Model Explanations with Differential Privacy

Neel Patel*
neelbpat@usc.edu
University of Southern California
USA

Reza Shokri
reza@comp.nus.edu.sg
National University of Singapore
Singapore

Yair Zick†
yzick@umass.edu
University of Massachusetts Amherst
USA

Abstract

Using machine learning models in critical decision-making processes has given rise to a call for *algorithmic transparency*. Model explanations, however, might leak information about the sensitive data used to train and explain the model, undermining data *privacy*. We focus on *black-box feature-based* model explanations, which locally approximate the model around the point of interest, using potentially sensitive data. We design differentially private local approximation mechanisms, and evaluate their effect on explanation quality. To protect training data, we use existing differentially private learning algorithms. However, to protect the privacy of data which is used during the local approximation, we design an adaptive differentially private algorithm, which finds the minimal privacy budget required to produce accurate explanations. Both empirically and analytically, we evaluate the impact of the randomness needed in differential privacy algorithms on the fidelity of model explanations.

Keywords

Differential Privacy, Model Explainations

ACM Reference Format:
Neel Patel, Reza Shokri, and Yair Zick. 2022. Model Explanations with Differential Privacy. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3531146.3533235

1 Introduction

Algorithmic transparency for black-box algorithms in high-stakes domains is a foundational element of trustworthy machine learning [19, 23]. This has given rise to methods that offer additional information about the underlying algorithmic decision-making process, attempting to answer the question: what induces the model to offer a particular prediction for a particular input data point? A large class of model explanations is based on model-agnostic *post-hoc*[1] feature influence measures, highlighting the most influential features in the input data [6, 12, 13, 24, 31, 39].

We focus on black-box model explanations based on local approximation [6, 16, 24, 31, 39, 41]. These methods generate an explanation for the point of interest by approximating the model around it, using the model's predictions on a set of data points (referred to as the *explanation dataset*).

Offering additional insight about the models can pose significant *data privacy risks*, by leaking information about their training/explanation data, that can be exploited by inference attacks [34]. The model explanations are based on model predictions on explanation data. Even though these predictions are not directly revealed to the adversary, they are used to generate explanations, which may indirectly leak information, either about the data used to train the model, or the data used to generate the explanation. While we distinguish between the training data and the explanation data, they are often the same or drawn from the same distribution (to maximize the explanation fidelity). Therefore, a comprehensive privacy protection with respect to all the used data is necessary.

Thus far, there has been little work on modeling the relation between algorithmic transparency and privacy [13, 20, 34], and on designing *model explanations that protect data privacy*. More importantly, the negative impact of randomized differential privacy algorithms on the *fidelity of model explanations* is not well studied.

1.1 Our Contributions

In this work, we propose a framework for generating differentially-private model-agnostic post-hoc explanations for black-box machine learning algorithms. We focus on model explanations based on local approximation [6, 16, 24, 31, 39, 41]. The objective is to protect sensitive information in all the data used throughout the process, including the *training* and *explanation* data, and to analyze the impact of privacy on explanation quality.

Information leakage about the training data has to be protected via differentially private (DP) model training [1, 7, 10], or the predictions need to be obtained through a DP mechanism [15]. Otherwise, an adversary who observes model predictions can extract private information, e.g. via membership inference attacks [35]. DP training guarantees that the black-box model explanations do not leak any further information about the training data[2]. In fact, Theorem 4.3 shows that for an (ϵ, δ)-differentially private model, the output of our explanation algorithm is $(\epsilon, \gamma\delta)$-differentially private for $\gamma < 1$ with respect to the training data. Hence, our private explanations leak strictly less information about the training dataset compared to model predictions. This implies that any training membership inference algorithm that relies on our private model explanations can be outperformed by a membership inference algorithm that relies only model predictions (See Theorem 3.2 and Theorem 4.3).

*The work was done when the author was at NUS
†The work was done when the author was at NUS
[1]Post-hoc model explanations are a class of explanation algorithms that generates explanations of the models' decisions post to the model training. Thus, such explanation algorithms are independent of the model training.

[2]This is due to the post-processing property of differential privacy. Any further computation of differentially private computations does not reveal any further information

FAccT 2022, June, 2022, Seoul, South Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533235

To protect information leakage about the explanation data via the explanation process, we propose a framework for generating differentially private model-agnostic post-hoc explanations for black-box machine learning algorithms. The main challenge that we solve is to minimize the total privacy loss, over all explanation requests, while maintaining a high explanation quality. We propose an adaptive differentially private explanation algorithm to approximate the underlying black-box decision model. We utilize the previously computed DP explanations effectively, to significantly reduce the spending of the privacy budget on new queries. We achieve this by selecting a better initialization point for our underlying optimization algorithm using the explanation history. To further reduce the total privacy loss, we design an algorithm that achieves a DP explanation by determining the number of optimization iterations required to minimize the privacy budget. This significantly reduces the privacy loss of explanation data for achieving high fidelity explanations. We show that our DP explanation algorithm amplifies the privacy of the differentially private training algorithm.

The protection of sensitive data with certified differential privacy comes at the cost of *explanation quality*. The imposed randomness of differential privacy algorithms might reduce explanation fidelity, as it increases the uncertainty of model predictions and local approximations. In the experimental section 5, we analyze the implication of privacy on explanation fidelity.

1.2 Related Work

Shokri et al. [34] show that various types of model explanations, including post-hoc methods, leak a significant amount of information about their training data. Some feature-based model explanations which depend on model parameters [3, 5, 36, 37, 41], can further increase their privacy vulnerabilities with respect to white-box inference attacks [29].

To the best of our knowledge, there is only one method, QII, that provably provides differentially private black-box post-hoc model explanations [13], protecting the explanation data. QII introduces Shapley value based model explanations, which have become a popular model explanation framework [24]; thus, their guarantees naturally translate to any Shapley-based model explanation. Datta et al. [13] bound the sensitivity of computing QII with respect to the explanation dataset, and propose an interactive differential privacy mechanism to protect the explanation dataset. Their method, however, is computationally intractable, and does not optimize the privacy budget over multiple queries. Furthermore, Datta et al. do not analyze the privacy-quality tradeoff of their framework.

Although we exclusively focus on black-box explanations, a complementary approach is to train a privacy-preserving, interpretable model. Recent work proposes constructing differentially private locally linear maps during training [20]. The model is a linear combination of multiple differentially private logistic regression algorithms. This can provide privacy-preserving model explanations for simple machine learning tasks (which can be modeled with compositions of linear functions). Furthermore, the algorithm can be significantly improved by using better DP algorithms [9], and by optimizing the spending of the privacy budget to obtain more accurate explanations. Harder et al. [20] do not provide any theoretical or empirical analysis of the fidelity of model explanations.

As a separate threat, model explanations can also magnify the risks of model reconstruction attacks [27], which allow the adversary to reconstruct model parameters. Protecting against such attacks, however, not the focus of this paper (and DP algorithms in general).

2 Problem Statement

Consider a model $f_{\mathcal{D}} : \mathbb{R}^n \to \mathcal{R}$ trained on a *training dataset* $\mathcal{D}$. We assume **black-box** access to the model f, i.e., given a point $\vec{x}$, we can obtain the predicted label for $\vec{x}$. Our goal is to generate a post-hoc *model explanation* for any given *point of interest* $\vec{z}$ (PoI) explaining the decision made by model $f_{\mathcal{D}}$.

In the black-box explanation setting, explanation algorithms do not have any access to the model parameters or hypothesis class of the underlying model $f_{\mathcal{D}}$. Thus, the only way such explanation algorithms can access the model's behavior is through its predictions. Therefore, numerous existing post-hoc explanation algorithms in the black-box setting use datasets labeled by the model $f_{\mathcal{D}}$ to approximate the model behavior and generate model explanations [2, 6, 12, 13, 16, 22, 24, 31, 32, 39, 41]. We refer to this dataset as the *explanation dataset* $\mathcal{X} = \{(x_1, f_{\mathcal{D}}(x_1)), \ldots, (x_m, f_{\mathcal{D}}(x_m))\}$. We further refer (black-box) model explanation of datapoint $\vec{z}$ as $\phi(\vec{z}, \mathcal{X}, f_{\mathcal{D}}(\mathcal{X}))$. To reduce the clutter, we denote explanation of $\vec{z}$ as $\phi(\vec{z})$ whenever it is clear from the context.

