Transfer Learning In Differential Privacy's Hybrid-Model

Refael Kohen [1] **Or Sheffet** [1]

Abstract

The hybrid-model (Avent et al., 2017) in Differential Privacy is a an augmentation of the local-model where in addition to N local-agents we are assisted by one special agent who is in fact a curator holding the sensitive details of n additional individuals. Here we study the problem of machine learning in the hybrid-model where the n individuals in the curator's dataset are drawn from a *different* distribution than the one of the general population (the local-agents). We give a general scheme – Subsample-Test-Reweigh – for this transfer learning problem, which reduces any curator-model DP-learner to a hybrid-model learner in this setting using iterative subsampling and reweighing of the n examples held by the curator based on a smooth variation of the Multiplicative-Weights algorithm (introduced by Bun et al. (2020)). Our scheme has a sample complexity which relies on the χ^2-divergence between the two distributions. We give worst-case analysis bounds on the sample complexity required for our private reduction. Aiming to reduce said sample complexity, we give two specific instances our sample complexity can be drastically reduced (one instance is analyzed mathematically, while the other - empirically) and pose several directions for follow-up work.

1. Introduction

Differential privacy (DP) has become modern era's de-facto gold standard for privacy-preserving data analysis. In particular, the local model of DP — in which each user interacts in a computation by sending randomized messages whose view yields at most ϵ-privacy loss — has gained much popularity due to its (relative) design simplicity. In particular, much work has dealt with the problem of machine learning in the local-model of DP: assuming the N users' details are drawn i.i.d. from some distribution T, how can we design local-DP protocols for finding an hypothesis h of small loss w.r.t. T. However, as opposed to the curator-model of DP — in which all data is held by a trusted curator in charge of executing the computation — machine learning in the local model of DP suffers from two drawbacks: (1) its has a much larger sample complexity and (2) it is of limited learning capabilities — where only problems that are SQ-learnable are learnable in the local model, in contrast to the curator model that allows us to learn (almost) any PAC-learnable problem (Kasiviswanathan et al., 2008). In order to address these problems, there has been an extensive study in recent years regarding augmentations of DP's local-model. Most notably, the *shuffle* model has gained much focus (Bittau et al., 2017; Cheu et al., 2019; Balle et al., 2019; 2020; Ghazi et al., 2021a;b) as it reduces the privacy-loss of local-model protocols drastically. Yet the focus of this work is on a different, far less studied, augmentation of local-DP.

We study the *hybrid*-model of differential privacy (Avent et al., 2017) — in which a local-model computation protocol with N users is augmented by the aid of one special agent who is in fact a curator holding the sensitive details of additional n individuals. (We refer to the former N users as the "local-agents" and the latter n individuals as the "curator-agents.") The hybrid-model models a situation in which some users (the n curator agents) trust a proposed curator and allow her unrestricted access to their sensitive details; while the remaining $N \gg n$ users trust solely themselves and opt for the local-model. Indeed, hybrid-model improves the learning capabilities of the local-model: the theoretical work of Beimel et al. (2020) have proven that in the hybrid-model certain problems, which are inefficiently solvable in the standard local-DP model, are efficiently and privately computable. Alas, in the analysis of Beimel et al. (2020), the n curator-agents come from the *exact same* distribution T as the remaining N local-agents.[1] In contrast, this work is motivated by a setting in which the n individuals *voluntarily* opt-in to the curator-model. Coping with such a "selection bias" was posed by Beimel et al. (2020) as an acute open problem since (quoting Beimel et al. (2020) verbatim:) "from a practical point of view, this ... is aligned with

[1]Faculty of Engineering, Bar-Ilan University, Israel. Correspondence to: Refael Kohen <refael.kohen@gmail.com>, Or Sheffet <or.sheffet@biu.ac.il>.

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

[1]Such a situation may arise when an extrinsic powerful agency, e.g. the census, randomly samples n individuals and mandates they provide the curator with their sensitive details.

current industry practices, and the ... individuals willing to contribute via a curator can be employees, technology enthusiasts, or individuals recruited as alpha- or beta-testers of products... (Merriman, Oct 7, 2014; Microsoft, Sep 15, 2017; Mozilla, June 4, 2019)"

In this work we model this selection bias as a particular type of a transfer learning problem. We no longer assume that the n curator-agents were drawn from the *same* distribution T as the local-agents, but rather that they were drawn from some *different* distribution S (the *source* distribution). Our goal, however, remains the same: to learn a good hypothesis w.r.t. T (the *target* distribution) while incurring at most ϵ privacy-loss for any of the agents involved in the computation. Naturally, should S and T be so different that they reside on disjoint support then the n curator agents provide no assistance our transfer learning problem. Thus, in our model S and T are of bounded χ^2-divergence (see details in Section 2). In other words, we study a particular variation of a *transfer learning* problem: in the process of finding h of small loss w.r.t. T we are allowed to conduct a DP computation over a set of examples drawn from S and also to conduct a local-DP computation over N additional examples drawn from T. And so we ask:

Are there learning tasks that are infeasible in the local-model, yet can be computed privately and efficiently when we have a DP curator-model access to samples drawn from a different distribution?

Our Contribution. Our transfer-learning problem has two naïve baselines. (1) Relying solely on the N local-agents and learning a small loss hypothesis w.r.t. T via some off-the-shelf local-model protocol; this is infeasible for certain problems (e.g. PARITY (Kasiviswanathan et al., 2008)) and costly for others (e.g. sparse problem in a d-dimensional setting, where known sample complexity bounds are $\geq d$, (Duchi et al., 2013; Smith et al., 2017)). (2) Relying solely on the n curator-agents and learning an hypothesis via the some off-the-shelf DP learning algorithm guaranteed to return an hypothesis of small loss, such as a loss upper bounded by $\mathrm{err}_{\mathsf{S}}(h) \leq \alpha/\mathrm{D}^{\infty}(\mathsf{T}\|\mathsf{S})$ or $\mathrm{err}_{\mathsf{S}}(h) \leq \alpha^2/\chi^2(\mathsf{T}\|\mathsf{S})+1$ (assuming these divergences are finite, see definition in Section 2).[2] This venue is feasible only when indeed an hypothesis h exists whose error over S is as small as we require; but fails to give a meaningful guarantee in an agnostic setting, even if the best hypothesis in $\mathcal{H}$ has error $\Theta(\alpha)$.

Our work proposes a third technique, whose sample complexity of the N local-agents is independent of d and whose sample complexity of the n curator-agents depends on $pdim(\mathcal{H})$ (so if $|\mathcal{H}| = \mathrm{poly}(d)$, e.g. the Example in Section 4, then $n = \mathrm{polylog}(d)$) and which may be feasible in the agnostic setting as well. Our proposed framework resembles 'Subsample & Aggregate' (Nissim et al., 2007) in broad brushstrokes, except that instead of subsampling in parallel in order to convert a non-private mechanism to a private one, we subsample sequentially in order to convert a curator-model learning mechanism over S to hybrid-model learner. The crux of our technique is that upon each time we subsample and produce an hypothesis of small-loss w.r.t. S yet high loss w.r.t. T, we make a *Multiplicative-Weights (MW)*-based update step to our subsampling distribution. With each MW-update we transition closer to a "target" distribution which is based on the *importance sampling* (IS) weights of the points drawn from S. We thus title our technique *Subsample-Test-Reweigh*.

For clarity, we partition our analysis into two parts. The first, detailed in Section 3, presents the main ideas of our techniques while tabling the notion of privacy for a later section. Specifically, seeing as we know that learning in the local-model is equivalent to SQ-learning and that the vast majority of PAC-learnable problems are learnt in the curator-model, we pose the following model. Fix an hypothesis ℓ which is PAC-learnable via hypothesis class $\mathcal{H}$ of pdim d through some learning algorithm $\mathcal{M}$. We are given n labeled examples drawn from S and only a SQ-oracle access to T (see formal definitions in Section 2). So we iteratively (1) set a distribution μ over the n examples, (2) use $\mathcal{M}$ to learn an hypothesis h of small loss w.r.t. μ, (3) query the SQ-oracle to see if h's loss is sufficiently small (halt if yes), (4) if h has large loss we reweigh the distribution using the MW-algorithm and proceed to the next iteration. In Section 3 we prove that w.h.p. this algorithm outputs in $T = \tilde{O}(\alpha^{-2})$ iterations an hypothesis h of error $\mathrm{err}_{\mathsf{T}}(h) = O(\alpha)$ provided $n = \tilde{\Omega}(d(\chi^2(\mathsf{T}\|\mathsf{S}) + 1)\alpha^{-2})$. The crux of the proof lies in showing (w.h.p.) the existence of a particular distribution $\bar{u}$ over the n drawn samples from S s.t. $\forall h \in \mathcal{H}$ the loss of h w.r.t. $\bar{u}$ and its loss w.r.t. T are close. Based on the seminal results of Cortes et al. (2010), it is not surprising that this $\bar{u}$ is the truncated IS weights $w(x) = \frac{\mathbb{P}_{\mathsf{T}}(x)}{\mathbb{P}_{\mathsf{S}}(x)}$ for each x drawn from S.

Next, in Section 4 we give the privacy-preserving version of our MW-based technique from Section 3. Replacing the SQ-oracle calls with simple applications of the Randomized-Response mechanism is trivial, but maintaining the privacy of the n curator-model agents is far trickier. First, it is evident that our off-the-shelf learner $\mathcal{M}$ must now be a privacy-preserving PAC-learner. More importantly, our MW-update step must also maintain bounded privacy-loss. Luckily this latter point was already addressed by Bun et al. (2020) who use the notion of κ-*dense distributions* to give a MW-based algorithm where any intermediate distribution μ is such that no one single point has probability mass exceeding $1/\kappa n$ (see details in Section 4). Using known several results re-

[2] This follows from the well known inequality, stating that for any event E we have $\Pr_{\mathsf{T}}[E] \leq \sqrt{(\chi^2(\mathsf{T}\|\mathsf{S}) + 1)\Pr_{\mathsf{S}}[E]}$.

garding subsampling and privacy (Karwa & Vadhan, 2018; Bun et al., 2018), we end up with the following reduction. Denote m_1 as the sample complexity of some off-the-shelf curator-model learner that outputs an hypothesis of loss $\leq \alpha$ under privacy loss parameter of $\epsilon = 1$. In order to successfully apply the learner T times under our κ-dense MW-update subsampling scheme and have a total privacy-loss of ϵ, we require a sample of n curator-agents drawn from S where $n = \tilde{\Omega}(m_1 \cdot \frac{\sqrt{T}}{\epsilon\kappa}) = \tilde{\Omega}(m_1 \cdot \frac{\chi^2(\mathsf{T}\|\mathsf{S})+1}{\epsilon\alpha^2})$ and $N = \tilde{\Omega}(\epsilon^{-2}\alpha^{-4})$ local-agents. While, admittedly, these bounds are less than ideal, this is the first result to prove the feasibility of private transfer learning using poly-size sample even for problems whose sample complexity in the local-model is exponential.

In Section 5 we pose suggestions as to how to reduce this bound as open problems, leveraging on the fact that the worst-case bounds for either the number of iterations T or the density parameter κ may be drastically reduced for specific hypothesis classes / instances. We give two particular examples of such instances: one proven rigorously (for the case of PARITY under the uniform distribution) and one based on empirical evaluations of our (non private version of the) transfer-learning technique in a two high-dimensional Gaussian settings, in which our algorithm makes far fewer iterations than our $O(\alpha^{-2})$ worst-case upper-bound.

1.1. Related Work

The bounds and limitations of private learning were first established by Kasiviswanathan et al. (2008) who proved that (roughly speaking) the learning capabilities of the local-model are equivalent to SQ-learning. This, together with the seminal result of Blum et al. (1994), gives an exponential lower bound on the number of local-agents required for learning PARITY in the local-model (fully formally proven in Beimel et al. (2020)). Other results regarding the power and limitations of the local-model were given in Beimel et al. (2008); Duchi et al. (2013); Bassily & Smith (2015); Smith et al. (2017); Duchi & Rogers (2019) culminating in Joseph et al. (2019). And yet, the classical SGD-algorithm is still applicable in the local-model (Smith et al., 2017).

The two main works about the hybrid model (Avent et al., 2017; Beimel et al., 2020) have been discussed already, as well as the elegant private boosting paradigm of Bun et al. (2020). Other works have also studied the applicability of the MW-algorithm for various tasks in DP (Hardt & Rothblum, 2010; Gupta et al., 2011). In the context of transfer learning, few works tied differential privacy with multi-task learning (Gupta et al., 2016; Xie et al., 2017; Li et al., 2020), hypthesis testing (Wang et al., 2018) and in a semi-supervised setting (Kumar, 2022); all focusing on empirical measuring of the utility and none giving any general framework with proven guarantees as this work. Also, it could be interesting to implement a version of our algorithm which isn't private w.r.t. the samples from S in a "PATE"-like setting (Papernot et al., 2018) where public (or partially public) datasets are also available.

