Private Realizable-to-Agnostic Transformation with Near-Optimal Sample Complexity

Bo Li BLI@CSE.UST.HK
The Hong Kong University of Science and Technology

Wei Wang WEIWA@CSE.UST.HK
The Hong Kong University of Science and Technology

Peng Ye PYEAC@CONNECT.UST.HK
The Hong Kong University of Science and Technology

Editors: Nika Haghtalab and Ankur Moitra

Abstract

The realizable-to-agnostic transformation (Beimel et al., 2015; Alon et al., 2020) provides a general mechanism to convert a private learner in the realizable setting (where the examples are labeled by some function in the concept class) to a private learner in the agnostic setting (where no assumptions are imposed on the data). Specifically, for any concept class $\mathcal{C}$ and error parameter α, a private realizable learner for $\mathcal{C}$ can be transformed into a private agnostic learner while only increasing the sample complexity by $\widetilde{O}(\mathrm{VC}(\mathcal{C})/\alpha^2)$, which is essentially tight assuming a constant privacy parameter $\varepsilon = \Theta(1)$. However, when ε can be arbitrary, one has to apply the standard privacy-amplification-by-subsampling technique (Kasiviswanathan et al., 2011), resulting in a suboptimal extra sample complexity of $\widetilde{O}(\mathrm{VC}(\mathcal{C})/\alpha^2\varepsilon)$ that involves a $1/\varepsilon$ factor.

In this work, we give an improved construction that eliminates the dependence on ε, thereby achieving a near-optimal extra sample complexity of $\widetilde{O}(\mathrm{VC}(\mathcal{C})/\alpha^2)$ for any $\varepsilon \le 1$. Moreover, our result reveals that in private agnostic learning, the privacy cost is only significant for the realizable part. We also leverage our technique to obtain a nearly tight sample complexity bound for the private prediction problem, resolving an open question posed by Dwork and Feldman (2018) and Dagan and Feldman (2020).

Keywords: Differential Privacy, Agnostic Learning, Private Prediction

1. Introduction

Differential privacy (DP) (Dwork et al., 2006b,a) has emerged as a popular notion for quantifying the disclosure of individual information and has been widely deployed to protect personal privacy (Apple Differential Privacy Team, 2017; Abowd, 2018). Informally, a randomized algorithm is said to be differentially private if changing a single input item does not significantly affect the output distribution. As a consequence, it safeguards against data inference through the algorithm's output.

Machine learning algorithms are usually trained on datasets that contain sensitive information, necessitating the need for privacy protection. The intersection of differential privacy and machine learning was first explored by Kasiviswanathan et al. (2011), who introduced *private learning* by integrating DP with two foundational learning models: probably approximately correct (PAC) learning (Vapnik and Chervonenkis, 1971; Valiant, 1984) and agnostic learning (Haussler, 1992; Kearns et al., 1994). The former assumes the data points are labeled by some function in a given concept class $\mathcal{C}$ and requires the learner to output a hypothesis with an error close to 0. This setting is often referred to as the *realizable* setting. In contrast, the latter is termed the *agnostic* setting, which can be seen as an extension of the realizable setting that removes the assumption on the existence of a labeling function. In agnostic learning, the output hypothesis is required to have an error close to optimal by any concept in $\mathcal{C}$, which might be much larger than 0.

It turns out that the two settings are closely related: any private PAC learner can be transformed into a private agnostic learner via the realizable-to-agnostic transformation (Beimel et al., 2015; Alon et al., 2020). More formally, given a private PAC learner for concept class $\mathcal{C}$, the transformation produces a private agnostic learner for $\mathcal{C}$ with an increase of $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha^2)$ in the sample complexity, where α is the error parameter. Such an increase is optimal since it matches the lower bound on (non-private) agnostic learning (Simon, 1996). However, this transformation results in an algorithm with a constant privacy parameter $\varepsilon = \Theta(1)$. If one pursues an arbitrary privacy level, the privacy-amplification-by-subsampling technique (Kasiviswanathan et al., 2011) has to be applied. This raises the extra sample complexity to $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha^2\varepsilon)$, incorporating an undesirable $1/\varepsilon$ factor. It is natural to ask if we can remove the $1/\varepsilon$ factor and still achieve an optimal extra sample complexity when ε is extremely small.

While private learning mandates that the entire output hypothesis must preserve privacy, this requirement might be excessively stringent. It has been shown that several concept classes, which can be easily learned in the non-private setting, pose significant challenges under differential privacy constraints (Beimel et al., 2010; Feldman and Xiao, 2014; Bun et al., 2015; Alon et al., 2019). To bypass these hardness results, Dwork and Feldman (2018) introduced the problem of *private prediction*. This framework is particularly relevant when the complete model, trained on sensitive data, cannot be fully released to users (who might be potential adversaries). Instead, users submit queries consisting of unlabeled data points, and the system provides predictions on these points while ensuring privacy. In the realizable setting, they showed that the sample complexity of answering a single query privately is $\widetilde{\Theta}(\mathtt{VC}(\mathcal{C})/\varepsilon\alpha)$, which is notably lower than that of private learning (Beimel et al., 2019; Alon et al., 2019; Bun et al., 2020; Ghazi et al., 2021). In the more challenging agnostic setting, the initial upper bound for sample complexity was $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha^3\varepsilon)$. This was subsequently refined by Dagan and Feldman (2020) to $\widetilde{O}(\min(\mathtt{VC}(\mathcal{C})/\alpha^2\varepsilon, \mathtt{VC}(\mathcal{C})^2/\alpha\varepsilon)+\mathtt{VC}(\mathcal{C})/\alpha^2)$. Despite these improvements, there remains a gap between this upper bound and the $\widetilde{\Omega}(\mathtt{VC}(\mathcal{C})/\alpha\varepsilon + \mathtt{VC}(\mathcal{C})/\alpha^2)$ lower bound.

1.1. Results

Our main contribution is a transformation that converts any realizable learning algorithm to an agnostic learning algorithm under (ε, δ)-differential privacy. Remarkably, this transformation only increases the sample complexity asymptotically by $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha^2)$ for any $\varepsilon \leq 1$. Such an extra sample complexity is near-optimal as it matches the $\widetilde{\Omega}(\mathtt{VC}(\mathcal{C})/\alpha^2)$ lower bound on agnostic learning even without privacy (Simon, 1996).

Theorem 1 (Informal Version of Theorem 18) *An (ε, δ)-differentially private realizable learner for $\mathcal{C}$ with error α and with sample complexity m can be transformed into an (ε, δ)-differentially private agnostic learner for $\mathcal{C}$ with excess error $O(\alpha)$ and with sample complexity $\widetilde{O}(m + \mathtt{VC}(\mathcal{C})/\alpha^2)$.*

Our methodology builds upon the foundational transformation proposed by Beimel et al. (2015), which was originally limited to only achieving a constant level of privacy. We observe that directly applying the privacy-amplification-by-subsampling method (Kasiviswanathan et al., 2011) suffers from a $1/\varepsilon$ blow-up because it runs the transformation only on a subsampled dataset whose size is approximately ε of the input size and naively discards unsampled data points. To effectively utilize all data points, we design a novel score function that estimates the generalization error using the entire dataset rather than only the subsampled dataset while still enjoying amplification of privacy, thus avoiding the privacy cost incurred by previous methods. Additionally, we adopt a technique due to Alon et al. (2020) to accommodate improper learners.

