Task-aware Privacy Preservation for Multi-dimensional Data

Jiangnan Cheng [1] **Ao Tang** [1] **Sandeep Chinchali** [2]

Abstract

Local differential privacy (LDP) can be adopted to anonymize richer user data attributes that will be input to sophisticated machine learning (ML) tasks. However, today's LDP approaches are largely *task-agnostic* and often lead to severe performance loss – they simply inject noise to all data attributes according to a given privacy budget, regardless of what features are most relevant for the ultimate task. In this paper, we address how to significantly improve the ultimate task performance with multi-dimensional user data by considering a task-aware privacy preservation problem. The key idea is to use an encoder-decoder framework to learn (and anonymize) a *task-relevant* latent representation of user data. We obtain an analytical near-optimal solution for the linear setting with mean-squared error (MSE) task loss. We also provide an approximate solution through a gradient-based learning algorithm for general nonlinear cases. Extensive experiments demonstrate that our task-aware approach significantly improves ultimate task accuracy compared to standard benchmark LDP approaches with the same level of privacy guarantee.

1. Introduction

In recent years, there has been a tremendous growth in the volume of available data for machine learning (ML) tasks, leading to increasing emphasis on protecting user privacy. Differential privacy (DP) (Dwork et al., 2006; 2014) is a state-of-the-art technique for data privacy, and its local variant – local differential privacy (LDP) (Kasiviswanathan et al., 2011) – provides stronger privacy guarantees for individual users without dependence on any trusted third party. In practice, LDP has been successfully deployed in the products of companies like Google (Erlingsson et al., 2014), Apple (Differential Privacy Team), and Microsoft (Ding et al., 2017) for some basic frequency or histogram estimation tasks where raw user data is restricted to an n-bit discrete variable.

In the future, LDP has the promising potential to be adopted in more complex scenarios (Hassan et al., 2019; Dankar & El Emam, 2013; Zhao et al., 2014; Cortés et al., 2016) (e.g., health care, power grids, Internet of Things) that feature richer user data attributes that feed into more sophisticated downstream ML tasks (Cheng et al., 2021; Nakanoya et al., 2021). In such cases, today's standard task-agnostic LDP approaches may not be ideal. For example, consider complex user data that must be anonymized before being passed into a ML task function, such as a neural network classifier for credit scores. A standard approach would be to simply perturb the data by adding artificial noise whose scale depends on the sensitivity of user data (i.e. worst-case variation among a user population) and a given privacy budget, *regardless* of what ultimate task the anonymized data will be used for. However, as the dimension and variability of user data inevitably grows, today's methods would generally have to increase the scale of noise to provide the same LDP guarantee, even though many data attributes might be highly variable across a user population, but minimally *relevant* for a task. As a consequence, one often adds excessive noise to all data attributes, which can severely degrade an ultimate task's performance.

To address these challenges, this paper introduces a fundamentally different *task-aware* LDP approach. Our method improves the performance of ML tasks that operate on multi-dimensional user data while still guaranteeing the same levels of privacy. Our key technical insight is to characterize the dependence of task performance on various user data attributes, which guides how we *learn* a concise, task-relevant encoding (i.e. latent representation) of user data. Then, for the same privacy budget, we can directly expose and perturb only the task-relevant encoding rather than raw user data, which often allows us to add less noise and thereby improve task accuracy (see Section 2 for a concrete example). Crucially, user privacy is guaranteed under the same privacy budget according to the post-processing immunity

[1]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY [2]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX. Correspondence to: Jiangnan Cheng <jc3377@cornell.edu>, Ao Tang <atang@cornell.edu>, Sandeep Chinchali <sandeepc@utexas.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

of DP (Dwork et al., 2014) (i.e., one cannot make the output of a privacy algorithm less differentially private without additional knowledge). As such, an adversary cannot decode the anonymized latent representation to reduce the level of privacy. Our method allows us to learn and expose only high-valued data attributes and flexibly adjust their signal-to-noise ratio based on their importance to a task. Moreover, when different data attributes are inter-dependent, task-aware LDP preservation is even more promising in that we can consider the utilities of the underlying orthogonal bases through principal component analysis (PCA) (Duntemman, 1989), instead of the raw data attributes.

Contributions. In light of prior work, our contributions are three-fold. First, we propose a task-aware privacy preservation problem in which the effect of noise perturbation to preserve LDP is effectively considered, based on an encoder-decoder framework (Section 3). Second, in terms of task-aware privacy preservation, we obtain an analytical near-optimal solution for a linear setting and mean-squared error (MSE) task loss, and provide a gradient-based learning algorithm for more general settings (Section 4). Third, we validate the effectiveness of our task-aware approach through three real-world experiments, which show our task-aware approach outperforms the benchmark approaches on overall task loss under various LDP budgets by as much as 70.0% (Section 5). All the proofs are given in the Appendix.

Related Work. 1) Utility maximization in DP/LDP. Most theoretical DP/LDP research either provides a utility upper bound for a given privacy budget under some weak assumptions (Alvim et al., 2011; Kenthapadi et al., 2012; Duchi et al., 2013; Makhdoumi & Fawaz, 2013; Hardt & Talwar, 2010; Wang et al., 2019b; Acharya et al., 2020), or designs optimal privacy preservation mechanisms in some specific use cases (McSherry & Talwar, 2007; Wasserman & Zhou, 2010; Friedman & Schuster, 2010; Xiao et al., 2010; Thakurta & Smith, 2013; Kairouz et al., 2014; Geng & Viswanath, 2015; Liu et al., 2016; Yiwen et al., 2018; Joseph et al., 2019; Gondara & Wang, 2020; Wang et al., 2020). To the best of our knowledge, with respect to *multi-dimensional user data*, Murakami & Kawamoto (2019), Wang et al. (2019a), Chen et al. (2021), Li et al. (2015) and McMahan et al. (2022) are the only similar works to ours. Murakami & Kawamoto (2019) develops a utility-maximizing LDP framework under the assumption that some data attributes may not be privacy-sensitive, and hence the utility improvement is mainly achieved by providing privacy guarantees for only a subset of attributes. Wang et al. (2019a) mainly focuses on developing novel LDP mechanisms for multi-dimensional data with an objective of minimizing the worst-case noise variance, which naturally improves utility. Chen et al. (2021) preserves DP for images by adding noise to a latent representation learned through back-propagation of task loss, where the effect of noise perturbation is not considered. Li et al. (2015) and McMahan et al. (2022) consider linear transformation and preserve DP through matrix factorization. The key differences of our work are: i) we don't make additional assumptions on the sensitivity of user data, ii) our task-aware approach achieves a better task performance than standard LDP benchmarks by directly studying the *dependencies* between the task objective and different attributes of user data, iii) we effectively capture the effect of noise perturbation resulting from privacy requirements, and we also show matrix factorization is not optimal for linear transformation. 2) End-to-end (E2E) learning. There have been a wide variety of works training a cascade of deep neural networks (DNNs) through E2E learning, where a task-specific output is directly predicted from the raw inputs (Muller et al., 2006; Wang et al., 2012; Donti et al., 2017; Zhou & Tuzel, 2018; Amos et al., 2018). Our work also follows such a practice, but introduces an LDP guarantee while improving the task performance. 3) DP in deep learning. The popularity of deep learning also draws great attention to DP preservation therein (Song et al., 2013; Shokri & Shmatikov, 2015; Abadi et al., 2016; Phan et al., 2016; Papernot et al., 2016; Phan et al., 2017; McMahan et al., 2018; Wang et al., 2018; Phan et al., 2019; Arachchige et al., 2019; Liu et al., 2020; Mireshghallah et al., 2020; Bu et al., 2020). Our work is fundamentally different. Instead of preserving LDP *during the learning process*, we use learning as a tool to find the salient representation that improves the task performance under a given privacy budget. In other words, we don't perturb the gradient for back-propagation but perturb the representation to guarantee LDP. Furthermore, we don't specifically deal with privacy preservation during the *offline* training process, which requires some ground truth user data (e.g., from a small set of consenting volunteers). However, LDP of user data is guaranteed after a trained model is deployed *online*.

2. Background and Motivating Example

2.1. Background

ϵ-LDP (Kasiviswanathan et al., 2011). Let $x \in \mathbb{R}^n$ be an individual data sample, and $\mathcal{X}$ be the domain of x, which is assumed to be a compact subset of $\mathbb{R}^n$. A randomized algorithm $\mathcal{M} : \mathcal{X} \mapsto \mathbb{R}^Z$ is said to satisfy ϵ-LDP with *privacy budget* $\epsilon > 0$, if $\forall x, x' \in \mathcal{X}, \mathcal{S} \subseteq \text{im } \mathcal{M}$, we have

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{M}(x') \in \mathcal{S}]. \tag{1}$$

Essentially, when ϵ is small, one cannot readily differentiate whether the input of $\mathcal{M}$ is an individual user x or x' based on $\mathcal{M}$'s outcome.

Laplace Mechanism (Dwork et al., 2014). To release a sensitive function $g : \mathcal{X} \mapsto \mathbb{R}^Z$ under ϵ-LDP, $\forall \epsilon > 0$, the Laplace mechanism is a widely used mechanism which adds

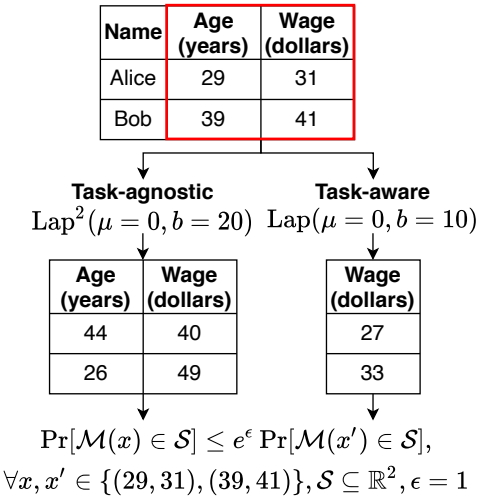

Figure 1. **Motivating example**. A task-aware approach (right) is more ideal than a task-agnostic approach (left) in terms of a mean wage estimation task, since the former perturbs the wage attribute with a smaller noise while guaranteeing the same LDP budget ϵ.

Laplace noise to function g:

$$\mathcal{M}_{\text{Lap}}(x, g, \epsilon) = g(x) + \text{Lap}^Z(\mu = 0, b = \frac{\Delta_1 g}{\epsilon}), \quad (2)$$

where $\text{Lap}^Z(\mu, b)$ is a Z-dimensional vector whose elements are i.i.d. Laplace random variables with mean μ and scale b, which leads its variance to be $2b^2$. And $\Delta_1 g = \max_{x,x' \in \mathcal{X}} \|g(x) - g(x')\|_1$ measures the sensitivity of g under the ℓ_1 norm.

For concreteness, the analysis and evaluation of this paper focus on the Laplace mechanism although the central idea of learning task-relevant data representations is applicable to other noise-adding mechanisms as well.

2.2. Motivating Example

Consider an example shown in Fig. 1. For simplicity, we only consider an example with two people, Alice and Bob, and two data attributes, age and wage. Suppose we need to preserve ϵ-LDP for each person with $\epsilon = 1$ and our task is to estimate the mean wage as accurate as possible. A straightforward task-agnostic approach will directly expose both the two data attributes, and add Laplace noise with scale $b = 20$ to each attribute. However, a task-aware approach will expose only the wage attribute and add Laplace noise with scale $b = 10$. Both the approaches guarantee LDP under the same budget ($\epsilon = 1$), but the wage attribute given by the task-aware approach is less noisy, and we can expect that the corresponding estimated mean wage (i.e. the ultimate task objective) is close to the real value with a higher probability.