The classic problem of transfer learning has been studied extensively since the 20th century, and has too long and too rich of a history to be surveyed properly here. We thus mention a few recent works that achieved sample complexity bounds based on importance sampling (IS) weights and show concentration bounds related to divergences between the source and that target distributions. Some works (Agapiou et al., 2017; Chatterjee & Diaconis, 2017) showed sample complexity bounds based, in part, on $\exp(\mathrm{KL}(\mathsf{T}\|\mathsf{S})) \leq \chi^2(\mathsf{T}\|\mathsf{S}) + 1$ as well as other properties of the distribution[3] which lack our desired sub-Gaussian behavior. The seminal result of Cortes et al. (2010) achieved a bound that is, in spirit, more similar to the desired sub-Gaussian behavior. They also proved that with sample size bounds depending on $\chi^2(\mathsf{T}\|\mathsf{S})$ we can achieve accurate estimation of the loss of any hypothesis (simultaneously) in a finite pdim hypothesis class $\mathcal{H}$. Recently, Metelli et al. (2021) suggested a method of correction of the weights that allow obtaining sub-Gaussian behavior, assuming prior knowledge of the second moment of the weights. Maia Polo & Vicente (2022) studied the case where access to distribution T is restricted to unlabeled examples, and propose methods for evaluating the IS weights. Yao & Doretto (2010) use a boosting algorithm (based on the MW algorithm) for transfer learning, yet in their work, the distributions S and T are identical but the features of the classes are different.

2. Preliminaries

PAC- and SQ-Learning. The seminal work of Valiant (1984) defined PAC-learnability, where the learner has access to poly-many examples drawn from a given distribution T over a domain $\mathcal{X}$ labeled by some ℓ and its goal is to approximate ℓ as well as the best hypothesis in some given hypothesis class $\mathcal{H}$. We measure our prediction success using some loss-function L where for every $x \in \mathcal{X}, h \in \mathcal{H}$ and any labelling function ℓ it holds that $\mathrm{L}(h(x), \ell(x)) \in [0, 1]$; and so for any distribution T we denote $\mathrm{err}_{\mathsf{T}}(h) = \mathbb{E}_{x\sim\mathsf{T}}[\mathrm{L}(h(x), \ell(x))]$. Thus, a (α, β)-PAC-learner outputs w.p.$\geq 1 - \beta$ an hypothesis $h \in \mathcal{H}$ where $\mathrm{err}_{\mathsf{T}}(h) \leq \alpha_{\mathcal{H}} + \alpha$, where $\alpha_{\mathcal{H}} \stackrel{\mathrm{def}}{=} \min_{h\in\mathcal{H}}\{\mathrm{err}_{\mathsf{T}}(h)\}$. The learner's ability to draw n i.i.d. examples is simulated via an *example-oracle* which upon a query returns such a labeled example. Thus, in the PAC-mode, a learner has full access to all of the details of any drawn example. In contrast, the *Statistical Query* (SQ) model (Kearns, 1998) restricts the

[3] Namely, probability of drawing a point of large weight.

operation of the algorithm to view solely statistical properties of the distribution T and the labeling function ℓ. This is simulated via access to a SQ_τ-*oracle* which upon a statistical query $\phi : \mathcal{X} \times \mathbb{R} \to [0,1]$ returns an estimation of $\mathbb{E}_{x\sim\mathsf{T}}[\phi(x,\ell(x))]$ up to a tolerance parameter τ (polynomially bounded away from 0).

Differential Privacy. Given a domain $\mathcal{X}$, two multi-sets $I, I' \in \mathcal{X}^n$ are called *neighbors* if they differ on a single entry. An algorithm (alternatively, mechanism) is said to be (ϵ,δ)-*differentially private* (DP) (Dwork et al., 2006b;a) if for any two neighboring I, I' and any set S of possible outputs we have: $\Pr[\mathcal{M}(I) \in S] \leq e^\epsilon \Pr[\mathcal{M}(I') \in S] + \delta$.

The Randomized-Response mechanism (Warner, 1965; Kasiviswanathan et al., 2008) is one of the classic (ϵ,δ)-DP mechanisms in the local-model. Given privacy parameters ϵ, δ, on an input $b \in [0,1]$ it outputs $RR_{\epsilon,\delta}(b) \sim \mathcal{N}(b, \frac{2\ln(2/\delta)}{\epsilon^2})$. When applied to N i.i.d. draws from a Bernoulli r.v. of mean μ, then we estimate μ by $\theta = \frac{1}{N}\sum_i RR_{\epsilon,\delta}(b_i)$. Standard application of the Hoeffding bound and Gaussian concentration bounds proves that applying the mechanism to $N(\alpha,\beta,\epsilon,\delta) = \frac{4\ln(2/\delta)\ln(4/\beta)}{\epsilon^2\alpha^2}$, we get that $\Pr[|\theta-\mu| \leq \alpha] \geq 1-\beta$.

Differentially Private Machine Learning is by now too large of a field to be surveyed properly. It was formally initiated by (Kasiviswanathan et al., 2008) who, as discussed above proved that the set of hypothesis-classes which is SQ-learnable is conceptually equivalent to the set hypothesis-classes learnable in the local-model of DP. It is also worth mentioning the DP techniques for Empirical Risk Minimization and especially private SGD (Chaudhuri et al., 2011; Bassily et al., 2014) which we use as our private learner in Appendix A.

Divergence Between Distributions. Given two distributions S and T over the same domain $\mathcal{X}$ that have a Radon-Nikodym derivative, we denote said derivative as the *importance sampling (IS) weight* at a point $x \in \mathcal{X}$ as $w(x) = \frac{\mathbb{P}_\mathsf{T}(x)}{\mathbb{P}_\mathsf{S}(x)}$. Given a convex function $f : (0,\infty) \to \mathbb{R}$ where $f(1) = 0$, the f-divergence between two distribution T and S is $D^f(\mathsf{T}||\mathsf{S}) = \mathbb{E}_{x\sim\mathsf{S}}[f(w(x))]$. In the specific case when $f(x) = \frac{1}{2}|x-1|$ we obtain the *total variation* distance, denoted $\mathrm{TV}(\mathsf{T},\mathsf{S})$; when $f(x) = x\log(x)$ we obtain the *Kullback-Leibler* (KL) divergence, denoted $\mathrm{KL}(\mathsf{T}\|\mathsf{S})$; and when $f(x) = x^2 - 1$ we obtain the χ^2-*divergence*, denoted $\chi^2(\mathsf{T}\|\mathsf{S})$. Thus, a finite χ^2-divergence between distributions implies a finite second moment for $w(x)$. It is indeed quite simple to see that $\mathbb{E}_{x\sim\mathsf{T}}[w(x)] = \mathbb{E}_{x\sim\mathsf{S}}[w(x)^2] = \chi^2(\mathsf{T}\|\mathsf{S}) + 1$. Moreover, a well-known result (Csiszár & Shields, 2004) states the for any twice-differentiable f it holds that $D^f(\mathsf{T}\|\mathsf{S}) \approx \frac{f''(1)}{2}\chi^2(\mathsf{T}\|\mathsf{S})$ (as follows from f's Taylor series). In the context of transfer learning, Cortes et al. (2010) proved that weighing all examples in a sufficiently large sample drawn from S according to their IS weights gives an accurate estimation of the loss of any hypothesis in a finite pdim hypothesis class $\mathcal{H}$ over T; where their sample size bounds depend on $\chi^2(\mathsf{T}\|\mathsf{S})$. Another well-used notion of divergence is the α-Reyni divergence between T and S, defined as $D_\alpha(\mathsf{T}\|\mathsf{S}) = \frac{1}{\alpha-1}\ln(\mathbb{E}_{x\sim\mathsf{S}}[w(x)^\alpha])$ for any $\alpha > 1$, which was also used to define similar notions of privacy (Bun & Steinke, 2016; Mironov, 2017; Bun et al., 2018). Note that $D_2(\mathsf{T}\|\mathsf{S}) = \ln(\chi^2(\mathsf{T}\|\mathsf{S})+1)$. We also denote $\mathrm{D}^\infty(\mathsf{T}\|\mathsf{S}) = \sup_{x\in\mathcal{X}} w(x)$.

Bernstein Inequality: In our work we use several standard concentration bounds (Markov- / Chernoff- / Hoeffding-inequalities), and also the slightly less familiar inequality of Bernstein (1954): Let $\{X_i\}_{i=1}^n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i. Then for any positive t,

$$\Pr\left(\sum_i X_i \geq t\right) \leq \exp\left(\frac{-t^2}{2\sum_i \mathbb{E}[X_i^2] + 2tM/3}\right) \quad (1)$$

3. A Non-Private Model

For the sake of clarity, we introduce our algorithm in stages — where first we disregard the privacy aspect of the problem. In this section we deal with a specific transfer learning model, whose details are as follows. We are given a known domain $\mathcal{X}$ and some unknown labelling function $\ell : \mathcal{X} \to \mathbb{R}$. We also have the following access oracles to two unknown distributions S and T over $\mathcal{X}$ — we are given an *example oracle* Ex access to S, that upon a query returns an example x drawn from S and labeled by ℓ; and we are given a *statistical query oracle* SQ_τ to T, that upon a query $\phi : \mathcal{X} \times \mathbb{R} \to [0,1]$ returns an answer in the range $\mathbb{E}_{x\sim\mathsf{T}}[\phi(x,\ell(x)] \pm \tau$. Our goal is learn ℓ through some hypothesis class $\mathcal{H}$; i.e. to find an hypothesis h whose loss w.r.t. ℓ over T is comparable to the loss of the best hypothesis in $\mathcal{H}$, whose $\mathrm{pdim}(\mathcal{H}) = d$. Formally, we introduce our algorithm in the realizable setting, so given a parameter $\alpha > 0$ and a loss function $\mathrm{L} : \mathbb{R}^2 \to [0,1]$ our goal is to find h such that $\mathrm{err}_\mathsf{T}(h) = \alpha$. (We later discuss extension to the agnostic case.) In addition, we are given an algorithm $\mathcal{M}$ which is a (α,β)-PAC learner for our hypothesis class $\mathcal{H}$. Namely, given $\alpha,\beta > 0$, our learner $\mathcal{M}$ takes a sample of $m(\alpha,\beta)$ i.i.d. examples drawn from some distribution μ and labeled by some ℓ and w.p. $\geq 1-\beta$ returns a function $h \in \mathcal{H}$ whose loss is upper bounded by $\mathrm{err}_\mu(h) \leq \alpha$. So, had we been able to draw i.i.d. examples from T, we would have just fed them to the algorithm and produce a good hypothesis h. Alas, in our model, we only have statistical query oracle access to T, SQ_τ, with error of $\pm\tau$.

In this section, we propose a general reduction of a PAC-learner over S to finding a good hypothesis w.r.t. T, which

Algorithm 1 Non-Private Subsample-Test-Reweigh

Input: parameters $0 < \alpha, \beta < 1/8$
Draw $n \geq \frac{800(\chi^2+1)\log(1/\alpha)}{\alpha^2}(d\log(\frac{400d}{\alpha^3}) + \log(\frac{8}{\beta}))$ labeled points i.i.d. from S, denoted $x_1, x_2, ..., x_n$.
Set weight $w_i^1 \leftarrow 1$ for each $1 \leq i \leq n$.
Set $T \leftarrow \frac{32\log_2(\frac{8(\chi^2+1)}{\alpha})}{\alpha^2}$.
for $(t = 1, 2, 3, ..., T)$ **do**
  Set μ^t as a distribution where $\mu_i^t \propto w_i^t$.
  Draw $m(\alpha, \beta/T)$ examples i.i.d from μ^t.
  Apply $\mathcal{M}$ to the drawn points and obtain a function h^t.
  Set a^t as the reply of the SQ_τ-oracle for the query $\phi^t(x, \ell(x)) = \mathrm{L}(h^t(x), \ell(x))$.
  if $(a^t > 2\alpha + \tau + \alpha_{\mathcal{H}})$ **then**
    $\forall i$ set $w_i^{t+1} \leftarrow w_i^t \cdot \exp(-\alpha/8 \cdot [1 - \mathrm{L}(h^t(x_i), \ell(x_i))])$
  else
    return h^t (and halt)
  end if
end for

depends solely on $\chi^2(\mathsf{T}\|\mathsf{S}) \stackrel{\text{def}}{=} \chi^2$. We argue that Algorithm 1 achieves our goal within $T = O(\log(\chi^2+1/\alpha)/\alpha^2)$-iterations.

Note that Algorithm 1 is presented for the realizable case. If we deal with an agnostic case, namely — where $\alpha_{\mathcal{H}} > 0$, then our algorithm iterates as long as $\mathrm{err}_{\mu^t}(h^t) < \alpha_{\mathcal{H}} + \alpha$. (A condition which ought to hold when we begin with h^1, the a good hypothesis w.r.t. S and $\mathrm{err}_{\mathsf{S}}(h^1) \ll \mathrm{err}_{\mathsf{T}}(h^1)$.) It ought to be clear that by returning h^t with the smallest error estimation given by the SQ-oracle in all T iterations, an hypothesis $h \in \mathcal{H}$ of small loss w.r.t. T is obtained.

