A Unified Analysis of Label Inference Attacks

Robert Busa-Fekete
Google
busarobi@google.com

Travis Dick
Google
tdick@google.com

Claudio Gentile
Google
cgentile@google.com

Andres Muñoz Medina
Google
ammedina@google.com

Marika Swanberg[*]
Boston University
marikas@bu.edu

Abstract

Randomized response and label aggregation are two common ways of sharing sensitive label information in a private way. In spite of their popularity in the privacy literature, there is a lack of consensus on how to compare the privacy properties of these two different mechanisms. In this work, we investigate the privacy risk of sharing label information for these privacy enhancing technologies through the lens of label reconstruction advantage measures. A reconstruction advantage measure quantifies the increase in an attacker's ability to infer the true label of an unlabeled example when provided with a *private* version of the labels in a dataset (e.g., averages of labels from different users or noisy labels output by randomized response), compared to an attacker that only observes the feature vectors, but may have prior knowledge of the correlation between features and labels. We extend the Expected Attack Utility (EAU) and Advantage of previous work to mechanisms that involve aggregation of labels across different examples. We theoretically quantify this measure for Randomized Response and random aggregates under various correlation assumptions with public features, and then empirically corroborate these findings by quantifying EAU on real-world data. To the best of our knowledge, these are the first experiments where randomized response and label proportions are placed on the same privacy footing. We finally point out that simple modifications to the random aggregate approach can provide extra DP-like protection.

1 Introduction

With the ubiquity of data collection, processing and sharing, users, companies, and regulators are becoming increasingly aware of the privacy risks associated with information sharing. This has led companies and governments to restrict and regulate sharing of user information across different entities. Some examples of these initiatives are the Digital Markets Act (DMA) [1], third-party cookie deprecation in Chrome [2] and intelligent tracking protection (ITP) in Safari. On the other hand, data sharing provides an undeniable utility to individuals and society at large. Indeed, it allows for faster advances in science and improves the economy and individuals' daily life through automation driven by machine learning models trained on (potentially sensitive) data. For this reason, most privacy initiatives do not fully disallow the sharing of information, but instead allow disclosure of user information as long as it is processed using a so-called Privacy Enhancing Technology (PET). While the technical specification of privacy is usually application dependent, the large majority of PETs are powered by two simple data processing techniques: data aggregation and data noising.

[*] Work done during an internship at Google.

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

One of the key factors in choosing which PET to use is the privacy-accuracy trade-off: *What accuracy does one get for a given level of privacy protection?* While these curves can be parameterized for each individual PET (ϵ for differential privacy, k for k-anonymity), seldom do people consider a metric that can put the privacy-utility trade-off of different PETs under the same footing. The lack of comparison metrics across PETs makes it hard for regulators and decision makers to evaluate the protections of different proposals for information sharing. Recent work in the privacy community has tried to solve this issue by considering empirical metrics such as inference attacks [12], reconstruction attacks [4], re-identification attacks [8] and label inference attacks [7, 14]. However the majority of this work, with the exception of [8] has focused on providing a better understanding of the protections provided by differential privacy only.

In this work, we extend results from [14] to quantify risks associated with PETs based on aggregation. While it is intuitive that data aggregation should provide some form of privacy protection, there has been very little work to try to understand how this protection compares to that provided by PETs such as differential privacy. Our work allows for such comparison and therefore provides decision makers with actionable tools for selecting the optimal privacy-utility trade-off.

We will focus on the simplest possible scenario for information sharing: the case where a user wishes to disclose a single bit of information (for instance a binary label in a classification problem). While this problem is extremely simple to present, we shall see that it already highlights a lot of the difficulties in providing a measure of privacy that is meaningful for both noise-based and aggregate-based tools. Moreover, understanding this simple version of the problem already has regulatory implications for industry. For example, for the design of conversion reporting APIs by Apple's Safari and Google's Chrome browsers, which can share conversion information by using randomized response or by aggregating conversions across random sets of users.

We show that the complexity of understanding the potential for privacy leaks requires also handling the more complicated scenario of known attributes that can be correlated with unknown sensitive labels. Similar to [14], we deal with this problem by measuring the advantage an attacker may have in reconstructing sensitive information after observing the output of a PET, compared to an attacker that knows these correlations only.

This privacy scenario is commonly observed in practice. For instance, with Chrome's proposed conversion reporting API, the event of a user converting after clicking on an online ad — buying a product, signing up for a newsletter, installing an app, etc. — or not is considered sensitive and therefore is reported only with some noise. However, once reported, ad tech providers can use features associated with an ad click (impression information, publisher information, ...) to train models that can predict future conversions.

As [14] shows, having a baseline is important for this setting: for example, if the labels are perfectly correlated with the features, and the features are public, then reconstruction need not indicate a failure of the PET we use to protect the labels. On the other hand, if the labels are independent of the features, a successful reconstruction reveals much more about the strength of this PET.

Our paper is organized as follows:

1. We introduce our setup and define the privacy problem that we are trying to solve.
2. We discuss the methods we consider for sharing user information: randomized response and aggregates on random partitions of data.
3. We introduce the *attack advantage measure* as a way to quantify the amount of leakage associated with these mechanisms. This is an extension of the Expected Attack Utility (EAU) and advantage of [14] to handle aggregate information as well.
4. We calculate the Advantage Measure of randomized response and random aggregates under different correlation assumptions with public data.
5. We conduct experiments that measure both the advantage measure and the utility of learning a model through either aggregate or noisy data. By using our measure, we present the first comparison of utility at the same privacy level for two previously incomparable PETs.
6. Finally, we discuss limitations of the framework and show that there are particular attacks that differential privacy can protect against that aggregates may not. On the other hand, we show that a very simple modification to the aggregate method can provide these protection as well.

2 Preliminaries

Let $\mathcal{X}$ denote a feature (or instance) space and $\mathcal{Y} = \{0, 1\}$ be a binary[2] label space. We assume the existence of a joint distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, encoding the correlation between the input features and the labels. We denote by $p = \Pr_{(x,y)\sim\mathcal{D}}(y = 1)$ the probability of drawing a sample $(x, y) \in \mathcal{X} \times \mathcal{Y}$ from $\mathcal{D}$ with label $y = 1$. For a natural number n, let $[n] = \{i \in \mathbb{N} \colon i \leq n\}$.

We define a dataset $S = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$ as a sequence of pairs (x_i, y_i), each one drawn i.i.d. from $\mathcal{D}$. We use $\mathbf{x} = (x_1, \ldots, x_m)$ to denote the features in S and $\mathbf{y} = (y_1, \ldots, y_m)$ to denote the labels.

Definition 2.1. Given $\epsilon, \delta \geq 0$, We say that a (randomized) algorithm A that takes as input S is (ϵ, δ)-Label Differentially Private (Label DP) if for any two datasets S and S' that differ in the label of a single sample we have $\Pr(A(S) \in B) \leq e^{\epsilon} \Pr(A(S') \in B) + \delta$, where B is any subset of the output space of A. When $\delta = 0$ the algorithm A is said to be ϵ-Label DP.

Randomized Response (RR) is a classical [13] way of achieving ϵ-Label DP. In the binary classification case, RR with privacy parameter $\pi = 1/(1 + e^{\epsilon})$ simply works by randomly flipping each label y_j in the dataset with independent probability π before revealing it to the learning algorithm. RR is especially appealing from practical point of view, since privatized data with label flipping can be handled by many prominent learning algorithms as is, with some tuning of their hyper-parameters. Often the theoretical guarantees of these learners in terms of sample complexity are only deteriorated by some constant that depends on the label noise level, see, for example [11].

A completely different label privacy criterion is one based on (random) label aggregation, sometimes called Learning from Label Proportions (LLP), whereby the dataset S gets organized in (i.i.d. random) *bags* of a given size k, $S = \langle (x_{11}, y_{11}), \ldots, (x_{1k}, y_{1k}), \ldots, (x_{n1}, y_{n1}), \ldots, (x_{nk}, y_{nk}) \rangle$, where examples are grouped together in groups of k, and only the fraction of positive labels in each group are revealed to the learning algorithm. In other words, the learning algorithm has access to S via a collection $\{(\mathcal{B}_i, \alpha_i),\ i \in [n]\}$ of n labeled bags of size k, with $m = nk$, where $\mathcal{B}_i = \{x_{ij} \colon j \in [k]\}$, $\alpha_i = \frac{1}{k}\sum_{j=1}^{k} y_{ij}$ is the label proportion of the i-th bag, and all the involved samples (x_{ij}, y_{ij}) are drawn i.i.d. from $\mathcal{D}$. Thus, the learner receives information about the m labels y_{ij} of the m instances x_{ij} from dataset S only in the aggregate form determined by the n label proportions α_i associated with the n labeled bags $(\mathcal{B}_i, \alpha_i)$. Note, however, that the feature vectors x_{ij} are individually observed.

A simple and very well known method for learning from aggregate labels is the one the authors of [15] call Empirical Proportion Risk Minimization. In fact, different versions of this algorithm are discussed in the literature without a clear reference to its origin. In [6], the authors simply call this algorithm the Proportion Matching algorithm (PROPMATCH), and we shall adopt their terminology here.

Given a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, a hypothesis set of functions $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, mapping $\mathcal{X}$ to a (convex) prediction space $\widehat{\mathcal{Y}} \subseteq \mathbb{R}$, and a collection $\{(\mathcal{B}_i, \alpha_i),\ i \in [n]\}$ of n labeled bags of size k, PROPMATCH minimizes the empirical *proportion matching loss*, i.e., it solves the following optimization problem: $\min_{h\in\mathcal{H}} \sum_{i=1}^{n} \ell\left(\frac{1}{k}\sum_{j=1}^{k} h(x_{ij}), \alpha_i\right)$.

3 Quantifying Privacy Loss via Reconstruction Advantage

In this section, we extend the advantage definition introduced by Wu et al. [14] to the aggregate setting and give analytical bounds on the reconstruction advantage for this setting.

The reconstruction advantage is grounded in the following natural privacy question: *How much does releasing the output of a PET increase the risk of label reconstruction compared to not releasing any private data?* In particular, we are interested in the reconstruction risk to an average user.

In order to unify the study of aggregation and noise-based PETs, for the rest of the section we fix $k > 0$ and model PETs as (possibly randomized) functions $\mathcal{M} : (\mathcal{X} \times \mathcal{Y})^k \to \mathcal{Z}$. These functions

[2] For ease of presentation, we restrict here to binary classification tasks, but the material contained in this paper can readily be lifted to more general classification or regression settings.

map a collection of k labeled examples to a privacy protected representation in the domain $\mathcal{Z}$.[3] The domain $\mathcal{Z}$ depends on the PET but, throughout the paper, the domain will be clear from context. For example, LLP corresponds to the function $\mathcal{M}_{\text{LLP}}(\mathbf{x}, \mathbf{y}) = \alpha$, where $\mathbf{x} \in \mathcal{X}^k$ represents the known feature vectors and $\alpha = \frac{1}{k}\sum_{i=1}^k y_i$ is the proportion of positive examples in the bag. Similarly, for a fixed $\epsilon > 0$, RR corresponds to the function $\mathcal{M}_{\text{RR}}(\mathbf{x}, \mathbf{y}) = \tilde{\mathbf{y}}$ where $\tilde{y}_i = 1 - y_i$ with probability $\pi = 1/(1 + e^{\epsilon})$ and equal to y_i with probability $1 - \pi$ independently for each $i \in [k]$. Finally, it will be helpful to consider a PET that reveals no label information at all: $\mathcal{M}_{\perp}(\mathbf{x}, \mathbf{y}) = \perp$.

The goal of the PETs we are considering is to hide individual's labels. To measure how well they preserve this privacy, we consider a label inference adversary whose goal is to predict the labels $\mathbf{y}$ given access to the features, $\mathbf{x}$, together with the output of a PET, $\mathcal{M}(\mathbf{x}, \mathbf{y})$. We model adversaries as functions $\mathcal{A} : \mathcal{X}^k \times \mathcal{Z} \to \mathcal{Y}^k$ that map the features and the output of a PET to a vector of predicted labels, one for each example. To measure the efficacy of an adversary, we define the *expected attack utility* of adversary $\mathcal{A}$ using information from PET $\mathcal{M}$ on a collection of k examples drawn iid from a distribution $\mathcal{D}$ as follows:

$$\text{EAU}_k(\mathcal{A}, \mathcal{M}, \mathcal{D}) = \underset{\substack{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^k \\ i\sim\text{Uniform}([k])}}{\mathbb{P}} \big(\mathcal{A}(\mathbf{x}, \mathcal{M}(\mathbf{x}, \mathbf{y}))_i = y_i\big) \,.$$

In other words, the expected attack utility of the adversary is the probability that they correctly guess the label of a random chosen example when provided the features and the output of $\mathcal{M}$. Equivalently, this is the expected fraction of the k examples that the adversary predicts the correct label for. The adversary's success rate may depend on the distribution over features and labels. For example, if labels are entirely determined by features, then our metric should reflect that privatized labels (for any mechanism) reveal no additional information about the true labels. To control for the information that features inherently reveal about labels, we always assume that the adversary has knowledge of the data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.

In order to measure the *increase* in risk incurred by releasing the output of a PET, we consider the expected attack utility of an optimal adversary in two scenarios: one in which the adversary gets the features, $\mathbf{x}$, together with the output of the PET, $\mathcal{M}(\mathbf{x}, \mathbf{y})$, and an alternate setting where the adversary gets only the features (which is equivalent to using $\mathcal{M}_{\perp}$). We call the difference in expected attack utility between the informed and uninformed adversary the *attack advantage*. Intuitively, the attack advantage measures the label reidentification risk that can be attributed to the PET rather than to correlations between the features $\mathbf{x}$ and labels $\mathbf{y}$ which are inherent in the distribution $\mathcal{D}$. While having a low reconstruction advantage does not necessarily guarantee that the mechanism poses no risks at all, having a high advantage is a clear sign that the published information increases the risk of reconstruction.

Definition 3.1. The *attack advantage* of a PET $\mathcal{M}$ for a set of k examples drawn from a data distribution $\mathcal{D}$ is defined by

$$\text{Adv}_k(\mathcal{M}, \mathcal{D}) = \sup_{\mathcal{A}_{\text{informed}}} \text{EAU}_k(\mathcal{A}_{\text{informed}}, \mathcal{M}, \mathcal{D}) - \sup_{\mathcal{A}_{\text{uninformed}}} \text{EAU}_k(\mathcal{A}_{\text{uninformed}}, \mathcal{M}_{\perp}, \mathcal{D}) \,. \quad (1)$$

Note that this definition is slightly unusual for RR, since RR is generally described in terms of its behavior on a single example (x, y), rather than a collection of k i.i.d. samples. This "type mismatch" seems unavoidable for any privacy measure that is applicable both to aggregation- and local DP-based PETs. However, in Section 3.2 we will see that for RR the attack advantage is independent of the number of examples k, confirming our intuition that the aggregation size k should play no role in the risk posed by RR.

