Position Paper: Open Research Challenges for Private Advertising Systems under Local Differential Privacy

Matilde Tullii, Solenne Gaucher, Hugo Richard, Eustache Diemert, Vianney Perchet, Alain Rakotomamonjy, Clément Calauzènes, Maxime Vono

Position Paper: Open Research Challenges for Private Advertising Systems under Local Differential Privacy

Matilde Tullii [*2], Solenne Gaucher[*2], Hugo Richard[*1], Eustache Diemert[1], Vianney Perchet[1, 2], Alain Rakotomamonjy[1], Clément Calauzènes[1], and Maxime Vono[1]

[1]Criteo AI Lab, France
[2]ENSAE, Crest, France

February 5, 2024

Abstract

Due to the ongoing deprecation of third-party cookies on mainstream browsers, the digital advertising industry is facing novel challenges regarding how to operate artificial intelligence (AI) systems. One of these bottlenecks lies in the tentative use of local differential privacy (LDP) to obfuscate granular user data, preventing from using standard machine learning pipelines to tackle the privacy/utility trade-off. This position paper reviews the main research directions that have been explored to cope with this issue and states the main positioning and research guidelines regarding how to operate an AI system under LDP, notably by pointing out the main limitations of existing work. More specifically, we highlight the importance of conducting research works focusing on multi-task learning under LDP schemes and of seeking prior information to help design privacy-preserving mechanisms. We hope this paper will incentivize the whole industry and academic research communities to address the open research questions we are underlying, which could also serve other industrial applications.

1 Introduction

Advertising Systems on the Open Internet. In the current web ecosystem, targeted or behavioral advertising is performed by allowing advertisers (*e.g.* e-commerce websites) to display personalized advertisement into ad banners

[*]Equal contribution

monetized by publishers' websites, potentially through a third-party entity such as an adtech company (Yuan et al., 2014). To build and show such personalized content on a specific publisher's webpage, advertisers need to (i) submit a bid to win an auction allowing them to show their ad, (ii) choose which products to recommend to the user, and (iii) build the ad creative that will maximize user intent. Such use-cases are performed by collecting and processing user personal data, including features and labels, to train machine learning (ML) models that are part of sophisticated artificial intelligence (AI) systems (McMahan et al., 2013; Agarwal et al., 2014; He et al., 2014; Chapelle et al., 2015). Indeed, the latter do not only consist of learning a single ML model but rather thousands having some dependencies between them and tackling precise goals, such as estimating the clearing price of an auction, the incremental revenue from an ad banner, or the time spent by a user on a specific webpage (Chapelle, 2014). In addition, a single ML model life cycle does not consist of a single learning step but of multiple sequential ones for the sake of monitoring, testing, and auditing. As an example, these steps are necessary for adaptive feature discovery and selection, or to perform model bias correction (Khalid et al., 2014; Hao et al., 2020).

Privacy-Preserving Browser Vendors. Among the techniques leveraged by advertisers to collect a sufficient amount of data, the most widespread one relies on so-called third-party cookies. The latter allows to gather features and labels associated with the same user by linking her activity on different websites, such as her last purchase across all visited retailer websites or the number of clicks performed in the past 24 hours. Yet, in past years, growing concerns regarding user privacy have led to substantial efforts from browser vendors to limit cross-website tracking (Erlingsson et al., 2014; Carey et al., 2023). As an example, Safari and Mozilla Firefox deprecated third-party cookies in 2017 and Google Chrome has started to deprecate them from early January 2024 for 1% of its users. To still support the industry that is funding the open web, browser vendors proposed a set of application programming interfaces (APIs) to mostly allow for reporting and measurement (*i.e.* data analytics) use cases. For instance, Apple, Meta/Mozilla Firefox, Microsoft Edge, and Google Chrome proposed the Private Ad Measurement system, the Interoperable Private Attribution, MaskedLark, and the Privacy Sandbox, respectively (Aksu et al., 2023). Unfortunately, the ML use-case and a fortiori how to operate an AI system, being much more complicated, are not specifically addressed by such proposals; *leaving room for innovation and for the ML community to work on associated open research challenges* that are described in the following paragraphs.

Differentially Private Learning Environment. In all current browser vendor proposals, the learning environment could be summarised as follows. Namely, an AI system (*e.g.* ran by an adtech company), is interacting with each user's data via a so-called *trusted entity* involving privacy and data governance restrictions. To provide *privacy* guarantees when releasing data, the trusted entity can obfuscate raw user data by (i) restricting the type of data queries that could be made (*e.g.* releasing only a pre-determined number of features per query), and (ii) transforming the outputs of such queries into differential private

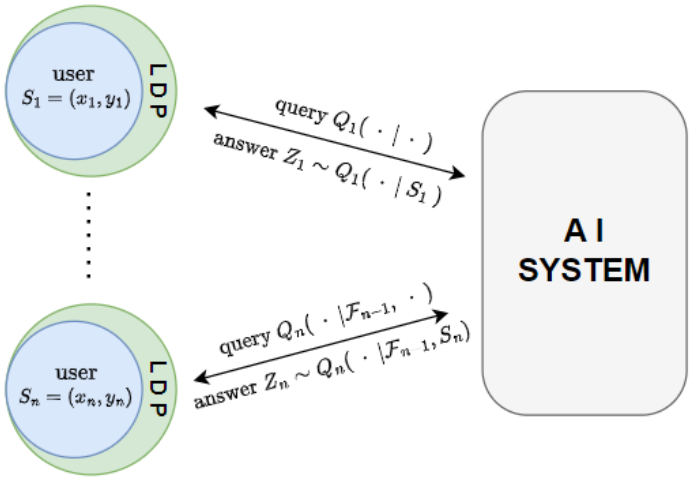

Figure 1: **The local differential privacy framework**

data, a commonly agreed standard within the advertising industry (Zucker-Scharff, 2023). A design proposition of such a system exists in the context of Google Chrome's Privacy Sandbox. However, it is mostly designed to handle data analysis use cases, but not learning ones. From a learning perspective, the main uncertainties and roadblocks are associated with the trade-offs browser vendors would be keen to make. More precisely, it is currently unclear whether (i) central and/or local differential privacy (DP) would be allowed, and (ii) on which type of data (granular or aggregated) model learning would be done. Indeed, one option is to learn under central DP which consists of perturbing, with DP, the output of an aggregation procedure performed by the trusted entity, either before or after model training. As an example, this includes prior-learning aggregation procedures such as outputting noisy contingency matrices (Zhang et al., 2020; Gilotte et al., 2022) or post-learning aggregation as offloading the model training computation to the trusted entity and getting back differentially-private model weights (Chaudhuri et al., 2011). An alternative option is to *implement DP in a decentralized way*: AI systems query raw granular data directly from each user's local data set (*e.g.* stored in browsers) after it has been adequately obfuscated using a layer of local DP (LDP), as depicted in Figure 1.

Operating an AI System under Decentralised LDP. Operating a full-fledged AI system in a decentralized environment under LDP is challenging. Actually, we need to perform the learning of potentially several models (associated with a respective use-case), but also to compare their performances, tune hyper-parameters or perform data analyses in order to assist human decision-making. Addressing each of these tasks individually is challenging under the large level of noise introduced by the LDP constraint. The latter has been the focus of many contributions for the past decade leading to a mature field, but unfortunately,

some important open research questions remain notably regarding how to operate sophisticated AI systems (Xiong et al., 2020; Ye & Hu, 2020; Yang et al., 2023). Indeed, all the downstream tasks that need to be addressed through an AI system might share the same common privacy budget if associated with the same training samples. This notably leads to challenging questions such as (i) **How to learn multiple tasks?** Namely, should we divide the initial raw global data set to perform each task in isolation and thus decrease the amount of noise added to each sample? Should we split the privacy budget? Is it possible to partially or fully learn multiple tasks under the same noisy data? (ii) **How to leverage data structure?** More precisely, how to perform optimally feature selection, dimension reduction, or binning? Is there a frontier between interactive and non-interactive mechanisms? When does learning the data structure give a gain? What are the exploration/exploitation trade-offs? Up to the authors' knowledge, the aforementioned non-exhaustive challenges have not been formulated and tackled yet, notably within a comprehensive paper.

