arXiv:1610.03577v1 [cs.LG] 12 Oct 2016

Minimax Filter: Learning to Preserve Privacy from Inference Attacks

Jihun Hamm HAMMJ@CSE.OHIO-STATE.EDU
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210, USA

Editor:

Abstract

Preserving privacy of continuous and/or high-dimensional data such as images, videos and audios, can be challenging with syntactic anonymization methods which are designed for discrete attributes. Differential privacy, which provides a more formal definition of privacy, has shown more success in sanitizing continuous data. However, both syntactic and differential privacy are susceptible to inference attacks, i.e., an adversary can accurately infer sensitive attributes from sanitized data. The paper proposes a novel filter-based mechanism which preserves privacy of continuous and high-dimensional attributes against inference attacks. Finding the optimal utility-privacy tradeoff is formulated as a min-diff-max optimization problem. The paper provides an ERM-like analysis of the generalization error and also a practical algorithm to perform the optimization. In addition, the paper proposes an extension that combines minimax filter and differentially-private noisy mechanism. Advantages of the method over purely noisy mechanisms is explained and demonstrated with examples. Experiments with several real-world tasks including facial expression classification, speech emotion classification, and activity classification from motion, show that the minimax filter can simultaneously achieve similar or better target task accuracy and lower inference accuracy, often significantly lower than previous methods.

Keywords: inference attack, empirical risk minimization, minimax optimization, differential privacy, k-anonymity

1. Introduction

Privacy is an important issue when data collected from or related to individuals are analyzed and released to a third party. In response to growing privacy concerns, various methods for privacy-preserving data publishing have been proposed (see Fung et al. (2010) for a review.)

Syntactic anonymization methods, such as k-anonymity (Sweeney, 2002) and l-diversity (Machanavajjhala et al., 2007) focus on anonymization of quasi-identifiers and protection of sensitive attributes in static databases. However, it is known that syntactic anonymization is susceptible to several types of attacks such as the DeFinetti attack (Kifer, 2009). An adversary may be able to accurately *infer* sensitive attributes of individuals from insensitive, sanitized attributes. High-dimensional data also poses a challenge for syntactic anonymization methods. For example, k-anonymity is known to be ineffective for high-dimensional sparse databases (Narayanan and Shmatikov, 2008). In addition, syntactic anonymization methods are designed for discrete attributes. Discretizing continuous attributes by binning

or clustering is unnatural for certain data types such as video, image, audio, and biometric data. Besides, conventional categorization of attributes as {identifiers, quasi-identifiers, sensitive informations} becomes ambiguous with multimedia-type data. For example, image can an identifier if it contains a face or fingerprints. Otherwise, it may be sensitive or insensitive depending on the content. Audio can also be an identifier if it contains identifiable voice or if there is a mention of identifying information such as names, or it can also be either sensitive or insensitive.

Differential privacy (Dwork and Nissim, 2004; Dwork et al., 2006; Dwork, 2006) was proposed to address many weaknesses of syntactic methods (see the discussion by Clifton and Tassa (2013).) Differential privacy has a more formal privacy guarantee than that of syntactic methods, and is applicable to many problems beyond database release (Dwork et al., 2014). In particular, differential privacy can be defined for continuous and/or high-dimensional attributes and for functions (Hall et al., 2013). However, similar to syntactic anonymization, differential privacy is not immune to inference attacks either (Cormode, 2011), as differentially privacy only prevents an adversary from gaining *additional* knowledge by inclusion/exclusion of a participant (Dwork et al., 2014), and not from gaining knowledge from released data itself. Therefore, an adversary can still be able to guess sensitive attributes of participants from differentially-private attributes with some confidence.

To preserve privacy of continuous high dimensional data from inference attacks, this paper proposes a *learning approach* which differs significantly from previous syntactic or differentially-private approaches. Suppose each participant uses a privacy mechanisms to sanitize her data before releasing them to third parities. The sanitization mechanism considered in the paper is non-invertible transformation–called *filter*–of original raw features. An adversary's goal is to infer sensitive or identifying attributes of participants from the released filtered data using machine learning algorithms. The problem can be viewed as a minimax game between participants and an adversary. Participants need a filter that minimizes the maximum accuracy any adversary may achieve in predicting sensitive or identifying attributes. In this context, privacy of filtered data is measured by *expected risk* of adversarial algorithms on specific inference tasks. However, if privacy is the only goal, near-perfect privacy is achievable with a simple mechanism that sends no or garbage data, which has no utility for data analysts at all. To avoid trivial solutions, utility of filtered data needs to be considered as the second goal, which is measured by expected risk of analysts' algorithms on target tasks. Finding an optimal tradeoff between utility and privacy is a central question in privacy research (see Related work.) The paper proposes to minimize the difference of two risks which is essentially a minimax problem, and the corresponding optimal solution will be referred to as **minimax filter**. Since the training procedure can only access empirical risks, the performance of the filter on test data is given in the form of expectation/probability. The paper presents generalization error analysis for empirical minimax optimizers in analogy with the analysis of empirical risk minimizers (ERM). Lastly, finding a minimax filter involves solving a continuous min-diff-max problem The paper discusses several optimization methods including a numerical optimizer by (Kiwiel, 1987) and a new alternating optimization method, based on the classic theorem of Danskin (1967).

The goal of minimax filter is to prevent inference attacks, and its privacy guarantee is quite different from other privacy mechanisms. It is task-dependent and is given in probability/expectation rather than given absolutely. This may be considered weaker than

some others such as differential privacy. Since minimax filter and differential privacy have almost independent goals, it is natural to ask if the two methods can be combined to take advantages of both methods. Consequently, this paper presents an extension of minimax filter called *noisy minimax filter*, which combines the filter with additive noise mechanism to satisfy the differential privacy criterion. Two methods of combination–preprocessing and postprocessing–are proposed (see Fig. 2.) In the preprocessing approach, minimax filter is applied before perturbation to reduce the sensitivity of transformed data, so that the same level of differential privacy is achieved with less noise. However, it requires a trusted curator who trains the minimax filter. In the postprocessing approach, a minimax filter is applied after perturbation. Since postprocessing cannot worsen differential privacy (Dwork et al., 2014), this approach has the advantages of not requiring a trusted curator and no leak of information through the released filter.

Minimax filter and its extensions are evaluated with several real-world tasks including facial expression classification, speech emotion classification, and activity classification from motion. Experiments show that publicly available continuous and high-dimensional data sets are surprisingly susceptible to subject identification attacks, and that minimax filters can reduce the privacy risks to near chance levels without sacrificing utility too much. Experiments with noisy minimax filter also yield intuitive results. Differential privacy and resilience to inference attack are indeed different goals, such that using differentially private mechanism alone to achieve the latter requires a large amount of noise that destroys utility of data. In contrast, minimax filters can suppress inference attack with little loss of utility with or without perturbation. Therefore, adding a small amount of noise to the minimax filter can provide a formal differential privacy to a degree and also high on-average task-dependent utility and privacy against inference attacks.

To summarize, the paper has the following contributions.

- The paper proposes a novel filtering approach which preserves privacy of continuous and high-dimensional attributes against inference attacks. This mechanism is different from previous mechanisms in many ways; in particular, it is a learning-based approach and is task-dependent.
- The paper measures utility and privacy by expected risks, and formulates the utility-privacy tradeoff as a min-diff-max optimization problem. The paper provides an ERM-like analysis of the generalization performance of empirical optimizers.
- The paper presents a practical algorithm which can find minimax filters for a broad family of filters and losses/classifiers. The proposed optimization algorithm and supporting classes are made public on an open-source repository[1].
- The paper proposes preprocessing and postprocessing approaches to combine minimax filter with noisy mechanisms. The resulting combination can achieve resilience to inference attacks and differential privacy at the same time.
- The paper evaluates proposed algorithms on real-world tasks and compare them with representative algorithms from the literature.

The rest of the paper is organized as follows. Sec. 2 compares the proposed approach further with previous approaches. Sec. 3 presents minimax filters and analyzes its general-

1. `https://github.com/jihunhamm/MinimaxFilter`

ization performance on test data. Sec. 4 explains the difficulty of solving general minimax problems, and present a simple alternating optimization algorithm. Sec. 5 presents noisy minimax filters and two methods of perturbation by additive noise. Sec. 6 evaluates minimax filters with three data sets compared to non-minimax approaches and also evaluates noisy minimax filters under varying levels of privacy. Sec. 7 concludes the paper with discussions.

2. Related work

Optimal utility-privacy tradeoff is one of the main questions in privacy research. Utility-privacy tradeoff has particularly been well-studied under differential privacy assumptions (Dwork and Nissim, 2004; Dwork et al., 2006; Dwork, 2006), in the context of the statistical estimation (Smith, 2011; Alvim et al., 2012; Duchi et al., 2013) and learnability (Kasiviswanathan et al., 2011).

Other measures of privacy and utility were also proposed. Information-theoretic quantities were proposed by Sankar et al. (2010); Rebollo-Monedero et al. (2010); du Pin Calmon and Fawaz (2012) who analyzed privacy in terms of the rate-distortion theory in communication. One problem with using mutual information or related quantity is that it is difficult to estimate mutual information of high-dimensional and continuous variables in practice without assuming a simple distribution. In contrast, this paper proposes to use classification or regression risks to measure privacy and utility, which is directly computable from data without making assumptions on the distribution. Regarding the use of risks in this paper, classification error-based quantities have been suggested in the literature (Iyengar, 2002; Brickell and Shmatikov, 2008; Li and Li, 2009). However, privacy in these works is measured either by syntactic anonymity or probabilistic divergence which are main appropriate for discrete attributes. In this paper, privacy and utility are both defined with risks and are therefore directly comparable when defining the tradeoff of the two. Furthermore, the proposed method explicitly preserves privacy against inference attacks, which both syntactic and differentially-private methods are known to be susceptible to (Cormode, 2011).