It is a very natural assumption that the explanation dataset follows the same distribution as the training dataset, and thus warrants privacy protection. Early proposals for model explanations, including LIME [31], QII [13], Int-GRAD [41] and LPR [28], sample random points around the PoI (known as perturbations) as the explanation dataset. Such explanation datasets are consists of out of training distribution datapoints. However, models' behavior can be meaningless on points outside the training data distribution, and using them as an explanation dataset might result in low fidelity explanations [11, 21, 40]. Another reason to choose in-distribution explanation data is providing model explanations that are resilient to (adversarial) noise. Recent works show that the explanations generated by LIME and SHAP are unstable and easily manipulatable, as they rely on the model behavior outside the data manifold [38]. Other recent works show that SHAP, Int-GRAD, and LPR are unstable due to off-manifold perturbations [4, 17]. Thus, most recent black-box model explanations use the training data or samples from the training data distribution, as their explanation dataset (see survey [19]). Hence, the explanation dataset contains information about actual data records rather than randomly generated (off-manifold) data points. Therefore, information leakage from the explanation dataset is a concern when explaining models' behaviour.

2.1 Privacy Risks of Black-box Explanations

We assume that an adversary queries the model $f_{\mathcal{D}}$ with a sequence of data points $\vec{z}_1, \ldots, \vec{z}_k$ and requests explanations for the model's decisions. The adversary observes explanations on the queried data points $\phi_1(\vec{z}_1), \ldots, \phi_k(\vec{z}_k)$. Allowing access to model explanations offers a new avenue for adversarial behavior. Given the obtained explanations, the adversary can reconstruct sensitive information about the explanation data $\mathcal{X} = \{(\vec{x}_1, f_{\mathcal{D}}(\vec{x}_1)), \ldots, (\vec{x}_m, f_{\mathcal{D}}(\vec{x}_m))\}$,

since the explanation algorithm performs computations on the explanation dataset to generate explanations. Moreover, the explanation algorithm relies on the labels (predictions) by the model f on the explanation dataset $f_{\mathcal{D}}(\vec{x}_j)$ to generate explanations, and $f_{\mathcal{D}}(\vec{x}_j)$ contains information about the training dataset $\mathcal{D}$. Hence, the adversary can potentially reconstruct sensitive information about the training dataset $\mathcal{D}$. Standard attack models assume access to model predictions as an avenue to access $\mathcal{D}$ [35], thus we focus on the mitigating the *additional information leakage* due to the adversary observing model explanations. Indeed, Milli et al. [27] and Shokri et al. [34] show that access to model explanations places significantly more predictive power at the hands of an adversary.

Our objective is to design and analyze algorithms that can provably protect the privacy of training and the explanation datasets against inference attacks that exploit explanations. We analyze privacy-preserving black-box explanations based on two major criteria: differential privacy loss, and utility loss of model explanations due to privacy. In other words, we want to have explanations that are both *private* and *good*.

Model explanations preserve privacy if they do not change much when we add a single datapoint to the explanation/training data. A model explanation $\phi(\cdot)$ is (ϵ, δ)-differentially private [14], with respect to the model's training data, if for any explanation dataset $\mathcal{X}$, for any sequence of PoIs (queries) $\vec{z}_1, \vec{z}_2, \cdots, \vec{z}_k$, any two neighboring training sets $\mathcal{D}$ and $\mathcal{D}'$ (i.e. ones who differ by a single point), and any subsets $S_1, \ldots, S_k \subseteq \mathbb{R}^n$, we have:

$$\Pr[\phi^1 \in S_1, \phi^2 \in S_2, \cdots, \phi^k \in S_k] \leq \mathrm{e}^{\epsilon} \cdot \Pr[\phi'^1 \in S_1, \phi'^2 \in S_2, \cdots, \phi'^k \in S_k] + \delta, \tag{1}$$

where, $\phi^i = \phi(\vec{z}_i, \mathcal{X}, f_{\mathcal{D}}(\mathcal{X}))$; $\phi'^i = \phi(\vec{z}_i, \mathcal{X}, f_{\mathcal{D}'}(\mathcal{X}))$ for all i. We can define similar guarantees with respect to the explanation set, by following (1) for two neighboring explanation datasets $\mathcal{X}$ and $\mathcal{X}'$, where $\phi_i = \phi(\vec{z}_i, \mathcal{X}, f_{\mathcal{D}}(\mathcal{X}))$, and $\phi'_i = \phi(\vec{z}_i, \mathcal{X}', f_{\mathcal{D}}(\mathcal{X}'))$.

Post-hoc explanation algorithms have no control over the model parameters or the training process. Hence, no post-hoc explanation algorithm can mitigate membership inference attacks on the training dataset that use information flow through model predictions. Therefore, our goal is to minimize any *further* information leakage of the training dataset through model explanations.

The explanation dataset $\mathcal{X}$ can be protected by making model explanation generation algorithms differentially private: these algorithms directly use the explanation dataset $\mathcal{X}$. However, the randomization introduced to achieve differential privacy negatively impacts explanation fidelity. Our goal is to maximize the accuracy of privacy-preserving randomized model explanations.

We emphasize that the differentially private training of f does not protect the explanation dataset $\mathcal{X}$, since explanation algorithms perform computations directly on the explanation dataset. Therefore, in order to protect the explanation dataset, we must ensure that explanation computations are conducted in a differentially private manner.

3 Differentially Private Model Explanations

In this work, we focus on *feature-based* model explanations: $\phi(\vec{z})$ is a vector in $\mathbb{R}^n$, where $\phi_i(\vec{z})$ measures the effect that the i-th feature has on the predicted label $f_{\mathcal{D}}(\vec{z})$. The model $f_{\mathcal{D}}$ is trained on the dataset $\mathcal{D}$ in a differentially private manner with privacy parameters $(\hat{\epsilon}, \hat{\delta})$. We omit the $\mathcal{D}$ subscript when it is clear from context. As previously mentioned, we focus on *local* linear model approximations in a region around the PoI $\vec{z}$. The objective of such explanation methods is to find a linear function ϕ, centered at a PoI $\vec{z}$, that minimizes the *local empirical model error* over the explanation dataset $\mathcal{X}$. We define the *empirical local loss* of ϕ, over $\mathcal{X}$ labeled by f, as

$$\mathcal{L}(\phi, \vec{z}, f(\mathcal{X})) \triangleq \frac{1}{|\mathcal{X}|} \sum_{\vec{x} \in \mathcal{X}} \alpha(\|\vec{x} - \vec{z}\|)(\phi^{\top} \cdot (\vec{x} - \vec{z}) - f(\vec{x}))^2, \tag{2}$$

where, $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is a weight function. For simplicity, we sometimes denote loss function as $\mathcal{L}(\phi, \vec{z})$ whenever it is clear from the context. We find a local explanation that minimizes $\mathcal{L}(\phi, \vec{z}, f(\mathcal{X}))$ within some region C. We set C to $\{\phi : \|\phi\|_2 \leq 1\}$. An *optimal* model explanation is thus one that minimizes loss around the PoI.

$$\phi^*(\vec{z}, \mathcal{X}, f(\mathcal{X})) = \arg\min_{\vec{\phi} \in C} \mathcal{L}(\phi, \vec{z}, f(\mathcal{X})) \tag{3}$$

We shorten $\phi^*(\vec{z}, \mathcal{X}, f_{\mathcal{D}}(\mathcal{X}))$ to $\phi^*(\vec{z})$ whenever it is clear from the context. The weight function α is a decreasing function in $\|\vec{z} - \vec{x}\|$: in other words, we reward explanations that correctly classify points closer to $\vec{z}$. For example, in the LIME framework, α takes a value of 0 for any points $\vec{x} \in \mathcal{X}$ that is more than a certain distance away from the PoI $\vec{z}$. Indeed, LIME, and several other model explanation frameworks can be cast in the language of our framework.