Positioning and Contributions. This position paper aims to fill this gap by providing researchers, practitioners, and the advertising industry with a unified formulation and perspectives on open research challenges associated with private advertising AI systems under local differential privacy constraints. For the sake of clarity, we formulate hereafter our main contributions:

- We provide a unified formulation of the problem of operating an AI system under LDP requirements, which notably encompasses a learning environment that is envisioned, by the advertising industry, to perform private ML model training.
- We perform an extensive literature review leading to a survey of existing works tackling user data querying under LDP. These related works could be loosely speaking divided into two main classes: (i) privacy-preserving mechanisms that can be used for multi-task learning, and (ii) mechanisms striving to identify relevant information in the data.
- We state our main positioning and research guidelines regarding how to operate an AI system under LDP, by pointing out the main limitations of existing works. More specifically, we highlight the importance of conducting research works focusing on multi-task learning under LDP schemes *tailored* to the set of tasks at hand and point out the missing privacy-preserving mechanisms needed to perform data exploration/exploitation efficiently.
- We finally support our positioning through illustrative examples, showing the importance of embracing the research challenges that are pointed out in this paper.

Notation and Conventions. The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|$ and we set $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. For $n \in \mathbb{N}^*$, we refer to $\{1, \dots, n\}$ with the notation $[n]$. $\mathcal{M}_+^1(\mathcal{S})$ denotes the set of probability measures on $\mathcal{S}$.

2 Problem Statement

We consider an environment where the AI system interacts with each user's local data set through a trusted entity implementing an LDP mechanism, see Figure 1. Since this *trusted entity* notion is not required to understand the LDP literature and associated challenges, and might be misleading with respect to the central DP framework, we will omit this layer in the rest of the paper. Instead, we will consider that *the AI system*, willing to get access to some data to train and maintain several ML models, is directly interacting with each user. We detail this learning environment hereafter, in a more formal manner.

User Data. Without loss of generality, we consider a global artificial data set of raw user data $(s_1, \ldots, s_n) \in \mathcal{S}^n$ where $n \in \mathbb{N}^*$. For instance, we could have for any $i \in [n]$, $s_i = (x_i, y_i)$ where $x_i \in \mathcal{X}$ stands for a feature vector and $y_i \in \mathcal{Y}$ is a label. For ease of exposure, each s_i refers to a local user data set and typically stands for a realization of the random variable $S = (X, Y)$ associated with a probability measure μ defined on $\mathcal{S}$.

AI System. The AI system aims at performing several ML-driven tasks, ideally based on the raw samples $\{s_i\}_{i=1}^n$ from μ. To this purpose, it estimates a parameter of this probability distribution $\theta(\mu) \in \mathcal{W}$ where $\theta \in \Theta \subseteq \left(\mathcal{M}_+^1(\mathcal{S}) \to \mathcal{W}\right)$. For instance, $\theta(\cdot)$ can formalize the process of learning a parametric ML model or computing some quantiles of the accuracy of a model, see Example 1 below.

Example 1. Given some loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ and parametric model $f_\omega : \mathcal{X} \to \mathcal{Y}$ where $\omega \in \mathcal{W}$, learning a parametric model is defined as the following risk minimization problem:

$$\theta^*(\mu) = \operatorname*{argmin}_{\omega \in \mathcal{W}} \mathbb{E}_\mu \left[\ell(f_\omega(X), Y)\right] \ . \tag{1}$$

Unfortunately, samples from μ cannot be used to build an estimator for (1), as these samples stand for private user information and cannot be released without considering an obfuscation mechanism. Instead, information from user $i \in \mathbb{N}$ is obtained from observation $Z_i \in \mathcal{Z}$ via a query Q_i. Similarly to Duchi et al. (2013b), Q_i is the conditional distribution of Z_i given the history observed by the agent so far $\mathcal{F}_i = \sigma((Z_i)_{0 \leq l \leq i-1})$ and the *true* user information S_i, that is

$$Z_i \sim Q_i(\ \cdot\ \mid \mathcal{F}_i, S_i = s_i) \ .$$

Note that the output of this query, $Z_i \in \mathcal{Z}$, corresponds to an obfuscated version of S_i as (i) it can refer to a partial view of S_i and (ii) Q_i is a conditional distribution and not a deterministic function since it has to include the randomization mechanism associated to LDP (Dwork et al., 2014), defined in Definition 1. The conditional distribution of $\{Z_i\}_{i=1}^n$ given $\{S_i\}_{i=1}^n$ is called a *mechanism*.

Definition 1. *For $\varepsilon, \delta > 0$, the mechanism Q is (ε, δ)-locally differentially private if for any $i \in \mathbb{N}^*$, for any $z_1, \ldots, z_{i-1} \in \mathcal{Z}$ such that $Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1}$, $s, s' \in \mathcal{S}$ and for any $A \subseteq \mathcal{Z}$ measurable, we have*

$$\frac{Q_i(Z_i \in A \mid \mathcal{F}_i, S_i = s) - \delta}{Q_i(Z_i \in A \mid \mathcal{F}_i, S_i = s')} \leq e^\varepsilon \ .$$

As an example, in the context of local label-DP with binary labels, we have $\mathcal{Z} = \{0, 1\}$ and $\mathcal{S} = \mathcal{Y} = \{0, 1\}$ (Busa-Fekete et al., 2023). A popular choice of query is the binary randomized response (RR) mechanism (Warner, 1965) where for $y_i \in \{0, 1\}$

$$Q_i(Z_i = y_i \mid Y_i = y_i) = \frac{\exp(\varepsilon)}{\exp(\varepsilon) + 1} \ ,$$
$$Q_i(Z_i \neq y_i \mid Y_i = y_i) = \frac{1}{\exp(\varepsilon) + 1} \ .$$

Constraints of a Real-Life AI System. Compared to the task of training one ML model under LDP data, operating and maintaining an AI system is much more involved and yields novel challenges, as detailed in what follows.
Privacy for multiple tasks: Most works on learning under privacy constraints consider that the agent only needs to perform one single task θ, which allows them to tailor querying to the task at hand. Yet, real-life systems often involve tens, if not hundreds of different learned models, and as many, if not more, tasks of model evaluation, comparison, or data analytics. It is therefore necessary to ensure that the queried dataset information can be re-used across multiple tasks, some of which may be unknown at the time of querying.
Interactivity: The large number of internet users and the pace of interaction with them makes it impossible to run a system in a fully synchronous manner. Yet, interactions can often be batched, leaning towards an environment that is not necessarily fully asynchronous. This is formalized by how much the queries are allowed to depend on each other. Two different levels of dependency have been studied in the literature. *Interactive* mechanisms are allowed to have a dependency on the history whereas *non-interactive* mechanisms assume that $Z_i \perp\!\!\!\perp Z_{1:i-1} \mid S_i$ [1].

Section 3 focuses on how to perform multi-task learning under LDP, while Section 4 aims at leveraging some knowledge about the target downstream task to build informed queries yielding better estimation performances.