We also obtain improved results for the private prediction problem in the agnostic setting by applying our transformation to a private prediction algorithm in the realizable setting due to Dwork and Feldman (2018) with some mild modification. The result is demonstrated as follows.

Theorem 2 *There is an ε-differentially private prediction algorithm that (α, β)-agnostically learns $\mathcal{C}$ using $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha\varepsilon + \mathtt{VC}(\mathcal{C})/\alpha^2)$ samples.*

This upper bound is tight up to logarithmic factors. In fact, a matching lower bound can be derived by combining a lower bound of $\Omega(\mathtt{VC}(\mathcal{C})/\alpha\varepsilon)$ on the sample complexity of private prediction in the realizable setting established by Dwork and Feldman (2018) with an $\widetilde{\Omega}(\mathtt{VC}(\mathcal{C})/\alpha^2)$ lower bound for agnostic learning (Simon, 1996).

1.2. Related Work

Realizable-to-agnostic transformation: The general approach of transforming realizable learners to agnostic learners under differential privacy was first introduced by Beimel et al. (2015). Their method, however, is only applicable to proper learners. This limitation was later addressed by Alon et al. (2020), who extended the applicability to improper learners using the generalization property of DP. Despite being general, their methods only satisfy a constant level of privacy. The connection between realizable and agnostic learning under privacy was also investigated by Hopkins et al. (2022). However, their reduction only works for a relaxation of private learning called semi-private learning, where the algorithms have access to a public unlabeled dataset, and cannot be applied to private learning, which treats the entire input dataset as private and does not rely on any public data points. The optimal sample complexity of private agnostic learning was also considered by Li et al. (2024) under pure differential privacy. They provided an algorithm building upon the pure private realizable learner in (Beimel et al., 2019). Nevertheless, the algorithm is not a general transformation and does not work for approximate differential privacy.

Private prediction: The study of private prediction was pioneered by Dwork and Feldman (2018), who established a near-optimal sample complexity of $\widetilde{\Theta}(\mathtt{VC}(\mathcal{C})/\alpha\varepsilon)$ for the realizable setting. In the agnostic setting, they gave a suboptimal upper bound of $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha^3\varepsilon)$ for any general concept class $\mathcal{C}$. When $\mathcal{C}$ is the class of unions of intervals, they presented an algorithm with an expected excess error of α using $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha\varepsilon + \mathtt{VC}(\mathcal{C})/\alpha^2)$ samples, which is nearly optimal with a constant success probability. The upper bound for general concept classes was further improved to $\widetilde{O}(\min(\mathtt{VC}(\mathcal{C})/\alpha^2\varepsilon, \mathtt{VC}(\mathcal{C})^2/\alpha\varepsilon) + \mathtt{VC}(\mathcal{C})/\alpha^2)$ by Dagan and Feldman (2020). Similar to our method, their algorithm combines the realizable-to-agnostic transformation (Beimel et al., 2015) and the privacy-amplification-by-subsampling technique (Kasiviswanathan et al., 2011) with some modification. However, the use of a VC argument in their proof introduces an extra multiplicative factor of $\mathtt{VC}(\mathcal{C})$, rendering their bound suboptimal.

2. Preliminaries

We start with some notations. Let $\mathcal{X}$ be an arbitrary domain. A concept/hypothesis is a function that labels each $x \in \mathcal{X}$ by either 0 or 1. A concept/hypothesis class is a set of concepts/hypotheses. We use $\mathcal{D}$ to denote a distribution over $\mathcal{X} \times \{0,1\}$, and $\mathcal{D}_\mathcal{X}$ to denote its marginal distribution over $\mathcal{X}$. Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \{0,1\})^n$ be a dataset consisting of n data points. The corresponding unlabeled dataset is denoted as $S_\mathcal{X} = \{x_1, \ldots, x_n\}$. For two hypotheses h_1 and h_2, we define $h_1 \oplus h_2$ as a hypothesis such that $(h_1 \oplus h_2)(x) = \mathbb{I}[h_1(x) \neq h_2(x)]$.

Given a hypothesis h, the *generalization error* of h with respect to a distribution $\mathcal{D}$ is defined as $\mathrm{err}_\mathcal{D}(h) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$. The *empirical error* of h with respect to a dataset S is defined as $\mathrm{err}_S(h) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}[h(x_i) \neq y_i]$.

For two hypotheses h_1 and h_2, the *generalization disagreement* between h_1 and h_2 with respect to a distribution $\mathcal{D}_\mathcal{X}$ is defined as $\mathrm{dis}_{\mathcal{D}_\mathcal{X}}(h_1, h_2) = \Pr_{x\sim\mathcal{D}_\mathcal{X}}[h_1(x) \neq h_2(x)]$. The *empirical disagreement* between h_1 and h_2 with respect to an unlabeled dataset $S_\mathcal{X}$ is defined as $\mathrm{dis}_{S_\mathcal{X}}(h_1, h_2) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}[h_1(x_i) \neq h_2(x_i)]$.

In the PAC learning framework, the learning algorithm receives as input a dataset sampled according to some underlying distribution $\mathcal{D}$, where the data points are labeled by some concept $c \in \mathcal{C}$. The objective is to output a hypothesis h with low generalization error $\mathrm{err}_\mathcal{D}(h)$.

Definition 3 (PAC Learning (Valiant, 1984)) *We say a learning algorithm $\mathcal{A}$ is an (α, β)-PAC learner for concept class $\mathcal{C}$ with sample complexity m if for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that $\Pr_{(x,y)\sim\mathcal{D}}[c(x) = y] = 1$ for some $c \in \mathcal{C}$, it takes a dataset $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ as input, where each (x_i, y_i) is drawn i.i.d. from $\mathcal{D}$, and outputs a hypothesis h satisfying*

$$\Pr[\mathrm{err}_\mathcal{D}(h) \leq \alpha] \geq 1 - \beta,$$

where the probability is taken over the random generation of S and the random coins of $\mathcal{A}$.

PAC learning focuses on the realizable case, which assumes that the underlying distribution $\mathcal{D}$ is labeled by some concept $c \in \mathcal{C}$. In contrast, agnostic learning (Haussler, 1992; Kearns et al., 1994) does not impose any assumptions on the distribution $\mathcal{D}$. Instead, the goal is to identify a hypothesis whose generalization error is close to that of the best possible concept in $\mathcal{C}$.

Definition 4 (Agnostic Learning) *We say a learning algorithm $\mathcal{A}$ is an (α, β)-agnostic learner for concept class $\mathcal{C}$ with sample complexity m if for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, it takes as input a dataset $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, where each (x_i, y_i) is drawn i.i.d. from $\mathcal{D}$, and outputs a hypothesis h satisfying*

$$\Pr[\mathrm{err}_\mathcal{D}(h) \leq \inf_{c\in\mathcal{C}} \mathrm{err}_\mathcal{D}(c) + \alpha] \geq 1 - \beta,$$

where the probability is taken over the random generation of S and the random coins of $\mathcal{A}$.