Most of the aforementioned works focused on the analyses of utility-privacy tradeoff using different measures and assumptions. Few studied efficient algorithms to actively find optimal tradeoff which this paper aims to do. For discrete variables, Krause and Horvitz (2008) studied the NP-hardness of optimal utility-privacy tradeoff in discrete attribute selection, and demonstrated near-optimality of greedy selection. In particular, they used a weighted difference of utility and privacy cost as the joint cost similar to this work. Ghosh et al. (2009) proposed geometric mechanism and linear programming to achieve near-optimal utility for unknown users. However, optimization problems for discrete distributions are quite different from the problems involving high-dimensional and/or continuous distributions.

Algorithms for preserving privacy of high-dimensional face images has been proposed previously. Newton et al. (2005) applies k-anonymity to images; Enev et al. (2012) learns a linear filter using Partial Least Squares to reduce the covariance between filtered data and private labels; Whitehill and Movellan (2012) also learns a linear filter using the log-ratio of the Fisher's Linear Discriminant Analysis metrics. This paper differs from these in

several aspects: it is not limited to linear filters and is applicable to arbitrary differentiable nonlinear filters such as multilayer neural networks; it directly optimizes the utility-privacy risk instead of optimizing heuristic criteria such as covariance differences or LDA log-ratios.

Lastly, the alternating optimization algorithm (Alg. 1) presented in this paper is related to the algorithm proposed by Goodfellow et al. (2014), which solves a different problem. The algorithm in this paper solves a min-diff-max problem to find an optimal utility-privacy tradeoff, while Goodfellow et al. (2014) solve a minimax problem to learn generative models.

Parts of this paper have appeared in conference proceedings (Hamm, 2015, 2016). New materials in this paper include reformulations of concepts and terms, ERM-like analysis of generalization error, new closed-form examples for minimax optimization, and a new alternating optimization algorithm to solve min-diff-max problems. All results in this paper are produced using the new algorithm.

3. Minimax Filter

In this section, minimax filter is formulated and discussed in detail, and its generalization performance is analyzed.

3.1 Formulation

Minimax filter is a non-invertible transformation of raw features/attributes such that the transformed data has optimal utility-privacy tradeoff. Non-invertibility is assumed so that original features are not always recoverable from the filtered data. Let's assume the filter is deterministic; randomize filters will be discussed in Sec. 5. Let $\mathcal{X} \subset \mathbb{R}^D$ be the space of features/attributes as real-valued vectors. Note that discrete attributes can also be represented by real vectors, e.g., by one-hot vector. Let the filter be a map

$$g(x; u) \in G : \mathcal{X} \times \mathcal{U} \to \mathbb{R}^d \tag{1}$$

which continuous in x and is continuously differentiable w.r.t. the parameter $u \in \mathcal{U}$. Given a filtered output $g(x)$, an adversary can make a prediction $h_p(g(x); v)$ of a private variable y which can be an identifier (e.g., subject number) or a sensitive attribute (e.g., religious beliefs). The prediction function $h_p(g(x); v)$ parameterized by $v \in \mathcal{V}$ is assumed to be continuous in v and continuously differentiable w.r.t. the input $g(x)$. The paper proposes to use the expected risk to measure the privacy of filtered output against adversarial inference:

$$f_{\text{priv}}(u, v) \triangleq E[l_p(h_p(g(x; u); v), y)], \tag{2}$$

where $l_p(\cdot)$ is a continuously differentiable loss function. From the assumptions above, f_{priv} is continuously differentiable w.r.t. the filter parameter u.

Trivial solutions to maximize privacy already exist, which are the filters that output random or constant numbers independent of actual data. However, such filters have no utility whatsoever for data analysts. To avoid trivial solutions, it is necessary to consider the secondary goal of maximizing utility. Let z be a target variable such as medical diagnosis that is of interest to the participants/analysts. An analyst can make a prediction $h_u(g(x); w)$ parameterized by $w \in \mathcal{W}$, which is assumed to be continuous in w and continuously differentiable w.r.t. the input $g(x)$. The 'disutility' of the filtered output for the

analyst is also measured by the expected risk

$$f_{\text{util}}(u, w) \triangleq E[l_u(h_u(g(x; u); w), z)], \tag{3}$$

where $l_u(\cdot)$ is a continuously differentiable loss function, such that f_{util} is continuously differentiable w.r.t. the filter parameter u. To facilitate analysis, the paper assumes that the constraint sets $\mathcal{W}$, $\mathcal{V}$, and $\mathcal{U}$ are compact and convex subsets of Euclidean spaces such as a ball with a large but finite radius. Along with the assumption that the filter g and the risks f_{priv} and f_{util} are all continuous, min and max values are bounded and attainable. In addition, the solutions to min or max problems are assumed to be in the interior of $\mathcal{W}$, $\mathcal{V}$, and $\mathcal{U}$, enforced by adding appropriate regularization (e.g, $\lambda\|w\|^2$) to the optimization problems if necessary. For this reason, min or max problems that appear in the paper will be treated as unconstrained and the notations $u \in \mathcal{U}$, $v \in \mathcal{V}$, and $w \in \mathcal{W}$ will be omitted.

Having defined the privacy and utility measure, the goal of a filter designer is to find a filter that achieves two objectives. The first objective is to *maximize privacy*

$$\max_u \min_v f_{\text{priv}}(u, v) \quad \text{(or equivalently,} \quad \min_u \max_v -f_{\text{priv}}(u, v)) \tag{4}$$

where $\min_v f_{\text{priv}}(u, v)$ represents the risk of the worst (i.e., most capable) adversary: the smaller the risk, the more accurately can she infer private variables. As mentioned above, this problem alone has a trivial solution such as a constant filter that outputs zeros. The second objective is to *minimize disutility*

$$\min_u \min_w f_{\text{util}}(u, w) \quad \text{(or equivalently,} \quad \min_u -\max_w -f_{\text{util}}(u, w)) \tag{5}$$

where $\min_w f_{\text{util}}(u, w)$ represents the risk of the best analyst: the smaller the risk, the more accurately can the analyst predict variables of interest. To achieve the two opposing goals, one solves the joint problem which is to minimize a weighted sum

$$\min_u \left[\max_v -f_{\text{priv}}(u, v) + \rho \min_w f_{\text{util}}(u, w)\right], \tag{6}$$

or equivalently a weighted difference

$$\min_u \left[\max_v -f_{\text{priv}}(u, v) - \rho \max_w -f_{\text{util}}(u, w)\right]. \tag{7}$$

The constant $\rho > 0$ determines the relative importance of utility versus privacy. For a small $\rho \ll 1$, the problem is close to a trivial privacy-only task, and for a large $\rho \gg 1$, the problem is close to a utility-only task. The solution to (6) or (7) will be referred to as **Minimax filter** [2] and is by definition an optimal filter for utility-privacy tradeoff in terms of expected risks given the family of filters and the family of losses/classifiers. Note that the choice of filter and loss/classifier families is very flexible, with the assumption of differentiability only. In practice, almost-everywhere differentiability suffices to use the algorithm in the paper. Fig. 1 shows an example filter/classifier from the class of multilayer neural networks. As an aside, the joint problem may be formulated as minimization of disutility with a hard constraint on privacy risk. When using interior-point methods, the procedure is roughly equivalent to solving (7) iteratively with an increasing ρ, which is more demanding than minimizing the weighted sum once as the paper proposes.

2. To be precise, the joint task (6) is a min-diff-max problem and the privacy-only task (4) is a minimax problem. However, both will be referred to as minimax when it is not important to distinguish the two.

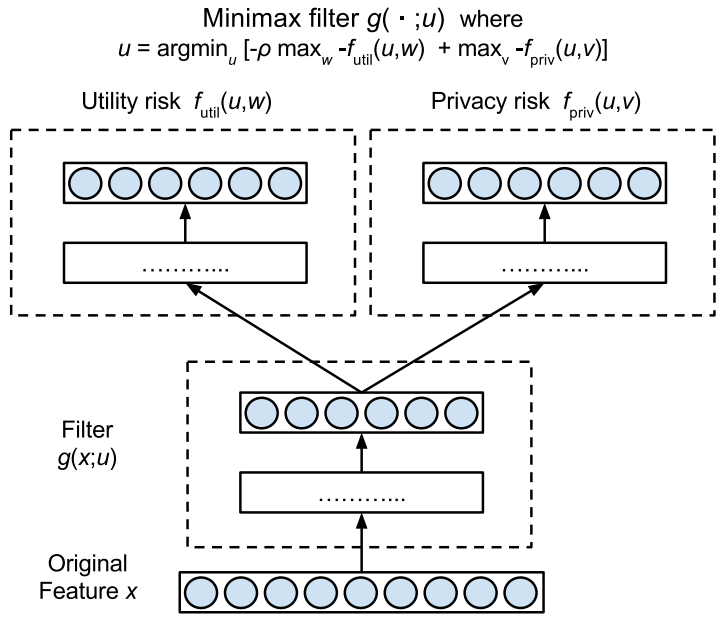

Figure 1: Minimax filter with a filter/classifiers from the class of multilayer neural networks.

3.2 Notes on private and utility tasks

The private variable y can be any attribute which is considered sensitive or identifying. For example, let y be any number or string unique to a person in the data set. Such identifiers are bijective with $\{1, ..., S\}$, where S is the total number of participants, so assume $y \in \{1, ..., S\}$. The private task for an adversary is then to predict the identity y from the filtered data $g(x)$, whose inaccuracy is measured by expected risk. That is, the higher the risk, the more anonymous the filtered output is. The identity variable can also be group identifiers, e.g., y is a demographic grouping based on age, sex, ethnicity, etc. Another example of private tasks is to single-out a particular subject among the rest, in which case y is binary: $y = -1$ means 'not the target subject' and $y = 1$ means 'target subject'. To summarize, anonymity of filtered data in this paper means resilience to inference attacks on any identity variable y that we choose, and is no different from attacks on other non-identity type variables. This *unifying* approach is convenient since we need not determine whether an attribute is an identifier, a quasi-identifier or a sensitive attribute as in syntactic anonymization. Any information hidden in the continuous high-dimensional features which are relevant to the private variable y–whatever it may be–will be filtered out by construction.

