Privacy in Metalearning and Multitask Learning: Modeling and Separations

Maryam Aliakbarpour
Department of Computer Science,
Ken Kennedy Institute
Rice University

Konstantina Bairaktari
Khoury College of Computer Sciences,
Northeastern University

Adam Smith
Department of Computer Science,
Boston University

Marika Swanberg
Google

Jonathan Ullman
Khoury College of Computer Sciences,
Northeastern University

Abstract

Model personalization allows a set of individuals, each facing a different learning task, to train models that are more accurate for each person than those they could develop individually. The goals of personalization are captured in a variety of formal frameworks, such as multitask learning and metalearning. Combining data for model personalization poses risks for privacy because the output of an individual's model can depend on the data of other individuals. In this work we undertake a systematic study of differentially private personalized learning. Our first main contribution is to construct a taxonomy of formal frameworks for private personalized learning. This taxonomy captures different formal frameworks for learning as well as different threat models for the attacker. Our second main contribution is to prove separations between the personalized learning problems corresponding to different choices. In particular, we prove a novel separation between private multitask learning and private metalearning.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

Model personalization allows a set of individuals, each facing a different learning task, to train models that are more accurate for each person than those they could develop individually. For example, consider a set of people, each of whom holds a relatively small dataset of photographs labeled with the names of their loved ones in the picture. Each person would like to build a classifier that labels future pictures with the names of people in the picture, but training such an image classifier would take more data than any individual person has. Even though the tasks they want to carry out are different—their photos have different subjects—those tasks share a lot of common structure. By pooling their data, a large group of people could learn the shared components of a good set of classifiers. Each individual could then train the subject-specific components on their own, requiring only a few examples for each subject. Other applications of personalization include next-word prediction on a mobile keyboard, speech recognition, and recommendation systems.

The goals of personalization are captured in a variety of formal frameworks, such as multitask learning and metalearning. Roughly, in *multitask learning* we are given data for t tasks and wish to find t different, but related, models $g_1, \dots, g_t$ that each perform well on one task. In *metalearning*, we aim to find a common *representation*, which is a summary (e.g. an embedding of the data into a lower-dimensional space) that can be adapted to new test tasks that are similar to the training tasks, and where similarity is formalized through some task distribution.

Although it offers many benefits, model personalization poses new risks for privacy because each individual's model depends on the data of other individuals. Thus, we undertake a systematic study of *differentially private (DP) personalized learning*. Informally, DP personalized learning requires that the components of the models made visible to *people other than you* do not reveal too much information about *your data*. As in the motivating example above, we think of your data as being synonymous with the training data for a single learning task.

Frameworks for private personalization. Our first contribution is a taxonomy of different formal frameworks for private personalized learning, which vary according to the learning objective, the privacy requirements, and the structure of the output.

Our main interest in investigating these models is sample complexity: How many individuals/tasks t need to contribute data to achieve a given learning objective, and how many samples n do we need from each individual/task? We focus on small n, so no one individual data can learn well on their own.

Because we focus on sample complexity (as opposed to computation or communication), we consider a centralized curator that can collect and process the data of all parties. Such a curator can typically be simulated by a secure multiparty computation without the need for an actual trusted party. Specifically, we consider the following output structures (omitting citations—see the more detailed discussion and related work sections below):

> *Billboard*: The outputs of the curator are visible to all individuals ("written on a billboard"), but may be adapted locally by each individual for their use. This is the model used in the overwhelming majority of papers on private personalized learning.
>
> *Separate Outputs*: The curator sends a separate output to each individual, that is only visible to them.

Each output structure leads to one or more corresponding privacy requirements. In the billboard structure, it is clear that the DP condition has to apply to everything the curator outputs. When individuals receive separate outputs, we consider two types of attackers. One type sees the outputs of *all but one* individual and we want this attacker to be unable to make inferences about the data of the remaining individual (*joint DP*). In the other (1-out-of-t DP), we require privacy only for an individual that sees *one* output.

Relating the frameworks. Our second contribution is a set of relationships among and separations between the different combinations of learning objectives and privacy requirements.

We introduce two basic problems, one involving estimation and one involving classification, and show a hierarchy of frameworks for both problems:

metalearning DP = multitask DP billboard $\subsetneq$ multitask JDP $\subsetneq$ multitask 1-out-of-t DP

Our results have a number of technical and conceptual implications for private personalized learning:

1. Multitask learning with billboard DP implies DP metalearning and vice-versa.
2. For some tasks, multitask learning with a general JDP algorithm can require much less data than for DP billboard algorithms. Thus, while most JDP learning algorithms to date use a billboard, there is potential to significantly reduce the sample complexity by exploring different algorithmic paradigms. As a consequence, there is a separation between DP multitask and DP metalearning, even though the two objectives are essentially equivalent in the absence of privacy constraints.
3. There is a separation between private multitask learning with JDP and with 1-out-of-t DP. Thus, private personalized learning requires much less data if we are in a setting where a realistic adversary can only see the outputs or behavior of a small number of individuals.

1.1 Problem Formulation and Results

In this section we describe the key concepts in our taxonomy and then describe our technical results.

Learning Objectives. We consider two learning objectives—*multitask learning* and *metalearning*. In both, data from t people is pooled together for learning. In multitask learning we simply want to learn a separate model for each of these tasks, but in metalearning, we want to extract some kind of representation that can later be specialized to another unseen task. Note that metalearning is intuitively the harder objective because the representation can also be specialized to the original tasks.

In *multitask learning*, we have t tasks, each of which is modeled by a distribution P_i, with training data for the task sampled i.i.d. from this distribution. The algorithm takes in training datasets S_i for each task and returns hypotheses g_i for each task. Our goal is for the g_i to have low error on their respective tasks, as measured by some loss function $\ell(P_i, g_i)$, averaged over the t tasks.

Definition 1.1 (Multitask Learning)**.** Let $\mathcal{P}$ be a set of t-tuples of distributions and ℓ be a loss function. An algorithm $\mathcal{M}$ *multitask learns $\mathcal{P}$ with error α with t tasks and n samples per task* if, when given datasets $S_1 \sim P_1^n, \ldots, S_t \sim P_t^n$ for any $(P_1, \ldots, P_t) \in \mathcal{P}$, it returns hypotheses $g_1, \ldots, g_t$ such that $\mathbb{E}[\frac{1}{t}\sum_{i\in[t]} \ell(P_i, g_i)] \leq \alpha$, where the expectation is taken over the datasets and the randomness of $\mathcal{M}$.

Note that the definition doesn't make assumptions about the individual tasks or relationships between them, but some assumptions will be necessary for any non-trivial multitask learning.

The other learning objective we consider is *metalearning*, which models learning a *representation* that can be used for some unseen task. Here, we have a collection of *t training tasks* and a separate *test task* drawn from a task distribution. The first algorithm takes training datasets S_i for each of the training tasks and outputs a representation[1] h. The second algorithm then takes h and training data for the unseen test task and returns a hypothesis g. Our goal is for the final hypothesis g to have low error on the test task.

Definition 1.2 (Metalearning)**.** Let $\mathcal{Q}$ be a distribution over $(t+1)$-tuples of distributions and ℓ be a loss function. A pair of algorithms $(\mathcal{M}_{\text{meta}}, \mathcal{M}_{\text{pers}})$ *metalearns $\mathcal{Q}$ with error α using t training tasks, n samples per training task, and a test task with n_{pers} personalization samples* if the following holds: Let $(P_1, \ldots, P_{t+1})$ be a tuple of tasks drawn from $\mathcal{Q}$. Let $S_i \sim P_i^n$ for $i \in [t]$ and give $S_1, \ldots, S_t$ to $\mathcal{M}_{\text{meta}}$ to obtain a representation h. Now give h and $S_{t+1} \sim P_{t+1}^{n_{\text{pers}}}$ to $\mathcal{M}_{\text{pers}}$ to obtain a hypothesis g_{t+1}. Then $\mathbb{E}[\ell(P_{t+1}, g_{t+1})] \leq \alpha$, where the expectation is taking over the choice of tasks, the training datasets, the test dataset, and the randomness of the algorithms.

Privacy Requirements for Personalized Models. We consider a range of privacy requirements based on *differential privacy*, which we summarize here informally (see §2 for precise definitions). Roughly, an algorithm $\mathcal{M}$ is differentially private if the distribution of its output is insensitive to changing one individual's data. In our context, this means that if we have $S = (S_1, \ldots, S_t)$, where each S_i is all the training data of a specific individual/task, and S' which differs on the data for one individual, then $\mathcal{M}(S)$ and $\mathcal{M}(S')$ have nearly the same distribution.

Our privacy requirements vary according to what we assume to be visible to an attacker. Some of these requirements are specific to one of the two learning objectives above, or assume particular structural constraints on the algorithms.

The privacy requirement for metalearning is easiest to describe, because we assume that the representation h is published, and thus require that h does not reveal too much about any of the individuals' data. Formally, this means simply that algorithm $\mathcal{M}_{\text{meta}}$ is differentially private.

Multitask learning offers a richer space of possible privacy requirements, because the learning framework allows for each person i to receive a different output g_i. In this context, the natural privacy requirement is *joint differential privacy (JDP)*. Here, we imagine that each individual i is given only their own model g_i. For any non-trivial learning, g_i must depend on individual i's dataset S_i. However, we do not want an attacker who can observe other players' outputs to learn about individual I's data set. So, we require that for every individual i, the collection of $t-1$ models g_{-i} is differentially private as a function of the dataset S_i belonging to individual i.

The above definition requires that S_i is protected even if every other individual colludes and combines their models. We can also consider a relaxation of this definition called *1-out-of-t differential privacy* where we do not allow individual's to collude. Here, we require that for every individual i, and for every other individual $i' \neq i$, the model $g_{i'}$ given to individual i' is differentially private as a function of the dataset S_i belonging to individual i.

The majority of algorithms that satisfy joint differential privacy have a particular form, called a *billboard algorithm*. This concept is general, but we will describe it in the context of multitask learning. In a billboard algorithm, we decompose $\mathcal{M}$ into two phases $\mathcal{M}_{\text{BB}}$ and $\mathcal{M}_{\text{pers}}$. $\mathcal{M}_{\text{BB}}$ takes the training data $S_1, \ldots, S_t$ and outputs a representation h, and then

[1] We do not constrain the form of h in any way, and use "representation" to mean any summary of the training data that can be specialized to future tasks. In fact, if privacy were not a concern, the representation could consist of raw training data.

for each individual i, we give them (or they compute for themselves) the model $g_i = \mathcal{M}_{\text{pers}}(S_i, h)$. So far we have described billboard algorithms as a constraint on the *structure* of the algorithm. However, conceptually, we think of the representation h as being published, and thus publicly available, while the individual models g_i are computed secretly by each individual and not published. Thus, when defining privacy for a billboard algorithm, we require that the algorithm $\mathcal{M}_{\text{BB}}$ be differentially private. Any private billboard algorithm also satisfies joint DP.

Intuitively, there is a hierarchy of frameworks for private personalized learning: the easiest is multitask learning with 1-out-of-t-DP, then multitask learning with joint DP, then multitask learning with a private billboard algorithm, and finally the hardest is metalearning.

Technical Contributions. We prove a set of relationships among these frameworks for private personalized learning, which we now summarize.

DP metalearning and DP billboard multitask learning are equivalent. While metalearning can be much harder than multitask learning in general, we show that private multitask learning with billboard algorithms actually implies metalearning. In other words, we recover (a version of) the nonprivate equivalence of Aliakbarpour et al. (2024) when the multitask learner is constrained to produce a publicly visible representation. However, the proof of the implication is quite different. To gain some intuition for the argument, first observe that the syntax of algorithms for the two settings is similar: in both cases, we look at all the training datasets to produce a private representation h and then use an individual's data to specialize h to a model in an arbitrary non-private way. The difference lies in which learning objective we expect this representation to satisfy. We prove the implication using the connection between differential privacy and generalization (Dwork et al., 2014; Bassily et al., 2016; Ligett et al., 2017; Jung et al., 2020). Intuitively, in multitask learning, the billboard depends on the training tasks $S_1, \ldots, S_t$ and produces a representation that can be specialized to one of those tasks. However, by privacy, the representation would have almost the same distribution if we had inserted data S_{t+1} from a fresh training task. Hence, the representation can also be specialized to this unseen test task as well. The formal result is stated and proved in §3.

Separating DP billboard multitask and DP metalearning from JDP multitask learning. Our first main contribution is to show a separation between JDP multitask learning and billboard multitask learning for an estimation problem that we introduce. Since our models satisfy a hierarchy from easier to hardest, this result implies a separation between JDP multitask learning and metalearning as well. Such a result has no nonprivate analogue since, absent a privacy constraint, multitask learning implies metalearning (Aliakbarpour et al., 2024). Perhaps not surprisingly, the nonprivate equivalence is inherently privacy-violating—it uses the concatenation of the training data sets as the "representation".

We prove the separation using what we call the ***indexed mean estimation*** problem. Here, each training datum has the form (x, j) where $x \in \{\pm 1\}^d$ is a vector and $j \in [d]$ is an index of a coordinate in the vector. For a task P, which is a distribution over pairs (x, j), the goal is to output an estimate of the mean of the j-th coordinate $\mathbb{E}_{(x,j)\sim P}[x_j]$ low mean squared error. In the multitask learning problem, we will consider t-tuples of tasks with the following constraints: (1) For every individual task P_i, the marginal distribution of the vector x is identical. (2) For every individual task P_i, the marginal distribution of j is deterministic. Under these assumptions, the x part of the distribution has some common mean vector p and and each individual has their own j_i, and they want an estimate of p_{j_i}.

Our separations show that DP billboard algorithms for this problem have much higher loss than general JDP algorithms, when the dimension d is larger than the number t of individuals. Intuitively, a JDP algorithm can simply give a private estimate of p_{j_i} to each individual, which is obtained by averaging x_{j_i} over all individuals and adding noise. Since we only compute t values in total, the noise variance can be proportional to t. In contrast, a billboard DP algorithm cannot depend on the specific values j_i held by each individual, and thus the billboard must contain enough information to estimate p_j for *most* coordinates j. Since the billboard must estimate every coordinate, it must add noise proportional to d. This last statement follows by adapting the lower bounds of Bun et al. (2014); Dwork et al. (2015) for marginal estimation.

Separating* 1*-out-of-*t *DP from JDP. We also use the indexed mean estimation problem to separate 1-out-of-t and JDP. In JDP, an attacker sees estimates of $t-1$ distinct coordinates of p, while in 1-out-of-t DP, each individual gets just a single coordinate of p so privacy intuitively requires much less noise.

Our results for indexed mean estimation are summarized in Table 1, and stated formally in Section 4.

Separating DP billboard and JDP for classification. The separation above applies for an estimation (unsupervised learning) problem. We also extend our results to prove a separation for a binary classification (supervised learning) problem that we call ***indexed classification***. In this problem, each task is a distribution over labeled examples $((x,j),y)$ where $x \in \{\pm 1\}^d$, $j \in [d]$, and $y \in \{\pm 1\}$. Our goal is to produce a classifier that predicts y while minimizing the excess classification error. We consider distributions in which y is strongly correlated with x_j but uncorrelated with other coordinates of x. We set these distributions up in such a way that finding a good classifier essentially requires estimating the mean of x_j. From here, we proceed similarly to indexed mean estimation. We give a careful argument that a DP billboard must estimate the mean vector of x, denoted p, but a JDP algorithm can get away with estimating the mean of a only small number of coordinates of p, and thereby introduce less noise.

This argument is more complex than for indexed mean estimation, since a good billboard algorithm only implies good estimates of the *sign* of each coordinate of p. We overcome this by showing that even this simpler problem is hard under differential privacy, via a novel extension of the fingerprinting technique (Bun et al., 2018; Dwork et al., 2015).

Our results for indexed classification are summarized in Table 1. See §5 for formal statements.

1.2 Related Work

Without privacy constraints, there is a large body of literature on both multitask learning and metalearning, including related concepts or alternative names such as *transfer learning*, *learning to learn*, and *few-shot learning*, which is too vast to survey here. Another related learning framework is *collaborative learning* (Blum et al., 2017), which considers multitask classification in a setting where each task has a different marginal distribution on features but there exists a single good labeling function for every task.

There is also related work looking at privacy for model personalization. The most directly related work is that of Jain et al. (2021), together with predecessors on private recommender systems (e.g. McSherry and Mironov (2009)), which fits into our framework of JDP multitask learning. Krichene et al. (2023) consider a private multitask learning setting where each individual can contribute data to one task, as opposed to our framework in which tasks perfectly correspond to individuals. Li et al. (2020) also present a taxonomy of frameworks for multitask and metalearning, but their focus is on more specialized learning procedures than those we consider in our work. Räisä et al. (2024) address a different metalerning scenario, aiming to obtain models that are private with respect to all the individuals' data by doing metalearning where the model is first trained on simulated data and then specialized to the real data that requires protection.

Private model personalization has some similarities to *federated learning* (see the survey in Kairouz et al. (2019)), in which the data for each task is stored on a different device, with devices coordinated by a central server, and we want the server to obtain a single good model while minimizing what they can learn about the training data. Federated learning refers to a distributed system architecture and not a particular learning objective or privacy model. The objective may still be to solve a single learning problem, and there are not necessarily any privacy constraints at all.

The concept of joint differential privacy was introduced in Kearns et al. (2014). Our notion of 1-out-of-t DP is also called *marginal differential privacy* (Kannan et al., 2018). The majority of JDP algorithms use the billboard model, which was first formally defined in Hsu et al. (2014), but was used implicitly in McSherry and Mironov (2009); Gupta et al. (2010).

Our lower bound arguments are based on the fingerprinting methodology that was introduced in Ullman (2013); Bun et al. (2014) and further refined in Dwork et al. (2015). In particular we use (and prove extensions of) the so-called fingerprinting lemmas from Bun et al. (2017); Peter et al. (2024). See Kamath and Ullman (2020) for a partial survey.

2 PRIVACY DEFINITIONS

We now present the formal definitions of privacy that we utilize throughout the paper. First we present the definition of standard (central) differential privacy, which is the natural privacy notion for billboard algorithms and for metalearning, as both of these produce a single public output.

Table 1: Asymptotic bounds on the squared-error for Indexed Mean Estimation and excess error for Indexed Classification (assuming $t \ll d$ and n constant). Here (ε, δ) are the privacy parameters, and we assume $\varepsilon \leq 1$ and $\delta = 1/\text{poly}(t)$ for simplicity.

Privacy Requirement and Learning Objective	Indexed Mean Estimation		Indexed Classification	
	Upper Bound	Lower Bound	Upper Bound	Lower Bound
Nonprivate baseline	$\frac{1}{t}$		$\frac{1}{\sqrt{t}}$	
1-out-of-t DP Multitask	$\frac{1}{\varepsilon^2 t^2} + \frac{1}{t}$	-	$\frac{1}{\varepsilon t} + \frac{1}{\sqrt{t}}$	-
JDP Multitask	$\frac{1}{\varepsilon^2 t} + \frac{1}{t}$	$\frac{1}{\varepsilon^2 t}$	$\frac{1}{\varepsilon\sqrt{t}} + \frac{1}{\sqrt{t}}$	$\frac{1}{\varepsilon\sqrt{t}}$
DP Billboard Multitask	$\frac{d}{\varepsilon^2 t^2} + \frac{1}{t}$	$\frac{d}{\varepsilon^2 t^2}$	$\frac{\sqrt{d}}{\varepsilon t} + \frac{1}{\sqrt{t}}$	$\frac{\sqrt{d}}{\varepsilon t}$
DP Metalearning	$\frac{d}{\varepsilon^2 t^2} + \frac{1}{t}$	$\frac{d}{\varepsilon^2 t^2}$	$\frac{\sqrt{d}}{\varepsilon t} + \frac{1}{\sqrt{t}}$	$\frac{\sqrt{d}}{\varepsilon t}$

Definition 2.1 (Differential Privacy (Dwork et al., 2006; Bun and Steinke, 2016))**.** A randomized algorithm $\mathcal{M} : \mathcal{Z}^t \to \mathcal{O}$ satisfies the following differential privacy notions at the person level if, for any pair of $D, D' \in \mathcal{Z}^t$ that differ in one person's dataset,

- (ε, δ)-DP: $\forall E \subset \mathcal{O}$,
  $\mathbb{P}[\mathcal{M}(D) \in E] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}(D') \in E] + \delta$.
- ρ-zCDP: $\forall \alpha \in (1, \infty), \mathrm{D}_{\alpha}(\mathcal{M}(D) \,\|\, \mathcal{M}(D')) \leq \rho\alpha$,

where $\mathrm{D}_{\alpha}(\cdot \,\|\, \cdot)$ denotes Rényi divergence of order α.

