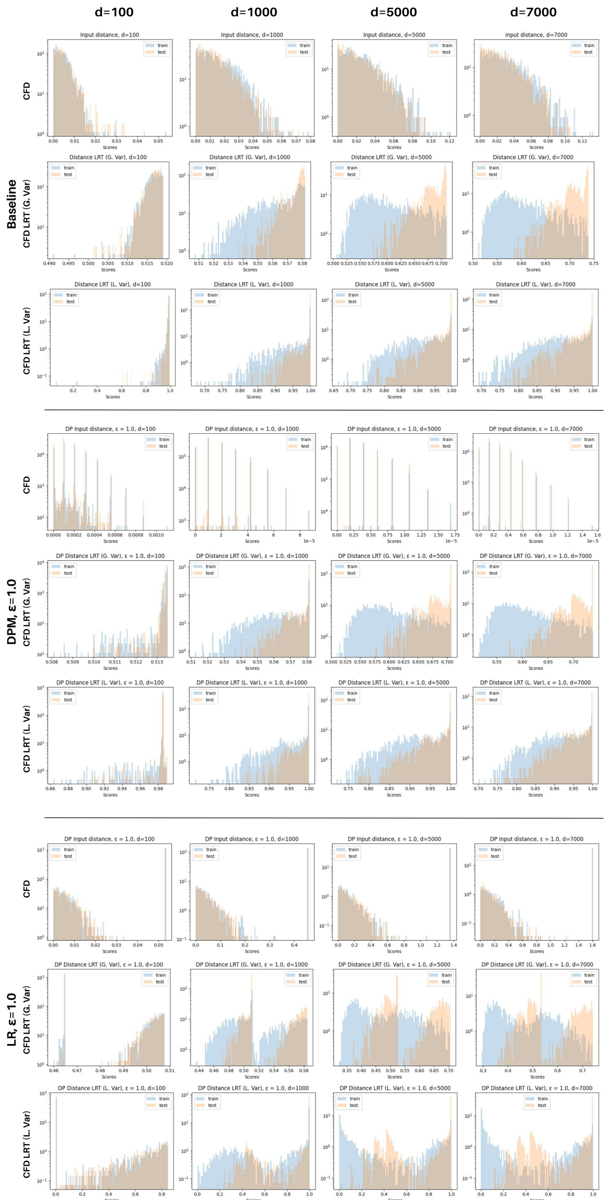

Figure 4. Histograms of train and test counterfactual distances (for the Input Distance (CFD) attack) and CFD LRTs (for the LRT attacks), under the synthetic data, $\varepsilon = 1.0$ setting. We see that as data dimension increases, train and test distributions differ, even under differentially private recourse generation. This difference is especially pronounced in CFD LRTs. These histograms also offer insight on DP recourse accuracy, explained in Section 5.3

.

corresponds to observations that received a negative label, and whose noisy predicted probabilities were originally negative and then clamped to 0. We proceed to test this hypothesis by examining the impact of privacy strength (ε) on the severity of this clamping-induced issue. Not only do these results reveal the privacy-accuracy trade-off (prevalent in and consistent with DP literature), but point to LR as promising for maintaining recourse accuracy.

On the Relationship Between Privacy Strength and Recourse Accuracy

The Laplace Mechanism (Definition 3.3) involves adding Laplace-distributed random noise to a sensitive numerical query, where the noise's scale parameter is proportional to $1/\varepsilon$. Lower ε corresponds to stronger privacy protection under DP, but for our LR method, it also means there is higher perturbation of the logistic regression's predicted probability scores, and higher chance that a noisy predicted probability score falls below 0 (upon which it is clamped to 0). The CFD distributions in Figure 4 illustrate a consequence of this clamping and is a source of accuracy loss of LR-generated recourse. In this appendix, we examine whether decreased privacy protection (higher ε) mitigates this consequence of clamping and yields better CFD distribution accuracy. Figure 5 shows CFD distributions for $d = 100$ synthetic data, under the LR, $\epsilon = \{5, 10, 20\}$ settings.