We note that any post-hoc model explanations for an $(\hat{\epsilon}, \hat{\delta})$-differentially private model f also satisfy $(\hat{\epsilon}, \hat{\delta})$ DP for the training dataset $\mathcal{D}$ due to the post-processing property of the differentially private mechanisms [14].

In order to protect the explanation dataset $\mathcal{X}$, we compute a differentially private model explanation $\phi^{Priv}(\vec{z}, f(\mathcal{X}))$ for a given PoI $\vec{z}$ by minimizing Equation (3) within the convex set C via a differentially private gradient descent algorithm described in the `DPGD-Explain()` procedure (Equation 4). The procedure uses the Gaussian mechanism [14] in each iteration to guarantee differential privacy for the explanation dataset.

In order to bound the privacy loss, we first need to bound the global sensitivity of the gradient of the loss function $\nabla\mathcal{L}(\cdot)$. Thus, we need to choose from a family of decreasing weight functions $\alpha(\cdot)$ that result in a bounded sensitivity for $\nabla\mathcal{L}(\cdot)$, required by our DP mechanisms. In Lemma 3.1 characterizes a family of weight functions offering such a guarantee.

Procedure `DPGD-Explain`(ϕ, σ, T) :

$\phi^{\{t\}} \leftarrow \phi$

for $t = 1, \ldots T - 1$ **do** :

$\xi_t \leftarrow \left(\phi^{\{t\}} - \eta(t)\left[\nabla\mathcal{L}(\phi, \vec{z}) + \mathcal{N}(0, \sigma^2 \mathrm{I})\right]\right)$

$\phi^{\{t+1\}} \leftarrow \arg\min_{\phi \in C_{2,1}} \|\phi - \xi_t\|$

Return : ϕ^T (4)

Lemma 3.1 (Conditions for Bounded Sensitivity for $\nabla\mathcal{L}(\cdot)$). *Given an explanation dataset $\mathcal{X}$ of size m, a neighboring dataset $\mathcal{X}'$,*

a PoI $\vec{z}$, and $\phi \in C$, we have that

$$\|\nabla\mathcal{L}(\phi,\vec{z},f(\mathcal{X})) - \nabla\mathcal{L}(\phi,\vec{z},f(\mathcal{X}'))\|_2 \leq (\frac{c}{m})$$

iff $\alpha(\|\vec{x}-\vec{z}\|) \leq \frac{c}{2\|\vec{x}-\vec{z}\|_2(\|\vec{x}-\vec{z}\|_2+1)}$ for every $\vec{x},\vec{z} \in \mathbb{R}^n$.

Proof. We first note that

$$\nabla\mathcal{L}(\phi,\vec{z},f(\mathcal{X})) = \frac{2}{m}\sum_{\vec{x}\in\mathcal{X}} \alpha(\|\vec{x}-\vec{z}\|)(\phi^\top(\vec{x}-\vec{z}) - f(\vec{x}))(\vec{x}-\vec{z})$$

Now, for $\alpha(\|\vec{x}-\vec{z}\|) \leq \frac{c}{2\|\vec{x}-\vec{z}\|_2(\|\vec{x}-\vec{z}\|_2+1)}$ and any $\vec{z} \in \mathcal{X}$, $\|\nabla\mathcal{L}(\phi,\vec{z},f(\mathcal{X})) - \nabla\mathcal{L}(\phi,\vec{z},f(\mathcal{X}\setminus\{\vec{z}\}))\|_2$

$$\leq \frac{2}{m}\alpha(\|\vec{z}-\vec{x}\|)\big((\phi^\top(\vec{x}-\vec{z}) - f(\vec{x})\big)$$
$$\leq \frac{2}{m}\alpha(\|\vec{z}-\vec{x}\|)\big(\|\phi\|_2\|\vec{x}-\vec{z}\|_2 - f(\vec{x})\big) \leq \frac{c}{m}$$

all inequalities are tight,which concludes the proof. □

Lemma 3.1 characterizes all weight functions $\alpha(\cdot)$ for which the sensitivity of the gradient $\nabla\mathcal{L}(\phi,\mathcal{X})$ is bounded. We define the family of desirable weight functions as $\mathcal{F}(c,\vec{z}) :=$

$$\left\{\alpha(\cdot): \begin{array}{l}\alpha(\cdot) \text{ is non-increasing and}\\ \forall\vec{x}\in\mathbb{R}^n, \alpha(\|\vec{x}-\vec{z}\|) \leq \frac{c}{2\|\vec{x}-\vec{z}\|(\|\vec{x}-\vec{z}\|+1)}\end{array}\right\}.$$

In Section 4, we propose an efficient algorithm to explain a sequence of quires using `DPGD-Explain()` procedure that optimizes privacy spending for each query using an adaptive differential private algorithm.

3.1 Evaluation Metrics

In order to protect the training and explanation datasets, we must utilize randomized differentially private mechanisms during model training and explanation generation; however, randomness introduced to achieve differential privacy might reduce the explanation quality. To quantify these tradeoffs, we utilize the following evaluation criteria.

Information Leakage of the Training Datasets Given sequence of queries, $\vec{z}_1 \ldots \vec{z}_k$, we say that post-hoc explanation algorithm is ***training data safe*** if for any underlying $(\hat{\epsilon},\hat{\delta})$-differentially private model f,

$$\Pr[\phi(\vec{z}_i,\mathcal{X},f_{\mathcal{D}}(\mathcal{X})) : i = 1,\ldots k] \leq e^{\tilde{\epsilon}}\cdot\Pr[\phi(\vec{z}_i,\mathcal{X},f_{\mathcal{D}'}(\mathcal{X})) : \forall i = 1,\ldots,k] + \tilde{\delta}, \quad (5)$$

for any $k < \infty$, where, $\tilde{\epsilon} \leq \hat{\epsilon}$ and $\tilde{\delta} \leq \hat{\delta}$ and at least one of the inequalities is strict. Intuitively, this definition simply states that even if the adversary has access to model explanations, using them *in addition* to model predictions offers no additional benefit. We formalize this claim in the following theorem (whose proof is in the Appendix-D).

Theorem 3.2 (Training Data Safe Model Explanations Leak no Further Information about the Training Data). *Let $\mathcal{A}$ be a membership inference attack on the training dataset that adaptively queries model explanations; if the model explanations are training data safe, then there exists a membership inference attack $\mathcal{A}'$ that queries model predictions, and performs strictly better than $\mathcal{A}$.*

Utilization of the Privacy Budget for Explanation Dataset Given a sequence of explanation queries, $\vec{z}_1 \ldots, \vec{z}_k$, we fix the total privacy budget for explaining all queries to (ϵ,δ). Moreover, we fix the privacy requirement for each individual query to be $(\epsilon_{\min},\delta_{\min})$ to ensure that no particular query leaks too much information. Hence, if an explanation algorithm explains $\vec{z}_1 \ldots, \vec{z}_k$ queries then we require that it does not leak too much information globally, i.e. that

$$\Pr[\phi_1 \in S_1, \phi_2 \in S_2, \cdots, \phi_k \in S_k] \leq e^{\epsilon}\cdot\Pr[\phi'_1 \in S_1, \phi'_2 \in S_2, \cdots, \phi'_k \in S_k] + \delta,$$

and that for every query $\vec{z}_j$,

$$\Pr[\phi(\vec{z}_j,\mathcal{X},f_{\mathcal{D}}(\mathcal{X}))] \leq e^{\epsilon_{min}}\cdot\Pr[\phi(\vec{z}_j,\mathcal{X},f_{\mathcal{D}}(\mathcal{X}'))] + \delta_{min}.$$

We aim to design an explanation algorithm that explains as many queries as possible given the privacy budget, while maintaining provable *quality* guarantees.