In PAC and agnostic learning, if the learner $\mathcal{A}$ always produces a hypothesis that is a concept in $\mathcal{C}$, then we say $\mathcal{A}$ is a *proper* learner. Otherwise, we say $\mathcal{A}$ is an *improper* learner.

We next introduce some useful tools and results from learning theory.

Definition 5 (The Growth Function) *Let $S_\mathcal{X} = (x_1, \ldots, x_n)$ be an unlabeled dataset of size n. The set of all dichotomies on $S_\mathcal{X}$ realized by $\mathcal{C}$ is denoted by*

$$\Pi_\mathcal{C}(S_\mathcal{X}) = \{\{(x_1, c(x_1)), \ldots, (x_n, c(x_n))\} \mid c \in \mathcal{C}\}.$$

The growth function of $\mathcal{C}$ is defined as $\Pi_{\mathcal{C}}(n) = \max_{S_{\mathcal{X}} \in \mathcal{X}^n} |\Pi_{\mathcal{C}}(S_{\mathcal{X}})|$.

The Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971) of a concept class $\mathcal{C}$ is defined as the largest number d such that $\Pi_{\mathcal{C}}(d) = 2^d$ (or infinity, if no such maximum exists), denoted by $\mathtt{VC}(\mathcal{C})$. Sauer's lemma (Sauer, 1972) states that the number of dichotomies is polynomially bounded if $\mathcal{C}$ has a finite VC dimension.

Lemma 6 (Sauer's Lemma) *For any $n \geq \mathtt{VC}(\mathcal{C})$, we have $\Pi_{\mathcal{C}}(n) \leq \left(\frac{en}{\mathtt{VC}(\mathcal{C})}\right)^{\mathtt{VC}(\mathcal{C})}$.*

The following realizable generalization bound (Vapnik and Chervonenkis, 1971; Blumer et al., 1989) relates the generalization disagreement and the empirical disagreement simultaneously for every pair of concepts: if one is small, then the other will also be small.

Lemma 7 (Realizable Generalization Bound) *Let $\mathcal{C}$ be a concept class and $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$. Suppose $S_{\mathcal{X}} = \{x_1, \ldots, x_n\}$, where each x_i is drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$. Define the following two events:*

- *E_1: For any $h_1, h_2 \in \mathcal{C}$ such that $\mathrm{dis}_{S_{\mathcal{X}}}(h_1, h_2) \leq \alpha$, it holds that $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(h_1, h_2) \leq 2\alpha$.*
- *E_2: For any $h_1, h_2 \in \mathcal{C}$ such that $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(h_1, h_2) \leq \alpha$, it holds that $\mathrm{dis}_{S_{\mathcal{X}}}(h_1, h_2) \leq 2\alpha$.*

Then we have

$$\Pr[E_1 \cap E_2] \geq 1 - \beta$$

given that

$$n \geq C \cdot \frac{\mathtt{VC}(\mathcal{C}) \ln(1/\alpha) + \ln(1/\beta)}{\alpha}$$

for some universal constant C.

We also have the following agnostic generalization bound (Talagrand, 1994), which ensures that for every concept $c \in \mathcal{C}$, its generalization error and empirical error are close. Unlike the realizable generalization bound, the agnostic generalization bound does not require the error to be small. However, this relaxation increases the sample complexity by roughly a factor of $1/\alpha$.

Lemma 8 (Agnostic Generalization Bound) *Let $\mathcal{C}$ be a concept class and $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0,1\}$. Suppose $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where each (x_i, y_i) is drawn i.i.d. from $\mathcal{D}$. Then we have*

$$\Pr[\forall c \in \mathcal{C}, |\mathrm{err}_S(c) - \mathrm{err}_{\mathcal{D}}(c)| \leq \alpha] \geq 1 - \beta$$

given that

$$n \geq C \cdot \frac{\mathtt{VC}(\mathcal{C}) + \ln(1/\beta)}{\alpha^2}$$

for some universal constant C.

In this work, we consider learning algorithms that preserve differential privacy. We say that two datasets S_1 and S_2 are neighboring if they differ by a single entry. A private algorithm is required to produce similar outputs for every pair of neighboring datasets. The similarity between the output distributions is quantified by two parameters ε and δ. We refer to the case when $\delta = 0$ as *pure* differential privacy, and when $\delta > 0$ we term it *approximate* differential privacy.

Definition 9 (Differential Privacy (Dwork et al., 2006b,a)) *A randomized algorithm $\mathcal{A}$ is said to be (ε, δ)-differentially private if for any two neighboring datasets S_1 and S_2 and any set O of outputs, we have*

$$\Pr[\mathcal{A}(S_1) \in O] \le e^{\varepsilon} \Pr[\mathcal{A}(S_2) \in O] + \delta.$$

When $\delta = 0$, we may omit the parameter δ and simply say that $\mathcal{A}$ is ε-differentially private.

In private learning, the learning algorithm produces a hypothesis that contains the prediction result for every $x \in \mathcal{X}$. However, in the private prediction problem introduced by Dwork and Feldman (2018), the algorithm $\mathcal{A}$ receives a dataset S along with a query x and outputs only the prediction on x. For privacy, we require it to satisfy differential privacy with respect to the dataset S. For utility, we treat $\mathcal{A}(S, \cdot)$ as a (randomized) classifier and require it to exhibit a low error in the PAC/agnostic setting. The formal definition is provided below.

Definition 10 (Private Prediction Dwork and Feldman (2018)) *Let $\mathcal{A}$ be an algorithm that takes a labeled dataset S and an unlabeled data point x as input and produces a prediction value in $\{0, 1\}$. We say $\mathcal{A}$ is an (ε, δ)-differentially private prediction algorithm if for any $x \in \mathcal{X}$, the output $A(S, x)$ is (ε, δ)-differentially private with respect to S.*

Define $\mathrm{err}_{\mathcal{D}}(\mathcal{A}(S, \cdot)) = \Pr_{(x,y)\sim\mathcal{D},\mathcal{A}}[\mathcal{A}(S, x) \neq y]$. We say $\mathcal{A}$ (α, β)-PAC learns $\mathcal{C}$ if for any distribution $\mathcal{D}$ such that $\Pr_{(x,y)\sim\mathcal{D}}[c(x) = y] = 1$ for some $c \in \mathcal{C}$, we have

$$\Pr_{S\sim\mathcal{D}^n}[\mathrm{err}_{\mathcal{D}}(\mathcal{A}(S, \cdot)) \le \alpha] \ge 1 - \beta.$$

Similarly, we say $\mathcal{A}$ (α, β)-agnostically learns $\mathcal{C}$ if for any distribution $\mathcal{D}$, we have

$$\Pr_{S\sim\mathcal{D}^n}[\mathrm{err}_{\mathcal{D}}(\mathcal{A}(S, \cdot)) \le \inf_{c\in\mathcal{C}} \mathrm{err}_{\mathcal{D}}(c) + \alpha] \ge 1 - \beta.$$