Theorem 3.1. *In the above-described setting, w.p.* $\geq 1 - 2\beta$ *Algorithm 1, halts and outputs an hypothesis* h *with* $\mathrm{err}_{\mathsf{T}}(h) \leq 2\alpha + 2\tau$.

The proof of Theorem 3.1 relies on proving the following lemma, and builds on — and extends — a theorem of Cortes et al. (2010).

Lemma 3.2. *Given* $0 < \alpha, \beta \leq 1/8$ *and two distributions* S *and* T *whose* χ^2*-divergence is* $\chi^2(\mathsf{T}||\mathsf{S}) = \chi^2$. *Then, if* $n = \Omega((\chi^2+1)\frac{\log(1/\alpha)}{\alpha^2}(d\log(\frac{d}{\alpha}) + \log(\frac{1}{\beta})))$ *w.p.* $\geq 1 - \beta$ *it holds that there exists a distribution* $\bar{u}$ *over the* n *drawn points such that (i) its divergence to the uniform distribution over the* n *drawn point,* $U_{[n]}$, *satisfies* $\mathrm{KL}(\bar{u}\|U_{[n]}) \leq \log_2(8(\chi^2+1)/\alpha)$ *and also (ii)* $\forall h \in \mathcal{H}$, $\mathrm{err}_u(h) \geq \mathrm{err}_{\mathsf{T}}(h) - 5\alpha/8$.

Proof of Theorem 3.1. Based on Lemma 3.2, we now simply apply the characterization of the MW-algorithm from the seminal work of Arora et al. (2012). Setting the 'cost' of example i w.r.t. hypothesis h^t as $m_i^t = 1 - \mathrm{L}(h^t(x_i), \ell(x_i))$ we have that applying the MW-algorithm with costs $\bar{m}^1, ..., \bar{m}^T$ and with an update rate of $\eta = \alpha/8$, we get that for any fixed distribution ν it holds that

$$\sum_t \mathop{\mathbb{E}}_{i\sim\mu^t}[m_i^t] \leq \sum_t \mathop{\mathbb{E}}_{i\sim\nu}[m_i^t] + \frac{\alpha T}{8} + \frac{8\mathrm{KL}(\nu\|U_{[n]})}{\alpha}$$

Note that w.p. $\geq 1-\beta$ our algorithm $\mathcal{M}$ returns an hypothesis of $\mathrm{err}_{\mu^t} \leq \alpha$ in all T iterations. In contrast, we know that each MW-update happens since $\mathrm{err}_{\mathsf{T}}(h^t) \geq 2\alpha + \tau - \tau = 2\alpha$; and so, w.p. $\geq 1-\beta$, we have a distribution $\nu = \bar{u}$ given by Lemma 3.2, which we plug-in to the above equation and have

$$\sum_t (1-\alpha) \leq \sum_t [1 - (2\alpha - \tfrac{5\alpha}{8})] + \frac{\alpha T}{8} + \frac{8\log_2(\frac{8(\chi^2+1)}{\alpha})}{\alpha}$$

Rearranging we get the bound $\frac{3\alpha T}{8} \leq \frac{\alpha T}{8} + \frac{8\log_2(\frac{8(\chi^2+1)}{\alpha})}{\alpha}$. Hence, w.p.$\geq 1-2\beta$ Algorithm 1 halts within some iteration $t^* \leq \frac{32\log_2(\frac{8(\chi^2+1)}{\alpha})}{\alpha^2}$ steps, which means $\mathrm{err}_{\mathsf{T}}(h^{t^*}) \leq (2\alpha + \tau) + \tau = 2\alpha + 2\tau$. □

Proof of Lemma 3.2. Let $w(x) : \mathcal{X} \to \mathbb{R}_+$ be the function defined by $w(x) = \frac{\mathbb{P}_{\mathsf{T}}(x)}{\mathbb{P}_{\mathsf{S}}(x)}$, where $\mathbb{P}_{\mathsf{T}}$ and $\mathbb{P}_{\mathsf{S}}$ denotes the PDFs of T and S resp. We call $w(x)$ the *weight of* x. Lets $u(x) = \min(w(x), \frac{4(\chi^2+1)}{\alpha})$ be the *truncated weight* of x. Recall that $\mathop{\mathbb{E}}_{x\sim\mathsf{S}}[w(x)] = 1$ and that $\mathop{\mathbb{E}}_{x\sim\mathsf{S}}[w(x)^2] = \mathop{\mathbb{E}}_{x\sim\mathsf{T}}[w(x)] = \chi^2 + 1$. Thus, since for every $x \in \mathcal{X}$ it holds that $0 \leq u(x) \leq w(x)$ then we have $\mathop{\mathbb{E}}_{x\sim\mathsf{S}}[u(x)] \leq 1$ and that $\mathop{\mathbb{E}}_{x\sim\mathsf{S}}[u(x)^2] \leq \mathop{\mathbb{E}}_{x\sim\mathsf{T}}[u(x)] \leq \chi^2 + 1$. Denoting $\mathcal{A} = \{x \in \mathcal{X} : w(x) > \frac{4(\chi^2+1)}{\alpha}\}$, we can apply the Markov inequality to infer that $\Pr_{x\sim\mathsf{T}}[x \in \mathcal{A}] = \mathop{\mathbb{E}}_{x\sim\mathsf{T}}[\mathbb{1}_{\mathcal{A}}(x)] \leq \alpha/4$ with $\mathbb{1}_{\mathcal{A}}(x)$ denoting the indicator of whether $x \in \mathcal{A}$ or not. And so:

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathsf{S}}[w(x) - u(x)] &= \mathop{\mathbb{E}}_{x\sim\mathsf{S}}[(w(x) - u(x))\mathbb{1}_{\mathcal{A}}(x)] \\ &\leq \int_{\mathcal{X}} w(x)\mathbb{1}_{\mathcal{A}}(x)\mathbb{P}_{\mathsf{S}}(x)dx \\ &= \int_{\mathcal{X}} \mathbb{1}_{\mathcal{A}}(x)\mathbb{P}_{\mathsf{T}}(x)dx = \mathop{\mathbb{E}}_{x\sim T}[\mathbb{1}_{\mathcal{A}}(x)] \leq \alpha/4\end{aligned}$$

which implies that $\mathbb{E}_{x\sim\mathsf{S}}[u(x)] \geq 1 - \alpha/4$.

We continue to bounding $U = \sum_{i=1}^n u(x_i)$ for the n points $x_1, x_2, ..., x_n$ in our sample taken i.i.d. from S, provided $n \geq \frac{150(\chi^2+1)\ln(4/\beta)}{\alpha^2}$. To that end, we apply the Bernestein

inequality. First, we have that

$$\begin{aligned}\Pr[U > (1+\alpha/8)n] &\le \Pr[U - n\,\mathbb{E}[u(x_i)] > \alpha n/8] \\ &\le \exp\left(-\frac{\alpha^2 n^2/8^2}{2\sum_i \mathrm{Var}[u(x_i)] + \frac{2}{3}(\frac{4(\chi^2+1)}{\alpha})\cdot \alpha n/8}\right) \\ &\le \exp\left(-\frac{\alpha^2 n^2}{128n(\chi^2+1) + \frac{64n}{3}(\chi^2+1)}\right) \\ &\le \exp\left(-\frac{\alpha^2 n}{150(\chi^2+1)}\right) \le \beta/4\end{aligned}$$

and similarly we can prove $\Pr[n(1-\alpha/4) - U > \alpha n/8] \le \beta/4$. Hence, w.p. $\ge 1-\beta/2$ it holds that $(1-3\alpha/8)n \le U \le (1+\alpha/8)n$.

Now, setting $\bar{u}$ as the distribution where point i is sampled w.p. $\frac{u(x_i)}{U}$ we can use the above bound on U to infer that

$$\begin{aligned}\mathrm{KL}(\bar{u}\|U_{[n]}) &= \sum_i \tfrac{u(x_i)}{U}\log_2(\frac{u(x_i)}{U}/\frac{1}{n}) \\ &\le \sum_i \tfrac{u(x_i)}{U}\log_2\left(\frac{u(x_i)}{1-3\alpha/8}\right) \\ &\le \sum_i \tfrac{u(x_i)}{U}\log_2(\tfrac{4(\chi^2+1)}{\alpha}\cdot\tfrac{8}{5}) \le \log_2(\tfrac{8(\chi^2+1)}{\alpha})\end{aligned}$$

proving the first part of the claim. As for the second part of the claim, we simply use the result of Cortes et al. (2010) which shows universal convergence w.r.t. any unnormalized weights function. Namely, by setting $\alpha' = \frac{\alpha}{\sqrt{50(\chi^2+1)\log(1/\alpha)}}$,[4] we have that for any hypothesis class $\mathcal{H}$ of $\mathrm{pdim}(\mathcal{H}) = d$, the following holds

$$\begin{aligned}&\Pr[\sup_{h\in\mathcal{H}}\{\mathop{\mathbb{E}}_{x\sim\mathsf{S}}[u(x)\mathrm{L}(h(x),\ell(x))] - \textstyle\sum_i \frac{u(x_i)}{n}\mathrm{L}(h(x_i),\ell(x_i))\} > \frac{\alpha}{4}] \\ &\le \Pr[\sup_{h\in\mathcal{H}}\{\tfrac{\mathbb{E}_{x\sim\mathsf{S}}[u(x)\mathrm{L}(h(x),\ell(x))] - \sum_i \frac{u(x_i)}{n}\mathrm{L}(h(x_i),\ell(x_i))}{\sqrt{\mathbb{E}_{x\sim\mathsf{S}}[w^2(x)\mathrm{L}^2(h(x),\ell(x))]}}\} > \alpha'\sqrt{2+\log(\tfrac{1}{\alpha'})}] \\ &\le \Pr[\sup_{h\in\mathcal{H}}\{\tfrac{\mathbb{E}_{x\sim\mathsf{S}}[u(x)\mathrm{L}(h(x),\ell(x))] - \sum_i \frac{u(x_i)}{n}\mathrm{L}(h(x_i),\ell(x_i))}{\sqrt{\mathbb{E}_{x\sim\mathsf{S}}[u^2(x)\mathrm{L}^2(h(x),\ell(x))]}}\} > \alpha'\sqrt{2+\log(\tfrac{1}{\alpha'})}] \\ &\overset{(*)}{\le} 4\exp\left(d\log(\tfrac{2en}{d}) - \tfrac{n\alpha'^2}{4}\right) \\ &= 4\exp\left(d\log(\tfrac{2en}{d}) - \tfrac{n\alpha^2}{200(\chi^2+1)\log(1/\alpha)}\right) \le \tfrac{\beta}{2}\end{aligned}$$

when $n \ge \frac{800(\chi^2+1)\log(1/\alpha)}{\alpha^2}(d\log(\frac{400d}{\alpha^3}) + \log(\frac{8}{\beta}))$. Note that $(*)$ is taken verbatim from Cortes et al. (2010) Theorem 8. Thus, w.p. $\ge 1-\beta$ both the above bounds on U hold and we have that for any $h \in \mathcal{H}$ it holds that

$$\begin{aligned}\mathop{\mathrm{err}}_{\mathsf{T}}(h) &= \mathop{\mathbb{E}}_{x\sim\mathsf{T}}[\mathrm{L}(h(x),\ell(x))] = \mathop{\mathbb{E}}_{x\sim\mathsf{S}}[w(x)\mathrm{L}(h(x),\ell(x))] \\ &\le \mathop{\mathbb{E}}_{x\sim\mathsf{S}}[u(x)\mathrm{L}(h(x),\ell(x))] + \mathop{\mathbb{E}}_{x\sim\mathsf{S}}[w(x)-u(x)] \\ &\le \left[\tfrac{\alpha}{4} + \tfrac{1}{n}\sum_i u(x_i)\mathrm{L}(h(x_i),\ell(x_i))\right] + \tfrac{\alpha}{4} \\ &\le \tfrac{\alpha}{2} + \tfrac{(1+\alpha/8)}{U}\sum_i u(x_i)\mathrm{L}(h(x_i),\ell(x_i)) \\ &\le \tfrac{5\alpha}{8} + \mathop{\mathbb{E}}_{x_i\sim\bar{u}}[\mathrm{L}(h(x_i),\ell(x_i)] \qquad \square\end{aligned}$$

4. The Private Boosting Paradigm

We now turn our attention to the full hybrid-model, and to our need to privatize Algorithm 1. In order to design a private version of Algorithm 1 one required multiple changes. First, and perhaps the easiest, is the fact that instead of a SQ_τ oracle access to T we estimate the error of h^t using RR (in the local-model). Assuming our algorithm makes at most T iterations, standard argument shows that each such query requires $\Omega(\frac{1}{\epsilon^2\alpha^2}\log(T/\beta))$ users so that w.p. $\ge 1-\beta/T$ we get a α-esimation of $\mathrm{err}_T(h^t)$; so, the number of local-users required for our paradigm is $\Omega(\frac{T\log(T/\beta)}{\epsilon^2\alpha^2}) = \Omega(\frac{\log((\chi^2+1)/\alpha)\log(\log(\chi^2+1)/\alpha\beta)}{\epsilon^2\alpha^4})$.