First, we bound the EAU_k achievable by any adversary for a given PET $\mathcal{M}$, data distribution $\mathcal{D}$, and number of examples, k. The bound is tight and we derive the prediction rule for an optimal attacker that achieves the bound with equality.

Lemma 3.2. *The following expected attack utility bound holds for any PET $\mathcal{M}$, data distribution $\mathcal{D}$, number of examples k and adversary $\mathcal{A}$ that observes the output of $\mathcal{M}$*

$$\text{EAU}_k(\mathcal{A}, \mathcal{M}, \mathcal{D}) \le 1 - \underset{\substack{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^k \\ i\sim\textit{Uniform([k])}}}{\mathbb{E}} \big[\min\{\mathbb{P}(y_i = 1 \mid \mathbf{x}, \mathcal{M}(\mathbf{x}, \mathbf{y})), \mathbb{P}(y_i = 0 \mid \mathbf{x}, \mathcal{M}(\mathbf{x}, \mathbf{y}))\}\big].$$

[3] Note that we allow the PET to have access to the features $\mathbf{x}$ in addition to the labels $\mathbf{y}$. None of the PETs we study use $\mathbf{x}$, but this would allow PETs that, for example, output multiple aggregations on subsets of the k examples. All of our general theory can handle PETs that also depend on the features.

Moreover, the following optimal adversary $\mathcal{A}^*$ achieves this upper bound with equality:

$$\mathcal{A}^*(\mathbf{x}, \mathcal{M}(\mathbf{x}, \mathbf{y}))_i := \begin{cases} 1 & \textit{if } \mathbb{P}(y_i = 1 \mid \mathbf{x}, \mathcal{M}(\mathbf{x}, \mathbf{y})) \geq 1/2 \\ 0 & \textit{otherwise.} \end{cases}$$

The proof of the previous lemma can be found in Appendix A.1.

Next, we leverage Lemma 3.2 to bound the advantage for any PET $\mathcal{M}$, data distribution $\mathcal{D}$, and number of examples k.

Theorem 3.3. *For any PET $\mathcal{M}$, any data distribution $\mathcal{D}$, and number of examples k, the attack advantage is given by*

$$\begin{aligned}\mathrm{Adv}_k(\mathcal{M}, \mathcal{D}) = & \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \big[\min\{\eta(x), 1 - \eta(x)\}\big] \\ & - \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\mathbf{y}) \sim \mathcal{D}^k \\ i \sim \mathit{Uniform}([k])}} \Big[\min\big\{\mathbb{P}(y_i = 1 \mid \mathbf{x}, \mathcal{M}(\mathbf{x}, \mathbf{y})), \mathbb{P}(y_i = 0 \mid \mathbf{x}, \mathcal{M}(\mathbf{x}, \mathbf{y}))\big\}\Big]\end{aligned}$$

where $\eta : x \mapsto \mathbb{P}(y = 1 \mid x)$ is the Bayes optimal predictor.

The proof of Theorem 3.3 is in Appendix A.2.

In the following sections, we prove bounds on the attack advantage for LLP and RR that are more directly interpretable than Theorem 3.3.

3.1 Bounding Attack Advantage for Learning from Label Proportions

In this section we provide explicit bounds on the attack advantage for LLP. Recall that LLP corresponds to the function $\mathcal{M}_{\mathrm{LLP}}(\mathbf{x}, \mathbf{y}) = \alpha := \frac{1}{k}\sum_{i=1}^k y_i$.

We begin by studying the attack advantage for LLP when the labels are independent of the features.

Theorem 3.4. *Fix a data distribution $\mathcal{D}$, let $p = \mathbb{P}_{(x,y)\sim\mathcal{D}}(y = 1)$, and fix an arbitrary threshold $\beta \in [0, 1/2]$. If labels are independent of features (i.e., $\mathcal{D}$ is a product of distributions over $\mathcal{X}$ and $\mathcal{Y}$), then for all bag sizes $k \geq 1$ we have:*

$$\mathrm{Adv}_k(\mathcal{M}_{LLP}, \mathcal{D}) = \min\{p, 1 - p\} - \mathop{\mathbb{E}}_{\alpha}[\min\{\alpha, 1 - \alpha\}] \leq \begin{cases} \sqrt{\frac{p(1-p)}{k}} & \textit{if } p \in [0, 1] \\ e^{-\Omega(\beta^2 k)} & \textit{if } |p - 1/2| \geq \beta, \end{cases}$$

where the Ω notation hides constants independent of β and k.

A couple of remarks are in order. First, observe that the advantage $\mathrm{Adv}_k(\mathcal{M}_{\mathrm{LLP}}, \mathcal{D})$ is always non-negative, as can be easily derived by noting that $\mathbb{E}[\alpha] = p$ and then applying Jensen's inequality to the concave function $x \mapsto \min\{x, 1 - x\}$, for $x \in [0, 1]$. Second, despite being non-negative, Theorem 6.4 also proves the desirable property that $\mathrm{Adv}(\mathcal{M}_{\mathrm{LLP}}, \mathcal{D})$ goes to zero as the bag size k increases. The convergence rate is in general of the form $1/\sqrt{k}$, but it becomes *negative exponential* in k when p is bounded away from 1/2.

We now again consider the more general scenario and obtain a more concise bound for Theorem 3.3 when $\mathcal{M}$ is $\mathcal{M}_{\mathrm{LLP}}$. Let then $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Recall that, in this more general case, the distribution of random variable $k\alpha = \sum_{j=1}^k y_j$ conditioned on $\mathbf{x} = (x_1, \ldots, x_k)$ is Poisson Binomial (PBin) with parameters $\{\eta(x_j)\}_{j=1}^k$, that is, the distribution of the sum of k independent Bernoulli random variables y_j, each with its own bias $\eta(x_j)$.

Theorem 3.5. *Let $\mathcal{D}$ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and let $p = \mathbb{E}[\eta(x)]$. Then, for all bag sizes $k \geq 2$ we have:*

$$\mathrm{Adv}(\mathcal{M}_{LLP}, \mathcal{D}) = \widetilde{O}\left(\frac{\mathbb{E}[\eta(x)(1 - \eta(x)]^{1/4}(p(1 - p))^{1/4}}{\sqrt{k}} + \frac{\mathbb{E}[\eta(x)(1 - \eta(x))]^{1/4}}{k}\right),$$

where $\widetilde{O}$ hides logarithmic factors in k.

Hence, also in this more general case of the label proportion privacy mechanism, the advantage converges to zero as the bag size k grows large. Compared to the rate in Theorem 6.4, we are only losing the $\log k$ factors implicit in the $\widetilde{O}$ notation. This is because, when applied to the scenario where labels and features are independent, $\eta(x) = p$ is constant with x, so that $\mathbb{E}[\eta(x)(1-\eta(x))] = p(1-p)$, and the first term becomes $\sqrt{\frac{p(1-p)}{k}}$, while the second one reads $\frac{(p(1-p))^{1/4}}{k}$, which is lower order when k is large. We do not know whether the tighter gap-dependent analysis we carried out for Theorem 6.4 extends to the more general scenario of Theorem 3.5.

3.2 Bounding Attack Advantage for Randomized Response

In this section we provide explicit bounds on the attack advantage for RR. Recall that RR corresponds to the function

$$\mathcal{M}_{\text{RR}}(\mathbf{x}, \mathbf{y}) = \tilde{\mathbf{y}} \quad \text{where} \quad \tilde{y}_i = \begin{cases} y_i & \text{with probability } e^{\epsilon}/(1+e^{\epsilon}) \\ 1 - y_i & \text{with probability } 1/(1+e^{\epsilon}), \end{cases}$$

where the labels are flipped independently. For simplicity, we let $\pi = 1/(1+e^{\epsilon})$ denote the label flipping probability used by RR.

The results of Wu et al. [14] imply that every ϵ-label-DP PET $\mathcal{M}$ has advantage bounded by $\text{Adv}_k(\mathcal{M}, \mathcal{D}) \leq 1 - \frac{2}{1+e^{\epsilon}}$. However, one drawback of this bound is that it is distribution independent, yet the attack advantage depends heavily on the data distribution $\mathcal{D}$. For an extreme example, if we have $\mathbb{P}_{(x,y)\sim\mathcal{D}}(y = 1) = 1$, the attack advantage is zero for every PET, which is not captured by the bound derived using only properties of differential privacy.

In the remainder of this section, we derive an exact expression for the advantage of $\mathcal{M}_{\text{RR}}$ which we use to estimate the attack advantage of RR for various values of ϵ in our experiments in Section 5. Our exact expression is distribution dependent and leads to much tighter bounds on the attack advantage. For example at $\epsilon = 1$, the bound from Wu et al. [14] is $1 - \frac{2}{1+\epsilon} \approx 0.46$. However, for the dataset used in our experiments, we estimate the attack advantage for RR with $\epsilon = 1$ to be only 0.00095.

We first show that, since randomized response operates on each example independently[4], the advantage is independent of the number of examples k.

Lemma 3.6. *For all data distributions $\mathcal{D}$ and all k, $\text{Adv}_k(\mathcal{M}_{RR}, \mathcal{D}) = \text{Adv}_1(\mathcal{M}_{RR}, \mathcal{D})$.*

Next we derive a specialized version of Theorem 3.3 tailored to the case of RR and derive an expression for the optimal adversary under RR. Due to Lemma 3.6 we only need to consider the special case where $k = 1$.

Theorem 3.7. *For any data distribution $\mathcal{D}$, the attack advantage for randomized response with privacy parameter $\pi = \frac{1}{1+e^{\varepsilon}}$ is*

$$\text{Adv}_1(\mathcal{M}_{RR}, \mathcal{D}) = \mathbb{E}_x\Big[\big(\min\{\eta(x), 1-\eta(x)\} - \pi\big)\cdot\mathbb{I}\{\eta(x) \in [\pi, 1-\pi]\}\Big] .$$

The optimal adversary that maximizes $\text{EAU}_1(\cdot, \mathcal{M}_{RR}, \mathcal{D})$ is given below:

$$\mathcal{A}^*(x, \tilde{y}) = \begin{cases} 1, & \textit{if } \eta(x) > 1 - \pi \\ 0, & \textit{if } \eta(x) < \pi \\ \tilde{y}, & \textit{otherwise} \end{cases} .$$

4 Utility

Thus far we have focused on the privacy component of the privacy-utility trade-off. We now discuss our definition of utility which will be focused on the problem of learning a labeling hypothesis h^*.

Given a dataset $\mathcal{S} = (x_1, \ldots, x_{nk})$ let $\mathbf{x}_i = (x_{i1}, ..., x_{ik})$ and $\mathbf{y}_i = (y_{i1}, \ldots, y_{ik})$. Finally, let $\tilde{\mathcal{S}} = \big(\big(\mathbf{x}_1, \mathcal{M}(\mathbf{x}_1, \mathbf{y}_1)\big), \ldots, \big(\mathbf{x}_n, \mathcal{M}(\mathbf{x}_n, \mathbf{y}_n)\big)\big)$ be a *privatized* dataset. We define $\mathfrak{L}$ to be a learning algorithm that takes as input $\tilde{\mathcal{S}}$ and returns a hypothesis $\widehat{h} := \mathfrak{L}(\tilde{\mathcal{S}}) \in \mathcal{H}$.

[4] More generally, any PET $\mathcal{M} : (\mathcal{X} \times \mathcal{Y})^k \to \mathcal{Z}^k$ where $\mathcal{M}(\mathbf{x}, \mathbf{y})_i$ is a function of (x_i, y_i) and noise independent from the other examples has this property. In other words, if the PET does not do any aggregation across the examples, then the attack advantage is the same for all values of k.

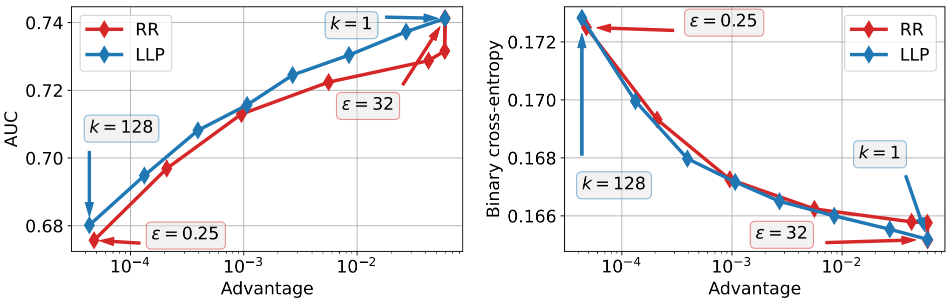

Figure 1: Privacy vs utility tradeoff curves for RR and LLP. Each point corresponds to a setting of the privacy parameter for the PET (i.e., ϵ for RR and k for LLP). The x coordinate is the reconstruction advantage for that PET, while the y coordinate is the utility of a model trained from the output of that PET (measured via AUC in the left figure, and binary cross-entropy in the right figure).

Given a loss function $\ell \colon \mathbb{R} \times \widehat{\mathcal{Y}} \to \mathbb{R}$ (for instance, binary cross-entropy), our definition of utility is given by the expected loss of $\widehat{h}$ over the data distribution $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\widehat{h}(x), y)]$, where smaller loss implies better utility. PropMatch is an instance of such a learning algorithm $\mathfrak{L}$.

5 Experiments

In this section we compare the privacy vs utility trade-off for RR and LLP in the context of training machine learning models for a click prediction task. At a high level, we train click prediction models from data that has been protected by either RR or LLP for various values of their privacy parameters. For each PET and privacy parameter, we measure the performance of the trained model via either the area under the receiver operating characteristic curve (AUC) or the binary cross-entropy loss and estimate the Attack Advantage bound provided by the PET. The curves relating Attack Advantage (privacy risk) to AUC (data utility) for RR and LLP are shown in Figure 1. We find that LLP provides a better Utility vs Attack Advantage tradeoff.

Dataset. We run our experiment on the click prediction data from the KDD Cup 2012, Track 2 [3] with the feature processing performed by Juan et al. [10]. The learning task for this data is to predict the click through rate for an advertisement based on a number of features related to the advertisement, the page that it appears on, and the user viewing it. There are 11 categorical features that are each one-hot encoded, resulting in a sparse feature vector with 11 non-zero entries in 54,686,452 dimensions. The label for the example is 1 if the ad was clicked, and 0 otherwise.

