Context-Aware Generative Adversarial Privacy

Chong Huang∗‡, Peter Kairouz†‡, Xiao Chen†, Lalitha Sankar∗, and Ram Rajagopal†

Abstract

Preserving the utility of published datasets, while providing provable privacy guarantees, is a well-known challenge. On the one hand, context-free privacy solutions, such as differential privacy, provide strong privacy guarantees, but often lead to a significant reduction in utility. On the other hand, context-aware privacy solutions, such as information theoretic privacy, achieve an improved privacy-utility tradeoff, but assume that the data holder has access to the dataset’s statistics. To circumvent this problem, we present a novel context-aware privacy framework called generative adversarial privacy (GAP). GAP leverages recent advancements in generative adversarial networks (GANs) to allow the data holder to learn “optimal” privatization schemes from the dataset itself. Under GAP, learning the privacy mechanism is formulated as a constrained minimax game between two players: a privatizer that sanitizes the dataset in a way that limits the risk of inference attacks on the individuals’ private variables, and an adversary that tries to infer the private variables from the sanitized dataset. To evaluate GAP’s performance, we investigate two simple (yet canonical) statistical dataset models: (a) the binary data model, and (b) the binary Gaussian mixture model. For both models, we derive game-theoretically optimal minimax privacy mechanisms, and show that the privacy mechanisms learned from data (in an iterative generative adversarial fashion) match the theoretically optimal ones. This demonstrates that our framework can be easily applied in practice, even in the absence of dataset statistics.

Keywords- Generative Adversarial Privacy; Generative Adversarial Networks; Privatizer Network; Adversarial Network; Statistical Data Privacy; Differential Privacy; Information Theoretic Privacy; Mutual Information Privacy; Error Probability Games; Machine Learning

1 Introduction

The explosion of information collection across a variety of electronic platforms is enabling the use of inferential machine learning (ML) and artificial intelligence to guide consumers through a myriad of choices and decisions in their daily lives. In this era of artificial intelligence, data is quickly becoming the most valuable resource [22]. Indeed, large scale datasets provide tremendous utility in helping researchers design state-of-the-art machine learning algorithms that can learn from and make predictions on real life data. Scholars and researchers are increasingly demanding access to larger datasets that allow them to learn more sophisticated models. Unfortunately, more often than not, in addition to containing public information that can be published, large scale datasets also contain confidential information about participating individuals. Thus, data collection and curation organizations are reluctant to release such datasets before carefully sanitizing them, especially in light of recent public policies on data sharing [24, 61] (see Figure 1).

To protect the privacy of individuals, datasets are typically anonymized before their release. This is done by stripping off personally identifiable information (e.g., first and last name, social security number, IDs, etc.) [11, 57, 65]. Anonymization, however, does not provide immunity against correlation and linkage attacks [30, 50]. Indeed, several successful attempts to re-identify individuals from anonymized datasets have been reported in the past ten years. For instance, [50] were able to successfully de-anonymize watch histories in the Netflix Prize, a public recommender system competition. In a more recent attack, [60] showed that participants of an anonymized DNA study were identified by linking their DNA data with the publicly available Personal Genome Project

∗C. Huang and L. Sankar are with the School of Electrical, Computer, and Energy Engineering at Arizona State University, Tempe, AZ
†P. Kairouz, X. Chen, and R. Rajagopal are with the Department of Electrical Engineering at Stanford University, Stanford, CA
‡Equal contributions
Addressing the shortcomings of anonymization techniques requires data randomization. In the recent years, two randomization-based approaches with provable statistical privacy guarantees have emerged: (a) context-free approaches that assume worst-case dataset statistics and adversaries; (b) context-aware approaches that explicitly model the dataset’s statistics and adversary’s capabilities.

**Context-free privacy.** One of the most popular context-free notions of privacy is differential privacy (DP) \[19, 20, 21\]. DP, quantified by a leakage parameter \(\epsilon\), restricts distinguishability between any two “neighboring” datasets from the published data. DP provides strong, context-free theoretical guarantees against worst-case adversaries. Thus, training machine learning models on randomized data with DP guarantees often leads to a significantly reduced utility and comes with a tremendous hit in sample complexity \[16, 17, 18, 25, 31, 35, 36, 40, 53, 70, 74, 80, 81, 82\] in the desired leakage regimes. For example, learning population level histograms under local DP suffers from a stupendous increase in sample complexity by a factor proportional to the size of the dictionary \[18, 35, 36\].

**Context-aware privacy.** Context-aware privacy notions have been so far studied by information theorists under the rubric of information theoretic (IT) privacy \[3, 4, 5, 7, 9, 10, 11, 12, 13, 37, 38, 39, 42, 44, 46, 54, 56, 58, 59, 60, 71, 79\]. IT privacy has predominantly been quantified by mutual information (MI) which models how well an adversary, with access to the released data, can refine their belief about a number of private features of the data. Recently, Isa et al. introduced maximal leakage (MaxL) to quantify leakage to a strong adversary capable of guessing any function of the dataset \[44\]. More recently, they showed that their adversarial model can be generalized to encompass local DP (wherein the mechanism ensures limited distinction for any pair of entries—a stronger guarantee without a neighborhood constraint \[15, 75\] \[83\]. When one restricts the adversary to guessing specific private features (and not all functions of these features), the resulting adversary is a maximum a posteriori (MAP) adversary that has been studied by Assoodeh et al. in \[45, 69, 84\]. Compared to context-free privacy notions, context-aware privacy notions achieve a better privacy-utility tradeoff by incorporating the statistics of the dataset and placing reasonable restrictions on the capabilities of the adversary. However, using information theoretic quantities (such as mutual information) as privacy metrics requires either learning the distribution of the data or learning the parameters of the privatization mechanism in a data driven fashion that involves minimizing an empirical information theoretic loss function. Both tasks are remarkably challenging in practice \[2, 27, 45, 69, 84\].

**Generative adversarial privacy.** Given the challenges of existing privacy approaches, we believe that there is a strong need for a tractable and verifiable privatization method that preserves the utility of published datasets, while still protecting the privacy of individuals as much as possible. As such, we take a fundamentally new approach towards enabling private data publishing with guarantees on both privacy and utility. Instead of adopting worst-case, context-free notions of data privacy (such as differential privacy), we introduce a novel, context-aware model of privacy that allows the designer to cleverly add noise where it matters. An inherent challenge in taking a context-aware privacy approach is that it requires having access to priors such as joint distributions of public and private variables. Such information is hardly ever present in practice. To overcome this issue, we take a data driven approach to context-aware privacy. We leverage recent advancements in generative adversarial networks (GANs) to introduce a unified framework for context-aware privacy.

\footnote{smaller \(\epsilon \in [0,\infty)\) implies smaller leakage and stronger privacy guarantees.}

---

**Figure 1:** An example privacy preserving mechanism for smart meter data

| Entry (row 1) | 0.140 | 0.233 | 70,000 | 1 |
| Entry (row 2) | 0.108 | 0.371 | 60,000 | 3 |
| Entry (row 3) | 0.248 | 0.192 | 200,000 | 4 |
| Entry (row K) | 0.210 | 0.182 | 150,000 | 3 |

**Database D**

| Original meter data \(X\) | Private features \(Y\) | Perturbed meter data \(\tilde{X}\) |
| --- | --- | --- |
| Entry (row 1) | 0.210 | 0.231 | 0.302 |
| Entry (row 2) | 0.182 | 0.158 | 0.350 |
| Entry (row 3) | 0.302 | 0.226 | 0.176 |
| Entry (row K) | 0.179 | 0.192 | 0.202 |
privacy called *generative adversarial privacy* (GAP). Under GAP, the parameters of a generative model, representing the privatization mechanism, are learned from the data itself.

### 1.1 Our Contributions

We investigate a setting where a data holder would like to publish a dataset $D$ in a privacy preserving fashion. Each row in $D$ contains both private variables (represented by $Y$) and public variables (represented by $X$). The goal of the data holder is to generate $\hat{X}$ in a way such that: (a) $\hat{X}$ is as good of a representation of $X$ as possible, and (b) an adversary cannot use $\hat{X}$ to reliably infer $Y$. To this end, we present GAP, a unified framework for context-aware privacy that includes existing information-theoretic privacy notions. Our formulation is inspired by GANs [28, 44, 61] and error probability games [47, 48, 49, 55, 62]. It includes two learning blocks: a *privatizer*, whose task is to output a sanitized version of the public variables (subject to some distortion constraints); and an *adversary*, whose task is to learn the private variables from the sanitized data. The privatizer and adversary achieve their goals by competing in a constrained minimax, zero-sum game. On the one hand, the privatizer (a conditional generative model) is designed to minimize the adversary’s performance in inferring $Y$ reliably. On the other, the adversary (a classifier) seeks to find the best inference strategy that maximizes its performance. This generative adversarial framework is represented in Figure 2.

At the core of GAP is a utility function that captures how well an adversary does in terms of inferring the private variables. Different utility functions lead to different adversarial models. We focus our attention on two types utility functions: (a) a 0-1 utility that leads to a *maximum a posteriori probability* (MAP) adversary, and (b) a negative cross entropy utility that leads to a cross entropy adversary. Ultimately, our goal is to show that our data driven approach can provide privacy guarantees against a MAP adversary. However, derivatives of a 0-1 utility function are ill-defined. To overcome this issue, the ML literature uses the more analytically tractable cross-entropy as a utility/loss function. We do the same by choosing negative cross entropy as the utility function in the data driven framework. We show that it leads to privacy mechanisms that perform as good as game-theoretic optimal mechanisms under a MAP adversary. We also show that GAP recovers mutual information privacy under a log information utility function.

To showcase the power of our context-aware, data driven framework, we investigate two simple, albeit canonical, statistical dataset models: (a) the binary data model, and (b) the binary Gaussian mixture model. Under the binary data model, both $X$ and $Y$ are binary. Under the binary Gaussian mixture model, $Y$ is binary whereas $X$ is conditionally Gaussian. For both models, we derive and compare the performance of game-theoretically optimal privatization mechanisms with those that are directly learned from data (using an iterative generative adversarial fashion).

For the above-mentioned statistical dataset models, we present two approaches towards designing privacy mechanisms: (i) private-data-dependent (PDD) mechanisms, where the privatizer uses both the public and private variables, and (ii) private-data-independent (PDI) mechanisms, where the privatizer only uses the public variables. We show that the PDD mechanisms lead to a superior privacy-utility tradeoff.

### 1.2 Related Work

In practice, a context-free notion of privacy (such as DP) is desirable because it places no restrictions on the dataset’s statistics or adversary’s strength. This explains why DP has been remarkably successful in the past ten years, and has been deployed in array of systems, including Google’s Chrome browser [23] and Apple’s iOS [77]. Nevertheless, because of its strong, context-free nature, DP has suffered from a sequence of impossibility results. These results have made...
the deployment of DP with a reasonable leakage parameter practically impossible. Indeed, it was recently reported that Apple’s DP implementation suffers from several limitations—most notably of which is Apple’s use of unacceptably large leakage parameters [67].

Context-aware privacy notions can exploit the structure and statistics of the dataset to design mechanisms matched to both the data and the worst adversarial model requirements. In this context, information-theoretic metrics for privacy are naturally well suited. In fact, the adversarial model determines the appropriate information metric: an estimating adversary that minimizes mean square error is captured by $\chi^2$-squared measures, a belief refining adversary is captured by MI [59], an adversary that can make a hard MAP decision for a specific set of private features is captured by the Arimoto MI of order $\infty$ [6, 8], and an adversary that can guess any function of the private features (i.e., MaxL) is captured by the maximal (over all distributions of the dataset for a fixed support) Sibson information of order $\infty$ [33, 54]. Under MI, Fano’s inequality and its variants [72] bound the rate of learning the private variables for a variety of learning metrics, such as error probability and minimum mean-squared error (MMSE).

In particular, despite the strength of MI in providing statistical utility as well as capturing a fairly strong adversary that involves refining beliefs, in the absence of priors on the dataset, using MI as an empirical loss function leads to computationally intractable procedures when learning the optimal parameters of the privatization mechanism from data. Indeed, training algorithms with empirical information theoretic loss functions is a challenging problem that has been explored in specific learning contexts, such as determining randomized encoders for the information bottleneck problem [2] and designing deep auto-encoders using a rate-distortion paradigm [27, 69, 84]. Even in these specific contexts, variational approaches were taken to minimize/maximize a surrogate function instead of minimizing/maximizing an empirical mutual information loss function directly [64]. In an effort to bring theory to practice, we present a general data driven framework to design privacy mechanisms that can capture a range of information-theoretic privacy metrics as loss functions. We will show how our framework leads to very practical (generative adversarial) data-driven formulations that match their corresponding theoretical formulations.

In the context of publishing datasets with privacy and utility guarantees, four approaches that are similar to our work have been recently considered. We briefly review them and clarify how our work is different from each one of them. In [29], the author presents a data-driven methodology to design filters in a way that allows non-malicious entities to learn some public features from the filtered data and prevents malicious entities from learning some private features. While this approach is the closest to ours, the privatizer model considered in this paper is quite restrictive: a deterministic, compressive mapping of the input data. Thus, it leaves out the entire class of randomization-based mechanisms, which we capture via a generative model. Another restriction of the paper is the approach considered to tradeoff utility and privacy. The authors present a weighted combination of utility and privacy using a Lagrangian formulation that is commonly used in machine learning literature (to avoid constrained optimization). Such an approach suffers from two important drawbacks: (i) it is often the case that the optimal solution lies on the boundary (this means that the Lagrangian formulation does not make much sense in practice for privacy applications because the Lagrangian objective is invariant/constant with respect to the Lagrange multiplier when the optimal solution lies on the boundary of the constraint), and (ii) the Lagrangian formulation (whenever applicable) necessitates an excruciating tuning phase where the privacy designer carefully selects the Lagrange multiplier to get a meaningful privacy-utility trade-off. Our formulation allows the designer to place a meaningful distortion constraint thereby directly capturing the privacy-utility trade-off.

