# (Poster) Safe and Sound: Approximately-Optimal Black-box Model Explanations with Differential Privacy (Extended Abstract)

Neel Patel, Reza Shokri, Yair Zick
National University of Singapore
{neel, reza, zick}@comp.nus.edu.sg

Machine learning models are currently applied in a variety of high-stakes domains, e.g. providing predictive healthcare analytics. These domains require high prediction accuracy over high-dimensional data, and as a result, adopt increasingly complex ML models. While simple rule-based systems and linear models are easily understood, modern ML models — which utilize complex mathematical structures, such as deep neural networks — are hard to interpret. What makes the problem even more challenging is that these models are usually used as black-box algorithms, where the model only outputs the decision without providing detailed information on its intermediate computations for producing the output (decision).

The implementation of black-box algorithms in high-stakes domains affects trust, raises concerns with respect to the model's reliability and fairness, and hinders model adoption [1], [2]. Recent years have seen a widespread call for *algorithmic transparency*. Broadly speaking, algorithmic transparency methods offer stakeholders additional information about the algorithmic decision-making process, to facilitate their understanding of the model. In this work, we focus on *feature-based* explanations, where the algorithm computes the influence of various features of the point of interest on the model's decision [3]–[9]. However, offering additional information can result in significant privacy risks, potentially exposing the underlying model [10] or the model's training data [11].

The analysis of the privacy risks of explanation algorithms is largely limited to these works, and their overall privacy implications are widely unexplored. Notably, there has been little work on designing *provably sound*, as well as *differentially private*, model explanations for black-box models. This calls for a rigorous approach towards modeling the privacy risks of model explanations, and the design of practical differentially-private model explanations with provable explanation accuracy.

We focus on protecting the privacy of *model-agnostic feature-based explanations for black-box models*, which do not have access to the model parameters, and make no assumptions on model structure. These techniques assume access to a labeled dataset, sampled from the same underlying distribution as the model's training set, which we refer to as the *explanation dataset*. The explanation algorithms locally approximate the target model in the vicinity of the point of interest, using the explanation dataset, in order to measure the features' influence on the model [4]–[6], [8], [12]. This popular approach can leak sensitive information about the explanation dataset through model explanations. This leakage is independent of privacy loss with respect to training data.

Almost all existing model explanations do not offer any proven privacy guarantees, and given their large/unbounded sensitivity, achieving differential privacy can result in an extremely low accuracy. As an exception, QII could be randomized to satisfy differential privacy to protect the explanation dataset [5]. However, there is no theoretical analysis of the composition of the privacy loss of QII over a sequence of queries, nor there is a bound on the utility loss of the algorithm due to the randomness of the differential privacy mechanism.

***Our Contributions.*** We propose provably sound, model-agnostic, and differentially private algorithms for computing model explanations. Our explanations are sound in the sense that they are provably similar to some standard model explanations, such as LIME [12]. Our methodology can be adapted to any explanation method that relies on generating accurate local models around the data point one tries to explain.

We first design a baseline *interactive* differentially private mechanism for model explanation (class of algorithms that require accessing the sensitive dataset for each query).This method generates a feature-based explanation — a scalar value for each data feature — in a differentially private manner, using an *explanation dataset*. Our basic DP model explanation generates an explanation for queried data points by optimizing a convex function using a differentially private gradient descent algorithm. We design our main algorithm on top of this baseline. The main challenge, that we solve in this paper, is to optimize the composed privacy loss of the model explanation over all queries, with low explanation error.

Our main theoretical contribution is designing an adaptive differentially private model explanation with bounded utility loss. The algorithm utilizes past information (DP explanations) effectively, and saves privacy spending significantly for model explanations on new queries. Our key idea is to achieve greater privacy saving by selecting a better initial point for the gradient descent algorithm for each query using past information. We improve upon bounds in [13] on the convergence of the DP gradient descent (under a minor assumption), offering faster convergence rates that depend on the initial point. Thus, by carefully selecting an initialization point for the gradient de-

scent algorithm, we obtain significant privacy savings, without compromising on explanation accuracy.

We show that when the initial point approaches the optimal point, the DP gradient descent algorithm oscillates around the optimal point (due to the algorithm's inherent noise), spending the privacy budget in vain. This insight leads to an enhanced adaptive algorithm, which offers far better privacy savings for only a minor accuracy loss.

All approaches described above are interactive. Despite their efficiency in controlling the privacy risk, a sufficiently large number of queries will eventually lead to unacceptable information leaks (i.e., using all the allocated differential privacy budget). To counteract this effect, we propose switching to a *non-interactive* explanation phase, once a sufficient amount of information has been released using the given privacy budget. In this phase, new explanation queries no longer use the private dataset, so do not impose any privacy loss; rather, they generate explanations using only past explanation queries. In other words, we continue to adaptively provide model explanations, but instead of using previous explanations to initialize the gradient descent algorithm, we use them directly to linearly approximate the model on new queries. We provide outline of our algorithms and statements of our theoretical results in the poster.

We extensively test our approaches on micro-data as well as text data, on classification machine learning models. Our adaptive algorithm utilizes the privacy budget efficiently and outperforms the non-adaptive approach in terms of both privacy and approximation accuracy. The adaptive algorithm spends a similar amount of its privacy budget on initial queries as the non-adaptive algorithm; however, it quickly begins to effectively utilize past information to answer new queries. For example, the adaptive algorithm can answer 4,500 queries using $52\%$ of the privacy budget required by the non-adaptive algorithm on the ACS13 dataset; the enhanced adaptive algorithm can answer the same number of queries using $21\%$ of the budget (neither compromise on explanation accuracy). This gap increases as the number of queries increases. We provide some basic experimental results for ACS13 dataset in Figure 1.