Error in Private Approximation We measure the error caused by randomness added when privately minimizing $\mathcal{L}(\cdot)$ for protecting $\mathcal{X}$ as the expected deviation of the randomized explanation from the best local approximation. More formally, we define *approximation loss* as

$$\mathcal{E}(\phi,\vec{z},f(\mathcal{X})) \triangleq \mathbb{E}\left[\mathcal{L}(\phi,\vec{z},f(\mathcal{X}))\right] - \mathcal{L}(\phi^*(\vec{z},f),f(\mathcal{X})). \quad (6)$$

4 An Adaptive Differentially Private Algorithm for Model Explanations

In this section, we design an efficient differentially private explanation algorithm that protects the privacy of the explanation dataset, explaining several queries while maintaining a low approximation loss defined in Equation (6). Given a sequence of explanation queries $\vec{z}_1,\ldots,$ a total privacy budget (ϵ,δ) and a minimum privacy requirement for each query $(\epsilon_{min},\delta_{min})$, we design an explanation algorithm that maximizes the number of queries explained without degrading the approximation error.

We sequentially explain each query in a differentially private manner, and use information from previously explained queries to optimize our computation. When computing a model explanation for a new query, it is safe to use previously released information, as it was computed in a differentially private manner. Our explanation algorithm is built upon this intuition and utilizes previous queries (generated with privacy guarantees) to reduce privacy spending. More formally, when we receive a new query $\vec{z}_{h+1}$, our explanation algorithm extracts information from previously explained queries $(\vec{z}_1,\phi^{Priv}(\vec{z}_1)),\ldots,(\vec{z}_h,\phi^{Priv}(\vec{z}_h))$ lowering the privacy budget spending for the new query $\vec{z}_{h+1}$.

We utilize two key insights to reduce the privacy budget used. First, if the underlying model f exhibits consistent behavior within a local region, then model explanations for nearby points should be similar. In other words, if we had already computed a model explanation $\phi(\vec{z})$ for a PoI $\vec{z}$ in some prior iteration, and are queried on a nearby point $\vec{v}$, we may as well release $\phi(\vec{z})$ as the explanation for $\vec{v}$ without compromising on its quality. The second insight is that we can ensure faster convergence of the `DPGD-Explain()` procedure by selecting a better initialization point for the fixed noise injection per iteration. With a better initialization point, `DPGD-Explain()` requires less iterations to converge and spend less of the privacy

budget. However, the process of selecting the initialization point should itself be privacy-preserving; hence, we need to balance the privacy budget required for finding a good initialization point, and the potential savings obtained by faster convergence. The above insights are used in our *Adaptive Private Interactive Explanation Protocol* described in Algorithm 1.

Our algorithm optimizes a convex function, which offers several indicators for a "good" starting point. Ideally, given historical queries, we would like to select a previous explanation $\phi^{Priv}(\vec{z}_j)$ minimizing $\|\phi^{Priv}(\vec{z}_j) - \phi^*(\vec{z}_{h+1})\|$. However, it is difficult to bound the sensitivity of $\min_j \|\phi^{Priv}(\vec{z}_j) - \phi^*(\vec{z}_{h+1})\|$; thus, searching for an initialization point results in a noisier selection process, requiring more undesirable privacy spending. As an alternative, we adopt a greedy approach: we search for a past query $\vec{z}_j$ minimizing $\|\nabla\mathcal{L}(\phi^{Priv}(\vec{z}_j), \vec{z}_{h+1})\|$. The sensitivity of $\|\nabla\mathcal{L}(\phi, \vec{z})\|$ is bounded by $O(\frac{1}{m})$ (as per Lemma 3.1), which allows us to efficiently search for the optimal point while incurring minor privacy spending.

4.1 Identifying Similar Points

In order to ensure that the process of finding similarly explained points has bounded sensitivity, we employ a specific type of weight function α. Consider the weight function α for $r = \frac{\sqrt{2c+1}-1}{2}$:

$$\alpha(\|\vec{x} - \vec{z}\|) = \begin{cases} 1, & \text{if } \|\vec{x} - \vec{z}\| \le r \\ \frac{c}{2\|\vec{x}-\vec{z}\|(1+\|\vec{x}-\vec{z}\|)}, & \text{else} \end{cases} \quad (7)$$

This weight function assigns equal weight to all points $\vec{x} \in \mathcal{X}$ close to the PoI $\vec{z}$, and decreases the weight for points further from $\vec{z}$ quadratically in their distance. This weight function (7) is bounded by 1, and belongs to $\mathcal{F}(c, \vec{z})$. Moreover, this weight function is *stable*: it preserves the consistency of the local explanation in a small region. The stable weight function assigns similar weights to neighboring datapoints. (See Appendix C.1, Lemma C.1). This property is highly desirable when the underlying black-box model is consistent in local regions: it implies that the behavior of our explanation algorithm is also consistent in local regions; this is captured in Theorem 4.1 (proof in Appendix D).

Theorem 4.1. (Consistent Explanations in Small Regions) *For $\alpha(\cdot) \in \mathcal{F}(c, \vec{z})$ described in (7), if $\phi^*(\vec{z}) \in \arg\min_{\phi\in C} \mathcal{L}(\phi, \vec{z}, f(\mathcal{X}))$, and $\phi^*(\vec{v}) = \arg\min_{\phi\in C} \mathcal{L}(\phi, \vec{v}, f(\mathcal{X}))$ and $\|\vec{x} - \vec{z}\| \le d \ll r$, then $\mathcal{L}(\phi^*(\vec{z}), \vec{v}, f(\mathcal{X})) - \mathcal{L}(\phi^*(\vec{v}), \vec{v}, f(\mathcal{X})) < 8d + O(d^2)$. Moreover, $|\mathcal{L}(\phi, \vec{v}, f(\mathcal{X})) - \mathcal{L}(\phi, \vec{z}, f(\mathcal{X}))| \in O(d^2)$ for all $\phi \in C$.*

If the current queried point $\vec{v}$ satisfies $\|\vec{v} - \vec{z}\| \le d < r$, and $\vec{z}$ is already explained in a differentially private manner, then Theorem 4.1 tells us that we can utilize the explanation for $\vec{z}$ in order to compute an explanation for $\vec{v}$ without spending any privacy budget, and with little approximation loss. However, this result only allows us to save the privacy budget if the query $\vec{v}$ is in the region of some previously explained query $\vec{z}$.

4.2 Reusing Prior Explanations for a Better Optimization.

Suppose that $\vec{v}$ is not within a small region of $\vec{z}$, but the model exhibits similar behavior around $\vec{z}$ and $\vec{v}$, i.e. $\|\nabla\mathcal{L}(\vec{\phi}^{Priv}(\vec{z}), \vec{v}, \mathcal{X})\|_2$ is sufficiently small. In this case, we can use the explanation of $\vec{z}$ as an initialization point $\phi^{\{0\}}$ for `DPGD-Explain()` when explaining $\vec{v}$. The selected initialization point $\phi^{\{0\}}$ guarantees faster convergence to a high quality explanation for the query $\vec{v}$, resulting in less privacy spending. Theorem 4.2 (proof in Appendix D) shows that if $\|\nabla\mathcal{L}(\vec{\phi}^{Priv}(\vec{z}), \vec{v})\|_2$ is sufficiently small, then the gradient descent method initialized with $\phi^{\{0\}} = \phi^{Priv}(\vec{z})$ converges significantly faster when computing an explanation for $\vec{v}$. We also bound the number of iterations required in order to achieve same approximation error for our explanation compared to a random starting point, as a function of $\|\nabla\mathcal{L}(\phi^{\{0\}}, \vec{v}, \mathcal{X})\|$.