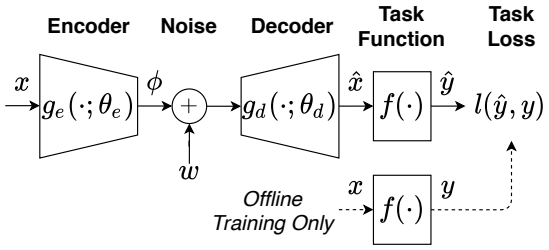

Figure 2. Overall architecture of the task-aware privacy preservation problem.

In more complex scenarios, such as when each data attribute is not redundant but is valued differently in terms of the considered task or data attributes are dependent but not perfectly correlated, etc., the optimal solution will not be as straightforward as the given example and will be explored in the following sections.

3. Problem Formulation

We now introduce the task-aware privacy preservation problem depicted in Fig. 2. Let $y = f(x) \in \mathbb{R}^m$ denote the task output associated with each ground truth data sample x, where f represents the task function. To guarantee ϵ-LDP for each data sample x, its true value should never be exposed to the task function. Instead, an estimate of x, denoted by $\hat{x}$, is used as the input to the task function with the corresponding task output $\hat{y} = f(\hat{x})$. The objective is to minimize the overall task loss $\mathcal{L} = \mathbb{E}[l(\hat{y}, y)]$ due to the difference between $\hat{x}$ and x, where x follows distribution $\mathcal{D}_x$, and l is a task loss function that captures the discrepancy between task output $\hat{y}$ and y, such as the common ℓ_2 and cross-entropy loss. Note that we don't specifically deal with privacy preservation during the *offline* training process, and the ground truth x is used to calculate y and $\mathcal{L}$. However, LDP of user data is guaranteed after a trained model is deployed *online*.

More concretely, x is first mapped to a latent representation $\phi \in \mathbb{R}^Z$ through an encoder function $\phi = g_e(x; \theta_e)$, where θ_e are a set of encoder parameters. ϕ is then perturbed by a Laplace noise vector $w \in \mathbb{R}^Z$. That is, g_e is treated as the sensitive function g in Eq. (2). Next, $\hat{x}$ is reconstructed from $\phi + w$ using a decoder function $\hat{x} = g_d(\phi + w; \theta_d)$ where θ_d are a set of decoder parameters. Moreover, in reality the encoder is deployed at the end of each individual user and in general has to be lightweight (e.g., linear or one hidden-layer neural network).

The optimal task-aware $\hat{x}$ minimizes $\mathcal{L}$ while preserving ϵ-LDP. In other words, the task-aware privacy preservation problem aims to co-design encoder and decoder, i.e., find proper values for Z, θ_e and θ_d, such that $\mathcal{L}$ is minimized

and ϵ-LDP is preserved. Formally, we have

$$\min_{Z,\theta_e,\theta_d} \quad \mathcal{L} = \mathbb{E}_{x,w}[l(\hat{y}, y)], \tag{3}$$

$$\text{s.t.} \quad y = f(x), \tag{4}$$

$$\hat{y} = f(g_d(g_e(x;\theta_e) + w;\theta_d)), \tag{5}$$

$$x \sim \mathcal{D}_x, w \sim \text{Lap}^Z(0, \frac{\Delta_1 g_e}{\epsilon}). \tag{6}$$

The difficulty of our task-aware LDP problem mainly comes from the discrepancy of the measurement of overall task loss $\mathcal{L}$, which depends on $\mathcal{D}_x$ and captures the *average* performance, and the mechanism of preserving LDP, which depends on X and focuses only on the *worst-case* privacy guarantee.

Benchmarks. We now describe two natural approaches to preserve ϵ-LDP. First, a **task-agnostic approach** adds noise directly to the normalized x[1]. For convenience we assume x is already normalized, and we have $Z = n$ and $g_e(x) = x$. Second, a **privacy-agnostic approach** adds noise to ϕ obtained by considering the problem defined in Eq. (3)-(6) with a pre-determined $Z \leq n$ and w being a zero vector instead. That is, the privacy-preservation part is neglected when designing the encoder, and hence a proper Z needs to be pre-determined or one would always conclude a larger Z (under which more information can be delivered when noise is absent) is better. Both two benchmark approaches still need to determine the optimal decoder parameters θ_d for input $\phi + w$. Note that for the task-agnostic approach even though g_e is an identity function, the corresponding optimal g_d is usually not an identity function, as exemplified in Section A.7.1.

4. Analysis

In this section, we solve the task-aware privacy preservation problem. Assuming a linear model and MSE task loss, we are able to find near-optimal analytical results, which shed clear insight on how to co-design an encoder and decoder. We then move to general settings and present a gradient-based learning algorithm which demonstrates strong empirical performance.

4.1. Linear Model with MSE Task Loss

In this subsection, we consider a linear model with MSE task loss. More specifically, encoder function g_e, decoder function g_d, and task function f are assumed to be linear functions in their corresponding inputs, and the loss function is $l = \|\hat{y} - y\|_2^2$. The task function f can then be expressed as $f(x) = Kx$, where $K \in \mathbb{R}^{m\times n}$ is the task matrix.

[1]We may either normalize each dimension independently when x is a multi-variate random variable, or normalize all the dimensions jointly when x is a uni-variate time-series.

Practicality of the Setting. Linear transformation is a common encoding and decoding approach for many dimensionality-reduction techniques, such as PCA. And ℓ_2 task loss is widely used in many application scenarios. For example, given N samples of x, suppose we want to estimate the mean value of these samples in a few directions, given by task matrix K. Then the sum of the variance of these estimates by using $\hat{x}$ instead of x will be $\frac{1}{N}\mathbb{E}_{x,w}[\|K(\hat{x} - x)\|_2^2]$, so $\mathcal{L} = \mathbb{E}_{x,w}[\|K(\hat{x} - x)\|_2^2] = \mathbb{E}_{x,w}[\|\hat{y} - y\|_2^2]$ is a natural objective function.

Contents. We first summarize the key results. For our task-aware approach, the optimal decoder is first determined in Proposition 4.1, and then the optimal encoder and the corresponding optimal loss is formulated in Proposition 4.4-4.6 under Assumption 4.3. We then relax Assumption 4.3 and provide lower and upper bounds for the optimal loss in Theorem 4.7 and an approximate solution. All the proofs are given in the Appendix.

Detailed Analysis. We start our analysis with a few definitions. First, without loss of generality, we assume the covariance matrix of $x - \mu_x$, i.e., $\mathbb{E}[(x - \mu_x)(x - \mu_x)^\top]$, is positive definite, where $\mu_x \in \mathbb{R}^n$ is the mean vector of x. This assumption guarantees x cannot be linearly transformed to a low-dimensional representation without losing information.

We then factorize $\mathbb{E}[(x - \mu_x)(x - \mu_x)^\top]$ into $LL^\top$ through Cholesky decomposition, where $L \in \mathbb{R}^{n\times n}$ is a lower triangular matrix with positive diagonal entries. For analytical convenience, we let $h = L^{-1}(x - \mu_x)$, which can be viewed as another representation of x, with mean $\mu_h = \mathbf{0}$ and covariance matrix $\Sigma_{hh} = I$. Let $\mathcal{D}_h$ denote the distribution of h, and $\mathcal{H} = \{L^{-1}(x - \mu_x) | x \in \mathcal{X}\}$ denote the compact set that contains all the possible values of $h \sim \mathcal{D}_h$. Since $K(\hat{x} - x) = P(\hat{h} - h)$, where $P = KL$, working with data representation h with task matrix P is equivalent to using x and K. Considering zero-centered h instead of original x saves us from considering the constant terms in the linear encoder and decoder functions[2].

Let $E \in \mathbb{R}^{Z\times n}$ and $D \in \mathbb{R}^{n\times Z}$ denote the encoder and decoder matrix associated with h, i.e., $\phi = Eh$ and $\hat{h} = D(Eh + w)$. Without loss of generality, we let $Z \geq n$ and allow some rows of E to be zero. Equivalently, based on the relationship between x and h, we have $\phi = EL^{-1}(x - \mu_x)$ and $\hat{x} - \mu_x = LD(EL^{-1}(x - \mu_x) + w)$. We denote the covariance matrix of w by Σ_{ww}, and it can be expressed as $\Sigma_{ww} = \sigma_w^2 I$, where σ_w^2 is the variance of the noise added to each dimension of ϕ.

We first determine the optimal decoder D that minimizes

[2]In particular, the ℓ_1-sensitivity of a linear encoder function doesn't change when the constant term is zero, i.e., $\Delta_1 g_e = \Delta_1(g_e + c), \forall c \in \mathbb{R}^Z$.

$\mathcal{L}$ for a given encoder E, which is given by the following proposition. In particular, the optimal decoder D is not the Moore-Penrose inverse[3] of encoder E.

Proposition 4.1 (Optimal decoder D that minimizes $\mathcal{L}$)**.** *An optimal decoder D that minimizes $\mathcal{L}$ for a given encoder E and σ_w^2 can be expressed as $D = E^\top(EE^\top + \sigma_w^2 I)^{-1}$, and corresponding $\mathcal{L}$ is*

$$\mathcal{L} = Tr(P^\top P) - Tr(P^\top P E^\top (EE^\top + \sigma_w^2 I)^{-1} E), \quad (7)$$

where $Tr(\cdot)$ denotes the trace of a matrix.

The next main step is to find an encoder E that minimizes Eq. (7). Since $\Delta_1 g_e = \max_{v,v' \in E(\mathcal{H})} \|v - v'\|_1$, where $E(\mathcal{H}) = \{Eh | h \in \mathcal{H}\}$ is the image of $\mathcal{H}$ under linear transformation E, the design of encoder E will affect $\Delta_1 g_e$ and therefore σ_w^2, and for different $\mathcal{H}$'s the effect of E is also different in general. So we need to carefully consider the relationship between E and $\mathcal{H}$ through geometric analysis[4].

When computing $\Delta_1 g_e$ we can actually use $\mathcal{H}$'s convex hull $\mathcal{S}$ instead of $\mathcal{H}$ itself, according to the following lemma.

Lemma 4.2 (Convex hull preserves $\Delta_1 g_e$)**.**

$$\Delta_1 g_e = \max_{v,v' \in E(\mathcal{S})} \|v - v'\|_1. \quad (8)$$

For encoder E, we consider its singular value decomposition (SVD) instead: $U\Sigma V^\top$, where $U \in \mathbb{R}^{Z \times Z}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{Z \times n}$ is a rectangular diagonal matrix. And the singular values are denoted by $\sigma_1, \cdots, \sigma_n$ with $|\sigma_1| \geq \cdots \geq |\sigma_n|$. Then designing E is equivalent to designing matrix U, V and Σ. The geometric interpretation of applying transform $E = U\Sigma V^\top$ to set $\mathcal{S}$ consists of three sub-transforms: 1) rotate $\mathcal{S}$ by applying rotation matrix $V^\top$; 2) scale $V^\top(\mathcal{S})$ by applying scaling matrix Σ; 3) rotate $\Sigma V^\top(\mathcal{S})$ by applying rotation matrix U. In general, the choice of any of U, Σ, and V will affect $\Delta_1 g_e$ and hence σ_w^2.