We empirically investigate the convergence of the adaptive protocol when starting from a good initialization point, and verify the foundations of our algorithm. The differntially private gradient descent updates tend to throw the iterative process away from the optimal point, followed by repeatedly converging back to similar points. We show that the enhanced adaptive protocol effectively does away with this step, which results in significant privacy savings, and only a minor loss in approximation accuracy.

We investigate the effect of data density on our approach. We show that, as expected, the adaptive algorithm works more efficiently in dense regions, which allow it to gather information more effectively as compared to sparse regions. We also analyze the effect of overfitting on our model explanations. We show that the overfitted models require slightly more privacy spending due to complex decision boundaries, which offer less opportunity for adaptive savings. Finally, we show that our algorithm can still provide accurate explanations in the non-interactive phase. We could not include all results of experiments in this abstract due to space constraint.

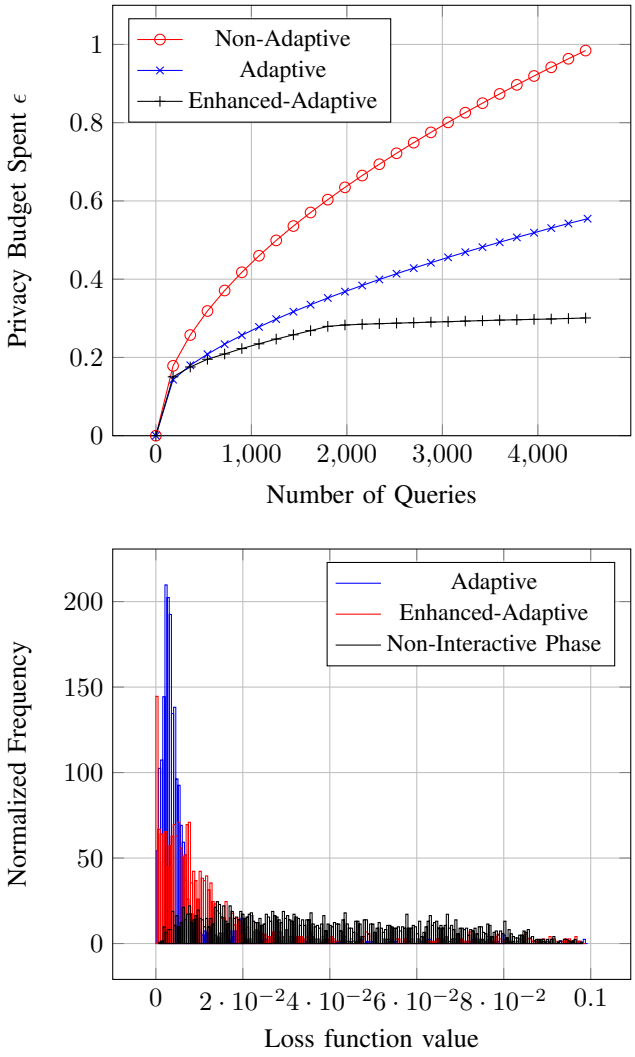


Fig. 1: Comparison of the total privacy budget spent after each query and histogram of explanation loss by different algorithms for ACS13.

TABLE I: Examples of influence measures generated for Text movie reviews dataset by DP-Explanation $\epsilon \approx 0.1$ and $\delta = 10^{-6}$, LIME and MIM. Upwards (downwards) arrows indicate a high positive (negative) influence of a word. Blue, red and green arrows correspond to words selected to DP-Explanation, LIME, and MIM.

| **Movie Review** |
|---|
| 1. ...**superb**↑↑↑ performance by Natalie Portman... saying script **bad**↓ at times but I **don't**↓↓... **good**↑ direction and **excellent**↑↑ **performances**↑↑...(Label:+1) |
| 2. 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**↑↑↑... (Lable:-1) |

## REFERENCES


[1] Z. C. Lipton, "The mythos of model interpretability," *CoRR*, vol. abs/1606.03490, 2016.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 93:1–93:42, Aug. 2018.

[3] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "A unified view of gradient-based attribution methods for deep neural networks," in *NIPS 2017-Workshop on Interpreting, Explaining and Visualizing Deep Learning*, Long Beach, CA, USA, 2017.

[4] A. Datta, A. Datta, A. D. Procaccia, and Y. Zick, "Influence in classification via cooperative game theory," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[5] A. Datta, S. Sen, and Y. Zick, "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, 2016, pp. 598–617.

[6] J. Sliwinski, M. Strobel, and Y. Zick, "A characterization of monotone influence measures for data classification," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 718–725.

[7] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.

[8] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[9] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018, pp. 1–16.

[10] S. Milli, L. Schmidt, A. D. Dragan, and M. Hardt, "Model reconstruction from model explanations," in *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019, pp. 1–9.

[11] R. Shokri, M. Strobel, and Y. Zick, "Privacy risks of explaining machine learning models," *CoRR*, vol. abs/1907.00164, 2019.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "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)*, 2016, pp. 1135–1144.

[13] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013, pp. 71–79.