Second, which is also a rather straight-forward change, is that we need to replace the learning mechanism $\mathcal{M}$ with a *privacy-preserving* learning mechanism whose sample complexity depends also on the privacy-loss parameter(s). This implies that the sample complexity of $\mathcal{M}$ is a function of the 4 parameter $m = m(\alpha,\beta,\epsilon,\delta)$.[5] The question of setting the privacy parameters of $\mathcal{M}$, thereby inferring the sample complexity of $\mathcal{M}$, will be discussed momentarily.

Lastly, the more challenging aspect of the problem is maintaining the privacy of the samples among the n examples that are drawn from S. To that end, we rely on the MW-variant of (Bun et al., 2020), which in turn requires we introduce one more parameter, namely κ, into the problem.

Definition 4.1. Fix some $0 < \kappa < 1$. Given n points and a set of weights $w_1, w_2, ..., w_n \ge 0$, we denote $w_{\mathrm{avg}} = \frac{1}{n}\sum_i w_i$ and $w_{\max} = \max_i\{w_i\}$. We say that the distribution induced by these weights, namely the distribution where $\mu_i \propto w_i$, is κ-*dense* if $\kappa w_{\max} \le w_{\mathrm{avg}}$.

From a privacy stand-point, it is clear why a dense distribution is a desired trait: it makes it so that in a random sample of m draws from μ we expect each point i to be drawn no more than $1/\kappa$ times. (Bun et al., 2020) showed that for any set of weights $\bar{w} = (w_1, w_2, ..., w_n)$ where $\forall i,\ w_i \in (0,\kappa]$,

[4] Which satisfies that $\alpha/4 > \alpha'\sqrt{(2+\ln(1/\alpha'))(\chi^2+1)} = \alpha'\sqrt{(2+\ln(1/\alpha'))\,\mathbb{E}_{x\sim S}[w^2(x)]}$, when $\alpha \le 1/8$.

[5] For brevity, we use (ϵ,δ) as the privacy parameters, even if $\mathcal{M}$ is a (possibly truncated) zCDP-mechanism or Réyni-DP.

by setting

$$\Pi_\kappa(\bar{w}) = (\min\{c\cdot w_1, 1\}, \min\{c\cdot w_2, 1\}, \ldots, \min\{c\cdot w_n, 1\}) \quad (2)$$

for the smallest c s.t. $\|\Pi_\kappa(w)\|_1 = \kappa n$, we obtain a set of weights whose induced distribution $\mu = \frac{\Pi_\kappa(\bar{w})}{\|\Pi_\kappa(\bar{w})\|_1}$ is κ-dense. Using this projection we get the following claim.

Claim 4.1. *Let S and S' be any two neighboring datasets of size n each. Let $\bar{w}$ and $\bar{w}'$ be two weight vectors in $(0,\kappa]^n$ which may differ on the only entry that differs between S and S', and let μ and μ' be the two distributions derived from $\bar{w}$ and $\bar{w}'$ resp. using the projection of (2). Let $\mathcal{H}$ be an hypothesis class and let $\mathcal{M}$ be a (ϵ,δ)-DP mechanism that takes as input a dataset of size $m = m(\alpha,\beta,\epsilon,\delta)$ and outputs some $h \in \mathcal{H}$. Then, for any $T \subset \mathcal{H}$, if we denote $\bar{X}$ (resp. $\bar{X}'$) as the result of m i.i.d. draws from S (resp. S') using μ (resp. μ'), then, setting $\epsilon^* = \frac{6\epsilon m}{\kappa n}$ and $\delta^* = \frac{4me^{\epsilon^*}\delta}{\kappa n}$ we have that*

$$\Pr_{\bar{X}\sim\mu^m}[\mathcal{M}(\bar{X}) \in T] \le e^{\epsilon^*} \Pr_{\bar{X}\sim\mu'^m}[\mathcal{M}(\bar{X}') \in T] + \delta^*$$

Proof. This follows immediately from applying Lemma 6.1 in (Karwa & Vadhan, 2018) to this setting, using the bound $\mathrm{TV}(\mu,\mu') \le 1/\kappa n$ proven by Bun et al. (2020). □

A reader familiar with sample complexity bounds in DP-literature, knows that usually the dependency of m in ϵ is inverse. Thus, the dependency of ϵ^* in ϵm suggests that ϵ^* ends up independent of the privacy-loss of parameter set to the private learning mechanism $\mathcal{M}$. That is why chose to apply $\mathcal{M}$ with $\epsilon = 1$, a parameter under which most private and non-private sample complexity bounds are asymptotically equivalent.[6] The flip side of it is that we now set n to be proportional to ϵ^{-1}. We are now ready to give our DP algorithm for transfer learning in the hybrid model.

Theorem 4.2. *Using the same notation as in Algorithm 2, Algorithm 2 is a hybrid-model (ϵ,δ)-DP algorithm provided $n \ge \frac{m_1\sqrt{288T\ln(2/\delta)}}{\epsilon\kappa} = \Omega(m_1\frac{(\chi^2+1)\sqrt{\log((\chi^2+1)/\alpha)\log(1/\delta)}}{\alpha^2\epsilon})$.*

Proof. First, it is clear that each local-user is asked a single query and replies using a mechanism that is (ϵ,δ)-DP. As for the privacy of the curator-agents, by setting $m_1 = m(\alpha, \beta/T, 1, \frac{\kappa\delta}{8eT})$, Claim 4.1 asserts that in each iteration we are (ϵ^*,δ^*)-DP for $\epsilon^* = \frac{6\cdot 1\cdot m_1}{\kappa n} \le \frac{\epsilon}{\sqrt{8T\ln(2/\delta)}} < 1$ and $\delta^* \le \frac{4m_1e^1}{\kappa n}\cdot\frac{\delta}{4\epsilon\sqrt{2T}} \le \frac{4e\cdot\epsilon}{6\sqrt{8T}}\cdot\frac{\delta}{\epsilon\sqrt{2T}} \le \frac{\delta}{2T}$. Applying the Advanced Composition theorem of (Dwork et al., 2010) we get that in all T iterations together we are (ϵ,δ)-DP w.r.t. to each of the n curator-agents. □

[6] Loosely speaking, a parameter under which we typically get "privacy for free."

Algorithm 2 Private Subsample-Test-Reweigh

Input: parameters $0 < \alpha,\beta < 1/8$, $0 < \epsilon,\delta$. A (ϵ,δ)-DP learning algorithm $\mathcal{M}$ of sample complexity $m(\alpha,\beta,\epsilon,\delta)$.

Set $\kappa \leftarrow \frac{\alpha}{8(\chi^2+1)}$, $T \leftarrow \frac{128\log_2(\frac{8(\chi^2+1)}{\alpha})}{\alpha^2}$, $N_0 \leftarrow \frac{4\ln(2/\delta)\ln(8T/\beta)}{\epsilon^2\alpha^2}$.

Draw a sample of $n \ge m(\alpha,\frac{\beta}{T},1,\frac{\delta}{\epsilon\sqrt{2T}})\frac{\sqrt{288T\ln(2/\delta)}}{\epsilon\kappa}$ points i.i.d. from S, denoted $x_1,x_2,...,x_n$, all labeled by ℓ. Similarly draw $N_0\cdot T$ local-users from T.

Set weight $w_i^1 \leftarrow \kappa$ for each $1 \le i \le n$.

for $(t = 1,2,3,...,T)$ **do**

  Set μ^t as a distribution where $\mu^t = \frac{\Pi_\kappa(\bar{w})}{\|\Pi_\kappa(\bar{w})\|_1}$.

  Apply $\mathcal{M}$ to a sample of $m_1 = m(\alpha,\frac{\beta}{T},1,\frac{\delta}{\epsilon\sqrt{2T}})$ examples drawn i.i.d. from μ^t, to obtain some hypo. h^t.

  Pick arbitrarily a new batch B of N_0 local-users, and set $a^t \leftarrow \frac{1}{N_0}\sum_{x\in B} RR_{\epsilon,\delta}\big(\mathrm{L}(h^t(x),\ell(x))\big)$.

  if $(a^t > 3\alpha)$ **then**

    $\forall i$ set $w_i^{t+1} \leftarrow w_i^t\cdot\exp(-\alpha/8\cdot[1-\mathrm{L}(h^t(x_i),\ell(x_i))])$

  else

    return h^t (and halt)

  end if

end for

Theorem 4.3. *W.p.$\ge 1-3\beta$, Algorithm 2 returns an hypothesis h such that $\mathrm{err}_T(h) \le 4\alpha + \alpha_{\mathcal{H}}$, provided the number of curator-agents is $n = \Omega((\chi^2+1)\frac{\log(1/\alpha)}{\alpha^2}(d\log(\frac{d}{\alpha}) + \log(\frac{1}{\beta})))$, and the number of local-agents is $N = N_0T = \Omega(\frac{\log((\chi^2+1)/\alpha)\log(\log(\chi^2+1)/\alpha\beta)}{\epsilon^2\alpha^4})$.*

Proof. Again, the proof relies on the characterization of the utility of the MW-mechanism w.r.t. k-dense set of weights. Again, let $\bar{w}^1$ be the initial weights vector where $w_i^1 = \kappa$ for all i. Let $\bar{w}^* \in [0,1]^n$ be any fixed set of weights which is κ-dense, and denote $\hat{w} = \frac{\bar{w}^*}{\|\bar{w}^*\|_1}$ as its induced distribution. Then, we have that for any sequence of costs $\bar{m}^1,\bar{m}^2,...,\bar{m}^t$, Bun et al. (2020) proved that this MW-mechanism with learning rate of $\eta = \frac{\alpha}{8}$ guarantees that

$$\sum_t \mathop{\mathbb{E}}_{i\sim\mu^t}[m_i^t] \le \sum_t \mathop{\mathbb{E}}_{i\sim\hat{w}}[m_i^t] + \frac{\alpha T}{8} + \frac{\frac{1}{\kappa n}\phi(\bar{w}^*,\bar{w}^1)}{\alpha/8} \quad (3)$$

where $\phi(\bar{w}^*,\bar{w}^1)$ is the Bregman divergence induced by the entropy function, namely

$$\phi(\bar{w}^*,\bar{w}^1) = \sum_i w_i^*\log(\frac{w_i^*}{w_i^1}) - w_i^* + w_i^1$$

As $\bar{w}^*$ we aim to use the weights which Lemma 3.2 guarantees whose nice properties hold w.p.$\ge 1-\beta$. Thus, we set for each x_i drawn from S the weight $w_i^* = \frac{u(x_i)}{4(\chi^2+1)/\alpha} \le 1$. As Lemma 3.2 asserts, it holds that $U = \sum u(x_i)$

lies in the range $[(1-{}^{3\alpha}/_{8})n, (1+{}^{\alpha}/_{8})n]$, thus $w_{\text{avg}} = \frac{1}{n}\sum \frac{\alpha u(x_i)}{4(\chi^2+1)} \geq \frac{\alpha(1-3\alpha/8)}{4(\chi^2+1)}$. Thus, by setting $\kappa = \frac{\alpha}{8(\chi^2+1)}$ we have that $w_{\text{avg}} \geq \kappa \cdot 1 \geq \kappa w_{\max}$, implying w^* is indeed κ-dense. Also, the same bounds on U imply that

$$\begin{aligned}\phi(\bar{w}^*, \bar{w}^1) &= \sum_i 2\kappa u(x_i)\log(\frac{2\kappa u(x_i)}{\kappa}) - 2\kappa u(x_i) + \kappa \\ &= \kappa\sum_i (2u(x_i)\log(2u(x_i)) - 2u(x_i) + 1) \\ &\leq \kappa\big(n - 2U + 2\sum_i u(x_i)\log({}^{8(\chi^2+1)}/_{\alpha})\big) \\ &\leq 2\kappa\log({}^{8(\chi^2+1)}/_{\alpha})\cdot U \leq 4\kappa n\log({}^{8(\chi^2+1)}/_{\alpha})\end{aligned}$$

The remainder of the proof is as in Theorem 3.1. In each iteration it must hold that $\text{err}_{\mu^t}(h^t) \leq \alpha$; whereas $a^t \geq 3\alpha$, which w.p.$\geq 1-\beta$ — using classic utility bounds for the Randomized Response mechanism — implies that $\text{err}_{\mathsf{T}}(h^t) \geq 2\alpha$. Plugging this bound to Equation (3) stating the regret of the κ-dense MW-algorithm we get

$$\textstyle\sum_t(1-\alpha)] \leq \sum_t[1-(2\alpha-\frac{5\alpha}{8})] + \frac{\alpha T}{8} + \frac{4\log_2({}^{8(\chi^2+1)}/_{\alpha})}{\alpha/8}$$

Rearranging yields the bound $\frac{3\alpha T}{8} \leq \frac{\alpha T}{8} + \frac{32\log_2({}^{8(\chi^2+1)}/_{\alpha})}{\alpha}$ which implies that we halt in $T \leq \frac{128\log_2({}^{8(\chi^2+1)}/_{\alpha})}{\alpha^2}$ iterations. Again, upon halting $a^t \leq 3\alpha$ so it holds $\text{err}_{\mathsf{T}}(h^t) \leq 4\alpha$. □

Thus in a realizable setting, given a finite $\mathcal{H}$, we can set $\mathcal{M}$ as the exponential mechanism over $\mathcal{H}$ returns w.p. $\geq 1-\beta$ an hypothesis of error $\Theta(\alpha)$ when the sample size it at least $\Omega({}^{\ln(|\mathcal{H}|)}/_{\alpha\epsilon})$. We thus obtain the following corollary.