Our overall strategy is to first minimize the loss $\mathcal{L}$ and the sensitivity value $\Delta_1 g_e$ over U and V. Under Assumption 4.3, the resulting U and V only depend on Σ (Propositions 4.4, 4.5). We then find the optimal Σ that minimizes the $\mathcal{L}$ within a given privacy budget ϵ (Proposition 4.6). By the end, we relax Assumption 4.3 to discuss the quality of our solution (Theorem 4.7).

We start by noting that for two points within a compact set, they must lay on the boundary to have the maximum ℓ_1 distance. We then make the following assumption in terms of $\partial\mathcal{S}$, which is the boundary of $\mathcal{S}$. It decouples the relationship between the choice of V and the value of $\Delta_1 g_e$.

Assumption 4.3 (Boundary $\partial\mathcal{S}$ is a centered hypersphere)**.** The boundary $\partial\mathcal{S}$ is a centered hypersphere of radius $r \geq 0$, which is expressed as $\{h \in \mathbb{R}^n | \|h\|_2^2 = r^2\}$.

This is a strong assumption, but at the end of this subsection we will give a lower and upper bound of $\mathcal{L}$ for any possible $\partial\mathcal{S}$ based on the results under this assumption. Since $\partial V(\mathcal{S}) = \{h \in \mathbb{R}^n | \|h\|_2^2 = r^2\} = \partial\mathcal{S}$ for any orthogonal V, this assumption gives us a nice property: the choice of V doesn't affect $\Delta_1 g_e$ and σ_w^2. Based on the above assumption, we can safely consider the optimal design of V that minimizes $\mathcal{L}$ when Σ and σ_w^2 are given, which leads to the following proposition.

Proposition 4.4 (Optimal rotation matrix V that minimizes $\mathcal{L}$ under Assumption 4.3)**.** *Suppose the eigen-decomposition of the Gram matrix $P^\top P$ is expressed as $P^\top PQ = Q\Lambda$, where $\Lambda = diag(\lambda_1, \cdots, \lambda_n) \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal elements are eigenvalues with $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, and $Q \in \mathbb{R}^{n \times n}$ is an orthogonal matrix whose columns are corresponding normalized eigenvectors. Then, when Σ and σ_w^2 are given, $\mathcal{L}$ is minimized for $V = Q$, any $Z \geq n$, and any orthogonal U. And the corresponding $\mathcal{L}$ can be expressed as:*

$$\mathcal{L} = \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \lambda_i \frac{\sigma_i^2}{\sigma_i^2 + \sigma_w^2}. \quad (9)$$

It is clear that choosing a $Z > n$ brings no additional benefit. Hence we can only consider $Z = n$ for simplicity.

After the first two sub-transforms Σ and $V^\top$, the boundary $\partial\mathcal{S} = \{h \in \mathbb{R}^n | \|h\|_2^2 = r^2\}$ becomes $\partial\Sigma V^\top(\mathcal{S}) = \{v \in \mathbb{R}^n | \sum_{i=1}^n v_i^2/\sigma_i^2 = r^2\}$, which is a hyperellipsoid. We then have the following proposition that gives the optimal U which minimizes $\Delta_1 g_e$.

Proposition 4.5 (Optimal rotation matrix U that minimizes $\Delta_1 g_e$ under Assumption 4.3)**.** *For a given Σ, $U = I$ minimizes $\Delta_1 g_e$, and the corresponding minimum value is*

$$\Delta_1 g_e = 2r\sqrt{\sum_{i=1}^{n} \sigma_i^2}. \quad (10)$$

Next, we need to consider how to design the scaling matrix Σ, or equivalently, the values of $\sigma_1^2, \cdots, \sigma_n^2$, to minimize Eq. (9), which is the only remaining piece. Clearly, for any given $\sigma_1^2, \cdots, \sigma_n^2$ if we increase them proportionally, σ_w^2 also needs to be increased proportionally to preserve the same ϵ-LDP. So without loss of generality, we impose an additional constraint $\sum_{i=1}^n \sigma_i^2 = M$, where M is a positive constant. And the following proposition gives the optimal

[3] In matrix factorization (Li et al., 2015; McMahan et al., 2022), the decoder matrix is however constrained to be the Moore-Penrose inverse of the encoder matrix.

[4] When using an **upper bound** of the sensitivity (Li et al., 2015; McMahan et al., 2022), such geometric analysis is not needed. We here however consider the **exact value** of $\Delta_1 g_e$.

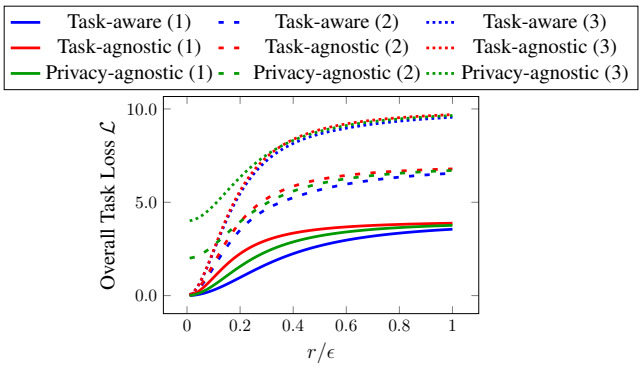

Figure 3. Theoretical overall task loss $\mathcal{L}$ comparison when $L = I$ and Assumption 4.3 holds. We consider three different settings which have $n = 4$ and $\lambda_1 = 4$ in common. Difference between settings: 1) $\lambda_{2:4} = 0$; 2) $\lambda_{2:4} = 1$; 3) $\lambda_{2:4} = 2$. For the privacy-agnostic approach we use $Z = 2$.

choice of Σ that minimizes $\mathcal{L}$ and preserves ϵ-LDP with the Laplace mechanism.

Proposition 4.6 (Optimal scaling matrix Σ that minimizes $\mathcal{L}$ and preserves ϵ-LDP with Laplace mechanism under Assumption 4.3)**.** *The optimal choice of $\sigma_1^2, \cdots, \sigma_n^2$ that minimize $\mathcal{L}$ and preserve ϵ-LDP with Laplace mechanism under constraint $\sum_{i=1}^n \sigma_i^2 = M$ is given by*

$$\sigma_i^2 = \begin{cases} M \cdot \left(\frac{\sqrt{\lambda_i}}{\sum_{i=1}^{Z'} \sqrt{\lambda_i}}\left(1 + Z' \cdot \frac{8r^2}{\epsilon^2}\right) - \frac{8r^2}{\epsilon^2}\right), & \forall i \le Z' \\ 0, & o.w. \end{cases} \tag{11}$$

where $Z' \le n$ is the largest integer such that:

$$\frac{\sqrt{\lambda_{Z'}}}{\sum_{i=1}^{Z'} \sqrt{\lambda_i}}\left(1 + Z' \cdot \frac{8r^2}{\epsilon^2}\right) - \frac{8r^2}{\epsilon^2} > 0, \tag{12}$$

and the corresponding $\mathcal{L}$ is

$$\mathcal{L} = \frac{8r^2/\epsilon^2}{1 + Z' \cdot 8r^2/\epsilon^2}\left(\sum_{i=1}^{Z'} \sqrt{\lambda_i}\right)^2 + \sum_{i=Z'+1}^{n} \lambda_i. \tag{13}$$

For our task-aware approach, Proposition 4.1-4.6 complete the optimal encoder and decoder design that preserves ϵ-LDP with the Laplace mechanism under Assumption 4.3.

Validation of Task Loss from Theoretical Results (Fig. 3). We now compare the performance of our task-aware approach and the benchmark approaches when $L = I$ and Assumption 4.3 holds. The derivations of the benchmark approaches are in Section A.7. We consider three different settings which have $n = 4$ and $\lambda_1 = 4$ in common. And in setting 1, 2, and 3 we let $\lambda_2 = \lambda_3 = \lambda_4$ be 0, 1, 2 respectively. For the privacy-agnostic approach[5], we use a pre-determined $Z = 2$. Our observations are: 1) Compared to the task-agnostic approach, our task-aware approach achieves the largest improvement in setting 1, because $\lambda_{2:4} = 0$ implies that $x_{2:4}$ are purely redundant. We can even expect higher gain than setting 1 when we have larger n and zero $\lambda_{2:n}$'s, and the gain will be zero if all the λ_i's are equal. 2) The privacy-agnostic approach completely missed the information carried by $x_{3:4}$, which explains the improvement of our task-aware approach for small r/ϵ in setting 2 and 3. We can expect higher gain than setting 3 when we have larger n and larger $\lambda_{2:n}$'s, and the gain will be zero if all the missed x_i's correspond to zero λ_i's and all the other λ_i's are equal.

Transition to general boundary $\partial\mathcal{S}$. Based on the results under Assumption 4.3, we can give a lower and upper bound of $\mathcal{L}^*$ for general $\partial\mathcal{S}$, which is not necessarily a centered hypersphere.

Theorem 4.7 (Lower and upper bound of $\mathcal{L}^*$ for general boundary $\partial\mathcal{S}$ when ϵ-LDP is preserved with Laplace mechanism)**.** *Suppose $\partial\mathcal{S} \subset \{h \in \mathbb{R}^n : r_{min}^2 \le \|h\|_2^2 \le r_{max}^2\}$. Then when ϵ-LDP is preserved with the Laplace mechanism, the optimal $\mathcal{L}^*$ is bounded by:*

$$\mathcal{L}(r_{min}; \lambda_{1:n}, \epsilon) \le \mathcal{L}^* \le \mathcal{L}(r_{max}; \lambda_{1:n}, \epsilon) \tag{14}$$

where $\mathcal{L}(r; \lambda_{1:n}, \epsilon)$ is the value of $\mathcal{L}$ determined by Eq. (12) and (13) for given radius r, eigenvalues $\lambda_{1:n}$ and privacy budget ϵ.

Therefore, to preserve ϵ-LDP with the Laplace mechanism, our task-aware solution for general $\partial\mathcal{S}$ is: 1) First, find the smallest $r_{\max}$ and the largest $r_{\min}$ that bound $\partial\mathcal{S}$; 2) Then assume $\partial\mathcal{S}$ is $\{h \in \mathbb{R}^n | \|h\|_2^2 = r_{\max}^2\}$, and choose the encoder E and decoder D based on Proposition 4.4-4.6. We don't use the corresponding σ_w^2 however, because it may guarantee a higher LDP than needed. 3) Next, compute σ_w^2 for real $\partial\mathcal{S}$ under decoder D and privacy budget ϵ.

The associated loss for the task-aware approach is at most $\mathcal{L}(r_{\max}; \lambda_{1:n}, \epsilon)$. Though in general not optimal, it differs from $\mathcal{L}^*$ by at most $\mathcal{L}(r_{\max}; \lambda_{1:n}, \epsilon) - \mathcal{L}(r_{\min}; \lambda_{1:n}, \epsilon)$. The difference is small when $\partial\mathcal{S}$ is "nearly" a hypersphere, i.e., $r_{\max} - r_{\min} \approx 0$.

4.2. General Settings

For more complex scenarios, it is challenging to give an analytical solution to the task-aware privacy preservation problem, especially when the encoder function g_e, decoder function g_d, and task function f correspond to neural networks. Thus, we present a gradient-based learning algorithm. (Benchmark algorithms are given in Section A.7.4.)