Intuitively, differential privacy protects against an adversary from inferring any person i's input even if the adversary knows all the other peoples' inputs as well as the algorithm output. We will say that a billboard algorithm is differentially private if the output of the curator is differentially private.

Next we consider different models of privacy for algorithms that may produce *different outputs* for different people. The definitions protect against an adversary who is trying to infer the input of some person i and who gets all other $t-1$ inputs as well as some outputs of the mechanism. We consider other privacy requirements that vary according to what we assume to be visible to an attacker (corresponding to weaker or stronger privacy guarantees).

Joint DP First, we define Joint DP, where the adversary gets all outputs *except the secret person's output*.

Definition 2.2 (Joint Differential Privacy (Kearns et al., 2014))**.** A randomized algorithm $\mathcal{M} : \mathcal{Z}^t \to \mathcal{O}^t$ satisfies (ε, δ)-JDP if for all people $i \in [t]$, all pairs of neighboring datasets for person i $S_i, S_i' \in \mathcal{Z}$, all datasets for everyone else $S_{-i} \in \mathcal{Z}^{t-1}$, and all $E \subset \mathcal{O}^{t-1}$, $\mathbb{P}[\mathcal{M}(S_i; S_{-i})_{-i} \in E] \leq e^{\varepsilon}\mathbb{P}\left[\mathcal{M}(S_i'; S_{-i})_{-i} \in E\right] + \delta$ If the analogous definition holds with $\mathrm{D}_{\alpha}(\mathcal{M}(S_i; S_{-i})_{-i} \,\|\, \mathcal{M}(S_i'; S_{-i})_{-i}) \leq \rho\alpha$. instead, then $\mathcal{M}$ is ρ-JCDP.

1-out-of-t DP We also consider a weaker privacy requirement in which the adversary only sees one person's output (not the secret person's). We call this 1-out-of-t differential privacy (instead of marginal DP (Kannan et al., 2018)).

Definition 2.3 (1-out-of-t Differential Privacy)**.** A randomized algorithm $\mathcal{M} : \mathcal{Z}^t \to \mathcal{O}^t$ satisfies (ε, δ)-1-out-of-t-DP (resp. ρ-1-out-of-t-zCDP) if for all pairs of distinct people $i, j \in [t]$, all pairs of neighboring datasets for person i $S_i, S_i' \in \mathcal{Z}$, all datasets for everyone else $S_{-i} \in \mathcal{Z}^{t-1}$, and all $E \subset \mathcal{O}$,

$$\mathbb{P}\left[\mathcal{M}(S_i; S_{-i})_j \in E\right] \leq e^{\varepsilon}\mathbb{P}\left[\mathcal{M}(S_i'; S_{-i})_j \in E\right] + \delta$$

Resp. $\mathrm{D}_{\alpha}(\mathcal{M}(S_i; S_{-i})_j \,\|\, \mathcal{M}(S_i'; S_{-i})_j) \leq \rho\alpha \ \forall \alpha \in (1, \infty)$.

For further details about our notation and preliminaries see Appendix A.

3 BILLBOARD MULTITASK LEARNING IMPLIES METALEARNING

In this section, we present a general reduction from multitask learning to metalearning. Suppose $\mathcal{P}^{(t)}$ denotes the set of all t-tuples that an algorithm $\mathcal{M}$ multitask learns. Without loss of generality, we assume that $\mathcal{P}^{(t)}$ is closed under permutation. For any tuple in $\mathcal{P}^{(t)}$, an unseen task could be a similar one that can be swapped with any of the t tasks, and the new tuple would still remain in the domain. We can apply metalearning to this set of tasks. In other words, consider a distribution $\mathcal{Q}$ over *exchangeable tuples* of length $t+1$. Let $\mathcal{Q}^{(t)}$ be the marginal distribution over t-tuples of distributions obtained by taking the first t distributions from a $(t+1)$-tuple drawn from $\mathcal{Q}$. If $\mathcal{P}^{(t)}$ contains all the t-tuples of distributions that are in the support of $\mathcal{Q}^{(t)}$, then $\mathcal{M}$ can metalearn as well. We formalize this argument in the following theorem (proven in Appendix B). The proof relies on Lemma A.5, which requires that the post-processing function (the loss) to be bounded.

Theorem 3.1. *Fix parameters $\alpha > 0$, $\varepsilon > 0$, $\delta \in (0,1)$, $t \in \mathbb{N}$, $n \in \mathbb{N}$, a loss function ℓ taking values in $[0,1]$, and a distribution $\mathcal{Q}$ over $(t+1)$-tuples of exchangeable distributions. Assume $\mathcal{P}^{(t)}$ is a set of t-tuples of distributions that contains all the t-tuples of distributions in the support of $\mathcal{Q}^{(t)}$ (as defined above). Let $\mathcal{M} = (\mathcal{M}_{BB}, \mathcal{M}_{pers})$ be an (ε,δ)-DP billboard algorithm. If $\mathcal{M}$ multitask learns $\mathcal{P}^{(t)}$ with t tasks and n samples per task with expected error α, then $\mathcal{M}$ also metalearns $\mathcal{Q}$ with t training tasks, n samples per training task, and n personalization samples, with expected error at most $e^{\varepsilon}\alpha + \delta$.*

4 INDEXED MEAN ESTIMATION

In this section we showcase how different frameworks for private personalization impact the error achieved by a learning algorithm in an estimation problem. The proofs are provided in Appendix C.

Indexed Mean Estimation For simplicity, we first describe the multitask learning version of the estimation problem we focus on. Every person i in $[t]$ gets n samples drawn from distribution P_X over domain $\{\pm 1\}^d$ and an index $j_i \in [d]$. P_X is a product of d independent distributions over $\{\pm 1\}$, with mean $\mathbf{p} = (p_1, \ldots, p_d)$ in $[-1,+1]^d$. Person i aims to find an estimate, $\hat{p}_{j_i}$, for the true mean of the j_i-th coordinate in distribution P_X, denoted by p_{j_i}. Their loss function is defined as the squared difference between the mean estimate for coordinate j_i and its true mean:

$$\ell_{\text{mean}}(P_X, \hat{p}_{j_i}) := \tfrac{1}{4}(\hat{p}_{j_i} - p_{j_i})^2,$$

(where the $\frac{1}{4}$ factor keeps the loss in $[0,1]$).

To cast this problem in the language of multitask learning, as in Definition 1.1, we can think of the equivalent setting where the input of each person comes from a distribution P_i. Instead of assuming that each person has n i.i.d. samples from a shared P_X and an index j_i, we define P_i to be the Cartesian product of P_X and a singleton distribution on $\{j_i\}$. In this new setting, the input of person i is n samples drawn from P_i.

We define the domain of t-tuples of distributions for t people that use the same feature distribution P_X, but potentially different indices as the following

$$\begin{aligned}\mathcal{P}_{\text{est},d,t} := \{(P_1, \ldots, P_t) \mid P_i := P_X(\mathbf{p}) \otimes \text{Det}(j_i) \\ \forall i \in [t], \forall \mathbf{p} \in [-1,+1]^d, \forall j_1, \ldots, j_t \in [d]^t\},\end{aligned} \tag{1}$$

where $P_X(\mathbf{p})$ is the distribution P_X with mean $\mathbf{p}$ and $\text{Det}(j_i)$ deterministically returns j_i.

We establish upper bounds for indexed mean estimation in the multitask learning frameworks. All three bounds result from applying the Gaussian mechanism to the empirical means with variance adjusted according to the privacy requirements.

Proposition 4.1 (Nonprivate upper bound)**.** *For any $d \in \mathbb{N}$ and $t \in \mathbb{N}$, and loss function ℓ_{mean}, there exists an algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error $1/4t$ with t tasks and one sample per task .*

Theorem 4.2 (1-out-of-t upper bound)**.** *For any $\rho \ge 0$, $d \in \mathbb{N}$, and $t \in \mathbb{N}$, and loss function ℓ_{mean}, there exists an algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error $\frac{1}{2\rho t^2} + \frac{1}{4t}$ with t tasks and one sample per task and satisfies ρ-1-out-of-t-zCDP. Furthermore, for any $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, and $t \in \mathbb{N}$, there exists an algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1}) \cdot \frac{1}{t^2} + \frac{1}{4t}$ with t tasks and one sample per task, and satisfies (ε,δ)-1-out-of-t-DP.*

Theorem 4.3 (JDP upper bound)**.** *For any $\rho \ge 0$, $d \in \mathbb{N}$, and $t \in \mathbb{N}$, and loss function ℓ_{mean}, there exists an algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error $O(\frac{1}{\rho t}) + \frac{1}{4t}$ with t tasks and one sample per task and satisfies ρ-JCDP. Moreover, for any $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, and $t \in \mathbb{N}$, and loss function ℓ_{mean}, there exists an algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1}) \cdot \frac{1}{t} + \frac{1}{4t}$ with t tasks and one sample per task, and satisfies (ε,δ)-JDP.*

Theorem 4.4 (Billboard upper bound)**.** *For any $\rho \ge 0, d \in \mathbb{N}, t \in \mathbb{N}$ and for loss function ℓ_{mean}, there exists a billboard algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error $\frac{d}{2\rho t^2} + \frac{1}{4t}$ with t tasks and 1 sample per task and satisfies ρ-zCDP. Moreover, for any $\varepsilon > 0, \delta \in (0,1), d \in \mathbb{N}, t \in \mathbb{N}$ and for loss function ℓ_{mean}, there exists an algorithm in the billboard model that multitask learns $\mathcal{P}_{est,d,t}$ to error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1}) \cdot \frac{d}{t^2} + \frac{1}{4t}$ with t tasks and 1 sample per task, and satisfies (ε,δ)-DP.*

This problem can easily be extended to metalearning for a metadistribution over $\mathcal{P}_{\text{est},d,t+1}$ with t training tasks and a test task with loss function ℓ_{mean}. Theorem 4.4, combined with our general reduction from metalearning to multitask learning in Theorem 3.1, also implies the existence of a metalearning algorithm within the billboard model.

Corollary 4.5 (Metalearning upper bound)**.** *For any $\varepsilon > 0, \delta \in (0,1), d \in \mathbb{N}, t \in \mathbb{N}$ and for loss function ℓ_{mean}, there exists a pair of algorithms $(\mathcal{M}_{meta}, \mathcal{M}_{pers})$ that metalearn a distribution $\mathcal{Q}$ over $\mathcal{P}_{est,d,t+1}$ with error $\Theta(e^{\varepsilon} \cdot \min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1}) \cdot \frac{d}{t^2} + \frac{e^{\varepsilon}}{4t} + \delta$ using t training tasks, one sample per training task and a test task with one personalization sample and $\mathcal{M}_{meta}$ satisfies (ε,δ)-DP.*

Next, we formally separate the privacy frameworks in the multitask setting by proving a number of lower bounds. The proofs for the JDP multitask lower bound and the metalearning lower bound use the fingerprinting lower bound for private mean estimation. Specifically, we adapt the tracing attack used in the proof of the standard lower bound to each of our threat models.

Theorem 4.6 (JDP lower bound)**.** *Fix $\alpha > 0$, $t \in \mathbb{N}$, $d \geq ct^2$ for a sufficiently large constant, $\varepsilon > 0$ and $\delta \in (0, \frac{1}{96t})$. Let ℓ_{mean} be the loss function. Let $\mathcal{M}$ be an algorithm that multitask learns $\mathcal{P}_{d,t}$ with error α with t tasks and one sample per task and satisfies (ε,δ)-JDP. Then, $\alpha \geq \Omega(\min\{\frac{1}{\varepsilon^2 t}, 1\})$.*

Theorem 4.7 (Metalearning lower bound)**.** *Fix $\alpha > 0$, $\varepsilon > 0$, $t \in \mathbb{N}$, $d \in \mathbb{N}$ and $\delta \in (0, \frac{1}{96t})$. Let ℓ_{mean} be the loss function. Let $\mathcal{M} = (\mathcal{M}_{meta}, \mathcal{M}_{pers})$ be a pair of algorithms that metalearn a distribution $\mathcal{Q}$ over $\mathcal{P}_{est,d,t+1}$ with error α with t training tasks, one sample per training task and a test task with one personalization sample and $\mathcal{M}_{meta}$ satisfies (ε,δ)-JDP. Then, $\alpha \geq \Omega(\min\{\frac{d}{\varepsilon^2 t^2}, 1\})$.*

For the multitask learning lower bound in the billboard model, we use Theorem 3.1 to reduce to the metalearning lower bound in Theorem 4.7.

Corollary 4.8 (Billboard lower bound)**.** *Fix $\alpha > 0$, $\varepsilon \in (0,1]$, $t \in \mathbb{N}$, $\delta \in (0, \frac{1}{8 \cdot 144t})$ and $d \geq ct$ for a sufficiently large constant c. Let ℓ_{mean} be the loss function. Let $\mathcal{M}$ be a billboard algorithm that multitask learns $\mathcal{P}_{d,t}$ with error α with t tasks and one sample per task, and satisfies (ε,δ)-JDP. Then, $\alpha \geq \Omega(\min\{\frac{d}{\varepsilon^2 t^2}, 1\})$.*

5 INDEXED CLASSIFICATION

In this section, we characterize the error achievable on a classification problem in different private personalization frameworks. For the proofs of the results presented here see Appendix D.

Indexed Classification Every individual i has an index $j_i \in [d]$ and n labeled samples $\{(x^{(i,k)}, y^{(i,k)})\}_{k \in [n]}$ drawn from a personal distribution $R_{\mathbf{p},j_i}$ over $\{\pm 1\}^d \times \{\pm 1\}$, where $\mathbf{p}$ is a parameter of the data distribution that is the same among individuals. For a fixed vector $\mathbf{p} = (p_1, \ldots, p_d)$ in $[-1,+1]^d$ and an index $j \in [d]$, we define $R_{\mathbf{p},j}$ as the distribution on the outcome of the following sampling process. First, draw an auxiliary random variable $w \in \{\pm 1\}^d$ from a product of d independent distributions with mean $\mathbf{p}$. Then, draw $y \in \{\pm 1\}$ uniformly. Finally, construct x as follows

$$x_\ell = \begin{cases} w_\ell, & \textbf{if } \ell \neq j \\ w_\ell \cdot y, & \textbf{if } \ell = j. \end{cases}$$

Individual i's goal is to use these samples, and in cooperation with other individuals learn a classifier of the form $\hat{f}_i(x_{j_i}) : \{\pm 1\} \to \{\pm 1\}$ that predicts labels y as correctly as possible using only the j_i-th feature. Depending on the output structure of the algorithm, $\hat{f}_i$ might be sent as a private message from the curator or it might be calculated locally by individual i using the representation broadcasted by the curator. The loss function of individual i is the difference between the misclassification error of $\hat{f}_i$ and the optimal classifier for distribution R_{j_i}:

$$\ell_{\text{class}}(R_{\mathbf{p},j_i}, \hat{f}_i) := \mathbb{P}\left[\hat{f}_i(x_{j_i}) \neq y\right] - \min_{f_i} \mathbb{P}\left[f_i(x_{j_i}) \neq y\right],$$

where the probability is taken over $(x,y) \sim R_{\mathbf{p},j_i}$.

We cast this problem in the "multitask learning" language of Definition 1.1 and specify the connection between the data distributions by considering that every individual has n copies of index j_i and that the n samples come from the following family of t-tuples of distributions

$$\mathcal{P}_{\text{class},d,t} := \{(P_1, \ldots, P_t) \mid P_i = R_{\mathbf{p},j_i} \otimes \text{Det}(j_i) \forall i \in [t], \\ \forall j_i \in [d], \forall \mathbf{p} \in [-1,+1]^d\},$$

In this definition, for all P_i that are in the same tuple, the distributions over the features, $R_{\mathbf{p},j_i}$ are parameterized by the same $\mathbf{p}$. However, the P_is can potentially have different j_is for different people. For convenience we write $(x,j,y) \sim P_i$, instead of (x,y,j).

We first prove upper bounds on the indexed classification problem by reducing to and from a related learning problem called indexed sign estimation (see Appendix D for details), encoding the sign estimation problem into a mean estimation problem, and finally applying the upper bounds from Section 4 for indexed mean estimation.

Lemma 5.1 (Non-private upper bound)**.** *For any $d \in \mathbb{N}$ and $t \in \mathbf{N}$, and loss function ℓ_{class}, there exists an* *algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error $\frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task.*

Theorem 5.2 (1-out-of-t upper bound)**.** *For any parameters $\rho \geq 0$, $d \in \mathbb{N}$, $t \in \mathbb{N}$ and for loss function ℓ_{class}, there exists an algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error $\frac{\sqrt{2}}{t\sqrt{\rho}} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task and satisfies ρ-1-out-of-t-zCDP. Moreover, it is implied that for any parameters $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, $t \in \mathbb{N}$ and for loss function ℓ_{class}, there exists an algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1/2}) \cdot \frac{1}{t} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies (ε,δ)-1-out-of-t-DP.*

Theorem 5.3 (JDP upper bound)**.** *For any parameters $\rho \geq 0$, $d \in \mathbb{N}$, $t \in \mathbb{N}$, and for loss function ℓ_{class}, there exists an algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error $O(\frac{1}{\sqrt{\rho t}}) + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies ρ-JCDP. Moreover, it is implied that for any $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, $t \in \mathbb{N}$, and for loss function ℓ_{class}, there exists an algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1/2}) \cdot \frac{1}{\sqrt{t}} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies (ε,δ)-JDP.*

Theorem 5.4 (Billboard upper bound)**.** *For any parameters $\rho \geq 0$, $d \in \mathbb{N}$, $t \in \mathbb{N}$, and for loss function ℓ_{class}, there exists a billboard algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error $\frac{\sqrt{2d}}{t\sqrt{\rho}} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies ρ-zCDP. Moreover, it is implied that for any parameters $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, $t \in \mathbb{N}$, and for loss function ℓ_{class}, there exists a billboard algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1/2}) \cdot \frac{\sqrt{d}}{t} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies (ε,δ)-DP.*

Finally, Theorem 5.4 and the reduction of metalearning to multitask learning in Theorem 3.1 imply that there exists a mealearning algorithm whose metalearning process is differentially private.

Corollary 5.5 (Metalearning upper bound)**.** *For any $\rho \geq 0$, $d \in \mathbb{N}$, $t \in \mathbb{N}$, and for loss function ℓ_{class}, there exists a pair of algorithms $(\mathcal{M}_{meta}, \mathcal{M}_{pers})$ that metalearn a metadistribution $\mathcal{Q}$ over $\mathcal{P}_{class,d,t+1}$ with error $\Theta(e^{\varepsilon} \cdot \min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1/2}) \cdot \frac{\sqrt{d}}{t} + \frac{e^{\varepsilon}}{\sqrt{t}} + \delta$ with t training tasks, 1 sample per training task and a test task with 1 personalization sample and $\mathcal{M}_{meta}$ satisfies (ε,δ)-DP.*

We also prove lower bounds for the error of indexed classification that separate the interaction and privacy models in the multitask learning setting. We initially prove lower bounds for indexed sign estimation (defined in Appendix D) using our fingerprinting lemma. Finally, we prove the lower bounds for indexed classification by using a reduction of indexed sign estimation to indexed classification.