Corollary 4.4. *For any* S, T *with bounded* χ^2*-divergence, there exists a* (ϵ,δ)*-DP hybrid-model learning for a finite* $\mathcal{H}$ *in the realizable case which returns w.p.* $\geq 1-\beta$ *an hypothesis* $h \in \mathcal{H}$ *with* $\text{err}_{\mathsf{T}}(h) = \Theta(\alpha)$*, provided that we have* $n = \tilde{\Omega}(\frac{(\chi^2+1)\ln(|\mathcal{H}|/\beta)}{\alpha^3\epsilon})$ *curator-agents drawn from* S *and* $N = \tilde{\Omega}(\frac{\ln({}^{(\chi^2+1)}/_{\alpha})}{\epsilon^2\alpha^4})$ *local-agents drawn from* T.

Example: Sparse Hypothses. It is worth noting that our sample complexity bounds are independent of the dimension d (assuming the domain $\mathcal{X} \subset \mathbb{R}^d$). So consider a specific case that deals with a s-sparse problem over a high d-dimensional set where $s \ll d$, say, where $\mathcal{H}$ is a linear seprartor that uses no more than $s = O(1)$ features out of the d features of each example. This suggests that $|\mathcal{H}| = d^{O(s)} = \text{poly}(d)$. We thus obtain $s\cdot\text{polylog(d)}$ sample complexity bounds (setting all other parameters to be come reasonable constants.) whereas learning solely over T requires a sample complexity $\geq d$ (see Duchi et al. (2013)).

Private SGD. The specific case where the private learning mechanism $\mathcal{M}$ is SGD discussed in Appendix A. This case is discussed separately for two main reasons. First, its algorithmic presentation is slightly different than in Algorithm 2, since in each iteration it makes successive draws from μ^t rather than a single draw of a subsample. Secondly, this is the one canonical case where L is continuous (rather than binary), and so the subject of scaling comes into play. Whereas thus far we assumed L is bounded by 1, it is straightforward to see that a L-Lipfshitz convex loss function over a convex set of diameter D has loss $\in [0, LD]$, thus our goal is now to obtain a loss $\leq \alpha$ (rather than $\alpha\cdot LD$). Lastly, aiming to give a tight analysis, we use the notion of (ρ,ω)-tCDP (Bun et al., 2018) which is then converted to (ϵ,δ)-DP scheme. Nonetheless, the sample complexity bounds we get are similar to those of Algorithm 2.

5. On Reducing Sample-Complexity Bounds

While our work is the first to show the feasibility of transfer learning using poly-size sample, its sample complexity bound for the curator-agents is a large multiplicative factor $\tilde{O}(\epsilon^{-1}\alpha^{-2})$ over the sample complexity of a (non-transfer) curator-model learner. We believe that it is possible to significantly improve said bound, at least for particular instances. Here we provide two specific instances where indeed the number of required iterations until convergence of our algorithm is $o(\alpha^{-2})$, leading to a much smaller sample complexity.

Transfer Learning for PARITY under the Uniform Distribution. Consider the domain $\mathcal{X} = \{0,1\}^d$ and the class of PARITY functions, where for any $S \subset [d]$ we have $c_S(x) = \bigoplus_{i\in S} x_i$. It is a well-known result that under the uniform distribution the PARITY class cannot be learnt in the local-model unless the number of local-agents is $N = \exp(d)$, yet a sample of size $n = \Theta({}^{d}/_{\epsilon\alpha})$ suffices to learn PARITY under any distribution in the curator-model (Kasiviswanathan et al., 2008). In Appendix B we show that in the hybrid-model one can learn PARITY in a single iteration, provided S and the uniform distribution T have polynomial χ^2-divergence. The crux of the proof lies in proving that w.h.p. a sufficiently large sample drawn from S is linearly independent over $\mathbb{F}_2^d$. This suggests that the private curator-model learner for PARITY (Kasiviswanathan et al., 2008) — which leverages on (multiple) Gaussian elimination over $\mathbb{F}_2^d$ — returns w.h.p. the true labeling function in the PARITY hypothesis class.

Empirical Experiments: When Both S and T are Gaussians. Next, we show empirically that in other settings, both the number of required iterations until convergence and our sample complexity bounds are far greater than required. First, we consider a setting where S is a simple spherical Gaussian in $d = 500$-dimensions, $\mathsf{S} = \mathcal{N}(\bar{0}, I_d)$ whereas for T we picked an arbitrary set of $k = 10$ coordinates and

set the standard deviation on these k as $\sigma = 0.02$ whereas the remaining $d-k$ coordinates have standard deviation of 1, i.e. $\mathsf{T} = \mathcal{N}(\bar{0}, I_{d-k} \otimes \sigma^2 I_k)$. It is a matter of simple calculation to show that $\chi^2(\mathsf{T}\|\mathsf{S}) + 1 = (\frac{1}{\sigma^2(2-\sigma^2)})^{k/2} > 3 \cdot 10^{15}$. Now, our target hyperplane separator is set such that its only non-zero coordinates are the k ones on which S and T have a different variance, and so that is classifies precisely $\alpha = 0.01$ of the mass of T as -1. Now, testing this setting with the *non-private version*[7] (Algorithm 1) over $n \geq 90,000$ we obtain an hypothesis of error $\leq 2\alpha$ in all $t = 50$ repetitions of our experiment. Moreover, our iterative algorithm runs for only $T \approx 1200 - 1300$ iterations, far below the $\approx \alpha^{-2}$ upper bound. The results appear in Figures 1a and 1b in Appendix C. Second, we consider a setting where the true hypothesis actually corresponds to a k-dimensional ball over the k coordinates that are more concentrated in T than in S and ran both the private and non-private version of our algorithm. While the non-private version converged very fast, the private version required a very large sample of curator-agents, and so we were able to conduct only preliminary experiments with it. The experiment is detailed in Appendix D where the results appear in Figures 2 and 3.

Open Problems. Our work is the first to present a general framework for transfer learning in the hybrid-model, thus providing an initial answer to the open problem of "selection bias" posed by Beimel et al. (2020). While our framework does surpasses certain naïve baselines in some specific cases, there is still much work to be done in order to reduce its sample complexity. Considering different divergences can be a promising direction, especially if we know that the 4th moment of the IS weights is bounded (as it may increase the value of κ). A different venue can be the study of repeated uses of Subsample-Test-Reweigh — when person A applied the paradigm until she finds a good hypothesis, then hands over her last set of weights to Person B who uses it for learning a different hypothesis. Lastly, we believe that there's more to be studied in general in the intersection between DP and transfer learning, as the problem can be tackled from the alternative approach of *discrepancy* between hypothesis classes (Ben-David et al., 2010; Mansour et al., 2009).

Acknowledgements

This work was done when the first author was advised by the second author. O.S. is supported by the BIU Center for Research in Applied Cryptography and Cyber Security in conjunction with the Israel National Cyber Bureau in the Prime Minister's Office, and by ISF grant no. 2559/20. Both authors thank the anonymous reviewers for many helpful suggestions in improving this paper.

[7] We thoroughly apologize, but experimenting with the private version becomes infeasible on a desktop computer due to its sample complexity constraints.

References

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. Importance sampling: Intrinsic dimension and computational cost, 2017.

Arora, S., Hazan, E., and Kale, S. The multiplicative weights update method: a meta-algorithm and applications. *Theory Comput.*, 8(1):121–164, 2012.

Avent, B., Korolova, A., Zeber, D., Hovden, T., and Livshits, B. BLENDER: Enabling local search with a hybrid differential privacy model. In *USENIX Security*, pp. 747–764. USENIX Association, 2017.

Balle, B., Bell, J., Gascón, A., and Nissim, K. The privacy blanket of the shuffle model. In *CRYPTO*, volume 11693, pp. 638–667. Springer, 2019.

Balle, B., Bell, J., Gascón, A., and Nissim, K. Private summation in the multi-message shuffle model. In *CCS*, pp. 657–676. ACM, 2020.

Bassily, R. and Smith, A. D. Local, private, efficient protocols for succinct histograms. In *STOC*, pp. 127–135. ACM, 2015.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*, 2014.

Beimel, A., Nissim, K., and Omri, E. Distributed private data analysis: Simultaneously solving how and what. In *CRYPTO*, volume 5157, pp. 451–468. Springer, 2008.

Beimel, A., Korolova, A., Nissim, K., Sheffet, O., and Stemmer, U. The power of synergy in differential privacy: Combining a small curator with local randomizers. In *ITC*, volume 163 of *LIPIcs*, pp. 14:1–14:25. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Bernstein, S. Collected works ii. *Moscow: Akad. Nauk SSSR*, 1954.

Bittau, A., Erlingsson, Ú., Maniatis, P., Mironov, I., Raghunathan, A., Lie, D., Rudominer, M., Kode, U., Tinnés, J., and Seefeld, B. Prochlo: Strong privacy for analytics in the crowd. In *SOSP*, pp. 441–459. ACM, 2017.

Blum, A., Furst, M. L., Jackson, J. C., Kearns, M. J., Mansour, Y., and Rudich, S. Weakly learning DNF and characterizing statistical query learning using fourier analysis. In *STOC*, pp. 253–262. ACM, 1994.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *TCC*, volume 9985 of *Lecture Notes in Computer Science*, pp. 635–658, 2016.

Bun, M., Dwork, C., Rothblum, G. N., and Steinke, T. Composable and versatile privacy via truncated CDP. In *STOC*, pp. 74–86. ACM, 2018.

Bun, M., Carmosino, M. L., and Sorrell, J. Efficient, noise-tolerant, and private learning via boosting. In *COLT*, volume 125, pp. 1031–1077. PMLR, 2020.

Chatterjee, S. and Diaconis, P. The sample size required in importance sampling, 2017.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12, 2011.

Cheu, A., Smith, A. D., Ullman, J. R., Zeber, D., and Zhilyaev, M. Distributed differential privacy via shuffling. In *EUROCRYPT*, volume 11476 of *Lecture Notes in Computer Science*, pp. 375–403. Springer, 2019.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *NIPS*, pp. 442–450. Curran Associates, Inc., 2010.

Csiszár, I. and Shields, P. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

Duchi, J. C. and Rogers, R. Lower bounds for locally private estimation via communication complexity. In Beygelzimer, A. and Hsu, D. (eds.), *COLT*, volume 99, pp. 1161–1191. PMLR, 2019.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *FOCS*, pp. 429–438. IEEE Computer Society, 2013.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006a.

Dwork, C., Mcsherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006b.

Dwork, C., Rothblum, G., and Vadhan, S. Boosting and differential privacy. In *FOCS*, 2010.

Ghazi, B., Golowich, N., Kumar, R., Pagh, R., and Velingker, A. On the power of multiple anonymous messages: Frequency estimation and selection in the shuffle model of differential privacy. In *EUROCRYPT*, volume 12698, pp. 463–488. Springer, 2021a.

Ghazi, B., Kumar, R., Manurangsi, P., Pagh, R., and Sinha, A. Differentially private aggregation in the shuffle model: Almost central accuracy in almost a single message. In *ICML*, volume 139, pp. 3692–3701. PMLR, 2021b.

Gupta, A., Hardt, M., Roth, A., and Ullman, J. Privately releasing conjunctions and the statistical query barrier. In *STOC*, 2011.

Gupta, S. K., Rana, S., and Venkatesh, S. Differentially private multi-task learning. In *Intelligence and Security Informatics - 11th Pacific Asia Workshop, PAISI 2016, Auckland, New Zealand, April 19, 2016, Proceedings*, volume 9650 of *Lecture Notes in Computer Science*, pp. 101–113. Springer, 2016.

Hardt, M. and Rothblum, G. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 61–70. IEEE, 2010.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Joseph, M., Mao, J., Neel, S., and Roth, A. The role of interactivity in local differential privacy. In *FOCS*, pp. 94–105. IEEE Computer Society, 2019.

Karwa, V. and Vadhan, S. P. Finite sample differentially private confidence intervals. In Karlin, A. R. (ed.), *ITCS*, volume 94 of *LIPIcs*, pp. 44:1–44:9, 2018.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? In *FOCS*, 2008.

Kearns, M. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, 11 1998.

Kumar, M. Differentially private transferrable deep learning with membership-mappings, 2022.

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

Maia Polo, F. and Vicente, R. Effective sample size, dimensionality, and generalization in covariate shift adaptation. *Neural Computing and Applications*, pp. 1–13, 2022.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms, 2009.