Algorithm 1: Adaptive DP for Model Explanation

Input: Queries $\{\vec{z}_1, \dots\} \in \mathbb{R}^n$ arriving one by one, explanation dataset $\mathcal{X}$, privacy budget (ϵ, δ), the minimum per-query privacy loss $(\epsilon_{min}, \delta_{min})$, and the number of GD steps T;

$\mathcal{H} \leftarrow \emptyset$, $\epsilon_{spent}, \delta_{spent} \leftarrow 0$;
$\epsilon_{ite} \leftarrow \frac{\epsilon_{min}}{\sqrt{8T\log\frac{2}{\delta_{min}}}}$; // Privacy budget to spend per iteration
$\sigma_{min} = \frac{\sqrt{2\log(2.5T/\delta_{min})}}{m\cdot\epsilon_{ite}}$; // Variance needed for Gaussian mechanism
$d \leftarrow \frac{\log T}{\sqrt{T}}$; // Distance bound required according to Thm. 4.1
for $h = 1, \dots, \infty$ **do**
  if $\exists \vec{z}_j \in \mathcal{H}$ *with* $\|\vec{z}_h - \vec{z}_j\| \le d$ & `Flag`$(\vec{z}_j) = \top$ **then**
    $\phi^{Priv}(\vec{z}_h) \leftarrow \phi^{Priv}(\vec{z}_j)$; // Use a nearby point (Thm. 4.1)
    report: $\phi^{Priv}(\vec{z}_h)$; // Report explanation of $\vec{z}_h$
    $\mathcal{H}$.`append`$(\vec{z}_h : \phi^{Priv}(\vec{z}_h)$,`Flag`$(\vec{z}_h) = \perp)$
  else
    $\phi^{best}, \sigma, T' \leftarrow$ `Parameters-DPGD`$(\vec{z}_h, \mathcal{H}, \epsilon_{ite}, \sigma_{min}, \mathcal{X}, T)$;
    $\phi^{Priv}(\vec{z}_h) \leftarrow$ `DPGD-Explain`$(\phi^{best}, \sigma, T')$;
    Update $\epsilon_{spent}, \delta_{spent}$; // via the Strong Composition Theorem
    if $\epsilon_{spent} > \epsilon$ *or* $\delta_{spent} \ge \delta$ **then**
      break; // Privacy budget is exhausted
    end
    report: $\phi^{Priv}(\vec{z}_h)$;
    $\mathcal{H}$.`append`$(\vec{z}_h : \phi^{Priv}(\vec{z}_h)$,`Flag`$(\vec{z}_h) = \top)$;
  end
end

Theorem 4.2 (Number of Iterations). *Let $\phi^{\{0\}} = \phi^{Priv}(\vec{z})$ be the initialization point given to the* `DPGD-Explain()` *procedure, with $\|\nabla\mathcal{L}(\phi^{\{0\}}, \vec{v}, f(\mathcal{X}))\| = \beta$. Given the* `DPGD-Explain()` *, if $\max(\sqrt{n}\sigma, \beta) \le \frac{1}{(\log T)^a}$ for $a > \frac{1}{2}$, then for $T' = \max(\sqrt{n}\sigma, \beta)^{1-\frac{1}{2a}}T$, the approximation error satisfies $\mathcal{E}(\phi^{\{T\}}, \vec{v}, f(\mathcal{X})) \in O\left(\frac{\log T}{\sqrt{T}}\right)$.*

4.3 An Adaptive Explanation Algorithm

Using the insights of Theorem 4.1 and 4.2, we present an adaptive explanation algorithm (Algorithm 1). Algorithm 1 takes a total privacy budget (ϵ, δ), a minimum privacy required for each query $(\epsilon_{min}, \delta_{min})$ (the maximum information leakage allowed per query), an explanation dataset $\mathcal{X}$, a maximal number of iterations T (which implicitly defines the explanation quality requirement), and adaptively generates differentially private model explanations for a string of queries $\vec{z}_1, \dots$ until it exhausts its entire privacy budget (ϵ, δ).

Given the minimum privacy loss parameters $\epsilon_{min}, \delta_{min}$, we have a lower bound on the variance of Gaussian noise σ_{min} (the Gaussian Mechanism [14]), added at each iteration, where ϵ_{ite} is computed using the composition theorem. This is the minimum variance required to achieve the resultant δ parameter with at most δ_{min} spent per query.

At each query $\vec{z}_h$, Algorithm 1 first inspects whether it has already explained some other data point $\vec{z}_j$ using the differentially private gradient decent algorithm, such that $\|\vec{z}_h - \vec{z}_j\| < d$. If such $\vec{z}_j$ exists, it outputs $\phi^{Priv}(\vec{z}_j)$ for $\vec{z}_h$, spending none of the privacy budget. The produced explanation is guaranteed to be sufficiently accurate (Theorem 4.1). We note that if the explanation of $\vec{z}_j$ was not computed from scratch (used explanation for some other $\vec{z}_{j'}; j' < j$) then we cannot use the explanation of $\vec{z}_j$ for $\vec{z}_h$ because $\|\vec{z}_{j'} - \vec{z}_h\|$ might be $> d$. This is taken care by Flag().

If there is no such $\vec{z}_j$, then Parameters-DPGD selects parameters for `DPGD-Explain()`. This procedure adaptively exploits previously explained queries to ensure faster convergence with lower privacy spending and lower approximation loss using Theorem 4.2. The Parameters-DPGD procedure is explained in Algorithm 2. Given the current explanation query, it picks $\phi^{Priv}(\vec{z}_j)$ with minimum $\|\nabla\mathcal{L}(\phi^{Priv}(\vec{z}_j), \vec{z}_h)\|$ as an initialization point for the gradient descent algorithm using a differentially private exponential mechanism [14]. It then computes the number of iterations required depending on the selected initialization point according to Theorem 4.2.

Algorithm 2 spends ϵ_{ite} privacy budget to choose the starting point, however even if $\beta = \|\nabla\mathcal{L}(\phi^{Priv}(\vec{z}_j), \vec{z}_h)\| \leq 1/\log T$ we need to run only $\sqrt{\beta}$ fraction of total iterations of the differentially private gradient descent algorithm, saving at least a $\beta^{\frac{1}{4}}$ factor of the privacy budget and offering a lower approximation loss (Theorem 4.2).

Algorithm 2: Adaptive DP for Parameter Selection

Input: $\vec{z}_h \in \mathbb{R}^n$, explanation dataset $\mathcal{X}$, History $\mathcal{H}$, privacy spending for parameter selection ϵ_{para}, and the number of GD steps T, minimum variance σ_{min};

Output: Input variables for `DPGD-Explain()` for the query $\vec{z}_h$;

Procedure `Parameters-DPGD`($\vec{z}_h$, $\mathcal{H}$, ϵ_{para}, σ_{min}, $\mathcal{X}$, T)

- **if** $\mathcal{H} == \emptyset$ **then**
  - Arbitrary $\phi \in C_{2,1}$;
  - **return** ϕ, σ_{min}, T;
- **else**
  - $\phi^{best} \leftarrow \phi^{Priv}(\vec{z}_j)$ with $\Pr \propto exp\Big(-m \cdot \epsilon_{para} \cdot \frac{\|\nabla\mathcal{L}(\phi^{Priv}(\vec{z}_j), \vec{z}_h)\|}{2}\Big)$ for $\vec{z}_j \in \mathcal{H}$
  - $\beta \leftarrow \|\nabla\mathcal{L}(\phi^{best}, \vec{z}_h)\|$;
  - $\sigma \leftarrow \max\Big(\frac{\beta}{\sqrt{n}}, \sigma_{min}\Big)$;
  - $a \leftarrow \frac{\log\frac{1}{\sqrt{n}\sigma}}{\log\log T}$; // Thm. 4.2
  - **if** $a > \frac{1}{2}$ **then**
    - $T' \leftarrow (\sqrt{n}\sigma)^{1-\frac{1}{2a}} T$;
  - **else**
    - $T' \leftarrow T$;
  - **end**
- **end**
- **return** ϕ^{best}, σ, T';

4.4 Training data safeness and Utility Analysis of Our Explanations

Algorithm 1 is (ϵ, δ)-differentially private for the explanation dataset $\mathcal{X}$: all computations are privacy preserving, and respect the (ϵ, δ) privacy budget.