In [43], the authors focus on inferences in mobile sensing applications by presenting an algorithmic approach to preserving utility and privacy; their approach relies on using auto-encoders to determine the relevant feature space to add noise to, and thus, avoid adding noise to the original data (which can be very high dimensional). Thus, the data is compressed via a deep auto-encoder which extracts the necessary features to enable learning of select public features. After extracting those low dimensional features, differentially private noise is added to all the features and the original signal is reconstructed. It is worthwhile to note that the autoencoder parameters are carefully selected not just to minimize the $\ell_2$ loss between the original and reconstructed signal but also to maximize the performance of a linear classifier that attempts to learn the public features from the reconstructed signal. This novel approach leverages deep auto-encoders to incorporate some notion of context-aware privacy and achieve a better privacy-utility tradeoff while using DP is used to enforce privacy. However, fundamentally, differential privacy will still incur an insurmountable
utility cost! Our approach follows similar steps but replaces differential privacy (a context free notion) with a much more meaningful context-aware notion of privacy (Generative Adversarial Privacy).

In [78], the authors consider linear adversarial models, and therefore, linear privatizers. Specifically, they ensure privacy by adding noise in directions that are orthogonal to the public features in the hope that the “spaces” of the public and private features are orthogonal (or nearly orthogonal). Ideally, if the public and private features are statistically orthogonal, one can add noise to those public features that overlap with the private features and achieve full privacy without sacrificing utility. However, this work provides no rigorous quantification of privacy.

Finally, in [53], the authors take a similar approach to ours in considering an adversarial nets formulation to share images between consumers and data curators with privacy guarantees. Their framework is not precisely a GANs-like one but more analogous to [29] in that it takes a specific learning function for the attacker (adversary), which in turn is the loss function for the obfuscator (privatizer) and considers a Lagrangian formulation for the utility-privacy tradeoff that the obfuscator computes.

We use conditional generative models to represent privatization schemes. Generative models have recently received a lot of attention in the machine learning community [28, 32, 44, 61, 63]. Ultimately, deep generative models hold the promise of discovering and efficiently internalizing the statistics of the target signal to be generated. State-of-the-art generative models are trained in an adversarial fashion [28, 44]: the generated signal is fed into a discriminator which attempts to distinguish whether the data is real (i.e., sampled from the true underlying distribution) or synthetic (i.e., generated from a low dimensional noise sequence). Training generative models in an adversarial fashion has proven to be successful in the computer vision community and enabled several exciting applications. Similarly, we train the generator, which models the privatizer, in an iterative fashion by making it compete with a discriminator, which models the adversary.

1.3 Outline

The remainder of our paper is organized as follows. We formally present our GAP model in Section 2. We also show how, as a special case, it can recover several information theoretic notions of privacy. We then study a simple (but canonical) binary dataset model in Section 3. In particular, we present theoretically optimal privatization schemes for both PDD and PDI scenarios and show how these schemes can be learned from data using a generative adversarial network. In Section 4, we investigate Gaussian mixture dataset models, and provide a variety of privatization schemes. We comment on their theoretical performance and show how their parameters can be learned from data in a generative adversarial fashion. Our proofs are deferred to sections A, B, and C of the Appendix. We conclude our paper in Section 5 with a few remarks and interesting extensions.

2 Generative Adversarial Privacy Model

We consider a dataset \( \mathcal{D} \) which contains both public and private variables for \( K \) individuals (see Figure 1). We model the public variable as a random variable \( X \in \mathcal{X} \). The private variable, which is correlated with the public variable \( X \), is modeled by a random variable \( Y \in \mathcal{Y} \). The dataset can be viewed as a collection of public and private variables denoted by \((X, Y)\). Each instance of \( X \) and \( Y \) is denoted by \( x \) and \( y \), respectively. We assume each entry pair \((X, Y)\) is distributed as \( P(X, Y) \), and is independent and identically distributed (i.i.d.) across entry pairs in the dataset. Since each entry pair is independent with other entry pairs in the dataset, we consider an approach of using memoryless mechanisms (a privacy mechanism which is applied on each data entry one at a time) to model the privacy mechanism of the privatizer. Formally, we define the privacy mechanism generated by the privatizer to be a memoryless mapping given by

\[
g(X, Y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}.
\]

In the proposed GAP framework, the privatizer, whose goal is to generate the optimal privacy mechanism, is pitted against an adversary. We model the interactions between the privatizer and the adversary as a non-cooperative game. The strategy of the adversary is to improve its accuracy of inferring the private variable \( Y \) from the public variable \( X \). To preserve the privacy of the dataset \( \mathcal{D} \) from inference attacks, the privatizer takes in both public and private variables as input.
and attempts to perturb the public variable $X$ such that it minimizes the adversary’s capability of inferring the private variable from the perturbed data. The optimal privacy mechanism is obtained at an equilibrium point at which both the privatizer and the adversary can not improve its strategy by unilaterally deviating from the equilibrium point.

### 2.1 Formulation

Given the output $\hat{X} = g(X,Y)$ of a privacy mechanism $g(X,Y)$, we define $h(g(X,Y))$ to be the adversary’s inference of the private variable $Y$ from $\hat{X}$. To quantify the effect of adversarial inference, for a given public-private variable pair $(x,y)$, we model the utility of the adversary as a function

$$u(h(g(X = x, Y = y)), Y = y) : \mathcal{Y} × \mathcal{Y} → \mathbb{R}.$$  

Therefore, the utility function of the adversary with respect to $(w.r.t.)$ $X$ and $Y$ is defined to be

$$U(h(g(X,Y)), Y) \triangleq \mathbb{E}_{g(X,Y), Y}[u(h(g(X,Y)), Y)],$$

where the expectation is taken over the joint distribution of $g(X,Y)$ and $Y$. The goal of the adversary is to find the function $h(g(X,Y))$ that maximizes its utility function given by $[1]$.

Note that in this GAP model, the private variable $Y$ can be any attribute in the dataset. For example, if $Y$ is continuous, the task for a privatizer is then to minimize the utility of the adversary, which can be captured by the accuracy of the predicted value of $Y$ from the privatized data $g(X,Y)$. One possible function capturing such utility of the adversary is the expected Euclidean distance

$$U_{ED}(h(g(X,Y)), Y) = -\mathbb{E}_{g(X,Y), Y}[\|h(g(X,Y), Y)\|^2].$$

For this utility choice, the smaller the distance, the higher accuracy the adversary achieves. The private variable can also be a discrete variable, e.g., age, sex, political bias, etc. In this case, the task for the privatizer is to minimize the classification accuracy of the adversary. A meaningful choice of classification accuracy is the 0-1 utility function [52] given by

$$u(h(g(X = x, Y = y), Y = y)) = \begin{cases} 1 & \text{if } h(g(X = x, Y = y)) = y \\ 0 & \text{otherwise} \end{cases}$$

for which the utility of the adversary is given by

$$U_{MAP}(h(g(X,Y)), Y) = \mathbb{E}_{g(X,Y), Y}[u(h(g(X,Y)), Y)] = \mathbb{E}_{g(X,Y)}[P(Y = h(g(X,Y))|g(X,Y))].$$

Although the GAP mechanism obtained with 0-1 utility provides privacy guarantees against a strong MAP adversary, in practice, taking derivatives of a 0-1 utility function w.r.t. the parameters of a machine learning algorithm is intractable. To overcome this, as often done in ML literature, we use the (analytically tractable) negative cross entropy $[65, 83]$ as the utility function in our GAP framework. However, we develop information-theoretic mechanisms for the 0-1 utility function (and therefore, adversarial model) to use as a benchmark to compare the performance of the GAP mechanisms.

An alternative model for the adversary is the log information utility function measured in bits (often referred to as information density in IT literature; see, for example, [3])

$$U_{MI}(h(g(X,Y)), Y) = \mathbb{E}_{h(g(X,Y)), Y} \log \frac{P(h(g(X,Y), Y))}{P(h(g(X,Y)))}$$

(5)

to capture the correlation between the recovered private variable $h(g(X,Y))$ and $Y$.

Intuitively, the privacy mechanism must minimize the utility of the adversary. In this case, there exists a trivial solution in which the privatizer just outputs $\hat{X}$ independent of $X$. However, such a privacy mechanism provides no utility for data analysts who want to learn non-private variables from $\hat{X}$. In order to circumvent this issue, we model the loss incurred by privatizing the original data as a distortion function

$$d(\hat{x}, x) : \mathcal{X} × \mathcal{X} → \mathbb{R},$$

(6)
which measures the difference between the original data $x$ and the privatized data $\hat{x}$. Thus, the average distortion w.r.t. $g(X,Y)$ and $X$ is

\[ \mathbb{E}_{g(X,Y),X}[d(g(X,Y), X)]. \]

Given an individual’s public variable $X \in \mathcal{X}$, the objective of the privatizer is to find a mechanism $g(X,Y)$ that maps the input data $X$ to an output data $\hat{X} \in \mathcal{X}$ such that it is both privacy preserving (in the sense that it is very difficult for an adversary to learn $Y$ from $\hat{X}$) and does not distort the original data too much (helps data analysts to learn useful features of the individuals).

To achieve the two opposing goals, we model the problem as a constrained minimax game between the privatizer and the adversary. The objective of the privatizer is to minimize the utility of the adversary subject to the distortion constraint:}

\[
\min_{g(\cdot)} \max_{h(\cdot)} \mathbb{E}_{g(X,Y),X}[d(g(X,Y), X)] \\
\text{s.t. } \mathbb{E}_{g(X,Y),X}[d(g(X,Y), X)] \leq D,
\]

where the constant $D \geq 0$ determines the allowable distortion for the privatizer. If $D$ is very large, the problem is close to a trivial privacy-only task. However, if $D$ is very small, the privatizer can not distort the data too much. This GAP framework is universal since any information hidden in the dataset which are relevant to the private variable will be privatized.

We define the private mechanism to be private data dependent (PDD) if its output is dependent on the private variable $Y$. Otherwise, we call the privacy mechanism private data independent (PDI). In general, the optimal privacy mechanism that solves (6) depends on the public variable $X$ and the private variable $Y$. However, modeling this relationship is not straightforward or known a priori. Given this limitation, we consider two simple yet meaningful distributions. In Section 3 we consider binary private and public variables with Bernoulli distributions. In Section 4 we study the case in which the private variable follows a Bernoulli distribution while the public variable follows a Gaussian distribution conditioned on the the private variable.

### 2.2 Generative Adversarial Privacy under MAP Adversary

Consider a MAP adversary, i.e., an adversary that guesses $\hat{Y}$ as the most likely value of the private variable $Y$ given its observation $\hat{X} = g(X,Y)$. We assume the privatizer has access to the joint distribution of $X$ and $Y$. Furthermore, we consider a strong adversary who has knowledge of the joint distribution $P(X,Y)$ and the privacy mechanism of the privatizer. Let $\hat{Y} = h(g(X,Y))$ be the adversary’s inference of $Y$ from the privatized public variable $\hat{X}$. Recall that since $Y$ is binary, the utility of the MAP adversary can be written as the classification accuracy in (4), i.e., as an expected 0-1 utility given by

\[
U_{MAP}(h(g(X,Y)), Y) = \mathbb{E}_{g(X,Y)} \mathbb{P}(Y = h(g(X,Y)) | g(X,Y)) \\
= \mathbb{E}_{g(X,Y)} \mathbb{P}(Y = h(g(X,Y)) | g(X,Y)).
\]

For a fixed privacy mechanism $g(X,Y)$, the MAP adversary chooses $h(g(X,Y))$ that maximizes $P(Y = h(g(X,Y)), g(X,Y))$.

Having defined the utility of the adversary, the privatizer’s objective is to use the privacy mechanism $g(X,Y)$ to minimize the expected probability of correctly inferring $Y$ from $g(X,Y)$ subject to the distortion constraint $\mathbb{E}_{g(X,Y),X}[d(g(X,Y), X)] \leq D$. Therefore, the privatizer solves the following optimization problem

\[
\min_{g(\cdot)} \max_{h(\cdot)} \mathbb{E}_{g(X,Y),X}[d(g(X,Y), X)] \\
\text{s.t. } \mathbb{E}_{g(X,Y),X}[d(g(X,Y), X)] \leq D.
\]

### 2.3 Generative Adversarial Privacy under Log Information Utility

In this section, we consider the log information utility function in (4). Let $h(g(X,Y))$ be an arbitrary function of $g(X,Y)$. Thus, $Y \rightarrow g(X,Y) \rightarrow h(g(X,Y))$ forms a Markov chain. The expected utility of the adversary is given by

\[
U_{MI}(h(g(X,Y)), Y) = \mathbb{E}_{h(g(X,Y)), Y} \log \frac{P(h(g(X,Y)), Y)}{P(h(g(X,Y)))P(Y)} = I(h(g(X,Y)); Y),
\]
2.4 Data-driven Generative Adversarial Privacy

If the privatizer only has access to the dataset \( \mathcal{D} \) but not the joint distribution of \( X \) and \( Y \), finding the optimal privacy mechanism becomes a learning problem. The learned privacy mechanism \( g(X,Y;\theta_p) \) is defined to be a function of \( (X,Y) \) parameterized by \( \theta_p \). This mechanism takes the public and private variables \( (X,Y) \) as input and outputs privatized variable \( \hat{X} \). On the other hand, the adversary learns a classification function \( h(g(X,Y;\theta_p);\theta_a) \) parameterized by \( \theta_a \), which uses the privatized data \( \hat{X} \) to classify the private variable \( Y \).

We note that the functions \( h(g(X,Y;\theta_p);\theta_a) \) and \( g(X,Y;\theta_p) \) are arbitrary, i.e., they capture all possible learning algorithms. Figure 3 shows an example of the GAP model in which the privatizer and adversary are modeled as multi-layer neural networks. For a chosen privatizer and adversary, we define the utility function of the adversary to be a continuous and differentiable function \( U_{ML}(h(g(X,Y;\theta_p);\theta_a), Y) \), which captures the accuracy of predicting the private variable \( Y \) from the privatized data \( \hat{X} \). In keeping with state-of-the-art machine learning loss functions, we choose \( U_{ML}(h(g(X,Y;\theta_p);\theta_a), Y) \), as the negative of the cross entropy function given by

\[
U_{ML}(h(g(X,Y;\theta_p);\theta_a), Y) = \frac{1}{K} \sum_{i=1}^{K} [y(i) \log h(g(x(i),y(i);\theta_p);\theta_a) \\
+ (1 - y(i)) \log(1 - h(g(x(i),y(i);\theta_p);\theta_a))],
\]

where \( x(i) \) and \( y(i) \) are the \( i \)th row of \( X \) and \( Y \) in the dataset \( \mathcal{D} \), respectively and \( K \) is the size of the dataset. Thus, the adversary learns the optimal parameter \( \theta_a^* \) such that

\[
\theta_a^* = \arg \max_{\theta_a} U_{ML}(h(g(X,Y;\theta_p);\theta_a), Y).
\]

To prevent the adversary from inferring the private variable \( Y \), the privatizer generates the privatized data \( \hat{X} \) based on \( X \) such that it minimizes the classification accuracy of the adversary without distorting the original data \( X \) too much. Therefore, the optimal parameters for the privacy mechanism \( (\theta_p, \theta_a) \) are given by the solution to

\[
\begin{align*}
\min_{\theta_p} & \max_{\theta_a} U_{ML}(h(g(X,Y;\theta_p);\theta_a), Y) \\
\text{s.t.} & \quad \mathbb{E}_{D}[d(g(X,Y;\theta_p), X)] \leq D,
\end{align*}
\]
where the expectation is taken over the dataset $D$.