We next describe the exponential mechanism (McSherry and Talwar, 2007) and its properties. Let H be a finite set and $q : (\mathcal{X} \times \{0, 1\})^n \times H \to \mathbb{R}$ be a score function. We say q has sensitivity Δ if $\max_{h\in H}|q(S_1, h) - q(S_2, h)| \le \Delta$ for any neighboring datasets S_1 and S_2 of size n. The exponential mechanism outputs an element $h \in H$ with probability

$$\frac{\exp(-\varepsilon \cdot q(S, h)/2\Delta)}{\sum_{f\in H} \exp(-\varepsilon \cdot q(S, f)/2\Delta)}.$$

Lemma 11 (Properties of the Exponential Mechanism (McSherry and Talwar, 2007)) *The exponential mechanism is ε-differentially private. Moreover, with probability $1 - \beta$, it outputs an h such that*

$$q(S, h) \le \min_{f\in H} q(S, f) + \frac{2\Delta}{\varepsilon} \ln(|H|/\beta).$$

An important property of differential privacy is that it implies generalization (Dwork et al., 2015a,b; Bassily et al., 2016; Rogers et al., 2016; Feldman and Steinke, 2017; Jung et al., 2020): a differentially private learning algorithm with a low empirical error also exhibits a low generalization error. The following version of the generalization property was used by Alon et al. (2020) in their transformation to handle improper learners. Note that this bound holds even for $\varepsilon > 1$.

Lemma 12 (DP Generalization) *Let $\mathcal{A}$ be an (ε, δ)-differentially private learning algorithm that takes $S_{\mathcal{X}} \in \mathcal{X}^n$ as input and outputs a predicate $h : \mathcal{X} \to \{0, 1\}$. Suppose each element in $S_{\mathcal{X}}$ is drawn i.i.d. from some distribution $\mathcal{D}_{\mathcal{X}}$, then we have*

$$\Pr\left[\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[h(x)\right] > e^{2\varepsilon}\left(\frac{\sum_{x\in S_{\mathcal{X}}} h(x)}{n} + \frac{10}{\varepsilon n}\log\left(\frac{1}{\varepsilon\delta n}\right)\right)\right] < O\left(\frac{\varepsilon\delta n}{\log\left(\frac{1}{\varepsilon\delta n}\right)}\right),$$

where the probability is taken over the random generation of $S_{\mathcal{X}}$ and the random coins of $\mathcal{A}$.

A learner that outputs a hypothesis with low empirical error is called an empirical learner (Bun et al., 2015), which can be constructed from a PAC learner while preserving privacy.

Definition 13 (PAC Empirical Learner) *An algorithm $\mathcal{A}$ is said to be an (α, β)-PAC empirical learner for concept class $\mathcal{C}$ with sample complexity m if for any $c \in \mathcal{C}$ and any dataset $S = \{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\}$, it takes S as input and outputs a hypothesis h such that*

$$\Pr[\mathrm{err}_S(h) \leq \alpha] \geq 1 - \beta.$$

Lemma 14 (Private Empirical Learner) *Let $\varepsilon \leq 1$. Suppose $\mathcal{A}$ be an (ε, δ)-differentially private (α, β)-PAC learner for $\mathcal{C}$ with sample complexity m. Then there exists an $(1, O(\delta/\varepsilon))$-differentially private (α, β)-PAC empirical learner $\mathcal{A}'$ for $\mathcal{C}$ with sample complexity $O(\varepsilon m)$. Moreover, if $\mathcal{A}$ is proper, then $\mathcal{A}'$ is also proper.*

3. The Transformation

In this section, we present our realizable-to-agnostic transformation. We will start by describing a relabeling procedure proposed by Beimel et al. (2015), which serves as the key component of our transformation.

Let $\mathcal{C}$ be a concept class and $S \in (\mathcal{X} \times \{0, 1\})^n$ be a dataset. In the agnostic setting, S may not be consistent with any $c \in \mathcal{C}$. The idea of Beimel et al. (2015) is to first relabel S by some concept $h \in \mathcal{C}$. After that, realizable learning algorithms can be applied.

Their method first constructs a candidate set H such that for every labeling of S there is one concept in H consistent with that labeling. Then it initiates the exponential mechanism with score function $q(S, h) = \mathrm{err}_S(h)$ to select a concept h for relabeling. Though the selection of h is not private since the construction of H depends on S, they proved that running a private algorithm on the relabeled dataset S^h still preserves differential privacy. Moreover, the agnostic generalization bound and the property of the exponential mechanism ensure that $\mathrm{err}_{\mathcal{D}}(h)$ is close to the optimal error achieved by concepts in $\mathcal{C}$. Thus, if we can find some hypothesis g such that $\mathrm{err}_{\mathcal{D}}(g) \approx \mathrm{err}_{\mathcal{D}}(h)$ by running a realizable learner on S^h, the resulting algorithm is an agnostic learner as desired.

However, the data points in S^h are no longer i.i.d. from some distribution because the selection of h depends on S. Therefore, directly running a private PAC learner on the relabeled dataset S^h might not produce a good hypothesis. One should instead convert it to an empirical learner and apply the empirical learner to obtain some hypothesis g whose empirical error is small on S^h. When the learner is proper (i.e., $g \in \mathcal{C}$), the realizable generalization bound implies that $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(g, h)$ is small, which indicates $\mathrm{err}_{\mathcal{D}}(g) \approx \mathrm{err}_{\mathcal{D}}(h)$. However, when the given private PAC learner is improper (and so is the resulting empirical learner), the realizable generalization bound cannot provide any guarantee on $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(g, h)$.

To deal with improper learners, we next discuss a technique due to Alon et al. (2020). In their work, they split the input dataset S into two parts $S = U \circ V$ and showed that a simple variant of Beimel et al. (2015)'s algorithm, which outputs V^h as well, still preserves privacy with respect to U (i.e., the V portion is regarded as public). They then considered an auxiliary algorithm that outputs $g \oplus \bar{h}$ for some $\bar{h} \in \mathcal{C}$ consistent with V^h and used the generalization property of DP to derive an upper bound on $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(g, \bar{h})$. Since $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(h, \bar{h})$ can be controlled by applying the realizable generalization bound to $V_{\mathcal{X}}$, the triangle inequality yields a bound on $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(g, h)$. Therefore, the generalization error of g can be successfully bounded.

The above transformation incurs an extra sample complexity of $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha^2)$ when converting a private PAC learner to a private agnostic learner (Beimel et al., 2015; Alon et al., 2020). However, it only provides a constant level of privacy (i.e., $\varepsilon = \Theta(1)$) even if the PAC learner $\mathcal{A}$ is (ε, δ)-differentially private for some $\varepsilon \ll 1$. To achieve an arbitrary privacy level ε, one has to apply the privacy-amplification-by-subsampling technique (Kasiviswanathan et al., 2011): first subsample a dataset T from S with size $|T| \approx \varepsilon |S| = \varepsilon n$, then perform the transformation on T only. Since the size of T should be at least $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha^2)$ to ensure the agnostic generalization of every $h \in H \subseteq \mathcal{C}$, the overall transformation results in an extra sample complexity of $\widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha^2 \varepsilon)$.