Similarly, the target variable z of interest can be any attribute that is not the same as the private variable y. When they are the same ($z = y$), the objective (7) becomes

$$\min_u \left[(1-\rho) \max_v -f(u,v)\right] \tag{8}$$

which is either a trivial privacy-only problem when $0 \leq \rho < 1$, or a utility-only problem when $\rho > 1$. In general, z and y will be correlated to certain degrees, and the minimax filter

will find the best compromise of utility and privacy risks. Also, private and target tasks need not be classification tasks only. Regression tasks can also be used as a target task, as well as unsupervised tasks that do not require label z. Unsupervised tasks are useful when the participants/curator do not know beforehand what tasks analysts will perform at a later time. Without such knowledge, one can use least-squares autoencoding of original features as the target task

$$f_{\text{util}}(u, w) = E[\|h(g(x; u); w) - x\|^2], \tag{9}$$

where $g(x; u)$ and $h(\cdot; w)$ are a encoder and a decoder respectively, such as a nonlinear autoencoder constructed from multilayer neural networks.

3.3 Multiple tasks

Extension to multiple private and target tasks is straightforward. Suppose there are N_p private tasks $f^1_{\text{priv}}(u, v_1), ..., f^{N_p}_{\text{priv}}(u, v_{Np})$ associated with private random variables $y^1, ..., y^{Np}$. Note that $f^i_{\text{priv}}(u, v_i) = E[l^i_p(y^i(g(x; u); v_i), y^i)]$. Similarly, suppose there are N_u target tasks $f^1_{\text{util}}(u, w_1), ..., f^{N_u}_{\text{util}}(u, w_{Nu})$ associated with target random variables $z_1, ..., z_{Nu}$. If $\kappa_1, ..., \kappa_{Np}$ are the coefficients representing relative importance of private tasks, and $\rho_1, ..., \rho_{Nu}$ are the coefficients for utility tasks, then the joint goal is to solve the following problem

$$\min_u \left[\sum_{i=1}^{N_p} \kappa_i \max_{v_i} -f^i_{\text{priv}}(u, v_i) \; + \; \sum_{i=1}^{N_u} \rho_i \min_{w_i} f^i_{\text{util}}(u, w_i) \right]. \tag{10}$$

Since this can be rewritten as

$$\min_u \left[\max_{v_1,...,v_{N_p}} - \sum_{i=1}^{N_p} \kappa_i f^i_{\text{priv}}(u, v_i) \; + \; \min_{w_1,...,w_{N_u}} \sum_{i=1}^{N_u} \rho_i f^i_{\text{util}}(u, w_i) \right], \tag{11}$$

a multiple task problem is nearly identical to the original single task problem (6), with the new utility and privacy tasks defined as

$$\hat{f}_{\text{priv}}(u, v{=}(v_1, ..., v_{Np})) \triangleq \sum_i \kappa_i f^i_{\text{priv}}(u, v_i), \text{ and} \tag{12}$$

$$\hat{f}_{\text{util}}(u, w{=}(w_1, ..., w_{Nu})) \triangleq \sum_i \rho_i f^i_{\text{util}}(u, w_i), \tag{13}$$

with $\rho = 1$. Using this, it is straightforward to extend the analysis and algorithms developed for single tasks to those for multiple tasks.

3.4 Generalization performance of minimax filter

The proposed privacy mechanism is a learning approach. An optimal filter is one that solves the expected risk optimization (6). However, in reality, an optimal filter has to be estimated from finite training samples, and we need a guarantee on the performance of the learned filter on unseen test samples. This section derives generalization bounds for empirical minimax filter, similarly to the bounds for empirical risk minimization (ERM).

The joint problem for expected risks was

$$\min_u \left[\max_v -f_{\text{priv}}(u,v) + \rho \min_w f_{\text{util}}(u,w)\right] = \min_u \left[\max_v -E[l_p(u,v)] + \rho \min_w E[l_u(u,w)]\right]. \tag{14}$$

A joint loss l_J is introduced for convenience:

$$l_J(u,v,w) \triangleq -l_p(u,v) + \rho\, l_u(u,w). \tag{15}$$

Let (u^*, v^*, w^*) be a solution to the expected risk optimization problem:

$$E_D[l_J(u^*,v^*,w^*)] = \min_u \left[\max_v E_D[-l_p(u,v)] + \rho \min_w E_D[l_u(u,w)]\right], \tag{16}$$

where $E_D[\cdot]$ is the expected value w.r.t. the unknown data distribution $P(x,y)$. Similarly, Let $(\hat{u}, \hat{v}, \hat{w})$ be a solution to the empirical risk minimax problem:

$$E_S[l_J(\hat{u},\hat{v},\hat{w})] = \min_u \left[\max_v E_S[-l_p(u,v)] + \rho \min_w E_S[l_u(u,w)]\right], \tag{17}$$

where the empirical mean $E_S[\cdot]$ for $S = \{(x_1,y_1),\cdots,(x_N,y_N)\}$ is

$$E_S[l(x,y)] \triangleq \frac{1}{|S|} \sum_{(x,y)\in S} l(x,y). \tag{18}$$

The goal in this analysis is to show that the expected and the empirical optimizers perform equally well in expectation/probability given enough training samples:

$$E_D[l_J(\hat{u},\hat{v},\hat{w})] \simeq E_D[l_J(u^*,v^*,w^*)], \quad \text{as} \quad |S| \to \infty. \tag{19}$$

The main result is the Theorem 4 which is proved in the remainder of this section. Let's define optimal parameters $v(u)$ and $w(u)$ given u as

$$v^*(u) \triangleq \arg\max_v E_D[-l_p(u,v)], \qquad \hat{v}(u) \triangleq \arg\max_v E_S[-l_p(u,v)], \tag{20}$$

$$w^*(u) \triangleq \arg\min_w E_D[l_u(u,w)], \qquad \hat{w}(u) \triangleq \arg\min_w E_S[l_u(u,w)]. \tag{21}$$

One can then write

$$E_D[l_J(u^*,v^*,w^*)] = \min_u \left[E_D[-l_p(u,v^*(u))] + \rho E_D[l_u(u,w^*(u))]\right] \tag{22}$$

$$= \min_u E_D[l_j(u,v^*(u),w^*(u))], \tag{23}$$

and similarly

$$E_S[l_J(\hat{u},\hat{v},\hat{w})] = \min_u \left[E_S[-l_p(u,\hat{v}(u))] + \rho E_S[l_u(u,\hat{w}(u))]\right] = \min_u E_S[l_j(u,\hat{v}(u),\hat{w}(u))]. \tag{24}$$

From these definitions we have for all u,

$$E_D[l_J(u^*,v^*,w^*)] \leq E_D[l_J(u,v^*(u),w^*(u))], \tag{25}$$

$$E_S[l_J(\hat{u},\hat{v},\hat{w})] \leq E_S[l_J(u,\hat{v}(u),\hat{w}(u))]. \tag{26}$$

Also from definition, for all (u, v, w),

$$E_D[l_J(u, v, w^*(u))] \leq E_D[l_J(u, v, w)] \leq E_D[l_J(u, v^*(u), w)], \tag{27}$$

$$E_S[l_J(u, v, \hat{w}(u))] \leq E_S[l_J(u, v, w)] \leq E_S[l_J(u, \hat{v}(u), w)]. \tag{28}$$

These observations imply the following theorem.

Theorem 1 *The risk difference of expected and empirical optimizers is at most twice of the largest difference of expected and empirical risks of any set of parameters:*

$$|E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_D[l_J(u^*, v^*, w^*)]| \leq 2 \sup_{u,v,w} |E_D[l_J(u, v, w)] - E_S[l_J(u, v, w)]| . \tag{29}$$

Proof The expected risk of empirical risk optimizers $(\hat{u}, \hat{v}, \hat{w})$ is upper-bounded by the risk of expected risk optimizers (u^*, v^*, w^*) as follows:

$$\begin{aligned}
& E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_D[l_J(u^*, v^*, w^*)] \\
= \; & E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})] - (E_D[l_J(u^*, v^*, w^*)] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})]) \\
\leq \; & E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})] - (E_D[l_J(u^*, v^*, w^*)] - E_S[l_J(u^*, \hat{v}(u^*), \hat{w}(u^*))]) \quad \text{(from (26))} \\
\leq \; & E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})] - (E_D[l_J(u^*, \hat{v}(u^*), w^*)] - E_S[l_J(u^*, \hat{v}(u^*), \hat{w}(u^*))]) \quad \text{(from (27))} \\
\leq \; & E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})] - (E_D[l_J(u^*, \hat{v}(u^*), w^*)] - E_S[l_J(u^*, \hat{v}(u^*), w^*)]) \quad \text{(from (28))} \\
\leq \; & 2 \sup_{u,v,w} |E_D[l_J(u, v, w)] - E_S[l_J(u, v, w)]| .
\end{aligned}$$

The difference can also be lower-bounded as follows:

$$\begin{aligned}
& E_D[l_J(u^*, v^*, w^*)] - E_D[l_J(\hat{u}, \hat{v}, \hat{w})] \\
= \; & E_D[l_J(u^*, v^*, w^*)] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})] - (E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})]) \\
\leq \; & E_D[l_J(\hat{u}, v^*(\hat{u}), w^*(\hat{u}))] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})] - (E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})]) \quad \text{(from (25))} \\
\leq \; & E_D[l_J(\hat{u}, v^*(\hat{u}), w^*(\hat{u}))] - E_S[l_J(\hat{u}, v^*(\hat{u}), \hat{w})] - (E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})]) \quad \text{(from (28))} \\
\leq \; & E_D[l_J(\hat{u}, v^*(\hat{u}), \hat{w})] - E_S[l_J(\hat{u}, v^*(\hat{u}), \hat{w})] - (E_D[l_J(\hat{u}, \hat{v}, \hat{w})] - E_S[l_J(\hat{u}, \hat{v}, \hat{w})]) \quad \text{(from (27))} \\
\leq \; & 2 \sup_{u,v,w} |E_D[l_J(u, v, w)] - E_S[l_J(u, v, w)]| .
\end{aligned}$$

■

To bound the RHS of (29), one can use the Rademacher complexity theory (e.g., Lemma 26.2 of Shalev-Shwartz and Ben-David (2014).)