The minimax optimization problem in (12) can be considered as a two-player non-cooperative game between the privatizer and the adversary. The strategies of the privatizer and adversary are given by $\theta_p$ and $\theta_a$, respectively. Each player chooses the strategy that optimizes its utility function w.r.t. what its opponent does. In particular, the privatizer must expect that if it chooses $\theta_p$, the adversary chooses $\theta_a$ that maximizes its own utility function based on the choice of the privatizer. The optimal privacy mechanism is given by the equilibrium of the privatizer-adversary game.

In practice, we can learn an equilibrium of the game using an iterative algorithm presented in Algorithm 1. We first optimize the utility function of the adversary in the inner loop to compute the parameters of the adversary for fixed privatizer parameters. Then, we optimize the utility function of the privatizer in the outer loop by fixing the parameters of the adversary. To avoid over-fitting the parameters, we alternate between $k$ steps of optimizing the adversary and one step of the privatizer. This results in the adversary moving towards its optimal solution for small perturbations of the privatizer. The mini-batch version of the alternating minimax privacy preserving algorithm is formally presented in Algorithm 1.

**Algorithm 1** Alternating minimax privacy preserving algorithm

**Input:** data $(X,Y)$, privatizer $g$, adversary $h$, distortion parameter $D$, iteration number $T$

**Output:** Optimal privatizer parameters $\theta_p$

**procedure** ALCERNATE MINIMAX($X,Y,D,T$)

Initialize $\theta_p^1$ and $\theta_a^1$

for $t = 1,\ldots,T$ do

Random minibatch of $M$ datapoints $\{x^{(1)},\ldots,x^{(M)}\}$ drawn from full dataset

Generate $\{\hat{x}^{(1)},\ldots,\hat{x}^{(M)}\}$ via $\hat{x}^{(i)} = g(x^{(i)},y^{(i)};\theta_p^t)$

Compute the parameter $\theta_a^{t+1}$ for the adversary

$$\theta_a^{t+1} = \arg \max_{\theta_a} \frac{1}{M} \sum_{i=1}^{M} u(h(\hat{x}^{(i)};\theta_a),y^{(i)})$$

Compute the descent direction $\nabla_{\theta_p} l(\theta_p,\theta_a^{t+1})$, where

$$l(\theta_p,\theta_a^{t+1}) = \frac{1}{M} \sum_{i=1}^{M} u(h(g(x^{(i)},y^{(i)};\theta_p);\theta_a^{t+1}),y^{(i)})$$

subject to $\frac{1}{M} \sum_{i=1}^{M} [d(g(x^{(i)},y^{(i)};\theta_p),x^{(i)})] \leq D$

Perform line search along $\nabla_{\theta_p} l(\theta_p,\theta_a^{t+1})$ and update

$$\theta_p^{t+1} = \theta_p^{t} - \alpha_t \nabla_{\theta_p} l(\theta_p,\theta_a^{t+1}), \quad \alpha_t > 0$$

Exit if solution converged

return $\theta_p^{t+1}$

To incorporate the distortion constraint in the learning algorithm, we use a penalty method to replace the constrained optimization problem by a series of unconstrained problems whose solutions
asymptotically converge to the solution of the constrained problem. The new unconstrained problem is formed by adding a penalty function, which consists of a penalty parameter $\rho$, multiplied by a measure of violation of the constraint. The measure of violation is non-zero when the constraint is violated and is zero if the constraint is not violated. Therefore, in Algorithm 1, the constrained optimization problem of the privatizer can be approximated by a series of unconstrained optimization problems with the loss function

$$l(\theta_p, \theta_a^{t+1}) = \frac{1}{M} \sum_{i=1}^{M} u(h(g(x(i), y(i); \theta_p); \theta_a^{t+1}), y(i))$$

$$\quad + \rho_t \left( \max \{0, \frac{1}{M} \sum_{i=1}^{M} d(g(x(i), y(i); \theta_p), x(i)) - D \} \right),$$

where $\rho_t$ is a penalty coefficient which increases with the number of iterations $t$. The solutions of the series of unconstrained problems will eventually converge to the solution of the original constrained problem.

The augmented Lagrangian method is yet another approach to enforce equality constraint by penalizing the objective function when the constraint is not satisfied. Different from the penalty method, the augmented Lagrangian method combines the use of Lagrange multiplier and a quadratic penalty term. Note that this method is designed for equality constraint. Therefore, we introduce a slack variable $\delta$ to convert the inequality distortion constraint into an equality constraint. Using the augmented Lagrangian method, the constrained optimization problem of the privatizer can be replaced by a series of unconstrained problem with the loss function given by

$$l(\theta_p, \theta_a^{t+1}, \delta) = \frac{1}{M} \sum_{i=1}^{M} u(h(g(x(i), y(i); \theta_p); \theta_a^{t+1}), y(i)) + \frac{\rho_t}{2} \left( \frac{1}{M} \sum_{i=1}^{M} d(g(x(i), y(i); \theta_p), x(i)) + \delta - D \right)^2$$

$$\quad - \lambda_t \left( \frac{1}{M} \sum_{i=1}^{M} d(g(x(i), y(i); \theta_p), x(i)) + \delta - D \right),$$

where $\rho_t$ is a penalty coefficient which increases with the number of iterations $t$ and $\lambda_t$ is updated according to the rule $\lambda_{t+1} = \lambda_t - \rho_t \left( \frac{1}{M} \sum_{i=1}^{M} d(g(x(i), y(i); \theta_p), x(i)) + \delta - D \right)$. Solutions to the series of unconstrained problems formulated by the augmented Lagrangian method will also converge to the solution of the original constrained problem.

## 3 Binary Data Model

We consider both public and private variables to be single valued binary random variables. Let $p_{i,j}$ denote joint probability of $(X, Y) = (i, j)$, where $i, j \in \{0, 1\}$. To prevent an adversary from correctly inferring the private variable $Y$ from the public variable $X$, for each public variable $X = x$, the privatizer applies a randomized mechanism on $x$ to generate the privatized data $\hat{X} = \hat{x}$ that preserves privacy while restricting the distortion of the original data to an acceptable level. Since both the original and privatized public variables are binary random variables, the distortion between $X$ and $\hat{X}$ can be modeled by the Hamming distortion, i.e. $d(\hat{X}, X) = 1$ if $\hat{X} \neq X$ and $d(\hat{X}, X) = 0$ if $\hat{X} = X$. Therefore, the expected distortion is given by $E_{\hat{X}, X}[d(\hat{X}, X)] = P(X \neq \hat{X})$, where $P(X \neq \hat{X})$ denotes the probability of the event $X \neq \hat{X}$.

### 3.1 Theoretical Approach for Binary Data Model

The adversary’s objective is to guess $Y$ from $\hat{X}$. We consider a MAP adversary who has access to the joint distribution $P(X, Y)$ and the privacy mechanism. Thus, the privatizer’s goal is to privatize the dataset $D$ such that it minimizes the adversary’s probability of correctly guessing the private variable $Y$ from the privatized public variable $\hat{X}$ given by $P(Y \mid \hat{X})$ subject to the distortion constraint. We first focus on the PDD privacy mechanism, which depends on both $Y$ and $X$. Later we consider a more tractable PDI privacy mechanism in which the mechanism only depends on $X$. 

10
3.1.1 PDD Privacy Mechanism

Let \( g(X,Y) \) denote the randomizing PDD mechanism. Since we map a binary \( X \) to a binary \( \hat{X} \), the function \( g(X,Y) \) is a conditional probability distribution which maps \( (X,Y) \) to an output \( \hat{X} \). Let \( P(\hat{X}|X,Y) \) be the conditional distribution that maps the public and private variable pair \( (X,Y) \) to an output \( \hat{X} \) such that

\[
P(\hat{X} = 0|X = 0, Y = 0) = s_{0,0}, \quad P(\hat{X} = 0|X = 0, Y = 1) = s_{0,1}
\]

\[
P(\hat{X} = 1|X = 1, Y = 0) = s_{1,0}, \quad P(\hat{X} = 1|X = 1, Y = 1) = s_{1,1}.
\]

Thus, the marginal distribution of \( \hat{X} \) is given by

\[
P(\hat{X} = 0) = \sum_X P(\hat{X} = 0|X,Y)P(X,Y) = s_{0,0}p_{0,0} + s_{0,1}p_{0,1} + (1 - s_{1,0})p_{1,0} + (1 - s_{1,1})p_{1,1},
\]

\[
P(\hat{X} = 1) = \sum_X P(\hat{X} = 1|X,Y)P(X,Y) = (1 - s_{0,0})p_{0,0} + (1 - s_{0,1})p_{1,0} + s_{1,0}p_{0,1} + s_{1,1}p_{1,1}.
\]

If \( \hat{X} = 0 \), the adversary’s utility of guessing \( \hat{Y} = h(\hat{X}) = 1 \) is

\[
P(\hat{Y} = 1, \hat{X} = 0) = \sum_X P(Y = 1, X)P(\hat{X} = 0|X) = p_{1,1}(1 - s_{1,1}) + p_{0,1}s_{0,1}
\]

and the utility of guessing \( \hat{Y} = h(\hat{X}) = 0 \) is

\[
P(\hat{Y} = 0, \hat{X} = 0) = \sum_X P(Y = 0, X)P(\hat{X} = 0|X) = p_{1,0}(1 - s_{1,0}) + p_{0,0}s_{0,0}.
\]

We define \( s = \{s_{0,0}, s_{0,1}, s_{1,0}, s_{1,1}\} \) to be a probability vector which describes the privacy mechanism. Thus, for \( \hat{X} = 0 \), the adversary’s utility function is given by

\[
U_{MAP}^{(B)}(s, \hat{X} = 0) = \max\{P(\hat{Y} = 1, \hat{X} = 0), P(\hat{Y} = 0, \hat{X} = 0)\}.
\]

Similarly, if \( \hat{X} = 1 \), the adversary’s utility function is given by

\[
U_{MAP}^{(B)}(s, \hat{X} = 1) = \max\{P(\hat{Y} = 1, \hat{X} = 1), P(\hat{Y} = 0, \hat{X} = 1)\},
\]

where

\[
P(\hat{Y} = 1, \hat{X} = 1) = \sum_X P(Y = 1, X)P(\hat{X} = 1|X,Y) = p_{1,1}s_{1,1} + p_{0,1}(1 - s_{0,1});
\]

\[
P(\hat{Y} = 0, \hat{X} = 1) = \sum_X P(Y = 0, X)P(\hat{X} = 1|X) = p_{1,0}s_{1,0} + p_{0,0}(1 - s_{0,0}).
\]

Therefore, we have \( \max U_{MAP}^{(B)} = U_{MAP}^{(B)}(s, \hat{X} = 1) + U_{MAP}^{(B)}(s, \hat{X} = 0) \). The optimal privacy mechanism is given by the solution to

\[
\min_s U_{MAP}^{(B)}(s, \hat{X} = 1) + U_{MAP}^{(B)}(s, \hat{X} = 0)
\]

\[
s.t. \quad P(\hat{X} = 0, X = 1) + P(\hat{X} = 1, X = 0) \leq D
\]

\[
s \in [0, 1]^4.
\]

Note that the above optimization problem is a four dimensional linear minimax optimization problem parameterized by \( p, q \), and \( D \), it can be formulated as a linear program

\[
\min_{s_{1,1}, s_{0,1}, s_{1,0}, s_{0,0}, t_0, t_1} t_0 + t_1
\]

\[
s.t. \quad 0 \leq s_{1,1}, s_{0,1}, s_{1,0}, s_{0,0} \leq 1
\]

\[
p_{1,1}(1 - s_{1,1}) + p_{0,1}s_{0,1} \leq t_0
\]

\[
p_{1,0}(1 - s_{1,0}) + p_{0,0}s_{0,0} \leq t_0
\]

\[
p_{1,1}s_{1,1} + p_{0,1}(1 - s_{0,1}) \leq t_1
\]

\[
p_{1,0}s_{1,0} + p_{0,0}(1 - s_{0,0}) \leq t_1
\]

\[
p_{1,1}(1 - s_{1,1}) + p_{0,1}(1 - s_{0,1}) + p_{1,0}(1 - s_{1,0}) + p_{0,0}(1 - s_{0,0}) \leq D,
\]

where \( t_0 \) and \( t_1 \) are two slack variables representing the maximum in (15) and (16), respectively. The optimal mechanism can be obtained numerically by solving the above linear programming.
3.1.2 PDI Privacy Mechanism

In the previous formulation, we assume that the privacy mechanism is dependent on both \( Y \) and \( X \). Although we can formulate the problem as an linear programming with four variables, determining the closed form solution for such a highly parameterized problem is not analytically tractable. Thus, we consider a simple yet meaningful PDI privacy mechanism in which we assume that given the public variable \( X \), the output of the privacy mechanism is independent of the private variables. Thus, the Markov chain \( Y \to X \to \hat{X} \) holds. As a result, the joint distribution of \( P(Y, X = \hat{x}) \) can be written as

\[
P(Y, \hat{X} = \hat{x}) = \sum_X P(Y, \hat{X} = \hat{x} | X)P(X)
\]