Merriman, C. Microsoft reminds privacy-concerned Windows 10 beta testers that they're volunteers. In *The Inquirer,* http://www.theinquirer.net/2374302, Oct 7, 2014.

Metelli, A. M., Russo, A., and Restelli, M. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Microsoft. Windows Insider Program Agreement. https://insider.windows.com/en-us/program-agreement, Sep 15, 2017.

Mironov, I. Rényi differential privacy. In *CSF*, pp. 263–275, 2017.

Mozilla. Firefox Privacy Notice. https://www.mozilla.org/en-US/privacy/firefox/#pre-release, June 4, 2019.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM Symposium on Theory of Computing*, pp. 75–84. ACM, 2007. Full version in: http://www.cse.psu.edu/~asmith/pubs/NRS07.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, Ú. Scalable private learning with PATE. In *ICLR*. OpenReview.net, 2018.

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

Valiant, L. G. A theory of the learnable. *Commun. ACM*, 27 (11):1134–1142, 11 1984.

Wang, Y., Gu, Q., and Brown, D. E. Differentially private hypothesis transfer learning. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part II*, volume 11052 of *Lecture Notes in Computer Science*, pp. 811–826. Springer, 2018.

Warner, S. L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309), March 1965.

Xie, L., Baytas, I. M., Lin, K., and Zhou, J. Privacy-preserving distributed multi-task learning with asynchronous updates. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pp. 1195–1204. ACM, 2017.

Yao, Y. and Doretto, G. Boosting for transfer learning with multiple sources. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1855–1862, 2010. doi: 10.1109/CVPR.2010.5539857.

A. Private Stochastic Gradient Descent

The Private SGD Algorithm. In this section, we discuss our private Subsample-Test-Reweigh paradigm where the private learning mechanism applied in each iteration is the standard private SGD (see Hazan (2016); Bassily et al. (2014)). For simplicity, we give the algorithm here.

Algorithm 3 Online SGD

Input: Parameters α, β, σ Lipfshitz constant L and a convex set $\mathcal{H} \subset \mathbb{R}^d$ of diameter D. A distribution μ over a sample S.
Let $h^1 \in \mathcal{H}$ arbitrarily.
Set $R \leftarrow \max\left\{\frac{9D^2(L^2+\sigma^2 d)}{\alpha^2}, \frac{8D^2L^2\ln(2/\beta)}{\alpha^2}\right\}$; or set $R \leftarrow \max\left\{\frac{4(L^2+\sigma^2 d)}{\lambda\alpha}\ln(\frac{L^2+\sigma^2 d}{\lambda\alpha}), \frac{8D^2L^2\ln(2/\beta)}{\alpha^2}\right\}$ if L is λ-strongly convex.
for $(r = 1, 2, 3, ..., R)$ **do**
  Draw a labeled example $(x_r, \ell(x_r)) \sim \mu$
  Draw a random vector $v \sim \mathcal{N}(0, \sigma^2 I_d)$
  Set $\eta^r \leftarrow \frac{D}{\sigma\sqrt{r}}$; or set $\eta^r \leftarrow \frac{D}{\lambda r}$ if L is λ-strongly convex.
  Set $h^{r+1} \leftarrow \Pi_{\mathcal{H}}\big(h^r - \eta^r(\nabla \mathrm{L}(h^r(x_r), \ell(x_r)) + v)\big)$
end for
Return $\bar{h} = \frac{1}{R}\sum_{r=1}^{R} h^r$.

Standard arguments from Hazan (2016) give the following utility theorem.

Theorem A.1. *Let $\mathcal{H} \subset \mathbb{R}^d$ be a convex set of diameter D and let* L *be a L-Lipfshitz function and denote $\alpha_{\mathcal{H}} = \min_{h\in\mathcal{H}} \mathbb{E}_{(x,\ell(x))\sim\mu}[\mathrm{L}(h(x), \ell(x))]$. Then, w.p. $\geq 1-\beta$ after R iterations,* $\mathrm{err}(\bar{h}) \leq \alpha_{\mathcal{H}} + \alpha$.

Proof. The proof follows from the usual analysis of SGD, under the observation that in each iterations

$$\mathbb{E}_{\substack{(x_r,\ell(x_r))\sim\mu \\ v\sim\mathcal{N}(0,\sigma^2 I_d)}}[\nabla \mathrm{L}(h^r(x_r), \ell(x_r)) + v] = \mathbb{E}_{(x,y)\sim\mu}[\nabla \mathrm{L}(h^r(x), \ell(x))]$$

and that

$$\mathbb{E}[\|\nabla \mathrm{L}(h^r(x_r), \ell(x_r)) + v\|^2] \leq \mathbb{E}[\|\nabla \mathrm{L}(h^r(x_r), \ell(x_r))\|^2] + 2\,\mathbb{E}[\nabla \mathrm{L}(h^r(x_r), \ell(x_r)) \cdot v] + \mathbb{E}[\|v\|^2] \leq L^2 + \sigma^2 d$$

Plugging those into the bounds of the generalization properties of the SGD for convex functions we obtain:

$$\underset{\mu}{\mathrm{err}}(\bar{h}) \leq \underset{\mu}{\mathrm{err}}(h^*) + \frac{3D\sqrt{(L^2+d\sigma^2)}}{2\sqrt{R}} + LD\sqrt{\frac{8\ln(2/\beta)}{R}}$$

So setting $R \geq \max\left\{\frac{9D^2(L^2+\sigma^2 d)}{\alpha^2}, \frac{8D^2L^2\ln(2/\beta)}{\alpha^2}\right\}$ yields that $\mathrm{err}_\mu(\bar{h}) \leq \alpha_{\mathcal{H}} + \frac{\alpha}{2} + \frac{\alpha}{2}$ as required. Similarly, for a λ-strongly convex function we get

$$\underset{\mu}{\mathrm{err}}(\bar{h}) \leq \underset{\mu}{\mathrm{err}}(h^*) + \frac{(L^2+d\sigma^2)(1+\ln(T))}{2\lambda R} + LD\sqrt{\frac{8\ln(2/\beta)}{R}}$$

Algorithm 4 Private Subsample-Test-Reweigh with SGD

Input: parameters $0 < \alpha, \beta < 1/8$, $0 < \epsilon < 1$,$0 < \delta < 1/4$. Private SGD mechanism over a convex set $\mathcal{H} \in \mathbb{R}^d$ of diameter D with convex L-Lipfshitz loss function L making at most R SGD-iterations.

Set $\kappa \leftarrow \frac{\alpha}{8(\chi^2+1)}$, $T \leftarrow \frac{128L^2D^2\log_2(\frac{8(\chi^2+1)}{\alpha})}{\alpha^2}$, $N_0 \leftarrow \frac{4\ln(2/\delta)\ln(8T/\beta)}{\epsilon^2\alpha^2}$, $\sigma^2 \leftarrow \max\left\{20L^2, \frac{16\epsilon^{-1}L^2\ln(1/\delta)+8L^2}{\ln(\kappa n)}\right\}$.
Set $R \leftarrow \max\left\{\frac{9D^2(L^2+\sigma^2 d)}{\alpha^2}, \frac{8D^2L^2\ln(2/\beta)}{\alpha^2}\right\}$; or set $R \leftarrow \max\left\{\frac{4(L^2+\sigma^2 d)}{\lambda\alpha}\ln(\frac{L^2+\sigma^2 d}{\lambda\alpha}), \frac{8D^2L^2\ln(2/\beta)}{\alpha^2}\right\}$ if L is λ-strongly convex.
Draw a sample of $n \geq \sqrt{\frac{52RT}{8\kappa^2\epsilon}\ln(\frac{52RT}{8\kappa^2\epsilon})}$ points i.i.d. from S, denoted $x_1, x_2, ..., x_n$, all labeled by ℓ.
Draw $N_0 \cdot T$ local-users from T.
Set weight $w_i^1 \leftarrow \kappa$ for each $1 \leq i \leq n$.
for $(t = 1, 2, 3, ..., T)$ **do**
  Set μ^t as a distribution where $\mu^t = \frac{\Pi_\kappa(\bar{w})}{\|\Pi_\kappa(\bar{w})\|_1}$.
  Apply private SGD using μ^t and using σ^2 for R iterations, and obtain an hypothesis h^t.
  Pick arbitrarily a new batch B of N_0 local-users, and set $a^t \leftarrow \frac{1}{N_0}\sum_{x\in B} RR_{\epsilon,\delta}\big(\mathrm{L}(h^t(x), \ell(x))\big)$.
  if $(a^t > 3\alpha + \alpha_{\mathcal{H}})$ **then**
    $\forall i$ set $w_i^{t+1} \leftarrow w_i^t \cdot \exp(-\alpha/8LD \cdot [LD - \mathrm{L}(h^t(x_i), \ell(x_i))])$
  else
    return h^t (and halt)
  end if
end for

So setting $R \geq \max\left\{\frac{4(L^2+\sigma^2 d)}{\lambda\alpha}\ln(\frac{L^2+\sigma^2 d}{\lambda\alpha}), \frac{8D^2L^2\ln(2/\beta)}{\alpha^2}\right\}$ yields that $\mathrm{err}_\mu(\bar{h}) \leq \alpha_{\mathcal{H}} + \alpha$ as required. □

While we can apply a privacy analysis of Algorithm 3, we table it to the privacy analysis of our full algorithm.

The Subsample-Test-Reweigh Using SGD. We now transition to the full analysis of STR when we apply SGD as an intermediate procedure.

The utility analysis of Algorithm 4 is just as the one of Algorithm 2 modulo the fact that the non-negative loss is now bounded by LD rather than 1, which implies we must increase the number of MW-iterations by L^2D^2. It requires a sample complexity of $n = \Omega(L^2D^2(\chi^2+1)\frac{\log(1/\alpha)}{\alpha^2}(d\log(\frac{d}{\alpha}) + \log(\frac{1}{\beta})))$ curator-agents, and $N = N_0T = \Omega(\frac{L^2D^2\log((\chi^2+1)/\alpha)\log(\log(\chi^2+1)/\alpha\beta)}{\epsilon^2\alpha^4})$ local-agents to return, w.p.$\geq 1-3\beta$ an hypothesis of error $\leq \alpha_{\mathcal{H}} + \alpha$. (The increase in the number of curator-agents is due to the fact that the loss is now in the range $[0, LD]$ rather than $[0, 1]$.) The privacy analysis of Algorithm 4 requires

we transition of (ρ, ω)-tCDP given in Bun et al. (2018). Its definition as well as some of its basic properties (proven in Bun & Steinke (2016); Mironov (2017); Bun et al. (2018)) are provided below.

Definition A.1. A mechanism $\mathcal{M}$ is said to be (ρ, ω)-tCDP if for two neighboring inputs I and I' the α-Reyni divergence of the two distributions is bounded: $D_\alpha(\mathcal{M}(I)||\mathcal{M}(I')) \leq \alpha\rho$ for any $1 < \alpha \leq \omega$.

Fact A.2.
- *Let f be a d-dimensional function of L_2-global sensitivity of $\max_{I,I' \text{ neighbors}} \|f(I) - f(I')\| \leq L$. Then the mechanism that outputs for any instance I an output drawn from $\mathcal{N}(f(I), \sigma^2 I_d)$ is $(\frac{2L^2}{\sigma^2}, \infty)$-tCDP.*
- *Let $\mathcal{M}_1, \mathcal{M}_2$ be two (ρ_1, ω_1)-tCDP (resp. (ρ_2, ω_2)-tCDP) mechanisms. Then the mechanism that applies them to the same instance but using independent coin toss for each is $(\rho_1 + \rho_2, \min\{\omega_1, \omega_2\})$-tCDP.*
- *A mechanism which is (ρ, ω)-tCDP is also $(\rho\omega + \frac{\ln(1/\delta)}{\omega - 1}, \delta)$-DP for any $\delta < 1/e$.*

Perhaps, however, the most important property of tCDP is that it is amplified by subsampling.