Lemma 2 *Let F be a class of real-valued functions, and let S be a set of N samples $S = \{(x_1, y_1), ..., (x_N, y_N)\}$. Then,*

$$E_{S \sim D^N} \left[\sup_{f \in F} |E_D[f] - E_S[f]| \right] \leq 2 E_{S \sim D^N} [\mathfrak{R}(F \circ S)], \tag{30}$$

where $\mathfrak{R}(F \circ S)$ is the empirical Rademacher complexity

$$\mathfrak{R}(f \circ S) \triangleq \frac{1}{N} E_{\sigma \sim \{-1,+1\}^N} \left[\sup_{f \in F} \sum_{i=1}^{N} \sigma_i f(x_i, y_i) \right] \tag{31}$$

for the class of real-valued functions $\{(x, y) \mapsto f(x, y) \; : \; \forall f \in F\}$.

Consider the class of real-valued functions defined from the joint loss (15):

$$\begin{aligned} l_J \circ H_J \circ S &\triangleq \{(x,y,z) \mapsto l_J(x,y,z;u,v,w) \ : \ u \in \mathcal{U}, v \in \mathcal{V}, w \in \mathcal{W}\} \quad (32) \\ &= \{(x,y,z) \mapsto -l_p(h_v(g_u(x)),y) + \rho \, l_u(h_w(g_u(x)),z) \ : \ u \in \mathcal{U}, v \in \mathcal{V}, w \in \mathcal{W}\}. \end{aligned}$$

Let $\mathfrak{R}(l_J \circ H_J \circ S)$ denote the empirical Rademacher complexity of the joint loss class. Furthermore, the Rademacher complexity of sum of functions can be upper-bounded by the sum of complexities:

Lemma 3 *The empirical Rademacher complexity of the joint privacy-utility loss is upper-bounded as*

$$\mathfrak{R}(l_J \circ H_J \circ S) \leq \mathfrak{R}(l_p \circ H_p \circ G \circ S) + \rho \, \mathfrak{R}(l_u \circ H_u \circ G \circ S), \quad (\rho > 0) \quad (33)$$

where

$$\begin{aligned} l_p \circ H_p \circ G \circ S &\triangleq \{(x,y,z) \mapsto l_p(h_p(g_u(x)),y) \ : \ u \in \mathcal{U}, v \in \mathcal{V}\}, \quad (34) \\ l_u \circ H_u \circ G \circ S &\triangleq \{(x,y,z) \mapsto l_u(h_u(g_u(x)),z) \ : \ u \in \mathcal{U}, w \in \mathcal{W}\}. \quad (35) \end{aligned}$$

Proof

$$\begin{aligned} \mathfrak{R}(l_J \circ H_J \circ S) &= \frac{1}{N} E_\sigma \left[\sup_{u,v,w} \left| \sum_{i=1}^N \sigma_i l_J(x_i,y_i,z_i;u,v,w) \right| \right] \quad (36) \\ &= \frac{1}{N} E_\sigma \left[\sup_{u,v,w} \left| \sum_{i=1}^N \sigma_i (-l_p(x_i,y_i;u,v) + \rho \, l_u(x_i,z_i;u,w)) \right| \right] \quad (37) \\ &\leq \frac{1}{N} E_\sigma \left[\sup_{u,v} \left| \sum_{i=1}^N \sigma_i l_p(x_i,y_i;u,v) \right| + \rho \sup_{u,w} \left| \sum_{i=1}^N \sigma_i l_u(x_i,z_i;u,w) \right| \right] \quad (38) \\ &= \mathfrak{R}(l_p \circ H_p \circ G \circ S) + \rho \, \mathfrak{R}(l_u \circ H_u \circ G \circ S). \quad (39) \end{aligned}$$

■

From Theorem 1 and Lemmas 2 and 3, we get the following generalization bounds in terms of the Rademacher complexity.

Theorem 4

$$\begin{aligned} &E_{S \sim D^m} \left[|E_D[l_J(u^*,v^*,w^*)] - E_D[l_J(\hat{u},\hat{v},\hat{w})]| \right] \\ &\quad \leq 4 E_{S \sim D^m} \left[\mathfrak{R}(l_p \circ H_p \circ G \circ S) + \rho \, \mathfrak{R}(l_u \circ H_u \circ G \circ S) \right]. \quad (40) \end{aligned}$$

A probabilistic bound instead of expected value can also be obtained by applying McDiarmid's inequality, which is omitted.

The Rademacher complexity of privacy and utility losses depends on our choice of loss functions, hypothesis classes, and filter classes. For the simple case of linear filters and linear classifiers, one can compute the complexity using the following lemmas (26.9 and 26.10 from Shalev-Shwartz and Ben-David (2014)):

Lemma 5 *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is α-Lipschitz, i.e., $|\phi(a) - \phi(b)| \leq \alpha|a - b|, \ \forall a, b \in \mathbb{R}$. Then,*

$$\mathfrak{R}(\phi \circ F) = \alpha \mathfrak{R}(F). \tag{41}$$

Lemma 6 *For the class of linear classifiers $H = \{x \mapsto w^T x \ : \ \|w\|_2 \leq 1\}$,*

$$\mathfrak{R}(H \circ S) \leq \frac{1}{\sqrt{N}} \sup_{x \in S} \|x\|_2. \tag{42}$$

.

From these lemmas and Theorem 4, we have a corollary for a simple case of linear filters and classifiers.

Corollary 7 *Let the loss functions l_u and l_p be α-Lipschitz (e.g., $\alpha = 1$ for logistic regression.) Suppose U is a $d \times D$ real matrix with a bounded norm $\|U\|_2 \leq 1$, and w and v are vectors with bounded norms ($\|w\|_2 \leq 1$ and $\|v\|_2 \leq 1$). If the feature domain $\mathcal{X}$ is also bounded with a radius $r = \max_{x \in \mathcal{X}} \|x\|_2$, then we have*

$$|E_D[l_J(u^*, v^*, w^*)] - E_D[l_J(\hat{u}, \hat{v}, \hat{w})]| \leq \frac{4(1+\rho)\ \alpha\ r}{\sqrt{N}}. \tag{43}$$

Alternatively, one can use the VC dimension to specify the bound. In any case, the generalization bounds in this section justify the claim that minimax filter can preserve utility-privacy of not only the current data but also the unseen data in expectation/probability generated from the same unknown distribution.

4. Minimax Optimization

This section presents theoretical and numerical solutions of the joint problem (7), which is a variant of unconstrained continuous minimax problems (see Rustem and Howe (2009) for a review.) The problem (7) can be written in an equivalent form

$$\min_u \Phi(u) = \min_u [\Phi_{\text{priv}}(u) - \rho\ \Phi_{\text{util}}(u)] \tag{44}$$

$$= \min_u [\max_v -f_{\text{priv}}(u, v) - \rho\ \max_w -f_{\text{util}}(u, w)] \tag{45}$$

The optimization above is a min-diff-max problem and can be considered as simultaneously solving two subproblems $\min_u[\max_v -f_{\text{priv}}(u, v)]$ and $\min_u[-\max_w -f_{\text{util}}(u, w)]$, but is evidently not the same as summing individual solutions

$$\min_u \Phi(u) \neq \min_u [\max_v -f_{\text{priv}}(u, v)] \ + \ \min_u [-\rho \max_w -f_{\text{util}}(u, w)]. \tag{46}$$

Since the second subproblem $\min_u[-\max_w -f_{\text{util}}(u, w)] = \min_{u,w} f_{\text{util}}(u, w)$ is a standard minimization problem, let's focus only on the first subproblem $\min_u[\max_v -f_{\text{priv}}(u, v)]$ which is a continuous minimax problem. Continuous minimax problems are in general more challenging to solve than standard minimization problems, as the inner optimization $\Phi_{\text{priv}}(u) = \max_v -f_{\text{priv}}$ does not usually have a closed-form solution; when it does, the whole problem can be treated as a standard minimization problem. Furthermore, there can be more than one solution to $\Phi_{\text{priv}}(u) = \max_v -f_{\text{priv}}$. To better understand minimax problems, we look at several examples starting from a simple case where Φ_{priv} and Φ_{util} have closed-form solutions.

4.1 Simple case: eigenvalue problem

Consider finding a minimax filer for the following problem. The filter class is a linear dimensionality reduction ($g(x;u) = U^T x$) parameterized by the matrix $U \in \mathbb{R}^{D\times d}$, and the private and target tasks are least-squares regressions parameterized by the matrices V and W:

$$f_{\text{priv}}(U,V) = \frac{1}{N}\sum_i \|V^T U^T x_i - y_i\|^2, \text{ and} \tag{47}$$

$$f_{\text{util}}(U,W) = \frac{1}{N}\sum_i \|W^T U^T x_i - z_i\|^2. \tag{48}$$

In this case, $\Phi_{\text{priv}}(U) = \max_V -f_p(U,V)$ and $\Phi_{\text{util}}(U) = \max_W -f_u(U,W)$ are both concave problems with closed-form solutions

$$\hat{V} = \arg\min_V f_{\text{priv}} = (U^T C_{xx} U)^{-1} U^T C_{xy} \text{ and} \tag{49}$$

$$\hat{W} = \arg\min_W f_{\text{util}} = (U^T C_{xx} U)^{-1} U^T C_{xz}, \tag{50}$$

where

$$C_{xy} = \frac{1}{N}\sum_i x_i y_i^T, \;\; C_{xz} = \frac{1}{N}\sum_i x_i z_i^T, \text{ and } C_{xx} = \frac{1}{N}\sum_i x_i x_i^T. \tag{51}$$

The corresponding min values are

$$\Phi_{\text{priv}}(U) = -f_{\text{priv}}(U,\hat{V}) = \text{Tr}\left[(U^T C_{xx} U)^{-1} U^T C_{xy} C_{xy}^T U\right] + \text{const, and} \tag{52}$$

$$\Phi_{\text{util}}(U) = -f_{\text{util}}(U,\hat{W}) = \text{Tr}\left[(U^T C_{xx} U)^{-1} U^T C_{xz} C_{xz}^T U\right] + \text{const.} \tag{53}$$