Thus, the Markov chain \( Y \to X \to \hat{X} \) holds. As a result, the joint distribution of \( P(Y, X = \hat{x}) \) can be written as

\[
P(Y, \hat{X} = \hat{x}) = \sum_X P(Y | X)P(\hat{X} = \hat{x} | X)P(X)
\]

where the third equality is due to the property of the Markov chain \( Y \to X \to \hat{X} \). For a PDI privatizer, the privacy mechanism \( P(X | Y) \) can be simplified to \( P(\hat{X} | X) \), which is a conditional distribution that maps the input data \( X \) to the output \( \hat{X} \). To make the problem more tractable, we focus on a slightly simpler setting in which \( Y = X \oplus N \) is modeled as a mod two sum of \( X \) and \( N \), where \( N \in \{0, 1\} \) is independent of \( X \) and follows a Bernoulli distribution with parameter \( q \). In this setting, \( P(X, Y) \) can be computed by

\[
P(X = 1, Y = 1) = P(Y = 1 | X = 1)P(X = 1) = p(1 - q)
\]

\[
P(X = 0, Y = 1) = P(Y = 1 | X = 0)P(X = 0) = (1 - p)q
\]

\[
P(X = 1, Y = 0) = P(Y = 0 | X = 1)P(X = 1) = pq
\]

\[
P(X = 0, Y = 0) = P(Y = 0 | X = 0)P(X = 0) = (1 - p)(1 - q).
\]

Define \( s = \{s_0, s_1\} \) to be the PDI privacy mechanism in which \( s_0 = P(\hat{X} = 0 | X = 0) \) and \( s_1 = P(\hat{X} = 1 | X = 1) \). Thus, we have \( P(\hat{X} = 1 | X = 0) = 1 - s_0 \) and \( P(\hat{X} = 0 | X = 1) = 1 - s_1 \). The joint distribution \( P(\hat{Y}, \hat{X}) \) can be computed by

\[
P(\hat{Y} = 1, \hat{X} = 0) = p(1 - q)(1 - s_1) + (1 - p)qs_0,
\]

\[
P(\hat{Y} = 0, \hat{X} = 0) = pq(1 - s_1) + (1 - p)(1 - q)s_0,
\]

\[
P(\hat{Y} = 1, \hat{X} = 1) = p(1 - q)s_1 + (1 - p)q(1 - s_0),
\]

\[
P(\hat{Y} = 0, \hat{X} = 1) = pqs_1 + (1 - p)(1 - q)(1 - s_0).
\]

Therefore, the optimal PDI privacy mechanism is given by the solution to

\[
\min_s \max_{h(\cdot)} U^{(B)}_{MAP}(s, h(\cdot))
\]

\[
s.t. \quad P(\hat{X} = 0, X = 1) + P(\hat{X} = 1, X = 0) \leq D
\]

\[
s \in [0, 1]^2,
\]

where \( \max_{h(\cdot)} U^{(B)}_{MAP} = \max\{P(\hat{Y} = 1, \hat{X} = 0), P(\hat{Y} = 0, \hat{X} = 0)\} + \max\{P(\hat{Y} = 1, \hat{X} = 1), P(\hat{Y} = 0, \hat{X} = 1)\} \). Furthermore, the distortion in \( U^{(B)}_{MAP} \) is given by \( P(X \neq \hat{X}) = (1 - s_0)(1 - p) + (1 - s_1)p \). Note that \( \max_{h(\cdot)} U^{(B)}_{MAP} \) can be considered as a sum of two functions, where each function is a maximum of two linear functions. Therefore, it is convex in \( s_0 \) and \( s_1 \) for different values of \( p, q \) and \( D \).

**Theorem 1.** For fixed \( p, q \) and \( D \), there exists infinitely many privacy mechanisms that achieve the optimal privacy-utility tradeoff. If \( q = \frac{1}{2} \), any privacy mechanism that satisfies \( \{s_0, s_1 | ps_1 + (1 - p)s_0 \geq 1 - D, s_0, s_1 \in [0, 1]\} \) is optimal. If \( q \neq \frac{1}{2} \), the optimal PDI privacy mechanism is given as follows:
In practice, the joint distribution between public and private variables is often unknown to both the privatizer and adversary. We then compute the optimal privacy mechanism of the privatizer in each subproblem. Theorem 1. Summarizing the optimal solution to the subproblems for different values of the distortion constraint, we define the utility given by (11). Since both the distortion constraint and adversary guessing are binary, we use two simple neural networks to model the privatizer and adversary, respectively. On the other hand, the adversary is also modeled by a two-layer neural network to represent the privacy mechanism directly. For the PDD privacy mechanism, we have the optimal privacy parameters that solves (12) with the utility given by (11). Since both X and Y are binary variables as in the MAP analysis, here too we can use the privatizer parameters θp to represent the privacy mechanism s directly. For the adversary, we define θa = (θa,0, θa,1), where θa,0 = P(Y = 0|X = 0) and θa,1 = P(Y = 1|X = 1). Thus, given a privatized public variable input g(x(i); y(i); θp) ∈ {0, 1}, the output belief of the adversary guessing y(i) = 1 can be written as (1 − θa,0)(1 − g(x(i); y(i); θp)) + θa,1g(x(i); y(i); θp).

For the PDD privacy mechanism, we have θp = s = {s0,0, s0,1, s1,0, s1,1}. Given the fact that both x(i) and y(i) are binary, we use two simple neural networks to model the privatizer and the adversary. As shown in Figure 4 the privatizer is modeled by a single-layer neural network parameterized by s. On the other hand, the adversary is also modeled by a two-layer neural network classifier. From the perspective of the privatizer, the expected privacy mechanism of an adversary guessing y(i) = 1 conditional on the input (x(i), y(i)) is given by

$$h(g(x(i), y(i); s); θa) = θa,1P(̂x(i) = 1) + (1 − θa,0)P(̂x(i) = 0),$$

where

$$P(̂x(i) = 1) = x(i)y(i)s_{1,1} + (1 − x(i))y(i)(1 − s_{0,1}) + x(i)(1 − y(i))s_{1,0} + (1 − x(i))(1 − y(i))(1 − s_{0,0}),$$

$$P(̂x(i) = 0) = x(i)y(i)(1 − s_{1,1}) + (1 − x(i))y(i)s_{0,1} + x(i)(1 − y(i))(1 − s_{1,0}) + (1 − x(i))(1 − y(i))s_{0,0}.$$
Furthermore, the expected distortion is given by

$$\mathbb{E}_{g(X,Y),X}[d(g(X, Y; s), X)] = \frac{1}{K} \sum_{i=1}^{K} [(x(i)y(i)(1 - s_{1,1}) + x(i)(1 - y(i))(1 - s_{1,0})$$

$$+ (1 - x(i))y(i)(1 - s_{0,1}) + (1 - x(i))(1 - y(i))(1 - s_{0,0})].$$

(31)

Similar to the PDD case, we can also compute the expected probability of guessing $y(i) = 1$ conditional on the input $(x(i), y(i))$ for the PDI privatizer. Note that in the PDI case, $\theta_p = s = \{s_0, s_1\}$. Therefore, we have

$$h(g(x(i), y(i); \theta_p); s) = \theta_{a,1}[x(i)s_1 + (1 - x(i))(1 - s_0)] + (1 - \theta_{a,0}][(1 - x(i))s_0 + x(i)(1 - s_1)].$$

(32)

The expected distortion is given by

$$\mathbb{E}_{g(X,Y),X}[d(g(X, Y; s), X)] = \frac{1}{K} \sum_{i=1}^{K} [x(i)(1 - s_1) + (1 - x(i))(1 - s_0)].$$

(33)

Thus, we can use Algorithm 1 proposed in Section 2.4 to learn the optimal PDD and PDI privacy mechanisms from the dataset.

### 3.3 Illustration of Results

We now evaluate our proposed GAP framework using synthetic datasets. We focus on a slightly simpler setting in which $Y = X \oplus N$, where $N \in \{0, 1\}$ is independent of $X$ and follows a Bernoulli distribution with parameter $q$. We generate two synthetic datasets with $(p, q)$ equals to $(0.75, 0.25)$ and $(0.5, 0.25)$, respectively. Each synthetic dataset used in this experiment contains 10,000 training samples and 2,000 testing samples. We use Tensorflow to train both the privatizer and the adversary using Adam optimizer with learning rate equals to 0.01 and minibatch size equals to 200. The privacy performance is measured by the accuracy of correctly inferring the private variable by a strong MAP adversary, assuming that the adversary: (a) has access to $P(X,Y)$, (b) has knowledge of the learned privacy mechanism, and (c) can compute the MAP detection rule. Firstly, the optimal privacy mechanisms against the MAP adversary obtained via both theoretical and learning approaches are compared in terms of privacy vs. distortion. For comparison’s sake, we also compute the privacy guarantees of our GAP mechanisms under a log information utility function adversary (i.e., MI privacy measure).

Figure 3(a) illustrates the performance of both optimal PDD and PDI privacy mechanisms against the strong theoretical MAP adversary when $(p, q) = (0.5, 0.25)$. It can be seen that the classification accuracy of the MAP adversary reduces as the distortion increases for both optimal PDD and PDI privacy mechanisms. As one would expect, the PDD privacy mechanism achieves a lower classification accuracy for the adversary, i.e., better privacy, than the PDI mechanism. Furthermore, when the distortion is greater than some threshold value, the classification accuracy of the MAP adversary saturates regardless of the distortion. This is due to the fact that the correlation between the private variable and the privatized public variable can be reduced no further once the distortion goes beyond the saturation threshold. Therefore, increasing distortion will not further reduce the accuracy of the MAP adversary when it has knowledge of the priors on $(X,Y)$. We also observe that the privacy mechanism obtained through the data-driven approach performs very well when pitted against the MAP adversary (maximum accuracy difference around 3% compared with theoretical approach). In other words, for the binary data model, the data-driven approach for GAP can generate privacy mechanisms that perform as well as the mechanisms computed from the theoretical approach, which assumes the privatizer has access to the underlying distribution of the data.

Figure 3(b) shows the performance of both optimal PDD and PDI privacy mechanisms against the MAP adversary for $(p, q) = (0.75, 0.25)$. Similar to the equal prior case, we observe that both PDD and PDI privacy mechanisms reduce the accuracy of the MAP adversary as the distortion increases and saturate after the distortion goes beyond certain threshold. It can be seen that the saturation thresholds for both PDD and PDI privacy mechanisms in Figure 3(a) are lower than the equal prior case plotted in Figure 3(a). The reason is that when $(p, q) = (0.75, 0.25)$, the correlation between $Y$ and $X$ is weaker than the $(p, q) = (0.5, 0.25)$ case. Therefore, it requires less distortion...
to achieve the same privacy than the equal prior case. We also observe that the performance of the GAP mechanism obtained via the data-driven approach is comparable to the theoretical approach against the MAP adversary.

The performance of the GAP mechanism obtained using log information utility (i.e., MI privacy) is plotted in Figure 5c and 5d. Similar to the MAP adversary case, as the distortion increases, the mutual information between the private variable and the privatized public variable achieved by the optimal PDD and PDI mechanisms decreases if the distortion is smaller than some threshold. If the distortion is larger than the threshold, the optimal privacy mechanism is able to make the private variable and the privatized public variable independent regardless of the distortion. Furthermore, the values of the saturation distortion thresholds are very close to what we observe in Figure 5a and 5b.

4 Binary Gaussian Mixture Model

Thus far, we have studied binary private and public variables. However, in many real datasets, the sample space of variables may contain countably infinite number of possible values or even be continuous. It is well known that the Gaussian distribution is a flexible approximate for many distributions [76]. Therefore, in this section, we study a setting where $Y \in \{0, 1\}$ and $X$ is a Gaussian random variable whose mean and variance are dependent on $Y$. Without loss of generality, let $E[X|Y=1] = -E[X|Y=0] = \mu$ and $P(Y=1) = p$. Thus, $X|Y=0 \sim N(-\mu, \sigma_0^2)$ and $X|Y=1 \sim N(\mu, \sigma_1^2)$.

Similar to the binary public variable case, we study two privatization schemes: (a) private data independent (PDI) schemes (where $\hat{X} = g(X)$), and (b) private data dependent (PDD) schemes (where $\hat{X} = g(X, Y)$). In order to have a tractable model for the privatizer, we assume $g(X, Y)$ is realized by adding an affine function of an independently generated Gaussian noise to the public variable $X$. The affine function enables controlling both the mean and variance of the privatized data. In particular, we consider $g(X, Y) = X + (1 - Y)\beta_0 - Y\beta_1 + (1 - Y)\gamma_0 N + Y\gamma_1 N$, in which $N$ is a one dimensional random variable and $\beta_0, \beta_1, \gamma_0, \gamma_1$ are constant parameters. The goal of the
privatizer is to privatize the public data $X$ subject to the distortion constraint $E_{X,X'}||X - X'||_2^2 \leq D$.

4.1 Theoretical Approach for Binary Gaussian Mixture Model

We study the theoretical approach for noise adding privacy mechanisms for the binary Gaussian mixture model. To make the problem more tractable, let us consider a slightly simpler setting where the variances of the public variables are the same regardless of the private variable $Y$ (i.e., $\sigma_0 = \sigma_1 = \sigma$). We will relax this assumption later when we take a data driven approach. Furthermore, we assume $N$ to be a standard Gaussian random variable. One might, rightfully, question our choice of focusing our attention on adding (potentially $Y$-dependent) Gaussian noise. Though other distributions can be considered, our approach is motivated by the following two reasons:

- (a) Although it is shown that adding Gaussian noise is not the worst case noise adding mechanism for non-Gaussian input [62], identifying the correct noise distribution is very difficult. Thus, for tractability and ease of analysis, we choose Gaussian noise.
- (b) Adding Gaussian noise to each data entry conditioned on its private variable preserves the Gaussianity of the released dataset.

In what follows, we will analyze a variety of PDI and PDD mechanisms.