Theorem 5.6 (classification JDP lower bound)**.** *Fix $\alpha \in (0, \frac{1}{16})$, $t \in \mathbb{N}$, $d \geq ct^2$ for a sufficiently large constant c, $\varepsilon > 0$ and $\delta \in (0, \frac{1}{2t})$. Let $\mathcal{M}$ be an algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error α, for loss function ℓ_{class}, with t tasks and 1 sample per task, and satisfies (ε,δ)-JDP. Then, $\alpha \geq \Omega(\min\{\frac{1}{\varepsilon\sqrt{t}}, 1\})$.*

Theorem 5.7 (Metalearning lower bound)**.** *Fix $\alpha \in (0, \frac{1}{8})$, $t \in \mathbb{N}$, $d \in \mathbb{N}$, $\varepsilon \in (0,1]$ and $\delta \in (0, \frac{1}{2t})$. Let $\mathcal{M} = (\mathcal{M}_{meta}, \mathcal{M}_{pers})$ be a pair of algorithms that metalearn a distribution $\mathcal{Q}$ over $\mathcal{P}_{class,d,t+1}$ with error α, for loss function ℓ_{class}, using t training tasks, 1 sample per training task and a test task with 1 personalization sample, and $\mathcal{M}_{meta}$ satisfies (ε,δ)-DP. Then, $\alpha \geq \Omega(\min\{1, \frac{\sqrt{d}}{\varepsilon t}\})$.*

Theorem 5.8 (Billboard lower bound)**.** *Fix $\alpha \in (0, \frac{1}{8})$, $t \in \mathbb{N}$, $d \geq \frac{\varepsilon^2 t}{4}$, $\varepsilon \in (0,1]$ and $\delta \in (0, \frac{1}{32^2 t})$. Let $\mathcal{M}$ be a billboard algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error α, for loss function ℓ_{class}, with t tasks and 1 sample per task and satisfies (ε,δ)-DP. Then, $\alpha \geq \Omega(\min\{\frac{\sqrt{d}}{\varepsilon t}, 1\})$.*

6 CONCLUSION

Our work gave a taxonomy of formal frameworks for private personalized learning and proved separations and equivalences between these frameworks. Our work leaves open several future directions. First, can we more thoroughly understand what types of relationships between tasks yield separations? In our separations for classification, we assume that the x part of the data points of all participating individuals is drawn from the same distribution. A natural follow-up question is what upper bounds can be achieved with privacy when the distributions of the individuals are related in more realistic ways. Second, can we relate private multitask learning to learning from a combination of public and private data. From an individual's perspective, private multitask learning can be viewed as a form of public-private learning, where their own samples are treated as public and the samples of others as private. These samples must be used in a specific order, first the private samples and then the public ones. Investigating how this perspective can lead to new algorithms for private multitask learning would be an interesting direction for future research.

Acknowledgements

M.A. was supported by NSF awards CNS-2120667, CNS-2120603, CCF-1934846, and the Hariri Institute for Computing at Boston University. Part of this work was conducted while M.A. was affiliated with Boston University and Northeastern University, and while serving as a research fellow at the Simons Institute for the Theory of Computing. K.B. was supported by NSF awards CNS-2232692 and CCF-2311649. The work of A.S. at BU was supported by NSF awards CCF-1763786 and CNS-2120667, Faculty Awards from Google and Apple, and Cooperative Agreement CB16ADR0160001 with the Census Bureau. This work was done while M.S. was affiliated with Boston University, supported by NSF awards CCF-1763786 and CNS-2120667. J.U. was supported by NSF awards CNS-2232692 and CNS-2247484. The views expressed in this paper are those of the authors and not those of the U.S. Census Bureau or any other sponsor.

References

Aliakbarpour, M., Bairaktari, K., Brown, G., Smith, A., Srebro, N., and Ullman, J. (2024). Metalearning with very few samples per task. In *Conference on Learning Theory*, COLT '24, pages 46–93. PMLR.

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. (2016). Algorithmic stability for adaptive data analysis. In *ACM Symposium on the Theory of Computing*, STOC '16, Cambridge, MA.

Blum, A., Haghtalab, N., Procaccia, A. D., and Qiao, M. (2017). Collaborative pac learning. *Advances in Neural Information Processing Systems*, 30.

Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, TCC '16, pages 635–658, Beijing, China. https://arxiv.org/abs/1605.02065.

Bun, M., Steinke, T., and Ullman, J. (2017). Make up your mind: The price of online queries in differential privacy. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1306–1325, Philadelphia, PA. SIAM.

Bun, M., Ullman, J., and Vadhan, S. (2014). Fingerprinting codes and the price of approximate differential privacy. In *ACM Symposium on the Theory of Computing*, STOC '14, pages 1–10, New York, NY, USA. https://arxiv.org/abs/1311.3158.

Bun, M., Ullman, J., and Vadhan, S. (2018). Fingerprinting codes and the price of approximate differential privacy. *SIAM Journal on Computing*, 47(5):1888–1938.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2014). Preserving statistical validity in adaptive data analysis. *arXiv preprint arXiv:1411.2664*.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Conference on Theory of Cryptography*, TCC '06, pages 265–284, New York, NY, USA.

Dwork, C., Smith, A., Steinke, T., Ullman, J., and Vadhan, S. (2015). Robust traceability from trace amounts. In *IEEE Symposium on Foundations of Computer Science*, FOCS '15.

Gupta, A., Ligett, K., McSherry, F., Roth, A., and Talwar, K. (2010). Differentially private combinatorial optimization. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 1106–1125.

Hsu, J., Huang, Z., Roth, A., Roughgarden, T., and Wu, Z. S. (2014). Private matchings and allocations. In *ACM symposium on Theory of computing*, pages 21–30.

Jain, P., Rush, J. K., Smith, A., Song, S., and Thakurta, A. (2021). Differentially private model personalization. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2021)*.

Jung, C., Ligett, K., Neel, S., Roth, A., Sharifi-Malvajerdi, S., and Shenfeld, M. (2020). A new analysis of differential privacy's generalization guarantees. In *Innovations in Theoretical Computer Science (ITCS)*.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., and Cummings, R. (2019). Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*.

Kamath, G. and Ullman, J. (2020). A primer on private statistics. *arXiv preprint arXiv:2005.00010*.

Kannan, S., Morgenstern, J., Rogers, R., and Roth, A. (2018). Private pareto optimal exchange. *ACM Transactions on Economics and Computation (TEAC)*, 6(3-4):1–25.

Kearns, M., Pai, M. M., Roth, A., and Ullman, J. (2014). Mechanism design in large games: incentives and privacy. In *Proceedings of the 5th ACM*

Conference on Innovations in Theoretical Computer Science, ITCS '14, pages 403–410, Princeton, NJ. ACM.

Krichene, W., Jain, P., Song, S., Sundararajan, M., Thakurta, A. G., and Zhang, L. (2023). Multi-task differential privacy under distribution skew. In *International Conference on Machine Learning*, pages 17784–17807. PMLR.

Li, J., Khodak, M., Caldas, S., and Talwalkar, A. (2020). Differentially private meta-learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ligett, K., Neel, S., Roth, A., Waggoner, B., and Wu, S. Z. (2017). Accuracy first: Selecting a differential privacy level for accuracy constrained erm. In *Advances in Neural Information Processing Systems 30*, NIPS '17, pages 2566–2576.

McSherry, F. and Mironov, I. (2009). Differentially private recommender systems: Building privacy into the netflix prize contenders. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–636.

Peter, N., Tsfadia, E., and Ullman, J. (2024). Smooth lower bounds for differentially private algorithms via padding-and-permuting fingerprinting codes. In *Conference on Learning Theory*, COLT '24, pages 4207–4239. PMLR.

Räisä, O., Markou, S., Ashman, M., Bruinsma, W. P., Tobaben, M., Honkela, A., and Turner, R. E. (2024). Noise-aware differentially private regression via meta-learning. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C., editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Steinke, T. A. (2016). *Upper and Lower Bounds for Privacy and Adaptivity in Algorithmic Data Analysis*. PhD thesis.

Ullman, J. (2013). Answering $n^{2+o(1)}$ counting queries with differential privacy is hard. In *ACM Symposium on the Theory of Computing*, STOC '13.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A PRELIMINARIES

Definition A.1 (Rényi Divergence). For two probability distributions P and Q defined over the same domain, the Rényi divergence of order $\alpha > 1$ is

$$\mathrm{D}_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q} \left(\frac{P(x)}{Q(x)} \right)^\alpha .$$

Definition A.2 (Exchangeable distributions). A t-tuple of distributions $(P_1, \ldots, P_t) \sim \mathcal{Q}$ is exchangeable with respect to $\mathcal{Q}$ if for every permutation $\pi : [t] \to [t]$, the tuple $(P_{\pi(1)}, \ldots, P_{\pi(t)})$ is equally likely to occur in $\mathcal{Q}$ as the unpermuted one. When the metadistribution is clear from the context, we simply refer to the tuple as exchangeable.

Definition A.3 (Sensitivity). A function $q : \mathcal{X}^n \to \mathbb{R}^d$ has sensitivity Δ if for all $x, x' \in \mathcal{X}^n$ differing in a single entry, we have $\|q(x) - q(x')\|_2 \leq \Delta$.

Fact A.4 (Gaussian Mechanism for zCDP (Bun and Steinke, 2016)). *Let $q : \mathcal{X}^n \to \mathbb{R}^d$ be a sensitivity-Δ query. Consider the mechanism $M : \mathcal{X}^n \to \mathbb{R}^d$ that on input x, releases a sample from $\mathcal{N}(q(x), \mathbb{I}_d \sigma^2)$. Then, M satisfies $(\Delta^2/2\sigma^2)$-zCDP.*

Lemma A.5 (Post-processing). *For any (ε, δ)-DP algorithm $\mathcal{M}$ and any (possibly randomized) function $g : \mathcal{O} \to [0,1]$, for any D and D' that differ in the dataset of one person*

$$\mathbb{E}_{g,\mathcal{M}}\big[\, g(\mathcal{M}(D)) \big] \leq e^\varepsilon \mathbb{E}_{g,\mathcal{M}}\big[\, g(\mathcal{M}(D')) \big] + \delta.$$

Fact A.6 (Adapted from Bun and Steinke (2016)). *Let $\mathcal{M} : \mathcal{Z}^t \to \mathcal{O}$ satisfy ρ-zCDP, then $\mathcal{M}$ satisfies (ε, δ)-DP for all $\delta > 0$ and*

$$\varepsilon = \rho + 2\sqrt{\rho \log\left(\frac{1}{\delta}\right)}.$$

In other words, for any two parameters ε and $\delta > 0$, there exists an absolute constant c, such that if $\mathcal{M}$ satisfies $c \cdot \left(\min\left(\varepsilon, \frac{\varepsilon^2}{\log(1/\delta)}\right)\right)$-zCDP, then $\mathcal{M}$ satisfies (ε, δ)-DP.

Lemma A.7 (Birthday Paradox). *Suppose we toss t balls uniformly at random into d bins. If $d \geq ct^2$ for a sufficiently large constant c, the probability of any two balls falling into the same bin is at most $\frac{1}{2}$.*

B PROOFS FROM SECTION 3

Theorem 3.1 (Restated). Fix parameters $\alpha > 0$, $\varepsilon > 0$, $\delta \in (0,1)$, $t \in \mathbb{N}$, $n \in \mathbb{N}$, a loss function ℓ taking values in $[0,1]$, and a distribution $\mathcal{Q}$ over $(t+1)$-tuples of exchangeable distributions. Assume $\mathcal{P}^{(t)}$ is a set of t-tuples of distributions that contains all the t-tuples of distributions in the support of $\mathcal{Q}^{(t)}$ (as defined above). Let $\mathcal{M} = (\mathcal{M}_{\mathrm{BB}}, \mathcal{M}_{\mathrm{pers}})$ be an (ε, δ)-DP billboard algorithm. If $\mathcal{M}$ multitask learns $\mathcal{P}^{(t)}$ with t tasks and n samples per task with expected error α, then $\mathcal{M}$ also metalearns $\mathcal{Q}$ with t training tasks, n samples per training task, and n personalization samples, with expected error at most $e^\varepsilon \alpha + \delta$.

Proof. We focus on the error of $\mathcal{M}$ in the metalearning framework. As we have described the error in Definition 1.2, the error of this algorithm corresponds to its performance on the $t+1$-st distribution. More precisely, we have:

$$\mathop{\mathbb{E}}_{\substack{\mathcal{M}_{\text{pers}},\mathcal{M}_{\text{BB}},\\(P_1,\dots,P_{t+1})\sim\mathcal{Q},\\S_1\sim P_1^n,\dots,S_t\sim P_t^n,S_{t+1}\sim P_{t+1}^n}}\Big[\ell(P_{t+1},\mathcal{M}_{\text{pers}}(S_{t+1},\mathcal{M}_{\text{BB}}(S_1,\dots,S_t))\Big].$$

We know that the distributions in the tuple that we draw from $\mathcal{Q}$ are exchangeable, so the distribution of $(t+1)$-tuples of distributions remains identical when we swap P_{t+1} with any P_i. Therefore, in the metalearning error of $\mathcal{M}$ we can swap P_{t+1}, S_{t+1} with any P_i, S_i for $i \in [t]$. Let D be the datasets of the first t people, $D = (S_1, \dots, S_t)$, and D_{-i} be the t datasets from D after we swap the dataset of person i, S_i, with the dataset of person $t+1$, S_{t+1}. Thus, we have that

$$\mathop{\mathbb{E}}_{\substack{\mathcal{M}_{\text{pers}},\mathcal{M}_{\text{BB}},\\(P_1,\dots,P_{t+1})\sim\mathcal{Q},\\S_1\sim P_1^n,\dots,S_t\sim P_t^n,S_{t+1}\sim P_{t+1}^n}}\Big[\ell(P_{t+1},\mathcal{M}_{\text{pers}}(S_{t+1},\mathcal{M}_{\text{BB}}(D))\Big] = \frac{1}{t}\sum_{i\in[t]}\mathop{\mathbb{E}}_{\substack{\mathcal{M}_{\text{pers}},\mathcal{M}_{\text{BB}},\\(P_1,\dots,P_{t+1})\sim\mathcal{Q},\\S_1\sim P_1^n,\dots,S_t\sim P_t^n,S_{t+1}\sim P_{t+1}^n}}\Big[\ell(P_i,\mathcal{M}_{\text{pers}}(S_i,\mathcal{M}_{\text{BB}}(D_{-i}))\Big].$$

The error term in the equation above is analogous to the error term in the multitask learning algorithm, with one key distinction. In multitask learning, the i-th error term is computed when S_i is included in the dataset provided to $\mathcal{M}_{\text{BB}}$. We demonstrate that these two terms are closely related due to the privacy properties of the server's algorithm.

For every individual $i \in [t]$, D and D_{-i} are neighboring datasets that differ only by the replacement of one person's data. Consequently, the privacy of $\mathcal{M}_{\text{BB}}$ ensures that $\mathcal{M}_{\text{BB}}(D)$ and $\mathcal{M}_{\text{BB}}(D_{-i})$ are nearly indistinguishable in distribution. It is well established that post-processing does not alleviate this indistinguishability. That is, for any function applied to $\mathcal{M}_{\text{BB}}(D)$ and $\mathcal{M}_{\text{BB}}(D_{-i})$, the resulting outputs remain indistinguishable. Specifically, for each individual $i \in [t]$, given a fixed distribution P_i and dataset S_i, let $g_i(b)$ be a post-processing function.

$$g_i(b) = \mathop{\mathbb{E}}_{\mathcal{M}_{\text{pers}}}\Big[\ell(P_i,\mathcal{M}_{\text{pers}}(S_i,b))\Big]. \tag{2}$$

We apply Lemma A.5 for function g_i and obtain that

$$\mathop{\mathbb{E}}_{\substack{\mathcal{M}_{\text{pers}},\mathcal{M}_{\text{BB}},\\(P_1,\dots,P_{t+1})\sim\mathcal{Q},\\S_1\sim P_1^n,\dots,S_t\sim P_t^n,S_{t+1}\sim P_{t+1}^n}}\Big[\ell(P_i,\mathcal{M}_{\text{pers}}(S_i,\mathcal{M}_{\text{BB}}(D_{-i}))\Big] \le e^{\varepsilon}\cdot\mathop{\mathbb{E}}_{\substack{\mathcal{M}_{\text{pers}},\mathcal{M}_{\text{BB}},\\(P_1,\dots,P_{t+1})\sim\mathcal{Q},\\S_1\sim P_1^n,\dots,S_t\sim P_t^n,S_{t+1}\sim P_{t+1}^n}}\Big[\ell(P_i,\mathcal{M}_{\text{pers}}(S_i,\mathcal{M}_{\text{BB}}(D))\Big]+\delta$$

Therefore, the error of $\mathcal{M}$ in the metalearning setting is bounded by

$$\frac{1}{t}\sum_{i\in[t]}\mathop{\mathbb{E}}_{\substack{\mathcal{M}_{\text{pers}},\mathcal{M}_{\text{BB}},\\(P_1,\dots,P_{t+1})\sim\mathcal{Q},\\S_1\sim P_1^n,\dots,S_t\sim P_t^n,S_{t+1}\sim P_{t+1}^n}}\Big[\ell(P_i,\mathcal{M}_{\text{pers}}(S_i,\mathcal{M}_{\text{BB}}(D_{-i}))\Big] \le e^{\varepsilon}\frac{1}{t}\sum_{i\in[t]}\mathop{\mathbb{E}}_{\substack{\mathcal{M}_{\text{pers}},\mathcal{M}_{\text{BB}},\\(P_1,\dots,P_{t+1})\sim\mathcal{Q},\\S_1\sim P_1^n,\dots,S_t\sim P_t^n}}\Big[\ell(P_i,\mathcal{M}_{\text{pers}}(S_i,\mathcal{M}_{\text{BB}}(S_1,\dots,S_t)))\Big]+\delta.$$

In the second line, the billboard message is computed using data drawn from tasks 1 to t. This equation captures the multitask learning error for the first t tasks drawn from Q. Alternatively, this tuple is drawn from $Q^{(t)}$, making it a member of $\mathcal{P}^{(t)}$. Given the assumptions of the theorem—that $\mathcal{M}$ is a billboard algorithm that multitask learns $\mathcal{P}^{(t)}$ with an error bound of α—it follows that for all $(P_1, \dots, P_t) \in \mathcal{P}^{(t)}$, we have the following:

$$\frac{1}{t}\sum_{i\in[t]}\mathop{\mathbb{E}}_{\substack{\mathcal{M}_{\text{pers}},\mathcal{M}_{\text{BB}}\\S_1\sim P_1^n,\dots,S_t\sim P_t^n}}\Big[\ell(P_i,\mathcal{M}_{\text{pers}}(S_i,\mathcal{M}_{\text{BB}}(S_1,\dots,S_t)))\Big] \le \alpha. \tag{3}$$

Combining the inequalities above we obtain that

$$\mathop{\mathbb{E}}_{\substack{\mathcal{M}_{\text{pers}},\mathcal{M}_{\text{BB}},\\(P_1,\dots,P_{t+1})\sim\mathcal{Q},\\S_1\sim P_1^n,\dots,S_t\sim P_t^n,S_{t+1}\sim P_{t+1}^n}}\Big[\ell(P_{t+1},\mathcal{M}_{\text{pers}}(S_{t+1},\mathcal{M}_{\text{BB}}(S_1,\dots,S_t))\Big] \le e^{\varepsilon}\alpha+\delta.$$

Thus, the proof is complete. □

C PROOFS FROM SECTION 4

C.1 Upper bound proofs

Proposition 4.1 (Restated)**.** For any $d \in \mathbb{N}$ and $t \in \mathbb{N}$, and loss function ℓ_{mean}, there exists an algorithm that multitask learns $\mathcal{P}_{\text{est},d,t}$ with error $1/4t$ with t tasks and one sample per task .