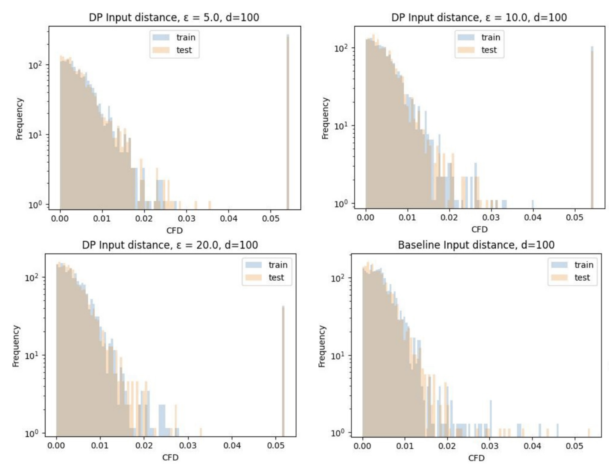

Figure 5. Histograms of LR-based train and test counterfactual distances from models trained with synthetic data, under the settings $d = 100, \varepsilon = \{5, 10, 20\}$. The LR-based CFD distribution approaches the baseline CFD distribution as we increase ε.

These findings highlight a privacy-accuracy tradeoff (also referred to as privacy-utility tradeoff in DP literature): the CFD distribution is more accurate for larger values of ε (i.e. weaker levels of privacy protection).

Even though the values of ϵ presented here are larger than those presented in main results (and those used as library default parameters), it is not uncommon for ε values of up to 10 (and at times up to 20) to be used in the literature: (Abadi et al., 2016), (Jayaraman & Evans, 2019). Overall, this is a promising result for the accuracy of LR-based recourse.

6. Future Work

The accuracy and utility of DPM and LR-generated recourse are worth further exploration; our work is a start. Our recourse accuracy analysis in Section 5.3 highlights privacy-accuracy tradeoff. Where on this tradeoff spectrum we choose to lie depends on factors such as the motivation behind generating recourse in the first place, the context and end users behind the model, and the risk of an adversary taking hold of data and our trained model. Further research is needed to assess the human interpretability of the new recourses.

Figures 3 and 4 show that as dimension of synthetic data increases—especially as dimension passes the interpolation threshold (i.e., when the number of training points equals the dimension: $d = n = 5000$)—MI attack success remains difficult to prevent, even under differentially private recourse generation methods. On high-dimensional datasets with too few training points, this finding highlights a continued privacy concern (even though such datasets are discouraged in the real world for accuracy and privacy purposes). Addressing this difficulty is worth further exploration.

There is practical value in exploring privacy risk mitigation in recourse-based membership inference attacks when the underlying model is a non-linear or non-inherently interpretable classifier, such as a neural network. (Abadi et al., 2016) propose a differentially private stochastic gradient descent (DP-SGD) mechanism by adding Gaussian distributed random noise to gradients during training. For neural network classifiers—and for general classifiers trained with gradient descent—we hypothesize that DP-SGD with carefully tuned hyperparameters can help us achieve recourse with lowered privacy risk.

7. Conclusion

We develop two methods to generate differentially private recourse, in order to protect against MI attacks leveraging explainability. We find particular success in preventing privacy leakage while maintaining model and recourse accuracy with LR (Laplace recourse), especially on CFD attacks, and especially when training dataset size is larger than dimension.

Our work leaves behind remaining difficulties in this field. CFD LRT attacks remain more effective than CFD attacks. On synthetic datasets of high dimensionality, attacks remain hard to prevent, as distances are simply greater in high dimensions. While our LR method appears promising, research is needed to determine if the Laplace mechanism harms the human interpretability of the new recourses, in particular under high strength levels of differential privacy. Finally, while we considered a logistic regression classifier, this is an inherently interpretable model; it would be worthwhile investigating whether our results generalize to a blackbox neural network scenario.