As the black-box model f is $(\hat{\epsilon}, \hat{\delta})$-differentially private with respect to the traning dataset $\mathcal{D}$, by the post-processing property of the differential privacy [14], explanations generated by our algorithm are $(\hat{\epsilon}, \hat{\delta})$-differentialy private with respect to the training dataset as well. However, in Theorem 4.3, we show that our explanations strictly improve the privacy guarantees for the training dataset. Therefore, our explanation algorithm is **training data safe**. Theorem 4.3 and Theorem 3.2 imply that our explanations do not offer information that can be used to design more powerful training data inference attacks.

Theorem 4.3 (Training data safe explanations). *Explanations computed by Algorithm 1 are **training data safe**. More formally, given a training dataset $\mathcal{D}$, if f is $(\hat{\epsilon}, \hat{\delta})$-differentially private w.r.t. $\mathcal{D}$, then the explanations computed by Algorithm 1 are $(\hat{\epsilon}, \gamma\hat{\delta})$-differentilly private for the training dataset, for some $\gamma < 1$. Moreover if f is trained using a non-private training process, each explanation is $(O(m\epsilon_{min}), \delta_{min})$-differentially private for the training dataset.*

Theorem 4.3 (proof in Appendix D) shows that our privacy guarantees with respect to the training set grow weaker as the size of the explanation dataset m increases; this presents a natural tradeoff between the amount of data used by the mechanisms generating model explanations, and the privacy guarantees we can offer with respect to the training data.

Theorem 4.4 (proof in Appendix D) shows that the approximation error for each explanation computed by Algorithm 1 converges at the rate of $O(\log T/\sqrt{T})$. The previous analysis for private gradient descent in [7, 33] requires T many iterations for the same σ for achieving the same level of $\mathcal{E}(\phi^{Priv}(z_j), \vec{z}_j)$. Therefore, even if $\beta = \|\nabla\mathcal{L}(\phi^{Priv}(\vec{z}_j), \vec{z}_h)\| \leq 1/\log T$ we need to run only $\sqrt{\beta}T$ iterations of the `DPGD-Explain()` whenever better initialization is found, which saves at least a $\log^{-1/4} T$ factor of the privacy budget.

Theorem 4.4. (Utility Loss) *Let $\phi^{Priv}(z_1), \ldots \phi^{Priv}(z_h)$ be the output of the Algorithm 1, then for all $j = 1, \ldots, h$; $\mathcal{E}(\phi^{Priv}(z_j), \vec{z}_j, f(\mathcal{X})) \in O\left(\frac{\log T}{\sqrt{T}}\right)$. Moreover, for lower dimensional setting ($m > n/\epsilon_{ite}$) and $T = \sqrt{m}$: $\mathcal{E}(\phi^{Priv}(z_j), \vec{z}_j, f(\mathcal{X})) \in O\left(\log m/m^{\frac{1}{4}}\right)$*

4.5 Early Termination and an Enhanced Adaptive Algorithm Implementation

During the gradient descent optimization phase, if $\|\nabla\mathcal{L}(\phi^{\{t\}}, \vec{z})\| \leq \sqrt{n}\sigma_{\min}$, then Gaussian noise starts dominating $\nabla\mathcal{L}(\phi^{\{t\}}, \vec{z})$. This domination by random Gaussian noise prevents further improvement in the loss function. Thus, the extra privacy budget spent once $\|\nabla\mathcal{L}(\phi^{\{t\}}, \vec{z})\| \leq \sqrt{n}\sigma_{min}$ does not significantly improve approximation error; rather, it starts oscillating around the optimal point, resulting in slower convergence/decrease in approximation

loss. Therefore, spending any additional privacy budget offers no benefits once $\|\mathcal{L}(\phi^{\{t\}}, \vec{z})\| < \sqrt{n}\sigma$; this observation is confirmed in Appendix E.5 Figure 6.

This motivates us to define an unrestricted version of Algorithm 1, which maximizes possible savings by the initial point via iterating for $\max\{\beta^{1-\frac{1}{2a}}T, 1\}$ times on all queries, where a is the solution to $\beta = \frac{1}{\log^a T}$. The main intuition behind this approach is that whenever Algorithm 1 finds an initialization with $\beta = \|\mathcal{L}(\phi^{\{0\}}, \vec{z})\| \ll \sqrt{n}\sigma_{min}$ then increasing the number of iterations does not result in faster convergence as Gaussian noise dominates: privacy spending in these cases offers little improvement in loss.

To implement the enhanced adaptive algorithm, we only change Line 8 to $T' \leftarrow \beta^{1-1/2a}T$ (instead of $\sqrt{n}\sigma^{1-1/2a}T$) in Algorithm 2 where $a = \frac{\log 1/\beta}{\log\log T}$. We do not maintain a similar general theoretical bound on the approximation error as in Algorithm 1; however, we empirically analyze the performance of the enhanced adaptive algorithm in Section 5 (see Figures 8 (in Appendix) and 2 for privacy spending and approximation loss).

4.6 Non-Interactive Differential Privacy Mechanisms for Model Explanation

While Algorithm 1 makes good use of its privacy budget, it will eventually exhaust it after explaining finitely many queries. At this point, a system designer would be wise to replace $\mathcal{X}$ with a new set of points, or risk information leaks. However, if this is not possible, our framework can still offer a reasonable compromise. Once the privacy budget has been exhausted, we have explained a sufficiently large number of queries, and have gathered enough information to generate explanations for new queries.

We propose a *non-interactive phase* for generating DP model explanations, which takes $\mathcal{H}$ (history of the explanation queries) — the output of Algorithm 1 — as input, and generates an explanation for new queries without spending any *additional* privacy budget. The main idea of this approach is that if $\mathcal{H}$ contains enough information about the black-box model then we can use the explanations already given to the user (adversary) and their own dataset to explain additional queries. We construct a *proxy explanation dataset* using the history $\mathcal{H}$ which contains explained datapoints and their corresponding differentially private explanation. The proxy explanation dataset is simply $\mathcal{X}' = \{\vec{z}_1, \ldots \vec{z}_h : \vec{z}_j \in \mathcal{H}\}$, i.e. the points queried by the user. Their labels are the corresponding model explanations, linear approximations of the original model: $\hat{f}(\mathcal{X}') = \{\phi^{Priv}(\vec{z}_j)^\top \cdot \vec{z}_j : j = 1, \ldots, h\}$. Given a new query $\vec{z}$, we generate a new query via a linear approximation of the black box model around $\vec{z}$ using $\mathcal{X}'$ and the corresponding differentially private approximation $\hat{f}(\mathcal{X}')$. $\phi^{Priv}(\vec{z}, \mathcal{H}) :=$

$$\arg\min_{\phi \in C} \sum_{\vec{z}_j \in \mathcal{X}'} \alpha(\|\vec{z}_j - \vec{z}\|)(\phi^\top \cdot (\vec{z}_j - \vec{z}) - \hat{f}(\vec{z}_j))^2 \qquad (8)$$

5 Empirical Analysis

We evaluate our model explanations on standard machine learning datasets: census data (ACS13)[3] and text corpus (IMDB/Amazon movie reviews) [25, 30].

ACS13: We use a scrubbed version of the dataset used in [8] containing 1,494,974 records and predict income (> 50k$ vs ≤ 50k$). We train a random forest classifier with 500 trees with maximum depth = 10, which achieves 85% training accuracy and 84% test accuracy. We use the entire datasset as an explanation dataset.