Proof. The algorithm takes as input one sample $(x^{(i)}, j^{(i)})$ drawn from P_i for every person $i \in [t]$ and then computes the empirical mean $\bar{p} \leftarrow \frac{1}{t}\sum_{i\in[t]} x^{(i)}$. Note that for all $\ell \in [t]$ and $j \in [d]$, $x_j^{(\ell)}$ is independent of the other coordinates and $(x_j^{(\ell)})^2 = 1$. Therefore, the error is

$$\frac{1}{4t}\sum_{i\in[t]} \mathbb{E}\left[\left(\bar{p}_{j_i} - p_{j_i}\right)^2\right] = \frac{1}{4t}\sum_{i\in[t]} \mathrm{Var}\left(\bar{p}_{j_i} - p_{j_i}\right) = \frac{1}{4t}\sum_{i\in[t]} \mathrm{Var}\left(\frac{1}{t}\sum_{\ell\in[t]} x_{j_i}^{(\ell)} - p_{j_i}\right)$$
$$= \frac{1}{4t}\sum_{i\in[t]} \frac{1}{t^2}\sum_{\ell\in[t]} \mathrm{Var}\left(x_{j_i}^{(\ell)}\right) \leq \frac{1}{4t}.$$

□

Theorem 4.2 (Restated)**.** For any $\rho \geq 0$, $d \in \mathbb{N}$, and $t \in \mathbb{N}$, and loss function ℓ_{mean}, there exists an algorithm that multitask learns $\mathcal{P}_{\text{est},d,t}$ with error $\frac{1}{2\rho t^2} + \frac{1}{4t}$ with t tasks and one sample per task and satisfies ρ-1-out-of-t-zCDP. Furthermore, for any $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, and $t \in \mathbb{N}$, there exists an algorithm that multitask learns $\mathcal{P}_{\text{est},d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1}) \cdot \frac{1}{t^2} + \frac{1}{4t}$ with t tasks and one sample per task, and satisfies (ε, δ)-1-out-of-t-DP.

Proof. We propose the following algorithm, which begins with the empirical mean and produces a private version of the mean for the coordinate of each person $i \in [t]$ using the Gaussian mechanism. More precisely, the personalized output for person i with index j_i is

$$\hat{p}_{j_i} \leftarrow \bar{p}_{j_i} + Z_{j_i},$$

where $\bar{p}_{j_i} \leftarrow \frac{1}{t}\sum_{k\in[t]} x_{j_i}^{(k)}$ and $Z_{j_i} \sim \mathcal{N}\left(0, \frac{2}{\rho t^2}\right)$.

To show that this algorithm satisfies ρ-1-out-of-t-zCDP, fix a pair of distinct people $i \neq j$ and focus on the personalized outputs received by person k while replacing the input dataset of person i with a neighboring one. Since each person has a dataset of only one sample (i.e., $n = 1$), two neighboring datasets correspond to having two different samples. Now, if we change the sample of person i, the sensitivity of the mean at index j_k, corresponding to the output of person k, is bounded by

$$\left|\bar{p}_{j_k}(D) - \bar{p}_{j_k}(D')\right| \leq \frac{2}{t}.$$

Thus, using Gaussian mechanism as outlined in Fact A.4, $\hat{p}_{j_k}$ satisfy ρ–zCDP for all distinct i and k. Therefore, our algorithm satisfies ρ-1-out-of-t-zCDP.

Next, we analyze the error of our estimates. Since the Gaussian noise and the sampled data are independent, we have

$$\frac{1}{4t}\sum_{i\in[t]} \mathbb{E}\left[\left(\hat{p}_{j_i} - p_{j_i}\right)^2\right] = \frac{1}{4t}\sum_{i\in[t]} \mathrm{Var}\left(\hat{p}_{j_i} - p_{j_i}\right)$$
$$= \frac{1}{4t}\sum_{i\in[t]} \left(\mathrm{Var}\left(\bar{p}_{j_i} - p_{j_i}\right) + \mathrm{Var}\left(Z_{j_i}\right)\right) \leq \frac{1}{4t} + \frac{1}{2\rho t^2}.$$

The last inequality follows from the error analysis in the proof of Theorem 4.1. Using the standard conversion between zCDP and approximate DP, from Fact A.6, we show that this algorithm also satisfies (ε, δ)-DP by setting $\rho = \Theta\Big(\max\Big(\frac{1}{\varepsilon}, \frac{\log(1/\delta)}{\varepsilon^2}\Big)\Big)$. □

Theorem 4.3 (Restated)**.** For any $\rho \geq 0$, $d \in \mathbb{N}$, and $t \in \mathbb{N}$, and loss function ℓ_{mean}, there exists an algorithm that multitask learns $\mathcal{P}_{\text{est},d,t}$ with error $O(\frac{1}{\rho t}) + \frac{1}{4t}$ with t tasks and one sample per task and satisfies ρ-JCDP. Moreover, for any $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, and $t \in \mathbb{N}$, and loss function ℓ_{mean}, there exists an algorithm that multitask learns $\mathcal{P}_{\text{est},d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1}) \cdot \frac{1}{t} + \frac{1}{4t}$ with t tasks and one sample per task, and satisfies (ε, δ)-JDP.

Proof. Here, for each person i with index j_i, the algorithm returns an estimate

$$\hat{p}_{j_i} \leftarrow \bar{p}_{j_i} + Z_{j_i},$$

where $\bar{p}_{j_i} \leftarrow \frac{1}{t}\sum_{k\in[t]} x_{j_i}^{(k)}$ and $Z_{j_i} \sim \mathcal{N}\Big(0, \frac{2(t-1)}{\rho t^2}\Big)$. This satisfies ρ-JCDP because the global ℓ_2 sensitivity of the outputs given to the $t-1$ colluding people $(\bar{p}_{j_1}, \ldots, \bar{p}_{j_{t-1}})$ is $\frac{2\sqrt{t-1}}{t}$. Hence, on average over the t people participating

$$\begin{aligned}\frac{1}{4t}\sum_{i\in[t]} \mathbb{E}\Big[\Big(\hat{p}_{j_1} - p_{j_1}\Big)^2\Big] &= \frac{1}{4t}\sum_{i\in[t]} \mathrm{Var}\Big(\hat{p}_{j_i} - p_{j_i}\Big) \\ &= \frac{1}{4t}\sum_{i\in[t]}\Big(\mathrm{Var}\Big(\bar{p}_{j_1} - p_{j_1}\Big) + \mathrm{Var}\Big(Z_{j_1}\Big)\Big) \leq \frac{1}{4t} + \frac{(t-1)}{2\rho t^2}.\end{aligned}$$

□

Theorem 4.4 (Restated)**.** For any $\rho \geq 0, d \in \mathbb{N}, t \in \mathbb{N}$ and for loss function ℓ_{mean}, there exists a billboard algorithm that multitask learns $\mathcal{P}_{\text{est},d,t}$ with error $\frac{d}{2\rho t^2} + \frac{1}{4t}$ with t tasks and 1 sample per task and satisfies ρ-zCDP. Moreover, for any $\varepsilon > 0, \delta \in (0,1), d \in \mathbb{N}, t \in \mathbb{N}$ and for loss function ℓ_{mean}, there exists an algorithm in the billboard model that multitask learns $\mathcal{P}_{\text{est},d,t}$ to error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1}) \cdot \frac{d}{t^2} + \frac{1}{4t}$ with t tasks and 1 sample per task, and satisfies (ε, δ)-DP.

Proof. The billboard algorithm runs the Gaussian mechanism once and broadcasts the output to all the people. Then every person can use their index j_i to get the estimate they are interested in. More specifically, the algorithm takes as input one sample $(x^{(i)}, j^{(i)})$ drawn from P_i from every person i and then computes an empirical mean $\bar{p} := \frac{1}{t}\sum_{i\in[t]} x^{(i)}$ and outputs on the billboard

$$\hat{p} \leftarrow \bar{p} + Z, \quad \text{where } Z \sim \mathcal{N}\left(0, \frac{2d}{\rho t^2}\mathbb{I}_d\right). \tag{4}$$

The global ℓ_2 sensitivity of $\bar{p}$ is $\frac{2\sqrt{d}}{t}$, so this satisfies ρ-zCDP and, consequently, ρ-JCDP.

To analyze the error, by symmetry it suffices to analyze person i. Since sampling the data and the Gaussian noise are independent, we have

$$\mathbb{E}\Big[\Big(\hat{p}_{j_i} - p_{j_i}\Big)^2\Big] = \mathrm{Var}\Big(\hat{p}_{j_i} - p_{j_i}\Big) = \mathrm{Var}\Big(Z_{j_i}\Big) + \mathrm{Var}\Big(\bar{p}_{j_i} - p_{j_i}\Big). \tag{5}$$

We have $\mathrm{Var}\Big(Z_{j_i}\Big) = \frac{d}{2\rho t^2}$ and $\mathrm{Var}\Big(\bar{p}_{j_i} - p_{j_i}\Big) \leq \frac{1}{t}$. Summing these variances gives you that the total error is

$$\frac{1}{4t}\sum_{i\in[t]} \mathbb{E}\Big[(\hat{p}_{j_i} - p_{j_i})^2\Big] \leq \frac{d}{2\rho t^2} + \frac{1}{4t}.$$

Using the standard conversion between zCDP and approximate DP, from Fact A.6, we show that this algorithm also satisfies (ε, δ)-DP by setting $\rho = \Theta\Big(\max\Big(\frac{1}{\varepsilon}, \frac{\log(1/\delta)}{\varepsilon^2}\Big)\Big)$. □

C.2 Lower bound proofs

Theorem 4.6 (Restated)**.** Fix $\alpha > 0$, $t \in \mathbb{N}$, $d \geq ct^2$ for a sufficiently large constant, $\varepsilon > 0$ and $\delta \in (0, \frac{1}{96t})$. Let ℓ_{mean} be the loss function. Let $\mathcal{M}$ be an algorithm that multitask learns $\mathcal{P}_{d,t}$ with error α with t tasks and one sample per task and satisfies (ε, δ)-JDP. Then, $\alpha \geq \Omega(\min\{\frac{1}{\varepsilon^2 t}, 1\})$.

Proof. For the proof of Theorem 4.6 we follow a series of steps similar to that presented in Kamath and Ullman (2020), which itself closely follows the presentation in Steinke (2016).

We have a JDP algorithm on t tasks/people and d dimensions. Each person $i \in [t]$, has a sample $(x^{(i)}, j^{(i)}) \in \{\pm 1\} \times [d]$ and gives it as input to the algorithm. Let $S = \{(x^{(i)}, j^{(i)})\}_{i \in [t]}$ denote this dataset. We will construct a hard distribution over the family of distributions $\mathcal{P}_{d,t}$ that the samples are drawn from. We draw a vector $\mathbf{p} \in [-1, +1]^d$ and a vector $\mathbf{j} \in [d]^t$, both uniformly at random. The datapoint $x^{(i)}$ of person i is drawn from the product distribution with mean $\mathbf{p}$ and the index $j^{(i)}$ is deterministically j_i. Let the expected error of $\mathcal{M}$ be

$$\mathop{\mathbb{E}}_{\mathbf{p},\mathbf{j}}\left[\sum_{i \in [t]} \mathop{\mathbb{E}}_{S,\mathcal{M}}\left[\frac{1}{4}(\hat{p}_{j_i} - p_{j_i})^2\right]\right] \leq \alpha.$$

Our goal now is to prove a lower bound on α.

We will analyze the case where $\mathbf{j}$ has no duplicated indices. So for now we assume that we have a fixed $\mathbf{j}$ with no repeats. For person i we define two test statistics

$$T_i \stackrel{\text{def}}{=} \sum_{\ell \in [t] \setminus \{i\}} (\hat{p}_{j_\ell} - p_{j_\ell})(x^{(i)}_{j_\ell} - p_{j_\ell})$$
$$T_i' \stackrel{\text{def}}{=} \sum_{\ell \in [t] \setminus \{i\}} (\hat{p}'_{j_\ell} - p_{j_\ell})(x^{(i)}_{j_\ell} - p_{j_\ell})$$

where $\hat{p}'_{j_\ell}$ denotes the output of algorithm $\mathcal{M}$ to person ℓ for input $S_{\sim i}$ where $\left(x^{(i)}, j^{(i)}\right)$ has been replaced with a fresh draw $\left(x^{(i)'}, j^{(i)'}\right)$ from P_i. Since the distribution over the indices is deterministic, $j^{(i)} = j^{(i)'} = j_i$.

Now, we will use the privacy guarantee of JDP to provide an upper bound for $\mathop{\mathbb{E}}_{\mathbf{p},S,\mathcal{M}}\left[\sum_{i \in [t]} T_i\right]$. Since $\mathcal{M}$ is (ε, δ)-JDP for any ε, δ in $(0,1)$, we have that $\{\hat{p}_{j_\ell}\}_{\ell \in [t] \setminus i}$ is (ε, δ)-DP with respect to i's dataset. As a result, the following inequality holds

$$\mathbb{E}[T_i] \leq \mathbb{E}\left[T_i'\right] + 2\varepsilon\sqrt{\operatorname{Var}\left(T_i'\right)} + 2\delta\|T_i'\|_\infty. \tag{6}$$

T_i' is a sum of $t-1$ values in $[-4, 4]$, so the bound $\|T_i'\|_\infty \leq 4(t-1) \leq 4t$ holds. For fixed $\mathbf{p}$, $\mathbf{j}$, and ℓ, $\hat{p}'_{j_\ell}$ is independent of $x^{(i)}_{j_\ell}$ and $\mathbb{E}\left[x^{(i)}_{j_\ell}\right] = p_{j_\ell}$. Thus, for any $\ell \in [t] \setminus \{i\}$

$$\begin{aligned}&\mathbb{E}\left[(\hat{p}'_{j_\ell} - p_{j_\ell})(x^{(i)}_{j_\ell} - p_{j_\ell})\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{p}}\left[\mathop{\mathbb{E}}_{S_{\sim i},\mathcal{M}}\left[(\hat{p}'_{j_\ell} - p_{j_\ell})\right]\mathop{\mathbb{E}}_{S,\mathcal{M}}\left[(x^{(i)}_{j_\ell} - p_{j_\ell})\right]\right] = 0.\end{aligned}$$

This means that $\mathbb{E}\left[T_i'\right] = 0$. We apply the same observation, that $\mathbb{E}\left[x^{(i)}_{j_\ell}\right] = p_{j_\ell}$, and that every coordinate is independent to show that the cross terms in the variance of T_i' cancel out, leaving us with

$$\operatorname{Var}\left(T_i'\right) = \mathbb{E}\left[(T_i')^2\right] = \mathbb{E}\left[\sum_{\ell \in [t] \setminus i} (\hat{p}'_{j_\ell} - p_{j_\ell})^2 (x^{(i)}_{j_\ell} - p_{j_\ell})^2\right].$$

Since $(x^{(i)}_{j_\ell} - p_{j_\ell})^2$ is at most 4, we have the upper bound

$$\begin{aligned}
\mathrm{Var}\left(T'_i\right) &= \mathbb{E}\left[\sum_{\ell\in[t]\setminus i} (\hat{p}'_{j_\ell} - p_{j_\ell})^2 (x^{(i)}_{j_\ell} - p_{j_\ell})^2\right] \\
&\leq 4\mathbb{E}\left[\sum_{\ell\in[t]\setminus i} (\hat{p}'_{j_\ell} - p_{j_\ell})^2\right] \\
&\leq 4\mathbb{E}\left[\sum_{\ell\in[t]} (\hat{p}'_{j_\ell} - p_{j_\ell})^2\right].
\end{aligned}$$

Plugging these into inequality 6 and summing up over all the t people, we obtain that

$$\mathbb{E}\left[\sum_{i\in[t]} T_i\right] \leq 4\varepsilon t \sqrt{\mathbb{E}\left[\sum_{\ell\in[t]} (\hat{p}_{j_\ell} - p_{j_\ell})^2\right]} + 8\delta t^2.$$

The next step is to show that accuracy implies a lower bound for $\mathop{\mathbb{E}}_{\mathbf{p},S,\mathcal{M}}\left[\sum_{i\in[t]} T_i\right]$. We notice that in the sum of T_i we can rearrange the terms.

$$\begin{aligned}
\sum_{i\in[t]} T_i &= \sum_{i\in[t]} \sum_{\ell\in[t]\setminus\{i\}} (\hat{p}_{j_\ell} - p_{j_\ell})(x^{(i)}_{j_\ell} - p_{j_\ell}) \\
&= \sum_{\ell\in[t]} \sum_{i\in[t]\setminus\{\ell\}} (\hat{p}_{j_\ell} - p_{j_\ell})(x^{(i)}_{j_\ell} - p_{j_\ell}). \\
&= \sum_{\ell\in[t]} (\hat{p}_{j_\ell} - p_{j_\ell}) \sum_{i\in[t]\setminus\{\ell\}} (x^{(i)}_{j_\ell} - p_{j_\ell}).
\end{aligned}$$

We can now apply Lemma D.6 to each coordinate j_ℓ.

$$\begin{aligned}
\mathbb{E}\left[\sum_{i\in[t]} T_i\right] &= \sum_{\ell\in[t]} \mathbb{E}\left[(\hat{p}_{j_\ell} - p_{j_\ell}) \sum_{i\in[t]\setminus\{\ell\}} (x^{(i)}_{j_\ell} - p_{j_\ell})\right] \\
&\geq \sum_{\ell\in[t]} \left(\frac{1}{3} - \mathbb{E}\left[(\hat{p}_{j_\ell} - p_{j_\ell})^2\right]\right) \\
&= \frac{t}{3} - \mathbb{E}\left[\sum_{\ell\in[t]} (\hat{p}_{j_\ell} - p_{j_\ell})^2\right]
\end{aligned}$$

For a vector of indices $\mathbf{j}$ let $\alpha_{\mathbf{j}}$ be the error

$$\alpha_{\mathbf{j}} = \mathop{\mathbb{E}}_{\mathbf{p},S,\mathcal{M}}\left[\sum_{\ell\in[t]} (\hat{p}_{j_\ell} - p_{j_\ell})^2\right].$$

Combining the two inequalities for $\mathbb{E}\left[\sum_{i\in[t]} T_i\right]$ we have shown that when J has no duplicates

$$\frac{t}{3} - \alpha_{\mathbf{j}} \leq \mathbb{E}\left[\sum_{i\in[t]} T_i\right] \leq 4\varepsilon t \alpha_J + 8\delta t^2.$$

If $\alpha_{\mathbf{j}} \leq \frac{t}{6}$, then $8\delta t^2 < \frac{t}{12}$ because we have assumed that $\delta < \frac{1}{96t}$. By rearranging the terms in the inequality we see that

$$\begin{aligned}\alpha_{\mathbf{j}} &\geq \frac{1}{16\varepsilon^2 t^2}\left(\frac{t}{3} - \alpha_{\mathbf{j}} - 8\delta t^2\right)\\ &\geq \frac{1}{16\varepsilon^2 t^2}\frac{t^2}{12^2} = \frac{1}{2304\varepsilon^2}.\end{aligned}$$

Thus, $\alpha_{\mathbf{j}} \geq \min\left\{\frac{t}{6}, \frac{1}{2304\varepsilon^2}\right\}$.

We now incorporate the randomness over the choice of $\mathbf{j}$. Let E be the event that the vector of target indices has no duplicates. Since $d \geq ct^2$ for a sufficiently large constant c, by Lemma A.7 event E occurs with probability at least $\frac{1}{2}$. Therefore,

$$\begin{aligned}4\alpha &\geq \mathop{\mathbb{E}}_{\mathbf{j}}\left[\alpha_{\mathbf{j}}\right]\\ &\geq \mathop{\mathbb{E}}_{\mathbf{j}}\left[\alpha_{\mathbf{j}} \mid E\right]\mathop{\mathbb{P}}_{\mathbf{j}}[E]\\ &\geq \frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{j}}\left[\alpha_{\mathbf{j}} \mid E\right]\mathop{\mathbb{P}}_{\mathbf{j}}[E].\end{aligned}$$

For each $\mathbf{j}$ without duplicates we have a lower bound on $\alpha_{\mathbf{j}}$, so

$$\alpha \geq \frac{1}{8}\min\left\{\frac{t}{6}, \frac{1}{2304\varepsilon^2}\right\}.$$

□

Corollary 4.8 (Restated)**.** Fix $\alpha > 0$, $\varepsilon \in (0,1]$, $t \in \mathbb{N}$, $\delta \in (0, \frac{1}{8\cdot 144 t})$ and $d \geq ct$ for a sufficiently large constant c. Let ℓ_{mean} be the loss function. Let $\mathcal{M}$ be a billboard algorithm that multitask learns $\mathcal{P}_{d,t}$ with error α with t tasks and one sample per task, and satisfies (ε, δ)-JDP. Then, $\alpha \geq \Omega(\min\{\frac{d}{\varepsilon^2 t^2}, 1\})$.