The outer minimization over u is then

$$\min_U \Phi(U) = \min_U \left[\Phi_{\text{priv}}(U) - \rho\Phi_{\text{util}}(U)\right] \tag{54}$$

$$= \min_U \left[-f_{\text{priv}}(U,\hat{V}) + \rho f_{\text{util}}(U,\hat{W})\right] = \text{Tr}\left[(U^T C_{xx} U)^{-1}\, U^T C_{xyz} U\right], \tag{55}$$

where

$$C_{xyz} = C_{xy} C_{xy}^T - \rho C_{xz} C_{xz}^T. \tag{56}$$

This special case problem is quite similar to the objective of Enev et al. (2012):

$$\max_u \; [-\lambda u^T C_{xy}^T C_{xy} u + u^T C_{xz}^T C_{xz} u], \text{ s. t. } u^T u = 1,$$

The problem (55) can be reformulated as a generalized eigenvalue problem. Let $Q = C_{xx}^{1/2} U$ be a $D \times d$ full-rank matrix. The problem can be rewritten as

$$\min_U \text{Tr}\left[(U^T C_{xx} U)^{-1}\, U^T C_{xyz} U\right] = \min_Q \text{Tr}\left[(Q^T Q)^{-1} Q^T C_{xx}^{-1/2} C_{xyz} C_{xx}^{-1/2} Q\right]. \tag{57}$$

Furthermore, note that min value (55) is invariant to the right multiplication of U by any $d \times d$ nonsingular matrix R. So chose R so that $Q^T Q = R^T U^T C_{xx} U R = I_d$ without loss of

generality. Let $A = (C_{xx}^{-1/2})^T C_{xyz} C_{xx}^{-1/2}$, and the minimax problem becomes the following eigenvalue problem:

$$\min_U \Phi(U) = \min_{\{Q \mid Q^T Q = I_d\}} \text{Tr } Q^T A Q, \tag{58}$$

which is the sum of the d smallest eigenvalues of A. Note that A may not be positive definite.

The paper also proposes a variant of the above eigenvalue problem, called **Privacy LDS** which is an analogue of linear discriminant analysis (LDS) for privacy-utility optimization problem. Define the symmetric positive semidefinite matrix C_u as

$$C_u = \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T, \tag{59}$$

where

$$z \in \{1, ..., K\}, \;\; N_k = \sum_{i=1}^{N} I[z_i = k], \;\; \mu_k = \frac{1}{N_k} \sum_{i=1}^{N} x_i I[z_i = k], \;\; \text{and} \;\; \mu = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{60}$$

Define C_p similarly with

$$y \in \{1, ..., K'\}, \;\; N'_k = \sum_{i=1}^{N'} I[y_i = k], \;\; \text{and} \;\; \mu'_k = \frac{1}{N'_k} \sum_{i=1}^{N'} x_i I[y_i = k]. \tag{61}$$

The proposed Privacy LDS is a linear filter $g(x; W) = W^T x$, where $W = [v_1, ..., v_d]$ is a matrix of top eigenvectors v_i's from the following generalized eigenvalue problem:

$$\max_{\|v\|=1} \frac{v^T (C_u + \lambda I) v}{v^T (C_p + \lambda I) v}. \tag{62}$$

This paper uses Privacy LDS as a heuristic to find the initial linear filter before fine-tuning the parameter u using a general optimization methods presented in the following sections. Note that this initialization applicable only to linear filters.

4.2 Saddle-point problem

Continuous minimax problems in general cannot be solved in closed form and require numerical solvers. There is a subclass of continuous minimax problems which are easier to solve than others. Saddle point problems are minimax problems for which $f(u, v)$ is convex in u and concave in v, such as

$$\min_u \max_v f(u, v) = \min_u \max_v \, [u^2 - v^2]. \tag{63}$$

Analogous to convex problems, $f(u, v)$ has a global optimum (u^*, v^*) which satisfies

$$f(u^*, v) \leq f(u^*, v^*) \leq f(u, v^*). \tag{64}$$

The presence of global optima in saddle-point problems makes numerical optimization practical. Convergence rate of a simple subgradient-descent method for saddle-point problems

was previously analyzed by Nedić and Ozdaglar (2009). Unfortunately, the minimax problem $\min_u \max_v -f_{\text{priv}}$ considered in this paper is not a saddle-point problem even for a relative simple case. Suppose one chooses linear filters, convex differentiable losses (e.g., least-squares, logistic, or exponential losses) and linear classifiers for the problem. Then

$$-f_{\text{priv}}(u, v) = -E[l(g(u); v)] = -E[l(yv^T g(U^T x))] \tag{65}$$

is the negative expected value of the composition of a convex $l(\cdot)$ and a linear $U^T x$, which is concave in u as well as concave in v, which cannot a saddle-point problem.

4.3 General problem

A general numerical solution to the optimization (44) is described in this section. Let $f(u, v)$ be a real-valued function $f : \mathbb{R}^d \times \mathcal{V} \to \mathbb{R}$, where $\mathcal{V}$ is a compact subset of the Euclidean space. Suppose f is jointly continuous and has a continuous partial derivative $\nabla_u f$ w.r.t. the first variable u. The maximum over v

$$\Phi(u) = \max_{v \in \mathcal{V}} f(u, v) \tag{66}$$

has a property that $\Phi(u)$ is in general not differentiable in u even if $f(u, v)$ is. The derivatives of $\Phi(u)$ was studied by Danskin (1967). Suppose $V(u)$ is the set of maximizers of f given u:

$$V(u) = \{\hat{v} \in \mathcal{V} \mid f(u, \hat{v}) = \max_{v \in \mathcal{V}} f(u, v)\}. \tag{67}$$

Danskin proved that the directional derivative $D_y \Phi(u)$ in any direction $y \in \mathbb{R}^d$ can be written as the maximum directional derivatives of $f(u, v)$ over all $\hat{v} \in V(u)$:

$$D_y \Phi(u) = \max_{\hat{v} \in V(u)} D_y f(u, \hat{v}), \tag{68}$$

where $D_y f(u, v)$ is the directional derivative of f w.r.t. u. Furthermore, in the case where $V(u)$ is a singleton $\{\hat{v}(u)\}$ for each u, we have

$$D_y \Phi(u) = D_y f(u, \hat{v}(u)). \tag{69}$$

There are several classic minimax optimization algorithms using this property. Suppose $f(u, v)$ is also continuously differentiable w.r.t. v, and $\nabla_u f$ is continuously differentiable w.r.t. v. A first-order method for minimax problems was proposed by Panin (1981) and was later refined by Kiwiel (1987). The latter uses a linear approximation of f at a fixed $\bar{u}$ along the direction q

$$f^l(q, v) = f(\bar{u}, v) + \langle \nabla_u f(\bar{u}, v),\ q \rangle, \tag{70}$$

and uses it to compute the approximate max value

$$\Phi^l(q) = \max_v f^l(q, v). \tag{71}$$

Using this approximation, a line search can be performed along the descent direction q that minimizes the max function $\Phi(\bar{u} + \alpha q)$. In particular, with additional assumptions of Lipschitz continuity of $\nabla_u f$ and compactness of $\mathcal{U}$ and $\mathcal{V}$, Kiwiel's algorithm monotonically

decreases f for each iteration and converges to a stationary point u^*, i.e., a point u for which $\max_v \langle \nabla_u f(u^*, v),\ q\rangle \geq 0$ for all directions q. Previously, Hamm (2015) used Kiwiel's algorithm to solve the optimization problems (44). However, one disadvantage of the method was that its slow speed in practice, due to the auxiliary routine of finding the descent direction q at each iteration (described in the supplementary material of Hamm (2015).)

Instead, this paper proposes a simple alternating algorithm (Alg. 1) for solving min-diff-max problem based directly on Danskin's theorem. The algorithm only assumes $f_{\text{priv}}(u, v)$ and $f_{\text{util}}(u, w)$ to be jointly continuous and have continuous partial derivatives $\nabla_u f_{\text{priv}}$ and $\nabla_u f_{\text{util}}$. Additionally, if $f_{\text{priv}}(u, v)$ and $f_{\text{util}}(u, w)$ are convex in v and w respectively, then the global minima

$$v_t = \arg\min_v f_{\text{priv}}(u_t, v) \ \text{ and } \ w_t = \arg\min_w f_{\text{util}}(u_t, w) \tag{72}$$

can be found easily, either approximately or accurately. Furthermore, if f_{util} and f_{priv} are strongly convex (e.g., due to regularization), the solutions are unique. Consequently, the descent direction q_t in Alg. 1 is truly the (negative) gradient of $\Phi(u)$ (44) as desired:

$$q_t = \nabla_u f_{\text{priv}}(u, v_t) - \rho \nabla_u f_{\text{util}}(u, w_t) = -\nabla_u \Phi_{\text{priv}}(u) + \rho\ \nabla_u \Phi_{\text{util}}(u) = -\nabla_u \Phi(u). \tag{73}$$

Note that it is still a heuristic for non-convex f_{util} and f_{priv} such as when using neural networks for the filter and/or the classifiers. A related heuristic for minimax problems was proposed by Goodfellow et al. (2014) for solving an unrelated problem of learning generative models.

Algorithm 1 Alternating algorithm for min-diff-max

Input: data $\{(x_i, y_i, z_i)\}$, filter g, loss l, classifier h, coefficient ρ, max iteration T
Output: optimal filter parameter u
Begin:
  Initialize u_1
  for $t = 1, ..., T$ **do**
    Solve (approximately)

$$v_t = \arg\min_v f_{\text{priv}}(u_t, v) \ \text{ and } \ w_t = \arg\min_w f_{\text{util}}(u_t, w), \ \text{ where} \tag{74}$$

$$f_{\text{priv}}(u, v) = \frac{1}{N}\sum_{i=1}^{N} l_p(h_p(g(x_i; u); v), y_i) \ \text{ and } \ f_{\text{util}}(u, w) = \frac{1}{N}\sum_{i=1}^{N} l_u(h_u(g(x_i; u); w), z_i). \tag{75}$$

    Compute the descent direction by

$$q_t = \nabla_u f_{\text{priv}}(u, v_t) - \rho \nabla_u f_{\text{util}}(u, w_t) \tag{76}$$

    Perform line search along q_t and update $u_{t+1} = u_t + \alpha_t \cdot q_t$
    Exit if solution converged
  end for

The proposed optimization algorithm and supporting classes are implemented in Python and are made available on an open-source repository[3].