3 Multi-Task Learning under LDP

In industrial applications, it often happens that multiple tasks should be performed on a data set, be it for model selection or data analytics. Due to the number of tasks the agent may need to perform, splitting the privacy budget allocated to the agent or the samples across the different tasks would result in prohibitively high noise, as illustrated in Figure 2. As such, the agent has to make a compromise between optimizing the querying to reduce the noise incurred to ensure privacy and the re-usability of the answer across many tasks.

[1] A user can only be queried when browsing the internet. *Fully interactive* mechanisms that can query a user several times are therefore unrealistic. These algorithms can be turned into *interactive* mechanisms with a constant factor increase in sample complexity Joseph et al. (2019), are therefore not studied in this work.

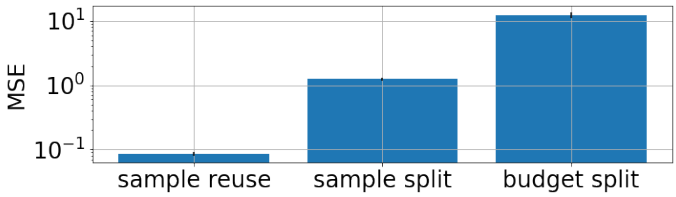

Figure 2: Advantage of re-using samples when performing $m = 10$ estimation tasks with $n = 1000$ and $\epsilon = 1$. One-shot private querying followed by multiple estimations (*sample reuse*) beats naive strategies based on privacy *budget splitting* or *sample splitting*. See Appendix A for details.

Given a set of tasks Θ, a mechanism Q, a loss function $\rho : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ and an estimator $\hat{\theta}$, a possible objective is to solve[2]

$$\min_{Q,\hat{\theta}} \max_{\theta \in \Theta} \mathbb{E}\left[\rho(\theta(\mu), \hat{\theta}(\mu_{Z_{1:n}}))\right] .$$

where $\mu_{Z_{1:n}}$ is the empirical distribution of $(Z_i)_{i \leq n}$, and the minimum is taken over the set of estimators $\hat{\theta}$ and of (ϵ, δ)-LDP mechanisms $\mathcal{Q}$. For instance, the set of tasks could be to estimate the mean $m^\star$ and the first decile $d^\star$ of μ, giving the following objective depending on the mechanism Q and the estimators $\hat{m}, \hat{d}$:

$$\min_{Q,\hat{m},\hat{d}} \max \left(\mathbb{E}\left[\|\hat{m} - m^\star\|_2^2\right], \mathbb{E}\left[\|\hat{d} - d^\star\|_2^2\right]\right) .$$

In this section, we review first learning with privatized data where data are noised via a known mechanism independent of the tasks. The challenge is then to perform all tasks as well as possible despite the noise. Second, we focus on density estimation of μ (independent of Θ) as all tasks can be performed from the knowledge of the density. Lastly, we study some special cases where mechanisms optimized for a set of tasks can be designed.

3.1 Learning with privatised data

The first road towards LDP is to choose Z_i as a noisy version of S_i. For instance, assuming $\mathcal{S} = \mathcal{X} \times \mathcal{Y} \subseteq [0, 1]^d$, a common way to do this is to set $Z_i = S_i + \eta_i$ with $\eta_i \sim \text{Lap}(d/\varepsilon)$. Then, any algorithm working with this database as input would be ε-LDP. Coming back to Example 1, a naïve solution is to perform empirical risk minimization using $Z_{1:n}$, *i.e.* to train the model on noisy data as if they were clean, as done in Yin et al. (2019). It can lead to a consistent estimator of $\theta^\star(\mu)$ in some particular cases when the model is linear Fukuchi et al. (2017); Kang et al. (2020), the noisy label is unbiased and the loss ℓ is a Bregman divergence Badanidiyuru et al. (2023), however, consistency does not

[2]Note that the max over Θ could be replaced by other aggregation operators, such as a weighted sum or an OWA.

hold in general. Note that, depending on the regime of n and d, the variance introduced by the noise may dominate the bias. For instance, it is known (Duchi et al., 2013c, Eq. 30) that the Laplace mechanism is sub-optimal for mean estimation in large dimension.

Many works Butucea et al. (2020); Farokhi (2020); Ju et al. (2022) have underlined the analogy between learning with privatized data and solving inverse problems. In the inverse problem literature Hadamard (1902); Engl et al. (1996); Benning & Burger (2018), the goal is to recover the unknown signal from the knowledge of the obfuscation mechanism and a set of realizations of a given random variable. In particular, if the mechanism is based on additive noise, inverse problems amount to deconvolution, which can be used to perform regression from noisy data Farokhi (2020).

In general, solving the inverse problem analytically is hard. For instance, Natarajan et al. (2013) builds a surrogate loss $\tilde{\ell}$ to be an unbiased estimator of ℓ w.r.t. the label noise by solving the following Fredholm integral equation given a non-interactive mechanism Q identical for all users i:

$$\forall y, \hat{y} \in \mathbb{R}, \quad \ell(y, \hat{y}) = \int_{\mathbb{R}} \tilde{\ell}(u, \hat{y}) \mathrm{d}Q(u|y)\,. \tag{2}$$

Similarly, Farokhi (2022) and Reshetova et al. (2023) (respectively for regression use-cases and generative adversarial networks) modify the loss ℓ into $\tilde{\ell}$ to (approximately) solve (2). Overall, solving the inverse problem for noisy labels is a well-studied topic. Yet, tackling both the label noise and the noise on the features is challenging, except for some specific parametric models such as linear models. The following open question formalizes this problem.

Open Question 1. *When is it possible to find a surrogate loss function acting on the noisy data Z so that its value stands for an unbiased estimator of the loss acting on clean data $S = (X, Y)$? If so, how to (approximately) estimate this surrogate loss function in practice?*

3.2 Density estimation

Another conceptually simple way to perform multiple tasks under LDP is to first build an estimate of the density of measure μ of the true data and then perform all the tasks in Θ on $\tilde{\mu}$. Formally, assuming S (resp. Z) have (potentially discrete) density μ (resp. $\tilde{\mu}$) and that Q has a conditional density q, it amounts to finding an estimate of μ by solving

$$\tilde{\mu}(z) = \int_{\mathcal{S}} q(z|s) \mathrm{d}\mu(s). \tag{3}$$

Density estimation for continuous random variables When S is a continuous random variable and Z is equal to S plus some zero-mean additive

noise, then (3) becomes a *deconvolution* problem. Works in private generative learning Cao et al. (2021); Reshetova et al. (2023) and kernel density estimation Farokhi (2020) learned on privatized data therefore fit in this framework. Here, a connection with the literature on generative modeling with diffusion models can also be highlighted Sohl-Dickstein et al. (2015); Ho et al. (2020); Weng (2021); Strümke & Langseth (2023). Indeed, in a diffusion process, noise (typically a Gaussian one) is sequentially added to data. Hence, the original data distribution (μ) is converted into a new one (here, $\tilde{\mu}$) through gradual addition of noise. The reverse diffusion process consists of recovering the data from its noisy counterpart, and this is exactly the objective of Equation (3).