8. Acknowledgements

The corresponding authors would like to thank Professor Hima Lakkaraju for motivating this work and for thoughtful feedback and discussions throughout the project. We also thank Postdoctoral Fellow Jiaqi Ma for their detailed guidance and feedback throughout the research process. We would like to also thank Martin Pawelczyk for his guidance on the implementation and evaluation portions of this work, and for clarifying to us ideas from his work (Pawelczyk et al.). Lastly, we thank the reviewers for their valuable comments and suggestions to improve our work.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Aïvodji, U., Bolot, A., and Gambs, S. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884*, 2020.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. Membership inference attacks from first principles. *CoRR*, abs/2112.03570, 2021. URL https://arxiv.org/abs/2112.03570.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private support vector machines. *CoRR*, abs/0912.0071, 2009. URL http://arxiv.org/abs/0912.0071.

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

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, aug 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

FICO Community. Explainable machine learning challenge. URL https://community.fico.com/s/explainable-machine-learning-challenge?tabset-158d9=3.

Holohan, N., Braghin, S., Mac Aonghusa, P., and Levacher, K. Diffprivlib: the IBM differential privacy library. *ArXiv e-prints*, 1907.02444 [cs.CR], July 2019.

Jayaraman, B. and Evans, D. Evaluating differentially private machine learning in practice, 2019.

Khodabakhsh, H., LeCun, Y., Cortes, C., and Burges, C. J. Mnist dataset, 2019. URL https://www.kaggle.com/datasets/hojjatk/mnist-dataset.

Milli, S., Schmidt, L., Dragan, A. D., and Hardt, M. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 1–9, 2019.

Pawelczyk, M., Lakkaraju, H., and Neel, S. On the privacy risks of algorithmic recourse. *arXiv*, abs/2211.05427, 2022. URL https://arxiv.org/abs/2211.05427.

Shokri, R., Strobel, M., and Zick, Y. On the privacy risks of model explanations. *arXiv*, abs/1907.00164, 2021. URL https://arxiv.org/abs/1907.00164.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Weller, A. Transparency: Motivations and challenges. *arXiv*, abs/1708.01870, 2019. URL https://arxiv.org/abs/1708.01870.

Yeom, S., Fredrikson, M., and Jha, S. The unintended consequences of overfitting: Training data inference attacks. *CoRR*, abs/1709.01604, 2017. URL http://arxiv.org/abs/1709.01604.

Appendix

.1. CFD LRT Attack Detailed Formulation

Section 3.2 introduces the CFD LRT attack that (Pawelczyk et al., 2022) show successfully leak sensitive training data membership. To fully estimate the likelihood ratio, the adversary uses maximum likelihood estimation (MLE) methods to model the distributions of counterfactual distances (CFDs) when **1)** x (the point in question) is in the training data, and when **2)** x is in the test data. These building blocks comprise the numerator and denominator, respectively, of the likelihood ratio Λ.

(Pawelczyk et al., 2022) models the distributions of CFDs as log-normal (parameterized by mean and standard deviation), meaning MLE estimates are of (μ_{in}, σ_{in}) and $(\mu_{out}, \sigma_{out})$. In the full estimation process, the adversary — who has access to training data distribution $\mathcal{D}$ — can estimate all four parameters by training shadow models with and without point x, and then computing the resulting CFDs.

However, this process would entail sampling an adversary training set and training a shadow model separately for each data entry x that we perform the attack on. This is computationally intractable in practice. To work around this, Pawelczyk et al. show and implement a *one-sided* version of the LRT, where we only estimate μ_{out}, σ_{out}, and the attack predicts MEMBER if $c(x, x')$'s likelihood ratio is sufficiently low under such parameters. Conveniently, since μ_{out}, σ_{out} do not depend on x, we need only train shadow models once.

Algorithm 1 shows a detailed formulation of the proposed one-sided CFD LRT attack, tailored to the linear classifier case.