IMDB/Amazon Movie Reviews (Text dataset) [25, 30]: This dataset consists of 8,765,568 movie reviews from the Amazon review dataset along with 50,000 movie reviews from IMDB large review dataset mapped to binary vector using the top 500 words. Each movie review is labeled as either a positive (+1) or a negative (−1) review. We use the entire dataset as an explanation dataset.

Facial expression dataset [18] This dataset consists of 12,156 48×48 pixel grayscale images of faces. We train a CNN with two convolution layers with 5×5 filters followed by max-pooling and a fully connected layer, achieving training and test accuracy of 86% and 84.3%, respectively. We use this dataset to demonstrate the visualization of our explanations.

5.1 Interactive DP Model Explanation

We use our private explanation algorithm to generate private model explanations for points in all datasets. We compare the explanations we generate with existing non-private model agnostic explanation methods. In the Text dataset, we generate differentially private model explanations for 1000 randomly sampled datapoints (movie reviews) from the dataset with strong privacy parameters $\epsilon = 0.1$ and $\delta = 10^{-6}$. We present a few examples in Table 2. The explanations generated by our protocol agree with well-established non-private model agnostic explanations: using LIME [31] and MIM [39], we extract the top 5 most influential words. According to our evaluation, our protocol and MIM share 2.6 of the top 5 influential words on average, whereas our protocol and LIME share 3.9 words on average, with a variance of less than 0.3 in both cases.

How does explanation quality degrade as we make our privacy requirements more stringent? This can be visually observed for the facial expression dataset [4] by generating explanation by Algorithm 1 with different ϵ values, and a fixed $\delta = 10^{-5}$ (Figure 1); Algorithm 1 generates explanations that appear meaningful with $\epsilon \geq 0.07$ and $\delta = 10^{-5}$.

We compare the estimated approximation loss algorithm defined in Equation (6) for different privacy parameters ϵ on 1000 randomly selected datapoints from the datasets. We compute the mean ± variance value of the approximation loss for ACS13 and Text dataset for $\epsilon = 0.01$ and $\epsilon = 0.1$ with $\delta = 10^{-6}$ (Table 1). The approximation loss decreases as we relax privacy requirements (this follows from Theorem B.1). We compute better explanations for the Text dataset than the ACS13 dataset; this can be explained by their different dimensionality (the $\sqrt{n}/m$ factor in Theorem B.1). Note that $\max \mathcal{L} = 1 + \frac{1}{m}\sum_{\vec{x} \in \mathcal{X}} \frac{1}{\|\vec{x} - \vec{z}\|} > 1$. It gives the idea of the magnitude of the approximation loss of explanations computed.

[3] http://www.census.gov/programs-surveys/acs/
[4] details about the dataset can be found in Appndix E

Table 1: Summary of mean ± Variance of loss in utility for each dataset for explanation generated by Algorithm 1 (Theorem B.1).

Dataset	$\epsilon = 0.01, \delta = 10^{-6}$	$\epsilon = 0.1, \delta = 10^{-6}$
Face	$1.4 \times 10^{-2} \pm 4.3 \times 10^{-3}$	$2.2 \times 10^{-3} \pm 2.1 \times 10^{-4}$
Text	$2.7 \times 10^{-3} \pm 7.8 \times 10^{-4}$	$4.3 \times 10^{-4} \pm 9.4 \times 10^{-6}$
ACS13	$5.7 \times 10^{-3} \pm 1.3 \times 10^{-3}$	$2.6 \times 10^{-4} \pm 6.3 \times 10^{-5}$

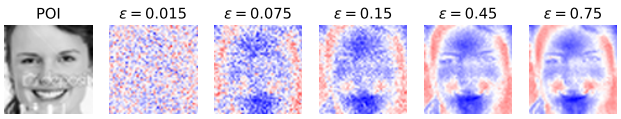


Figure 1: The effect of varying ϵ on explanation quality of Algorithm 1(In Appendix B). The color blue (red) indicates positive (negative) influence. Brighter colors indicate greater influence.

5.2 Saving Privacy Budget via Adaptive Algorithm

We generate explanations for randomly sampled 4500 datapoints from Text and ACS dataset using adaptive Algorithm 1, enhanced adaptive algorithm (described in Section 4.5), and baseline non-adaptive algorithm (Algorithm 1 in Appendix-B). After explaining 4500 data points, we sample 4500 datapoints from Text and ACS datasets and generate explanations via the non-interactive algorithm described in Section 4.

We run parallel composition [26] for our private explanations by separately using the three disjoint explanation datasets to explain three sequences of 1500 queries. We predefine the desired level of loss guarantee; in other words, both model explanations must achieve the same explanation quality, to ensure a valid comparison. We set the per-query ϵ parameter to $\epsilon = 0.01$ for the ACS13 dataset, and to $\epsilon = 0.006$ for the Text dataset. In both evaluations we set $\delta = 10^{-7}$, and set the maximal number of `DPGD-Explain()` iterations to $T = 300$. This upper bound is only reached for 94/4500 (73/4500) queries for the Text(ACS13) dataset. In all other cases, the adaptive algorithms were able to find a better initialization point. See Figure 7 in Appendix E for more details.

Interestingly, for the Text (ACS13) data, on average, over 1850 (2220) instances had a better initialization point, which failed to satisfy the $\sqrt{n}\sigma_{min} \leq \beta$ condition, and thus had to proceed with $T' = 91$ ($T' = 95$) (the minimum iterations theoretically required according to Theorem 4.2. We note that this minimum T' improves as we relax our privacy requirements (Theorem 4.2), allowing the adaptive algorithm to save more of the privacy budget. The improved performance for the ACS13 dataset may be explained by the fact that it consists of dense regions, as compared to the Text dataset.

We analyze privacy spending and loss values (Equation 2) by different algorithms in Figure 8 (in Appendix E) (ACS13) and Figure 2 (Text). The adaptive protocols spend most of their privacy budget on the initial $\sim$ 100 queries; after they generate enough information, they capitalize on it to explain the remaining queries using a smaller per-query privacy budget.

Furthermore, the adaptive protocols achieve the same explanation quality as the non-adaptive protocol, while utilizing a much smaller privacy budget (as predicted by Theorem 4.4). Figures 6 (in Appendix) and 2b show that the Adaptive and Enhanced-Adaptive protocols exhibit a similar distribution of loss values, with a significant difference in the privacy spending rate. This can be explained by the dominance of Gaussian noise over the "good" initialization point (See Appendix E).

Once Algorithm 1 explains 4500 queries, we use its output to generate additional explanations in the Non-Interactive Phase (without spending any further privacy budget) for another randomly sampled 4500 datapoints for both datasets. The non-interactive phase exhibits higher loss, but spends none of the privacy budget (Figure 2).

5.3 Tradeoffs between Training and Explanation Data Privacy

Differentially private model training often leads to noisy regions where the model behaves in an arbitrary manner. In such regions, differently labeled data points can be uniformly distributed, potentially leading to extremely convoluted decision boundaries. Therefore, explaining the behavior of such models might result in more privacy spending for protecting the explanation dataset.

To evaluate our hypotheses, we generate private explanations for models $f^1_{\mathcal{D},0.5}, f^1_{\mathcal{D},2}, f^1_{\mathcal{D},10}$ on the same randomly sampled 500 datapoints for the IMDB dataset where $f^1_{\mathcal{D},\hat{\epsilon}}$ denotes model trained on IMDB dataset with training epsilon $\hat{\epsilon}$. We use the non-adaptive Algorithm 1 described in the Appendix B, that does not utilize past information to reduce privacy spending and satisfies similar approximation loss guarantees, with $\epsilon = 0.1, \delta = 10^{-5}$ and $T = 200$. We further compute the approximation loss for each explanation described in Equation 6.