Theorem A.3. *[Thm. 12 of Bun et al. (2018) reworded] Fix $\rho \in (0, 0.1]$. Let s be a constant satisfying $\log(1/s) \geq 3\rho(2 + \ln(1/\rho))$. Let $\mathcal{M}$ be a (ρ, ∞)-tCDP mechanism. Let I and I' be two neighboring instance and let μ be a distribution where the probability of sampling the one different entry between the two instances is s. Then the mechanism that samples entries from the input and then applies $\mathcal{M}$ on the subsample is $(13s^2\rho, \frac{\log(1/s)}{4\rho})$-tCDP.*

Now, throughout the execution of Algorithm 4 it holds that we subsample a point and apply the Gaussian mechanism for RT iterations. In all of these iterations we apply a distribution where the probability of subsampling any point into $\mathcal{M}$ is at most $1/\kappa n$. Moreover, our private-SGD mechanism when applied to a L-lipfshitz loss function is $(\frac{2L^2}{\sigma^2}, \infty)$-tCDP. Thus, if $\frac{2L^2}{\sigma^2} \leq 0.1$ and if

$$\ln(\kappa n) \geq \frac{6L^2(2 + \ln(\sigma^2/2L^2))}{\sigma^2} \tag{4}$$

then all conditions of Theorem A.3 hold. This suggests that each time we execute $\mathcal{M}$ over a randomly drawn sample we are $(\frac{26L^2}{\sigma^2\kappa^2 n^2}, \frac{\sigma^2 \ln(\kappa n)}{8L^2})$-tCDP; thus, by composition, we are $(\frac{26L^2 RT}{\sigma^2\kappa^2 n^2}, \frac{\sigma^2 \ln(\kappa n)}{8L^2})$-tCDP. Thus, w.r.t. the curator-agents, we are (ϵ, δ)-DP for any $\delta < 1/4$ for

$$\epsilon = \frac{26RT \ln(\kappa n)}{8\kappa^2 n^2} + \frac{8L^2 \ln(1/\delta)}{\sigma^2 \ln(\kappa n) - 8L^2}$$

This suggests that in order to achieve (ϵ, δ)-DP w.r.t. the curator agents, we set $\sigma^2 = \frac{16\epsilon^{-1}L^2 \ln(1/\delta) + 8L^2}{\ln(\kappa n)}$, and we must set n so that $n \geq \sqrt{\frac{52RT[\ln(\kappa) + \ln(n)]}{8\kappa^2\epsilon}}$. Seeing as $\kappa < 1$ it thus suffices for us to set $n = \sqrt{\frac{52RT}{8\kappa^2\epsilon} \ln(\frac{52RT}{8\kappa^2\epsilon})}$. It remains to check that under these values (4) holds; but indeed, note that $\frac{\sigma^2}{2L^2} \geq 10$ and so

$$\frac{3(2 + \ln(\sigma^2/2L^2))}{\sigma^2/2L^2} \leq \frac{3(2 + \ln(10))}{10} < 1.3$$

whereas under any reasonable set of parameters we get that $n \geq \sqrt{\frac{52RT}{8\kappa^2\epsilon} \ln(\frac{52RT}{8\kappa^2\epsilon})} > \frac{4}{\kappa}$, implying that $\ln(\kappa n) \geq \ln(4) > 1.3$.

Theorem A.4. *For any given $\epsilon \in (0, 1)$, $\delta \in (0, 1/4)$, $0 < \alpha < 1/8$, $0 < \beta < 1/3$, two distribution S, T with bounded χ^2-divergence and a convex loss-function L over a convex set $\mathcal{H} \subset \mathbb{R}^d$ of diameter D, we have that Algorithm 4 is (ϵ, δ)-DP algorithm in the hybrid model provided that*

$$n = \tilde{\Omega}\left(\frac{(\chi^2 + 1)LD^2\sqrt{L^2 + \sigma^2 d} \cdot \sqrt{\ln(2/\beta)}}{\alpha^3\sqrt{\epsilon}}\right)$$
$$= \tilde{\Omega}\left(\frac{(\chi^2 + 1)D^2L^2\sqrt{d\ln(1/\delta)} \cdot \sqrt{\ln(2/\beta)}}{\alpha^3\epsilon}\right)$$

if the loss-function is L-Lipfshitz and convex, or provided that

$$n = \tilde{\Omega}\left(\frac{LD(\chi^2 + 1)\sqrt{\frac{L^2 + \sigma^2 d}{\lambda\alpha} + \frac{D^2L^2\ln(1/\beta)}{\alpha^2}}}{\alpha^2\sqrt{\epsilon}}\right)$$
$$= \tilde{\Omega}\left(\frac{LD(\chi^2 + 1)}{\alpha^2} \cdot \left(\frac{L\sqrt{d\ln(1/\delta)}}{\epsilon\sqrt{\alpha\lambda}} + \frac{DL\sqrt{\ln(1/\beta)}}{\sqrt{\epsilon\alpha}}\right)\right)$$

if the loss-function is L-Lipfshitz and λ-strongly convex. Furthermore, if the number of local-agents is $N = \Omega(\frac{L^2D^2\log((\chi^2+1)/\alpha)\log(\log(\chi^2+1)/\alpha\beta)}{\epsilon^2\alpha^4})$ then w.p. $\geq 1 - 3\beta$ it returns an hypothesis $h \in \mathcal{H}$ where $\mathrm{err}_{\mathsf{T}}(h) \leq \alpha_{\mathcal{H}} + 4\alpha$.

B. Transfer Learning for the PARITY-Problem Under the Uniform Distribution

Transfer Learning of PARITY for the uniform distribution. Consider the domain $\mathcal{X} = \{0, 1\}^d$ and the class of PARITY functions, where for any $S \subset [d]$ we have $c_S(x) = \bigoplus_{i \in S} x_i$. It is a well-known result that under the uniform distribution the PARITY class cannot be learnt in the local-model unless the number of local-agents is $N = \exp(d)$, yet a sample of size $n = \Theta(d/\epsilon\alpha)$ suffices to learn PARITY under any distribution in the curator-model (Kasiviswanathan et al., 2008). Here we show that in the hybrid-model one can learn the PARITY class with a single iteration, provided S and the uniform distribution T have polynomial χ^2-divergence.

To establish this, we prove the following sequence of claims and corollaries.

Proposition B.1. *Fix $0 < \beta < 1/2$. Let S be a sample of $n \geq d\log_2(d/\beta)$ points drawn i.i.d. from the uniform distribution over $\{0,1\}^d$. Then the probability that S isn't linearly independent is $\leq \beta$.*

Proof. We prove the claim inductively as we iterate over all vectors in S. Due to simple counting argument, it is easy to see that the probability to draw an vector that is linearly dependent of given set of i vectors in a the d-dimension space $\mathbb{F}_2^d$ is at most 2^{i-d}. So fix any $0 \leq i \leq d-1$. At each step we have a set of i linearly independent vectors spanning a subspace of dimension i. We argue that the probability that among the next $t = \log_2(d/\beta)$ vectors in S the probability that all t vectors are linearly dependent of these i basis vectors is at most $(2^{i-d})^t \leq 2^{-t} = \beta/d$. And so, after d iterations we have found a set of d linearly independent vectors in S w.p. $\geq 1-\beta$. □

Claim B.2. *Fix $0 < \beta < 1/2$. Let T be the uniform distribution over $\{0,1\}^d$ and let S be a distribution over the same domain s.t. $\chi^2(\mathsf{T}\|\mathsf{S}) = \chi^2$ is finite. Let S be a sample of $n \geq 4(\chi^2+1)d\ln(d/\beta)$ points drawn i.i.d. from S. Then the probability that S isn't linearly independent is $\leq \beta$.*

Proof. The claim is proven in a similar inductive fashion to Proposition B.1. Fix any $0 \leq i \leq d-1$. At each step we have a set of i linearly independent vectors spanning a subspace of dimension i. We argue that the probability that among the next $t = \log_2(d/\beta)$ vectors in S the probability that all t vectors are linearly dependent of these i basis vectors is at most $\leq \beta/d$, from which the claim follows immediately.

So now, given i linearly independent vectors, let E be the event that we draw a vector not in their span. Under the uniform distribution $\Pr_{\mathsf{T}}[E] \geq 1 - 2^{i-d} \geq 1/2$. Standard bounds on the χ^2-divergence give that

$$\sqrt{\Pr_{\mathsf{S}}[E](\chi^2+1)} \geq \Pr_{\mathsf{T}}[E] \geq 1/2$$

implying that $\Pr_{\mathsf{S}}[E] \geq \frac{1}{4(\chi^2+1)}$ and so $\Pr_{\mathsf{S}}[\bar{E}] \leq 1 - \frac{1}{4(\chi^2+1)}$. It follows that the probability that among the next $t = 4(\chi^2+1)\ln(d/\beta)$ draws, not a single one lies outside the span of these i vectors is at most $(1-\frac{1}{4(\chi^2+1)})^t \leq \exp(-\frac{t}{4(\chi^2+1)}) = \beta/d$ as required. □

Corollary B.3. *Fix $\beta > 0$. Set $k = \log_{4/3}(2/\beta)$. Under the same notation as in Claim B.2 let $S_1, S_2, ..., S_k$ be k independently drawn batches from S s.t. each S_i contains at least $n \geq \frac{32(\chi^2+1)d\ln(2dk/\beta)}{\epsilon}$. Then w.p. $\geq 1-\beta$ it holds that when we drawn from each S_i a subsample S_i' where each $x \in S_i$ is put in the subsample w.p. $\epsilon/4$ independently of all other examples, then all S_i' are linearly independent.*

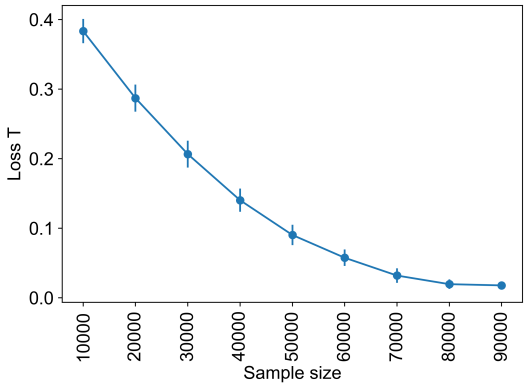
Proof. Straight-forward application of the Chernoff bound gives that w.p.$\geq 1-\beta/2$ each of the k S_i'-s contains at least $\epsilon/8|S_i|$ many points. This suffices for us to apply Claim B.2 and have that w.p. $\geq 1-\beta/2$ all S_i' are linearly independent. □

Based on Corollary B.3 we can now apply the same PARITY learning algorithm from Kasiviswanathan et al. (2008) on the kn curator-agents *just once* and then test its correctness over the N. This algorithm outputs for each S_i' either a $\perp$ or a solution in the affine subspace that solves a system of equations over $\mathbb{F}_2$. But due to the linear independence of each S_i', this solution must be the indicating vector of the relevant features of the true classifying function $c_S^* \in$ PARITY. It follows that for each S_i outputs the true classifying function w.p.$\geq 1/4$; and so, w.p. $\geq 1-\beta$ all S_i return either $\perp$ or c_S^* where at least one of the S_i-s outputs the true classifier. Note that the true classifier's loss – under any distribution S or T – must be 0. Using additional $O(\epsilon^{-2}\alpha^{-2})$ we can test and see that indeed we have an hypothesis $c \in$ PARTIY which is of small loss.

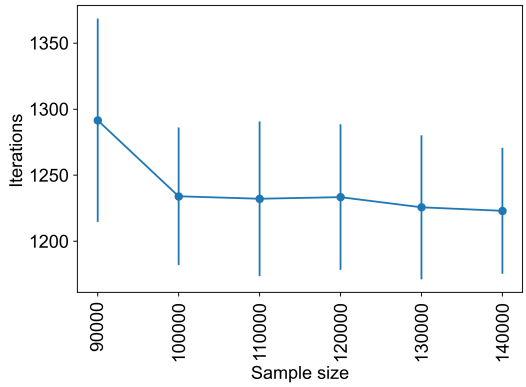
C. Experimental Evaluation of Non-Private Subsample-Test-Reweigh.

In this section, we show empirically that in another setting, both the number of required iterations until convergence and our sample complexity bounds are far greater than required. We consider a setting where S is a simple spherical Gaussian in $d = 500$-dimensions, $\mathsf{S} = \mathcal{N}(\bar{0}, I_d)$ whereas for T we picked an arbitrary set of $k = 10$ coordinates and set the standard deviation on these k as $\sigma = 0.02$ whereas the remaining $d-k$ coordinates have standard deviation of 1, i.e. $\mathsf{T} = \mathcal{N}(\bar{0}, I_{d-k} \otimes (0.01)^2 I_k)$. It is a matter of simple calculation to show that $\chi^2(\mathsf{T}\|\mathsf{S}) + 1 = (\frac{1}{\sigma^2(2-\sigma^2)})^{k/2} > 3 \cdot 10^{15}$. Now, true hypothesis is a hyperplane separator set on the k coordinates on which S and T have a different variance, so that is classifies precisely $\alpha = 0.01$ of the mass of T as -1.

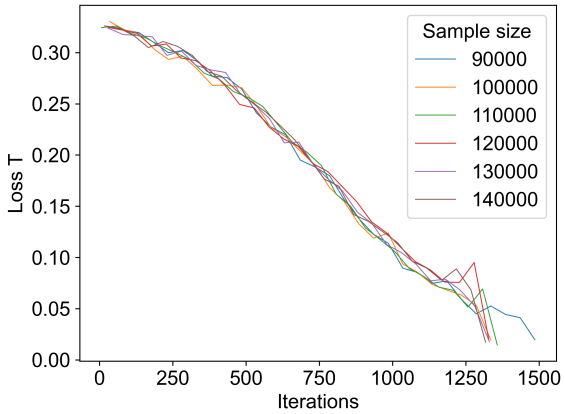
We applied the non-private version of our algorithm (Algorithm 1). In the non-private version, in order to learn in each iteration a hyperplane separator over a subsample of examples from S we used SVM, where our optimization goal is $\frac{1}{2}\|w\|^2 + C \cdot \sum_x \max\{0, 1-\langle w, x\rangle\}$ with a very large $C = 10^{30}$ (aiming to find an exact hyperplane as possible) over a subsample whose size is set to be $\frac{d+\ln(0.05/T)}{\alpha}$. First, aiming to find the sample complexity of the curator-agents that already yields an hypothesis of error $\leq 2\alpha$, we ran our experiments with varying values of $n = \{10^4, 2\cdot 10^4, ..., 8\cdot 10^4, 9\cdot 10^4\}$, repeating each experiment $t = 50$ times. We observed that for $n = 80,000$ we consistently return an hypothesis of small-loss. Results appear in Figure 1a. Then, for even larger values of



(a) **Loss on** T: The loss on distribution T until convergence (in samples 90K-140K) or (in samples 10K-80K) until arriving to early stopping condition (the average loss on T in last 200 iterations not improved in more 0.01 compared to the best average loss in the previous 200 iterations).