[5] One can use another value of Z, under which the result may be slightly different but our task-aware approach will still outperform.

Algorithm 1 Task-aware Algorithm for ϵ-LDP Preservation in General Settings

Require: Privacy budget ϵ and Z

1: Initialize encoder/decoder parameters θ_e, θ_d and noise vector w
2: **for** $\tau \in \{0, 1, \cdots, N_{\text{epochs}} - 1\}$ **do**
3: Update θ_e and θ_d with $-(\nabla_{\theta_e}\mathcal{L} + 2\eta\theta_e)$ and $-\nabla_{\theta_d}\mathcal{L}$, respectively, by one or multiple steps
4: Recompute $\Delta_1 g_e$, and re-sample w from $\text{Lap}^Z(0, \Delta_1 g_e/\epsilon)$
5: **end for**
6: Return θ_e, θ_d and $\Delta_1 g_e$

Algorithm 1 is our proposed task-aware algorithm for general settings. First, the privacy budget ϵ and the latent dimension Z are required inputs for the algorithm. In general, Z should be proper, i.e., it is neither too small (we can find a better solution by choosing a larger Z) nor too big (which introduces unnecessary complexity). In practice a practitioner may need to determine a proper Z on a case-by-case basis (See Section A.9 of the Appendix for more details). Next, the algorithm adopts an alternating iteration approach, where in each epoch, we first update parameters θ_e, θ_d by their corresponding negative gradients in line 3, and then recompute $\Delta_1 g_e$ and re-sample w from $\text{Lap}^Z(0, \Delta_1 g_e/\epsilon)$ in line 4. Note that, in terms of encoder parameter θ_e, instead of considering the gradient of $\mathcal{L}$, we add an ℓ_2 regularization term $\eta\|\theta_e\|_F^2$ where η is a positive constant. Therefore, we update θ_e with the negative gradient $-(\nabla_{\theta_e}\mathcal{L} + 2\eta\theta_e)$. Without regularization, the $\|\theta_e\|_F^2$ will grow to infinity since we can always achieve a smaller $\mathcal{L}$ by increasing the scale of ϕ proportionally. But it is not a direction we are looking for, since σ_w^2 will also increase proportionally to guarantee ϵ-LDP. Moreover, the time complexity of computing $\Delta_1 g_e$ is quadratic in the number of data samples, and when necessary one can split the samples into mini-batches or use parallel processing to reduce the computational time.

5. Evaluation

Our evaluation compares the performance of the proposed task-aware approach and the benchmark approaches (as defined in Section 3). Three applications and corresponding datasets from the standard UCI Machine Learning Repository (Dua & Graff, 2017) are considered: mean estimation of hourly household power consumption, real estate valuation, and breast cancer detection. Configuration and training details are provided in the appendix due to space limitations. Moreover, to show the generality of our task-aware approach with respect to **high-dimensional** image datasets, we provide strong experimental results for MNIST dataset (LeCun et al., 1998) as well, given in Section A.10. Our code is publicly available at https://github.com/chengjiangnan/task_aware_privacy.

5.1. Mean Estimation of Hourly Household Power Consumption

We first consider a mean estimation problem, based on measurements of individual household electric power consumption over four years (Hebrail & Berard, 2012). Each data sample $x \in \mathbb{R}^{24}$ is a time-series that contains the hourly household power consumption for one single day, and our objective is to estimate the mean of the hourly household power consumption for N days. As discussed in Section 4.1, we can define the overall task loss in the following way:

$$\mathcal{L} = \mathbb{E}_{x\sim\mathcal{D}_x}[\|K(\hat{x} - x)\|_2^2] = \sum_{i=1}^{24} k_i^2 \mathbb{E}_{x\sim\mathcal{D}_x}[(\hat{x}_i - x_i)^2]$$

where $K = \text{diag}(k_1, k_2, \cdots, k_{24})$ factors the importance of the mean estimation for each hour. In our experiment we set $k_i = 2$ for $i \in \{9, 10, \cdots, 20\}$ (i.e., day-time hours) and $k_i = 1$ for other i's (i.e., night-time hours). And we adopt a linear encoder and decoder model. As the considered problem is based on a linear model with MSE task loss, we adopt the solutions developed in Section 4.1 for the three approaches (we choose $Z = 3$ for the privacy-agnostic approach).

Fig. 4 shows our experimental results. First, on the left, we compare the task loss $l(\hat{y}, y)$ for the three approaches under different LDP budgets. For each approach, the overall task loss $\mathcal{L}$ decreases when a larger LDP budget ϵ is given. Besides, for a given LDP budget, our task-aware approach always outperforms the benchmark approaches on overall task loss $\mathcal{L}$, and the maximum improvements against the task-agnostic and privacy-agnostic approach are 22.9% ($\epsilon = 10$) and 11.7% ($\epsilon = 5$), respectively. Second, on the right, we select $\epsilon = 5$ and compare the MSE of power consumption for each hour. We see that our task-aware approach achieves a lower MSE for all the day-time hours, and a similar MSE for the night-time hours. This observation can be explained by three reasons: 1) We select a higher k_i for the day-time hours, so our task-aware approach gives higher priority to minimizing the loss for those dimensions in x; 2) Although x has 24 dimensions, the variance in each dimension can be mostly explained by several common latent dimensions, so our task-aware approach still achieves a similar MSE for the night-time hours; 3) Our task-aware approach is able to adopt different scales to different latent dimensions according to their task relevance while the privacy-agnostic approach cannot.

5.2. Real Estate Valuation and Breast Cancer Detection

Next, we consider a real estate valuation problem and a breast cancer detection problem. Both the problems are not based on the linear model with MSE task loss, so we use

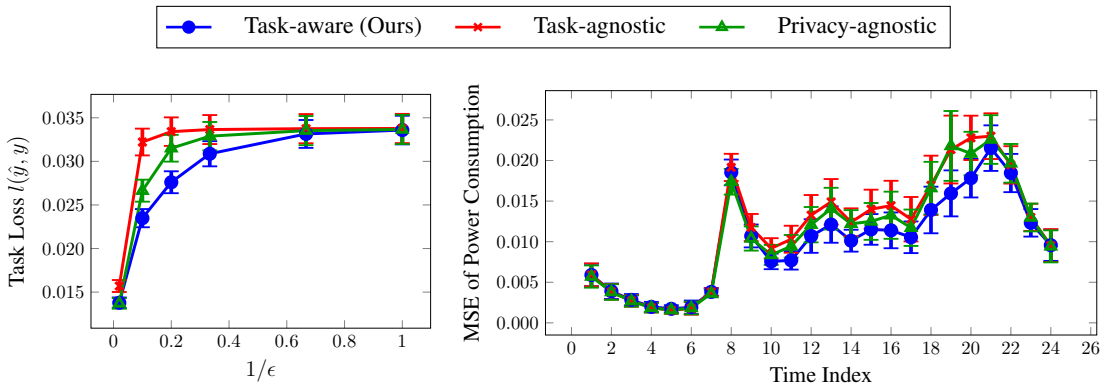

Figure 4. Results of Hourly Household Power Consumption. Left: Task loss $l(\hat{y}, y)$ under different LDP budgets. Right: MSE of power consumption for each hour, when $\epsilon = 5$. Our task-aware approach achieves a lower MSE for all the day-time hours.

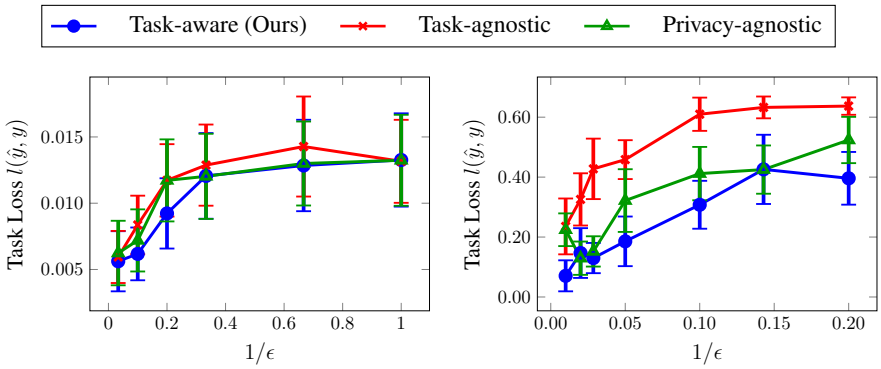

Figure 5. Task loss $l(\hat{y}, y)$ under different LDP budgets for real estate valuation (left) and breast cancer detection (right).

Algorithm 1 developed in Section 4.2 to solve them (we use $Z = 3$ for both our task-aware approach and the privacy-agnostic approach for fair comparison; and the performance of the task-aware approach under different Z's can be found in Section A.9 of the Appendix).

Real Estate Valuation. For this problem, we use historical real estate valuation data collected from Taiwan (Yeh & Hsu, 2018), which contains 400+ instances. Here, $x \in \mathbb{R}^6$ contains 6 attributes that are highly related to the value of a house, including transaction date, house age, geographic coordinates, etc. And $y \in \mathbb{R}$ represents the valuation of a house. We first train a one-hidden-layer feedforward neural network regression model using the ground truth x and y, to serve as our task function f. Then, we minimize the ℓ_2 loss of $\hat{y}$ and y, based on a linear encoder and decoder model.

Breast Cancer Detection. For this problem, we use a well-known breast cancer diagnostic dataset (Street et al., 1993) from Wisconsin, which contains 500+ instances. Here, $x \in \mathbb{R}^{30}$ contains 30 attributes that measure 10 features of a cell nucleus. And y is a binary variable that represents a diagnosis result (malignant or benign). We first train a one-hidden-layer feedforward neural network classification model using the ground truth x and y, to serve as our task function f. Then we aim to minimize the cross-entropy loss of $\hat{y}$ and y, with encoder and decoder both being one-hidden-layer feedforward neural networks.

Fig. 5 shows the evaluation results. For both problems, we can see our task-aware approach nearly always outperforms the benchmark approaches on overall task loss $\mathcal{L}$ under different LDP budgets, which demonstrates the effectiveness of our proposed solution. The maximum improvements against the task-agnostic and privacy-agnostic approach are 26.1% ($\epsilon = 10$) and 21.2% ($\epsilon = 5$) for real estate valuation, and are 70.0% ($\epsilon = 100$) and 68.5% ($\epsilon = 100$) for breast cancer detection.

Limitations. Further research is required to learn anonymized representations that are useful for *multiple* tasks. Moreover, our theoretical guarantee does not apply to general deep neural networks.

6. Conclusion and Future Work

This paper provides a principled task-aware privacy preservation method to improve the privacy-utility trade-off

for ML tasks that increasingly operate on rich, multi-dimensional user data. We gave an analytical near-optimal solution for a general linear encoder-decoder model and MSE task loss, and developed a gradient-based learning algorithm for more general nonlinear settings. Our evaluation showed that our task-aware approach outperforms the benchmark approaches on overall task loss under various LDP budgets.

There are several directions along which we can extend this work. First, it is worthwhile to extend our analysis of the task-aware privacy preservation problem to other LDP mechanisms as well, including mechanisms for approximate LDP. Second, another direction is to consider task-aware privacy preservation for different groups of users in a distributed setting. Lastly, we also plan to find task-aware anonymized representations for multi-task learning.