Proof. By Theorem 3.1 if we have a billboard algorithm for multitask learning with error α, then we have metalearning algorithm with error $e^{\varepsilon}\alpha + \delta$. Then, by Theorem 4.7

$$e^{\varepsilon}\alpha + \delta \geq \min\left\{\frac{1}{4\cdot 144}, \frac{11d}{8^2\cdot 144^2\varepsilon^2 t^2}\right\}.$$

We have made the assumption that $d \geq \frac{16\cdot 144}{11}t$.

If $\frac{11d}{8^2\cdot 144^2\varepsilon^2 t^2} < \frac{1}{4\cdot 144}$, then for $t \leq \frac{11d}{8\cdot 144}$, $\delta < \frac{1}{8\cdot 144 t}$ and $\varepsilon \leq 1$, we have

$$\alpha \geq \frac{11d}{e^{\varepsilon}2\cdot 8^2\cdot 144^2\varepsilon^2 t^2} \geq \frac{11d}{6\cdot 8^2\cdot 144^2\varepsilon^2 t^2}.$$

If $\frac{11d}{8^2\cdot 144^2\varepsilon^2 t^2} \geq \frac{1}{4\cdot 144}$, since $t \geq 1$, $\delta \leq \frac{1}{8\cdot 144 t}$ and $\varepsilon \in [0,1]$, we have that $\alpha \geq \frac{1}{3\cdot 8\cdot 144}$. Therefore,

$$\alpha \geq \Omega\left(\min\left\{1, \frac{d}{\varepsilon^2 t^2}\right\}\right)$$

□

Theorem 4.7 (Restated)**.** Fix $\alpha > 0$, $\varepsilon > 0$, $t \in \mathbb{N}$, $d \in \mathbb{N}$ and $\delta \in (0, \frac{1}{96t})$. Let ℓ_{mean} be the loss function. Let $\mathcal{M} = (\mathcal{M}_{\text{meta}}, \mathcal{M}_{\text{pers}})$ be a pair of algorithms that metalearn a distribution $\mathcal{Q}$ over $\mathcal{P}_{\text{est},d,t+1}$ with error α with t training tasks, one sample per training task and a test task with one personalization sample and $\mathcal{M}_{\text{meta}}$ satisfies (ε, δ)-JDP. Then, $\alpha \geq \Omega(\min\{\frac{d}{\varepsilon^2 t^2}, 1\})$.

Proof. We consider that the metalearning algorithm $\mathcal{M}_{\text{meta}}$ takes as input 1 sample $(x^{(i)}, j^{(i)})$ per person $i \in [t]$. Then the personalization algorithm $\mathcal{M}_{\text{pers}}$ gets as input the output of $\mathcal{M}_{\text{meta}}$ and the sample of the $(t+1)$-th person $(x^{(t+1)}, j^{(t+1)})$ and outputs an estimate $\hat{p}_{j^{(t+1)}}$ of the mean of coordinate $j^{(t+1)}$. By the definition of $\mathcal{P}_{d,t+1}$, the index of person $t+1$ is deterministically j_{t+1}. Hence, we will write $\hat{p}_{j_{t+1}}$ instead of $\hat{p}_{j^{(t+1)}}$ for simplicity.

To prove this theorem we construct a hard metadistribution $\mathcal{Q}$ where we draw a vector of means $\mathbf{p} \in [-1,+1]^d$ and a vector of indices $\mathbf{j} \in [d]^{t+1}$ both uniformly at random. Let the error of $\mathcal{M}$ be

$$\begin{aligned} 4\alpha &\geq \mathop{\mathbb{E}}_{\substack{\mathcal{M}, \\ (P_1,\ldots,P_{t+1})\sim\mathcal{Q}, \\ x^{(1)},\ldots,x^{(t+1)}}} \left[(\hat{p}_{j_{t+1}} - p_{j_{t+1}})^2\right] \\ &= \mathop{\mathbb{E}}_{\substack{\mathcal{M}, \\ \mathbf{p}, j_1,\ldots,j_t, \\ x^{(1)},\ldots,x^{(t+1)}}} \left[\mathop{\mathbb{E}}_{j_{t+1}}\left[(\hat{p}_{j_{t+1}} - p_{j_{t+1}})^2\right]\right] \\ &= \mathop{\mathbb{E}}_{\substack{\mathcal{M}, \\ \mathbf{p}, j_1,\ldots,j_t, \\ x^{(1)},\ldots,x^{(t+1)}}} \left[\frac{1}{d}\sum_{j\in[d]} (\hat{p}_j - p_j)^2\right]. \end{aligned}$$

We consider a tracing attack that uses the following test statistics for $i \in [t]$

$$\begin{aligned} T_i &\overset{\text{def}}{=} \sum_{j\in[d]} (\hat{p}_j - p_j)(x_j^{(i)} - p_j) \text{ and} \\ T_i' &\overset{\text{def}}{=} \sum_{j\in[d]} (\hat{p}_j' - p_j)(x_j^{(i)} - p_j), \end{aligned}$$

where $\hat{p}_j'$ denotes the output of algorithm $\mathcal{M}$ for $j_{t+1} = j$ when the input of person i to $\mathcal{M}_{\text{meta}}$ has been replaced with a fresh draw from P_i. For $i = t+1$ we construct only test statistic

$$T_{t+1} \overset{\text{def}}{=} \sum_{j\in[d]} (\hat{p}_j - p_j)(x_j^{(t+1)} - p_j).$$

Since $\mathcal{M}_{\text{meta}}$ is (ε,δ)-DP with respect to i's dataset for every $i \in [t]$, $\hat{p}_j$ is (ε,δ)-DP with respect to the same dataset for every $j \in [d]$. Therefore,

$$\mathbb{E}[T_i] \leq \mathbb{E}\left[T_i'\right] + 2\varepsilon\sqrt{\operatorname{Var}\left(T_i'\right)} + 2\delta\|T_i'\|_\infty.$$

We will now analyze each term of the right hand side of the inequality. We see that

$$\|T_i'\|_\infty \leq 4d$$

because T_i' is the sum of d entries of value at most 4. Next, since $\hat{p}_j'$ is independent of $x_j^{(i)}$ conditioned on $\mathbf{p}$, we get that

$$\begin{aligned} \mathbb{E}\left[T_i'\right] &= \mathop{\mathbb{E}}_{\mathcal{M},\mathbf{p}}\left[\sum_{j\in[d]} \mathop{\mathbb{E}}_{\substack{\mathcal{M}, j_1,\ldots,j_t \\ x^{(1)},\ldots,x^{(t+1)}, x^{(i)\prime}}} \left[(\hat{p}_j' - p_j)\right] \mathop{\mathbb{E}}_{x^{(i)}}\left[(x_j^{(i)} - p_j)\right]\right] \\ &= 0 \end{aligned}$$

Finally, by the same observation the crosse terms in the variance of T_i' cancel out and we obtain that

$$\begin{aligned}\mathrm{Var}\left(T_i'\right) &= \mathbb{E}\left[(T_i')^2\right]\\ &= \mathbb{E}\left[\sum_{j\in[d]} (\hat{p}_j' - p_j)^2 (x_j^{(i)} - p_j)^2\right]\\ &\le 4\mathbb{E}\left[\sum_{j\in[d]} (\hat{p}_j' - p_j)^2\right] \le 16 d\alpha\end{aligned}$$

Combining the inequalities above we conclude that

$$\mathbb{E}\left[T_i\right] \le 8\varepsilon\sqrt{\alpha}\sqrt{d} + 8\delta d.$$

For the $i = t+1$, we have that

$$\begin{aligned}\mathbb{E}\left[T_{t+1}\right] &= \mathbb{E}\left[\sum_{j\in[d]} (\hat{p}_j - p_j)(x_j^{(t+1)} - p_j)\right]\\ &\le 2\mathbb{E}\left[\sum_{j\in[d]} |\hat{p}_j - p_j|\right]\\ &\le 2\mathbb{E}\left[\sqrt{d}\sqrt{\sum_{j\in[d]} (\hat{p}_j - p_j)^2}\right]\\ &\le 2\sqrt{d}\sqrt{\mathbb{E}\left[\sum_{j\in[d]} (\hat{p}_j - p_j)^2\right]}\\ &\le 4d\sqrt{\alpha}.\end{aligned}$$

Therefore, by summing up the test statistict T_i

$$\mathbb{E}\left[\sum_{i\in[t+1]} T_i\right] \le 8\varepsilon\sqrt{\alpha}t\sqrt{d} + 8\delta d t + 4\sqrt{\alpha}d.$$

The next step is to show that accuracy implies a lower bound for the test statistics in terms of error α. We apply Lemma D.6 to every coordinate $j \in [d]$ of the estimate

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in[t+1]} T_i\right] &= \mathbb{E}\left[\sum_{i\in[t+1]}\sum_{j\in[d]} (\hat{p}_j - p_j)(x_j^{(i)} - p_j)\right]\\ &= \sum_{j\in[d]} \mathbb{E}\left[\sum_{i\in[t+1]} (\hat{p}_j - p_j)(x_j^{(i)} - p_j)\right]\\ &= \sum_{j\in[d]} \mathop{\mathbb{E}}_{j_1,\ldots,j_t}\left[\mathbb{E}\left[(\hat{p}_j - p_j)\sum_{i\in[t+1]} (x_j^{(i)} - p_j)\right]\right]\\ &\ge \sum_{j\in[d]} \left(\frac{1}{3} - \mathbb{E}\left[(\hat{p}_j - p_j)^2\right]\right) = \frac{d}{3} - 4d\alpha.\end{aligned}$$

Combining the bounds on $\mathbb{E}\left[\sum_{i\in[t+1]} T_i\right]$, we have the following inequality

$$\frac{d}{3} - 4d\alpha \leq 8\varepsilon\sqrt{\alpha}t\sqrt{d} + 8\delta dt + 4\sqrt{\alpha}d.$$

If we rearrange the terms we get that

$$\alpha \geq \frac{1}{16\varepsilon^2 t^2 d}\left(\frac{d}{3} - 4\sqrt{\alpha}d - 4\alpha d - 8\delta dt\right)^2.$$

If $\alpha \leq \frac{1}{4\cdot 144}$, then since $\delta \leq \frac{1}{96t}$

$$\begin{aligned}\alpha &\geq \frac{1}{8^2\varepsilon^2 t^2}\left(\frac{d}{3} - \frac{d}{6} - \frac{d}{144} - \frac{d}{12}\right)^2 \\ &\geq \frac{11^2 d}{8^2\cdot 144^2\varepsilon^2 t^2}.\end{aligned}$$

Therefore,

$$\alpha \geq \min\left\{\frac{1}{4\cdot 144}, \frac{11d}{8^2\cdot 144^2\varepsilon^2 t^2}\right\}$$

□

D PROOFS FROM SECTION 5

For the results of this section, we first prove a reduction to and from an estimation problem we call *indexed sign estimation*. We then obtain upper and lower bounds for the error of indexed sign estimation that imply bounds for indexed classification.

Indexed sign estimation problem In the multitask learning version of indexed sign estimation there are t people and every person $i \in [t]$ has n samples drawn from the same distribution P_X over domain $\{\pm 1\}^d$ and an index $j_i \in [d]$. As in Section 4, P_X is a product of d independent distributions over $\{\pm 1\}^d$ with mean $\mathbf{p} \in [-1,+1]^d$. Person i's goal is to estimate the sign $\hat{s}_{j_i}$ of the mean of coordinate j_i. The error for person i is $|p_{j_i}|$ if they make a wrong prediction of the sign of p_{j_i} and 0 otherwise. This means that they get penalized more for the sign of coordinates that have mean further from zero and are, thus, easier to predict. The loss function for person i is

$$\ell_{\text{sign}}(P_X, \hat{s}_{j_i}) := \mathbb{I}\{\text{sign}(p_{j_i}) \neq \hat{s}_{j_i}\}|p_{j_i}|,$$

where p_{j_i} is the true mean of coordinate j_i and $\hat{s}_{j_i}$ is the sign estimate of person i. The overall error for multitask learning is

$$\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\mathbb{I}\{\text{sign}(p_{j_i}) \neq \hat{s}_{j_i}\}|p_{j_i}|\right],$$

where the expectation is taken over the samples of the t people and the randomness of the algorithm. Following the same process as in Section 4 to cast this problem in the "multitask learning" language, we consider that every person draws n samples from a distribution P_i that is in a t-tuple of distributions from family $\mathcal{P}_{\text{est},d,t}$, as defined in Section 4.

We start by showing that we can reduce indexed classification to indexed sign estimation in the multitask learning framework, and vice versa.

Lemma D.1 (Reduction of indexed classification to indexed sign estimation)**.** *Fix $a \in [0,1]$. Let $\mathcal{M}$ be an algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error α, for loss function $\ell + sign$, with t tasks and 1 sample per task. Then, there exists an algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error α, for loss function ℓ_{class}, with t tasks and 1 sample per task.*

Proof. Let $\{(x^{(i)}, j^{(i)}, y^{(i)})\}_{i\in[t]}$ be the input to the multitask classification algorithm that is drawn from a tuple of distributions $(P^{(1)}_{\text{class}}, \dots, P^{(t)}_{\text{class}})$ in $\mathcal{P}_{\text{class},d,t}$. We can transform this dataset to an input dataset for algorithm $\mathcal{M}$. For person i we can use $j^{(i)} = j_i$ to recover the value of the auxiliary random variable $w^{(i)}$ as follows

$$w^{(i)}_\ell \leftarrow \begin{cases} x^{(i)}_\ell, & \text{if } \ell \neq j^{(i)}, \\ x^{(i)}_\ell y^{(i)}, & \text{if } \ell = j^{(i)}. \end{cases}$$

We see that the tuple of the t distributions $(P^{(1)}_{\text{est}}, \dots, P^{(t)}_{\text{est}})$ over $(w^{(i)}, j^{(i)})$ for all $i \in [t]$ is in $\mathcal{P}_{\text{est},d,t}$. For this reason, we give $\{(w^{(i)}, j^{(i)})\}_{i\in[t]}$ as input to algorithm $\mathcal{M}$.

Let $\hat{s}_{j_i}$ be the output of $\mathcal{M}$ for person i. Then, for person $i \in [t]$ the classification algorithm ouputs the labeling function

$$\hat{f}_i(x_{j_i}) = \hat{s}_{j_i} x_{j_i}.$$

The misclassification error of $\hat{f}_i$ for distribution $P^{(i)}_{\text{class}}$ is

$$\begin{aligned}
\mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[\hat{f}_i(x_j) \neq y\right] &= \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[\hat{s}_{j_i} x_{j_i} \neq y\right] \\
&= \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[\hat{s}_{j_i} \neq y x_{j_i}\right] \\
&= \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[\hat{s}_{j_i} \neq w_{j_i}\right] \\
&= \mathop{\mathbb{P}}_{(w,j)\sim P^{(i)}_{\text{est}}}\left[\hat{s}_{j_i} \neq w_{j_i}\right] \\
&= \mathbb{I}\{\hat{s}_{j_i} = 1\} \mathop{\mathbb{P}}_{(w,j)\sim P^{(i)}_{\text{est}}}\left[w_{j_1} = -1\right] + \mathbb{I}\{\hat{s}_{j_i} = -1\} \mathop{\mathbb{P}}_{(w,j)\sim P^{(i)}_{\text{est}}}\left[w_{j_1} = 1\right] \\
&= \frac{1 - p_{j_i}}{2}\mathbb{I}\{\hat{s}_{j_i} = 1\} + \frac{1 + p_{j_i}}{2}\mathbb{I}\{\hat{s}_{j_i} = -1\}.
\end{aligned}$$

The misclassification error of a fixed classifier f_i for distribution $P^{(i)}_{\text{class}}$ is

$$\begin{aligned}
\mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i(x_j) \neq y\right] &= \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i(x_j) = -1 \mid x_{j_i} = 1, y = 1\right]\mathbb{P}\left[x_{j_i} = 1, y = 1\right] \\
&+ \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i(x_j) = 1 \mid x_{j_i} = 1, y = -1\right]\mathbb{P}\left[x_{j_i} = 1, y = -1\right] \\
&+ \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i(x_j) = -1 \mid x_{j_i} = -1, y = 1\right]\mathbb{P}\left[x_{j_i} = -1, y = 1\right] \\
&+ \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i(x_j) = 1 \mid x_{j_i} = -1, y = -1\right]\mathbb{P}\left[x_{j_i} = -1, y = -1\right] \\
&= \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i(x_j) = -1 \mid x_{j_i} = 1, y = 1\right]\frac{1 + p_{j_i}}{4} \\
&+ \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i(x_j) = 1 \mid x_{j_i} = 1, y = -1\right]\frac{1 - p_{j_i}}{4} \\
&+ \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i(x_j) = -1 \mid x_{j_i} = -1, y = 1\right]\frac{1 - p_{j_i}}{4} \\
&+ \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i(x_j) = 1 \mid x_{j_i} = -1, y = -1\right]\frac{1 + p_{j_i}}{4}.
\end{aligned}$$

Given p_{j_i}, the function that minimizes this error is $f_i^*(x_{j_i}) = \mathrm{sign}(p_{j_i})x_{j_i}$. The misclassification error of f_i^* for distribution $P^{(i)}_{\text{class}}$ is

$$\begin{aligned}\mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i^*(x_j) \neq y\right] &= \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[\mathrm{sign}(p_{j_i})x_{j_i} \neq y\right] \\ &= \mathop{\mathbb{P}}_{(w,j)\sim P^{(i)}_{\text{est}}}\left[\mathrm{sign}(p_{j_i}) \neq w_{j_i}\right] \\ &= \mathbb{I}\{\mathrm{sign}(p_{j_i}) = 1\}\frac{1-p_{j_i}}{2} + \mathbb{I}\{\mathrm{sign}(p_{j_i}) = -1\}\frac{1+p_{j_i}}{2}.\end{aligned}$$

Taking the difference between the misclassification error of $\hat{f}_i$ and $f*_i$, we obtain that

$$\begin{aligned}\mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[\hat{f}_i(x_j) \neq y\right] - \mathop{\mathbb{P}}_{(x,j,y)\sim P^{(i)}_{\text{class}}}\left[f_i^*(x_j) \neq y\right] &= \frac{1-p_{j_i}}{2}\mathbb{I}\{\hat{s}_{j_i} = 1\} + \frac{1+p_{j_i}}{2}\mathbb{I}\{\hat{s}_{j_i} = -1\} \\ &\quad - \mathbb{I}\{\mathrm{sign}(p_{j_i}) = 1\}\frac{1-p_{j_i}}{2} - \mathbb{I}\{\mathrm{sign}(p_{j_i}) = -1\}\frac{1+p_{j_i}}{2} \\ &= \mathbb{I}\{\hat{s}_{j_i} = 1, \mathrm{sign}(p_{j_i}) = -1\}(-p_{j_i}) + \mathbb{I}\{\hat{s}_{j_i} = -1, \mathrm{sign}(p_{j_i}) = 1\}p_{j_i} \\ &= \mathbb{I}\{\hat{s}_{j_i} \neq \mathrm{sign}(p_{j_i})\}|p_{j_i}|.\end{aligned}$$

Therefore, if we take the expectation of the error over the samples $\{(x^{(i)}, j^{(i)}, y^{(i)})\}_{i\in[t]}$ and the randomness of algorithm $\mathcal{M}$ we have that the error of multitask learning for indexed classification is

$$\begin{aligned}\frac{1}{t}\sum_{i\in[t]}\mathbb{E}\left[\left(\mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[\hat{f}_i(x_{j_i}) \neq y\right] - \min_{f_i}\mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[f_i(x_{j_i}) \neq y\right]\right)\right] &\leq \frac{1}{t}\sum_{i\in[t]}\mathbb{E}\left[\mathbb{I}\{\hat{s}_{j_i} \neq \mathrm{sign}(p_{j_i})\}|p_{j_i}|\right] \\ &\leq \alpha.\end{aligned}$$

This concludes our proof. □

Lemma D.2 (Reduction of indexed sign estimation to indexed classification)**.** *Fix $\alpha \in [0,1]$. Let $\mathcal{M}$ be an algorithm that multitask learns $\mathcal{P}_{class,d,t}$ with error α, for loss function ℓ_{class}, with t tasks and* 1 *sample per task. Then, there exists an algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error α, for loss function ℓ_{est}, with t tasks and* 2 *samples per task.*

Proof. Let $\{(x^{(i,1)}, j^{(i,1)}), (x^{(i,2}, j^{(i,2)})\}$ be the dataset of task $i \in [t]$ gives to the indexed estimation algorithm. Every sample $(x^{(i,k)}, j^{(i,k)})$, for $k \in \{1,2\}$, is drawn from the corresponding distribution $P^{(i)}_{\text{est}}$ from a tuple of t distributions $(P^{(1)}_{\text{est}}, \ldots, P^{(t)}_{\text{est}}) \in \mathcal{P}_{\text{est},d,t}$.