Model description and training setup. For both RR and LLP we train a deep embedding network using gradient descent. For RR, we perform regular stochastic gradient descent on a dataset obtained by flipping each label with probability $1/(1 + e^{\epsilon})$ together with a debaising procedure that produces unbiased estimates of the gradient of the model's loss w.r.t. its parameters. For LLP, we use stochastic gradient descent to optimize the Empirical Proportion Risk defined in Section 2.

The model architecture in both cases is as follows: First, we reduce the dimension of each example from 54,686,452 dimensions to 100,000 dimensions by hashing feature indices. Each hashed feature index is associated with a learned embedding vector in $\mathbb{R}^{50}$, and the representation vector for an example is the sum of the learned embeddings for each of its non-zero hashed feature indices. This representation vector is passed through two dense layers with 100 and 50 units, respectively, and ReLU activation functions. The final output is a single unit with sigmoid activation that is interpreted as the click probability for each example.

Experimental Setup. For RR and LLP, we train the above model using the Adam optimizer to minimize the binary cross-entropy loss. For RR, we use privacy parameters ϵ in $\{2^{-4}, 2^{-3}, \ldots, 2^{5}\}$ and for LLP we use bag sizes k in $\{2^0, 2^1, \ldots, 2^9\}$. For both RR and LLP, for every value of the privacy parameter, we train the model with each learning rate in $\{10^{-6}, 5 \cdot 10^{-6}, 10^{-5}, 10^{-4}, 5 \cdot 10^{-4}, 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}\}$. Finally, for each combination of privacy parameter and learning rate, we train the model 10 times (each trial corresponds to different model initialization and data shuffling). For each privacy parameter, we report the mean AUC of the learning rate with the highest mean AUC.

Estimating Attack Advantage. The attack advantage of the optimal attacker is estimated and reported for every PET with various parameters. The optimal attacker and its advantage have been characterized for RR as well as for LLP, in Theorem 3.3 and 3.7, respectively. Crucially, notice that the advantage can be easily calculated with knowledge of the parameters of each PET (label flipping probability or the bag size), and the class conditional distribution $\eta(x)$ — or at least an accurate estimate of it. We considered two estimates of the class conditionals: one based on Deep Neural Networks (DNN) as described above, and one based of k-Nearest Neighbors (kNN). The kNN estimator can produce accurate estimates for class conditional in the large-scale data regime, due to its strong consistency properties that has been studied in the 70s and 80s and became the part of machine learning folklore. DNN can also produce accurate estimates if the training process and architecture is tuned carefully enough. We found that these two approaches result in very similar results on our large scale benchmark datasets, therefore we relied on the estimate provided by DNN, and we estimated the attack advantage of the optimal attacker based on the output score of DNN.

Results. Figure 1 shows the utility vs advantage tradeoff for RR and LLP. We find that across the advantage spectrum, LLP provides a slightly higher utility than RR when measured either by AUC or by the binary cross-entropy loss.

6 Protecting data with DP and Aggregation

Thus far we have focused on understanding the privacy risks through the average notion of attack advantage. It is worth mentioning that this notion does not take into account potential side information from an attacker in addition to feature vectors and knowledge of the distribution $\mathcal{D}$. In particular, the distributional assumption may be broken if the attacker has knowledge of some of the true labels in the dataset.

In order to provide more extensive privacy protection, one could combine both LLP and Differential Privacy. One way to achieve this is to design a PET that releases a differentially private estimate of the label proportion. For example, we can provide ϵ-label-DP by adding $\text{Laplace}(\frac{1}{k\epsilon})$ noise to the true label proportion [9]. Formally, after fixing the privacy parameter ϵ, this corresponds to the following PET: $\mathcal{M}_{\text{Lap-LLP}}(\mathbf{x}, \mathbf{y}) = \tilde{\alpha} := \frac{1}{k}\sum_{i=1}^{y} + Z, \quad \text{where} \quad Z \sim \text{Laplace}\left(\frac{1}{k\epsilon}\right).$

From a privacy perspective, $\mathcal{M}_{\text{Lap-LLP}}$ is ϵ-label-DP and enjoys the full suite of formal guarantees that this implies. At the same time, for a fixed value of k, the attack advantage of $\mathcal{M}_{\text{Lap-LLP}}$ is never larger than that of $\mathcal{M}_{\text{LLP}}$, since adding noise to the label proportion cannot increase advantage. Next, since $\mathcal{M}_{\text{LLP}}$ is a special case of $\mathcal{M}_{\text{Lap-LLP}}$ (where $\epsilon = \infty$), the utility-vs-advantage trade-off for $\mathcal{M}_{\text{Lap-LLP}}$ cannot be worse than that of $\mathcal{M}_{\text{LLP}}$.

In the remainder of the section we argue that PROPMATCH is still an effective learning algorithm for the PET $\mathcal{M}_{\text{Lap-LLP}}$. Recall that PROPMATCH generally learns by using stochastic gradient descent (SGD) to minimize the empirical proportion matching loss. Our first result shows that the expected value of the gradient of proportion matching loss is the same whether Laplace noise is added to the label proportion or not. In other words, this implies that the expected trajectory of SGD for PROPMATCH when using $\mathcal{M}_{\text{Lap-LLP}}$ is the same for all values of ϵ, including $\epsilon = \infty$ which corresponds to the speical case of $\mathcal{M}_{\text{LLP}}$.

Theorem 6.1. *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function, and let $\tilde{\alpha} = \alpha + Z$ where Z is unbiased and independent of the data. Then, for the square loss function defined by $\ell(p, X) = (p - X)^2$ and the binary cross entropy loss function defined by $\ell(p, X) = -X\log(p) - (1 - X)\log(1 - p)$, we have $\mathbb{E}\left[\nabla_\theta \ell\left(q, \tilde{\alpha}\right)\right] = \mathbb{E}\left[\nabla_\theta \ell\left(q, \alpha\right)\right]$.*

The main way that decreasing the value of ϵ (resulting in stronger differential privacy guarantees) affects utility is that the *variance* of the gradients increases. The following result characterizes how much the variance increases compared to the case where $\epsilon = \infty$ (or equivalently, compared to PROPMATCH run with $\mathcal{M}_{\text{LLP}}$).

Theorem 6.2. *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be the binary cross entropy loss function defined by $\ell(p, X) = -X\log(p) - (1 - X)\log(1 - p)$, and let $\tilde{\alpha} = \frac{1}{k}\sum_{i=1}^{k} y_i + Z$, where $Z \sim \text{Laplace}(\frac{1}{k\epsilon})$ is the output*

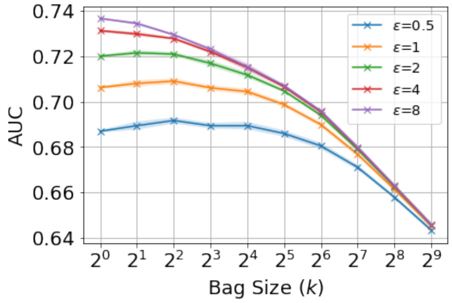

Figure 2: This plot depicts the AUC for PROPMATCH when run with PET $\mathcal{M}_{\text{Lap-LLP}}$ as a function of the privacy parameters ϵ and k. Note that the utility gap between $\epsilon = 0.5$ and $\epsilon = 8$ shrinks as the bag size grows, which is what we would expect given the variance decomposition of Theorem 6.2.

of $\mathcal{M}_{\textit{Lap-LLP}}$. Then,

$$\mathbb{E}\left[\|\nabla_\theta \ell\left(q, \tilde{\alpha}\right)\|_2^2\right] = \mathbb{E}\left[\|\nabla_\theta \ell(q, \alpha)\|^2\right] + \frac{2}{k^2 \varepsilon^2} \cdot \mathbb{E}\left[\|\nabla_\theta \log\left(q/(1-q)\right)\|^2\right]$$

where $q = \frac{1}{k}\sum_{x \in \mathcal{B}} h_\theta(x)$ is the average prediction.

Intuitively, Theorem 6.2 demonstrates that when the bag size k is large, adding differential privacy has a smaller impact than when the bag size k is small. Figure 2 plots the AUC for the same dataset and model as in Section 5 when training PROPMATCH using $\mathcal{M}_{\text{Lap-LLP}}$. It shows that the decrease in AUC due to decreasing ϵ shrinks as k grows.

Furthermore, we characterize the optimal attacker for $\mathcal{M}_{\text{Lap-LLP}}$ and we bound the advantage.

Theorem 6.3. *For any privacy parameters $k \geq 1$, $\varepsilon > 0$, and any data distribution $\mathcal{D}$, $\mathrm{Adv}_k(\mathcal{M}_{\textit{Lap-LLP}}, \mathcal{D})$ is maximized by the following adversary:*

$$\mathcal{A}^*(\mathbf{x}, \mathcal{M}_{\textit{Lap-LLP}}(\mathbf{x}, \mathbf{y}))_i := \begin{cases} 1 & \textit{if} \quad \frac{\eta(\mathbf{x}_i)}{1-\eta(\mathbf{x}_i)} \cdot \frac{\sum_{b=1}^{k} \mathbb{P}(\textit{PBin}(\{\eta(x_j)\}_{i \neq j}) = b-1) \cdot e^{-k\varepsilon|\tilde{\alpha} - b/k|}}{\sum_{b=0}^{k-1} \mathbb{P}(\textit{PBin}(\{\eta(x_j)\}_{i \neq j}) = b) \cdot e^{-k\varepsilon|\tilde{\alpha} - b/k|}} \geq 1 \\ 0 & \textit{otherwise}. \end{cases}$$

Alternatively, using properties of the Laplace distribution, this adversary is approximately equivalent to the following for small ε

$$\mathcal{A}^{**}(\mathbf{x}, \mathcal{M}_{\textit{Lap-LLP}}(\mathbf{x}, \mathbf{y}))_i := \begin{cases} 1 & \textit{if} \quad \frac{\eta(x_i)}{1-\eta(x_i)} < e^{-\varepsilon} \\ 0 & \textit{if} \quad \frac{\eta(x_i)}{1-\eta(x_i)} > e^{\varepsilon} \\ y_i & \textit{otherwise} \end{cases}.$$

Theorem 6.4. *Let $\mathcal{D}$ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and let $p = \mathbb{E}[\eta(x)]$. Then for all bag sizes $k \geq 1$ we have:*

$$\begin{aligned} &\mathrm{Adv}_k(\mathcal{M}_{\textit{Lap-LLP}}, \mathcal{D}) \\ &\leq \min\left\{2(1 - e^{-\epsilon})\,\mathbb{E}[\eta(x)(1-\eta(x)], \widetilde{O}\left(\frac{\mathbb{E}[\eta(x)(1-\eta(x)]^{1/4}(p(1-p))^{1/4}}{\sqrt{k}} + \frac{\mathbb{E}[\eta(x)(1-\eta(x))]^{1/4}}{k}\right)\right\}. \end{aligned}$$

7 Conclusion

We have extended the notion of expected attach utility of [14]. This extension allowed us, for the first time, to compare two commonly used PETs: randomzied response and LLP. We show that in some scenarios LLP provides a better privacy-utility tradeoff and that the protections to LLP can also be extended to the more common scenario of differential privacy while preserving most of its algorithmic properties. Our intention is that regulators and decision makers can make more informed selections of PETs to preserve user privacy while maintaining the high utility of their products.

References

[1] The digital markets act: ensuring fair and open digital markets. URL https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en.

[2] Prepare for phasing out third-party cookies. URL https://developer.chrome.com/en/docs/privacy-sandbox/third-party-cookie-phase-out/.

[3] Yi Wang Aden. Kdd cup 2012, track 2, 2012. URL https://kaggle.com/competitions/kddcup2012-track2.

[4] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pages 1138–1156. IEEE, 2022.

[5] John J Benedetto. *Harmonic analysis and applications*. CRC Press, 2020.

[6] R. Busa-Fekete, H. Choi, T. Dick, C. Gentile, and A. Munos Medina. Easy learning from label proportions. *arXiv preprint arXiv:2302.03115*, 2023.

[7] Róbert Istvan Busa-Fekete, Andrés Muñoz Medina, Umar Syed, and Sergei Vassilvitskii. Label differential privacy and private training data release. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 3233–3251. PMLR, 2023.

[8] CJ Carey, Travis Dick, Alessandro Epasto, Adel Javanmard, Josh Karlin, Shankar Kumar, Andres Muñoz Medina, Vahab Mirrokni, Gabriel Henrique Nunes, Sergei Vassilvitskii, and Peilin Zhong. Measuring re-identification risk. *Proc. ACM Manag. Data*, 1(2):149:1–149:26, 2023.

[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[10] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. Field-aware factorization machines for ctr prediction. In *Proceedings of the 10th ACM conference on recommender systems*, pages 43–50, 2016.

[11] Henry Reeve and Kabán. Classification with unknown class-conditional label noise on non-compact feature spaces. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2624–2651. PMLR, 25–28 Jun 2019.

[12] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society, 2017.

[13] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *JASA*, pages 63–69, 1965.

[14] Ruihan Wu, Jin Peng Zhou, Kilian Q Weinberger, and Chuan Guo. Does label differential privacy prevent label inference attacks? *arXiv preprint arXiv:2202.12968*, 2022.

[15] F. Yu, D. Liu, S. Kumar, T. Jebara, and S.F. Chang. α-svm for learning with label proportions. In *ICML*, Proceedings of the 30th International Conference on Machine Learning, pages 504–512, 2013.