A large part of the literature on inverse problems Donoho (1995); Mair & Ruymgaart (1996); Natterer (2001); Cohen et al. (2004) is about leveraging sparse decomposition of the signal over fixed or learned basis functions to solve Equation (3). Building upon these works, Butucea et al. (2020) and Duchi et al. (2013a) perform density estimation of continuous random variables by computing a noisy representation of their basis coefficients. Bucketing the features and then applying techniques from discrete density estimation Sun et al. (2019); Xia et al. (2020); Berrett & Butucea (2019); Berrett et al. (2021) is another path towards continuous density estimation. Lastly, in Xu et al. (2019), authors learn a binary representation of the features (privatized by randomized response) via an auto-encoder so that the ℓ_2 norm of the reconstruction error is preserved.

Density estimation for discrete random variables The case of estimating the distribution of a categorical feature is particularly interesting as it has inspired the design of more advanced algorithms. Consider the database $(S_i)_{i=1}^n$ where for all $i \in [n]$, $k \in [d]$, $S_i = k$ with probability μ_k. The goal is to estimate $(\mu_k)_{k\in[d]}$. A popular idea is to use a generalized Randomized Response mechanism Q to ensure local privacy. Let $(Z_i)_{i=1}^n$ be the output of the mechanism, calling $p = Q(Z_i = k|S_i = k)$ and $q = Q(Z_i = k|Z_i \neq k)$, an unbiased estimate of μ_k is given by:

$$\tilde{\mu}_k = \frac{\sum_{i=1}^n \mathbb{1}\{Z_i = k\} - nq}{n(p-q)}.$$

Then the dominant term in the variance of $\tilde{\mu}_k$ scales as $\frac{d-2+\exp(\epsilon))}{(\exp(\epsilon)-1)^2}$ Wang et al. (2017). Erlingsson et al. (2014) introduces the idea of using unary encoding which first encodes the value k as a binary vector corresponding to the k-th vector of the d-dimensional canonical basis , and then to each of them applies a randomized response mechanism parameterized by $Q(Z = 1|S = 1)$ and $Q(Z = 1|S = 0)$. Choosing $Q(Z = 1|S = 1) = Q(Z = 1|S = 0)$ gives Basic Rapport Erlingsson et al. (2014), while optimizing for $Q(Z = 1|S = 1)$ and $Q(Z = 1|S = 0)$ gives optimal unary encoding Wang et al. (2017). In the last case, the variance is $4\frac{\exp(\epsilon)}{(\exp(\epsilon)-1)^2}$, hence independent from d, however it suffers a communication cost of order d, which can be prohibitive. Other variants based on random matrix projections Bassily & Smith (2015); Acharya et al. (2019) or hashing Wang et al. (2017) reach the same variance with a smaller communication cost. Wang et al. (2016) suggests optimizing the mutual information between a uniform

distribution over possible feature value and its transformation via a randomized response mechanism.

Be it for discrete or continuous densities, error bounds from plug-in density estimators are generally looser than those from task-specific mechanisms when tackling a known parametric estimation task, as the problem of density estimation is typically harder than that of parameter estimation. For instance, in Duchi et al. (2013a), locally differentially private estimation of densities in an elliptical Sobolev space with smoothness 1 has mean squared error scaling as $\frac{1}{(\epsilon^2 n)^{\frac{2}{3}}}$. In contrast, in locally differentially private mean estimation Duchi et al. (2013c), the squared error scales as $\frac{1}{\epsilon^2 n}$.

3.3 Towards not being agnostic to the set of tasks

In central DP, the problem of performing multiple tasks is well studied (Dwork et al., 2014, Ch. 5,6). To handle correlated tasks, the idea is to build and update a *fake database* or *synopsis* on which the outcome of queries is close to the one obtained on the true data. Such methods are not as well-studied under LDP. Furthermore, AI systems often need different information than what synopses provide. First, we show that in some cases, multi-task learning can be achieved by computing sufficient statistics or by density estimation. Then, we focus on the case of linear tasks, the only instance of multi-task learning studied in LDP literature.

Learning sufficient statistics / densities Consider first the task of finding the right hyper-parameter in a ridge regression. We want to solve

$$\operatorname*{argmin}_{w \in \mathbb{R}^d} \mathbb{E}[(\langle X, w \rangle - Y)^2] + \lambda \|w\|^2$$

which only depends on μ through $\mathbb{E}[XY]$ and $\mathbb{E}[XX^\top]$. Therefore local differentially private estimates of $\mathbb{E}[XY]$ and $\mathbb{E}[XX^\top]$ allow us to solve the above optimization problem for many values of λ without additional cost. Fukuchi et al. (2017) suggests to obtain these estimates by noising x_i and $y_i x_i$ with an additive Gaussian noise large enough to ensure ϵ-local DP. Unfortunately, they do not obtain utility bounds decreasing with n. The more general task of finding the best binary classifier is solved if the density of $Y|X$ is known. In Berrett & Butucea (2019); Berrett et al. (2021) an estimation of the density $Y|X$ is used to obtain a universally consistent classifier. Overall, however, finding a statistical quantity that is fast to compute and summarizes all tasks is not easy.

The case of linear tasks In local-DP, the framework of linear tasks has been studied by Bassily (2019); Edmonds et al. (2019); McKenna et al. (2020). Consider the database $(S_i)_{i=1}^n$ introduced for discrete density estimation, where for any user $i \in [n]$ and value $k \in [d]$, $S_i = k$ with probability μ_k. A set of m linear tasks is a set of vectors $q_1, \ldots, q_m \in \mathbb{R}^d$ and the goal is to estimate the quantities $\langle q_1, \mu \rangle, \ldots, \langle q_m, \mu \rangle$. In the offline setting, all tasks are available

before the algorithm starts and can be stacked in a matrix $A \in \mathbb{R}^{m \times d}$. The goal is to produce an estimate $\hat{u} \in \mathbb{R}^m$ of $A\mu$. Bassily (2019) shows that when $\forall i \in [m], \|q_i\|_2 \leq r$, querying $A[\cdot, S_i]$ via the Gaussian mechanism reaches (up to constants):

$$\mathbb{E}[\|\hat{u} - A\mu\|_2] \leq r \min\left(\left(\frac{\log(m)\log(n)}{n\epsilon^2}\right)^{\frac{1}{2}}, \frac{d\log(n)}{n\epsilon^2}\right)^{\frac{1}{2}}.$$

While, up to some log factors, this achieves optimality in the minimax sense, better data-dependent bounds can be obtained. For instance, removing tasks that are identical to others should not change the performance of the algorithm. The mechanism in Edmonds et al. (2019) therefore factorizes the matrix of tasks yielding data-dependent bounds depending on the structure of the tasks matrix. More generally McKenna et al. (2020) remarks that most mechanisms with discrete output Bassily & Smith (2015); Acharya et al. (2019); Wang et al. (2016); Warner (1965) can be represented by a matrix $Q \in \mathbb{R}^{m \times n}$ such that $Q[i, j] = P(Z = i|S = j)$. Based on this remark, McKenna et al. (2020) optimized Q to obtain a minimum variance estimator for uniform μ. In the online setting, Bassily (2019) obtains, via a mechanism based on sample splitting, the minimax optimal bound

$$\mathbb{E}[\|\hat{u} - A\mu\|_\infty] \leq r\sqrt{\frac{(\exp(\epsilon)+1)^2 d\log(d)}{n(\exp(\epsilon)-1)^2}}$$

under the assumption $\forall i \in [m], \|q_i\|_2 \leq r$. These tools rely on splitting the privacy budget in the offline setting and splitting samples in the online one. Whether they can be generalized for other types of tasks and other settings is unclear.