Funding Disclosure

This material is based upon work supported by the National Science Foundation under Grant No. 2133481. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Acharya, J., Bonawitz, K., Kairouz, P., Ramage, D., and Sun, Z. Context aware local differential privacy. In *International Conference on Machine Learning*, pp. 52–62. PMLR, 2020.

Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., and Palamidessi, C. Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, pp. 39–54. Springer, 2011.

Amos, B., Rodriguez, I. D. J., Sacks, J., Boots, B., and Kolter, J. Z. Differentiable mpc for end-to-end planning and control. *arXiv preprint arXiv:1810.13400*, 2018.

Arachchige, P. C. M., Bertok, P., Khalil, I., Liu, D., Camtepe, S., and Atiquzzaman, M. Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7): 5827–5842, 2019.

Bu, Z., Dong, J., Long, Q., and Su, W. J. Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23), 2020.

Casey, J. *A TREATISE OF THE ANALYTICAL GEOMETRY OF THE POINT, LINE, CIRCLE, AND CONIC SECTIONS.* 1893.

Chen, J.-W., Chen, L.-J., Yu, C.-M., and Lu, C.-S. Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6478–6487, 2021.

Cheng, J., Pavone, M., Katti, S., Chinchali, S., and Tang, A. Data sharing and compression for cooperative networked control. *Advances in Neural Information Processing Systems*, 34, 2021.

Cortés, J., Dullerud, G. E., Han, S., Le Ny, J., Mitra, S., and Pappas, G. J. Differential privacy in control and network systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 4252–4272. IEEE, 2016.

Dankar, F. K. and El Emam, K. Practicing differential privacy in health care: A review. *Trans. Data Priv.*, 6(1): 35–67, 2013.

Differential Privacy Team, A. Learning with privacy at scale. URL https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf.

Ding, B., Kulkarni, J., and Yekhanin, S. Collecting telemetry data privately. *arXiv preprint arXiv:1712.01524*, 2017.

Donti, P., Amos, B., and Kolter, J. Z. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, pp. 5484–5494, 2017.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Dunteman, G. H. *Principal components analysis*. Number 69. Sage, 1989.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

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

Erlingsson, Ú., Pihur, V., and Korolova, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014.

Friedman, A. and Schuster, A. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 493–502, 2010.

Geng, Q. and Viswanath, P. The optimal noise-adding mechanism in differential privacy. *IEEE Transactions on Information Theory*, 62(2):925–951, 2015.

Gondara, L. and Wang, K. Differentially private small dataset release using random projections. In *Conference on Uncertainty in Artificial Intelligence*, pp. 639–648. PMLR, 2020.

Hardt, M. and Talwar, K. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 705–714, 2010.

Hassan, M. U., Rehmani, M. H., and Chen, J. Differential privacy techniques for cyber physical systems: a survey. *IEEE Communications Surveys & Tutorials*, 22(1):746–789, 2019.

Hebrail, G. and Berard, A. Individual household electric power consumption data set, 2012. URL https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption.

Joseph, M., Mao, J., Neel, S., and Roth, A. The role of interactivity in local differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 94–105. IEEE, 2019.

Kairouz, P., Oh, S., and Viswanath, P. Extremal mechanisms for local differential privacy. *arXiv preprint arXiv:1407.1338*, 2014.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Kenthapadi, K., Korolova, A., Mironov, I., and Mishra, N. Privacy via the johnson-lindenstrauss transform. *arXiv preprint arXiv:1204.2606*, 2012.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, C., Miklau, G., Hay, M., McGregor, A., and Rastogi, V. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB journal*, 24 (6):757–781, 2015.

Liu, C., Chakraborty, S., and Mittal, P. Dependence makes you vulnberable: Differential privacy under dependent tuples. In *NDSS*, volume 16, pp. 21–24, 2016.

Liu, R., Cao, Y., Yoshikawa, M., and Chen, H. Fedsel: Federated sgd under local differential privacy with top-k dimension selection. In *International Conference on Database Systems for Advanced Applications*, pp. 485–501. Springer, 2020.

Makhdoumi, A. and Fawaz, N. Privacy-utility tradeoff under statistical uncertainty. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1627–1634. IEEE, 2013.

McMahan, B., Rush, K., and Thakurta, A. G. Private online prefix sums via optimal matrix factorizations. *arXiv preprint arXiv:2202.08312*, 2022.

McMahan, H. B., Andrew, G., Erlingsson, U., Chien, S., Mironov, I., Papernot, N., and Kairouz, P. A general approach to adding differential privacy to iterative training procedures. *arXiv preprint arXiv:1812.06210*, 2018.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Mireshghallah, F., Taram, M., Jalali, A., Elthakeb, A. T., Tullsen, D., and Esmaeilzadeh, H. A principled approach to learning stochastic representations for privacy in deep neural inference. *arXiv preprint arXiv:2003.12154*, 2020.

Muller, U., Ben, J., Cosatto, E., Flepp, B., and Cun, Y. L. Off-road obstacle avoidance through end-to-end learning. In *Advances in neural information processing systems*, pp. 739–746. Citeseer, 2006.

Murakami, T. and Kawamoto, Y. Utility-optimized local differential privacy mechanisms for distribution estimation. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 1877–1894, 2019.

Nakanoya, M., Chinchali, S., Anemogiannis, A., Datta, A., Katti, S., and Pavone, M. Co-design of communication and machine inference for cloud robotics. *Robotics: Science and Systems XVII, Virtual Event*, 2021.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Phan, N., Wang, Y., Wu, X., and Dou, D. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Phan, N., Wu, X., Hu, H., and Dou, D. Adaptive laplace mechanism: Differential privacy preservation in deep learning. In *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 385–394. IEEE, 2017.

Phan, N., Vu, M., Liu, Y., Jin, R., Dou, D., Wu, X., and Thai, M. T. Heterogeneous gaussian mechanism: Preserving differential privacy in deep learning with provable robustness. *arXiv preprint arXiv:1906.01444*, 2019.

Ruhe, A. Perturbation bounds for means of eigenvalues and invariant subspaces. *BIT Numerical Mathematics*, 10(3): 343–354, 1970.

Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248. IEEE, 2013.

Street, W. N., Wolberg, W. H., and Mangasarian, O. L. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pp. 861–870. International Society for Optics and Photonics, 1993.

Thakurta, A. G. and Smith, A. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pp. 819–850. PMLR, 2013.

Von Neumann, J. *Some matrix-inequalities and metrization of matric space*. 1937.

Wang, D., Gaboardi, M., Smith, A., and Xu, J. Empirical risk minimization in the non-interactive local model of differential privacy. *Journal of machine learning research*, 21(200), 2020.

Wang, J., Zhang, J., Bao, W., Zhu, X., Cao, B., and Yu, P. S. Not just privacy: Improving performance of private deep learning in mobile cloud. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2407–2416, 2018.

Wang, N., Xiao, X., Yang, Y., Zhao, J., Hui, S. C., Shin, H., Shin, J., and Yu, G. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 638–649. IEEE, 2019a.

Wang, T., Wu, D. J., Coates, A., and Ng, A. Y. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pp. 3304–3308. IEEE, 2012.

Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1226–1235. PMLR, 2019b.

Wasserman, L. and Zhou, S. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Xiao, X., Wang, G., and Gehrke, J. Differential privacy via wavelet transforms. *IEEE Transactions on knowledge and data engineering*, 23(8):1200–1214, 2010.

Yeh, I.-C. and Hsu, T.-K. Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65:260–271, 2018.

Yiwen, N., Yang, W., Huang, L., Xie, X., Zhao, Z., and Wang, S. A utility-optimized framework for personalized private histogram estimation. *IEEE Transactions on Knowledge and Data Engineering*, 31(4):655–669, 2018.

Zhao, J., Jung, T., Wang, Y., and Li, X. Achieving differential privacy of data disclosure in the smart grid. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pp. 504–512. IEEE, 2014.

Zhou, Y. and Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2018.

A. Appendix

A.1. Proof of Proposition 4.1

Proof. We have

$$\begin{aligned}\mathcal{L} =&\mathbb{E}[\|K(\hat{x}-x)\|_2^2] = \mathbb{E}[\|P(\hat{h}-h)\|_2^2] && (15)\\ =&\mathbb{E}[\|P((DE-I)h+Dw)\|_2^2] && (16)\\ =&\mathbb{E}[\mathrm{Tr}(P(DE-I)hh^\top(DE-I)^\top P^\top)\\ &+\mathrm{Tr}(PDww^\top D^\top P^\top)] && (17)\\ =&\mathrm{Tr}(P(DE-I)(DE-I)^\top P^\top)\\ &+\mathrm{Tr}(PD\Sigma_{ww}D^\top P^\top). && (18)\end{aligned}$$

And we can verify that $D = E^\top(EE^\top + \sigma_w^2 I)^{-1}$ is a zero point of

$$\frac{\nabla\mathcal{L}}{\nabla D} = 2P^\top P((DE-I)E^\top + D\cdot\sigma_w^2 I). \quad (19)$$

Then we plug the expression of D into Eq. (18) and get Eq. (7). □

A.2. Proof of Lemma 4.2

In this subsection we give the proof of Lemma 4.2. Here we treat E as a mapping instead of a matrix, and its inverse mapping, which is also linear, is denoted by E^{-1}.

We first prove the following lemma.

Lemma A.1 (Convex hull after linear transformation)**.** *$E(\mathcal{S})$ is the convex hull of $E(\mathcal{H})$.*

Proof. Since $\mathcal{H}\subseteq\mathcal{S}$, we have $E(\mathcal{H})\subseteq E(\mathcal{S})$. And for any $v, v'\in E(\mathcal{S})$ and $\delta\in[0,1]$, we have $\delta h+(1-\delta)h'\in\mathcal{S}$, where $h\in E^{-1}(v)\cap\mathcal{S}$ and $h'\in E^{-1}(v')\cap\mathcal{S}$. Thus $\delta v+(1-\delta)v' = E(\delta h+(1-\delta)h')\in E(\mathcal{S})$. Therefore, $E(\mathcal{S})$ is a convex set containing $E(\mathcal{H})$.

If $E(\mathcal{S})$ is not the convex hull, then we can find a convex set $\mathcal{B}$ such that $E(\mathcal{H})\subseteq\mathcal{B}\subset E(\mathcal{S})$. Then we have $\mathcal{H}\subseteq E^{-1}(\mathcal{B})\cap\mathcal{S}\subset\mathcal{S}$ (If $E^{-1}(\mathcal{B})\cap\mathcal{S}=\mathcal{S}$, we will have $\mathcal{B}=E(E^{-1}(\mathcal{B})\cap\mathcal{S})=E(\mathcal{S})$, which is not true). Besides, for any $h,h'\in E^{-1}(\mathcal{B})\cap\mathcal{S}$ and $\delta\in[0,1]$, we have $\delta v+(1-\delta)v'\in\mathcal{B}$, where $v=E(h)$ and $v'=E(h')$. Thus $\delta h+(1-\delta)h'\in E^{-1}(\delta v+(1-\delta)v')\cap\mathcal{S}\in E^{-1}(\mathcal{B})\cap\mathcal{S}$. Therefore, $E^{-1}(\mathcal{B})\cap\mathcal{S}$ is a convex set containing $\mathcal{H}$. This is contradictory to the fact that $\mathcal{S}$ is the convex hull of $\mathcal{H}$. So $E(\mathcal{S})$ must be the convex hull of $E(\mathcal{H})$. □

Now we can proceed to the proof of Lemma 4.2.