4.1.1 PDI Gaussian Noise Adding Privacy Mechanism

We consider a PDI noise adding privatizing scheme which adds an affine function of the standard Gaussian noise to the public variable. Since the privacy mechanism is PDI, we have $g(X,Y) = \gamma N + \beta$, where $\gamma$ and $\beta$ are constant parameters and $N \sim N(0,1)$. Using the classical Gaussian hypothesis testing analysis, it is straightforward to verify that the utility of a MAP adversary is given by

$$U_{\text{MAP}}^{(G)} = pQ\left(-\frac{\alpha}{2} + \frac{1}{\alpha} \ln \left(\frac{1-p}{p}\right)\right) + (1-p)Q\left(-\frac{\alpha}{2} - \frac{1}{\alpha} \ln \left(\frac{1-p}{p}\right)\right),$$

where $\alpha = \frac{2\mu}{\sqrt{\gamma^2 + \sigma^2}}$. Furthermore, since $E_{X,X'}[d(\hat{X},X)] = \beta^2 + \gamma^2$, the distortion constraint implies that $\beta^2 + \gamma^2 \leq D$.

**Theorem 2.** For a PDI noise adding privatizing scheme given by $g(X,Y) = X + \beta + \gamma N$, $\beta, \gamma \geq 0$, the optimal parameters are given by

$$\beta = 0, \gamma = \sqrt{D}. \quad (35)$$

Let $\alpha = \frac{2\mu}{\sqrt{D} + \sigma}$. For this optimal scheme, the accuracy of the MAP adversary is

$$U_{\text{MAP}}^{(G)} = pQ\left(-\frac{\alpha}{2} + \frac{1}{\alpha} \ln \left(\frac{1-p}{p}\right)\right) + (1-p)Q\left(-\frac{\alpha}{2} - \frac{1}{\alpha} \ln \left(\frac{1-p}{p}\right)\right).$$

The proof of Theorem 2 is provided in Appendix [3]. We observe that the privatizing scheme which minimizes the probability of correct guessing under a MAP adversary with distortion upper-bounded by $D$ is to add a Gaussian noise with zero mean and $\sqrt{D}$ variance.

4.1.2 PDD Gaussian Noise Adding Privacy Mechanism

For the PDD privatizing scheme, we first consider a simple case in which $\gamma_0 = \gamma_1 = 0$. Without loss of generality, we assume $\beta_0$ and $\beta_1$ are both non-negative. Therefore, the privatized data is given by $\hat{X} = X + (1-Y)\beta_0 - Y\beta_1$. This is a PDD mechanism because $\hat{X}$ depends on both $X$ and $Y$. Intuitively, this mechanism privatizes the data by shifting the two Gaussian distributions (under $Y = 0$ and $Y = 1$) closer to each other. Under this mechanism, it is easy to show that the MAP probability of classifying the private variable $Y$ from $\hat{X}$ is given by $U_{\text{MAP}}^{(G)}$ in (34) with $\alpha = \frac{2\mu - (\beta_1 + \beta_0)}{\sigma}$. Observe that since $d(\hat{X},X) = ((1-Y)\beta_0 - Y\beta_1)^2$, we have $E[d(\hat{X},X)] = (1-p)\beta_0^2 + p\beta_1^2$. Thus, the distortion constraint implies that $(1-p)\beta_0^2 + p\beta_1^2 \leq D$. 

16
Theorem 3. For a PDD privatizing scheme given by \( g(X, Y) = X + (1 - Y)\beta_0 - Y\beta_1 \), \( \beta_0, \beta_1 \geq 0 \), the optimal parameters are given by

\[
\beta_0 = \sqrt{\frac{pD}{1-p}}, \quad \beta_1 = \sqrt{\frac{(1-p)D}{p}}.
\]

Let \( \alpha = 2^{\mu-(\sqrt{\frac{(1-p)D}{\sigma}} + \sqrt{\frac{pD}{0}})} \). For this optimal scheme, the accuracy of the MAP adversary is

\[
U^{(G)}_{MAP} = pQ\left( -\frac{\alpha}{2} + \frac{1}{\alpha} \ln \left( \frac{1-p}{p} \right) \right) + (1-p)Q\left( -\frac{\alpha}{2} - \frac{1}{\alpha} \ln \left( \frac{1-p}{p} \right) \right).
\]

The proof of Theorem 3 is provided in Appendix C. When \( P(Y = 1) = P(Y = 0) = \frac{1}{2} \), we have \( \beta_0 = \beta_1 = \sqrt{D} \), which implies the optimal privacy mechanism for this particular case is to shift the two Gaussian distribution closer to each other equally by \( \sqrt{D} \) regardless of the variance \( \sigma^2 \). When \( P(Y = 1) = P(Y = 0) = \frac{1}{2} \), the Gaussian distribution with a lower prior probability, in this case, \( X|Y = 0 \), gets shifted \( \frac{p}{1-p} \) times more than \( X|Y = 1 \).

Next, we consider a slightly more complicated case in which \( \beta_0 = \beta_1 = \sqrt{D} \), where \( N \sim N(0, 1) \). Intuitively, this mechanism privatizes the data by shifting the two Gaussian distributions (under \( Y = 0 \) and \( Y = 1 \)) closer to each other and add another Gaussian noise \( N \in N(0, 1) \) scaled by a constant \( \gamma \).

In this case, the MAP probability of classifying the private variable \( Y \) from \( X \) is given by

\[
U^{(G)}_{MAP} = pQ\left( -\frac{\alpha}{2} + \frac{1}{\alpha} \ln \left( \frac{1-p}{p} \right) \right) + (1-p)Q\left( -\frac{\alpha}{2} - \frac{1}{\alpha} \ln \left( \frac{1-p}{p} \right) \right),
\]

where \( \alpha = 2^{\mu-(\beta_0 + \beta_1)} \). Furthermore, the distortion constraint implies that \( (1-p)\beta_0^2 + p\beta_1^2 + \gamma^2 \leq D \).

Theorem 4. For a PDD privatizing scheme given by \( g(X, Y) = X + (1 - Y)\beta_0 - Y\beta_1 + \gamma N \), \( \beta_0, \beta_1 \geq 0, \gamma > 0 \), the optimal parameters \( \beta_0, \beta_1, \gamma \) are given by the solution to

\[
\min_{\beta_0, \beta_1, \gamma} \frac{(2\mu - \beta_0 - \beta_1)}{\gamma^2 + \sigma^2} \quad \text{s.t.} \quad (1-p)\beta_0^2 + p\beta_1^2 + \gamma^2 \leq D
\]

\[
\beta_0, \beta_1 \geq 0
\]

\[
\gamma > 0.
\]

Let \( \alpha = \frac{(2\mu - \beta_0 - \beta_1)}{\gamma^2 + \sigma^2} \). Using this optimal scheme, the accuracy of the MAP adversary is

\[
U^{(G)}_{MAP} = pQ\left( -\frac{\alpha}{2} + \frac{1}{\alpha} \ln \left( \frac{1-p}{p} \right) \right) + (1-p)Q\left( -\frac{\alpha}{2} - \frac{1}{\alpha} \ln \left( \frac{1-p}{p} \right) \right).
\]

Proof. Similar to the proofs of Theorem 2 and 3, we can compute the derivative of \( U^{(G)}_{MAP} \) with respect to \( \alpha \). It is easy to show that \( U^{(G)}_{MAP} \) is also monotonically increasing with \( \alpha \). Therefore, the optimal mechanism is given by the solution to (38). Substituting the optimal solution to (38) into (37) yields the MAP probability of classifying the private variable \( Y \) from \( X \).

Remark: Note that the objective function in (38) is only dependent on \( \beta_1 + \beta_2 \) and \( \gamma \). We define \( \beta = \beta_1 + \beta_2 \). Thus, the above objective function can be written as

\[
\min_{\beta, \gamma} \frac{(2\mu - \beta)}{\gamma^2 + \sigma^2}
\]

It is easy to verify that the determinant of the Hessian of (39) is always equal to a negative value \( -\frac{1}{(\gamma^2 + \sigma^2)^2} \). Therefore, the above optimization problem is non-convex in \( \beta \) and \( \gamma \).

Finally, we consider the PDD Gaussian noise adding privatizing scheme given by \( g(X, Y) = X + (1 - Y)\beta_0 - Y\beta_1 + (1 - Y)\gamma_0 N + Y\gamma_1 N \), where \( N \sim N(0, 1) \). This PDD mechanism is the most general one in the Gaussian noise adding setting and includes the two previous mechanisms. The...
objective of the privatizer is to minimizing the adversary’s probability of correctly guessing $Y$ from $g(X,Y)$ subject to the distortion constraint given by $p((\beta_0)^2 + (\gamma_0)^2) + (1 - p)((\beta_0)^2 + (\gamma_0)^2) \leq D$. As we have discussed in the remark after Theorem 4 the problem becomes non-convex even for the simpler case in which $\gamma_0 = \gamma_1 = \gamma$. In order to obtain the optimal parameters for this case, we first show that the optimal privacy mechanism lies on the boundary of the distortion constraint.

**Proposition 1.** For the privacy mechanism given by $g(X,Y) = X + (1 - Y)\beta_0 - Y \beta_1 + (1 - Y)\gamma_0 N + Y \gamma_1 N$, the optimal parameters $\beta_0^*, \beta_1^*, \gamma_0^*, \gamma_1^*$ satisfy $p((\beta_0^*)^2 + (\gamma_0^*)^2) + (1 - p)((\beta_0^*)^2 + (\gamma_0^*)^2) = D$

**Proof.** We prove by contradiction. Assume that the optimal parameters satisfy $p((\beta_1^*)^2 + (\gamma_1^*)^2) + (1 - p)((\beta_1^*)^2 + (\gamma_1^*)^2) < D$. Let $\tilde{\beta}_1 = \beta_1^* + c$, where $c > 0$ is chosen so that $p((\tilde{\beta}_1)^2 + (\gamma_1^*)^2) + (1 - p)((\beta_0)^2 + (\gamma_0)^2) = D$. Since the probability of correct guessing is monotonically decreasing with $\beta_1$, the resultant probability of correct guessing can only be smaller for replacing $\beta_1^*$ with $\tilde{\beta}_1$. This contradicts with the assumption that $p((\beta_1^*)^2 + (\gamma_1^*)^2) + (1 - p)((\beta_0^*)^2 + (\gamma_0^*)^2) < D$. Using the same type of analysis, we can show that any parameter that deviates from $p((\beta_1)^2 + (\gamma_1)^2) + (1 - p)((\beta_0)^2 + (\gamma_0)^2) = D$ is suboptimal. $\square$

Let $c_0^2 = (\beta_0^*)^2 + (\gamma_0^*)^2$ and $c_1^2 = (\beta_1^*)^2 + (\gamma_1^*)^2$. Since the optimal parameters of the privatizer lie on the boundary of the distortion constraint, we have $p c_1^2 + (1 - p) c_0^2 = D$. This implies $c_1$ and $c_2$ lie on the boundary of an ellipse parametrized by $p$ and $D$. Thus, we have $c_1 = \sqrt{D/p} \frac{\epsilon}{1+c^2}$ and $c_0 = 2 \sqrt{D/p} \frac{\epsilon}{1+c^2}$, where $\epsilon \in [0,1]$. Therefore, the optimal parameters satisfy

$$\frac{(\beta_0^*)^2}{(\beta_1^*)^2} + \frac{(\gamma_0^*)^2}{(\gamma_1^*)^2} = \left[\frac{D}{1-p} \frac{\epsilon}{1+c^2}\right]^2,$$

This implies $\beta_0^*, \gamma_0^*, \gamma_0^*, \gamma_1^*$, $i \in \{0,1\}$ lie on the boundary of two circles parametrized by $D, p, \epsilon$. Thus, we can write $\beta_0^*, \beta_1^*, \gamma_0^*, \gamma_1^*$ as

$$\beta_0^* = 2 \sqrt{D/p} \frac{\epsilon}{1+c^2} \frac{1-w_0}{1+w_0} \quad \beta_1^* = \sqrt{D/p} \frac{\epsilon}{1+c^2} \frac{1-w_1}{1+w_1},$$

$$\gamma_0^* = 2 \sqrt{D/p} \frac{\epsilon}{1+c^2} \frac{w_0}{1+w_0} \quad \gamma_1^* = 2 \sqrt{D/p} \frac{\epsilon}{1+c^2} \frac{w_1}{1+w_1},$$

where $\epsilon, w_0, w_1 \in [0,1]$. The optimal parameters $\beta_0^*, \beta_1^*, \gamma_0^*, \gamma_1^*$ can be computed by a grid search on a cube parametrized by $\epsilon, w_0, w_1 \in [0,1]$ that minimizes the accuracy of the MAP adversary. In the following section, we will use this general PDD Gaussian noise adding privatizing scheme in our-data driven simulations and compare the performance of the privacy mechanisms obtained by both theoretical and data-driven approaches.

### 4.2 Data-driven Approach for Binary Gaussian Mixture Model

To illustrate our data-driven GAP approach, we assume the privatizer only has access to the dataset $D$ but does not know $P(X,Y)$. Finding the optimal privacy mechanism becomes a learning problem. Similar to the data-driven approach for the binary case, we use the negative cross entropy $U_{ML}(h(g(x,\theta),\theta_0),g)$ provided in (11) as a utility function for the adversary. Therefore, for a fixed privatizer parameter $\theta_p$, the adversary learns the optimal parameter $\theta_0^*$ that maximizes $U_{ML}(h(g(x,\theta),\theta_0),g)$. On the other hand, the parameters for the privacy mechanism is obtained by solving (10) with the loss function provided in (11).

We consider the PDD Gaussian noise adding privacy mechanism given by $g(X,Y) = X + (1 - Y)\beta_0 - Y \beta_1 + (1 - Y)\gamma_0 N - Y \gamma_1 N$. Similar to the binary setting, we use two neural networks to model the privatizer and the adversary. As shown in Figure 6 the privatizer is modeled by a two-layer neural network with parameters $\beta_0, \beta_1, \gamma_0, \gamma_1 \in \mathbb{R}$. The adversary, whose goal is to classify $Y$ from privatized data $X$, is modeled by a three-layer neural network classifier. The random noise is drawn from a standard Gaussian distribution $N \sim N(0,1)$.

In order to enforce the distortion constraint, we use the augmented Lagrangian method to penalize the learning objective when the constraint is not satisfied. In the binary Gaussian mixture model setting, the augmented Lagrangian method uses two hyper parameters, namely $\lambda_t$ and
\[ \rho_t \] to approximate the constrained optimization problem by a series of unconstrained problems. Intuitively, a large value of \( \rho_t \) enforces the distortion constraint to be binding, whereas \( \lambda_t \) is an estimate of the Lagrangian multiplier. With the increasing iteration steps \( t \), we update the value of \( \rho_t \) and \( \lambda_t \) to solve a series of unconstrained problems that will eventually converge to the constrained optimization problem.