5. Noisy Minimax Filter

The privacy guarantee that minimax filters provide is very different from that of differentially-private mechanisms. As the filter is learned from data, its privacy guarantee is given only in expectation/probability. Besides, it is a deterministic mechanism which cannot provide differential privacy. This section presents *noisy minimax filter* that combines minimax filter with additive noise mechanism to satisfy differential privacy criteria. Two methods of combination–preprocessing and postprocessing–are proposed and compared. For completeness, a brief introduction to differential privacy is given.

5.1 Differential privacy

A randomized algorithm that takes data $\mathcal{D}$ as input and outputs $\tilde{f}(\mathcal{D})$ is called ϵ-differentially private if

$$Pr(\tilde{f}(\mathcal{D}) \in \mathcal{S}) \leq e^{\epsilon} Pr(\tilde{f}(\mathcal{D}') \in \mathcal{S}) \tag{77}$$

for all measurable $\mathcal{S} \subset \mathcal{T}$ of the output range and for all data sets $\mathcal{D}$ and $\mathcal{D}'$ differing in a single item, denoted by $\mathcal{D} \sim \mathcal{D}'$. That is, even if an adversary knows the whole data set $\mathcal{D}$ except for a single item, she cannot infer much more about the unknown item from the output of the algorithm. A well-known mechanism for turning a non-private function f into a private function $\tilde{f}$ is the perturbation by additive noise. When an algorithm outputs a real-valued vector $f(\mathcal{D}) \in \mathbb{R}^D$, its global sensitivity (Dwork et al., 2006) is defined as

$$S(f) = \max_{\mathcal{D} \sim \mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\| \tag{78}$$

where $\|\cdot\|$ is a norm such as the Euclidean norm. An important result from Dwork et al. (2006) is that the perturbation by additive noise

$$\tilde{f}(\mathcal{D}) = f(\mathcal{D}) + \xi, \tag{79}$$

where ξ has the Laplace-like probability density whose scale parameter is proportional to $S(f)$

$$P(\xi) \propto e^{-\frac{\epsilon}{S(f)}\|\xi\|} \tag{80}$$

is ϵ-differentially private.

This paper considers local differential privacy (Duchi et al., 2013) of filtered output $g(x)$, that is, perturbation is applied by each participant before $g(x)$ is released to a third party. In addition, per-sample privacy is of interest, in which the neighboring data sets are $\mathcal{D} = \{x\}$ and $\mathcal{D}' = \{x'\}$, where x and x' are any two samples from the common feature space ($x, x' \subset \mathcal{X}$) of all participants. Consequently, a randomized filter $\tilde{g}(\cdot)$ is ϵ-differentially private if for all $x, x' \in \mathcal{X}$ and all measurable $\mathcal{S} \subset \mathcal{T}$ of the output range,

$$Pr(\tilde{g}(x) \in S) \leq e^{\epsilon} Pr(\tilde{g}(x') \in S). \tag{81}$$

3. `https://github.com/jihunhamm/MinimaxFilter`

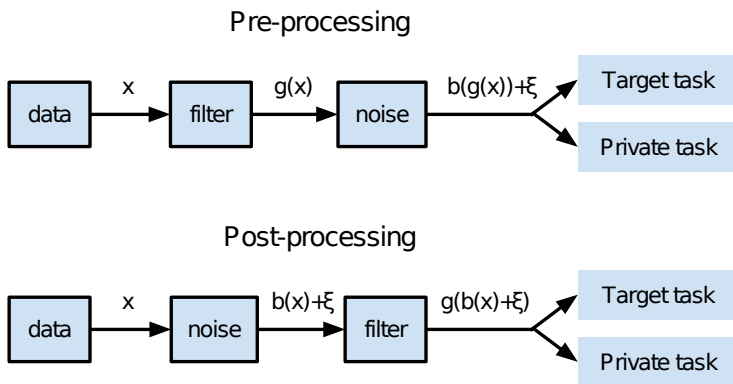

Figure 2: Pre vs postprocessing approach to differentially private minimax filtering.

To use additivity noise mechanism (79), the sensitivity (78) of the non-random $g(\cdot)$ needs to be determined:

$$S(g) = \sup_{x,x' \in \mathcal{X}} \|g(x) - g(x')\| \tag{82}$$

which is finite if $\mathcal{X}$ is compact and $g(\cdot)$ is continuous. Care must be taken when computing the sensitivity $S(g)$, since the filter $g(\cdot)$ is learned from training data and is dependent on them. This breaks the differential privacy guarantee if one simply uses a trained $g(\cdot)$ to compute the sensitivity. To avoid this problem without making more assumptions on $\mathcal{X}$ or $g(\cdot)$, the paper proposes to directly bound the diameter of $g(\mathcal{X})$ by *bounding functions*. Let $h = g(x)$ be the filtered value of any sample $x \in \mathcal{X}$. The bounding functions $b : \mathbb{R}^D \to \mathbb{R}^D$ are one of

1. Hard-bound by clipping: $b(h) = \min\{1,\ 1/\|h\|\}h$,
2. Soft-bound by squashing: $b(h) = \tanh(a\|h\|)$, or
3. Normalization: $b(h) = h/\|h\|$.

Note that these functions enforce the sensitivity $S(b(g))$ to be at most 2, regardless of $\mathcal{X}$ or $g(\cdot)$.

5.2 Preprocessing vs postprocessing

Minimax filters can be made locally differentially private by additive noise mechanism (79) in the signal chain of filtering. The paper proposes two approaches: filtering is performed before perturbation (called *preprocessing*) and after perturbation (called *postprocessing*)[4]. Fig. 2 show the signal chains of the two approaches. In preprocessing, the original feature x is first filtered by $g(x)$, and then made ϵ-differential private by bounding and perturbing $b(g(x)) + \xi$. In postprocessing, the original feature x is first made ϵ-differentially private by

4. Preprocessing and postprocessing are similar to but not exactly the same as output perturbation and input perturbation (Sarwate and Chaudhuri, 2013)).

bounding and perturbing $b(x) + \xi$ followed by filtering $g(b(x) + \xi)$. By adding an appropriate amount of noise, both approaches are locally differentially private regardless of data distribution. However, when the noisy mechanism is used in conjunction with a minimax filter which is dependent on data distribution $P(x, y, z)$, preprocessing and postprocessing approaches have different effects depending on the distribution.

A scenario when preprocessing is preferable to postprocessing is as follows. For convenience of explanation, let's assume the private task is subject identification. Let $y(x)$ be the subject identity label of sample x and let $z(x)$ be the target label of sample x. Define *between-subject diameter* as the max distance of two samples x, x' from different subjects that have the same target label:

$$S_b \triangleq \max_{x,x' \in \mathcal{X}} \|x - x'\| \quad \text{s.t.} \quad y(x) \neq y(x'),\ z(x) = z(x'). \tag{83}$$

Similarly, define *within-subject diameter* as the max distance of two samples x, x' from the same subject that have different target labels:

$$S_w \triangleq \max_{x,x' \mathcal{X}} \|x - x'\| \quad \text{s.t.} \quad y(x) = y(x'),\ z(x) \neq z(x'). \tag{84}$$

For a given data set $\mathcal{X}$, if the between-subject diameter is larger than the within-subject diameter ($S_b > S_w$) in the original feature space (Fig. 3a), then the sensitivity after minimax filtering $S(g) = \max_{x,x'} \|g(x) - g(x')\|$ can potentially be reduced a lot. This translates to less amount of noise to achieve the same ϵ-privacy than the amount of noise required before filtering, as the data diameter has shrunk. This results in better utility of the preprocessing approach over the postprocessing approach where noise added before filtering. From the same reasoning, if the opposite is true ($S_w > S_b$) (Fig. 3b), then the sensitivity after minimax filtering $S(g)$ will not change much, and the preprocessing approach will not offer much advantage over the postprocessing. However, there are other aspects to consider as well. In the preprocessing approach, training of minimax filter by solving (6) is not itself a differentially private procedure and requires trusted collector/curator. In addition, the learned filters contain information about private data analogous to learned PCA components revealing information about data (Chaudhuri et al., 2012), and require further sanitization if the filters are released to a third party. In contrast, the postprocessing approach makes the whole process much simpler. Any postprocessing–including the training of minimax filter–does not worsen differential privacy guarantees (Dwork et al., 2014), and there is no need for trusted collector/curator once original data are perturbed early in the signal chain. In the experiment, utility and privacy of two approaches are further compared using real data.

6. Experiments

In this section, algorithms proposed in the paper are evaluated using three real-world data sets: face data for gender/expression classification speech data for emotion classification, and motion data for activity classification. Firstly, minimax filters are compared with non-minimax methods in terms of privacy breach vs utility as measured by accuracy of private and target tasks classifiers on test data. Secondly, noisy minimax filters are tested under various conditions using the same data sets.

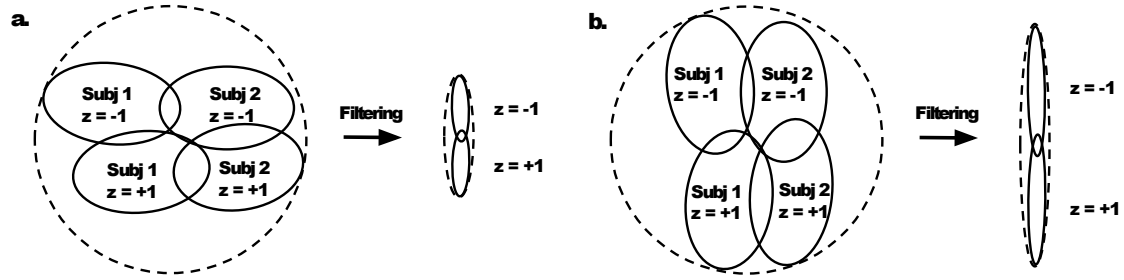

Figure 3: Two example data distributions with the same data diameter before filtering but different diameters after filtering. a. When between-subject diameter is large. b. When within-subject diameter is large. (See text for definitions.)