Proof. Notice that $\Delta_1 g_e = \max_{v,v'\in E(\mathcal{H})}\|v-v'\|_1$, so our target is to prove

$$\max_{v,v'\in E(\mathcal{S})}\|v-v'\|_1 = \max_{v,v'\in E(\mathcal{H})}\|v-v'\|_1. \quad (20)$$

First, since $E(\mathcal{H})\subseteq E(\mathcal{S})$, we have $\max_{v,v'\in E(\mathcal{S})}\|v-v'\|_1\geq\max_{v,v'\in E(\mathcal{H})}\|v-v'\|_1$.

Next, suppose v_1, v_2 are the two points in $E(\mathcal{S})$ such that the ℓ_1 distance between them is the largest. Since $E(\mathcal{S})$ is the convex hull of $E(\mathcal{H})$, we can express v_i as ($\forall i\in\{1,2\}$):

$$\begin{aligned}&v_i = \sum_{j=1}^{p(i)} t_{i,j}\tilde{v}_{i,j}, && (21)\\ \text{s.t.}\quad &\tilde{v}_{i,j}\in E(\mathcal{H}),\quad \forall j\in\{1,2,\cdots,p(i)\}, && (22)\\ &\sum_{j=1}^{p(i)} t_{i,j} = 1, && (23)\\ &t_{i,j}\geq 0,\quad \forall j\in\{1,2,\cdots,p(i)\}, && (24)\end{aligned}$$

where $p(i)$ is an positive integer, $\forall i\in\{1,2\}$.

For convenience, we let $A = \max_{v,v'\in\{\tilde{v}_{i,j}|\forall i,j\}}\|v-v'\|_1$. Clearly, $A\leq\max_{v,v'\in E(\mathcal{H})}\|v-v'\|_1$.

Notice that v_1-v_2 can be expressed as the linear combination of $\tilde{v}_{1,j}-\tilde{v}_{2,q}$, $\forall j\in\{1,2,\cdots,p(1)\}, q\in\{1,2,\cdots,p(2)\}$. That is, $\exists\gamma_{j,q}\geq 0$ such that:

$$\begin{aligned}&v_1-v_2 = \sum_{j=1}^{p(1)}\sum_{q=1}^{p(2)}\gamma_{j,q}(\tilde{v}_{1,j}-\tilde{v}_{2,q}), && (25)\\ &\sum_{j=1}^{p(1)}\sum_{q=1}^{p(2)}\gamma_{j,q} = 1. && (26)\end{aligned}$$

Then we have

$$\begin{aligned}&\|v_1-v_2\|_1 && (27)\\ =&\|\sum_{j=1}^{p(1)}\sum_{q=1}^{p(2)}\gamma_{j,q}(\tilde{v}_{1,j}-\tilde{v}_{2,q})\|_1 && (28)\\ \leq&\sum_{j=1}^{p(1)}\sum_{q=1}^{p(2)}\|\gamma_{j,q}(\tilde{v}_{1,j}-\tilde{v}_{2,q})\|_1 && (29)\\ =&\sum_{j=1}^{p(1)}\sum_{q=1}^{p(2)}\gamma_{j,q}\|(\tilde{v}_{1,j}-\tilde{v}_{2,q})\|_1 && (30)\\ \leq&\sum_{j=1}^{p(1)}\sum_{q=1}^{p(2)}\gamma_{j,q}A = A\leq\max_{v,v'\in E(\mathcal{H})}\|v-v'\|_1. && (31)\end{aligned}$$

So we also have $\max_{v,v'\in E(\mathcal{S})}\|v-v'\|_1\leq\max_{v,v'\in E(\mathcal{H})}\|v-v'\|_1$.

Thus $\Delta_1 g_e = \max_{v,v'\in E(\mathcal{H})}\|v-v'\|_1 = \max_{v,v'\in E(\mathcal{S})}\|v-v'\|_1$. □

A.3. Proof of Proposition 4.4

Proof. We have $\mathrm{Tr}(P^\top P) = \sum_{i=1}^n\lambda_i$, which is a fixed number. Therefore we focus on maximizing the second

trace term in Eq. (7). For any $Z \geq n$ we have

$$EE^\top + \sigma_w^2 I \tag{32}$$
$$= U\Sigma\Sigma^\top U^\top + \sigma_w^2 I \tag{33}$$
$$= U\text{diag}(\sigma_1^2 + \sigma_w^2, \cdots, \sigma_n^2 + \sigma_w^2, \underbrace{\sigma_w^2, \cdots, \sigma_w^2}_{Z-n \text{ in total}})U^\top. \tag{34}$$

Thus we have for any orthogonal U:

$$E^\top(EE^\top + \sigma_w^2 I)^{-1}E \tag{35}$$
$$= V\Sigma^\top U^\top \cdot U\text{diag}(\frac{1}{\sigma_1^2+\sigma_w^2}, \cdots, \frac{1}{\sigma_n^2+\sigma_w^2}, \frac{1}{\sigma_w^2}, \cdots, \frac{1}{\sigma_w^2})U^\top \cdot U\Sigma V^\top \tag{36}$$
$$= V\Sigma^\top \text{diag}(\frac{1}{\sigma_1^2+\sigma_w^2}, \cdots, \frac{1}{\sigma_n^2+\sigma_w^2}, \frac{1}{\sigma_w^2}, \cdots, \frac{1}{\sigma_w^2})\Sigma V^\top \tag{37}$$
$$= V\text{diag}(\frac{\sigma_1^2}{\sigma_1^2+\sigma_w^2}, \cdots, \frac{\sigma_n^2}{\sigma_n^2+\sigma_w^2})V^\top. \tag{38}$$

So $E^\top(EE^\top + \sigma_w^2 I)^{-1}E$ is a positive semi-definite matrix with eigen-values $\frac{\sigma_1^2}{\sigma_1^2+\sigma_w^2} \geq \cdots \geq \frac{\sigma_n^2}{\sigma_n^2+\sigma_w^2} \geq 0$. Then by Ruhe's trace inequality (Ruhe, 1970) (a corollary of Von Neumann's trace inequality (Von Neumann, 1937)):

$$\text{Tr}(P^\top PE^\top(EE^\top + \sigma_w^2 I)^{-1}E) \tag{39}$$
$$\leq \sum_{i=1}^{n} \lambda_i \frac{\sigma_i^2}{\sigma_i^2 + \sigma_w^2}, \tag{40}$$

and when $V = Q$ the equality holds. So we have Eq. (9) for $V = Q$, any $Z \geq n$ and any orthogonal U. □

A.4. Proof of Proposition 4.5

In this subsection we give the proof of Proposition 4.5. For convenience we let $a_i = r|\sigma_i|, \forall i \in \{1, 2, \cdots, n\}$, and then the hyperellipsoid $\{v \in \mathbb{R}^n | \sum_{i=1}^n v_i^2/\sigma_i^2 = r^2\}$ can be written as $\{v \in \mathbb{R}^n | \sum_{i=1}^n v_i^2/a_i^2 = 1\}$, which is the standard expression. And we also have $a_1 \geq a_2 \geq \cdots \geq a_n$.

Before proving Proposition 4.5, we first give two lemmas related to the properties of a hyperellipsoid.

Lemma A.2 (Tangent hyperplane of hyperellipsoid)**.** *Any tangent hyperplane of hyperellipsoid $\{v \in \mathbb{R}^n | \sum_{i=1}^n v_i^2/a_i^2 = 1\}$ can be expressed as:*

$$\sum_{j=1}^{n} u_j v_j = \sqrt{\sum_{j=1}^{n} a_j^2 u_j^2}, \tag{41}$$

where $u_1, u_2, \cdots, u_n$ are the coefficients.

Proof. It can be easily verified that point $\tilde{v} \in \mathbb{R}^n$ such that

$$\tilde{v}_j = \frac{a_j^2 u_j}{\sqrt{\sum_{q=1}^{n} a_q^2 u_q^2}}, \quad j \in \{1, 2, \cdots, n\}, \tag{42}$$

is located on the hyperellipsoid. And by adjusting the values of $u_1, u_2, \cdots, u_n$ we can express any point on the hyperellipsoid with Eq. (42). Moreover, the tangent hyperplane for point of tangency $\tilde{v}$ can be expressed as

$$\sum_{j=1}^{n} \frac{\tilde{v}_j v_j}{a_j^2} = 1. \tag{43}$$

Plugging Eq. (42) into Eq. (43) we get Eq. (41). □

Lemma A.3 (Locus of the vertices of circumscribed orthotope)**.** *The vertices of any orthotope that circumscribes hyperellipsoid $\{v \in \mathbb{R}^n | \sum_{i=1}^n v_i^2/a_i^2 = 1\}$ is on hypersphere $\{v \in \mathbb{R}^n | \sum_{i=1}^n v_i^2 = \sum_{i=1}^n a_i^2\}$.*

Proof. When $n = 2$, the hypersphere reduces to a circle that is well-known as the orthoptic circle of an ellipse (Casey, 1893). We hereby generalize this result to any n.

A vertex of a circumscribed orthotope can be viewed as the intersection of n tangent hyperplanes. According to Lemma A.2, we can express them as:

$$\sum_{j=1}^{n} u_{i,j} v_j = \sqrt{\sum_{j=1}^{n} a_j^2 u_{i,j}^2}, \quad \forall i \in \{1, 2, \cdots, n\}, \tag{44}$$

where i is the index of the i-th hyperplane. Here we let $\sum_{j=1}^n u_{i,j}^2 = 1, \forall j \in \{1, 2, \cdots, n\}$. Besides, these hyperplanes are perpendicular to each other, so the coefficients also satisfy $\sum_{i=1}^n u_{i,j} u_{i,k} = 0, \forall k \neq j$. So if we let $\Omega \in \mathbb{R}^{n \times n}$ be a matrix with $u_{i,j}$ on the i-th row and j-th column, $\forall i, j$, then Ω is an orthogonal matrix, and we further have $\sum_{i=1}^n u_{i,j}^2 = 1, \forall i \in \{1, 2, \cdots, n\}$.

Thus the considered vertex satisfies Eq. (44), $\forall i \in \{1, 2, \cdots, n\}$, which implies it also satisfies:

$$\sum_{i=1}^{n} (\sum_{j=1}^{n} u_{i,j} v_j)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_j^2 u_{i,j}^2. \tag{45}$$

For the left hand side, we have

$$\sum_{i=1}^{n}(\sum_{j=1}^{n}u_{i,j}v_j)^2 \tag{46}$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}u_{i,j}^2v_j^2+2\sum_{i=1}^{n}\sum_{j\neq p}u_{i,j}u_{i,k}v_jv_p \tag{47}$$

$$=\sum_{j=1}^{n}(\sum_{i=1}^{n}u_{i,j}^2)v_j^2+2\sum_{j\neq p}(\sum_{i=1}^{n}u_{i,j}u_{i,k})v_jv_p \tag{48}$$

$$=\sum_{j=1}^{n}v_j^2. \tag{49}$$

And for the right hand side,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}a_j^2u_{i,j}^2=\sum_{j=1}^{n}(\sum_{i=1}^{n}u_{i,j}^2)a_j^2=\sum_{j=1}^{n}a_j^2. \tag{50}$$

Thus the vertex is on hypersphere $\{v\in\mathbb{R}^n|\sum_{i=1}^n v_i^2=\sum_{i=1}^n a_i^2\}$. □

Since the equation of a centered hypersphere after rotation remains unchanged, we have the following corollary.