We now illustrate how to eliminate the $1/\varepsilon$ factor. Let W denote the dataset containing the data points that are not in T, i.e., $S = T \circ W$. In the above process, we discard W after sampling and do not exploit any information contained in W, which seems too wasteful. Our idea is to utilize W so that we can apply the agnostic generalization bound to the entire dataset rather than T only. To be specific, we still construct H from the subsampled dataset T, but evaluate the score function of every $h \in H$ over S. Thus, we only require $|S| \geq \widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha^2)$ to ensure agnostic generalization.

The primary obstacle here is how to ensure that such a modification still preserves privacy. Let $S_1 = T_1 \circ W_1$ and $S_2 = T_2 \circ W_2$ be two neighboring datasets. There are two cases: $T_1 = T_2$ and $T_1 \neq T_2$. In the case where $T_1 = T_2$, we will construct the same candidate set H from them. Therefore, it is easy to achieve privacy for the selection of h (the hypothesis for relabeling the dataset) by launching an ε-differentially private exponential mechanism since W_1 and W_2 are neighboring datasets. According to the post-processing property of DP, running any algorithm on the relabeled dataset is also ε-differentially private.

However, it is not as simple in the case where $T_1 \neq T_2$, as the candidate sets constructed from T_1 and T_2 are different. To see why, let $\hat{T}$ be the overlapping portion of T_1 and T_2, which has size $|T| - 1$. The privacy analysis proposed by Beimel et al. (2015) requires $|q(S_1, h_1) - q(S_2, h_2)|$ to be small for any h_1 and h_2 that agree on $\hat{T}$. This naturally holds in the original transformation, where the score function is $q(S, h) = \mathrm{err}_T(h)$ (recall that we discard all data points in W) and the difference $|q(S_1, h_1) - q(S_2, h_2)|$ is only $1/|T|$ since h_1 and h_2 agree on $\hat{T}$. However, if we try to incorporate the W portion and set the score function to be $q(S, h) = \mathrm{err}_S(h)$, the difference can be close to 1 as h_1 and h_2 may totally disagree on $W_1 = W_2$, failing to provide a satisfactory privacy guarantee.

We overcome this issue by devising a score function that estimates the generalization error of h using the entire dataset while having a small "sensitivity". In particular, we run the exponential mechanism with the following score function:

$$q(T \circ W, h) = \min_{f \in \mathcal{C}} \{ \mathrm{dis}_{T_{\mathcal{X}}}(h, f) + \mathrm{err}_W(f) \}.$$

Algorithm 1: $\mathcal{A}_{\text{Relabel}}$

```
Input: Parameter $\varepsilon$, Datasets $T, W$
1. Initialize $H = \emptyset$.
2. For every possible labeling in $\Pi_{\mathcal{C}}(T_{\mathcal{X}})$, add to $H$ an arbitrary concept $h \in \mathcal{C}$ that is consistent
   with the labeling.
3. Define the following score function $q$:
       $q(T \circ W, h) = \min_{f \in \mathcal{C}} \{\mathrm{dis}_{T_{\mathcal{X}}}(h, f) + \mathrm{err}_W(f)\}.$
4. Choose $h \in H$ using the exponential mechanism with privacy parameter $\varepsilon$, score function $q$,
   and sensitivity parameter $\Delta = \frac{1}{|W|}$.
5. Relabel $T$ using $h$ and output the relabeled dataset $T^h$.
```

Algorithm 2: $\mathcal{A}_{\text{Agnostic}}$

```
Input: Parameter $\varepsilon$, Dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, Private Algorithm $\mathcal{A}$
1. Sample a subset $I \subseteq [n]$ of size $|I| = \lceil \varepsilon n \rceil$ uniformly at random.
2. Let $T = \{(x_i, y_i) \mid i \in I\}$ and $W = \{(x_i, y_i) \mid i \in [n] \setminus I\}$.
3. Execute $\mathcal{A}_{\text{Relabel}}$ (Algorithm 1) with parameter $\varepsilon$ and input datasets $T, W$ to obtain relabeled
   dataset $T^h$.
4. Output $\mathcal{A}(T^h)$.
```

The above score function can be interpreted as searching for a concept f that is close to h over $T_{\mathcal{X}}$ and also has a low empirical error on W. We describe the relabeling procedure in Algorithm 1, where we set the privacy parameter as ε and sensitivity parameter as $1/|W|$ to ensure that it preserves ε-differential privacy when $T_1 = T_2$. In the case that $T_1 \neq T_2$, one can verify that $|q(S_1, h_1) - q(S_2, h_2)| \leq 1/|T| = \Theta(1/\varepsilon n)$. Because we have set the privacy parameter as ε and sensitivity parameter $\Delta = \Theta(1/n)$, we can apply the analysis of Beimel et al. (2015) to show that running a private algorithm on the relabeled dataset is still private with a constant privacy parameter. Note that this case only happens with probability ε. We can apply the privacy-amplification-by-subsampling argument to deduce (actually, the formal proof requires a more delicate privacy analysis for this case) that the overall algorithm is private with privacy parameter ε. We formally describe the details of the entire agnostic learning algorithm in Algorithm 2 and state its privacy guarantee in the following lemma.

Lemma 15 (Privacy of $\mathcal{A}_{\text{Agnostic}}$) *Suppose $\mathcal{A}$ is $(1, \delta)$-differentially private. Then $\mathcal{A}_{\text{Agnostic}}$ (Algorithm 2) is $(O(\varepsilon), O(\varepsilon\delta))$-differentially private.*

```
Algorithm 3: $\mathcal{A}_{\text{Auxiliary}}$
Input: Parameter $\varepsilon$, Datasets $U, V, W$, Private Algorithm $\mathcal{A}$
1. Execute $\mathcal{A}_{\text{Relabel}}$ (Algorithm 1) with parameter $\varepsilon$ and input datasets $T = U \circ V, W$ to obtain
   relabeled dataset $T^h = U^h \circ V^h$.
2. Select an arbitrary $\bar{h} \in \mathcal{C}$ that is consistent with $V^h$.
3. Output $\mathcal{A}(T^h) \oplus \bar{h}$.
```

We then prove the utility guarantee of $\mathcal{A}_{\text{Agnostic}}$. The first step is to show that $\mathcal{A}_{\text{Relabel}}$ will relabel the dataset using some h whose generalization error is close to the optimal. We show that it suffices to set $|T| = \widetilde{O}(\mathtt{VC}(\mathcal{C})/\alpha)$ and $|W| = \widetilde{O}(\mathtt{VC}(\mathcal{C}) \cdot \max(1/\alpha^2, 1/\alpha\varepsilon))$.