6.1 Methods

Filters. The following minimax and non-minimax filters are compared:

- Rand: random subspace projection with $g(x;U) = U^T x$, where U is a random full rank $D \times d$ matrix.
- PCA: principal component analysis with $g(x;U) = U^T x$, where U is the eigenvectors corresponding to d largest eigenvalues of $\mathrm{Cov}(x)$.
- PPLS: private partial least squares, using Algorithm 1 from Enev et al. (2012).
- DDD: discriminately decreasing discriminability (DDD) from Whitehill and Movellan (2012) with a mask-type filter from the code[5].
- Minimax 1: linear filter $g(x;U) = U^T x$ where U is computed from Alg. (1).
- Minimax 2: nonlinear filter $g(x)$ from a two-layer sigmoid neural network with of hidden nodes of 20 and 10, computed from Alg. (1).

Remarks. DDD requires analytical solutions to eigenvalue problems which are unavailable for multiclass problems, and is used only in the binary problem with the face database. Also, DDD code uses a mask-type filter, and the dimension d is same as the image size. The dimension d is also irrelevant to nonlinear Minimax filter 2 since it does not use linear dimensionality reduction. The nonlinear filter is pretrained as a stacked denoising autoencoders (Vincent et al., 2008) followed by supervised backpropagation with the target task.
Classifier/loss. For all experiments, binary or multinomial logistic regression is used a classifier for both utility and privacy risks, where the loss $l(h(g(x;u);v),y)$ is the negative log-likelihood:

$$l = -v(y)^T g(x;u) + \log(\sum_{k=1}^{K} e^{v(k)^T g(x;u)}) + \frac{\lambda}{2}\sum_{k=1}^{K} \|v(k)\|^2 \tag{85}$$

where K is the number of classes. The regularization coefficient was $\lambda = 10^{-6}$) and the utility-privacy tradeoff coefficient was $\rho = 10$. The main iteration in Alg. 1 was stopped manually when the progress was slow, which was between $T = 20 - 200$.

5. `http://mplab.ucsd.edu/~jake`

6.2 Data sets

Gender/expression classification from face: The GENKI database (Whitehill and Movellan, 2012) consists of face images with varying poses and facial expressions. The original data set is used unchanged, which have $N = 1740$ training images (50% male and 50% female; 50% smile and 50% non-smile). The test set has 100 images (50 males and 50 females; 50 smiling and 50 non-smiling) not overlapping with the training set. The dimensionality of the original data is $D = 256$, and the filters are tested with $d = 10, 20, 50, 100$. The data set has gender and expression labels but no subject label. Consequently, gender classification is used as the private task and expression classification is used as the target task.

Emotion classification from speech: The ENTERFACE database (Martin et al., 2006) is an audiovisual emotion database of 43 speakers from 14 nations reading predefined English sentences in six induced emotions. From the raw speech signals sampled in 48 KHz, MFCC coefficients are computed using 20 ms windows with 50% overlap and 13 Mel-frequency bands. The mean, max, min, and standard deviation of the MFCC coefficients over the duration of each sentence are computed, resulting in $N = 427$ samples of $D = 52$ dimensional feature vectors from $S = 43$ subjects. Each subject's samples are randomly split to generate training (80%) and test (20%) sets. Average test accuracy over 10 such trials is reported. Filters are tested with $d = 10, 20, 30, 40$. The target task is binary classification of 'happy' and 'non-happy' emotions from speech, and the privacy risks is multiclass ($S = 43$) subject classification.

Activity classification from motion: The UCI Human Activity Recognition (HAR) data set (Anguita et al., 2012) is a collection of motion sensor data on a smartphone by 30 subjects performing six activities (*walking*, *walking upstairs*, *walking downstairs*, *sitting*, *standing*, *laying*). Various time and frequency domain variables are extracted from the signal, resulting in $N = 10299$ samples of $D = 561$ dimensional features from 30 subjects which are used unchanged. Out of 30 subjects, 15 subjects are chosen randomly. For each domain, each subject's samples are randomly split to generate training (50%) and test (50%) sets. At each trial, the subjects and the training/test sets are randomized, and the average test accuracy over 10 such trials is reported. Filters with dimensions $d = 10, 20, 50, 100$ are used. The target task is multiclass ($C = 6$) classification of activity, and the privacy risks is multiclass ($S = 15$) subject classification risk.

6.3 Result 1: Minimax filters

Fig. 4 shows the test accuracy with GENKI. The dotted lines are level sets of utility-privacy tradeoff (i.e., target task accuracy - private task accuracy) shown for reference. Minimax 2 achieves the best utility (i.e., most accurate expression classification) and Minimax 1 (linear) achieves the best privacy (i.e., least accurate gender classification). For all dimensions d, Minimax 1 achieves the best utility-privacy compromise (i.e., closest to the top-left corner of the plot), with Minimax 2 and DDD performing very closely. In terms of privacy risk, Minimax 1 achieves almost the chance level accuracy (0.5), which implies a strong privacy preservation. DDD comes close to Minimax 1, while another private method PPLS is not very successful in reducing the privacy risk. As expected, non-private methods Rand and PCA do not reduce the privacy risk. As dimension d increases from 10 to 100, the accuracy

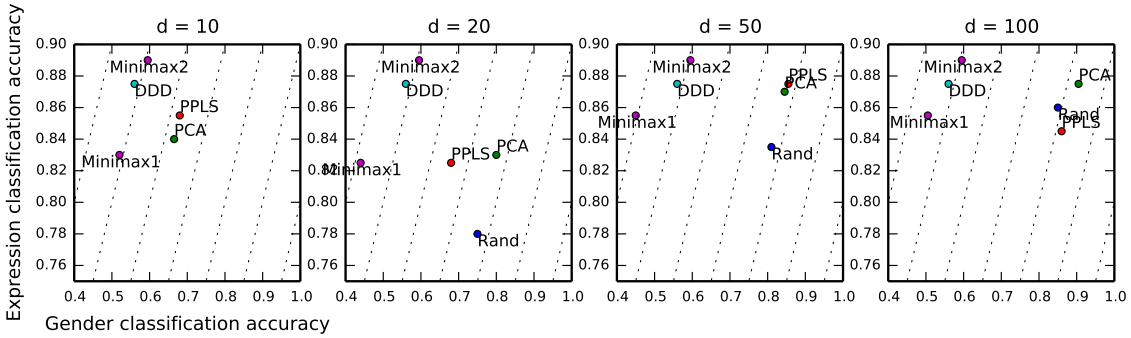

Figure 4: GENKI: Expression classification vs and gender classification from faces.

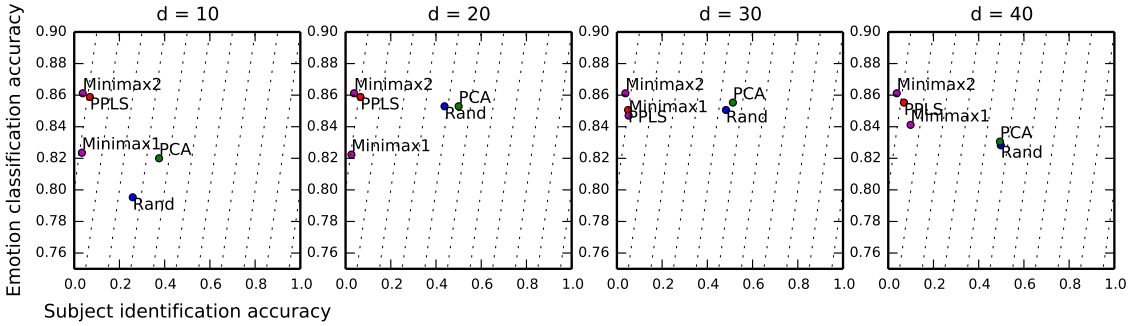

Figure 5: ENTERFACE: Emotion classification vs and subject identification from speech.

of both the target and the private tasks increase (toward the top-right corner of the plot) for PPLS, PCA and Rand, but the value of utility-privacy tradeoff (i.e., target task accuracy - private task accuracy) remains relatively similar even though d changes. Note that d is irrelevant to Minimax 2 and DDD.

Fig. 5 shows the test accuracy of ENTERFACE. Minimax 2 achieves the best utility (i.e., most accurate emotion classification) and the best privacy (i.e., least accurate subject classification) at the same time. PPLS performs well in this task; its private and target task accuracy is close to those of Minimax 2. The private task accuracy of Minimax 2 is near the chance level ($1/S = 0.02$) compared to $0.4 - 0.5$ of non-private methods, suggesting that seemingly harmless statistics (mean, max, min, s.d. of MFCC) are quite susceptible to identification attacks if no privacy mechanism is used. Similar to GENKI, the accuracy of both the target and the private tasks increases with the dimension d for PCA and Rand, and the value of utility-privacy tradeoff remains similar regardless of d.

Fig. 6 shows the test accuracy of HAR. Minimax 1 achieves the best utility (i.e., most accurate activity recognition) and the best privacy (i.e., least accurate subject classification), while Minimax 2 and PPLS performs similarly well. The private task accuracy of Minimax

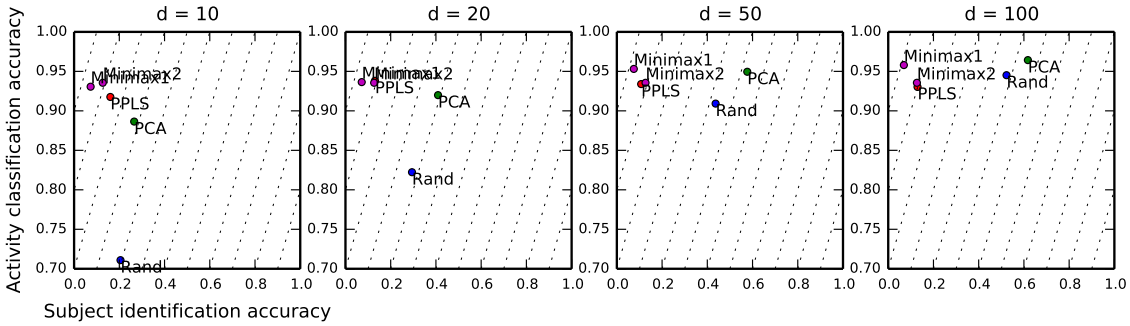

Figure 6: HAR: Activity classification vs subject identification from motion.