Corollary A.4 (Locus of the vertices of circumscribed orthotope after rotation)**.** *The vertices of any orthotope that circumscribes hyperellipsoid* $\{v\in\mathbb{R}^n|\sum_{i=1}^n v_i^2/a_i^2=1\}$ *after rotation is on hypersphere* $\{v\in\mathbb{R}^n|\sum_{i=1}^n v_i^2=\sum_{i=1}^n a_i^2\}$.

Now we can proceed to the proof of Proposition 4.5.

Proof. We first consider the case when n is a power of 2.

For the rotated hyperellipsoid, we consider whether there's any point $\tilde{v}\in\mathbb{R}^n$ s.t. $\|\tilde{v}\|_1\geq\sqrt{\sum_{j=1}^n a_j^2}$. Suppose we couldn't find such a point. Then consider any tangent hyperplane whose normal vector has the following form: $(u_1,u_2,\cdots,u_n)$, s.t. $u_i=\pm1,\forall i\in\{1,2,\cdots,n\}$. It can be expressed as:

$$\sum_{j=1}^{n}u_jv_j=W(u_1,u_2,\cdots,u_n), \tag{51}$$

where $W:(u_1,u_2,\cdots,u_n)\mapsto\mathbb{R}$ maps $(u_1,u_2,\cdots,u_n)$ to a corresponding constant. And we have $W(u_1,u_2,\cdots,u_n)<\sqrt{\sum_{j=1}^n a_j^2}$.

Since n is a power of 2, we can find a Hadamard matrix Ω in $\mathbb{R}^{n\times n}$ whose elements are either 1 or -1, such that $\Omega\Omega^\top=nI$. We let $u_{i,j}$ be the element of R on the i-th row and j-th column, $\forall i,j$, and consider the intersection of the following n tangent hyperplanes ($\forall i\in\{1,2,\cdots,n\}$):

$$\sum_{j=1}^{n}u_{i,j}v_j=W(u_{i,1},u_{i,2},\cdots,u_{i,n}), \tag{52}$$

whose intersection point must satisfy:

$$\sum_{i=1}^{n}(\sum_{j=1}^{n}u_{i,j}v_j)^2 \tag{53}$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}(W(u_{i,1},u_{i,2},\cdots,u_{i,n}))^2. \tag{54}$$

Similar to the proof of Lemma A.3, we can easily prove the left hand side equals $n\sum_{j=1}^n v_j^2$, and the right hand side is strictly less than $n\sum_{j=1}^n a_j^2$. Thus the intersection point doesn't locate on the hypersphere $\{v\in\mathbb{R}^n|\sum_{i=1}^n v_i^2=\sum_{i=1}^n a_i^2\}$. But since the considered n hyperplanes are also the surfaces of a circumscribed orthotope, the intersection point is hence a vertex and must be on the hypersphere $\{v\in\mathbb{R}^n|\sum_{i=1}^n v_i^2=\sum_{i=1}^n a_i^2\}$, according to Corollary A.4. This contradiction means that, the assumption that we cannot find a point $\tilde{v}\in\mathbb{R}^n$ s.t. $\|\tilde{v}\|_1\geq\sqrt{\sum_{j=1}^n a_j^2}$ is false.

Thus the point $\tilde{v}\in\mathbb{R}^n$ s.t. $\|\tilde{v}\|_1\geq\sqrt{\sum_{j=1}^n a_j^2}$ exists, and the ℓ_1 distance between $\tilde{v}$ and $-\tilde{v}$ (which both lie on the hyperepllisoid) is $2\sqrt{\sum_{j=1}^n a_j^2}$. Thus we have $\Delta_1 g_e\geq 2\sqrt{\sum_{j=1}^n a_j^2}$.

For $U=I$, which means we don't actually rotate the ellipsoid $\{v\in\mathbb{R}^n|\sum_{i=1}^n v_i^2/a_i^2=1\}$, we have for any point on the ellipsoid

$$(\sum_{i=1}^{n}|v_i|)^2=(\sum_{i=1}^{n}a_i\cdot\frac{|v_i|}{a_i})^2 \tag{55}$$

$$\leq(\sum_{i=1}^{n}a_i^2)(\sum_{i=1}^{n}\frac{v_i^2}{a_i^2})=\sum_{i=1}^{n}a_i^2, \tag{56}$$

according to the Cauchy-Schwartz inequality. This implies any point v on the hyperellipsoid has $\|v\|_1\leq\sqrt{\sum_{j=1}^n a_j^2}$. So $\Delta_1 g_e\leq 2\sqrt{\sum_{j=1}^n a_j^2}$. Combined with the result in the above paragraph we know $\Delta_1 g_e=2\sqrt{\sum_{j=1}^n a_j^2}$ for $U=I$.

Thus the proposition is proved for n being a power of 2. For other n's, we can treat the considered hyperellipsoid as a degenerated hyperellipsoid in space $\mathbb{R}^{\tilde{n}}$, where $\tilde{n}$ is the smallest power of 2 such that $\tilde{n}>n$. This implies the proposition still holds.

Therefore, the proposition is true for any n. □

A.5. Proof of Proposition 4.6

Proof. First, to preserve ϵ-LDP with Laplace mechanism, the minimum σ_w^2 required is:

$$\sigma_w^2=2\cdot\frac{(\Delta_1 g_e)^2}{\epsilon^2}=2\cdot\frac{4r^2\cdot\sum_{i=1}^n\sigma_i^2}{\epsilon^2}=\frac{8r^2M}{\epsilon^2}, \tag{57}$$

based on Proposition 4.5 and constraint $\sum_{i=1}^{n}\sigma_i^2 = M$.

Next we need to determine $\sigma_1^2, \cdots, \sigma_n^2$. The considered problem is an optimization problem which aims at minimizing $\mathcal{L}$ in Eq. (9) under constraint $\sum_{i=1}^{n}\sigma_i^2 = M$ and $\sigma_i^2 \geq 0$ (here we view σ_i^2 instead of σ_i as the decision variable). Note that though in Eq. (9) we also have $\sigma_1^2 \geq \sigma_n^2$, we don't need to explicitly consider this constraint, because minimizing $\mathcal{L}$ will implicitly guarantee that, according to the rearrangement inequality. This problem can be solved by Karush-Kuhn-Tucker (KKT) approach, with the following Lagrangian function:

$$F(\sigma_1^2, \cdots, \sigma_n^2, \alpha_1, \cdots, \alpha_n, \beta) \tag{58}$$

$$= \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\lambda_i \frac{\sigma_i^2}{\sigma_i^2 + \sigma_w^2} \tag{59}$$

$$+ \sum_{i=1}^{n}\alpha_i(-\sigma_i^2) + \beta(\sum_{i=1}^{n}\sigma_i^2 - M), \tag{60}$$

where $\alpha_1, \cdots, \alpha_n$ and β are Lagrangian multipliers. We know that the solution will automatically guarantee $\sigma_1^2 \geq \cdots \geq \sigma_n^2$, so we can safely assume there exists $Z' \leq n$ such that $\alpha_i = 0$ for $i \leq Z'$, and $\alpha_i > 0$ for $i > Z'$. Then for $i > Z'$ we have $\sigma_i^2 = 0$, and for $i \leq Z'$ we have

$$\frac{\nabla F}{\nabla \sigma_i^2} = -\frac{\sigma_w^2}{(\sigma_i^2 + \sigma_w^2)^2}\lambda_i = \beta. \tag{61}$$

Combining with $\sum_{i=1}^{Z'}\sigma_i^2 = M$ and Eq. (57) we eventually get Eq. (11). Enforcing $\sigma_{Z'}^2 > 0$ we get Eq. (12). And plugging Eq. (11) into Eq. (9) we get Eq. (13). □

A.6. Proof of Theorem 4.7

In this subsection we give the proof of Theorem 4.7.

We start with two definitions: $\mathcal{L}^*(\partial\mathcal{S}; P, \epsilon)$ denotes the optimal loss for any boundary $\partial\mathcal{S}$, task matrix P that preserves ϵ-LDP with Laplace mechanism; $R(r) = \{h \in \mathbb{R}^n | \|h\|_2^2 = r^2\}$ is the centered hypersphere with radius r.

Next we give the following lemma.

Lemma A.5 (Invariance of the optimal loss after scaling)**.**

$$\mathcal{L}^*(\partial\mathcal{S}; P, \epsilon) = \mathcal{L}^*(\partial\rho(\mathcal{S}); P, \rho\epsilon), \tag{62}$$

where $\rho > 0$ is a scalar and $\rho(\mathcal{S}) = \{\rho h | h \in \mathcal{S}\}$.

Proof. We only need to consider fixed task matrix P. For any encoder E, decoder D and noise variance σ_w^2, if they preserve ϵ-LDP with Laplace mechanism for boundary $\partial\mathcal{S}$, then we have

$$\sigma_w \geq \sqrt{2}\cdot\frac{\Delta_1 g_e}{\epsilon} = \sqrt{2}\cdot\frac{\Delta_1 \rho g_e}{\rho\epsilon}, \tag{63}$$

where

$$\Delta_1 \rho g_e = \rho \max_{h,h' \in \mathcal{S}} \|Eh - Eh'\|_1 \tag{64}$$

$$= \max_{v,v' \in \rho(\mathcal{S})} \|Ev - Ev'\|_1. \tag{65}$$

Thus we know encoder E, decoder D and noise variance σ_w^2 also preserve $\rho\epsilon$-LDP with Laplace mechanism for boundary $\partial\rho(\mathcal{S})$. And the reverse also holds true.

So we must have $\mathcal{L}^*(\partial\mathcal{S}; P, \epsilon) = \mathcal{L}^*(\partial\rho(\mathcal{S}); P, \rho\epsilon)$. □

Now we can proceed to the proof of Theorem 4.7.

Proof. We first construct a distribution $\mathcal{D}_{h'}$, s.t. points drawn from $\mathcal{D}_{h'}$ are uniformly distributed on $R(2^{n-1})$. Then $h' \sim \mathcal{D}_{h'}$ satisfies $\Sigma_{h'h'} = I$. And we also have $\partial\mathcal{S}' = R(2^{n-1})$, where $\mathcal{S}'$ is the convex hull of H' (H' is the domain of $h' \sim \mathcal{D}_{h'}$). According to Proposition 4.6, for $h' \sim \mathcal{D}_{h'}$ we have $\mathcal{L}^*(R(2^{n-1}); P, \rho\epsilon) = \mathcal{L}(2^{n-1}; \lambda_{1:n}, \rho\epsilon), \forall\rho, \epsilon > 0$.

We next consider scalar $\rho = 2^{n-1}/r_{\min}$ and the optimal loss $\mathcal{L}^*(\partial\rho(\mathcal{S}); P, \rho\epsilon)$. For any encoder E, decoder D and noise variance σ_w^2, if they preserve $\rho\epsilon$-LDP to boundary $\partial\rho(\mathcal{S})$, then they preserve at least $\rho\epsilon$-LDP to boundary $R(2^{n-1})$, since $R(2^{n-1}) \subset \rho(\mathcal{S})$. Thus we have $\mathcal{L}^*(\partial\rho(\mathcal{S}); P, \rho\epsilon) \geq \mathcal{L}^*(R(2^{n-1}); P, \rho\epsilon)$.