Claim 16 *Let T and W be two datasets with every data point sampled i.i.d. from $\mathcal{D}$. Suppose*

$$|T| \geq C_1 \cdot \frac{\mathtt{VC}(\mathcal{C}) \ln(1/\alpha) + \ln(1/\beta)}{\alpha}$$

and

$$|W| \geq \max\left(C_2 \cdot \frac{\mathtt{VC}(\mathcal{C}) + \ln(1/\beta)}{\alpha^2}, \frac{|T|}{6\varepsilon}\right),$$

where C_1 and C_2 are universal constants. Then with probability $1-\beta$, $\mathcal{A}_{\text{Relabel}}$ (Algorithm 1) will relabel T using some $h \in \mathcal{C}$ such that

$$\operatorname{err}_{\mathcal{D}}(h) \leq \inf_{c \in \mathcal{C}} \operatorname{err}_{\mathcal{D}}(c) + \alpha.$$

Now it remains to show that the output hypothesis $\mathcal{A}(T^h)$ is close to h on the underlying distribution $\mathcal{D}$. When $\mathcal{A}$ is a proper learner (i.e., $\mathcal{A}(T^h) \in \mathcal{C}$), this is directly implied by the realizable generalization.

To handle the case that $\mathcal{A}$ may be improper, we adopt the proof strategy of Alon et al. (2020). The key idea is to construct an auxiliary algorithm $\mathcal{A}_{\text{Auxiliary}}$ (Algorithm 3). It splits T into $T = U \circ V$ (we set $|U| = |V| = |T|/2$) and outputs $\mathcal{A}(T^h) \oplus \bar{h}$ for some $\bar{h} \in \mathcal{C}$ that is consistent with V^h. Given that $|V|$ is sufficiently large, the realizable generalization bound implies that $\bar{h}$ is close to h on the underlying distribution $\mathcal{D}$, hence on $U_{\mathcal{X}}$ as well. Since the output hypothesis $\mathcal{A}(T^h)$ is also close to h over $U_{\mathcal{X}}$, $\mathcal{A}(T^h)$ should be close to $\bar{h}$ on $U_{\mathcal{X}}$.

We prove that $\mathcal{A}_{\text{Auxiliary}}$ satisfies a constant level of differential privacy using an argument similar to part of the proof of Lemma 15. This allows us to bound the generalization disagreement between $\mathcal{A}(T^h)$ and $\bar{h}$ by the generalization property of DP. Therefore, we have $\operatorname{err}_{\mathcal{D}}(\mathcal{A}(T^h)) \approx \operatorname{err}_{\mathcal{D}}(\bar{h}) \approx \operatorname{err}_{\mathcal{D}}(h)$, which proves the utility guarantee of $\mathcal{A}_{\text{Agnostic}}$.

Lemma 17 (Utility of $\mathcal{A}_{\text{Agnostic}}$) *Suppose $\mathcal{A}$ is a $(1, \delta)$-differentially private (α, β)-PAC empirical learner with sample complexity m. Then $\mathcal{A}_{\text{Agnostic}}$ (Algorithm 2) is an $(O(\alpha), O(\beta + \varepsilon n \delta))$-agnostic learner with sample complexity*

$$n = O\left(\frac{m}{\varepsilon} + \frac{\mathtt{VC}(\mathcal{C}) \log(1/\alpha) + \log(1/\beta)}{\alpha\varepsilon} + \frac{\mathtt{VC}(\mathcal{C}) + \log(1/\beta)}{\alpha^2}\right).$$

Now we are able to prove our main theorem. In our transformation, we first invoke Lemma 14 to create a $(1, \delta' = O(\delta/\varepsilon))$-differentially private empirical learner from the given (ε, δ)-differentially private PAC learner. We then use this empirical learner as the private algorithm $\mathcal{A}$ in $\mathcal{A}_{\text{Agnostic}}$. Lemma 15 ensures that $\mathcal{A}_{\text{Agnostic}}$ is $(O(\varepsilon), O(\varepsilon\delta') = O(\delta))$-differentially private. The final sample complexity follows from Lemma 17.

Theorem 18 *Let $\varepsilon \leq O(1)$. Suppose there is an (ε, δ)-differentially private (α, β)-PAC learner for $\mathcal{C}$ with sample complexity m. Then there exists an (ε, δ)-differentially private $(O(\alpha), O(\beta + \delta n))$-agnostic learner for $\mathcal{C}$ with sample complexity*

$$n = O\left(m + \frac{\texttt{VC}(\mathcal{C})\log(1/\alpha) + \log(1/\beta)}{\alpha\varepsilon} + \frac{\texttt{VC}(\mathcal{C}) + \log(1/\beta)}{\alpha^2}\right).$$

Moreover, if the original learner is proper, then the resulting learner is also proper.

Proof By Lemma 14, there exists a $(1, \delta')$-differentially private (α, β)-PAC empirical learner $\mathcal{A}$ for $\mathcal{C}$ with sample complexity $O(\varepsilon m)$, where $\delta' = O(\delta/\varepsilon)$. Then by Lemma 15, we have that $\mathcal{A}_{\text{Agnostic}}$ is $(O(\varepsilon), O(\varepsilon\delta') = O(\delta))$-differentially private. Moreover, by Lemma 17, $\mathcal{A}_{\text{Agnostic}}$ is an $(O(\alpha), O(\beta + \varepsilon\delta' n) = O(\beta + \delta n))$-agnostic learner with sample complexity

$$n = O\left(m + \frac{\texttt{VC}(\mathcal{C})\log(1/\alpha) + \log(1/\beta)}{\alpha\varepsilon} + \frac{\texttt{VC}(\mathcal{C}) + \log(1/\beta)}{\alpha^2}\right).$$

The constant factors on the privacy parameters can be removed by the privacy-amplification-by-subsampling technique. Furthermore, if the original learner is proper, then Lemma 14 suggests that $\mathcal{A}$ is also proper. Therefore, $\mathcal{A}_{\text{Agnostic}}$ is proper since it outputs $\mathcal{A}(T^h)$. ■

Ignoring logarithmic factors, the second term in the resulting sample complexity is dominated by m (Bun et al., 2015). Hence, the extra sample complexity introduced by our transformation is $\widetilde{O}(\texttt{VC}(\mathcal{C})/\alpha^2)$, which is near-optimal.

4. Private Prediction

In this section, we provide an algorithm for private prediction in the agnostic setting with a nearly optimal sample complexity of $\widetilde{O}(\texttt{VC}(\mathcal{C})/\alpha\varepsilon + \texttt{VC}(\mathcal{C})/\alpha^2)$. We need the following algorithm for private prediction in the realizable setting (Dwork and Feldman, 2018).

Lemma 19 *Let $r = \lceil 6\ln(4/\alpha)/\varepsilon \rceil$ and $\mathcal{A}'$ be a PAC learning algorithm. Suppose hypotheses $g_1, \ldots, g_r$ are obtained by running r instances of $\mathcal{A}'$ on disjoint portions of the input dataset S. Then there exists an ε-differentially private prediction algorithm $\mathcal{A}$ that answers a prediction query x based on some aggregation mechanism over $\{g_1(x), \ldots, g_r(x)\}$ such that*

$$\forall i \in [r], \mathrm{err}_{\mathcal{D}}(g_i) \leq \alpha/4 \Rightarrow \mathrm{err}_{\mathcal{D}}(\mathcal{A}(S, \cdot)) \leq \alpha.$$

Like private agnostic learning, we run $\mathcal{A}_{\text{Agnostic}}$ with the private prediction algorithm $\mathcal{A}$ in the above lemma. The privacy analysis remains unchanged. However, the utility guarantee no longer holds because $\mathcal{A}$ only preserves privacy for a single prediction rather than the entire hypothesis, prohibiting us from applying the generalization property of DP.