A Proofs for Section 3

A.1 Proof of Lemma 3.2

Lemma 3.2. *The following expected attack utility bound holds for any PET $\mathcal{M}$, data distribution $\mathcal{D}$, number of examples k and adversary $\mathcal{A}$ that observes the output of $\mathcal{M}$*

$$\mathrm{EAU}_k(\mathcal{A},\mathcal{M},\mathcal{D}) \le 1 - \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^k \\ i\sim\textit{Uniform([k])}}}\big[\min\{\mathbb{P}(y_i = 1 \mid \mathbf{x}, \mathcal{M}(\mathbf{x},\mathbf{y})), \mathbb{P}(y_i = 0 \mid \mathbf{x}, \mathcal{M}(\mathbf{x},\mathbf{y}))\}\big].$$

Moreover, the following optimal adversary $\mathcal{A}^*$ achieves this upper bound with equality:

$$\mathcal{A}^*(\mathbf{x},\mathcal{M}(\mathbf{x},\mathbf{y}))_i := \begin{cases} 1 & \textit{if}\, \mathbb{P}(y_i = 1 \mid \mathbf{x},\mathcal{M}(\mathbf{x},\mathbf{y})) \ge 1/2 \\ 0 & \textit{otherwise}. \end{cases}$$

Proof. Fix any adversary $\mathcal{A}$. We begin by lower bounding the probability that the adversary makes a mistake on a fixed example j. Let $(\mathbf{x},\mathbf{y}) \sim \mathcal{D}^k$ be an i.i.d. sample of labeled examples, $z = \mathcal{M}(\mathbf{x},\mathbf{y})$ be the output of the PET, and $(\hat{y}_1,\dots,\hat{y}_k) = \mathcal{A}(\mathbf{x},z)$ be the output of the adversary. Since $\hat{y}_j$ is $(\mathbf{x},z)$-measurable, we have that

$$\begin{aligned}\mathbb{P}(\hat{y}_j \ne y_j \mid \mathbf{x},z) &= \mathbb{P}(\hat{y}_j = 0 \mid y_j = 1, \mathbf{x}, z)\,\mathbb{P}(y_j = 1 \mid \mathbf{x},z) \\ &\quad + \mathbb{P}(\hat{y}_j = 1 \mid y_j = 0, \mathbf{x}, z)\,\mathbb{P}(y_j = 0 \mid \mathbf{x},z) \\ &= \mathbb{1}\{\hat{y}_j = 0\}\,\mathbb{P}(y_j = 1 \mid \mathbf{x},z) + \mathbb{1}\{\hat{y}_j = 1\}\,\mathbb{P}(y_j = 0 \mid \mathbf{x},z).\end{aligned}$$

Next, let i be an index drawn uniformly at random from $[k]$. Then we can write the expected attack utility as follows (where the randomness is over the variables $(\mathbf{x},\mathbf{y})$, $z = \mathcal{M}(\mathbf{x},\mathbf{y})$, and i):

$$\begin{aligned}\mathrm{EAU}_k(\mathcal{A},\mathcal{M},\mathcal{D}) &= 1 - \mathbb{P}(\hat{y}_i \ne y_i) \\ &= 1 - \mathbb{E}[\mathbb{1}\{\hat{y}_i \ne y_i\}] \\ &= 1 - \mathbb{E}\big[\mathbb{E}[\mathbb{1}\{\hat{y}_i \ne y_i\} \mid \mathbf{x}, z, i]\big] \\ &= 1 - \mathbb{E}[\mathbb{1}\{\hat{y}_i = 0\}\,\mathbb{P}(y_i = 1 \mid \mathbf{x},z) + \mathbb{1}\{\hat{y}_i = 1\}\,\mathbb{P}(y_i = 0 \mid \mathbf{x},z)] \\ &\le 1 - \mathbb{E}\big[\min\{\mathbb{P}(y_i = 1 \mid \mathbf{x},z), \mathbb{P}(y_i = 0 \mid \mathbf{x},z)\}\big]\end{aligned}$$

where the inequality holds because the minimum probability term never exceeds the probability term selected by the indicator variables. Finally, to show that $\mathcal{A}^*$ is the optimal adversary, observe that the value of $\hat{y}_i$ chosen by $\mathcal{A}^*$ "selects" the smaller of the two probability terms with probability one. It follows that for $\mathcal{A}^*$, the inequality above holds with equality. □

A.2 Proof of Theorem 3.3

Theorem 3.3. *For any PET $\mathcal{M}$, any data distribution $\mathcal{D}$, and number of examples k, the attack advantage is given by*

$$\begin{aligned}\mathrm{Adv}_k(\mathcal{M},\mathcal{D}) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\big[\min\{\eta(x), 1-\eta(x)\}\big] \\ &\quad - \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^k \\ i\sim\textit{Uniform}([k])}}\Big[\min\big\{\mathbb{P}(y_i = 1 \mid \mathbf{x},\mathcal{M}(\mathbf{x},\mathbf{y})), \mathbb{P}(y_i = 0 \mid \mathbf{x},\mathcal{M}(\mathbf{x},\mathbf{y}))\big\}\Big]\end{aligned}$$

where $\eta : x \mapsto \mathbb{P}(y = 1 \mid x)$ is the Bayes optimal predictor.

Proof. Recall that the advantage is the difference in expected attack utility for an optimal attacker having access to $\mathcal{M}$ compared to $\mathcal{M}_\perp$. This theorem follows from Lemma 3.2 together with a simplification of the optimal expected uninformed attack utility.

Since $\mathcal{M}_\perp$ outputs a constant and the (x_i, y_i) pairs for $i \in [k]$ are independent, we have that

$$\mathbb{P}(y_i = 1 \mid \mathbf{x}, z) = \mathbb{P}(y_i = 1 \mid x_i) = \eta(x_i).$$

It follows that the expected attack utility of the optimal uninformed adversary is given by

$$1 - \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^k \\ i\sim\mathrm{Uniform}([k])}}\big[\min\{\eta(x_i), 1-\eta(x_i)\}\big] = 1 - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\big[\min\{\eta(x), 1-\eta(x)\}\big],$$

where the equality follows from the fact that $x_1, \ldots, x_k$ are i.i.d.

The proof is completed by taking the difference of the optimal expected attack utilities for the informed and uninformed adversaries. $\square$

A.3 Proof of Theorem 6.4

Theorem 3.4. *Fix a data distribution $\mathcal{D}$, let $p = \mathbb{P}_{(x,y)\sim\mathcal{D}}(y=1)$, and fix an arbitrary threshold $\beta \in [0, 1/2]$. If labels are independent of features (i.e., $\mathcal{D}$ is a product of distributions over $\mathcal{X}$ and $\mathcal{Y}$), then for all bag sizes $k \geq 1$ we have:*

$$\mathrm{Adv}_k(\mathcal{M}_{LLP}, \mathcal{D}) = \min\{p, 1-p\} - \mathop{\mathbb{E}}_{\alpha}[\min\{\alpha, 1-\alpha\}] \leq \begin{cases} \sqrt{\frac{p(1-p)}{k}} & \textit{if } p \in [0,1] \\ e^{-\Omega(\beta^2 k)} & \textit{if } |p - 1/2| \geq \beta, \end{cases}$$

where the Ω notation hides constants independent of β and k.

Proof. As in the proof of Theorem 3.3, we have that

$$\max_{\mathcal{A}} \mathrm{EAU}_k(\mathcal{A}, \mathcal{M}_\perp, \mathcal{D}) = \mathop{\mathbb{E}}_{x}[\min\{\eta(x), 1-\eta(x)\}].$$

Further, since this theorem studies the case where labels are independent of features, we have that $\eta(x) = \mathbb{P}(y = 1 \mid x) = \mathbb{P}(y=1) = p$, which implies that

$$\max_{\mathcal{A}} \mathrm{EAU}_k(\mathcal{A}, \mathcal{M}_\perp, \mathcal{D}) = \min\{p, 1-p\}.$$

As for $\mathrm{EAU}_k(\mathcal{A}^*, \mathcal{M}_{\mathrm{LLP}}, \mathcal{D})$, set for brevity $\Sigma = k\alpha = \sum_{i=1}^k y_i$ and let $\hat{y}_1, \ldots, \hat{y}_k$ be the predictions of the optimal adversary. We can write

$$\begin{aligned}
\mathbb{E}[\mathbb{I}\{\hat{y}_i \neq y_i\}] &= \mathbb{E}_\Sigma\, \mathbb{E}[\mathbb{I}\{\hat{y}_i \neq y_i\} \,|\, \Sigma] \\
&= \mathbb{E}_\Sigma\, \mathbb{E}[\mathbb{I}\{\hat{y}_i = 0,\, y_i = 1\} \,|\, \Sigma] + \mathbb{E}_\Sigma\, \mathbb{E}[\mathbb{I}\{\hat{y}_i = 1,\, y_i = 0\} \,|\, \Sigma] \\
&= \mathbb{E}_\Sigma[\mathbb{I}\{\hat{y}_i = 0\}\, \mathbb{E}[\mathbb{I}\{y_i = 1\} \,|\, \Sigma]] + \mathbb{E}_\Sigma[\mathbb{I}\{\hat{y}_i = 1\}\, \mathbb{E}[\mathbb{I}\{y_i = 0\} \,|\, \Sigma]]
\end{aligned}$$

which is minimized when

$$\hat{y}_i = \begin{cases} 1 & \text{if } \mathbb{E}[\mathbb{I}\{y_i = 1\} \,|\, \Sigma] \geq 1/2 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Since $\mathbb{E}[\mathbb{I}\{y_i = 1\} \,|\, \Sigma] = \alpha$ independent of i and p, the minimum value is thus

$$\mathbb{E}_\alpha[\min\{\alpha, 1-\alpha\}],$$

so that $\mathrm{EAU}_k(\mathcal{A}^*, \mathcal{M}_{\mathrm{LLP}}, \mathcal{D}) = 1 - \mathbb{E}_\alpha[\min\{\alpha, 1-\alpha\}]$, and the claimed equality for $\mathrm{Adv}_k(\mathcal{M}_{\mathrm{LLP}}, \mathcal{D})$ follows.

As for the inequality with general $p \in [0,1]$, note that, when $a, b \in [0,1]$,

$$\min\{a, 1-a\} - \min\{b, 1-b\} \leq |a - b|\,. \tag{3}$$

If applied to the expression

$$\min\{p, 1-p\} - \mathbb{E}_\alpha[\min\{\alpha, 1-\alpha\}]$$

this gives

$$\mathrm{Adv}_k(\mathcal{M}_{\mathrm{LLP}}, \mathcal{D}) \leq \mathbb{E}_\alpha[|\alpha - p|] \leq \sqrt{\mathbb{E}_\alpha[(\alpha - p)^2]} = \sqrt{\frac{p(1-p)}{k}}\,,$$

where the second inequality is Jensen's.

Finally, in the case where $|p - 1/2| \geq \beta$, for some gap $\beta > 0$, we can proceed through a more direct analysis. Assume $p \leq 1/2 - \beta$. We can write

$$\begin{aligned}
\min\{\alpha, 1-\alpha\} &= \min\{\alpha, 1-\alpha\}\mathbb{I}\{\alpha \leq 1/2\} + \min\{\alpha, 1-\alpha\}\mathbb{I}\{\alpha > 1/2\} \\
&\qquad + \alpha\mathbb{I}\{\alpha > 1/2\} - \alpha\mathbb{I}\{\alpha > 1/2\} \\
&= \alpha\mathbb{I}\{\alpha \leq 1/2\} + (1-\alpha)\mathbb{I}\{\alpha > 1/2\} + \alpha\mathbb{I}\{\alpha > 1/2\} - \alpha\mathbb{I}\{\alpha > 1/2\} \\
&= \alpha - (2\alpha - 1)\mathbb{I}\{\alpha > 1/2\} \\
&\geq \alpha - \mathbb{I}\{\alpha > 1/2\}\,.
\end{aligned}$$

Hence

$$\mathbb{E}_\alpha[\min\{\alpha, 1-\alpha\}] \geq p - \Pr(\alpha > 1/2) \ .$$

Now, $p < 1/2$ implies $\min\{p, 1-p\} = p$, which leads us to

$$\min\{p, 1-p\} - \mathbb{E}_\alpha[\min\{\alpha, 1-\alpha\}] \leq \Pr(\alpha > 1/2) \ .$$

Finally, by the standard Bernstein inequality we have

$$\mathbb{P}(\alpha > 1/2) \leq \exp\left(-\frac{k(1/2-p)^2}{2p(1-p) + (1-2p)/3}\right) = e^{-\Omega(k\beta^2)} \ ,$$

which gives the second inequality.

A similar argument holds if we reverse the assumption on p to $p \geq 1/2 + \beta$. □

A.4 Proof of Theorem 3.5

Notation. For simplicity of notation we let $\eta_i = \eta(x_i)$ be the conditional positive probability of a label given feature vector x and let $p = \mathbb{E}[\eta_i]$. For fixed feature vectors $\mathbf{x} = (x_1, \ldots, x_k, x_{k+1})$, let A_i denote a Bernoulli random variable with mean η_i, and let $Z_k = \sum_{j=1}^k A_j$

We recall the statement of the theorem we want to prove.

Theorem 3.5. *Let $\mathcal{D}$ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and let $p = \mathbb{E}[\eta(x)]$. Then, for all bag sizes $k \geq 2$ we have:*

$$\mathrm{Adv}(\mathcal{M}_{LLP}, \mathcal{D}) = \widetilde{O}\left(\frac{\mathbb{E}[\eta(x)(1-\eta(x)]^{1/4}(p(1-p))^{1/4}}{\sqrt{k}} + \frac{\mathbb{E}[\eta(x)(1-\eta(x))]^{1/4}}{k}\right) \ ,$$

where $\widetilde{O}$ hides logarithmic factors in k.

The above theorem is a direct consequence of the following Lemma.