We first use one sample from every task/person to construct an input for algorithm $\mathcal{M}$. For every person i we set $w^{(i,1)} \leftarrow x^{(i,1)}$, draw a $y^{(i,1)} \in \{\pm 1\}$ uniformly and set

$$\tilde{x}^{(i,1)}_\ell \leftarrow \begin{cases} w^{(i,1)}_\ell, & \text{if } \ell \neq j_i \\ w^{(i,1)}_\ell y^{(i,1)}, & \text{if } \ell = j_i. \end{cases}$$

We then give $\{(\tilde{x}^{(i,1)}, j^{(i,1)}, y^{(i,1)})\}_{i\in[t]}$ as input to the classification algorithm $\mathcal{M}$. Let $P^{(i)}_{\text{class}}$ be the distribution of $(\tilde{x}^{(i,1)}, j^{(i,1)}, y^{(i,1)})$. We notice that $(P^{(1)}_{\text{class}}, \ldots, P^{(t)}_{\text{class}})$ is in the family of tuples $\mathcal{P}_{\text{class},d,t}$.

Let $\hat{f}_1, \ldots \hat{f}_t$ be the functions that $\mathcal{M}$ outputs for this input. Then, we use the second sample of person i, $(x^{(i,2}, j^{(i,2)})$, and function $\hat{f}_i$ to get the sign estimate $\hat{s}_{j_i}$. In particular, we draw $y^{(i,2)} \in \{\pm 1\}$ uniformly, we set $w^{(i,2)} \leftarrow x^{(i,2)}$ and

$$\tilde{x}^{(i,2)}_\ell \leftarrow \begin{cases} w^{(i,2)}_\ell, & \text{if } \ell \neq j_i \\ w^{(i,2)}_\ell y^{(i,1)}, & \text{if } \ell = j_i. \end{cases}$$

We see that $P^{(i)}_{\text{class}}$ is also the distribution of $(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})$. For person i this indexed estimation algorithm outputs

$$\hat{s}_{j_i} = \hat{f}_i(\tilde{x}^{(i,2)}_{j_i})\tilde{x}^{(i,2)}_{j_i}.$$

The next step is to compute the error of $\hat{s}_{j_i}$. In the proof of Lemma D.1 we showed that for person i

$$\min_{f_i} \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{\text{class}}}{\mathbb{P}} \left[f_i(\tilde{x}^{(i,2)}_{j_i}) \neq y^{(i,2)}\right] = \mathbb{I}\{\text{sign}(p_{j_i}) = 1\}\frac{1-p_{j_i}}{2} + \mathbb{I}\{\text{sign}(p_{j_i}) = -1\}\frac{1+p_{j_i}}{2}.$$

The misclassification error of $\hat{f}_i$ is

$$\begin{aligned}
\underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} \left[\hat{f}_i(\tilde{x}^{(i,2)}_{j_i}) \neq y^{(i,2)}\right] &= \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} \left[\hat{f}_i(\tilde{x}^{(i,2)}_{j_i})\tilde{x}^{(i,2)}_{j_i} \neq \tilde{x}^{(i,2)}_{j_i} y^{(i,2)}\right] \\
&= \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} \left[\hat{s}_i \neq w^{(i,2)}_{j_i}\right] \\
&= \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} \left[\hat{s}_i \neq 1 | w^{(i,2)}_{j_i} = 1\right] \mathbb{P}\left[w^{(i,2)}_{j_i} = 1\right] \\
&+ \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} \left[\hat{s}_i \neq -1 | w^{(i,2)}_{j_i} = -1\right] \mathbb{P}\left[w^{(i,2)}_{j_i} = -1\right].
\end{aligned}$$

We can see that $\tilde{x}^{(i,2)}_{j_i}$ is independent of $w^{(i,2)}_{j_i}$. Thus, $\hat{s}_{j_i}$ is independent of $w^{(i,2)}_{j_i}$ and we have that

$$\begin{aligned}
\underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} \left[\hat{f}_i(\tilde{x}^{(i,2)}_{j_i}) \neq y^{(i,2)}\right] &= \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} [\hat{s}_i \neq 1]\frac{1+p_{j_i}}{2} + \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} [\hat{s}_i \neq -1]\frac{1-p_{j_i}}{2} \\
&= \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{E}} \left[\mathbb{I}\{\hat{s}_i \neq 1\}\frac{1+p_{j_i}}{2} + \mathbb{I}\{\hat{s}_i \neq -1\}\frac{1-p_{j_i}}{2}\right]
\end{aligned}$$

Combining the these steps, we get that the difference of the misclassification error between $\hat{f}_i$ and the optimal classifier is

$$\begin{aligned}
&\underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} \left[\hat{f}_i(\tilde{x}^{(i,2)}_{j_i}) \neq y^{(i,2)}\right] - \min_{f_i} \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{C}}{\mathbb{P}} \left[f_i(\tilde{x}^{(i,2)}_{j_i}) \neq y^{(i,2)}\right] \\
&= \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{E}} \left[\mathbb{I}\{\hat{s}_i \neq 1\}\frac{1+p_{j_i}}{2} + \mathbb{I}\{\hat{s}_i \neq -1\}\frac{1-p_{j_i}}{2} - \mathbb{I}\{\text{sign}(p_{j_i}) = 1\}\frac{1-p_{j_i}}{2} - \mathbb{I}\{\text{sign}(p_{j_i}) = -1\}\frac{1+p_{j_i}}{2}\right] \\
&= \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{E}} \left[\mathbb{I}\{\hat{s}_i \neq \text{sign}(p_{j_i})|p_{j_i}|\}\right].
\end{aligned}$$

The average error of the sign estimation algorithm, in expectation over the samples, the randomness of the input/output transformations and the randomness of algorithm $\mathcal{M}$ is

$$\begin{aligned}
&\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\mathbb{I}\{\text{sign}(p_{j_i}) \neq \hat{s}_{j_i}\}|p_{j_i}|\right] \\
&= \frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{\text{class}}}{\mathbb{E}} \left[\mathbb{I}\{\text{sign}(p_{j_i}) \neq \hat{s}_{j_i}\}|p_{j_i}|\right]\right] \\
&= \frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{class}}{\mathbb{P}} \left[\hat{f}_i(\tilde{x}^{(i,2)}_{j_i}) \neq y^{(i,2)}\right] - \min_{f_i} \underset{(\tilde{x}^{(i,2)}, j^{(i,2)}, y^{(i,2)})\sim P^{(i)}_{\text{class}}}{\mathbb{P}} \left[f_i(\tilde{x}^{(i,2)}_{j_i}) \neq y^{(i,2)}\right]\right] \\
&\leq \alpha.
\end{aligned}$$

□

D.1 Upper Bounds

Here we present the proofs of the upper bounds of the error of indexed classification for the different private personalization frameworks.

Proposition 5.1 (Restated)**.** For any $d \in \mathbb{N}$ and $t \in \mathbf{N}$, and loss function ℓ_{class}, there exists an algorithm that multitask learns $\mathcal{P}_{\text{class},d,t}$ with error $\frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task.

Proof. Fix any $d \in \mathbb{N}$, $t \in \mathbb{N}$. By Proposition 4.1 there is an algorithm $\mathcal{M}$ that multitask learns $\mathcal{P}_{\text{est},d,t}$ with error

$$\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[(\hat{p}_{j_i} - p_{j_i})^2\right] \le \frac{1}{t},$$

where $\hat{p}_{j_i}$ is the output of algorithm $\mathcal{M}$ to person i, with t tasks and 1 sample per task. We can use $\mathcal{M}$ to estimate the signs of the p_{j_i}s by setting $\hat{s}_{j_i} = \text{sign}(\hat{p}_{j_i})$. In this case the error for indexed sign estimation is

$$\begin{aligned}
\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\mathbb{I}\{\hat{s}_{j_i} \neq \text{sign}(p_{j_i})\}|p_{j_i}|\right] &\le \frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\mathbb{I}\{\hat{s}_{j_i} \neq \text{sign}(p_{j_i})\}|\hat{p}_{j_i} - p_{j_i}|\right] \\
&\le \frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\mathbb{I}\{\hat{s}_{j_i} \neq \text{sign}(p_{j_i})\}|\hat{p}_{j_i} - p_{j_i}| + \mathbb{I}\{\hat{s}_{j_i} = \text{sign}(p_{j_i})\}|\hat{p}_{j_i} - p_{j_i}|\right] \\
&= \frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[|\hat{p}_{j_i} - p_{j_i}|\right] \\
&\le \frac{1}{t}\mathbb{E}\left[\sqrt{\sum_{i\in[t]}(\hat{p}_{j_i} - p_{j_i})^2}\sqrt{t}\right] \\
&\le \sqrt{\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[(\hat{p}_{j_i} - p_{j_i})^2\right]} \\
&\le \frac{1}{\sqrt{t}}.
\end{aligned}$$

By applying the reduction of Lemma D.1 we obtain that for all $(P_1, \dots, P_t) \in \mathcal{P}_{\text{class},d,t}$

$$\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\left(\mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[\hat{f}_i(x_{j_i}) \neq y\right] - \min_{f_i} \mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[f_i(x_{j_i}) \neq y\right]\right)\right] \le \frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\mathbb{I}\{\hat{s}_{j_i} \neq \text{sign}(p_{j_i})\}|p_{j_i}|\right] \le \frac{1}{\sqrt{t}}.$$

□

Theorem 5.2 (Restated)**.** For any parameters $\rho \ge 0$, $d \in \mathbb{N}$, $t \in \mathbb{N}$ and for loss function ℓ_{class}, there exists an algorithm that multitask learns $\mathcal{P}_{\text{class},d,t}$ with error $\frac{\sqrt{2}}{t\sqrt{\rho}} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task and satisfies ρ-1-out-of-t-zCDP. Moreover, it is implied that for any parameters $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, $t \in \mathbb{N}$ and for loss function ℓ_{class}, there exists an algorithm that multitask learns $\mathcal{P}_{\text{class},d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1/2}) \cdot \frac{1}{t} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies (ε,δ)-1-out-of-t-DP.

Proof. Fix any $d \in \mathbb{N}$, $t \in \mathbb{N}$, and $\rho \ge 0$. By Theorem 4.2 there exists a ρ-1-out-of-t-zCDP algorithm $\mathcal{M}$ that multitask learns $\mathcal{P}_{\text{est},d,t}$ with error

$$\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[(\hat{p}_{j_i} - p_{j_i})^2\right] \le \frac{2}{\rho t^2} + \frac{1}{t},$$

with t tasks and 1 sample per task. In this algorithm the computes a mean estimate $\hat{p}_{j_i}$ that they send to person i. For the classification learning algorithm the curator can use $\hat{p}_{j_i}$ to construct an estimate of the sign of p_{j_i} by setting $\hat{s}_{j_i} = \mathrm{sign}(\hat{p}_{j_i})$ before sending anything to the people. Then, using the reduction of Lemma D.2 we get a function that the server can send to person i, by setting

$$\hat{f}_i(x,j) = \hat{s}_{j_i} x_{j_i}.$$

As $\hat{f}_i$ is a post-processing of $\hat{p}_{j_i}$ that only uses j_i, which is already in the dataset of person i, the output of this algorithm remains ρ-1-out-of-t-zCDP. By Lemma D.1 we obtain that the error of of $\hat{f}_1, \ldots, \hat{f}_t$ is

$$\begin{aligned}\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\left(\mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[\hat{f}_i(x_{j_i}) \neq y\right] - \min_{f_i} \mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[f_i(x_{j_i}) \neq y\right]\right)\right] &\le \frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\mathbb{I}\{\hat{s}_{j_i} \neq \mathrm{sign}(p_{j_i})\}|p_{j_i}|\right] \\ &\le \sqrt{\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[(\hat{p}_{j_i} - p_{j_i})^2\right]} \\ &\le \sqrt{\frac{2}{\rho t^2} + \frac{1}{t}} \\ &\le \sqrt{\frac{2}{\rho t^2}} + \sqrt{\frac{1}{t}}.\end{aligned}$$

□

Theorem 5.3 (Restated)**.** For any parameters $\rho \ge 0$, $d \in \mathbb{N}$, $t \in \mathbb{N}$, and for loss function ℓ_{class}, there exists an algorithm that multitask learns $\mathcal{P}_{\text{class},d,t}$ with error $O(\frac{1}{\sqrt{\rho t}}) + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies ρ-JCDP. Moreover, it is implied that for any $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, $t \in \mathbb{N}$, and for loss function ℓ_{class}, there exists an algorithm that multitask learns $\mathcal{P}_{\text{class},d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1/2}) \cdot \frac{1}{\sqrt{t}} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies (ε,δ)-JDP.

Proof. Fix any $d \in \mathbb{N}$, $t \in \mathbb{N}$ and $\rho \ge 0$. Similarly to the proof above we can base our algorithm on the algorithm of Theorem 4.3. Our classification algorithm takes the mean estimate $\hat{p}_{j_i}$ for person i and produces

$$\hat{f}_i(x_{j_i}) = \hat{s}_{j_i} x_{j_i},$$

with error

$$\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[\left(\mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[\hat{f}_i(x_{j_i}) \neq y\right] - \min_{f_i} \mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[f_i(x_{j_i}) \neq y\right]\right)\right] \le \sqrt{\frac{1}{t}\sum_{i\in[t]} \mathbb{E}\left[(\hat{p}_{j_i} - p_{j_i})^2\right]} \le \frac{\sqrt{2(t-1)}}{t\sqrt{\rho}} + \frac{1}{\sqrt{t}}.$$

The new algorithm also satisfies ρ-JCDP because the curator just post-processes the output of the ρ-JCDP mean estimation multitask learning algorithm using only j_i for person i, which is their personal information.

□

Theorem 5.4 (Restated)**.** For any parameters $\rho \ge 0$, $d \in \mathbb{N}$, $t \in \mathbb{N}$, and for loss function ℓ_{class}, there exists a billboard algorithm that multitask learns $\mathcal{P}_{\text{class},d,t}$ with error $\frac{\sqrt{2d}}{t\sqrt{\rho}} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies ρ-zCDP. Moreover, it is implied that for any parameters $\varepsilon > 0$, $\delta \in (0,1)$, $d \in \mathbb{N}$, $t \in \mathbb{N}$, and for loss function ℓ_{class}, there exists a billboard algorithm that multitask learns $\mathcal{P}_{\text{class},d,t}$ with error $\Theta(\min(\varepsilon, \varepsilon^2/\log(1/\delta))^{-1/2}) \cdot \frac{\sqrt{d}}{t} + \frac{1}{\sqrt{t}}$ with t tasks and 1 sample per task, and satisfies (ε,δ)-DP.

Proof. By Theorem 4.4 there exists a billboard algorithm $(\mathcal{M}_{\text{BB}}, \mathcal{M}_{\text{pers}})$ that multitask learns $\mathcal{P}_{\text{est},d,t}$ with t tasks and 1 sample per task, and $\mathcal{M}_{\text{BB}}$ is ρ-zCDP. Our classification algorithm keeps the billboard algorithm as it is and modifies the personalization part. In particular, person i takes their mean estimate $\hat{p}_{j_i}$ and produces

$$\hat{f}_i(x_{j_i}) = \hat{s}_{j_i} x_{j_i},$$

where $\hat{s}_{j_i} = \mathrm{sign}(\hat{p}_{j_i})$. By applying the reduction of D.2 we get error

$$\frac{1}{t}\sum_{i\in[t]}\mathbb{E}\left[\left(\mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[\hat{f}_i(x_{j_i}) \neq y\right] - \min_{f_i}\mathop{\mathbb{P}}_{(x,j,y)\sim P_i}\left[f_i(x_{j_i}) \neq y\right]\right)\right] \le \sqrt{\frac{1}{t}\sum_{i\in[t]}\mathbb{E}\left[(\hat{p}_{j_i} - p_{j_i})^2\right]} \le \frac{\sqrt{2d}}{t\sqrt{\rho}} + \frac{1}{\sqrt{t}}.$$

As we mentioned, the billboard part of the new algorithm remains the same. Thus, it still satisfies ρ-zCDP. □

D.2 Our Fingerprinting Lemma

Our error lower bounds for indexed classification are based on a generalization of the fingerprinting lemma by Bun et al. (2014). Our lemma differs in two ways; the mean p is drawn uniformly at random from an interval $[-\alpha,\alpha]$ for $\alpha \in (0,1]$ (instead of $[-1,1]$) and the error function is $2(|p| - f(x))p = 4|p|\mathbb{I}\{f(x) \neq \mathrm{sign}(p)\}$ (instead of the mean-squared-error). The key idea is that for a fixed α a function f cannot be bad at predicting the sign of the mean p and highly correlated with the sum of samples drawn from a distribution with mean p at the same time.

Lemma D.3 (Our Fingerprinting Lemma)**.** *Let $f : \{\pm1\}^t \to [\pm1]$ and $\alpha \in (0,1]$. If $p \in [-\alpha,+\alpha]$ is sampled uniformly at random and $x_1,\ldots,x_t \in \{\pm1\}^t$ are sampled independently with mean p, then*

$$\mathop{\mathbb{E}}_{\substack{p,\\ x_1,\ldots,x_t}}\left[r(\alpha,p)f(x)\sum_{i\in[t]}(x_i - p) + 2(|p| - f(x)p)\right] = \alpha,$$

where

$$r(\alpha,p) = \begin{cases}\frac{\alpha^2-p^2}{1-p^2}, & \text{if } \alpha^2 \neq 1,\\ 1, & \text{otherwise.}\end{cases}$$

We break the proof of our fingerprinting lemma into smaller lemmas following the proof in Bun et al. (2017).