To circumvent this issue, we choose the base learner $\mathcal{A}'$ to be an algorithm that outputs some concept in $\mathcal{C}$ that is consistent with the (relabeled) dataset. This allows us to leverage the realizable generalization bound to argue that with high probability, the generalization error (with respect to the relabeled distribution) of every instance of $\mathcal{A}'$ is small no matter which concept $h \in \mathcal{C}$ is selected to relabel the dataset, indicating that the generalization error of $\mathcal{A}$ is also small. That is to say, the generalization error of $\mathcal{A}$ is close to that of h with respect to the original distribution.

Theorem 20 (Theorem 2 Restated) *Let $\varepsilon \leq O(1)$. There exists an ε-differentially private prediction algorithm that (α, β)-agnostically learns $\mathcal{C}$ with sample complexity*

$$O\left(\frac{\mathtt{VC}(\mathcal{C})\log^2(1/\alpha) + \log(1/\alpha)\log(\log(1/\alpha)/\beta)}{\alpha\varepsilon} + \frac{\mathtt{VC}(\mathcal{C}) + \log(1/\beta)}{\alpha^2}\right).$$

Proof Consider an algorithm $\mathcal{A}'$ that arbitrarily selects a concept from $\mathcal{C}$ that is consistent with the input dataset. Let $\mathcal{A}$ be a 1-differentially private mechanism constructed using Lemma 19 with the base learner $\mathcal{A}'$. Then by Lemma 15, the overall algorithm $\mathcal{A}_{\text{Agnostic}}$ is $O(\varepsilon)$-differentially private.

We now prove the accuracy of $\mathcal{A}_{\text{Agnostic}}$. Let $\mathcal{D}$ be the underlying distribution and $\mathcal{D}_{\mathcal{X}}$ be its marginal distribution on $\mathcal{X}$. Set $n' \geq C \cdot \frac{\mathtt{VC}(\mathcal{C})\ln(1/\alpha)+\ln(r/\beta)}{\alpha}$ for some constant C and $r = \lceil 6\ln(4/\alpha)\rceil$. Let $T'_{\mathcal{X}} = \{x_1, \ldots, x_{n'}\}$ be some unlabeled dataset, where each x_i is drawn i.i.d. from $D_{\mathcal{X}}$. When C is sufficiently large, the realizable generalization bound (Lemma 7) suggests that with probability $1 - \beta/r$, it holds that $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(h_1, h_2) \leq \alpha/4$ for any $h_1, h_2 \in \mathcal{C}$ that are consistent on $T'_{\mathcal{X}}$. Let $S = T \circ W$ be the entire input dataset such that

$$|T| \geq rn' = O\left(\frac{\mathtt{VC}(\mathcal{C})\log^2(1/\alpha) + \log(1/\alpha)\log(\log(1/\alpha)/\beta)}{\alpha}\right)$$

and h be the concept chosen by $\mathcal{A}_{\text{Relabel}}$ for relabeling. Then by Lemma 19 and the union bound, it holds with probability $1 - \beta$ that $\mathrm{err}_{\mathcal{D}^h}(\mathcal{A}(T^h, \cdot)) \leq \alpha$, where $\mathcal{D}^h$ is the distribution obtained by labeling $\mathcal{D}_{\mathcal{X}}$ with h.

By Claim 16, we have $\mathrm{err}_{\mathcal{D}}(h) \leq \inf_{c \in \mathcal{C}} \mathrm{err}_{\mathcal{D}}(c) + \alpha$ with probability $1 - \beta$ given that $|T| \geq C_1 \cdot \frac{\mathtt{VC}(\mathcal{C})\ln(1/\alpha)+\ln(1/\beta)}{\alpha}$ and $|W| \geq C_2 \cdot \frac{\mathtt{VC}(\mathcal{C})+\ln(1/\beta)}{\alpha^2} + \frac{|T|}{6\varepsilon}$ for constants C_1, C_2. Therefore, the triangle inequality and the union bound imply that with probability $1 - 2\beta$, we have

$$\mathrm{err}_{\mathcal{D}}(\mathcal{A}_{\text{Agnostic}}(S, \cdot)) = \mathrm{err}_{\mathcal{D}}(\mathcal{A}(T^h, \cdot)) \leq \mathrm{err}_{\mathcal{D}^h}(\mathcal{A}(T^h, \cdot)) + \mathrm{err}_{\mathcal{D}}(h) \leq \inf_{c \in \mathcal{C}} \mathrm{err}_{\mathcal{D}}(c) + 2\alpha.$$

Adjusting $\varepsilon, \alpha, \beta$ by constant factors yields the desired result. ■

Acknowledgments

We thank anonymous COLT reviewers for helpful comments. The research was supported in part by a RGC RIF grant under contract R6021-20, RGC TRS grant under contract T43-513/23N-2, RGC CRF grants under contracts C7004-22G, C1029-22G and C6015-23G, NSFC project under contract 62432008, and RGC GRF grants under contracts 16207922, 16207423 and 16211123.

References

John M Abowd. The U.S. Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2867, 2018.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In *Proceedings of the 51st Annual ACM Symposium on Theory of Computing*, pages 852–860, 2019.

Noga Alon, Amos Beimel, Shay Moran, and Uri Stemmer. Closure properties for private classification and online prediction. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125, pages 119–152, 2020.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Apple Differential Privacy Team. Learning with privacy at scale. *Apple Machine Learning Journal*, 2017.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 1046–1059, 2016.

Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. In *Proceedings of the 7th International Conference on Theory of Cryptography*, volume 5978, pages 437–454, 2010.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Learning privately with labeled and unlabeled examples. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 461–477, 2015.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of pure private learners. *Journal of Machine Learning Research*, 20(146):1–33, 2019.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, pages 634–649, 2015.

Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science*, pages 389–402, 2020.

Yuval Dagan and Vitaly Feldman. PAC learning with stable and private predictions. In *Proceedings of the 33rd Annual Conference on Learning Theory*, volume 125, pages 1389–1410. PMLR, 2020.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pages 1693–1702, 2018.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, pages 265–284, 2006b.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the 47th Annual ACM symposium on Theory of Computing*, pages 117–126, 2015b.

Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In *Proceedings of the 27th Conference on Learning Theory*, pages 728–757, 2017.

Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pages 1000–1019, 2014.

Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Sample-efficient proper PAC learning with approximate differential privacy. In *Proceedings of the 53rd Annual ACM Symposium on Theory of Computing*, pages 183–196, 2021.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

Max Hopkins, Daniel M Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. In *Proceedings of the 35th Annual Conference on Learning Theory*, pages 3015–3069. PMLR, 2022.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *11th Innovations in Theoretical Computer Science Conference*, volume 151, page 31, 2020.

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

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

Bo Li, Wei Wang, and Peng Ye. Improved bounds for pure private agnostic learning: item-level and user-level privacy. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28870–28889, 2024.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94–103, 2007.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, pages 487–494. IEEE, 2016.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13 (1):145–147, 1972.

Hans Ulrich Simon. General bounds on the number of examples needed for learning probabilistic concepts. *Journal of Computer and System Sciences*, 52(2):239–254, 1996.