Lemma A.1. *Let $\mathcal{B}$ denote a bag with $k+1$ elements. Let*

$$c_k = \frac{1}{\sqrt{k}}\left(\frac{1}{3}\log 8k + \frac{1}{6}\sqrt{2\log 8k + 12kp(1-p)\log 8k}\right) = \widetilde{O}\left(\sqrt{p(1-p)} + \frac{1}{\sqrt{k}}\right) \ .$$

Then

$$\mathrm{Adv}_{k+1}(\mathcal{M}_{LLP}, \mathcal{D}) \leq k^{1/4}\sqrt{2c_k}\sqrt{\left(\frac{1}{e^{3/2}} + \frac{\pi}{4} + \frac{\pi}{e}\right)\frac{\mathbb{E}[\eta_1(1-\eta_1)]^{1/2}}{k^{3/2}} + \frac{\mathbb{E}[\eta_1(1-\eta_1)]}{k}} \ .$$

Proof. Set for brevity $\mathcal{M} = \mathcal{M}_{\mathrm{LLP}}(\mathbf{x}, \mathbf{y})$. By definition of Adv_{k+1} we have

$$\begin{aligned}
&\mathrm{Adv}_{k+1}(\mathcal{M}_{\mathrm{LLP}}, \mathcal{D}) \\
&\quad = \mathop{\mathbb{E}}_{x_{k+1}}[\min\{\eta(x_{k+1}), 1-\eta(x_{k+1})\}] - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})}[\min\{\mathbb{P}(y_{k+1} = 1|\mathbf{x}, \mathcal{M}), \mathbb{P}(y_{k+1} = 0|\mathbf{x}, \mathcal{M})\}] \\
&\quad \leq \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})}[|\eta(x_{k+1}) - \mathbb{P}(y_{k+1} = 1|\mathbf{x}, \mathcal{M})|] \ ,
\end{aligned}$$

where we have used the fact that $\min(a, 1-a) - \min(b, 1-b) \leq |a-b|$ for any real numbers a, b. We now focus on calculating $\mathbb{P}(y_{k+1} = 1|\mathbf{x}, \mathcal{M})$. Let $\Sigma = \sum_{j=1}^{k+1} y_j$. Note that for a given realization of feature vector $\mathbf{x}$, y_{k+1} is distributed like A_{k+1} and the output $\mathcal{M}$ is distributed like Z_{k+1}. Therefore:

$$\begin{aligned}
\mathbb{P}(y_{k+1} = 1 \,|\, \mathbf{x}, \mathcal{M}) &= \mathbb{P}(A_{k+1} = 1 \,|\, \mathbf{x}, Z_{k+1} = \Sigma) \\
&= \frac{\mathbb{P}(A_{k+1} = 1, Z_{k+1} = \Sigma \,|\, \mathbf{x})}{\mathbb{P}(Z_{k+1} = \Sigma \,|\, \mathbf{x})} \\
&= \eta_{k+1}\frac{\mathbb{P}(Z_k = \Sigma - 1 \,|\, \mathbf{x})}{\mathbb{P}(Z_{k+1} = \Sigma \,|\, \mathbf{x})} \ .
\end{aligned}$$

Using this expression in the original expectation we see that we can bound the advantage as

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}\left[\eta_{k+1}\left|1-\frac{\mathbb{P}(Z_k=\Sigma-1\,|\,\mathbf{x})}{\mathbb{P}(Z_{k+1}=\Sigma\,|\,\mathbf{x})}\right|\right]$$

Again, note that for a fixed $\mathbf{x}$, the variable Σ is distributed like Z_{k+1}. Therefore. taking expectation over $\mathbf{y}$ the above expression can be rewritten as

$$\mathbb{E}_{\mathbf{x}}\left[\eta_{k+1}\sum_{s=0}^{k+1}\left|1-\frac{\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})}{\mathbb{P}(Z_{k+1}=s)}\right|\mathbb{P}(Z_{k+1}=s\,|\,\mathbf{x})\right]=\mathbb{E}_{\mathbf{x}}\left[\eta_{k+1}\sum_{s=0}^{k+1}\Big|\mathbb{P}(Z_{k+1}=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})\Big|\right]$$

Finally, note that since $Z_{k+1}=Z_k+A_{k+1}$ we also have

$$\begin{aligned}\mathbb{P}(Z_{k+1}=s\,|\,\mathbf{x})&=\mathbb{P}(A_{k+1}=1\,|\,\mathbf{x})\,\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})+\mathbb{P}(A_{k+1}=0\,|\,\mathbf{x})\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})\\&=\eta_{k+1}\,\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})+(1-\eta_{k+1})\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})\;.\end{aligned}$$

Therefore, we conclude that the advantage can be bounded by

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}}\left[\eta_{k+1}(1-\eta_{k+1})\sum_{s=0}^{k+1}\Big|\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})\Big|\right]=\\&\qquad\mathbb{E}_{\mathbf{x}}\left[\eta_{k+1}(1-\eta_{k+1})\right]\,\mathbb{E}\left[\sum_{s=0}^{k+1}\Big|\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})\Big|\right]\;,\end{aligned}$$

where we have used the fact that the random variables η_i are independent from each other. Using also the fact that η_{k+1} has the same distribution as η_1 combined with Lemma A.2 below, we have that the above quantity is bounded by:

$$\begin{aligned}&k^{1/4}\sqrt{2c_k}\sqrt{\mathbb{E}[\eta_1(1-\eta_1)]^2\,\mathbb{E}[\eta_1^2+(1-\eta_1)^2]^k+\frac{\pi\,\mathbb{E}[\eta_1(1-\eta_1)]^{1/2}}{4k^{3/2}}+\frac{\pi\,\mathbb{E}[\eta_1(1-\eta_1)]}{ek^2}}\\&+\frac{\mathbb{E}[\eta_1(1-\eta_1)]}{k}\;.\end{aligned}\tag{4}$$

Moreover, notice that $\mathbb{E}[\eta_1^2+(1-\eta_1)^2]+2\,\mathbb{E}[\eta_1(1-\eta_1)]=1$. Therefore $\mathbb{E}[\eta_1^2+(1-\eta_1)^2]=1-2\,\mathbb{E}[\eta_1(1-\eta_1)$, and using the fact that $\eta_1(1-\eta_1)\le\frac{1}{4}$ we have

$$\begin{aligned}\mathbb{E}[\eta_1(1-\eta_1)]^2\,\mathbb{E}[\eta_1^2+(1-\eta_1)^2]^k&=\mathbb{E}[\eta_1(1-\eta_1)]^2(1-2\,\mathbb{E}[\eta_1(1-\eta_1)])^k\\&\le\max_{\frac{1}{4}\ge x\ge 0}x^2(1-2x)^k\;.\end{aligned}$$

But a simple calculation shows that the above function is maximized at $x_k^\star=\min\{\frac{1}{k+2},1/4\}$, thus we must have

$$\mathbb{E}[\eta_1(1-\eta_1)]^2\,\mathbb{E}[\eta_1^2+(1-\eta_1)^2]^k\le(x_k^\star)^2(1-2x_k^\star)^k\le\frac{1}{(ek)^2}$$

the last inequality holding for all $k\ge1$. In addition, we have the trivial bound $\mathbb{E}[\eta_1(1-\eta_1)]^2\,\mathbb{E}[\eta_1^2+(1-\eta_1)^2]\le\mathbb{E}[\eta_1(1-\eta_1)]^2$, so that

$$\begin{aligned}\mathbb{E}[\eta_1(1-\eta_1)]^2\,\mathbb{E}[\eta_1^2+(1-\eta_1)^2]^k&\le\min\left\{\mathbb{E}[\eta_1(1-\eta_1)]^2,\frac{1}{e^2k^2}\right\}\\&\le\frac{\mathbb{E}[\eta_1(1-\eta_1)]^2}{\mathbb{E}[\eta_1(1-\eta_1)]^2\,e^2\,k^2+1}\\&\quad\text{(using }\min\{a,b\}\le\tfrac{ab}{a+b}\text{, with }a=\mathbb{E}[\eta_1(1-\eta_1)]^2\text{ and }b=\tfrac{1}{e^2k^2}\text{)}\\&\le\frac{\sqrt{\mathbb{E}[\eta_1(1-\eta_1)]}}{e^{3/2}k^{3/2}}\\&\quad\text{(using }x^2-x^{3/2}+1\ge0\text{, with }x=ek\,\mathbb{E}[\eta_1(1-\eta_1)]\text{)}\;.\end{aligned}$$

Replacing this bound in (4) we obtain the following upper bound on the advantage:

$$
\begin{aligned}
&k^{1/4}\sqrt{2c_k}\sqrt{\left(\frac{1}{e^{3/2}}+\frac{\pi}{4}\right)\frac{\mathbb{E}[\eta_1(1-\eta_1)]^{1/2}}{k^{3/2}}+\frac{\pi\,\mathbb{E}[\eta_1(1-\eta_1)]}{ek^2}+\frac{\mathbb{E}[\eta_1(1-\eta_1)]}{k}} \\
&\qquad\le k^{1/4}\sqrt{2c_k}\sqrt{\left(\frac{1}{e^{3/2}}+\frac{\pi}{4}+\frac{\pi}{e}\right)\frac{\mathbb{E}[\eta_1(1-\eta_1)]^{1/2}}{k^{3/2}}+\frac{\mathbb{E}[\eta_1(1-\eta_1)]}{k}}\,,
\end{aligned}
$$

as claimed. □

Lemma A.2. *Let c_k be as in Lemma A.1. Then the following bound holds:*

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\sum_{s=0}^{k+1}|\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})|\right]\le \\
&\quad\frac{1}{k}+k^{1/4}\sqrt{2c_k}\sqrt{\mathbb{E}[\eta_1^2+(1-\eta_1)^2]^k+\frac{\pi}{(4\,\mathbb{E}[\eta_1(1-\eta)_1)k)^{3/2}}+\frac{\pi}{e\,\mathbb{E}[\eta_1(1-\eta_1)]k^2}}
\end{aligned}
$$

Proof. Let $a>0$ and $b<k$, and let $[a,b]=\{j\in\mathbb{N}|a\le j\le b\}$. For any $\mathbf{x}$ we then have

$$
\begin{aligned}
&\sum_{s=0}^{k+1}|\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})| \\
&\quad=\sum_{s\in[a,b]}|\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})|+\sum_{s\notin[a,b]}|\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1)| \\
&\quad\le\sum_{s\in[a,b]}|\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})|+\sum_{s\notin[a,b]}\mathbb{P}(Z_k=s\,|\,\mathbf{x})+\mathbb{P}(Z_k=s-1\,|\,\mathbf{x}) \\
&\quad\le\sum_{s\in[a,b]}|\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})|+2\,\mathbb{P}(Z_k\notin[a,b]\,|\,\mathbf{x})
\end{aligned}
$$

Taking expectation over both sides with respect to $\mathbf{x}$ we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\sum_{s=0}^{k+1}|\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})|\right] \\
&\quad\le\mathbb{E}_{\mathbf{x}}\left[\sum_{s\in[a,b]}|\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})|\right]+2\mathbb{E}_{\mathbf{x}}[\mathbb{P}(Z_k\notin[a,b]\,|\,\mathbf{x})] \qquad (5)
\end{aligned}
$$

Let now Q_k denote the probability measure associated with a binomial random variable with parameters (k,p). Since Z_k is a Poisson-Binomial random variable with parameters $\eta_1,\ldots,\eta_k$, the probability $\mathbb{P}(Z_k\notin[a,b]\,|\,\mathbf{x})$ is a *linear* function in each individual η_i, so that $\mathbb{E}_{\mathbf{x}}[\mathbb{P}(Z_k\notin[a,b]\,|\,\mathbf{x})]=Q_k([a,b]^c)$. We now proceed to bound the first expectation in (5). By Cauchy-Schwartz inequality we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\sum_{s\in[a,b]}|\,\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x})|\right] \\
&\quad\le\mathbb{E}_{\mathbf{x}}\left[\sqrt{(b-a)\sum_{s\in[a,b]}(\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x}))^2}\right] \\
&\quad\le\mathbb{E}_{\mathbf{x}}\left[\sqrt{(b-a)\sum_{s=0}^{k+1}(\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x}))^2}\right] \\
&\quad\le\sqrt{\mathbb{E}_{\mathbf{x}}\left[(b-a)\sum_{s=0}^{k+1}(\mathbb{P}(Z_k=s\,|\,\mathbf{x})-\mathbb{P}(Z_k=s-1\,|\,\mathbf{x}))^2\right]}\,,
\end{aligned}
$$

where the last inequality holds by Jensen's inequality. Let

$$A = \mathbb{E}[\eta_1^2 + (1-\eta_1)^2] \qquad \text{and} \qquad B = 2\,\mathbb{E}[\eta_1(1-\eta_1)]\ .$$

By Lemma A.3 below we have that the above term is bounded by

$$\sqrt{(b-a)\left(A^k + \frac{\pi}{(4Bk)^{3/2}} + \frac{\pi}{eBk^2}\right)}\ ,$$

so that (5) gives

$$\mathbb{E}_{\mathbf{x}}\left[\sum_{s=0}^{k+1} |\,\mathbb{P}(Z_k = s\,|\,\mathbf{x}) - \mathbb{P}(Z_k = s-1\,|\,\mathbf{x})|\right] \le 2Q_k([a,b]^c) + \sqrt{(b-a)\left(A^k + \frac{\pi}{(4Bk)^{3/2}} + \frac{\pi}{eBk^2}\right)}\ .$$

Let $a = \max\{kp - \sqrt{k}c_k, 0\}$ and $b = \min\{kp + \sqrt{k}c_k, 1\}$. By Bernstein's inequality applied to binomial random variables we have that $Q_k([a,b]^c) = \frac{1}{2k}$. Hence, with this choice of a and b we obtain

$$\mathbb{E}_{\mathbf{x}}\left[\sum_{s=0}^{k+1} |\,\mathbb{P}(Z_k = s\,|\,\mathbf{x}) - \mathbb{P}(Z_k = s-1\,|\,\mathbf{x})|\right] \le \frac{1}{k} + \sqrt{2\sqrt{k}c_k}\sqrt{A^k + \frac{\pi}{(4Bk)^{3/2}} + \frac{\pi}{eBk^2}}\ .$$

The lemma follows by replacing the values of A and B. □

Lemma A.3. *The following inequality holds*

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\sum_{s=0}^{k+1} (\mathbb{P}(Z_k = s\,|\,\mathbf{x}) - \mathbb{P}(Z_k = s-1\,|\,\mathbf{x}))^2\right] \\
&\quad \le \mathbb{E}[\eta_1^2 + (1-\eta_1)^2]^k + \frac{\pi}{(8\,\mathbb{E}[\eta_1(1-\eta_1)]k)^{3/2}} + \frac{\pi}{2e\,\mathbb{E}[\eta_1(1-\eta_1)]k^2}\ .
\end{aligned}$$

Proof. Let $p_s(\mathbf{x}) = \mathbb{P}(Z_k = s\,|\,\mathbf{x}) - \mathbb{P}(Z_k = s-1\,|\,\mathbf{x})$, for $s = 0, \dots, k+1$. Further, for $k+1 \ge u \ge 0$ let $g_u(\mathbf{x}) = \frac{1}{\sqrt{k+2}} \sum_{s=0}^{k+1} p_s(\mathbf{x}) e^{2\pi i \frac{us}{k+2}}$ denote the discrete Fourier transform. Since, for any $\mathbf{x}$, the mapping

$$\mathbf{p}(\mathbf{x}) := (p_0(\mathbf{x}), \dots, p_{k+1}(\mathbf{x})) \mapsto (g_1(\mathbf{x}), \dots, g_{k+1}(\mathbf{x}) := \mathbf{g}(\mathbf{x})$$

is a unitary linear transformation [5] we can write:

$$\sum_{s=0}^{k+1} (\mathbb{P}(Z_k = s\,|\,\mathbf{x}) - \mathbb{P}(Z_k = s-1\,|\,\mathbf{x}))^2 = \|\mathbf{p}(\mathbf{x})\|^2 = \|\mathbf{g}(\mathbf{x})\|^2 = \sum_{u=0}^{k+1} |g_u(\mathbf{x})|^2.$$