1 is lower than others close to the chance level ($1/S = 0.067$). The figure also shows that motion data are susceptible ($0.2 - 0.7$) to identification attacks when no privacy mechanism is used. For all dimensions d, Minimax 1 achieves the best compromise of all methods similar to previous experiments. Also the accuracy of both the target and the private tasks roughly increases with d for PCA and Rand, but the value of utility-privacy tradeoff remains similar.

6.4 Result 2: Noisy minimax filters

The same data sets from the previous section are used to demonstrate the effect of noisy mechanism on minimax filters. Four types of noisy filters are compared: PCA-pre, PCA-post, Minimax-pre, and Minimax-post. PCA is chosen as a non-minimax reference filter which preserves the original signal in the least mean-squared-error sense. PCA-pre/post means that PCA is applied before/after the perturbation similarly to Minimax-pre/post from Fig. 2. For Minimax-pre/post, a linear filter of the same dimension d as PCA-pre/post is used. Tests are performed for $d = 20$. Optimization of 6 is done similarly to the previous section. All tests are repeated 10 times for different noise samples of (80), for each of 10 random training/test splits.

Fig. 7 shows the following results. Firstly, within each plot, increasing the privacy level from left (ϵ^{-1}=0) to right (ϵ^{-1}=10) lowers the accuracy of both target and private tasks for all filter types and data sets, which is intuitively correct. Secondly, target task accuracy (top row) shows that the four filters are equally accurate with no noise (ϵ^{-1}=0), with Minimax-pre/post slightly more accurate than PCA-pre/post. This observation is consistent with the results in Sec. 6.3. In GENKI and HAR, preprocessing is better than postprocessing for both PCA and Minimax, and Minimax-pre performs the best. In ENTERFACE, preprocessing and postprocessing approaches perform similarly, and all four filters is perform similarly on the target task. This result may be ascribed to the discussion of different data distribution in Sec. 5.2. Thirdly, and most importantly, private task accuracy (bottom row) is quite different between Minimax-pre/post and non-minimax PCA-pre/post. For both Minimax-pre and Minimax-post, the private task accuracy is almost as low as the chance accuracy of each data set (0.5, 0.03, 0.07) regardless of the noise level ϵ. This demonstrates that minimax

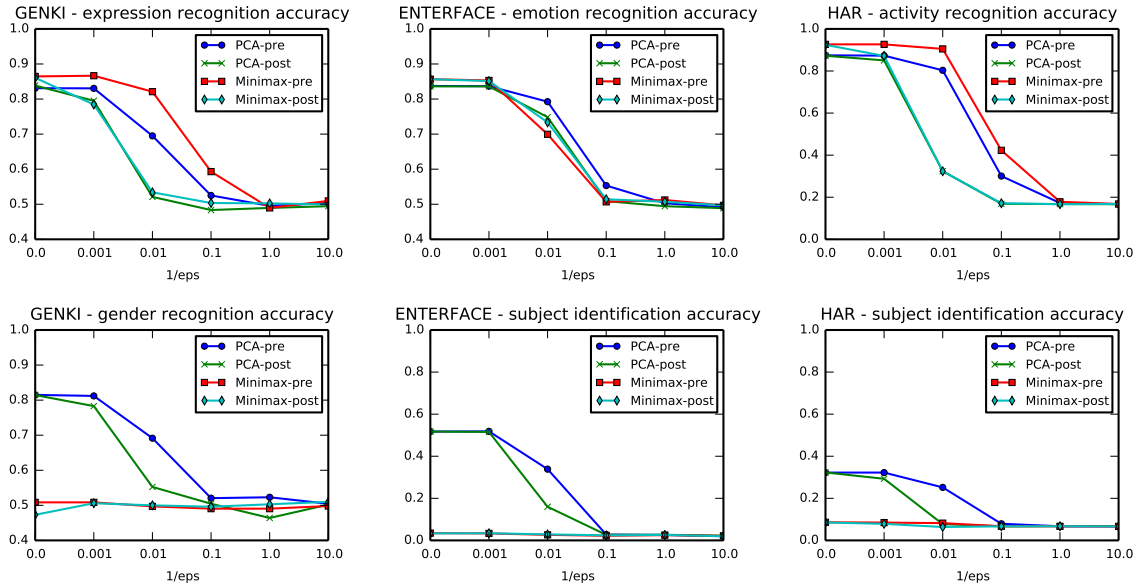

Figure 7: Impact of four noisy filters (PCA-pre/post and Minimax-pre/post) on the accuracy of target and private tasks for three data sets (GENKI, ENTERFACE, HAR), over the range of $\epsilon^{-1} = \{0, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$. Top row is the target task accuracy (higher the better) and bottom row is the private task accuracy (lower the better.) Minimax-pre/post can limit the accuracy of inference attack (bottom row) to almost chance levels regardless of the value of ϵ, while PCA-pre/post requires a significantly high ϵ to prevent inference attacks which also destroy the utility.

filter can prevent inference attacks with little help of noise. In contrast, the non-minimax filters (PCA-pre/post) allow an adversary to infer private variables quite accurately (0.8, 0.5, 0.3) when no noise is used. Preventing such attacks for non-minimax filters requires a significant amount of additive noise (e.g., $\epsilon^{-1} \geq 0.1$) which destroys the utility of data. These results show that differentially privacy is indeed different from privacy against inference attacks and the combination of two methods is beneficial.

7. Conclusion

This work presents a new privacy-preserving mechanism for preventing inference attacks on continuous and high-dimensional data. In this mechanism, a filter transforms continuous and high-dimensional raw features to dimensionality-reduced representation of data. After filtering, information on target tasks remains but information on identifying/sensitive attributes is removed which makes it difficult for an adversary to accurately infer such attributes from the released filtered output. Minimax filters are designed to achieve the optimal utility-privacy tradeoff in terms of expected risks. The paper proves that a filter

learned from empirical risks is not far from an ideal filter as the number of samples increases. This property and its dependency on the task make this mechanism quite different from previous mechanisms, including syntactic anonymization and differential privacy. Algorithms for finding minimax filters are presented and evaluated on real-world data sets to show its practical usages. Experiments show that publicly available multisubject data sets are surprisingly susceptible to subject identification attacks, and that even simple linear minimax filters can reduce the privacy risks close to chance level without sacrificing target task accuracy by much.

This work also presents preprocessing and postprocessing approaches to combine minimax privacy and differential privacy. While differential privacy has become a popular definition of privacy, it is not without limitations, in particular against inference attacks as empirically demonstrated in the paper. This leaves room for development of new mechanisms such as the noisy minimax filter presented in the paper, which aims to achieve high on-average utility and protection against inference attacks, and a formal privacy guarantee in the worst case. The results from experiments encourage further research on potential benefits of combining different notions and mechanisms of privacy, which is left as future work.

Acknowledgments

References

Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *Formal Aspects of Security and Trust*, pages 39–54. Springer, 2012.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *Ambient assisted living and home care*, pages 216–223. Springer, 2012.

Justin Brickell and Vitaly Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 70–78. ACM, 2008.

Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In *Advances in Neural Information Processing Systems*, pages 989–997, 2012.

Chris Clifton and Tamir Tassa. On syntactic anonymity and differential privacy. *Transactions on Data Privacy*, 6(2):161–183, 2013.

Graham Cormode. Personal privacy vs population privacy: learning to attack anonymization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1253–1261. ACM, 2011.

John M Danskin. *The theory of max-min and its application to weapons allocation problems*. Springer, 1967.

Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1401–1408. IEEE, 2012.

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.

Cynthia Dwork. Differential privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

Cynthia Dwork and Kobbi Nissim. Privacy-Preserving Data Mining on Vertically Partitioned Databases. In *Proc. CRYPTO*. Springer, 2004.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Miro Enev, Jaeyeon Jung, Liefeng Bo, Xiaofeng Ren, and Tadayoshi Kohno. Sensorsift: balancing sensor data privacy and utility in automated face understanding. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 149–158. ACM, 2012.

Benjamin Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comp. Surveys (CSUR)*, 42(4):14, 2010.

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 351–360. ACM, 2009.

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.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *Journal of Machine Learning Research*, 14(Feb):703–727, 2013.

Jihun Hamm. Preserving privacy of continuous high-dimensional data with minimax filters. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Jihun Hamm. Enhancing utility and privacy with noisy minimax filters. In *Under review*, 2016.

Vijay S Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288. ACM, 2002.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Daniel Kifer. Attacks on privacy and definetti's theorem. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 127–138. ACM, 2009.

Krzysztof C Kiwiel. A direct method of linearization for continuous minimax problems. *Journal of optimization theory and applications*, 55(2):271–287, 1987.

Andreas Krause and Eric Horvitz. A utility-theoretic approach to privacy and personalization. In *AAAI*, volume 8, pages 1181–1188, 2008.

Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526. ACM, 2009.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface´05 audio-visual emotion database. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 8–8. IEEE, 2006.

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.

Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *Knowledge and Data Engineering, IEEE Transactions on*, 17 (2):232–243, 2005.

VM Panin. Linearization method for continuous min-max problem. *Cybernetics and Systems Analysis*, 17(2):239–243, 1981.

David Rebollo-Monedero, Jordi Forne, and Josep Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *Knowledge and Data Engineering, IEEE Transactions on*, 22(11):1623–1636, 2010.

Berc Rustem and Melendres Howe. *Algorithms for worst-case design and applications to risk management*. Princeton University Press, 2009.

Lalitha Sankar, Raj Rajagopalan, and Vincent Poor. An information-theoretic approach to privacy. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1220–1227. IEEE, 2010.

Anand Sarwate and Kamalika Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *Signal Processing Magazine, IEEE*, 30(5):86–94, 2013.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 1107057132, 9781107057135.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

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

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of ICML*, pages 1096–1103. ACM, 2008.

Jacob Whitehill and Javier Movellan. Discriminately decreasing discriminability with learned image filters. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2488–2495. IEEE, 2012.