Lemma D.4. *Let $f : \{\pm1\}^t \to \mathbb{R}$. Define $g : [\pm1] \to \mathbb{R}$ by*

$$g(p) = \mathop{\mathbb{E}}_{x_1,\ldots,x_t\sim p}[f(x)].$$

Then,

$$\mathop{\mathbb{E}}_{x_1,\ldots,x_t}\left[f(x)\sum_{i\in[t]}(x_i - p)\right] = g'(p)(1-p)^2.$$

Proof. We can write

$$g(p) = \mathop{\mathbb{E}}_{x_1,\ldots,x_t}[f(x)] = \sum_{x_1,\ldots,x_t\in\{\pm1\}^t} f(x)\prod_{i\in[t]}\frac{1+x_ip}{2},$$

where $x_1,\ldots,x_t \in \{\pm1\}^t$ are sampled independently with mean p. We can now compute its derivative

$$\begin{aligned} g'(p) &= \sum_{x_1,\ldots,x_t\in\{\pm1\}^t} f(x)\frac{d}{dp}\prod_{i\in[t]}\frac{1+x_ip}{2} \\ &= \sum_{x_1,\ldots,x_t\in\{\pm1\}^t} f(x)\sum_{i\in[t]}\frac{x_i - p}{1-p^2}\prod_{j\in[t]\frac{1+x_jp}{2}} \\ &= \mathop{\mathbb{E}}_{x_1,\ldots,x_t}\left[f(x)\sum_{i\in[t]}\frac{x_i - p}{1-p^2}\right] \end{aligned}$$

□

Lemma D.5. *Let $g : [\pm 1] \to \mathbb{R}$ be a polynomial and $\alpha \in (0,1]$. If $p \in [-\alpha, +\alpha]$ is drawn uniformly at random, then*

$$\mathop{\mathbb{E}}_{p}\Big[r(\alpha,p)g'(p)(1-p^2)\Big] = 2\mathop{\mathbb{E}}_{p}[g(p)p],$$

where

$$r(\alpha,p) = \begin{cases} \frac{\alpha^2-p^2}{1-p^2}, & \text{if } \alpha^2 \neq 1, \\ 1, & \text{otherwise.} \end{cases}$$

Proof. If $\alpha^2 \neq 1$,

$$\begin{aligned} &\mathop{\mathbb{E}}_{p}\Big[r(\alpha,p)g'(p)(1-p^2)\Big] \\ &= \frac{1}{2\alpha}\int_{-\alpha}^{\alpha} r(\alpha,p)g'(p)(1-p^2)dp \\ &= \frac{1}{2\alpha}\int_{-\alpha}^{\alpha} (\alpha^2-p^2)g'(p)dp \\ &= \frac{1}{2\alpha}\int_{-\alpha}^{\alpha} \frac{d}{dp}\Big[(\alpha^2-p^2)g(p)\Big] - g(p)(-2p)dp \\ &= \frac{1}{2\alpha}\int_{-\alpha}^{\alpha} 2g(p)pdp = 2\mathop{\mathbb{E}}_{p}[g(p)p]. \end{aligned}$$

Similarly, we can show that when $\alpha^2 = 1$

$$\mathop{\mathbb{E}}_{p}\Big[g'(p)(1-p^2)\Big] = 2\mathop{\mathbb{E}}_{p}[g(p)p].$$

□

Proof of Lemma D.3. Applying Lemmas D.4 and D.5 we have

$$\begin{aligned} &\mathop{\mathbb{E}}_{p,x_1,\ldots,x_t}\left[r(\alpha,p)f(x)\sum_{i\in[t]}(x_i-p) + 2(|p| - f(x)p)\right] \\ &= \mathop{\mathbb{E}}_{p}\left[r(\alpha,p)\mathop{\mathbb{E}}_{x_1,\ldots,x_t}\left[f(x)\sum_{i\in[t]}(x_i-p)\right]\right] + \mathop{\mathbb{E}}_{p,x_1,\ldots,x_t}[2(|p|-f(x)p)] \\ &= \mathop{\mathbb{E}}_{p,x_1,\ldots,x_t}[2f(x)p] + \mathop{\mathbb{E}}_{p,x_1,\ldots,x_t}[2(|p|-f(x)p)] \\ &= \mathop{\mathbb{E}}_{p}[2|p|] = \alpha \end{aligned}$$

□

We can show that the fingerprinting lemma in Bun et al. (2014) follows from our fingerprinting lemma for $\alpha = 1$.

Lemma D.6 (Fingerprinting Lemma Bun et al. (2014))**.** *Let $f : \{\pm 1\}^t \to [\pm 1]$. If $p \in [-1,+1]$ is sampled uniformly at random and $x_1,\ldots,x_t \in \{\pm 1\}^t$ are sampled independently with mean p, then*

$$\mathop{\mathbb{E}}_{p,x}\left[(f(x)-p)\sum_{i\in[t]}(x_i-p) + (f(x)-p)^2\right] \geq \frac{1}{3}.$$

Proof. We have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{p,x_1,\dots,x_t}\left[(f(x)-p)\sum_{i\in[t]}(x_i-p)+(f(x)-p)^2\right]\\
&=\mathop{\mathbb{E}}_{p,x_1,\dots,x_t}\left[f(x)\sum_{i\in[t]}(x_i-p)+f(x)^2-2f(x)p\right]+\mathop{\mathbb{E}}_{p}\left[p^2-p\mathop{\mathbb{E}}_{x_1,\dots,x_t}\left[\sum_{i\in[t]}(x_i-p)\right]\right]\\
&\geq\mathop{\mathbb{E}}_{p,x_1,\dots,x_t}\left[f(x)\sum_{i\in[t]}(x_i-p)-2f(x)p\right]+\mathop{\mathbb{E}}_{p}[2|p|]-\frac{2}{3}\\
&=\mathop{\mathbb{E}}_{p,x_1,\dots,x_t}\left[f(x)\sum_{i\in[t]}(x_i-p)+2(|p|-f(x)p)\right]-\frac{2}{3}\\
&\geq 1-\frac{2}{3}=\frac{1}{3}.
\end{aligned}$$

The last inequality follows from Lemma D.3. □

D.3 Lower Bounds

In this section we provide the proofs for the lower bounds of the error of indexed classification in the private personalization frameworks we consider. We first prove lower bounds for indexed sign estimation and then prove the corresponding lower bounds for indexed classification by using the reduction of Lemma D.2.

Lemma D.7 (Indexed sign estimation JDP lower bound)**.** *Fix parameters $\alpha\in(0,1/16)$, $t\in\mathbb{N}$, $\varepsilon>0$, $\delta\in(0,\frac{1}{2t})$, and let ℓ_{sign} be the loss function. Let $\mathcal{M}$ be an algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error α, for loss function ℓ_{sign}, with t tasks and 2 samples per task and satisfies (ε,δ)-JDP. If $d\geq ct^2$ for a sufficiently large constant c, then $\alpha\geq\Omega(\min\{\frac{\sqrt{t}}{\varepsilon},1\})$.*

Proof. Fix parameters $\alpha\in(0,1/16)$, $t\in\mathbb{N}$, $\varepsilon>0$, $\delta\in(0,\frac{1}{2t})$. Let $\mathcal{M}$ be a JDP algorithm that gets as input the d-dimensional samples of t tasks/people. Each person $i\in[t]$, gives samples $(x^{(i,1)},j^{(i,1)}),(x^{(i,2)},j^{(i,2)})$ to the algorithm. We start by constructing a hard distribution over the family of distributions $\mathcal{P}_{\text{est},d,t}$ that the samples are drawn from. We draw a vector $\mathbf{p}\in[-\lambda,\lambda]^d$ for $\lambda=16\alpha$ and a vector $\mathbf{j}\in[d]^t$, both uniformly. The k-th datapoint $x^{(i,k)}$ of person i is drawn from the product distribution with mean $\mathbf{p}$ and the index $j^{(i,k)}$ is deterministically j_i.

$\mathcal{M}$ returns an estimate $\hat{s}_{j_i}$ of the sign of p_{j_i} to each person $i\in[t]$. If the expected error of $\mathcal{M}$ is

$$\mathop{\mathbb{E}}_{\mathbf{p},\mathbf{j}}\left[\frac{1}{t}\sum_{i\in[t]}\mathbb{E}\left[\mathbb{I}\{\text{sign}(p_{j_i})\neq\hat{s}_{j_i}\}|p_{j_i}|\right]\right]\leq\alpha.$$

our goal is to prove a lower bound on α.

We will first analyze the case where $\mathbf{j}$ has no duplicated indices. So for now we assume that we have a fixed $\mathbf{j}$ with no repeats. For person $i\in[t]$ and sample $k\in\{1,2\}$ we define two test statistics

$$\begin{aligned}
T_{i,k}&\stackrel{\text{def}}{=}\sum_{\ell\in[t]\setminus\{i\}}\frac{\lambda^2-p_{j_\ell}^2}{1-p_{j_\ell}^2}\hat{s}_{j_\ell}(x_{j_\ell}^{(i,k)}-p_{j_\ell})\\
T'_{i,k}&\stackrel{\text{def}}{=}\sum_{\ell\in[t]\setminus\{i\}}\frac{\lambda^2-p_{j_\ell}^2}{1-p_{j_\ell}^2}\hat{s}_{j_\ell}^{(i,k)}(x_{j_\ell}^{(i,k)}-p_{j_\ell})
\end{aligned}$$

where $\hat{s}_{j_\ell}^{(i,k)}$ denotes the output of algorithm $\mathcal{M}$ to person ℓ when the input $\left(x^{(i,k)}, j^{(i,k)}\right)$ if person i has been replaced with a fresh draw from P_i. Since the distribution over the indices is deterministic, $j^{(i,k)}$ is always j_i.

Now, we will use the privacy guarantee of JDP to provide an upper bound for $\mathbb{E}\left[\sum_{i\in[t]}(T_{i,1}+T_{i,2})\right]$. Since $\mathcal{M}$ is (ε,δ)-JDP, we have that $\{\hat{s}_{j_\ell}\}_{\ell\in[t]\setminus i}$ is (ε,δ)-DP with respect to i's dataset. As a result, the following inequality holds

$$\mathbb{E}\left[T_{i,k}\right] \le \mathbb{E}\left[T_{i,k}{}'\right] + 2\varepsilon\sqrt{\mathrm{Var}\left(T_{i,k}{}'\right)} + 2\delta\|T'_{i,k}\|_\infty.$$

T'_i is a sum of $t-1$ values that are at most $2\lambda^2$, so the bound $\|T'_i\|_\infty \le 2\lambda^2(t-1) \le 2\lambda^2 t$ holds. For fixed $\mathbf{p}$, and ℓ, $\hat{s}_{j_\ell}^{(i,k)}$ is independent of $x_{j_\ell}^{(i,k)}$ and $\mathbb{E}\left[x_{j_\ell}^{(i,k)}\right] = p_{j_\ell}$. Thus, for any $\ell \in [t]\setminus\{i\}$

$$\mathbb{E}\left[\frac{\lambda^2 - p_{j_\ell}^2}{1-p_{j_\ell}^2}\hat{s}_{j_\ell}^{(i,k)}(x_{j_\ell}^{(i)} - p_{j_\ell})\right] = \mathbb{E}_{\mathbf{p}}\left[\frac{\lambda^2 - p_{j_\ell}^2}{1-p_{j_\ell}^2}\mathbb{E}\left[\hat{s}_{j_\ell}^{(i,k)}\right]\mathbb{E}_{x_{j_\ell}^{(i,k)}}\left[(x_{j_\ell}^{(i,k)} - p_{j_\ell})\right]\right] = 0.$$

This means that $\mathbb{E}\left[T'_{i,k}\right] = 0$. We apply the same observation, that $\mathbb{E}\left[x_{j_\ell}^{(i,k)}\right] = p_{j_\ell}$, and that every coordinate is independent to show that the cross terms in the variance of $T'_{i,j}$ cancel out, leaving us with

$$\mathrm{Var}\left(T'_{i,k}\right) = \mathbb{E}\left[(T'_i)^2\right] = \mathbb{E}\left[\sum_{\ell\in[t]\setminus i}\frac{(\lambda^2 - p_{j_\ell}^2)^2}{(1-p_{j_\ell}^2)^2}(\hat{s}_{j_\ell}^{(i,k)})^2(x_{j_\ell}^{(i)} - p_{j_\ell})^2\right].$$

Since $(x_{j_\ell}^{(i)} - p_{j_\ell})^2$ is at most 4 and $(\hat{s}_{j_\ell}^{(i,k)})^2 = 1$, we have the upper bound

$$\mathrm{Var}\left(T'_{i,k}\right) \le 4(t-1)\lambda^4 \le 4t\lambda^4.$$

Plugging these into inequality above and summing up over all the t people and $k \in \{1,2\}$, we obtain that

$$\mathbb{E}\left[\sum_{i\in[t]}\sum_{k\in\{1,2\}} T_{i,k}\right] \le 4\varepsilon\lambda^2\sqrt{t} + 4\delta\lambda^2 t.$$

The next step is to show that accuracy implies a lower bound for $\mathbb{E}\left[\sum_{i\in[t]}(T_{i,1}+T_{i,2})\right]$. We notice that in the sum of T_i we can rearrange the terms.

$$\begin{aligned}\sum_{i\in[t]}(T_{i,1}+T_{i,2}) &= \sum_{i\in[t]}\sum_{\ell\in[t]\setminus\{i\}}\left(\frac{\lambda^2 - p_{j_\ell}^2}{1-p_{j_\ell}^2}\hat{s}_{j_\ell}(x_{j_\ell}^{(i,1)} - p_{j_\ell}) + \frac{\lambda^2 - p_{j_\ell}^2}{1-p_{j_\ell}^2}\hat{s}_{j_\ell}(x_{j_\ell}^{(i,2)} - p_{j_\ell})\right)\\ &= \sum_{\ell\in[t]}\sum_{i\in[t]\setminus\{\ell\}}\left(\frac{\lambda^2 - p_{j_\ell}^2}{1-p_{j_\ell}^2}\hat{s}_{j_\ell}(x_{j_\ell}^{(i,1)} - p_{j_\ell}) + \frac{\lambda^2 - p_{j_\ell}^2}{1-p_{j_\ell}^2}\hat{s}_{j_\ell}(x_{j_\ell}^{(i,2)} - p_{j_\ell})\right)\\ &= \sum_{\ell\in[t]}\frac{\lambda^2 - p_{j_\ell}^2}{1-p_{j_\ell}^2}\hat{s}_{j_\ell}\sum_{i\in[t]\setminus\{\ell\}}\sum_{k\in\{1,2\}}(x_{j_\ell}^{(i,k)} - p_{j_\ell}).\end{aligned}$$

We can now apply Lemma D.3 to each coordinate j_ℓ.

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in[t]}(T_{i,1}+T_{i,2})\right] &= \sum_{\ell\in[t]}\mathbb{E}\left[\frac{\lambda^2-p_{j_\ell}^2}{1-p_{j_\ell}^2}\hat{s}_{j_\ell}\sum_{i\in[t]\setminus\{\ell\}}\sum_{k\in\{1,2\}}(x_{j_\ell}^{(i,k)}-p_{j_\ell})\right]\\ &\geq \sum_{\ell\in[t]}\left(\lambda-\mathbb{E}\left[2(|p_{j_\ell}|-\hat{s}_{j_\ell}p_{j_\ell})\right]\right)\\ &= \lambda t-\mathbb{E}\left[\sum_{\ell\in[t]}2(|p_{j_\ell}|-\hat{s}_{j_\ell}p_{j_\ell})\right].\end{aligned}$$

For vector of indices $\mathbf{j}$ let $\alpha_{\mathbf{j}}$ be the error

$$\alpha_{\mathbf{j}}=\mathbb{E}\left[\sum_{\ell\in[t]}\mathbb{I}\{\mathrm{sign}(p_{j_\ell})\neq\hat{s}_{j_\ell}\}|p_{j_\ell}|\right]=\frac{1}{4}\mathbb{E}\left[\sum_{\ell\in[t]}2(|p_{j_\ell}|-\hat{s}_{j_\ell}p_{j_\ell})\right].$$

Combining the two inequalities for $\mathbb{E}\left[\sum_{i\in[t]}(T_{i,1}+T_{i,2})\right]$ we have shown that when $\mathbf{j}$ has no duplicates

$$\lambda t-4\alpha_{\mathbf{j}}\leq\mathbb{E}\left[\sum_{i\in[t]}(T_{i,1}+T_{i,2})\right]\leq 4\varepsilon\lambda^2\sqrt{t}+4\delta\lambda^2 t.$$

By rearranging the terms we get that

$$\alpha_{\mathbf{j}}\geq\frac{1}{4}(\lambda t-4\varepsilon\lambda^2\sqrt{t}-4\delta\lambda^2 t).$$

We now incorporate the randomness over the choice of $\mathbf{j}$. Let E be the event that the set of target indices has no duplicates. Since $d\geq ct^2$ for a sufficiently large constant c, by Lemma A.7 event E occurs with probability at least $\frac{1}{2}$. Therefore,

$$\alpha=\mathop{\mathbb{E}}_{\mathbf{p},\mathbf{j}}\left[\frac{1}{t}\sum_{i\in[t]}\mathbb{E}\left[\mathbb{I}\{\mathrm{sign}(p_{j_i})\neq\hat{s}_{j_i}\}|p_{j_i}|\right]\right]=\frac{1}{t}\mathop{\mathbb{E}}_{\mathbf{j}}\left[\alpha_{\mathbf{j}}\right]\geq\frac{1}{t}\mathop{\mathbb{E}}_{\mathbf{j}}\left[\alpha_{\mathbf{j}}\mid E\right]\mathop{\mathbb{P}}_{\mathbf{j}}[E]\geq\frac{1}{2t}\mathop{\mathbb{E}}_{\mathbf{j}}\left[\alpha_{\mathbf{j}}\mid E\right].$$

For each $\mathbf{j}$ without duplicates we have a lower bound on $\alpha_{\mathbf{j}}$, so substituting this bound for $\mathop{\mathbb{E}}_{\mathbf{j}}\left[\alpha_{\mathbf{j}}\mid E\right]$, we get

$$\alpha\geq\frac{1}{8t}(\lambda t-4\varepsilon\lambda^2\sqrt{t}-4\delta\lambda^2 t)$$

Since $\lambda=16\alpha$, $\delta<\frac{1}{2t}$ and $t\geq 1$, we get that

$$\alpha\geq\frac{1}{32}\frac{1}{\frac{4\varepsilon}{\sqrt{t}}+4\delta}\geq\frac{1}{32}\frac{1}{\frac{4\varepsilon}{\sqrt{t}}+2}\geq\frac{1}{32}\min\left\{\frac{\sqrt{t}}{4\varepsilon},\frac{1}{2}\right\}=\min\left\{\frac{\sqrt{t}}{4\cdot 32\varepsilon},\frac{1}{64}\right\}.$$

□

Lemma D.8 (Indexed sign estimation metalearning lower bound)**.** *Fix parameters $\alpha\in(0,\frac{1}{8})$, $t\in\mathbb{N}$, $\varepsilon\in(0,1]$, $\delta\in(0,\frac{1}{2t})$, and error function ℓ_{est}. Let $\mathcal{M}=(\mathcal{M}_{meta},\mathcal{M}_{pers})$ be a pair of algorithms that metalearn a metadistribution $\mathcal{Q}$ over $\mathcal{P}_{est,d,t+1}$ with error α using t training tasks, 2 samples per training task and a test task with 2 personalization samples and $\mathcal{M}_{meta}$ satisfies (ε,δ)-DP. Then, $\alpha\geq\Omega(\min\{1,\frac{\sqrt{d}}{\varepsilon t}\})$.*

Proof. Fix parameters $\alpha \in (0, \frac{1}{8})$, $t \in \mathbb{N}$, $\varepsilon \in (0,1]$, $\delta \in (0, \frac{1}{2t})$, and error function ℓ_{est}. The metalearning algorithm $\mathcal{M}_{\text{meta}}$ takes as input 2 samples $(x^{(i,1)}, j^{(i,1)}), (x^{(i,2)}, j^{(i,2)})$ per person $i \in [t]$. The personalization algorithm $\mathcal{M}_{\text{pers}}$ gets as input the output of $\mathcal{M}_{\text{meta}}$ and the two samples of the $(t+1)$-th person $(x^{(t+1,1)}, j^{(t+1,1)}), (x^{(t+1,2)}, j^{(t+1,2)})$. It then outputs an estimate $\hat{s}_{j^{(t+1,1)}}$ of the sign of the mean of coordinate $j^{(t+1,1)}$. By the definition of $\mathcal{P}_{\text{est},d,t+1}$, the index of person $t+1$ in both samples is deterministically j_{t+1}. Hence, we will write $\hat{s}_{j_{t+1}}$ instead of $\hat{s}_{j^{(t+1,1)}}$ for simplicity.