### Table 1: Datasets

| Dataset | \( P(Y = 1) \) | \( X|Y = 0 \) | \( X|Y = 1 \) |
|---------|----------------|----------------|----------------|
| 1       | 0.5            | \( \mathcal{N}(-3, 1) \) | \( \mathcal{N}(3, 1) \) |
| 2       | 0.5            | \( \mathcal{N}(-3, 4) \) | \( \mathcal{N}(3, 1) \) |
| 3       | 0.75           | \( \mathcal{N}(-3, 1) \) | \( \mathcal{N}(3, 1) \) |
| 4       | 0.75           | \( \mathcal{N}(-3, 4) \) | \( \mathcal{N}(3, 1) \) |

#### 4.3 Illustration of Results

We use synthetic datasets to evaluate our proposed GAP framework. We consider four synthetic datasets with parameters shown in Table 1. Each synthetic dataset used in this experiment contains 20,000 training samples and 2,000 testing samples. We use Tensorflow to train both the privatizer and the adversary using Adam optimizer with learning rate equals to 0.01 and minibatch size equals to 200. The optimal GAP mechanisms obtained by both theoretical and learning approaches are tested by a MAP adversary who is assumed to have access to \( P(X, Y) \) and the privacy mechanism. The performance of each optimal mechanism is measured by the MAP adversary’s probability of correct guessing \( Y \) from \( \hat{X} \).

![Optimal probability of detection w.r.t. different value of D for p=0.5](image1.png)

![Optimal probability of detection w.r.t. different value of D for p=0.75](image2.png)

Figure 7: Privacy-distortion tradeoff for binary Gaussian mixture model
Figure 7a and 7b illustrate the performance of the optimal PDD Gaussian noise adding mechanism against the strong theoretical MAP adversary when $P(Y = 1) = 0.5$ and $P(Y = 1) = 0.75$, respectively. Both theoretical and data-driven approaches are considered in this experiment. It can be seen that the optimal mechanisms obtained by both theoretical and data-driven approaches reduce the classification accuracy of the MAP adversary as the distortion increases. Similar to the binary case, we observe the saturation behavior of the accuracy of the adversary w.r.t. distortion. Furthermore, it is worth pointing out that for the Gaussian private variable case, we also observe that the privacy mechanism obtained through the data-driven approach performs very well when pitted against the MAP adversary (maximum accuracy difference around 6% compared with theoretical approach). In other words, for the binary Gaussian mixture model, the data-driven approach for GAP can generate privacy mechanisms that are comparable in terms of performance, to the theoretical approach, which assumes the privatizer has access to the underlying distribution of the data.

Figure 8 to 13 show the privatizing schemes for different datasets. The intuition of this Gaussian noise adding mechanism is to shift distributions of $\hat{X}|Y = 0$ and $\hat{X}|Y = 1$ closer and scale the variance to preserve privacy. When $P(Y = 0) = P(Y = 1)$ and $\sigma_0 = \sigma_1$, the privatizer shifts and scales the two distributions almost equally. Furthermore, the resultant $\hat{X}|Y = 0$ and $\hat{X}|Y = 1$ have very similar distributions. We also observe that if $P(Y = 0) \neq P(Y = 1)$, the public variable whose corresponding private variable has a lower prior probability gets shifted more. It is also worth mentioning that when $\sigma_0 \neq \sigma_1$, the public variable with a lower variance gets scaled more.

The optimal privacy mechanism obtained from the data-driven approach under different datasets are presented in Table 2 to 5. In each table, $D$ is the maximum allowable distortion. $\beta_0$, $\beta_1$, $\gamma_0$ and $\gamma_1$ are the parameters of privatizer neural network. These learned parameters dictate the statistical model of the privatizer, which is used to privatize the dataset. We use $\text{acc}$ to denote the classification accuracy of the adversary using a test dataset and $\text{xent}$ to denote the converged cross entropy loss of the adversary. Distance is the average distortion $E_P\|X - \hat{X}\|^2$ that results from sanitizing the test dataset via the learned privatization scheme. $P_{\text{detect}}$ is the MAP adversary’s probability of detection under the learned privatization scheme, assuming that the adversary: (a) has access to $P(X,Y)$, (b) has knowledge of the learned privatization scheme, and (c) can compute the MAP detection rule. $P_{\text{detect-theory}}$ is the “smallest” probability of detection we get if the privatizer had access to $P(X,Y)$ and used this information to compute the parameters of the privatization scheme based on the approach provided at the end of Section 4.1.2. As discussed in the previous subsections, the MAP adversary is aware that the privatizer has access to $P(X,Y)$.

![Histogram of X](image)

Figure 8: raw test samples, equivalent variance
Figure 9: prior $P(Y = 1) = 0.5$, $X|Y = 1 \sim N(3, 1)$, $X|Y = 0 \sim N(-3, 1)$

Figure 10: prior $P(Y = 1) = 0.75$, $X|Y = 1 \sim N(3, 1)$, $X|Y = 0 \sim N(-3, 1)$

Figure 11: raw test samples, inequivalent variance

Figure 12: prior $P(Y = 1) = 0.5$, $X|Y = 1 \sim N(3, 1)$, $X|Y = 0 \sim N(-3, 4)$
Figure 13: prior $P(Y = 1) = 0.75$, $X|Y = 1 \sim N(3, 1)$, $X|Y = 0 \sim N(-3, 4)$

Table 2: prior $P(Y = 1) = 0.5$, $X|Y = 1 \sim N(3, 1)$, $X|Y = 0 \sim N(-3, 1)$

| $D$ | $\beta_0$ | $\beta_1$ | $\gamma_0$ | $\gamma_1$ | acc | xent distance | $P_{\text{detect}}$ | $P_{\text{detect-theory}}$ |
|-----|------------|------------|-------------|-------------|-----|---------------|-----------------|-------------------|
| 1   | 0.5214     | 0.5214     | 0.7797      | 0.7797      | 0.9742 | 0.0715        | 0.9776          | 0.9747            |
| 2   | 0.9861     | 0.9861     | 1.0028      | 1.0029      | 0.9169 | 0.1974        | 1.9909          | 0.9225            |
| 3   | 1.3819     | 1.3819     | 1.0405      | 1.0403      | 0.8633 | 0.3130        | 3.0013          | 0.8689            |
| 4   | 1.5713     | 1.5713     | 1.2249      | 1.2249      | 0.8123 | 0.4066        | 4.0136          | 0.8169            |
| 5   | 1.8199     | 1.8199     | 1.3026      | 1.3024      | 0.7545 | 0.4970        | 4.9894          | 0.7638            |
| 6   | 1.9743     | 1.9745     | 1.436       | 1.4359      | 0.7122 | 0.5564        | 5.9698          | 0.7211            |
| 7   | 2.5332     | 2.5332     | 0.7499      | 0.7500      | 0.6391 | 0.6326        | 7.0149          | 0.6384            |
| 8   | 2.8284     | 2.8284     | 0.0044      | 0.0028      | 0.5727 | 0.6787        | 7.9857          | 0.6384            |
| 9   | 2.9999     | 3.0000     | 0.0003      | 0.0004      | 0.4960 | 0.6938        | 8.9983          | 0.5000            |

Table 3: prior $P(Y = 1) = 0.75$, $X|Y = 1 \sim N(3, 1)$, $X|Y = 0 \sim N(-3, 1)$

| $D$ | $\beta_0$ | $\beta_1$ | $\gamma_0$ | $\gamma_1$ | acc | xent distance | $P_{\text{detect}}$ | $P_{\text{detect-theory}}$ |
|-----|------------|------------|-------------|-------------|-----|---------------|-----------------|-------------------|
| 1   | 0.8094     | 0.2698     | 0.844       | 0.8963      | 0.9784 | 0.0591        | 0.9533          | 0.9731            |
| 2   | 1.4998     | 0.5000     | 0.9676      | 1.1612      | 0.9314 | 0.1635        | 1.9098          | 0.9271            |
| 3   | 0.9808     | 0.3269     | 1.3630      | 1.5762      | 0.911  | 0.2054        | 2.9833          | 0.9205            |
| 4   | 2.2611     | 0.7536     | 1.1327      | 1.6225      | 0.8559 | 0.3519        | 4.0559          | 0.8355            |
| 5   | 2.5102     | 0.8368     | 1.0724      | 1.8666      | 0.792  | 0.401         | 5.0445          | 0.7963            |
| 6   | 2.8238     | 0.9412     | 1.2894      | 1.9752      | 0.7627 | 0.4559        | 6.0843          | 0.7643            |
| 7   | 3.2148     | 1.0718     | 0.6938      | 2.1403      | 0.7500 | 0.4468        | 7.0131          | 0.7500            |
| 8   | 3.3955     | 1.1320     | 1.0256      | 2.2789      | 0.7500 | 0.4799        | 8.0484          | 0.7500            |
| 9   | 4.1639     | 1.3878     | 0.0367      | 2.0714      | 0.7500 | 0.4745        | 8.9343          | 0.7500            |

Table 4: prior $P(Y = 1) = 0.5$, $X|Y = 1 \sim N(3, 1)$, $X|Y = 0 \sim N(-3, -4)$

| $D$ | $\beta_0$ | $\beta_1$ | $\gamma_0$ | $\gamma_1$ | acc | xent distance | $P_{\text{detect}}$ | $P_{\text{detect-theory}}$ |
|-----|------------|------------|-------------|-------------|-----|---------------|-----------------|-------------------|
| 1   | 0.8660     | 0.8660     | 0.0079      | 0.7074      | 0.9122 | 0.2103        | 1.0078          | 0.9107            |
| 2   | 1.2781     | 1.2781     | 0.0171      | 0.8560      | 0.8595 | 0.3239        | 2.0181          | 0.8550            |
| 3   | 1.5146     | 1.5146     | 0.0278      | 1.1352      | 0.8084 | 0.4211        | 3.0264          | 0.8042            |
| 4   | 1.7587     | 1.7587     | 0.0330      | 1.2857      | 0.7557 | 0.4970        | 4.0274          | 0.7554            |
| 5   | 2.0923     | 2.0923     | 0.0142      | 1.0028      | 0.7057 | 0.5589        | 5.0082          | 0.7113            |
| 6   | 2.3079     | 2.2572     | 0.0211      | 1.1185      | 0.6650 | 0.5999        | 6.0377          | 0.6676            |
| 7   | 2.5351     | 2.5351     | 0.0567      | 1.0715      | 0.6100 | 0.6509        | 7.0125          | 0.6225            |
| 8   | 2.7056     | 2.7056     | 0.0358      | 1.1665      | 0.5770 | 0.6738        | 8.0088          | 0.5868            |
| 9   | 2.8682     | 2.8682     | 0.0564      | 1.2435      | 0.5445 | 0.6844        | 9.0427          | 0.5601            |

22
Table 5: prior $P(Y = 1) = 0.75$, $X|Y = 1 \sim N(3, 1)$, $X|Y = 0 \sim N(-3, 4)$

| $D$ | $\beta_0$ | $\beta_1$ | $\gamma_0$ | $\gamma_1$ | $\text{acc}$ | $\text{xent}$ | $\text{distance}$ | $P_{\text{detect}}$ | $P_{\text{detect-theory}}$ |
|-----|-----------|-----------|-----------|-----------|-----------|-----------|----------------|----------------|----------------|
| 1   | 0.8214    | 0.2739    | 0.0401    | 1.0167    | 0.9514    | 0.1357    | 0.9909 | 0.9448       | 0.9328        |
| 2   | 1.4164    | 0.4722    | 0.0583    | 1.2959    | 0.9026    | 0.2402    | 2.0257 | 0.9033       | 0.8891        |
| 3   | 2.2354    | 0.7450    | 0.0246    | 1.3335    | 0.8665    | 0.3354    | 2.9617 | 0.8514       | 0.8481        |
| 4   | 2.6076    | 0.8693    | 0.0346    | 1.5199    | 0.8269    | 0.4034    | 3.9522 | 0.8148       | 0.8120        |
| 5   | 2.9919    | 0.9977    | 0.0143    | 1.6399    | 0.7885    | 0.4625    | 5.0034 | 0.7833       | 0.7824        |
| 6   | 3.3079    | 1.1027    | 0.0094    | 1.7707    | 0.7616    | 0.5013    | 6.0022 | 0.7606       | 0.7500        |
| 7   | 3.1458    | 1.0488    | 0.0565    | 2.1606    | 0.7496    | 0.4974    | 7.0091 | 0.7500       | 0.7500        |
| 8   | 3.9707    | 1.3237    | 0.0142    | 1.9129    | 0.7500    | 0.5470    | 7.9049 | 0.7500       | 0.7500        |
| 9   | 4.0835    | 1.3613    | 0.0625    | 2.1364    | 0.7500    | 0.5489    | 8.8932 | 0.7500       | 0.7500        |

5 Concluding Remarks

We have presented a unified framework for context-aware privacy called generative adversarial privacy (GAP). GAP allows the data holder to learn the privatization mechanism directly from the dataset (to be published) without requiring access to the data distribution. Under GAP, finding the optimal privacy mechanism is formulated as a game between two players: a privatizer and an adversary. An iterative minimax algorithm is proposed to obtain the optimal mechanism under the GAP framework.

To evaluate the performance of the proposed GAP model, we have mostly focused on two types of datasets: (i) binary data model; and (ii) binary Gaussian mixture model. For both cases, the optimal GAP mechanisms are learned using cross-entropy as the adversarial loss function. For each type of dataset, both private data dependent and private data independent mechanisms are studied. These results are cross-validated against the privacy guarantees obtained by computing the information-theoretically optimal mechanism for a strong MAP adversary. In the MAP adversary setting, we have shown that for the binary data model, the optimal GAP mechanism is obtained by solving a linear programming. For the binary Gaussian mixture model, the optimal additive Gaussian noise privatizing scheme is determined. Simulations with synthetic datasets for both types (i) and (ii) show that the privacy mechanisms learned via the GAP framework perform as well as the mechanisms obtained from theoretical computation.