Open Question 2. *Given a set of tasks Θ, can we design a mechanism that performs well on Θ?*

What is the exact quantity to optimize? What makes a set of tasks easy or difficult to perform jointly? What if tasks arrive online? Is sample splitting or budget privacy splitting necessary? What are the trade-offs between re-usability, communication costs, and performance? These are follow-up questions that we believe are of interest. Another interesting question is how to distribute the privacy budget among different features and labels, given prior knowledge of their relevance, as efficient budget allocation could enhance the accuracy of collected information. More broadly, in the following section, we argue that understanding the distribution's structure could enable more efficient information querying, while also exploring methods for acquiring such knowledge.

4 Task-Specific LDP Mechanisms for Informed Query Design

The accuracy of the privacy-preserving methods discussed earlier is notably compromised in datasets with high dimensions, a large number of discrete feature values, or broad variable ranges. In this section, we first review works that showcase how gaining information about the data (exploration) can help the learning phase (exploitation). Then, we focus on exploration, highlighting the current advances in prominent exploratory tasks, and tasks for which no satisfactory LDP mechanism is currently known.

4.1 Leveraging Prior Information

We start by emphasizing the benefits of utilizing public data with prior information to design queries. Next, we discuss obtaining this information from private data under LDP.

Using Public Datasets. One research direction has focused on using available prior information to query data more effectively in scenarios where knowledge about the data distribution is accessible to the learner. This information may come from domain-specific knowledge, aggregated census data, historical records, or be provided by "opt-in" users willingly contributing or selling their data. For instance, Ghazi et al. (2021) utilizes prior knowledge of label distribution to refine the label set, enhancing response accuracy for relevant labels. Experiments in Jia & Gong (2019) suggest a calibration method to incorporate prior knowledge to estimate histograms. Bassily et al. (2020a) and Alon et al. (2019) show that combining public and private data significantly improves the performance in terms of effective sample size. More precisely, Alon et al. (2019) investigates the number of private and public samples necessary for choosing a classifier $\widehat{h}$ in a class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ with low error, that is

$$\mathbb{P}(\widehat{h}(X) \neq Y) \leq \inf_{h \in \mathcal{H}} \mathbb{P}(h(X) \neq Y) + \alpha \ .$$

They show that d/α public samples and $d/(\alpha^2 \vee \alpha\epsilon)$ private samples are sufficient, where d is the VC-dimension of $\mathcal{H}$ (by contrast, d/α^2 public samples would be required if no private samples were available). Other approaches include using public data to train a representation of the distribution through a decision tree Ma et al. (2023), to estimate low-dimensional gradients Kairouz et al. (2021); Zhou et al. (2021), or to pre-train a model before fine-tunning it on private data Ganesh et al. (2023).

However, in most situations, apart from well-defined specific cases, the expected performance improvements, the requirements in terms of the necessary volume of public data, and the algorithms to be implemented are still unknown. Moreover, with the exception of Ma et al. (2023), all methods presented above are in the central DP framework.

Using private Data Sets. As public data may not always be accessible, and its distribution might diverge from that of private data Mangold et al. (2023), one can

resort to a two-round approach to query the private dataset. Such an approach has been successfully applied in various problems including classification Ghazi et al. (2021), regression Ghazi et al. (2021), frequency estimation Qin et al. (2016) or clustering Stemmer & Kaplan (2018). It is still unclear in theory whether the best way to implement this two-phase procedure is to use sample splitting or privacy budget splitting or if other strategies than explore-then-commit are possible.

Open Question 3. *How can we leverage prior information from public or private data to help learn in the local DP framework? What are the potential gains, the potential trade-offs?*

4.2 Task-Specific Mechanisms for Complexity Reduction

In the described AI system, striving for an accurate dataset representation across various tasks, the initial focus should be on identifying relevant information and features, as well as the way to leverage them. Among the exploratory tasks of interest, the detection of the most frequent labels (sometimes called the heavy hitters problem) has attracted significant attention in the local DP framework. Various methods address this issue, with some utilizing hash-tables or unary encoding techniques (Bun et al., 2019; Bassily & Smith, 2015; Hsu et al., 2012; Bassily et al., 2020b). An alternative research direction focuses on randomized response-based algorithms, illustrated by methods like LDPMiner (Qin et al., 2016).

By contrast, other prominent exploratory tasks, like feature selection, principal component analysis, and adaptive binning have been mostly studied in the central DP model. Adapting these tasks to the local DP framework through non-interactive mechanisms remains largely unexplored.

Feature Selection. For feature selection, Kifer et al. (2012) suggests revealing through an exponential mechanism a support chosen by a central agent with access to the clean dataset. Alternatively, McKenna & Sheldon (2020) introduces permute-and-flip as an alternative to the exponential mechanism. Another approach from Barrientos et al. (2019) proposes feature selection using a privacy-preserving significance test, while Chu et al. (2023) suggests using decision tree-based importance measures for feature selection.

These methods are tailored to the central DP framework. Feature selection under LDP is harder because of the cost of introducing noise at the individual data level. In the related problem of sparse mean estimation, worst-case error in the non-interactive LDP framework scales as $sd/(\epsilon^2 n)\log(ed/s)$, where s represents the number of non-zero components of the mean, and d denotes the dimension of the feature space Duchi et al. (2018); Acharya et al. (2022) which becomes large if $n\epsilon^2/d \leq 1$. However, recent work Butucea et al. (2023a) has demonstrated that support recovery remains achievable in specific regimes under LDP, provided specific conditions on the magnitude of non-zero entries.

These results pave the way for new valuable research directions in LDP feature selection.

Adaptive Binning. Adaptive binning minimizes noise in the histogram representation of features by seeking a feature space partition where observations are roughly evenly distributed in each bin. Li et al. (2014) reduces this problem to minimizing the cost $\text{pcost}(x, \mathcal{B})$ associated to an histogram dataset $s \in \mathcal{S}^{\mathbb{N}}$ and a partition $\mathcal{B}$ of size k

$$\text{pcost}(x, \mathcal{P}) = \sum_{b \in \mathcal{B}} \sum_{j \in b} \left| s_j - \frac{\sum_{j' \in b} s_{j'}}{|b|} \right| + \frac{k}{\epsilon}$$

where $|b|$ is the length of bin b. Subsequent efforts, like Zhang et al. (2016), propose approximating the optimal partition through recursive feature space partitioning. While adaptive binning is closely linked to the heavy hitter problem, research (including the articles above) has predominantly focused on the central DP framework.

Principal Component Analysis. Principal component analysis (PCA) is crucial for obtaining a low-dimensional representation of a dataset. Yet, most research on private PCA has predominantly focused on the central DP framework. Under central DP, the covariance matrix is computed with the trusted data-holder adding noise, either directly to the covariance matrix (Blum et al., 2005) or its rank-k approximation (Cai et al., 2024). The noisy matrix is then transmitted to the learner, who applies singular value decomposition. An alternative approach estimates the subspace with the leading k eigenvectors and transmits it to the learner using an exponential mechanism Chaudhuri et al. (2012). Other strategies, like those in Ge et al. (2018) and Liu et al. (2022), address PCA in a distributed optimization framework, where each data holder applies an independent privacy-preserving mechanism to their batch of data. In the LDP framework, Wang & Xu (2019a) introduces a novel approach adding (Gaussian) noise to individual covariance matrices before transmission, estimating PCA based on the average matrix received by the learner. The authors also tackle sparse PCA, presenting a local interactive algorithm that outperforms their local, non-interactive mechanism, prompting consideration of the trade-off associated with non-interactivity.