Moreover, by Lemma A.4 below we have that

$$g_u(\mathbf{x}) = (1 - e^{\frac{2\pi i u}{k+2}}) \frac{1}{\sqrt{k+2}} \prod_{j=1}^{k} (1 - \eta_j + \eta_j e^{\frac{2\pi i u}{k+2}})$$

and therefore

$$|g_u(\mathbf{x})|^2 = g_u(\mathbf{x})\overline{g_u(\mathbf{x})} = \frac{1}{k+2}\left(1 - \cos\frac{2\pi u}{k+2}\right) \prod_{j=1}^{k} \left((1-\eta_j)^2 + \eta_j^2 + 2\eta_j(1-\eta_j)\cos\frac{2\pi u}{k+2}\right)\ .$$

Therefore we can write

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\sum_{s=0}^{k+1} (\mathbb{P}(Z_k = s\,|\,\mathbf{x}) - \mathbb{P}(Z_k = s-1\,|\,\mathbf{x}))^2\right] \\
&\quad = \frac{1}{k+2}\mathbb{E}_{\mathbf{x}}\left[\sum_{u=0}^{k+1}\left(1 - \cos\frac{2\pi u}{k+2}\right) \prod_{j=1}^{k} \left((1-\eta_j)^2 + \eta_j^2 + 2\eta_j(1-\eta_j)\cos\frac{2\pi u}{k+2}\right)\right] \\
&\quad = \frac{1}{k+2}\sum_{u=0}^{k+1}\left(1 - \cos\frac{2\pi u}{k+2}\right) \prod_{j=1}^{k} \mathbb{E}_{\mathbf{x}}\left[\left((1-\eta_j)^2 + \eta_j^2 + 2\eta_j(1-\eta_j)\cos\frac{2\pi u}{k+2}\right)\right]\ , \qquad (6)
\end{aligned}$$

Where we have used the fact that the random variables η_j are independent. Finally by linearity of expectation and the fact that η_j is distributed as η_1 for all j, we have

$$\mathbb{E}_{\mathbf{x}}\left[\sum_{s=0}^{k+1}(\mathbb{P}(Z_k = s \,|\, \mathbf{x}) - \mathbb{P}(Z_k = s-1 \,|\, \mathbf{x}))^2\right]$$

$$\frac{1}{k+2}\sum_{u=0}^{k+1}\left(1-\cos\frac{2\pi u}{k+2}\right)\left(\mathbb{E}[\eta_1^2+(1-\eta_1)^2]+2\,\mathbb{E}[\eta_1(1-\eta_1)]\cos\frac{2\pi u}{k+2}\right)^k . \tag{7}$$

Applying Proposition A.5 below with $a = \mathbb{E}[\eta_1^2 + (1-\eta_1)^2]$ and $b = 2\,\mathbb{E}[\eta_1(1-\eta_1)]$ we have that the above expression is bounded by

$$\mathbb{E}[\eta_1^2+(1-\eta_1)^2]^k + \frac{\pi}{(8\,\mathbb{E}[\eta_1(1-\eta_1)]k)^{3/2}} + \frac{\pi}{2e\,\mathbb{E}[\eta_1(1-\eta_1)]k^2}$$

which gives the claimed result. □

Lemma A.4. *Let* $p_s(\mathbf{x}) = \mathbb{P}(Z_k = s \,|\, \mathbf{x}) - \mathbb{P}(Z_k = s-1 \,|\, \mathbf{x})$ *and let* $g_u(\mathbf{x}) = \frac{1}{\sqrt{k+2}}\sum_{s=0}^{k+1} p_s(\mathbf{x})e^{\frac{2\pi ius}{k+2}}$ *denote the discrete Fourier transform of* $\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), \ldots, p_s(\mathbf{x}))$*. Then*

$$g_u(\mathbf{x}) = \frac{1}{\sqrt{k+2}}(1-e^{\frac{2\pi iu}{k+2}})\prod_{j=1}^{k}(1-\eta_j+\eta_j e^{\frac{2\pi iu}{k+2}})$$

Proof. By definition of $g_u(\mathbf{x})$ we have:

$$\begin{aligned}
&\frac{1}{\sqrt{k+2}}\left(\sum_{s=0}^{k+1}e^{\frac{2\pi ius}{k+2}}\,\mathbb{P}(Z_k = s \,|\, \mathbf{x}) - \sum_{s=0}^{k+1}e^{\frac{2\pi ius}{k+2}}\,\mathbb{P}(Z_k = s-1 \,|\, \mathbf{x})\right)\\
&\quad= \frac{1}{\sqrt{k+2}}\left(\sum_{s=0}^{k+1}e^{\frac{2\pi ius}{k+2}}\,\mathbb{P}(Z_k = s \,|\, \mathbf{x}) - e^{\frac{2\pi iu}{k+2}}\sum_{s=0}^{k+1}e^{\frac{2\pi iu(s-1)}{k+2}}\,\mathbb{P}(Z_k = s-1 \,|\, \mathbf{x})\right)\\
&\quad= \frac{1}{\sqrt{k+2}}(1-e^{\frac{2\pi iu}{k+2}})\mathbb{E}_{Z_k}[e^{\frac{2\pi iu}{k+2}Z_k} \,|\, \mathbf{x}]\\
&\quad= \frac{1}{\sqrt{k+2}}(1-e^{\frac{2\pi iu}{k+2}})\phi_{Z_k \,|\, \mathbf{x}}\left(\frac{2\pi u}{k+2}\right)
\end{aligned}$$

where $\phi_{Z_k \,|\, \mathbf{x}}$ denotes the characteristic function of Z_k conditioned on $\mathbf{x}$. Using the fact that $Z_k = \sum_{j=1}^{k} A_j$ and that $A_1, \ldots, A_k$ are independent given $\mathbf{x}$, we have $\phi_{Z_k \,|\, \mathbf{x}} = \prod_{j=1}^{k}\phi_{A_j \,|\, x_j}$. The result follows from the fact that A_j is a Bernoulli random variable and therefore $\phi_{A_j \,|\, x_j}(z) = (1-\eta_j+\eta_j e^{iz})$. □

Proposition A.5. *For any* $a, b, k > 0$ *such that* $a + b = 1$ *we have*

$$\frac{1}{k+2}\sum_{u=0}^{k+1}\left(1-\cos\frac{2\pi u}{k+2}\right)\left(a+b\cos\frac{2\pi u}{k+2}\right)^k \leq a^k + \frac{\pi}{(4kb)^{3/2}} + \frac{\pi}{ebk^2}$$

Proof. Using the fact that $\cos\frac{2\pi u}{k+2} \leq 0$ for $u \in [(k+2)/4, 3(k+2)/4]$ we have that

$$\begin{aligned}
&\sum_{u=0}^{k+1} \left(1 - \cos\frac{2\pi u}{k+2}\right)\left(a + b\cos\frac{2\pi u}{k+2}\right)^k \\
&= \sum_{u=0}^{k+2/4} \left(1 - \cos\frac{2\pi u}{k+2}\right)\left(a + b\cos\frac{2\pi u}{k+2}\right)^k \\
&\quad + \sum_{u=(k+2)/4+1}^{3(k+2)/4} \left(1 - \cos\frac{2\pi u}{k+2}\right)\left(a + b\cos\frac{2\pi u}{k+2}\right)^k \\
&\quad + \sum_{u=3(k+2)/4+1}^{k+1} \left(1 - \cos\frac{2\pi u}{k+2}\right)\left(a + b\cos\frac{2\pi u}{k+2}\right)^k \\
&\leq (k+2)a^k + \sum_{u=0}^{(k+2)/4} \left(1 - \cos\frac{2\pi u}{k+2}\right)\left(a + b\cos\frac{2\pi u}{k+2}\right)^k \\
&\quad + \sum_{u=3(k+2)/4+1}^{k+1} \left(1 - \cos\frac{2\pi u}{k+2}\right)\left(a + b\cos\frac{2\pi u}{k+2}\right)^k \\
&= (k+2)a^k + 2\sum_{u=0}^{(k+2)/4} \left(1 - \cos\frac{2\pi u}{k+2}\right)\left(a + b\cos\frac{2\pi u}{k+2}\right)^k \ ,
\end{aligned}$$

where we used the fact that $(1 - \cos t) \leq 2$ for the first inequality and the symmetry of the cosine function for the last equality. We now apply the result of Proposition A.7 to the above expression to see that

$$\begin{aligned}
\sum_{u=0}^{(k+2)/4} \left(1 - \cos\frac{2\pi u}{k+2}\right)\left(a + b\cos\frac{2\pi u}{k+2}\right)^k &\leq 2\pi^2 \sum_{u=0}^{(k+2)/4} \frac{u^2}{(k+2)^2}\left(1 - 4\pi b\frac{u^2}{(k+2)^2})\right)^k \\
&\leq 2\pi^2 \sum_{u=0}^{(k+2)/4} \frac{u^2}{(k+2)^2} e^{-4\pi kb\frac{u^2}{(k+2)^2}}
\end{aligned}$$

Therefore we conclude that

$$\frac{1}{k+2}\sum_{u=0}^{k+1} \left(1 - \cos\frac{2\pi u}{k+2}\right)\left(a + b\cos\frac{2\pi u}{k+2}\right)^k \leq a^k + 4\pi^2 \sum_{u=0}^{(k+2)/4} \frac{u^2}{(k+2)^2} e^{-4\pi kb\frac{u^2}{(k+2)^2}} \frac{1}{k+2} \ .$$

Finally, applying Proposition A.6 with $m = k+2$ and $\alpha = 4\pi kb$ we can upper bound the above quantity by:

$$a^k + 4\pi^2\left(\frac{\sqrt{\pi}}{4(4\pi kb)^{3/2}} + \frac{1}{4e\pi bk(k+2)}\right) \leq a^k + \frac{\pi}{(4kb)^{3/2}} + \frac{\pi}{ebk^2} \ ,$$

which concludes the proof. □

Proposition A.6. *Let* $m > 0$ *and* $\alpha > 0$. *Then*

$$\sum_{u=0}^{m/4} \frac{u^2}{m^2} e^{-\alpha\frac{u^2}{m^2}} \frac{1}{m} \leq \frac{\sqrt{\pi}}{4\alpha^{3/2}} + \frac{1}{e\alpha m}$$

Proof. Let $f\colon \mathbb{R} \to \mathbb{R}$ be given by $x \mapsto x^2 e^{-\alpha x^2}$. Note that the sum we are attempting to bound is then given by:

$$\sum_{u=0}^{m/4} f\left(\frac{u}{m}\right)\frac{1}{m}$$

Note also that f has a maximum at $x_0 = \frac{1}{\sqrt{\alpha}}$. Thus f is increasing for $x < x_0$ and decreasing otherwise. In particular if

$$\sum_{u=0}^{m/4} f\left(\frac{u}{m}\right)\frac{1}{m} = \sum_{u=0}^{\lfloor x_0 \rfloor} f\left(\frac{u}{m}\right)\frac{1}{m} + \sum_{u=\lceil x_0 \rceil}^{m/4} f\left(\frac{u}{m}\right)\frac{1}{m} := L + U,$$

then L corresponds to a lower Riemman sum for f and $L \leq \int_0^{\frac{\lfloor x_0 \rfloor + 1}{m}} f(x)dx$. Similarly U is an upper Riemman sum for f and $U \leq \int_{\frac{\lceil x_0 \rceil - 1}{m}}^{1/4-\frac{1}{m}} f(x)dx$. Therefore we have

$$\begin{aligned}
\sum_{u=0}^{m/4} f\left(\frac{u}{m}\right)\frac{1}{m} &\leq \int_0^{1/4-1/m} x^2 e^{-\alpha x^2} dx + \int_{\frac{\lceil x_0 \rceil - 1}{m}}^{\frac{\lfloor x_0 \rfloor + 1}{m}} f(x)dx \\
&\leq \int_0^{\infty} x^2 e^{-\alpha x^2} + \frac{1}{m}\max_x f(x) \\
&= \frac{\sqrt{\pi}}{4\alpha^{3/2}} + \frac{1}{e\alpha m}\ ,
\end{aligned}$$

as claimed. □

Proposition A.7. *The following inequality holds for any $t \in [0, 1/4]$:*

$$1 - 2(\pi t)^2 \leq \cos 2\pi t \leq 1 - 4\pi t^2$$

Proof. For the lower bound we start from the fact that for any $x \geq 0$ it is well known that

$$\sin 2\pi x \leq 2\pi x.$$

Integrating this inequality from $[0, t]$ we have that $\int_0^t \sin 2\pi x \leq \pi t^2$. Since $\int_0^t \sin 2\pi x = \frac{1}{2\pi}(1 - \cos 2\pi t)$ the lower bound follows.

For the upper bound we proceed in a similar fashion. By the fact that $\sin 2\pi x$ is concave for $x \in [0, 1/4]$ we have that

$$\sin 2\pi x = \sin 2\pi((1-4x)\cdot 0 + 4x \cdot \frac{1}{4}) \geq (1-4x)\sin 0 + 4x \sin\frac{\pi}{2} = 4x.$$

Again integrating the above inequality from 0 to t we have $\frac{1-\cos 2\pi t}{2\pi} \geq 2t^2$. □

A.5 Proof of Lemma 3.6

Lemma 3.6. *For all data distributions $\mathcal{D}$ and all k,* $\mathrm{Adv}_k(\mathcal{M}_{\mathit{RR}}, \mathcal{D}) = \mathrm{Adv}_1(\mathcal{M}_{\mathit{RR}}, \mathcal{D})$.

Proof. Fix any k and any data distribution $\mathcal{D}$. Then,

$$\begin{aligned}
\mathrm{Adv}_k(\mathcal{M}_{\mathrm{RR}}, \mathcal{D}) = &\mathop{\mathbb{E}}_{\substack{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^k \\ i\sim\mathrm{Uniform}([k])}}\left[\max\{\mathbb{P}(y_i = 1 \mid \mathbf{x}, \mathcal{M}_{\mathrm{RR}}(\mathbf{x},\mathbf{y})), \mathbb{P}(y_i = 0 \mid \mathbf{x}, \mathcal{M}_{\mathrm{RR}}(\mathbf{x},\mathbf{y}))\}\right] \\
&- \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\max\{\eta(x), 1-\eta(x)\}\right]
\end{aligned} \tag{8}$$

The second term already has no dependence on k. We focus on the first term.