To prove this theorem we construct a hard metadistribution $\mathcal{Q}$ where we draw a vector of means $\mathbf{p} \in [-\lambda, +\lambda]^d$, for $\lambda = 8\alpha \in (0,1)$ and a vector of indices $\mathbf{j} \in [d]^{t+1}$ both uniformly at random. Let the error of $\mathcal{M}$ be

$$\begin{aligned}
\alpha &\geq \mathop{\mathbb{E}}_{\substack{\mathcal{M}, \\ (P_1,\ldots,P_{t+1})\sim\mathcal{Q}, \\ x^{(1,1)},x^{(1,2)},\ldots,x^{(t+1,1)},x^{(t+1,2)}}} \left[\mathbb{I}\{\text{sign}(p_{j_{t+1}}) \neq \hat{s}_{j_{t+1}}\}|p_{j_{t+1}}|\right] \\
&= \mathop{\mathbb{E}}_{\substack{\mathcal{M}, \\ \mathbf{p},j_1,\ldots,j_t, \\ x^{(1,1)},x^{(1,2)},\ldots,x^{(t+1,1)},x^{(t+1,2)}}} \left[\mathop{\mathbb{E}}_{j_{t+1}} \left[\mathbb{I}\{\text{sign}(p_{j_{t+1}}) \neq \hat{s}_{j_{t+1}}\}|p_{j_{t+1}}|\right]\right] \\
&= \mathop{\mathbb{E}}_{\substack{\mathcal{M}, \\ \mathbf{p},j_1,\ldots,j_t, \\ x^{(1,1)},x^{(1,2)},\ldots,x^{(t+1,1)},x^{(t+1,2)}}} \left[\frac{1}{d}\sum_{j\in[d]} \mathbb{I}\{\text{sign}(p_j) \neq \hat{s}_j\}|p_j|\right].
\end{aligned}$$

We construct a tracing attack that uses the following test statistics for $i \in [t]$, $k \in \{1,2\}$

$$\begin{aligned}
T_{i,k} &\stackrel{\text{def}}{=} \sum_{j\in[d]} \frac{\lambda^2 - p_j^2}{1 - p_j^2} \hat{s}_j (x_j^{(i,k)} - p_j) \text{ and} \\
T'_{i,k} &\stackrel{\text{def}}{=} \sum_{j\in[d]} \frac{\lambda^2 - p_j^2}{1 - p_j^2} \hat{s}_j^{(i,k)} (x_j^{(i,k)} - p_j),
\end{aligned}$$

where $\hat{s}_j^{(i,k)}$ denotes the output of algorithm $\mathcal{M}_{\text{pers}}$ for $j_{t+1} = j$ when the input $(x^{(i,k)}, j^{(i,k)})$ of person i to $\mathcal{M}_{\text{meta}}$ has been replaced with a fresh draw from P_i. For $i = t+1$ we construct only test statistics

$$T_{t+1,k} \stackrel{\text{def}}{=} \sum_{j\in[d]} \frac{\lambda^2 - p_j^2}{1 - p_j^2} \hat{s}_j (x_j^{(i,k)} - p_j),$$

for $k \in \{1,2\}$. Since $\mathcal{M}_{\text{meta}}$ is (ε,δ)-DP with respect to i's dataset for every $i \in [t]$, $\hat{s}_j$ is (ε,δ)-DP with respect to the same dataset for every $j \in [d]$. Therefore, for $k \in \{1,2\}$

$$\mathbb{E}\left[T_{i,k}\right] \leq \mathbb{E}\left[T'_{i,k}\right] + 2\varepsilon\sqrt{\text{Var}\left(T'_{i,k}\right)} + 2\delta\|T'_{i,k}\|_\infty.$$

We will now analyze each term of the right hand side of the inequality. We see that

$$\begin{aligned}
\|T'_{i,k}\|_\infty &\leq \sum_{j\in[d]} \lambda^2 \left| \frac{1 - \frac{p_j^2}{\lambda^2}}{1 - p_j^2} \hat{s}_j^{(i,k)} \left(x_j^{(i,k)} - p_j\right) \right| \\
&\leq \sum_{i\in[d]} 2\lambda^2 = 2\lambda^2 d,
\end{aligned}$$

because $1 - \frac{p_j^2}{\lambda^2} \leq 1 - p_j^2$. Next, since $\hat{s}_j^{(i,k)}$ is independent of $x_j^{(i)}$ conditioned on $\mathbf{p}$, we get that

$$\begin{aligned}
\mathbb{E}\left[T_i'\right] &= \mathop{\mathbb{E}}_{\mathcal{M},\mathbf{p}}\left[\sum_{j\in[d]} \frac{\lambda^2 - p_j^2}{1-p_j^2} \mathop{\mathbb{E}}_{\substack{j_1,\ldots,j_t \\ x^{(1,1)},\ldots,x^{(t+1,2)},x^{(i,k)\prime}}}\left[\hat{s}_j^{(i,k)}\right] \mathop{\mathbb{E}}_{x^{(i,k)}}\left[(x_j^{(i,k)} - p_j)\right]\right] \\
&= 0
\end{aligned}$$

Finally, by the same observation the cross terms in the variance of $T_{i,k}'$ cancel out and we obtain that

$$\begin{aligned}
\mathrm{Var}\left(T_{i,k}'\right) &= \mathbb{E}\left[(T_{i,k}')^2\right] \\
&= \mathbb{E}\left[\sum_{j\in[d]} \frac{(\lambda^2 - p_j^2)^2}{(1-p_j^2)^2} (\hat{s}_j^{(i,k)})^2 (x_j^{(i,k)} - p_j)^2\right] \\
&\leq 4\mathbb{E}\left[\sum_{j\in[d]} \frac{(\lambda^2 - p_j^2)^2}{(1-p_j^2)^2}\right] = 4d\lambda^4
\end{aligned}$$

Combining the inequalities above we conclude that

$$\mathbb{E}\left[T_{i,k}\right] \leq 4\varepsilon\lambda^2\sqrt{d} + 4\delta d\lambda^2.$$

For the $i = t+1$, we have that

$$\begin{aligned}
\mathbb{E}\left[T_{t+1,k}\right] &= \mathbb{E}\left[\sum_{j\in[d]} \frac{\lambda^2 - p_j^2}{1-p_j^2} \hat{s}_j (x_j^{(t+1,k)} - p_j)\right] \\
&\leq 2d\lambda^2.
\end{aligned}$$

Therefore, by summing up the test statistics $T_{i,k}$

$$\mathbb{E}\left[\sum_{i\in[t+1]} (T_{i,1} + T_{i,2})\right] \leq 8\varepsilon\lambda^2 t\sqrt{d} + 8\delta d t\lambda^2 + 4\lambda^2 d.$$

The next step is to show that accuracy implies a lower bound for the test statistics in terms of error λ. We apply Lemma D.3 to every coordinate $j \in [d]$ of the estimate

$$\begin{aligned}
\mathbb{E}\left[\sum_{i\in[t+1]}\sum_{k\in\{1,2\}} T_{i,k}\right] &= \mathbb{E}\left[\sum_{i\in[t+1]}\sum_{k\in\{1,2\}}\sum_{j\in[d]} \frac{\lambda^2 - p_j^2}{1-p_j^2} \hat{s}_j (x_j^{(i,k)} - p_j)\right] \\
&= \sum_{j\in[d]} \mathbb{E}\left[\sum_{i\in[t+1]}\sum_{k\in\{1,2\}} \frac{\lambda^2 - p_j^2}{1-p_j^2} \hat{s}_j (x_j^{(i,k)} - p_j)\right] \\
&\geq \sum_{j\in[d]} \left(\lambda - \mathbb{E}\left[4\mathbb{I}\{\mathrm{sign}(p_j) \neq \hat{s}_j\}|p_j|\right]\right) = d\lambda - 4d\alpha.
\end{aligned}$$

Combining the bounds on $\mathbb{E}\left[\sum_{i\in[t+1]}\sum_{k\in\{1,2\}} T_{i,k}\right]$, we have the following inequality

$$d\lambda - 4d\alpha \leq 8\varepsilon\lambda^2 t\sqrt{d} + 8\delta d t\lambda^2 + 4\lambda^2 d.$$

By rearranging the terms and replacing λ with 8α we have that

$$\alpha \geq \frac{1}{16} \frac{1}{\frac{8\varepsilon t}{\sqrt{d}} + 8\delta t + 4}.$$

Since $\delta < \frac{1}{2t}$,

$$\alpha \geq \frac{1}{16} \frac{1}{\frac{8\varepsilon t}{\sqrt{d}} + 8}.$$

Finally, if $\frac{8\varepsilon t}{\sqrt{d}} \leq 8$, then $\alpha \geq \frac{\sqrt{d}}{16^2 \varepsilon t}$. Otherwise, $\alpha \geq \frac{1}{16^2}$. Therefore, we have shown that

$$\alpha \geq \min\left\{\frac{\sqrt{d}}{16^2 \varepsilon t}, \frac{1}{16^2}\right\}$$

□

Lemma D.9 (Indexed sign estimation billboard lower bound)**.** *Fix parameters $t \in \mathbb{N}$, $\varepsilon \in (0,1]$, $\delta \in (0, \frac{1}{32^2 t})$, $d \geq \frac{\varepsilon^2 t}{4}$. Let ℓ_{sign} be the loss function . Let $\mathcal{M}$ be a billboard algorithm that multitask learns $\mathcal{P}_{est,d,t}$ with error α with t tasks and 2 samples per task and satisfies (ε, δ)-DP. Then, $\alpha \geq \Omega(\min\{\frac{\sqrt{d}}{\varepsilon t}, 1\})$.*

Proof. By Theorem 3.1 if we have a billboard algorithm for multitask learning with error α, then we have a metalearning algorithm with error $e^{\varepsilon}\alpha + \delta$ for metadistributions over $P_{\text{est},d,t+1}$. Then, by Lemma D.8

$$e^{\varepsilon}\alpha + \delta \geq \min\left\{\frac{\sqrt{d}}{16^2 \varepsilon t}, \frac{1}{16^2}\right\}.$$

If $\frac{\sqrt{d}}{16^2 \varepsilon t} < \frac{1}{16^2}$, then since $\delta < \frac{1}{32^2 t}$, $d \geq \frac{\varepsilon^2 t}{4}$, $t \geq 1$, and $\varepsilon \leq 1$, we have that

$$\alpha \geq \frac{1}{e^{\varepsilon}}\left(\frac{\sqrt{d}}{16^2 \varepsilon t} - \delta\right) \tag{7}$$

$$\geq \frac{\sqrt{d}}{e^{\varepsilon} 2 \cdot 16^2 \varepsilon t} \geq \frac{\sqrt{d}}{6 \cdot 16^2 \varepsilon t}. \tag{8}$$

If $\frac{\sqrt{d}}{16^2 \varepsilon t} \geq \frac{1}{16^2}$, since $\delta < \frac{1}{32^2 t}$ we get that

$$\alpha \geq \frac{3}{e^{\varepsilon} 4 \cdot 16^2} > \frac{1}{4 \cdot 16^2}.$$

Therefore,

$$\alpha \geq \Omega\left(\min\left\{1, \frac{\sqrt{d}}{\varepsilon t}\right\}\right)$$

□

Theorem 5.6 (Restated)**.** Fix $\alpha \in (0, \frac{1}{16})$, $t \in \mathbb{N}$, $d \geq ct^2$ for a sufficiently large constant c, $\varepsilon > 0$ and $\delta \in (0, \frac{1}{2t})$. Let $\mathcal{M}$ be an algorithm that multitask learns $\mathcal{P}_{\text{class},d,t}$ with error α, for loss function ℓ_{class}, with t tasks and 1 sample per task, and satisfies (ε, δ)-JDP. Then, $\alpha \geq \Omega(\min\{\frac{1}{\varepsilon\sqrt{t}}, 1\})$.

Proof. If there exists an (ε, δ)-JDP algorithm $\mathcal{M}$ that multitask learns $\mathcal{P}_{\text{class},d,t}$ with t tasks and 1 sample per task to error $\alpha < \min\left\{\frac{\sqrt{t}}{4 \cdot 32\varepsilon}, \frac{1}{64}\right\}$, then by the reduction in Lemma D.1 we get that there exists a sign estimation algorithm that multitask learns $\mathcal{P}_{\text{est},d,t}$ with t tasks and 2 samples per task and has error $\alpha < \min\left\{\frac{\sqrt{t}}{4 \cdot 32\varepsilon}, \frac{1}{64}\right\}$, for loss function ℓ_{sign}. The algorithm we get from the reduction is also (ε, δ)-JDP due to post-processing. This contradicts the statement of Lemma D.7. Therefore, we get that α must be at least $\min\left\{\frac{\sqrt{t}}{4 \cdot 32\varepsilon}, \frac{1}{64}\right\}$. □

Theorem 5.7 (Restated)**.** Fix $\alpha \in (0, \frac{1}{8})$, $t \in \mathbb{N}$, $d \in \mathbb{N}$, $\varepsilon \in (0,1]$ and $\delta \in (0, \frac{1}{2t})$. Let $\mathcal{M} = (\mathcal{M}_{\text{meta}}, \mathcal{M}_{\text{pers}})$ be a pair of algorithms that metalearn a distribution $\mathcal{Q}$ over $\mathcal{P}_{\text{class},d,t+1}$ with error α, for loss function ℓ_{class}, using t training tasks, 1 sample per training task and a test task with 1 personalization sample, and $\mathcal{M}_{\text{meta}}$ satisfies (ε, δ)-DP. Then, $\alpha \geq \Omega(\min\{1, \frac{\sqrt{d}}{\varepsilon t}\})$.

Proof. We notice that a metadistribution $\mathcal{Q}_{\text{class}}$ over $\mathcal{P}_{\text{class},d,t+1}$ where we draw a vector of means $\mathbf{p} \in [-\lambda, \lambda]^d$, for $\lambda \in (0,1)$ and a vector of indices $\mathbf{j} \in [d]^{t+1}$ uniformly at random corresponds to metadistribution $\mathcal{Q}_{\text{est}}$ over $\mathcal{P}_{\text{est},d,t+1}$ where we draw $\mathbf{p}$ and $\mathbf{j}$ in the same way.

We will show that we can reduce metalearning for metadistribution $\mathcal{Q}_{\text{est}}$ for indexed sign estimation with t training tasks, 1 sample per training task and a test task with 2 personalization samples to metalearning for the corresponding distribution $\mathcal{Q}_{\text{class}}$ for indexed classification using t training tasks with 1 sample per training task and a test task with 1 personalization samples by following the steps of the proof of Lemma D.2 with some small changes.

In indexed sign estimation, let $(x^{(i)}, j^{(i)})$ be the sample of individual i that is drawn from $P^{(i)}_{\text{class}}$, for $(P^{(1)}_{\text{class}}, \dots, P^{(t+1)}_{\text{class}})$ drawn from $\mathcal{Q}_{\text{class}}$.

We can transform every sample $(x^{(i)}, j^{(i)})$, for $i \in [t]$ to a sample for indexed classification by setting $w^{(i)} \leftarrow x^{(i)}$, drawing a $y^{(i)} \in \{\pm 1\}$ uniformly and setting

$$\tilde{x}^{(i)}_{\ell} = \begin{cases} w^{(i)}_{\ell}, & \text{if } \ell \neq j_i \\ w^{(i)}_{\ell} y, & \text{if } \ell = j_i \end{cases}$$

The distribution of $(\tilde{x}^{(i)}, j^{(i)}, y^{(i)})$ is $P^{(i)}_{\text{class}}$ and the distribution of $(P^{(1)}_{\text{class}}, \dots, P^{(t+1)}_{\text{class}})$ is $\mathcal{Q}_{\text{class}}$. Every person $i \in [t]$ sends their new sample, $(\tilde{x}^{(i)}, j^{(i)}, y^{(i)})$, to the metalearning algorithm for indexed classification.

Person $t+1$ transform their two datapoints from $\{(x^{(t+1,1)}, j^{(t+1,1)})), (x^{(t+1,2)}, j^{(t+1,2)}))\}$ to $\{(\tilde{x}^{(t+1,1)}, j^{(t+1,1)}, y^{(t+1,1)}), (\tilde{x}^{(t+1,2)}, j^{(t+1,2)}, y^{(t+1,2)})\}$ using the same procedure as the people with the training tasks. They then run the personalization part of that algorithm with sample $(\tilde{x}^{(t+1,1)}, j^{(t+1,1)}, y^{(t+1,1)})$ and get a $\hat{f}(x_{j_{t+1}})$ with error

$$\mathbb{E}\left[\underset{(x,j,y)\sim P^{(t+1)}_{\text{class}}}{\mathbb{P}}\left[\hat{f}(x_{j_{t+1}}) \neq y\right] - \min_{f}\left\{\underset{(x,j,y)\sim P^{(t+1)}_{\text{class}}}{\mathbb{P}}\left[f(x_{j_{t+1}}) \neq y\right]\right\}\right]$$

where the expectation is taken over the t-tuple of classification distributions, the samples of the t training tasks, the randomness of the algorithm and the first sample of individual $t+1$. Then, person $t+1$ can get an estimate of the sign by postprocessing $\hat{f}$ using their second transformed sample $(\tilde{x}^{(t+1,2)}, j^{(t+1,2)}, y^{(t+1,2)})$:

$$\hat{s}_{j_t} = \hat{f}(\tilde{x}^{(t+1,2)}_{j_{t+1}})\tilde{x}^{(t+1,2)}_{j_{t+1}}.$$

Following the same calculations as in the proof of Lemma D.2 we get that

$$\mathbb{E}\left[\mathbb{I}\{\text{sign}(p_{j_{t+1}}) \neq \hat{s}_{j_{t+1}}\}|p_{j_{t+1}}|\right] = \mathbb{E}\left[\mathbb{P}\left[\hat{f}(x_{j_{t+1}}) \neq y\right] - \min_{f}\{\mathbb{P}\left[f(x_{j_{t+1}}) \neq y\right]\}\right] \leq \alpha.$$

where the expectation of the LHS of the equation is taken over the t-tuple of indexed sign estimation distributions, the initial samples of the t training tasks, the randomness of the algorithm we described and the two samples of individual $t+1$. By Lemma D.8 we have that $\alpha \geq \Omega(\min\{1, \frac{\sqrt{d}}{\varepsilon t}\})$.

□

Theorem 5.8 (Restated)**.** Fix $\alpha \in (0, \frac{1}{8})$, $t \in \mathbb{N}$, $d \geq \frac{\varepsilon^2 t}{4}$, $\varepsilon \in (0,1]$ and $\delta \in (0, \frac{1}{32^2 t})$. Let $\mathcal{M}$ be a billboard algorithm that multitask learns $\mathcal{P}_{\text{class},d,t}$ with error α, for loss function ℓ_{class}, with t tasks and 1 sample per task and satisfies (ε, δ)-DP. Then, $\alpha \geq \Omega(\min\{\frac{\sqrt{d}}{\varepsilon t}, 1\})$.

Proof. If there exists an (ε, δ)-DP billboard algorithm $\mathcal{M}$ that multitask $\mathcal{P}_{\text{class},d,t}$ with t tasks and 1 sample per task with error $\alpha < \min\left\{\frac{\sqrt{d}}{6 \cdot 16^2 \varepsilon t}, \frac{1}{32^2}\right\}$, then by the reduction in Lemma D.1 we get that there exists a sign estimation algorithm that multitask learns $\mathcal{P}_{\text{est},d,t}$ with t tasks and 2 samples per task and has error $\alpha < \min\left\{\frac{\sqrt{d}}{6 \cdot 16^2 \varepsilon t}, \frac{1}{32^2}\right\}$. The billboard algorithm we get from the reduction is also (ε, δ)-DP due to post-processing. This contradicts the statement of Lemma D.9. Therefore, we get that α must be at least $\min\left\{\frac{\sqrt{d}}{6 \cdot 16^2 \varepsilon t}, \frac{1}{32^2}\right\}$. □