> **Open Question 4.** *What are the regimes where feature selection, adaptive binning, and PCA are possible in the LDP framework? What are the optimal algorithms for these tasks?*

Trade-Offs Arising from Non-Interactivity. Given the added complexity and cost associated with interactive mechanisms, we need to determine when non-interactive mechanisms are suitable and assess whether the benefits of interactive mechanisms outweigh their costs. The question of whether interactive mechanisms can surpass non-interactive ones in an LDP framework is complex. Recent studies have identified scenarios where non-interactive mechanisms

achieve minimax-optimal performance, such as in classification with Hölder-continuous score functions (Berrett & Butucea, 2019), density estimation over Besov ellipsoids (Butucea et al., 2020), or one-dimensional mean estimation (Duchi et al., 2018). Yet, other research indicates that non-interactive mechanisms may be sub-optimal compared to interactive ones in certain settings. For example, Butucea et al. (2023b) and Butucea & Issartel (2021) establish this for the estimation of (linear functionals of) discrete distributions. Similarly, Berrett & Butucea (2020) demonstrates that interaction is necessary for minimax optimal estimation of the quadratic functional $D(f) = \int_0^1 f^2(x)dx$ of a density f. In the context of high-dimensional sparse mean estimation, interaction helps improve the rate for estimating s-sparse mean in dimension d from $\frac{sd}{n\epsilon^2}\log(\frac{ed}{s})$ to $\frac{sd}{n\epsilon^2}$, as shown by Acharya et al. (2022). Interestingly, in these examples, optimal rates can be obtained with just two rounds of queries.

Apart from these few examples, the potential gain from interactive mechanisms is not well understood. For instance, many algorithms for convex loss minimization in the LDP framework are variants of gradient descent algorithms, and each iteration requires interacting with the users (Duchi et al., 2018; Wang & Xu, 2019b). In multi-dimensional settings, it has been argued that the sample size of non-interactive algorithms should grow exponentially with the dimension, while interactive mechanisms achieve polynomial dependence Smith et al. (2017). However, this result only holds for the class of algorithms relying on neighborhood-based oracle, such as gradient descent. It is unclear whether different algorithms could achieve better convergence rates.

Open Question 5. *Can exploratory tasks be performed using solely non-interactive mechanisms, and if yes, does this result in a loss of accuracy?*

5 Conclusion

In this position paper, we shed light on the technical challenges of local differential privacy (LDP), a proposed solution for maintaining AI systems while safeguarding user privacy. For the computational advertising industry, in an era marked by growing privacy concerns and the ongoing deprecation of third-party cookies, addressing the challenges associated with operating a full-fledged AI system under LDP becomes crucial. Specifically, we advocate for developing mechanisms that allow data reusability across multiple tasks and facilitate the estimation and leverage of data structure. These research directions are rarely explored in LDP and raise numerous questions. How can we design LDP mechanisms optimized for a set of tasks? How can we train algorithms on noisy features and labels? How can we perform data exploration optimally, and can we restrict ourselves to non-interactive mechanisms? How can we incorporate prior information into LDP mechanisms? Answering these questions is essential for the development of AI systems that respect the privacy of their users.

Impact Statement

Nowadays, major internet services, which support the open internet economy, are based on learning from users behavioral data: recommender systems, online marketing... In the last years, browser vendors have been working on improving the users privacy, especially by deprecating the use of third-party cookies. The question of how to replace the function of third-party cookies, which are/were the backbone of data collection to learn AI models, while preserving the users privacy is a key challenge. In particular, the major browser vendor, Google Chrome, is currently in the process of testing possible implementations of APIs aimed at enabling computing statistics and learning AI models in a privacy-preserving way, with a final implementation planned for 2026.

This paper aims at highlighting, to the AI community, directions of research that would enable better practical compromises between the performance of AI systems (that support the internet economy) and the users privacy online. Providing answers to such questions could have a strong impact as the practical design of privacy mechanisms in browsers is under (very) active development.

References

Acharya, J., Sun, Z., and Zhang, H. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1120–1129. PMLR, 2019.

Acharya, J., Canonne, C. L., Sun, Z., and Tyagi, H. The role of interactivity in structured estimation. In *35th Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1328–1355, 2022. URL `https://arxiv.org/abs/2203.06870`.

Agarwal, A., Chapelle, O., Dudík, M., and Langford, J. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15(32): 1111–1133, 2014. URL `http://jmlr.org/papers/v15/agarwal14a.html`.

Aksu, H., Ghazi, B., Kamath, P., Kumar, R., Manurangsi, P., Sealfon, A., and Varadarajan, A. V. Summary reports optimization in the privacy sandbox attribution reporting api. *arXiv preprint arXiv:2311.13586*, 2023.

Alon, N., Bassily, R., and Moran, S. Limits of private learning with access to public data. *Advances in neural information processing systems*, 32, 2019.

Badanidiyuru, A., Ghazi, B., Kamath, P., Kumar, R., Leeman, E., Manurangsi, P., Varadarajan, A. V., and Zhang, C. Optimal unbiased randomizers for regression with label differential privacy. *arXiv preprint arXiv:2312.05659*, 2023.

Barrientos, A. F., Reiter, J. P., Machanavajjhala, A., and Chen, Y. Differentially private significance tests for regression coefficients. *Journal of Computational*

and Graphical Statistics, 28(2):440–453, 2019. doi: 10.1080/10618600.2018.1538881. URL https://doi.org/10.1080/10618600.2018.1538881.

Bassily, R. Linear queries estimation with local differential privacy. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 721–729. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/bassily19a.html.

Bassily, R. and Smith, A. Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 127–135, 2015.

Bassily, R., Cheu, A., Moran, S., Nikolov, A., Ullman, J., and Wu, S. Private query release assisted by public data. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 695–703. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/bassily20a.html.

Bassily, R., Nissim, K., Stemmer, U., and Thakurta, A. Practical locally private heavy hitters. *Journal of Machine Learning Research*, 21(16):1–42, 2020b. URL http://jmlr.org/papers/v21/18-786.html.

Benning, M. and Burger, M. Modern regularization methods for inverse problems. *Acta numerica*, 27:1–111, 2018.

Berrett, T. and Butucea, C. Classification under local differential privacy. *Annales de l'ISUP*, 63(2-3):191–204, 2019. URL https://hal.science/hal-03603855. 12 pages.

Berrett, T. and Butucea, C. Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. *Advances in Neural Information Processing Systems*, 33:3164–3173, 2020.

Berrett, T. B., Györfi, L., and Walk, H. Strongly universally consistent nonparametric regression and classification with privatised data. *Electronic Journal of Statistics*, 15:2430–2453, 2021.

Blum, A., Dwork, C., McSherry, F., and Nissim, K. Practical privacy: the sulq framework. In *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pp. 128–138, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930620. doi: 10.1145/1065167.1065184. URL https://doi.org/10.1145/1065167.1065184.

Bun, M., Nelson, J., and Stemmer, U. Heavy hitters and the structure of local privacy. *ACM Trans. Algorithms*, 15(4), oct 2019. ISSN 1549-6325. doi: 10.1145/3344722. URL https://doi.org/10.1145/3344722.

Busa-Fekete, R. I., Medina, A. M., Syed, U., and Vassilvitskii, S. Label differential privacy and private training data release. In *International Conference on Machine Learning*, pp. 3233–3251. PMLR, 2023.

Butucea, C. and Issartel, Y. Locally differentially private estimation of functionals of discrete distributions. *Advances in Neural Information Processing Systems*, 34:24753–24764, 2021.

Butucea, C., Dubois, A., Kroll, M., and Saumard, A. Local differential privacy: Elbow effect in optimal density estimation and adaptation over besov ellipsoids. *Bernoulli*, 26(3):1727–1764, 2020.