The key to this proof is that a noisy label $RR(y_i) = \tilde{y}_i$ is independent of the other true labels $y_j, j \neq i$. Thus, we have

$$\Pr(y_i = 1 \mid \mathbf{x}, \mathcal{M}_{\mathrm{RR}}(\mathbf{x},\mathbf{y})) = \Pr(y_i = 1 \mid \mathbf{x}, \mathcal{M}_{\mathrm{RR}}(\mathbf{x},\mathbf{y})_i) = \Pr(y_i = 1 \mid x_i, \mathcal{M}_{\mathrm{RR}}(x_i, y_i)).$$

Applying this to (8), we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^k\\ i\sim\text{Uniform}([k])}}\left[\max\left\{\mathbb{P}(y_i=1\mid\mathbf{x},\mathcal{M}_{\text{RR}}(\mathbf{x},\mathbf{y})),\mathbb{P}(y_i=0\mid\mathbf{x},\mathcal{M}_{\text{RR}}(\mathbf{x},\mathbf{y}))\right\}\right]\\
&=\mathop{\mathbb{E}}_{\substack{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^k\\ i\sim\text{Uniform}([k])}}\left[\max\left\{\mathbb{P}(y_i=1\mid x_i,\mathcal{M}_{\text{RR}}(x_i,y_i)),\mathbb{P}(y_i=0\mid x_i,\mathcal{M}_{\text{RR}}(x_i,y_i))\right\}\right]\\
&=\frac{1}{k}\sum_{i=1}^{k}\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^k}\left[\max\left\{\mathbb{P}(y_i=1\mid x_i,\mathcal{M}_{\text{RR}}(x_i,y_i)),\mathbb{P}(y_i=0\mid x_i,\mathcal{M}_{\text{RR}}(x_i,y_i))\right\}\right]\\
&=\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\max\left\{\mathbb{P}(y_1=1\mid x_1,\mathcal{M}_{\text{RR}}(x_1,y_1)),\mathbb{P}(y_1=0\mid x_1,\mathcal{M}_{\text{RR}}(x_1,y_1))\right\}\right],
\end{aligned}$$

where the final equality holds because the k features and labels are identically distributed. Plugging this back into (8) gives us $\text{Adv}_1(\mathcal{M}_{\text{RR}},\mathcal{D})$ as desired. □

A.6 Proof of Theorem 3.7

Theorem 3.7. *For any data distribution $\mathcal{D}$, the attack advantage for randomized response with privacy parameter $\pi=\frac{1}{1+e^{\varepsilon}}$ is*

$$\text{Adv}_1(\mathcal{M}_{RR},\mathcal{D})=\mathbb{E}_x\Big[\big(\min\{\eta(x),1-\eta(x)\}-\pi\big)\cdot\mathbb{I}\{\eta(x)\in[\pi,1-\pi]\}\Big]\,.$$

The optimal adversary that maximizes $\text{EAU}_1(\cdot,\mathcal{M}_{RR},\mathcal{D})$ is given below:

$$\mathcal{A}^*(x,\tilde{y})=\begin{cases}1, & \textit{if } \eta(x)>1-\pi\\ 0, & \textit{if } \eta(x)<\pi\\ \tilde{y}, & \textit{otherwise}\end{cases}\,.$$

Proof. First we characterize the optimal attacker for which the advantage is maximal. The Bayes optimal decision which minimizes the loss $\mathbb{I}\{A(x,\tilde{y})\neq y\}$ conditioned on x and $\tilde{y}$ is

$$A'(x,\tilde{y})=\mathbb{I}\{\Pr(y=1|x,\tilde{y})>\Pr(y=0|x,\tilde{y})\}$$

which can be written as

$$1<\frac{\Pr(y=1|x,\tilde{y})}{\Pr(y=0|x,\tilde{y})}=\frac{\Pr(\tilde{y}|y=1)\Pr(y=1|x)}{\Pr(\tilde{y}|y=0)\Pr(y=0|x)} \tag{9}$$

Assume that $\tilde{y}=1$, in which case (9) becomes

$$\frac{\pi}{1-\pi}<\frac{\eta(x)}{1-\eta(x)} \tag{10}$$

which is true whenever $\eta(x)>\pi$, and on the other hand, if $\eta(x)<\pi$ then (10) does not hold anymore, thus $A'(x,\tilde{y})=0$. The same argument holds for $\tilde{y}=0$ which implies that the optimal attack is.

$$A'(x,\tilde{y})=\begin{cases}1, & \text{if } \eta(x)>1-\pi\\ 0, & \text{if } \eta(x)<\pi\\ \tilde{y}, & \text{otherwise}\end{cases}$$

To compute the reconstruction advantage of optimal attacker A', we may decompose the feature set into parts as

$$G(\pi)=\{x:\eta(x)\in[\pi,1-\pi]\}$$

and its complement set $G^C(\pi)$. It is clear the advantage of the optimal attacker A' restricted to $G^C(\pi)$ is 0 since $A'(x)=A^*(x)$ if $x\in G^C(\pi)$. The advantage of A' for any $x\in G(\pi)$ is

$$(1-\pi)-1+\min\{\eta(x),1-\eta(x)\}=\min\{\eta(x),1-\eta(x)\}-\pi$$

which concludes the proof. □

B Proofs for Section 6

B.1 Proof of Theorem 6.1

Proof. We begin with the square loss $\ell(p, X) = (p - X)^2$. The proof follows from linearity of expectation and that Z is independent of x and y with expectation 0.

$$\begin{aligned}
\mathbb{E}\left[\nabla_\theta \ell\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x), \tilde{\alpha}\right)\right] &= \mathbb{E}\left[\nabla_\theta \left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x) - \alpha - Z\right)^2\right] \\
&= \mathbb{E}\left[\nabla_\theta \left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x) - \alpha\right)^2 - \nabla_\theta 2Z\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x) + \alpha\right) + \nabla_\theta Z^2\right] \\
&= \mathbb{E}\left[\nabla_\theta \ell\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x), \alpha\right)\right] + \nabla_\theta\, \mathbb{E}[-2Z]\cdot\mathbb{E}\left[\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x) + \alpha\right] + \nabla_\theta\, \mathbb{E}[Z^2] \\
&= \mathbb{E}\left[\nabla_\theta \ell\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x), \alpha\right)\right].
\end{aligned}$$

Now, consider the binary cross entropy loss $\ell(p, X) = -X\log(p) - (1 - X)\log(1 - p)$. Using the fact that $\mathbb{E}[Z] = 0$ and Z is independent of $x, \alpha, \mathcal{B}$, we have the following

$$\begin{aligned}
&\mathbb{E}\left[\nabla_\theta \ell\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x), \tilde{\alpha}\right)\right] \\
&= \mathbb{E}\left[\nabla_\theta - \tilde{\alpha}\cdot\log\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right) - (1-\tilde{\alpha})\log\left(1 - \frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right)\right] \\
&= \mathbb{E}\left[\nabla_\theta - (\alpha + Z)\cdot\log\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right) - (1-\alpha-Z)\cdot\log\left(1 - \frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right)\right] \\
&= \mathbb{E}\left[\nabla_\theta - \alpha\cdot\log\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right) - (1-\alpha)\cdot\log\left(1 - \frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right)\right] \\
&\quad + \mathbb{E}\left[\nabla_\theta - Z\cdot\log\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right) + Z\cdot\log\left(1 - \frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right)\right] \\
&= \mathbb{E}\left[\nabla_\theta \ell\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x), \alpha\right)\right] + \mathbb{E}\left[\nabla_\theta - Z\cdot\log\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right) + Z\cdot\log\left(1 - \frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)\right)\right] \\
&= \mathbb{E}\left[\nabla_\theta \ell\left(\frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x), \alpha\right)\right].
\end{aligned}$$

□

B.2 Proof of Theorem 6.2

Proof. Let $\ell(q, X) = -X\log(q) - (1 - X)\log(1 - q)$ be the binary cross entropy loss, and let $q = \frac{1}{k}\sum_{x\in\mathcal{B}} h_\theta(x)$, and since $\tilde{\alpha}$ is an unbiased estimate of α generated using a noise-addition mechanism, $\tilde{\alpha} = \alpha + Z$ for some unbiased Z.

Notice that $\ell(q, \tilde{\alpha}) = \ell(q, \alpha) - Z\log\left(\frac{q}{1-q}\right)$. Thus,

$$\begin{aligned}\mathbb{E}\left[\|\nabla_\theta \ell\left(q,\tilde{\alpha}\right)\|_2^2\right] &= \mathbb{E}\left[\left\|\nabla_\theta \ell(q,\alpha) - \nabla_\theta Z \log\left(\frac{q}{1-q}\right)\right\|^2\right] \\ &= \mathbb{E}\left[\|\nabla_\theta \ell(q,\alpha)\|^2\right] + \mathbb{E}\left[\left\|\nabla_\theta Z \log\left(\frac{q}{1-q}\right)\right\|^2\right] - \mathbb{E}\left[2Z\left\langle \nabla_\theta \ell(q,\alpha), \nabla_\theta \log\left(\frac{q}{1-q}\right)\right\rangle\right] \\ &= \mathbb{E}\left[\|\nabla_\theta \ell(q,\alpha)\|^2\right] + \mathrm{Var}[Z]\cdot \mathbb{E}\left[\left\|\nabla_\theta \log\left(\frac{q}{1-q}\right)\right\|^2\right].\end{aligned}$$

where we have used the fact that $\mathbb{E}[Z] = 0$ and Z is independent of q and α.

Taking $Z \sim Lap(1/k\varepsilon)$ gives the desired upper bound. □

C Computing Advantage for DP + LLP

In this section, we derive the optimal LIA attacker for a PET $\mathcal{M}_{\text{Lap-LLP}}$ that computes label proportions for groups examples in bags of size k (as in $\mathcal{M}_{\text{LLP}}$) and adds Laplace noise with scale $1/k\varepsilon$ to the true binary label proportion. For a given bag, $\mathcal{M}_{\text{Lap-LLP}}(\mathbf{x},\mathbf{y}) = Z + \frac{1}{k}\sum_{i=1}^{k} y_i$ where $Z \sim \text{Lap}(1/k\varepsilon)$.

Theorem 6.3. *For any privacy parameters $k \geq 1$, $\varepsilon > 0$, and any data distribution $\mathcal{D}$, $\mathrm{Adv}_k(\mathcal{M}_{\textit{Lap-LLP}}, \mathcal{D})$ is maximized by the following adversary:*

$$\mathcal{A}^*(\mathbf{x}, \mathcal{M}_{\textit{Lap-LLP}}(\mathbf{x},\mathbf{y}))_i := \begin{cases} 1 & \textit{if} \quad \frac{\eta(\mathbf{x}_i)}{1-\eta(\mathbf{x}_i)} \cdot \frac{\sum_{b=1}^{k} \mathbb{P}(\textit{PBin}(\{\eta(x_j)\}_{i\neq j})=b-1)\cdot e^{-k\varepsilon|\tilde{\alpha}-b/k|}}{\sum_{b=0}^{k-1} \mathbb{P}(\textit{PBin}(\{\eta(x_j)\}_{i\neq j})=b)\cdot e^{-k\varepsilon|\tilde{\alpha}-b/k|}} \geq 1 \\ 0 & \textit{otherwise.}\end{cases}$$

Alternatively, using properties of the Laplace distribution, this adversary is approximately equivalent to the following for small ε

$$\mathcal{A}^{**}(\mathbf{x}, \mathcal{M}_{\textit{Lap-LLP}}(\mathbf{x},\mathbf{y}))_i := \begin{cases} 1 & \textit{if} \quad \frac{\eta(x_i)}{1-\eta(x_i)} < e^{-\varepsilon} \\ 0 & \textit{if} \quad \frac{\eta(x_i)}{1-\eta(x_i)} > e^{\varepsilon} \\ y_i & \textit{otherwise}\end{cases}.$$

Proof. Fix any k, any $\varepsilon > 0$, and any data distribution $\mathcal{D}$. By Lemma 3.2, the $\mathrm{EAU}_k(\mathcal{A}_{\text{informed}}, \mathcal{M}_{\text{Lap-LLP}}, \mathcal{D})$ is maximized by the following adversary

$$\mathcal{A}^*(\mathbf{x}, \mathcal{M}(\mathbf{x},\mathbf{y}))_i := \begin{cases} 1 & \text{if } \mathbb{P}(y_i = 1 \mid \mathbf{x}, \mathcal{M}_{\text{Lap-LLP}}(\mathbf{x},\mathbf{y})) \geq 1/2 \\ 0 & \text{otherwise.}\end{cases}$$

In particular, we denote the output $\mathcal{M}_{\text{Lap-LLP}}(\mathbf{x},\mathbf{y})$ as $\tilde{\alpha} = \alpha + Z$ where $Z \sim \text{Lap}(1/k\varepsilon)$. We now characterize the conditional distribution that the attacker computes. Let $c \in \{0,1\}$.

$$\begin{aligned}\mathbb{P}(y_i = c \mid \mathbf{x}, \mathcal{M}_{\text{Lap-LLP}}(\mathbf{x},\mathbf{y})) &= \mathbb{P}(y_i = c \mid \mathbf{x}, \tilde{\alpha}) && (11)\\ &= \frac{\mathbb{P}(\tilde{\alpha} \mid \mathbf{x}, y_i = c)\,\mathbb{P}(y_i = c \mid \mathbf{x})}{\mathbb{P}(\tilde{\alpha} \mid \mathbf{x})} && (12)\end{aligned}$$

Now, when we take the ratio for $c = 1$ and $c = 0$, the term $\mathbb{P}(\tilde{\alpha} \mid \mathbf{x})$ cancels, so we do not need to compute it.