We plot the histogram of the approximation loss in Figure 4(b) in the Appendix and observe that the approximation error is significantly higher for the model $f^1_{\mathcal{D},0.5}$ where $f^1_{\mathcal{D},10}$ has nearly optimal approximation error. This reflects that $f^1_{\mathcal{D},0.5}$ is more difficult to explain compared with $f^1_{\mathcal{D},10}$, and explanations for the model $f^1_{\mathcal{D},0.5}$ does not converge to the optimal explanation in 200 iterations.

5.4 Directions to Additional Experiments

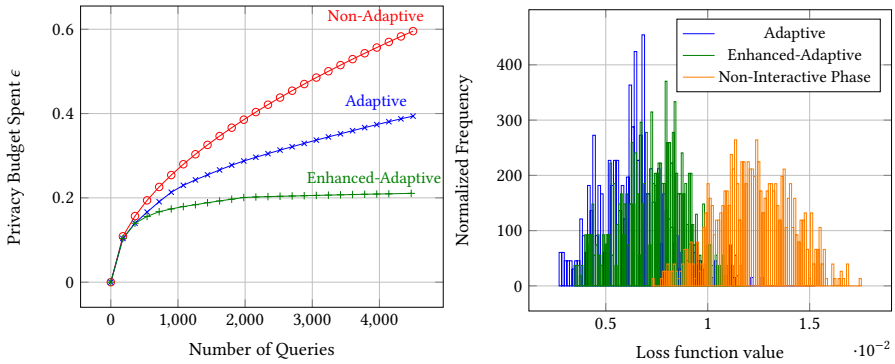
We refer the reader to Appendix E for additional experiments and details about the datasets and models. Appendix E.2 analyses the effect of differential private training via model explanations; Appendix E.3,E.4 highlights how data density and overfitting affect model explanation quality; finally, Appendix E.5 shows the effects of Gaussian noise on the convergence of the `DPGD-Explain()` procedure.

6 Conclusions

This paper provides a formal framework for achieving and analyzing differential privacy in model explanations. We highlight the possible tradeoffs between fidelity of explanations and data privacy, and show that sparse data regions might lead to poorer performance,

Table 2: Examples of influence measures generated for Text movie reviews dataset by Algorithm 1 (in Appendix-B) with $\epsilon \approx 0.1$ and $\delta = 10^{-6}$, LIME and MIM. Upwards (downwards) arrows indicate a high positive (negative) influence of a word. Moreover blue, red and green arrows correspond to words selected to Algorithm 1 (in Appendix-B)), LIME, and MIM.

	Movie Review	label
1.	... year Batman ... attempted make **well**↑↑ acted De Vito ... **bad**↓↓↓ intentions ... searching past **would**↓↓ like ... given **bad**↓↓↓ reviews...	+1
2.	...**superb**↑↑↑ performance by Natalie Portman... saying script **bad**↓ at times but I **don't**↓↓...The film **look**↑ bad, don't **good**↑ direction and **excellent**↑↑ **performances**↑↑...	+1
3.	Yeah adults may find **stupid**↑↑↑... **don't**↑↑ think really bad.... The **story**↓ aAlvin gang... across world search jewels **bad**↑↑↑ .... with... So animation **good**↓ ...	-1
4.	I never seen such **horrible**↑↑ special affects or acting... I **laughed**↓↓↓ so hard on this its just **stupid**↑ I mean the movie is so **awful**↑↑↑...	-1



(a) Total privacy budget spent by the non-adaptive (Algorithm 1 in Appendix B), adaptive (Algorithm 1) and enhanced adaptive algorithms on the same sample for the Text dataset.

(b) Normalized loss histogram for queries explained by the adaptive, enhanced adaptive and non-interactive mechanisms for the Text dataset.

Figure 2: Sub-figure 2a shows the total privacy budget spent by the non-adaptive (Algorithm 1 in Appendix B), adaptive (Algorithm 1) and enhanced adaptive algorithms on the same sample for the Text dataset. The x-axis represents the number of queries answered, and the y-axis shows the value of privacy parameter ϵ. Sub-figure 2b is the normalized loss histogram for queries explained by adaptive, enhanced adaptive and non-interactive phase.

either in terms of explanation accuracy, or in terms of required privacy budget. Such data regions often correspond to underrepresented population groups. Assessing the privacy/explainability tradeoffs for minority groups is an important direction for future exploration.

Acknowledgments

This research is supported by the National Research Foundation Singapore under its AI Singapore Program (Award Number: AISG-RP-2018-009). The authors would also like to thank Martin Strobel for very helpful technical discussions.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 308–318.

[2] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54, 1 (2018), 95–122.

[3] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 1–16.

[4] Christopher J Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. 2020. Fairwashing Explanations with Off-Manifold Detergent. *arXiv preprint arXiv:2007.09969* (2020).

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[6] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11, Jun (2010), 1803–1831.

[7] Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th Symposium on Foundations of Computer Science (FOCS)*. 464–473.

[8] Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. 2017. Plausible deniability for privacy-preserving data synthesis. *Proceedings of the VLDB Endowment* 10, 5 (2017), 481–492.

[9] Kamalika Chaudhuri and Claire Monteleoni. 2009. Privacy-preserving logistic regression. In *Advances in neural information processing systems*. 289–296.

[10] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12, Mar (2011), 1069–1109.

[11] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. 2020. True to the Model or True to the Data? *CoRR* abs/2006.16234 (2020). arXiv:2006.16234

[12] Amit Datta, Anupam Datta, Ariel D Procaccia, and Yair Zick. 2015. Influence in classification via cooperative game theory. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*. 511–517.

[13] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Proceedings of the 37th IEEE Symposium on Security and Privacy (Oakland)*. IEEE, 598–617.

[14] Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*. 1–12.

[15] Cynthia Dwork and Vitaly Feldman. 2018. Privacy-preserving prediction. *arXiv preprint arXiv:1803.10266* (2018).

[16] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3429–3437.

[17] Christopher Frye, Damien de Mijolla, Laurence Cowton, Megan Stanley, and Ilya Feige. 2020. Shapley-based explainability on the data manifold. *arXiv preprint arXiv:2006.01272* (2020).

[18] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks* 64 (2015), 59–63.

[19] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5, Article 93 (Aug. 2018), 93:1–93:42 pages.

[20] Frederik Harder, Matthias Bauer, and Mijung Park. 2020. Interpretable and Differentially Private Predictions. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*. 4083–4090.

[21] Dominik Janzing, Lenon Minorics, and Patrick Bloebaum. 2020. Feature relevance quantification in explainable AI: A causal problem. In *Proceedings of the 33rd International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2907–2916.

[22] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

[23] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490

[24] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*. 4765–4774.

[25] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT)*. 142–150.

[26] Frank D McSherry. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM Conference on Management of Data (SIGMOD)*. 19–30.

[27] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. 2019. Model Reconstruction from Model Explanations. In *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency (FAT*)*. 1–9.

[28] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* 65 (2017), 211–222.

[29] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Proceedings of the 40th IEEE Symposium on Security and Privacy (Oakland)*. 739–753.

[30] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fined-grained aspects. In *Proceedings of the 15th Conference on Empirical Methods on Natural Language Processing (EMNLP)*. 188––197.

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (KDD)*. 1135–1144.

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*. 1527–1535.

[33] Ohad Shamir and Tong Zhang. 2013. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*. 71–79.

[34] Reza Shokri, Martin Strobel, and Yair Zick. 2021. Privacy Risks of Model Explanations. In *Proceedings of the 4th AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES)*.

[35] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of the 38th IEEE Symposium on Security and Privacy (Oakland)*. 3–18.

[36] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 3145–3153.

[37] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[38] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[39] Jakub Sliwinski, Martin Strobel, and Yair Zick. 2019. A Characterization of Monotone Influence Measures for Data Classification. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*. 718–725.

[40] Mukund Sundararajan and Amir Najmi. 2020. The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 9269–9278.

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 3319–3328.