This further implies:

$$\mathcal{L}^*(\partial\mathcal{S}; P, \epsilon) = \mathcal{L}^*(\partial\rho(\mathcal{S}); P, \rho\epsilon) \tag{66}$$

$$\geq \mathcal{L}^*(R(2^{n-1}); P, \rho\epsilon) \tag{67}$$

$$= \mathcal{L}(2^{n-1}; \lambda_{1:n}, \rho\epsilon) = \mathcal{L}(r_{\min}; \lambda_{1:n}, \epsilon). \tag{68}$$

So the lower bound of Theorem 4.7 is proved. The upper bound can be proved in the same way. □

A.7. Benchmarks

A.7.1. Task loss for task-agnostic approach

One can obtain the resultant optimal $\mathcal{L}$ for the task-agnostic approach by letting $E = L$ and using Eq. (7) as stated in Proposition 4.1. It is worth noting that the associated decoder $D = L^\top(LL^\top + \sigma_w^2 I)^{-1}$ is not an identity matrix in general.

Corollary A.6 (Optimal $\mathcal{L}$ for the task-agnostic approach that preserves ϵ-LDP)**.** *For the task-agnostic approach, the optimal $\mathcal{L}$ that preserves ϵ-LDP is*

$$\mathcal{L} = Tr(P^\top P) - Tr(P^\top P L^\top (LL^\top + \sigma_w^2 I)^{-1} L), \tag{69}$$

where $\sigma_w^2 = 2(\Delta_1 g_e)^2/\epsilon^2$ with $g_e(x) = x$.

Table 1. Evaluation Details

APPLICATION	NUM OF SAMPLES	TRAIN/TEST SPLIT	TRAINING EPOCHS	RUNTIME
HOUSEHOLD POWER	1417	0.7/0.3	NA	$<$ 1 MIN
REAL ESTATE	414	0.7/0.3	2000	$<$ 2 HRS
BREAST CANCER	569	0.7/0.3	2000	$<$ 2 HRS

A.7.2. TASK LOSS FOR PRIVACY-AGNOSTIC APPROACH

Through similar analysis as Proposition 4.4, one can obtain the resultant optimal $\mathcal{L}$ for the privacy-agnostic approach, which has a pre-determined $Z \leq n$.

Corollary A.7 (Optimal $\mathcal{L}$ for the privacy-agnostic approach that preserves ϵ-LDP)**.** *For the privacy-agnostic approach with a pre-determined $Z \leq n$, the optimal $\mathcal{L}$ that preserves ϵ-LDP is*

$$\mathcal{L} = \sum_{i=1}^{Z} \lambda_i \frac{\sigma_w^2}{\sigma_i^2 + \sigma_w^2} + \sum_{i=Z+1}^{n} \lambda_i. \tag{70}$$

where $\sigma_w^2 = 2(\Delta_1 g_e)^2/\epsilon^2$.

Proof. For privacy-agnostic approach, we have $Z \leq n$. Since for encoder we need to select the top-Z principal components, we have $V = Q$. Similar to the proof of Proposition 4.4, for a given $Z \leq n$, we have

$$EE^\top + \sigma_w^2 I \tag{71}$$
$$= U\Sigma\Sigma^\top U^\top + \sigma_w^2 I \tag{72}$$
$$= U\text{diag}(\sigma_1^2 + \sigma_w^2, \cdots, \sigma_Z^2 + \sigma_w^2)U^\top, \tag{73}$$

and through similar derivations we eventually get

$$\text{Tr}(P^\top P E^\top (EE^\top + \sigma_w^2 I)^{-1} E) \tag{74}$$
$$= \sum_{i=1}^{Z} \lambda_i \frac{\sigma_i^2}{\sigma_i^2 + \sigma_w^2}. \tag{75}$$

Combined with Eq. (7) we get Eq. (70). □

A.7.3. TASK LOSS FOR BENCHMARK APPROACHES WHEN $L = I$ AND ASSUMPTION 4.3 HOLDS

When $L = I$ and Assumption 4.3 holds, for the task-agnostic approach, according to Eq. (69), we have

$$\mathcal{L} = \frac{\sigma_w^2}{1 + \sigma_w^2} \cdot \text{Tr}(P^\top P) = \frac{n \cdot 8r^2/\epsilon^2}{1 + n \cdot 8r^2/\epsilon^2} \sum_{i=1}^{n} \lambda_i. \tag{76}$$

For the privacy-agnostic approach, according to Eq. (70) and Eq. (10), and assuming we have equal σ_i's and minimum $\Delta_1 g_e$, we get

$$\mathcal{L} = \frac{Z \cdot 8r^2/\epsilon^2}{1 + Z \cdot 8r^2/\epsilon^2} \sum_{i=1}^{Z} \lambda_i + \sum_{i=Z+1}^{n} \lambda_i, \tag{77}$$

where the value of Z is pre-determined.

A.7.4. BENCHMARK ALGORITHMS UNDER GENERAL SETTINGS

For the privacy-agnostic approach, we first train the encoder and decoder without considering privacy preservation by updating θ_e and θ_d with $-\nabla_{\theta_e}\mathcal{L}$ and $-\nabla_{\theta_d}\mathcal{L}$, respectively. Next, we fix encoder parameters θ_e and train the decoder with input $\phi + w$ (a modification of Algorithm 1 line 3-4). The task-agnostic approach trains the decoder in the same way, but fixes g_e to an identity mapping function.

A.8. Configuration and Training Details of Evaluation

Our evaluation runs on a personal laptop with 2.7 GHz Intel Core I5 processor and 8-GB 1867 MHz DDR3 memory. Our code is based on Pytorch. We use the Adam optimizer and learning rate 10^{-3} for all the applications. The number of samples, train/test split, training epochs, and resulting runtime are summarized in Table 1. (Note that the evaluation for hourly household power consumption is based on the theoretical solutions, so "training epochs", which is associated with the gradient-based method, doesn't apply.) All three datasets cited in the evaluation are publicly-available from the standard UCI Machine Learning Repository (Dua & Graff, 2017) and anonymized using standard practices. The individual dataset licenses are not available.

For task function f, we use a one-hidden-layer feedforward neural network with input size n, hidden size $1.5n$ and output size 1 in both the real estate valuation and breast cancer detection experiments. The activation function used by the hidden layer and output layer is a Rectified Linear Unit (ReLU). In our experiments, we find that the chosen network architecture is good enough to yield near-zero loss with ground truth x and y, and to **avoid overfitting** we don't choose a deep neural network. For example, we do not see any task improvement using a two layer network.

For the encoder/decoder, we use a one-layer neural network (linear model) with input and output size n in the real estate valuation experiment. For this experiment, a linear encoder/decoder model is already enough to provide good performance. We use one-hidden-layer feedforward neural network with input size n, hidden size n and output size n in the breast cancer detection experiment. The activation functions used by the hidden layer and output layer are a

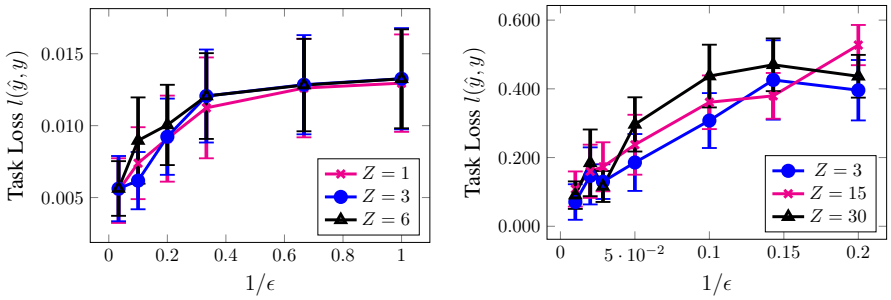

Figure 6. Task loss $l(\hat{y}, y)$ of task-aware approach under different Z's for real estate valuation (left) and breast cancer detection (right).

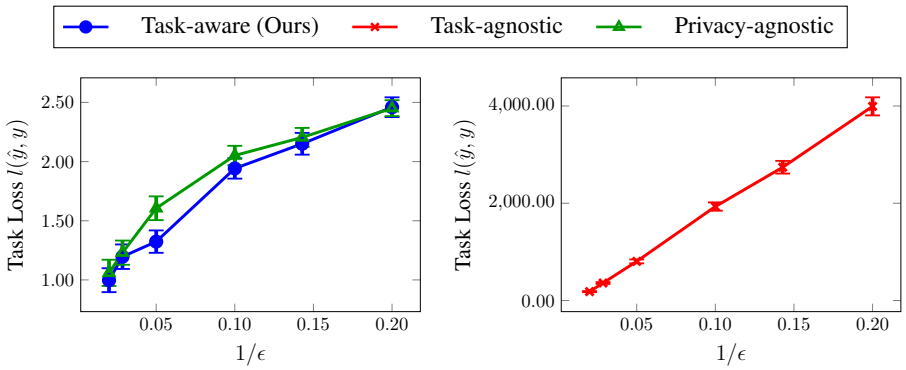

Figure 7. Task loss $l(\hat{y}, y)$ under different LDP budgets for handwritten digit recognition.

logistic and identity function, respectively.

For the gradient-based learning algorithm developed in Section 4.2, we set $\eta = 0.2$ and $\eta = 0.001$ in real estate valuation and breast cancer detection experiments, respectively; and in both experiments, for each epoch we update θ_e and θ_d by 15 steps.

A.9. A Proper Z should be Determined on a Case-by-case Basis

As mentioned in Section 4.2, a practitioner may need to determine a proper Z on a case-by-case basis for our task-aware approach. Fig. 6 illustrates the performance of our task-aware approach under different Z's for real estate valuation and breast cancer detection experiments. We can observe that in both two experiments, we obtained the best performance on average when $Z = 3$, i.e., $n/2$ for real estate valuation and $n/10$ for breast cancer detection[6].

A.10. Experiment with High-dimensional Data

To illustrate our task-aware approach in Section 4.2 also works well for high-dimensional data, such as image data, we consider a handwritten digit recognition problem with well known MNIST dataset (LeCun et al., 1998). Here, $x \in \mathbb{R}^{784}$ represents a 28×28 image of handwritten digit. And $y \in \{0, 1, \cdots, 9\}$ is a discrete variable represents the digit in the image. We first train a convolutional neural network (CNN) classification model using the ground truth x and y, to serve as our task function f. (The CNN classifier is composed of two consecutive convolution layers and a final linear layer. The number of input channels, the number of output channels, kernel size, stride and padding for two convolution layers are 1, 16, 5, 1, 2 and 16, 32, 5, 1, 2 respectively, and ReLU activation and max pooling with kernel size 2 are used after each convolution layer. The final linear layer has input size 1568 and output size 1.) Then we aim to minimize the cross-entropy loss of $\hat{y}$ and y, with linear encoder and decoder. We use $Z = 3$ for both our task-aware approach and the privacy-agnostic approach.

Fig. 7 shows the evaluation result. Since the task loss $\mathcal{L}$ of the task-agnostic approach is much larger than the other two approaches, we put it in a separate sub-figure on the left. On the right, we can see our task-aware approach always outperforms the privacy-agnostic approach on overall task loss $\mathcal{L}$ under different LDP budgets, which demonstrates the effectiveness of our proposed solution. The maximum improvement against the privacy-agnostic approach is 21.3% ($\epsilon = 20$).

[6]The privacy-agnostic approach also achieves the best performance on average under the chosen Z's.