Butucea, C., Dubois, A., and Saumard, A. Phase transitions for support recovery under local differential privacy. *Mathematical Statistics and Learning*, 2023a.

Butucea, C., Rohde, A., and Steinberger, L. Interactive versus noninteractive locally differentially private estimation: Two elbows for the quadratic functional. *The Annals of Statistics*, 51(2):464–486, 2023b.

Cai, T. T., Xia, D., and Zha, M. Optimal differentially private pca and estimation for spiked covariance matrices. *arXiv preprint arXiv:2401.03820*, 2024.

Cao, T., Bie, A., Vahdat, A., Fidler, S., and Kreis, K. Don't generate me: Training differentially private generative models with sinkhorn divergence. *Advances in Neural Information Processing Systems*, 34:12480–12492, 2021.

Carey, C., Dick, T., Epasto, A., Javanmard, A., Karlin, J., Kumar, S., Muñoz Medina, A., Mirrokni, V., Nunes, G. H., Vassilvitskii, S., and Zhong, P. Measuring re-identification risk. *Proc. ACM Manag. Data*, 1(2), jun 2023. doi: 10.1145/3589294. URL https://doi.org/10.1145/3589294.

Chapelle, O. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 1097–1105, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623634. URL https://doi.org/10.1145/2623330.2623634.

Chapelle, O., Manavoglu, E., and Rosales, R. Simple and scalable response prediction for display advertising. *ACM Trans. Intell. Syst. Technol.*, 5(4), dec 2015. ISSN 2157-6904. doi: 10.1145/2532128. URL https://doi.org/10.1145/2532128.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011. URL http://jmlr.org/papers/v12/chaudhuri11a.html.

Chaudhuri, K., Sarwate, A., and Sinha, K. Near-optimal differentially private principal components. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

URL https://proceedings.neurips.cc/paper_files/paper/2012/file/f770b62bc8f42a0b66751fe636fc6eb0-Paper.pdf.

Chu, Z., He, J., Zhang, X., Zhang, X., and Zhu, N. Differential privacy high-dimensional data publishing based on feature selection and clustering. *Electronics*, 12(9), 2023. ISSN 2079-9292. doi: 10.3390/electronics12091959. URL https://www.mdpi.com/2079-9292/12/9/1959.

Cohen, A., Hoffmann, M., and Reiss, M. Adaptive wavelet galerkin methods for linear inverse problems. *SIAM Journal on Numerical Analysis*, 42(4): 1479–1501, 2004.

Donoho, D. L. Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Applied and computational harmonic analysis*, 2(2):101–126, 1995.

Duchi, J., Wainwright, M. J., and Jordan, M. I. Local privacy and minimax bounds: Sharp rates for probability estimation. *Advances in Neural Information Processing Systems*, 26, 2013a.

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, 2013b.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy, data processing inequalities, and statistical minimax rates. *arXiv preprint arXiv:1302.3203*, 2013c.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018. doi: 10.1080/01621459.2017.1389735. URL https://doi.org/10.1080/01621459.2017.1389735.

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.

Edmonds, A., Nikolov, A., and Ullman, J. The power of factorization mechanisms in local and central differential privacy. *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2019. URL https://api.semanticscholar.org/CorpusID:208158365.

Engl, H. W., Hanke, M., and Neubauer, A. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

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.

Farokhi, F. Deconvoluting kernel density estimation and regression for locally differentially private data. *Scientific Reports*, 10(1):21361, 2020.

Farokhi, F. Distributionally-robust machine learning using locally differentially-private data. *Optimization Letters*, 16(4):1167–1179, 2022.

Fukuchi, K., Tran, Q. K., and Sakuma, J. Differentially Private Empirical Risk Minimization with Input Perturbation, October 2017. URL `http://arxiv.org/abs/1710.07425`. arXiv:1710.07425 [cs, stat].

Ganesh, A., Haghifam, M., Nasr, M., Oh, S., Steinke, T., Thakkar, O., Thakurta, A. G., and Wang, L. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, pp. 10611–10627. PMLR, 2023.

Ge, J., Wang, Z., Wang, M., and Liu, H. Minimax-optimal privacy-preserving sparse pca in distributed systems. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1589–1598. PMLR, 09–11 Apr 2018. URL `https://proceedings.mlr.press/v84/ge18a.html`.

Ghazi, B., Golowich, N., Kumar, R., Manurangsi, P., and Zhang, C. Deep learning with label differential privacy. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=RYcgfqmAOHh`.

Gilotte, A., Yahmed, A. B., and Rohde, D. Learning from aggregated data with a maximum entropy model. *arXiv preprint arXiv:2210.02450*, 2022.

Hadamard, J. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton university bulletin*, pp. 49–52, 1902.

Hao, B., Lattimore, T., and Szepesvari, C. Adaptive exploration in linear contextual bandit. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3536–3545. PMLR, 26–28 Aug 2020. URL `https://proceedings.mlr.press/v108/hao20b.html`.

He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., Shi, Y., Atallah, A., Herbrich, R., Bowers, S., and Candela, J. Q. n. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, ADKDD'14, pp. 1–9, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329996. doi: 10.1145/2648584.2648589. URL `https://doi.org/10.1145/2648584.2648589`.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hsu, J., Khanna, S., and Roth, A. Distributed private heavy hitters. In *Automata, Languages, and Programming: 39th International Colloquium, ICALP 2012, Warwick, UK, July 9-13, 2012, Proceedings, Part I 39*, pp. 461–472. Springer, 2012.

Jia, J. and Gong, N. Z. Calibrate: Frequency estimation and heavy hitter identification with local differential privacy via incorporating prior knowledge. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 2008–2016. IEEE, 2019.

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.

Ju, N., Awan, J., Gong, R., and Rao, V. Data augmentation mcmc for bayesian inference from privatized data. *Advances in neural information processing systems*, 35:12732–12743, 2022.

Kairouz, P., Diaz, M. R., Rush, K., and Thakurta, A. (Nearly) Dimension Independent Private ERM with AdaGrad Rates\{via Publicly Estimated Subspaces. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 2717–2746. PMLR, July 2021. URL `https://proceedings.mlr.press/v134/kairouz21a.html`. ISSN: 2640-3498.

Kang, Y., Liu, Y., Niu, B., Tong, X., Zhang, L., and Wang, W. Input Perturbation: A New Paradigm between Central and Local Differential Privacy, February 2020. URL `http://arxiv.org/abs/2002.08570`. arXiv:2002.08570 [cs, stat].

Khalid, S., Khalil, T., and Nasreen, S. A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pp. 372–378, 2014. doi: 10.1109/SAI.2014.6918213.

Kifer, D., Smith, A., and Thakurta, A. Private Convex Empirical Risk Minimization and High-dimensional Regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, pp. 25.1–25.40. JMLR Workshop and Conference Proceedings, June 2012. URL `https://proceedings.mlr.press/v23/kifer12.html`. ISSN: 1938-7228.

Li, C., Hay, M., Miklau, G., and Wang, Y. A data- and workload-aware algorithm for range queries under differential privacy. *Proc. VLDB Endow.*, 7 (5):341–352, jan 2014. ISSN 2150-8097. doi: 10.14778/2732269.2732271. URL `https://doi.org/10.14778/2732269.2732271`.

Liu, X., Kong, W., Jain, P., and Oh, S. Dp-pca: Statistically optimal and differentially private pca. *Advances in Neural Information Processing Systems*, 35:29929–29943, 2022.

Ma, Y., Zhang, H., Cai, Y., and Yang, H. Decision tree for locally private estimation with public data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=F5FVsfCxt8.