Binary and Gaussian models are canonical models with a wide range of applications. However, moving next, we would like to consider more sophisticated dataset models that can capture real life signals (such as time series data and images). The generative models we have considered in this paper were tailored to the statistics of the datasets. In the future, we would like to experiment with the idea of using a deep generative model to automatically generate the sanitized data. Another straightforward extension to our work is to use the GAP framework to obtain data-driven mutual information privacy mechanisms. Finally, it would be interesting to develop adversarial utility functions that allow sliding from weak to strong adversaries.

A Proof of Theorem 1

Proof. If $q = \frac{1}{2}$, then $X$ is independent of $Y$, the optimal solution is given by any $s_0$, $s_1$ that satisfies the distortion constraint $\{(s_0, s_1)|ps_1 + (1 - p)s_0 \geq 1 - D, s_0, s_1 \in [0, 1]\}$ since $X$ and $Y$ are already independent. If $q \neq \frac{1}{2}$, since each maximum in (27) can only be one of the two values (i.e., the utility of guessing $Y = 0$ or $Y = 1$), the objective function in of the privatizer is determined by the relationship between $P(Y = 1, \hat{X} = i)$ and $P(Y = 0, \hat{X} = i), i \in \{0, 1\}$. Therefore, the optimization problem in (27) can be decomposed into the following four subproblems:

**Subproblem 1:** $P(Y = 1, \hat{X} = 0) \geq P(Y = 0, \hat{X} = 0)$ and $P(Y = 1, \hat{X} = 1) \leq P(Y = 0, \hat{X} = 1)$, which implies $p(1 - 2q)(1 - s_1) - (1 - p)(1 - 2q)s_0 \geq 0$ and $(1 - p)(1 - 2q)(1 - s_0) - p(1 - 2q)s_1 \geq 0$. As a result, the objective of the privatizer is given by $P(Y = 1, \hat{X} = 0) + P(Y = 0, \hat{X} = 1)$. Thus,
the optimization problem in (27) can be written as

$$\min_{s_0, s_1} \ (2q - 1)[ps_1 + (1 - p)s_0] + 1 - q$$

s.t. \quad 0 \leq s_0 \leq 1
\quad 0 \leq s_1 \leq 1
\quad p(1 - 2q)s_1 + (1 - p)(1 - 2q)s_0 \leq p(1 - 2q)
\quad p(1 - 2q)s_1 + (1 - p)(1 - 2q)s_0 \leq (1 - p)(1 - 2q)
\quad -ps_1 - (1 - p)s_0 \leq D - 1.

- If $1 - 2q > 0$, i.e., $q < \frac{1}{2}$, we have $ps_1 + (1 - p)s_0 \leq p$ and $ps_1 + (1 - p)s_0 \leq 1 - p$. The privatizer must maximize $ps_1 + (1 - p)s_0$ to reduce the probability of correctly guessing the private variable. Thus, if $1 - D \leq \min\{p, 1 - p\}$, the optimal value is given by $(2q - 1) \min\{p, 1 - p\} + 1 - q$; the corresponding optimal solution is given by $\{s_0, s_1|ps_1 + (1 - p)s_0 = \min\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$ otherwise, the problem is infeasible.

- If $1 - 2q < 0$, i.e., $q > \frac{1}{2}$, we have $ps_1 + (1 - p)s_0 \geq p$ and $ps_1 + (1 - p)s_0 \geq 1 - p$. In this case, the privatizer has to minimize $ps_1 + (1 - p)s_0$. Thus, if $1 - D > \max\{p, 1 - p\}$, the optimal value is given by $(2q - 1)(1 - D) + 1 - q$; the corresponding optimal solution is $\{s_0, s_1|ps_1 + (1 - p)s_0 = 1 - D, 0 \leq s_0, s_1 \leq 1\}$. Otherwise, the optimal value is $(2q - 1) \max\{p, 1 - p\} + 1 - q$ and the corresponding optimal solution is given by $\{s_0, s_1|ps_1 + (1 - p)s_0 = \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$.

**Subproblem 2:** $P(Y = 1, \hat{X} = 0) \leq P(Y = 0, \hat{X} = 0)$ and $P(Y = 1, \hat{X} = 1) \geq P(Y = 0, \hat{X} = 1)$ which implies, $p(1 - 2q)(1 - s_1) - (1 - p)(1 - 2q)s_0 \leq 0$ and $(1 - p)(1 - 2q)(1 - s_0) - p(1 - 2q)s_1 \leq 0$. Thus, the objective of the privatizer is given by $P(Y = 0, \hat{X} = 0) + P(Y = 1, \hat{X} = 1)$. Therefore, the optimization problem in (27) can be written as

$$\min_{s_0, s_1} \ (1 - 2q)[ps_1 + (1 - p)s_0] + q$$

s.t. \quad 0 \leq s_0 \leq 1
\quad 0 \leq s_1 \leq 1
\quad -p(1 - 2q)s_1 - (1 - p)(1 - 2q)s_0 \leq -p(1 - 2q)
\quad -p(1 - 2q)s_1 - (1 - p)(1 - 2q)s_0 \leq -(1 - p)(1 - 2q)
\quad -ps_1 - (1 - p)s_0 \leq D - 1.

- If $1 - 2q > 0$, i.e., $q < \frac{1}{2}$, we have $ps_1 + (1 - p)s_0 \geq p$ and $ps_1 + (1 - p)s_0 \geq 1 - p$. The privatizer need to minimize $ps_1 + (1 - p)s_0$ to reduce the probability of correctly guessing the private variable. Thus, if $1 - D > \max\{p, 1 - p\}$, the optimal value is given by $(1 - 2q)(1 - D) + q$; the corresponding optimal solution is $\{s_0, s_1|ps_1 + (1 - p)s_0 = 1 - D, 0 \leq s_0, s_1 \leq 1\}$. Otherwise, the optimal value is $(1 - 2q) \max\{p, 1 - p\} + q$ and the corresponding optimal solution is given by $\{s_0, s_1|ps_1 + (1 - p)s_0 = \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$.

- If $1 - 2q < 0$, i.e., $q > \frac{1}{2}$, we have $ps_1 + (1 - p)s_0 \leq p$ and $ps_1 + (1 - p)s_0 \leq 1 - p$. In this case, the privatizer need to maximize $ps_1 + (1 - p)s_0$. Thus, if $1 - D \leq \min\{p, 1 - p\}$, the optimal value is given by $(1 - 2q)\min\{p, 1 - p\} + q$; the corresponding optimal solution is given by $\{s_0, s_1|ps_1 + (1 - p)s_0 = \min\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$. Otherwise, the problem is infeasible.

**Subproblem 3:** $P(Y = 1, \hat{X} = 0) \geq P(Y = 0, \hat{X} = 0)$ and $P(Y = 1, \hat{X} = 1) \geq P(Y = 0, \hat{X} = 1)$, we have $p(1 - 2q)(1 - s_1) - (1 - p)(1 - 2q)s_0 \geq 0$ and $(1 - p)(1 - 2q)(1 - s_0) - p(1 - 2q)s_1 \leq 0$. Under this scenario, the objective function in (27) is given by $P(Y = 1, \hat{X} = 0) + P(Y = 1, \hat{X} = 1)$. Thus, the privatizer solves

$$\min_{s_0, s_1} \ p(1 - q) + (1 - p)q$$

s.t. \quad 0 \leq s_0 \leq 1
\quad 0 \leq s_1 \leq 1
\quad p(1 - 2q)s_1 + (1 - p)(1 - 2q)s_0 \leq p(1 - 2q)
\quad -p(1 - 2q)s_1 - (1 - p)(1 - 2q)s_0 \leq -(1 - p)(1 - 2q)
\quad -ps_1 - (1 - p)s_0 \leq D - 1.
If $1 - 2q > 0$, i.e., $q < \frac{1}{2}$, the problem becomes infeasible for $p < \frac{1}{2}$. For $p > \frac{1}{2}$, if $1 - D > \max\{p, 1 - p\}$, the problem is also infeasible; if $\min\{p, 1 - p\} \leq 1 - D \leq \max\{p, 1 - p\}$, the optimal value is given by $p(1 - q) + (1 - p)q$ and the corresponding optimal solution is $\{s_0, s_1|1 - D \leq p s_1 + (1 - p) s_0 \leq \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$; otherwise, the optimal value is $p(1 - q) + (1 - p)q$ and the corresponding optimal solution is given by $\{s_0, s_1|\min\{p, 1 - p\} \leq p s_1 + (1 - p) s_0 \leq \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$.

If $1 - 2q < 0$, i.e., $q > \frac{1}{2}$, the problem becomes infeasible for $p > \frac{1}{2}$. For $p < \frac{1}{2}$, if $1 - D > \max\{p, 1 - p\}$, the problem is also infeasible; if $\min\{p, 1 - p\} \leq 1 - D \leq \max\{p, 1 - p\}$, the optimal value is given by $p(1 - q) + (1 - p)q$ and the corresponding optimal solution is $\{s_0, s_1|1 - D \leq p s_1 + (1 - p) s_0 \leq \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$; otherwise, the optimal value is $p(1 - q) + (1 - p)q$ and the corresponding optimal solution is given by $\{s_0, s_1|\min\{p, 1 - p\} \leq p s_1 + (1 - p) s_0 \leq \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$.

**Subproblem 4:** $P(Y = 1, \hat{X} = 0) \leq P(Y = 0, \hat{X} = 0)$ and $P(Y = 1, \hat{X} = 1) \leq P(Y = 0, \hat{X} = 1)$, which implies $p(1 - 2q)(1 - s_1) - (1 - p)(1 - 2q)s_0 \leq 0$ and $(1 - p)(1 - 2q)(1 - s_0) - p(1 - 2q)s_1 \geq 0$. Thus, the optimization problem in (27) is given by

\[
\begin{align*}
\min_{s_0, s_1} & \quad pq + (1 - p)(1 - q) \\
\text{s.t.} & \quad 0 \leq s_0 \leq 1 \\
& \quad 0 \leq s_1 \leq 1 \\
& \quad -p(1 - 2q)s_1 - (1 - p)(1 - 2q)s_0 \leq -p(1 - 2q) \\
& \quad p(1 - 2q)s_1 + (1 - p)(1 - 2q)s_0 \leq (1 - p)(1 - 2q) \\
& \quad ps_1 - (1 - p)s_0 \leq D - 1.
\end{align*}
\]

If $1 - 2q > 0$, i.e., $q < \frac{1}{2}$, the problem becomes infeasible for $p > \frac{1}{2}$. For $p < \frac{1}{2}$, if $1 - D > \max\{p, 1 - p\}$, the problem is also infeasible; if $\min\{p, 1 - p\} \leq 1 - D \leq \max\{p, 1 - p\}$, the optimal value is given by $pq + (1 - p)(1 - q)$ and the corresponding optimal solution is $\{s_0, s_1|1 - D \leq p s_1 + (1 - p) s_0 \leq \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$; otherwise, the optimal value is $pq + (1 - p)(1 - q)$ and the corresponding optimal solution is given by $\{s_0, s_1|\min\{p, 1 - p\} \leq p s_1 + (1 - p) s_0 \leq \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$.

If $1 - 2q < 0$, i.e., $q > \frac{1}{2}$, the problem becomes infeasible for $p < \frac{1}{2}$. For $p > \frac{1}{2}$, if $1 - D > \max\{p, 1 - p\}$, the problem is also infeasible; if $\min\{p, 1 - p\} \leq 1 - D \leq \max\{p, 1 - p\}$, the optimal value is given by $pq + (1 - p)(1 - q)$ and the corresponding optimal solution is $\{s_0, s_1|1 - D \leq p s_1 + (1 - p) s_0 \leq \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$; otherwise, the optimal value is $pq + (1 - p)(1 - q)$ and the corresponding optimal solution is given by $\{s_0, s_1|\min\{p, 1 - p\} \leq p s_1 + (1 - p) s_0 \leq \max\{p, 1 - p\}, 0 \leq s_0, s_1 \leq 1\}$.

Summarizing the analysis above yields Theorem[1] \[ \Box \]

**B Proof of Theorem[2]**

**Proof.** Let us consider $\hat{X} = X + \beta + \gamma N$, where $\beta$ and $\gamma$ are both non-negative. Given the MAP adversary’s probability of detection in [34], the objective of the privatizer is to

\[
\begin{align*}
\min_{\beta, \gamma} & \quad U_{MAP}^{(G)} \\
\text{s.t.} & \quad \beta^2 + \gamma^2 \leq D \\
& \quad \beta, \gamma \geq 0.
\end{align*}
\]
Define $\eta = \frac{1-p}{p}$. The gradient of $U_{MAP}^{(G)}$ w.r.t. $\alpha$ is given by

$$\frac{\partial U_{MAP}^{(G)}}{\partial \alpha} = p \left( - \frac{1}{\sqrt{2\pi}} e^{-\left(-\frac{\alpha + \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2} \right) \left( - \frac{1}{\alpha^2} \ln \eta \right)$$

$$+ (1-p) \left( - \frac{1}{\sqrt{2\pi}} e^{-\left(-\frac{\alpha - \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2} \right) \left( \frac{1}{\alpha^2} \ln \eta \right)$$

$$= \frac{1}{2\sqrt{2\pi}} \left( p e^{-\left(-\frac{\alpha + \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2} + (1-p)e^{-\left(-\frac{\alpha - \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2} \right)$$

$$+ \frac{\ln \eta}{\alpha^2 \sqrt{2\pi}} \left( p e^{-\left(-\frac{\alpha + \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2} - (1-p)e^{-\left(-\frac{\alpha - \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2} \right)$$

Note that

$$\frac{pe^{-\left(-\frac{\alpha + \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2}}{(1-p)e^{-\left(-\frac{\alpha - \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2}} = \frac{p}{1-p} \frac{e^{-\left(-\frac{\alpha + \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2} - e^{-\left(-\frac{\alpha - \frac{1}{2} \ln \eta}{\sqrt{2}}\right)^2}}{e^{2\ln \eta}} = \frac{p}{1-p} e^{2\ln \eta} = 1.$$

Therefore, the second term in (48) is 0. Furthermore, the first term in (48) is always positive. Thus, $U_{MAP}^{(G)}$ is monotonically increasing in $\alpha$. As a result, the optimization problem in (46) is equivalent to

$$\max_{\beta, \gamma} \sqrt{\gamma^2 + \sigma^2}$$

$$\text{s.t.} \quad \beta^2 + \gamma^2 \leq D$$

$$\beta, \gamma \geq 0.$$

Therefore the optimal solution is obtained at $\beta = 0$ and $\gamma = \sqrt{D}$. Substituting the optimal solution back into (54) yields the MAP probability of correctly guessing the private variable $Y$ from $X$. \(\square\)

### C Proof of Theorem 3

**Proof.** Let us consider $X = X + (1 - Y)\beta_0 - Y\beta_1$, where $\beta_0$ and $\beta_1$ are both non-negative. Given the MAP adversary’s probability of detection $U_{MAP}^{(G)}$, the objective of the privatizer is to

$$\min_{\beta_0, \beta_1} U_{MAP}^{(G)}$$

$$\text{s.t.} \quad (1-p)\beta_0^2 + p\beta_1^2 \leq D$$

$$\beta_0, \beta_1 \geq 0.$$

As we have shown in Appendix B, $U_{MAP}^{(G)}$ is monotonically increasing in $\alpha = \frac{2\mu - (\beta_1 + \beta_0)}{\sqrt{\gamma^2 + \sigma^2}}$. As a result, the optimization problem in (51) is equivalent to

$$\max_{\beta_0, \beta_1} \beta_1 + \beta_0$$

$$\text{s.t.} \quad (1-p)\beta_0^2 + p\beta_1^2 \leq D$$

$$\beta_0, \beta_1 \geq 0.$$

Note that the above optimization problem is convex. Therefore, using the KKT conditions, we obtain the optimal solution

$$\beta_0 = \sqrt{\frac{pD}{1-p}}, \quad \beta_1 = \sqrt{\frac{(1-p)D}{p}}.$$

Substituting the above optimal solution into $U_{MAP}^{(G)}$ yields the MAP probability of classifying the private variable $Y$ from $X$. \(\square\)
References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro,
Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale
machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467,
2016.