Algorithm 1 CFD-based Likelihood Ratio Test (CFD LRT) for Linear Classifier

Inputs: x: point in question. $t_0 = c(x, x')$: x's CFD in the model trained on owner training data. $\mathcal{D}$: training data distribution. α: FPR. N: number of shadow models.
estimatedCFD = []
Compute: $t_0 = c(x, x')$
for $i \leftarrow 1, \ldots, N$ **do**
  Sample $\mathcal{D}_t^{(i)} \sim \mathcal{D}$ {Adversary's training set for training shadow model i.}
  $f_{\theta^{(i)}} = \texttt{TrainShadowClassifier}(\mathcal{D}_t^{(i)})$ {ith shadow model.}
  $c(x, x'^{(i)}) = \texttt{GetCFD}(x, f_{\theta^{(i)}})$
  estimatedCFD $\leftarrow c(x, x'^{(i)})$
end for
$\hat{\mu}_{\text{out, MLE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\log c(x, \mathbf{x'}^{(i)}) \right)$
$\hat{\sigma}^2_{\text{out, MLE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{\mu}_{\text{out, MLE}} - \left(\log c(x, \mathbf{x'}^{(i)}) \right) \right)^2$ {$z_{1-\alpha}$ is the $1 - \alpha$ quantile of $Z \sim \text{LogNormal}(\hat{\mu}_{\text{out, MLE}}, \hat{\sigma}^2_{\text{out, MLE}})$.}
if $t_0 > z_{1-\alpha}$ **then**
  Output: NON-MEMBER
else
  Output: MEMBER
end if

.2. Experimental Results on Adult Dataset

See Fig. 6.

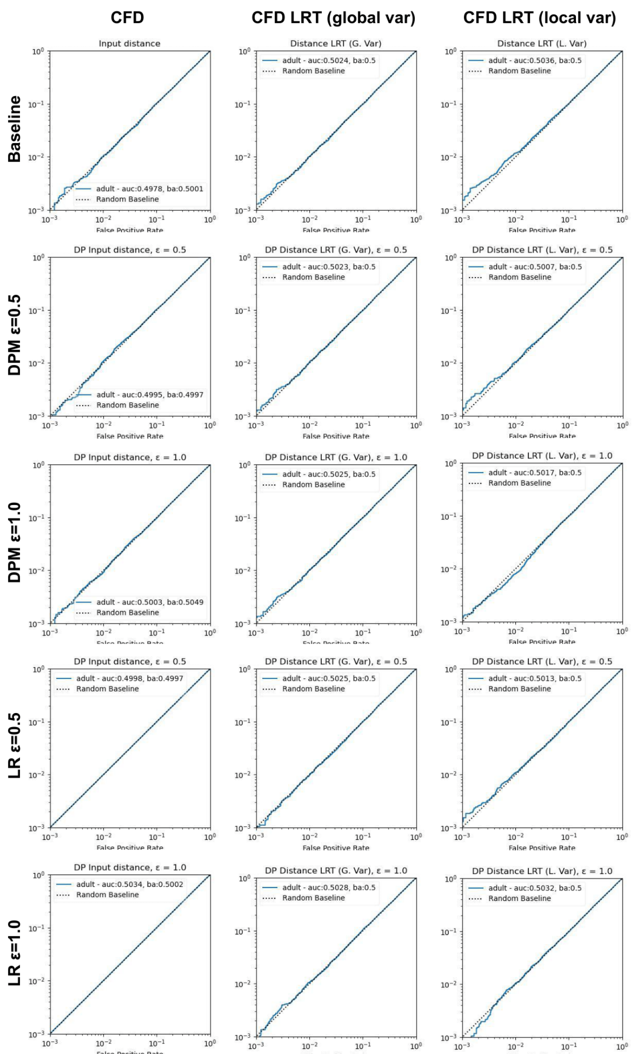

Figure 6. Log-scaled ROC curves (TPR v. FPR), AUC, and BA for all attacks on all settings using Adult dataset. ROC curves are near-random and AUC is around 0.5 even on baseline, signifying a lack of success of the attacks and therefore a negligible potential for our DP methods to help.