Mair, B. A. and Ruymgaart, F. H. Statistical inverse estimation in hilbert scales. *SIAM Journal on Applied Mathematics*, 56(5):1424–1444, 1996.

Mangold, P., Bellet, A., Salmon, J., and Tommasi, M. High-dimensional private empirical risk minimization by greedy coordinate descent. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4894–4916. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/mangold23a.html.

McKenna, R. and Sheldon, D. R. Permute-and-flip: A new mechanism for differentially private selection. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 193–203. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/01e00f2f4bfcbb7505cb641066f2859b-Paper.pdf.

McKenna, R., Maity, R. K., Mazumdar, A., and Miklau, G. A workload-adaptive mechanism for linear queries under local differential privacy. *Proc. VLDB Endow.*, 13(12):1905–1918, jul 2020. ISSN 2150-8097. doi: 10.14778/3407790.3407798. URL https://doi.org/10.14778/3407790.3407798.

McMahan, H. B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D., Chikkerur, S., Liu, D., Wattenberg, M., Hrafnkelsson, A. M., Boulos, T., and Kubica, J. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 1222–1230, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321747. doi: 10.1145/2487575.2488200. URL https://doi.org/10.1145/2487575.2488200.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Natterer, F. *The mathematics of computerized tomography*. SIAM, 2001.

Qin, Z., Yang, Y., Yu, T., Khalil, I. M., Xiao, X., and Ren, K. Heavy hitter estimation over set-valued data with local differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016. URL https://api.semanticscholar.org/CorpusID:3890155.

Reshetova, D., Chen, W.-N., and Özgür, A. Training generative models from privatized data. *arXiv preprint arXiv:2306.09547*, 2023.

Smith, A., Thakurta, A., and Upadhyay, J. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 58–77. IEEE, 2017.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Stemmer, U. and Kaplan, H. Differentially private k-means with constant multiplicative error. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/32b991e5d77ad140559ffb95522992d0-Paper.pdf.

Strümke, I. and Langseth, H. Lecture notes in probabilistic diffusion models. *arXiv preprint arXiv:2312.10393*, 2023.

Sun, L., Zhao, J., and Ye, X. Distributed clustering in the anonymized space with local differential privacy. *arXiv preprint arXiv:1906.11441*, 2019.

Wang, D. and Xu, J. Principal component analysis in the local differential privacy model. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 4795–4801. International Joint Conferences on Artificial Intelligence Organization, 7 2019a. doi: 10.24963/ijcai.2019/666. URL https://doi.org/10.24963/ijcai.2019/666.

Wang, D. and Xu, J. On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*, pp. 6628–6637. PMLR, 2019b.

Wang, S., Huang, L., Wang, P., Nie, Y., Xu, H., Yang, W., Li, X.-Y., and Qiao, C. Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*, 2016.

Wang, T., Blocki, J., Li, N., and Jha, S. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pp. 729–745, Vancouver, BC, August 2017. USENIX Association. ISBN 978-1-931971-40-9. URL https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, pp. 63–69, 1965.

Weng, L. What are diffusion models? *lilianweng.github.io*, Jul 2021. URL https://lilianweng.github.io/posts/2021-07-11-diffusion-models/.

Xia, C., Hua, J., Tong, W., and Zhong, S. Distributed k-means clustering guaranteeing local differential privacy. *Computers & Security*, 90:101699, 2020.

Xiong, X., Liu, S., Li, D., Cai, Z., and Niu, X. A comprehensive survey on local differential privacy. *Security and Communication Networks*, 2020:1–29, 2020.

Xu, C., Ren, J., She, L., Zhang, Y., Qin, Z., and Ren, K. Edgesanitizer: Locally differentially private deep inference at the edge for mobile data analytics. *IEEE Internet of Things Journal*, 6(3):5140–5151, 2019.

Yang, M., Guo, T., Zhu, T., Tjuawinata, I., Zhao, J., and Lam, K.-Y. Local differential privacy and its applications: A comprehensive survey. *Computer Standards & Interfaces*, pp. 103827, 2023.

Ye, Q. and Hu, H. Local differential privacy: Tools, challenges, and opportunities. In *International conference on web information systems engineering*, pp. 13–23. Springer, 2020.

Yin, C., Zhou, B., Yin, Z., and Wang, J. Local privacy protection classification based on human-centric computing. *Human-centric Computing and Information Sciences*, 9:1–14, 2019.

Yuan, Y., Wang, F., Li, J., and Qin, R. A survey on real time bidding advertising. In *Proceedings of 2014 IEEE International Conference on Service Operations and Logistics, and Informatics*, pp. 418–423, 2014. doi: 10.1109/SOLI.2014.6960761.

Zhang, J., Xiao, X., and Xie, X. Privtree: A differentially private algorithm for hierarchical decompositions. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pp. 155–170, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450335317. doi: 10.1145/2882903.2882928. URL https://doi.org/10.1145/2882903.2882928.

Zhang, Y., Charoenphakdee, N., Wu, Z., and Sugiyama, M. Learning from aggregate observations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7993–8005. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5b0fa0e4c041548bb6289e15d865a696-Paper.pdf.

Zhou, Y., Wu, S., and Banerjee, A. Bypassing the ambient dimension: Private {sgd} with gradient subspace identification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=7dpmlkBuJFC.

Zucker-Scharff, A. Design Dimensions for Private Measurement Technical Specification MVP. https://github.com/patcg/docs-and-reports/blob/main/design-dimensions/Dimensions-with-General-Agreement.md#privacy-defined-at-least-by-differential-privacy, 2023. [Online].

A Experiments Details

We provide hereafter details on the illustrative experiments setup and additional numerical results.

A.1 Advantage of reusing samples when performing multiple estimation tasks on LDP data (Synthetic)

Data To illustrate the case of multiple correlated tasks under LDP we consider the following setting. Data is generated as $X \in \mathcal{M}_{n,d} \sim Ber(\mu)$ s.t. $n = 1000, d = 100, \mu \in \mathbb{R}^d \sim \beta(1,1)^d$. Estimation tasks are defined as $\{\theta_i = \mathbb{E}_X[X^T w_i]\}_{i=1\dots m}$ with $m = 10$, $w_i \sim Ber(1/2)^d$. Privacy level is controlled by LDP budget ϵ.

Methods *sample reuse* performs one-shot private querying followed by multiple estimations on the obtained private sample. *budget split* performs m parallel, private queries using ϵ/m budget each and estimates each θ_i separately. *sample split* performs m parallel, private queries asking $n_i = n/m$ samples with ϵ budget each and estimates each θ_i separately.

Metrics Estimation quality is computed using mean square error (MSE) of the private mean estimation versus the public (non noisy) estimation. Mean MSE and confidence intervals are obtained with 100 bootstrap resamples and shown on figure at the $\alpha = .1$ level.

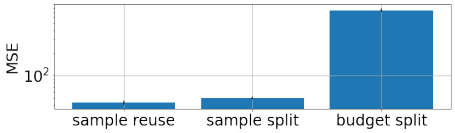

Figure 3: $\epsilon = .1$

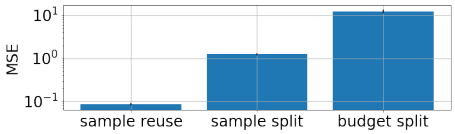

Figure 4: $\epsilon = 1$

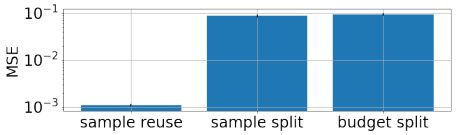

Figure 5: $\epsilon = 10$