[2] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information
bottleneck. In International Conference on Learning Representations, 2017. URL https://
arxiv.org/abs/1612.00410.

[3] S. Asoodeh, F. Alajaji, and T. Linder. Notes on information-theoretic privacy. In 2014 52nd
Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages
1272–1278, Sept 2014. doi: 10.1109/ALLERTON.2014.7028602.

[4] S. Asoodeh, F. Alajaji, and T. Linder. On maximal correlation, mutual information and data
privacy. In Information Theory (CWIT), 2015 IEEE 14th Canadian Workshop on, pages
27–31, July 2015.

[5] S. Asoodeh, F. Alajaji, and T. Linder. Privacy-aware MMSE estimation. In 2016 IEEE
International Symposium on Information Theory (ISIT), pages 1989–1993, July 2016. doi:
10.1109/ISIT.2016.7541647.

[6] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder. Privacy-aware guessing efficiency. In 2017
IEEE International Symposium on Information Theory (ISIT), pages 754–758, June 2017.
doi: 10.1109/ISIT.2017.800629.

[7] Shahab Asoodeh, Mario Diaz, Fady Alajaji, and Tamás Linder. Information extraction under
privacy constraints. Information, 7(1):15, 2016.

[8] Shahab Asoodeh, Mario Diaz, Fady Alajaji, and Tamas Linder. Estimation efficiency under
privacy constraints. arXiv:1707.02409, 2017.

[9] Y. O. Basciftci, Y. Wang, and P. Ishwar. On privacy-utility tradeoffs for constrained data
release mechanisms. In 2016 Information Theory and Applications Workshop (ITA), pages
1–6, Jan 2016. doi: 10.1109/ITA.2016.7888175.

[10] F. P. Calmon and N. Fawaz. Privacy against statistical inference. In Communication,
Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on, pages
1401–1408, 2012.

[11] Flávio Pin Calmon, Mayank Varia, Muriel Médard, Mark M. Christiansen, Ken R. Duffy, and
Stefano Tessaro. Bounds on inference. In Proc. 51st Annual Allerton Conf. on Commun.,
Control, and Comput., pages 567–574. IEEE, 2013.

[12] Flávio Pin Calmon, Mayank Varia, and Muriel Médard. On information-theoretic metrics for
symmetric-key encryption and privacy. In Proc. 52nd Annual Allerton Conf. on Commun.,
Control, and Comput., 2014.

[13] Flávio Pin Calmon, Ali Makhdoumi, and Muriel Médard. Fundamental limits of perfect
privacy. In Proc. International Symp. on Info. Theory, 2015.

[14] Thomas M Cover and Joy A Thomas. Elements of information theory. John Wiley & Sons,
2012.

[15] John Duchi, Michael Jordan, and Martin Wainwright. Local privacy and statistical minimax
rates. In Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on,
pages 429–438. IEEE, 2013.

[16] John Duchi, Martin J Wainwright, and Michael I Jordan. Local privacy and minimax bounds:
Sharp rates for probability estimation. In Advances in Neural Information Processing Systems,
pages 1529–1537, 2013.

[17] John Duchi, Martin Wainwright, and Michael Jordan. Minimax optimal procedures for locally
private estimation. arXiv preprint arXiv:1604.02390, 2016.
[18] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 429–438. IEEE, 2013.

[19] C. Dwork. Differential privacy. In *Proc. 33rd Intl. Colloq. Automata, Lang., Prog.*, Venice, Italy, July 2006.

[20] C. Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation: Lecture Notes in Computer Science*. New York:Springer, April 2008.

[21] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL [http://dx.doi.org/10.1561/0400000042](http://dx.doi.org/10.1561/0400000042)

[22] The Economist. The world’s most valuable resource is no longer oil, but data. *The Economist*, 2017.

[23] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[24] EUGDPR. The EU general data protection regulation (GDPR). [http://www.eugdpr.org/](http://www.eugdpr.org/), 2017. [http://www.eugdpr.org/](http://www.eugdpr.org/).

[25] Stephen E. Fienberg, Alessandro Rinaldo, and Xiaolin Yang. *Differential Privacy and the Risk-Utility Tradeoff for Multi-dimensional Contingency Tables*, pages 187–199. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-15838-4. doi: 10.1007/978-3-642-15838-4_17. URL [https://doi.org/10.1007/978-3-642-15838-4_17](https://doi.org/10.1007/978-3-642-15838-4_17)

[26] Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci*, 18(11): 1664–1671, 2015. ISSN 1097-6256. Article.

[27] Luis G. Sanchez Giraldo and Jose C. Principe. Rate-distortion auto-encoders. arXiv:1312.7381, 2013.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[29] Jihun Hamm. Minimax filter: Learning to preserve privacy from inference attacks. *arXiv preprint arXiv:1610.03577*, 2016.

[30] Arif Harmanci and Mark Gerstein. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat Meth*, 13(3):251–256, 2016. ISSN 1548-7091. Article.

[31] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro. LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks. *ArXiv e-prints*, 2017.

[32] Geoffrey E Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009.

[33] I. Issa and A. B. Wagner. Operational definitions for some common information leakage metrics. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 769–773, June 2017. doi: 10.1109/ISIT.2017.8006632.

[34] Ibrahim Issa, Sudeep Kamath, and Aaron B. Wagner. An operational measure of information leakage. In *2016 Annual Conference on Information Science and Systems, CISS 2016, Princeton, NJ, USA, March 16-18, 2016*, pages 234–239, 2016. doi: 10.1109/CISS.2016.7460507. URL [http://dx.doi.org/10.1109/CISS.2016.7460507](http://dx.doi.org/10.1109/CISS.2016.7460507)

[35] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML’16*, pages 2436–2444. JMLR.org, 2016. URL [http://dl.acm.org/citation.cfm?id=3045390.3045647](http://dl.acm.org/citation.cfm?id=3045390.3045647)
[36] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Journal of Machine Learning Research*, 2016.

[37] K. Kalantari, O. Kosut, and L. Sankar. On the fine asymptotics of information theoretic privacy. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 532–539, Sept 2016. doi: 10.1109/ALLERTON.2016.7832277.

[38] K. Kalantari, O. Kosut, and L. Sankar. Information-theoretic privacy with general distortion constraints. arXiv:1708.05468, August 2017.

[39] K. Kalantari, L. Sankar, and O. Kosut. On information-theoretic privacy with general distortion cost functions. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2865–2869, June 2017. doi: 10.1109/ISIT.2017.8007053.

[40] Vishesh Karwa and Aleksandra Slavković. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *The Annals of Statistics*, 44(1):87–112, 2016.

[41] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[42] J. Liao, L. Sankar, V. F. Tan, and F. du Pin Calmon. Hypothesis testing in the high privacy regime. In *54th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, Sep. 28-30, 2016*, 2016.

[43] Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Deepprotect: Enabling inference-based access control on mobile sensing applications. arXiv:1702.06159, 1987.

[44] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[45] K. R. Moon, K. Sricharan, and A. O. Hero. Ensemble estimation of mutual information. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3030–3034, June 2017. doi: 10.1109/ISIT.2017.8007086.

[46] B. Moraffah and L. Sankar. Privacy-guaranteed two-agent interactions using information-theoretic mechanisms. *IEEE Transactions on Information Forensics and Security*, 12(9):2168–2183, Sept 2017. ISSN 1556-6013. doi: 10.1109/TIFS.2017.2701278.

[47] J Morris. On single-sample robust detection of known signals with additive unknown-mean amplitude-bounded random interference. *IEEE Transactions on Information Theory*, 26(2):199–209, 1980.

[48] J Morris. On single-sample robust detection of known signals with additive unknown-mean amplitude-bounded random interference–ii: The randomized decision rule solution (corresp.). *IEEE Transactions on Information Theory*, 27(1):132–136, 1981.

[49] Joel M Morris and Neville E Dennis. A random-threshold decision rule for known signals with additive amplitude-bounded nonstationary random interference. *IEEE Transactions on Communications*, 38(2):160–164, 1990.

[50] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[51] National Science and Technology Council Networking and Information Technology Research and Development Program. National privacy research strategy. Technical report, Executive Office of the President of The United States, June 2016.

[52] Tan Nguyen and Scott Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *International Conference on Machine Learning*, pages 1085–1093, 2013.

[53] Nisarg Raval, Ashwin Machanavajjhala, and Landon P Cox. Protecting visual secrets using adversarial nets. In *CVPR Workshop Proceedings*, 2017.
[54] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer. From t-Closeness-Like Privacy to Postrandomization via Information Theory. *IEEE Transactions on Knowledge and Data Engineering*, 22(11):1623–1636, November 2010. ISSN 1041-4347. doi: 10.1109/TKDE.2009.190.

[55] William L. Root. Communications through unspecified additive noise. *Information and Control*, 4(1):15–29, 1961.

[56] S. Salamatian, A. Zhang, F. P. Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft. Managing your private and public data: Bringing down inference attacks against your privacy. 9(7):1240–1255, 2015. doi: 10.1109/JSTSP.2015.2442227. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7118663.

[57] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, Technical report, SRI International, 1998.

[58] L. Sankar, S. K. Kar, R. Tandon, and H. V. Poor. Competitive privacy in the smart grid: An information-theoretic approach. In *Smart Grid Communications*, Brusells, Belgium, Oct. 2011.

[59] L. Sankar, S. R. Rajagopalan, and H. V. Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852, 2013.

[60] L. Sankar, S. Raj Rajagopalan, S. Mohajer, and H. V. Poor. Smart meter privacy: A theoretical framework. *IEEE Transactions on Smart Grid*, 4(2):837–846, 2013. doi: 10.1109/TSG.2012.2211046.

[61] Jürgen H Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 1992.

[62] Shlomo Shamai and Sergio Verdu. Worst-case power-constrained noise for binary-input channels. *IEEE Transactions on Information Theory*, 38(5):1494–1511, 1992.

[63] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado University at Boulder Department of Computer Science, 1986.

[64] Mahito Sugiyama and Karsten M Borgwardt. Measuring statistical dependence via the mutual information dimension. *dim*, 10:1, 2013.

[65] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[66] Latanya Sweeney, Akua Abu, and Julia Winn. Identifying participants in the personal genome project by name (a re-identification experiment). 2013.

[67] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple’s implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

[68] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[69] L. Theis, W. Shi, A. Cunningham, and F. Huszár. Lossy image compression with compressive autoencoders. In *International Conference on Learning Representations*, 2017.

[70] Caroline Uhlerop, Aleksandra Slavković, and Stephen E Fienberg. Privacy-preserving data sharing for genome-wide association studies. *The Journal of privacy and confidentiality*, 5(1):137, 2013.

[71] D. Varodayan and A. Khisti. Smart meter privacy using a rechargeable battery: Minimizing the rate of information leakage. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1932–1935, May 2011. doi: 10.1109/ICASSP.2011.5946886.
[72] Sergio Verdú. α-mutual information. In 2015 Information Theory and Applications Workshop (ITA), 2015.

[73] Sergio Verdú. α-mutual information. In Information Theory and Applications Workshop (ITA), 2015, pages 1–6. IEEE, 2015.

[74] Yue Wang, Jaewoo Lee, and Daniel Kifer. Differentially private hypothesis testing, revisited. arXiv preprint arXiv:1511.03376, 2015.

[75] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309):63–69, 1965.

[76] Eric W Weisstein. Normal distribution. 2002.

[77] WWDC 2016. Engineering privacy for your user. https://developer.apple.com/videos/play/wwdc2016/709/, 2016.

[78] Ke Xu, Tongyi Cao, Swair Shah, Crystal Maung, and Haim Schweitzer. Cleaning the null space: A privacy mechanism for predictors. In Proc. AAAI Conference on Artificial Intelligence, 2017. URL https://aaaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14971/14477

[79] H. Yamamoto. A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers. IEEE Trans. Inform. Theory, 29(6):918–923, November 1983.

[80] M. Ye and A. Barg. Optimal schemes for discrete distribution estimation under local differential privacy. In 2017 IEEE International Symposium on Information Theory (ISIT), pages 759–763, June 2017.

[81] Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. arXiv:1702.00610, Feb. 2017.

[82] Fei Yu, Stephen E Fienberg, Aleksandra B Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. Journal of biomedical informatics, 50:133–141, 2014.

[83] Guoqiang Peter Zhang. Neural networks for classification: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30(4):451–462, 2000.

[84] Yan Zhang, Mete Ozay, Zhun Sun, and Takayuki Okatani. Information potential auto-encoders. arXiv:1706.04635, 2017.