(b) **Iterations number**: The iterations number until convergence decreases with size of the sample.



(c) **Loss on** T **along the run**: the loss on T decreases with the iterations.

Figure 1: Empirical Experiment Results

$n = \{9 \cdot 10^4, 1 \cdot 10^5, .., 1.4 \cdot 10^5\}$ we ran our experiment to see at which iteration do we halt. For $n = 80,000$ we halt after ≈ 1300 iterations whereas for $n = 140,000$ we halt after $T \approx 1200$; yielding the rather surprising result that T isn't greatly affected by the increasing sample complexity. But regardless, this is very far below than the $O(\alpha^{-2})$-upper bound. Results appear in Figure 1b. In fact, looking at the error of the resulting hypothesis along the run itself returns roughly the same values, as seen in Figure 1c.

D. Experimental Evaluation of Private Subsample-Test-Reweigh

In order to implement the private algorithm, we used another example with bounded examples and hypotheses.

Similarly, to the previous experiment, we consider a setting where S is a simple spherical Gaussian in $d = 200$-dimensions, $\mathsf{S} = \mathcal{N}(\bar{0}, I_d)$ whereas for T we picked an arbitrary set of $k = 6$ coordinates and set the standard deviation on these k as $\sigma = 0.4$ whereas the remaining $d - k$ coordinates have standard deviation of 1, i.e. $\mathsf{T} = \mathcal{N}(\bar{0}, I_{d-k} \otimes (0.4)^2 I_k)$. It is a matter of simple calculation to show that $\chi^2(\mathsf{T}\|\mathsf{S}) + 1 = (\frac{1}{\sigma^2(2-\sigma^2)})^{k/2} > 39.1$.

However, we now set the true labeling function as one that *correlated* to points with large importance sampling weight. It is fairly simple to see that by looking at the k-coordinates with smaller variance in T, any origin-centered ball has more probability mass in T then in S. And so, our hypothesis class is the set of origin-centered ellipses $\mathcal{H} = \{w_1, ..., w_d, b \in [0,1] : \sum_{i=1}^{d} w_i x_i^2 \leq b\}$ where so that $x_i \sim \mathsf{S}$ and thus $\bar{y} \sim \chi_k^2$. The true hypothesis ℓ is the hyperplane separator with $\bar{w} = 0 \cdot I_{d-k} \otimes 1 \cdot I_k$ and $b = \sigma^2 r^0$, where r^0 is the a numerically set threshold under which a $\Pr_{X \sim \chi_k^2}[X < r^0] = 0.3$. Thus ℓ labels precisely 30% of the probability mass of T as -1, whereas it labels a significantly smaller fraction of S as -1.

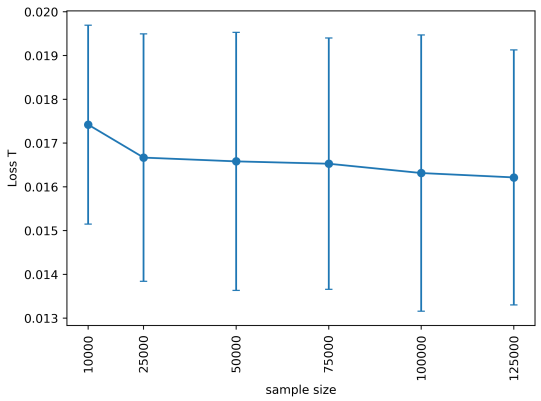
As the curator-model learning algorithm we used the online SGD as presented in Hazan (2016), after mapping each example $x \in \mathbb{R}^d$ to the vector $y \in \mathbb{R}_+^d$ where $y_i = x_i^2$ for each coordinate y. This allows us to use the Hinge-loss function - $\mathrm{L}(y) = \max\{0, \ell^*(y)(\langle w, y\rangle - b)\}$. The learning rate of the algorithm depends on a bounded diameter of the hypothesis and Lipschitzness of the loss that upper-bound the $\|\bar{y}\|$. So we set a value $B = 27.9$ - the empirically the max value of $\|\bar{y}\|$ of 90% of the examples drawn from S, and projected each example with norm $> B$ onto this simplex: $\{y \in \mathbb{R}_+^d : \|y\| \leq B\}$. Therefore, including the additional intercept coordinate, we get: $\|\nabla \mathrm{L}(\bar{y})\|^2 \leq B^2 + 1$, so we can use the Lipschitz parameter of $L = \sqrt{B^2 + 1}$. We also verify that the diameter is lower than $\sqrt{d+1}$ by verifying that $\sigma \leq 1/\sqrt{r^0}$.

Non-privately, we run the online SGD with a maximum of 10^4 iterations or until finding an exact hyperplane with a loss of at most $\alpha = 0.01$ with $\beta = {}^{10^{-6}}/_{T}$. Aiming to find the sample complexity of the curator-agents for which we get an hypothesis of error $\leq 2\alpha$ over T, we ran our experiments with varying values of $n = \{10K, 25K, 50K, 75K, 100K, 125K\}$, repeating each experiment $t = 50$ times. Note that all these values of n are below the worst-case bound (in our settings is $15 \cdot 10^{12}$), we consistently return a hypothesis of small-loss on T. Results appear in Figure 2a. Again, we see that the number of MW-iterations, T, isn't greatly affected by the increasing sample complexity, as seen in Figure 2c. Also in all values of n we halt after ≈ 1000 iterations (Results appear in Figure 2b) which is far below than the $O(\alpha^{-2})$-upper bound.

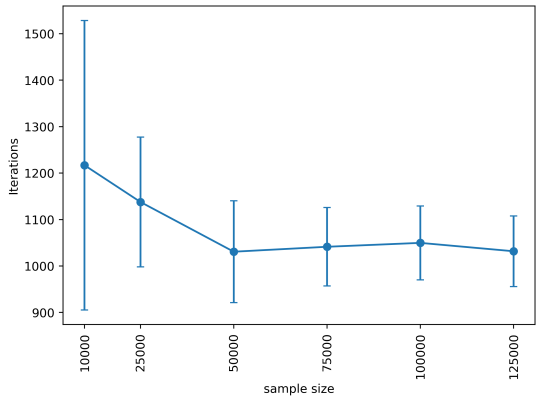
In the *private* version of the algorithm we applied Algorithm 4 with the private online SGD (Algorithm 3). The settings of the non-private version required a vast amount of memory because of the sample complexity bounds that are pretty big even for moderate values of α and ϵ, so we had to change some parameters and use fairly close distributions S and T. The changes from the non-private settings are as follows. We set the dimension to $d = 8$ and the number of "closer" coordinates to $k = 2$, so $\mathsf{S} = \mathcal{N}(\bar{0}, (0.1)^2 I_d)$ whereas $\mathsf{T} = \mathcal{N}(\bar{0}, (0.1)^2 I_{d-k} \otimes (0.07)^2 I_k)$. (It is a matter of simple calculation to show that $\chi^2(\mathsf{T}\|\mathsf{S}) + 1 = (\frac{\sigma_S^4}{\sigma_T^2(2\sigma_S^2 - \sigma_T^2)})^{k/2} > 1.35$, where σ_S, σ_T are the std in the k coordinates of S and T respectively.) We also set r^0 is a threshold for 0.4 of the examples, i.e. $\Pr_{X \sim \chi^2_k}[X < r^0] = 0.4$. We set $\alpha = 0.08$, and $\kappa = \frac{\alpha}{4(\chi^2+1)}$, and used the privacy parameters of $\epsilon = 0.5$ and $\delta = 0.0001$. Following similar calculations to before we set the bound on our hypothesis set's diameter as $B = 0.006$ and the Lipschitz value of $L = 0.1$.

We succeeded to get from the SGD a hypothesis whose loss is lower than $\alpha = 0.08$ with $\beta = {}^{10^{-6}}/_{T}$ after $50M$ iterations. So we run SGD with $R = 50M$, $R = 75M$, and $R = 100M$ iterations, repeating each experiment $t = 50$ time, to see their influence of them on the required number of MW-iterations. Our calculations lead to a sample complexity of $n = 324,700,000$ which we used in all runs. We can see (in Figure 3b) that the number of MW-iterations decreases as the iteration of SGD is increasing. In addition, the loss on T starts higher as the SGD iterations decrease (Figure 3c). However, in all these runs MW algorithm converged to 2α (Figure 3a).

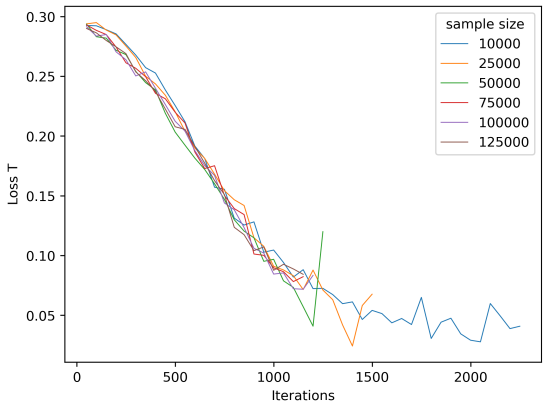
Due to the sample size, we were not able to experiment thoroughly with the private version of our algorithm, alas, our experiments do show that we succeed in implementing the algorithm. Furthermore, we also experimented with a SGD version which minimized the Lasso-regularized version of our algorithm. This version converged in a single iteration —



(a) **Loss on T**: The loss on distribution T until convergence.

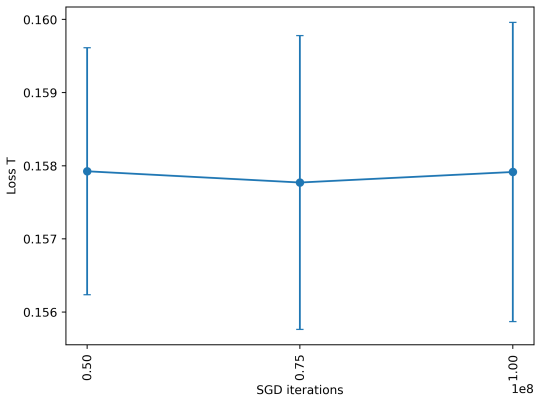


(b) **Iterations number**: The iterations number until the convergence.

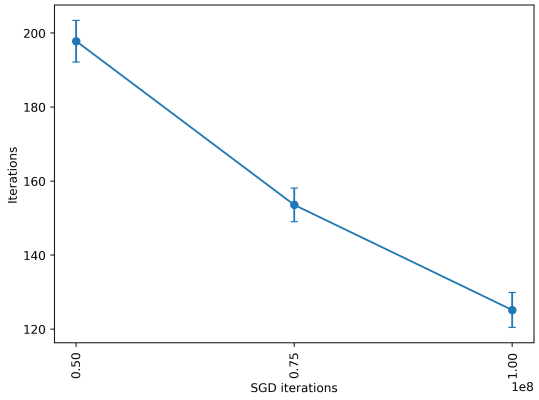


(c) **Loss on T along the run**: the loss on T decreases with the iterations.

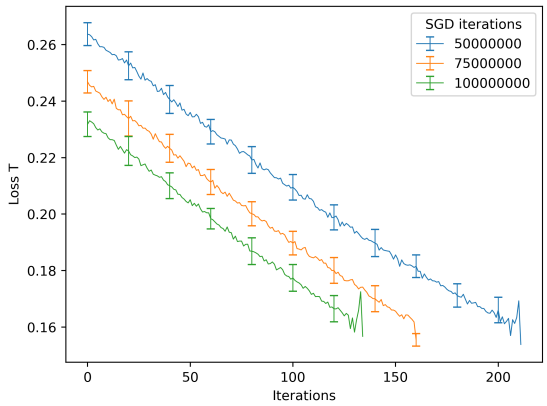
Figure 2: Empirical Experiment Results



(a) **Loss on** T: The loss on distribution T until convergence.



(b) **Iterations number**: The MW iterations number until convergence decreases as SGD iterations increases.



(c) **Loss on** T **along the run**: the loss on T decreases with the iterations.

Figure 3: Empirical Experiment Results

namely, it found the true separating coordinates on S. This is similar in spirit to the work of Avent et al. (2017) which also used the curator-agents to find the coordinates of the regression.