First, observe that

$$\mathbb{P}(y_i = c \mid \mathbf{x}) = \mathbb{P}(y_i = c \mid x_i) = c\cdot\eta(\mathbf{x}_i) + (1-c)\cdot(1-\eta(\mathbf{x}_i)).$$

Now, let's analyze $\mathbb{P}(\tilde{\alpha} \mid \mathbf{x}, y_i = c)$. Fix any a. We will use the fact that Z is independent of $\alpha, \mathbf{x}$, and $\mathbf{y}$ and apply the density function of the Laplace distribution.

$$\begin{aligned}
\mathbb{P}(\tilde{\alpha} = a \mid \mathbf{x}, y_i = c) &= \mathbb{P}(\alpha + Z = a \mid \mathbf{x}, y_i = c) && (13)\\
&= \sum_{b=c}^{k-1+c} \mathbb{P}\left(\alpha = \frac{b}{k}, Z = a - \frac{b}{k} \mid \mathbf{x}, y_i = c\right) && (14)\\
&= \sum_{b=c}^{k-1+c} \mathbb{P}\left(\alpha = \frac{b}{k} \mid \mathbf{x}, y_i = c\right) \cdot \mathbb{P}\left(Z = a - \frac{b}{k}\right) && (15)\\
&= \sum_{b=c}^{k-1+c} \mathbb{P}\left(\alpha = \frac{b}{k} \mid \mathbf{x}, y_i = c\right) \cdot \frac{k\varepsilon}{2} e^{-k\varepsilon|a-b/k|} && (16)\\
& && (17)
\end{aligned}$$

Now, we will analyze the first part of the summation.

$$\begin{aligned}
\mathbb{P}\left(\alpha = \frac{b}{k} \mid \mathbf{x}, y_i = c\right) &= \mathbb{P}\left(\sum_{j \in [k]} y_j = b \mid \mathbf{x}, y_i = c\right)\\
&= \mathbb{P}\left(\sum_{j \in [k], j \neq i} y_j = b - c \mid \mathbf{x}\right)\\
&= \mathbb{P}(\text{PBin}(\{\eta(x_j)\}_{i \neq j}) = b - c)
\end{aligned}$$

Plugging this back into (16), we have

$$\mathbb{P}(\tilde{\alpha} = a \mid \mathbf{x}, y_i = c) = \sum_{b=c}^{k-1+c} \mathbb{P}(\text{PBin}(\{\eta(x_j)\}_{i \neq j}) = b - c) \cdot \frac{k\varepsilon}{2} e^{-k\varepsilon|a-b/k|} \tag{18}$$

For simplicity of notation, let $f(b) = \mathbb{P}(\text{PBin}(\{\eta(x_j)\}_{i \neq j}) = b)$ and let $g(b) = \frac{k\varepsilon}{2} e^{-k\varepsilon|a-b/k|}$. Note that because g is the density function for $\text{Lap}(1/k\varepsilon)$, we have that $g(b) \leq e^{\varepsilon} \cdot g(b+1)$ for all $b \in \mathbb{R}$.

Then,

$$\begin{aligned}
\mathbb{P}(\tilde{\alpha} = a \mid \mathbf{x}, y_i = 0) = \sum_{b=0}^{k-1} f(b) \cdot g(b) &\leq \sum_{b=0}^{k-1} f(b) \cdot e^{\varepsilon} \cdot g(b+1)\\
&= e^{\varepsilon} \cdot \sum_{b=1}^{k} f(b-1) \cdot g(b)\\
&= e^{\varepsilon} \cdot \mathbb{P}(\tilde{\alpha} = a \mid \mathbf{x}, y_i = 1).
\end{aligned}$$

Similarly,

$$\mathbb{P}(\tilde{\alpha} = a \mid \mathbf{x}, y_i = 1) \leq e^{\varepsilon} \cdot \mathbb{P}(\tilde{\alpha} = a \mid \mathbf{x}, y_i = 0).$$

Now, plugging everything back into (12), and computing the proportion between $y_i = 0$ and $y_i = 1$, we have

$$\frac{\mathbb{P}(y_i = 0 \mid \mathbf{x}, \mathcal{M}_{\text{Lap-LLP}}(\mathbf{x}, \mathbf{y}))}{\mathbb{P}(y_i = 1 \mid \mathbf{x}, \mathcal{M}_{\text{Lap-LLP}}(\mathbf{x}, \mathbf{y}))} = \frac{\mathbb{P}(\tilde{\alpha} \mid \mathbf{x}, y_i = 0)}{\mathbb{P}(\tilde{\alpha} \mid \mathbf{x}, y_i = 1)} \cdot \frac{\mathbb{P}(y_i = 0 \mid \mathbf{x})}{\mathbb{P}(y_i = 1 \mid \mathbf{x})} \leq e^{\varepsilon} \cdot \frac{\eta(x_i)}{1 - \eta(x_i)} \tag{19}$$

Similarly,

$$e^{-\varepsilon} \cdot \frac{\eta(x_i)}{1 - \eta(x_i)} \leq \frac{\mathbb{P}(y_i = 0 \mid \mathbf{x}, \mathcal{M}_{\text{Lap-LLP}}(\mathbf{x}, \mathbf{y}))}{\mathbb{P}(y_i = 1 \mid \mathbf{x}, \mathcal{M}_{\text{Lap-LLP}}(\mathbf{x}, \mathbf{y}))} \leq e^{\varepsilon} \cdot \frac{\eta(x_i)}{1 - \eta(x_i)} \tag{20}$$

Thus, if $\frac{\eta(x_i)}{1-\eta(x_i)} < e^{-\varepsilon}$, then the ratio is greater than 1 and the algorithm outputs $\hat{y}_i = 0$. On the other hand, if $\frac{\eta(x_i)}{1-\eta(x_i)} > e^{\varepsilon}$, then the ratio is less than 1 and the algorithm outputs $\hat{y}_i = 1$. When $e^{-\varepsilon} \le \frac{\eta(x_i)}{1-\eta(x_i)} \le e^{\varepsilon}$, the bounds are not tight enough to determine whether the ratio is greater than or less than 1.

Plugging (18) into (20) gives the optimal adversary.

□

C.1 Bound on Adv

Theorem 6.4. *Let $\mathcal{D}$ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and let $p = \mathbb{E}[\eta(x)]$. Then for all bag sizes $k \ge 1$ we have:*

$$\begin{aligned}&\mathrm{Adv}_k(\mathcal{M}_{\textit{Lap-LLP}}, \mathcal{D})\\ &\le \min\left\{2(1-e^{-\epsilon})\,\mathbb{E}[\eta(x)(1-\eta(x)], \widetilde{O}\left(\frac{\mathbb{E}[\eta(x)(1-\eta(x)]^{1/4}(p(1-p))^{1/4}}{\sqrt{k}} + \frac{\mathbb{E}[\eta(x)(1-\eta(x))]^{1/4}}{k}\right)\right\}\ .\end{aligned}$$

Proof. With the optimal adversary, $\mathrm{Adv}_k(\mathcal{M}_{\text{Lap-LLP}}, \mathcal{D})$ can be upper bounded via (3) as

$$\mathrm{Adv}_k(\mathcal{M}_{\text{Lap-LLP}}, \mathcal{D}) \le \mathbb{E}[|\eta(x_1) - \mathbb{P}(y_1 = 1 \mid \mathbf{x}, \tilde{\alpha})|]\ .$$

We can write

$$\begin{aligned}\Pr(y_1 = 1 \,|\, \mathbf{x}, \tilde{\alpha} = a) &= \frac{\Pr(y_1 = 1, \tilde{\alpha} = a \,|\, \mathbf{x})}{\Pr(\tilde{\alpha} = a \,|\, \mathbf{x})}\\ &= \frac{\eta(x_1)\Pr(\tilde{\alpha} = a \,|\, y_1 = 1, \mathbf{x})}{\eta(x_1)\Pr(\tilde{\alpha} = a \,|\, y_1 = 1, \mathbf{x}) + (1-\eta(x_1))\Pr(\tilde{\alpha} = a \,|\, y_1 = 0, \mathbf{x})}\end{aligned}$$

so that

$$\begin{aligned}\mathrm{Adv}_k(\mathcal{M}_{\text{Lap-LLP}}, \mathcal{D}) &\le \mathbb{E}\left[\eta(x_1)(1-\eta(x_1))\frac{|\Pr(\tilde{\alpha} = a \,|\, y_1 = 0, \mathbf{x}) - \Pr(\tilde{\alpha} = a \,|\, y_1 = 1, \mathbf{x})|}{\eta(x_1)\Pr(\tilde{\alpha} = a \,|\, y_1 = 1, \mathbf{x}) + (1-\eta(x_1))\Pr(\tilde{\alpha} = a \,|\, y_1 = 0, \mathbf{x})}\right]\\ &= \mathbb{E}\left[\eta(x_1)(1-\eta(x_1))\frac{|\Pr(\tilde{\alpha} = a \,|\, y_1 = 0, \mathbf{x}) - \Pr(\tilde{\alpha} = a \,|\, y_1 = 1, \mathbf{x})|}{\Pr(\tilde{\alpha} \,|\, \mathbf{x})}\right]\ .\end{aligned}$$

Now, conditioned on $\mathbf{x}$, let $B_{k-1}(\mathbf{x})$ the Poisson-Binomial random variable with parameters $\{\eta(x_2), \ldots, \eta(x_k)\}$. Also, denote by $f_Z(z) = \frac{\epsilon k}{2}e^{-\epsilon k|z|}$ the (Laplace) density of Z. We can write

$$\begin{aligned}\Pr(\tilde{\alpha} = a \,|\, y_1 = 1, \mathbf{x}) &= \sum_{b=0}^{k}\Pr(\alpha = b/k \,|\, y_1 = 1, \mathbf{x})\ \Pr(Z = a - b/k)\\ &= \sum_{b=0}^{k-1}\Pr(B_{k-1}(\mathbf{x}) = b \,|\, \mathbf{x})\ \Pr(Z = a - (b+1)/k)\\ &= \mathbb{E}_{B_{k-1}(\mathbf{x})}\left[f_Z\left(a - \frac{B_{k-1}(\mathbf{x}) + 1}{k}\right)\right]\end{aligned}$$

And similarly,

$$\Pr(\tilde{\alpha} = a \,|\, y_1 = 0, \mathbf{x}) = \mathbb{E}_{B_{k-1}(\mathbf{x})}\left[f_Z\left(a - \frac{B_{k-1}(\mathbf{x})}{k}\right)\right]\ .$$

Thus

$$\begin{aligned}\mathrm{Adv}_k(\mathcal{M}_{\text{Lap-LLP}}, \mathcal{D}) &\le \mathbb{E}_{\tilde{\alpha},\mathbf{x}}\left[\eta(x_1)(1-\eta(x_1))\frac{\left|\mathbb{E}_{B_{k-1}(\mathbf{x})}\left[f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})}{k}\right)\right] - \mathbb{E}_{B_{k-1}(\mathbf{x})}\left[f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})+1}{k}\right)\right]\right|}{\Pr(\tilde{\alpha} \,|\, \mathbf{x})}\right]\\ &= \mathbb{E}[\eta(x)(1-\eta(x))]\\ &\quad \times \mathbb{E}_{\mathbf{x}}\left[\int_{-\infty}^{+\infty}\left|\mathbb{E}_{B_{k-1}(\mathbf{x})}\left[f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})}{k}\right) - f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x}) + 1}{k}\right)\right]\right|d\tilde{\alpha}\,\middle|\,\mathbf{x}\right]\end{aligned} \tag{21}$$

In the case where the features $\mathbf{x}$ are absent, the above can be simplified as

$$\mathrm{Adv}_k(\mathcal{M}_{\text{Lap-LLP}}, \mathcal{D}) \leq p(1-p) \int_{-\infty}^{+\infty} \left| \mathbb{E}_{B_{k-1}} \left[f_Z\left(\tilde{\alpha} - \frac{B_{k-1}}{k}\right) - f_Z\left(\tilde{\alpha} - \frac{B_{k-1}+1}{k}\right) \right] \right| d\tilde{\alpha} \,,$$

where B_{k-1} is a binomial random variable with parameters p and $k-1$.

We consider two ways of upper bounding (21). First, observe that

$$\frac{|f_Z(z) - f_Z(z-1/k)|}{\max\{f_Z(z), f_Z(z-1/k)\}} \leq 1 - e^{-\epsilon}$$

holds for every $k \geq 1$, $\epsilon \geq 0$, and $z \in \mathbb{R}$. Hence, for all $\mathbf{x}$,

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \left| \mathbb{E}_{B_{k-1}(\mathbf{x})} \left[f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})}{k}\right) - f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})+1}{k}\right) \right] \right| d\tilde{\alpha} \\
&\quad \leq \int_{-\infty}^{+\infty} \mathbb{E}_{B_{k-1}(\mathbf{x})} \left[\left| f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})}{k}\right) - f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})+1}{k}\right) \right| \right] d\tilde{\alpha} \\
&\quad = \mathbb{E}_{B_{k-1}(\mathbf{x})} \left[\int_{-\infty}^{+\infty} \left| f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})}{k}\right) - f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})+1}{k}\right) \right| d\tilde{\alpha} \right] \\
&\quad \leq (1-e^{-\epsilon}) \mathbb{E}_{B_{k-1}(\mathbf{x})} \left[\int_{-\infty}^{+\infty} \max\left\{ f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})}{k}\right), f_Z\left(\tilde{\alpha} - \frac{B_{k-1}(\mathbf{x})+1}{k}\right) \right\} d\tilde{\alpha} \right] \\
&\quad = (1-e^{-\epsilon}) \int_{-\infty}^{+\infty} \max\left\{ f_Z\left(\tilde{\alpha}\right), f_Z\left(\tilde{\alpha} - \frac{1}{k}\right) \right\} d\tilde{\alpha} \\
&\quad \leq (1-e^{-\epsilon}) \int_{-\infty}^{+\infty} \left(f_Z\left(\tilde{\alpha}\right) + f_Z\left(\tilde{\alpha} - \frac{1}{k}\right) \right) d\tilde{\alpha} \\
&\quad = 2(1-e^{-\epsilon}) \,.
\end{aligned}$$

Plugging back into (21), this allows us to conclude that

$$\mathrm{Adv}_k(\mathcal{M}_{\text{Lap-LLP}}, \mathcal{D}) \leq 2(1-e^{-\epsilon})\, \mathbb{E}[\eta(x)(1-\eta(x))] \,.$$

Next, we argue that the advantage of $\mathcal{M}_{\text{Lap-LLP}}$ never exceeds that of $\mathcal{M}_{\text{LLP}}$ (with the same value of k, the bag size). This is because given the output of $\mathcal{M}_{\text{LLP}}$, we can simulate the output of $\mathcal{M}_{\text{Lap-LLP}}$ by adding $\mathrm{Lap}(\frac{1}{k\epsilon})$ noise. In particular, any adversary that uses the output of $\mathcal{M}_{\text{Lap-LLP}}$ can be converted into one that uses the output of $\mathcal{M}_{\text{LLP}}$ (by noising the label proportion), and this new adversary has exactly the same attack utility. This implies that the advantage of $\mathcal{M}_{\text{Lap-LLP}}$ with parameters (ϵ, k) is at most the advantage of $\mathcal{M}_{\text{LLP}}$ with parameter k, for any $\epsilon \geq 0$. Combined with the above calculations, this gives

$$\begin{aligned}
&\mathrm{Adv}_k(\mathcal{M}_{\text{Lap-LLP}}, \mathcal{D}) \\
&\leq \min\left\{ 2(1-e^{-\epsilon})\, \mathbb{E}[\eta(x)(1-\eta(x)], \widetilde{O}\left(\frac{\mathbb{E}[\eta(x)(1-\eta(x)]^{1/4}(p(1-p))^{1/4}}{\sqrt{k}} + \frac{\mathbb{E}[\eta(x)(1-\eta(x))]^{1/4}}{k} \right) \right\} \,,
\end{aligned}$$

which concludes the proof.

□