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22:28–76, 1994.

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

VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

Appendix A. Proof of Lemma 14

The following lemma is due to Bun et al. (2015). We remark that although their original statement requires $n \geq 2m$, their proof actually works under the stronger conditions we present below.

Lemma 21 *Let $\varepsilon \leq 1$. Suppose $\mathcal{A}$ is an (ε, δ)-differentially private algorithm that takes a dataset of size m as input. For any n such that $n \geq 2$ and $6\varepsilon m/n \leq 1$, consider an algorithm $\mathcal{A}'$ works as follows:*

1. *Takes as input a dataset S of size n.*
2. *Constructs a dataset T of size m, where each data point is sampled independently and uniformly from S with replacement.*
3. *Runs $\mathcal{A}$ on the dataset constructed in the previous step.*

Then $\mathcal{A}'$ is (ε', δ')-differentially private for $\varepsilon' = 6\varepsilon m/n$ and $\delta' = \exp(6\varepsilon m/n)\frac{4m}{n}\delta$.

Proof [Proof of Lemma 14] Let $\mathcal{A}$ be an (ε, δ)-differentially private (α, β)-PAC learner for $\mathcal{C}$ with sample complexity m. Construct an algorithm $\mathcal{A}'$ as in Lemma 21 with $n = \lceil 6\varepsilon m \rceil$. Then Lemma 21 directly implies that $\mathcal{A}'$ is $(1, O(\delta/\varepsilon))$-differentially private.

Let $\mathcal{D}$ be the empirical distribution over S. Since $\mathcal{A}$ is an (α, β)-PAC learner for $\mathcal{C}$, we have $\mathrm{err}_{\mathcal{D}}(\mathcal{A}(T)) \leq \alpha$ with probability $1 - \beta$ over the random generalization of T and the internal randomness of $\mathcal{A}$. This is equivalent to $\Pr[\mathrm{err}_S(\mathcal{A}'(S)) \leq \alpha] \geq 1 - \beta$, where the probability is taken over the internal randomness of $\mathcal{A}'$. Thus, $\mathcal{A}'$ is an (α, β)-PAC empirical learner for $\mathcal{C}$.

Moreover, since $\mathcal{A}'$ runs A on some dataset T, $\mathcal{A}$ is proper implies that $\mathcal{A}'$ is proper. ■

Appendix B. Proof of Lemma 15

Proof Let S_1, S_2 be two neighboring datasets and O be any set of outputs. Without loss of generality, we assume S_1 and S_2 differ on the first element. That is, $S_1 = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ and $S_2 = \{(x_1', y_1'), (x_2, y_2), \dots, (x_n, y_n)\}$. Define

$$p_t(I) = \Pr[\mathcal{A}_{\text{Agnostic}}(S_t) \in O \mid \text{the sampled index set is } I]$$

for $t \in \{1, 2\}$ and $I \subseteq [n]$ of size $\lceil \varepsilon n \rceil$. Since I is sampled uniformly at random, we have

$$\Pr[\mathcal{A}_{\text{Agnostic}}(S_t) \in O] = \frac{1}{\binom{n}{|I|}} \sum_I p_t(I).$$

Now consider a fixed index set I. Let T_1, W_1 and T_2, W_2 be the corresponding partitions of S_1 and S_2. We will consider two cases: $1 \in I$ and $1 \notin I$.

When $1 \notin I$, we have $T_1 = T_2$. Therefore, $\mathcal{A}_{\text{Relabel}}(T_1, W_1)$ and $\mathcal{A}_{\text{Relabel}}(T_2, W_2)$ will construct the same candidate set H. For every $h \in H$, suppose its score function is minimized by f_1 on dataset $T_1 \circ W_1$, i.e., $q(T_1 \circ W_1, h) = \text{dis}_{(T_1)_{\mathcal{X}}}(h, f_1) + \text{err}_{W_1}(f_1)$. Since W_1 and W_2 are neighboring datasets, we can bound $q(T_2 \circ W_2, h)$ as follows:

$$\begin{aligned} q(T_2 \circ W_2, h) &= \min_{f \in \mathcal{C}} \left\{ \text{dis}_{(T_2)_{\mathcal{X}}}(h, f) + \text{err}_{W_2}(f) \right\} \\ &\le \text{dis}_{(T_2)_{\mathcal{X}}}(h, f_1) + \text{err}_{W_2}(f_1) \\ &\le \text{dis}_{(T_1)_{\mathcal{X}}}(h, f_1) + \text{err}_{W_1}(f_1) + \frac{1}{|W_1|} \\ &= q(T_1 \circ W_1, h) + \frac{1}{|W_1|}. \end{aligned}$$

By symmetry, $q(T_2 \circ W_2, h) \le q(T_1 \circ W_1, h) + \frac{1}{|W_1|}$. Therefore, the sensitivity of q is $\frac{1}{|W_1|}$. It then follows by Lemma 11 that

$$\Pr[\mathcal{A}_{\text{Relabel}}(T_1, W_1) = T_1^h] \le e^{\varepsilon} \Pr[\mathcal{A}_{\text{Relabel}}(T_2, W_2) = T_2^h]$$

for any $T_1^h = T_2^h$. The post-processing property of DP immediately implies

$$p_1(I) \le e^{\varepsilon} p_2(I).$$

We then turn to the case that $1 \in I$. We will prove the following conclusion for every $i \notin I$:

$$p_1(I) \le e^{O(1)} p_2((I \setminus \{1\}) \cup \{i\}) + O(\delta).$$

Let T_2' and W_2' be the partitions of S_2 using $(I \setminus \{1\}) \cup \{i\}$ as the index set. Since $1 \in I$, we have $T_1 \setminus \{(x_1, y_1)\} = T_2' \setminus \{(x_i, y_i)\} = \hat{T}$ for some $\hat{T}$ of size $|I| - 1$. Let H_1 and H_2' denote the candidate sets constructed during the execution of $\mathcal{A}_{\text{Relabel}}(T_1, W_1)$ and $\mathcal{A}_{\text{Relabel}}(T_2', W_2')$. For each possible labeling $\hat{T}^c$ of $\hat{T}_{\mathcal{X}}$, define $P_1(c) = \left\{ f \in H_1 : \text{err}_{\hat{T}^c}(f) = 0 \right\}$ and $P_2'(c) = \left\{ f \in H_2' : \text{err}_{\hat{T}^c}(f) = 0 \right\}$, i.e., the sets consisting of hypotheses in H_1 and H_2' that agree with c on $\hat{T}_{\mathcal{X}}$. Since the label set is $\{0, 1\}$, we have $1 \le |P_1(c)|, |P_2'(c)| \le 2$.

We next pick arbitrary $h_1 \in P_1(c)$ and $h_2' \in P_2'(c)$ and compare their scores. Suppose $q(T_1 \circ W_1, h_1)$ is minimized by f_1. Note that W_1 and W_2' are neighboring datasets (since $W_1 \setminus \{(x_i, y_i)\} = W_2' \setminus \{(x_1', y_1')\}$), and h_1 and h_2' agree on $\hat{T}_{\mathcal{X}}$, we have

$$\begin{aligned}
q(T_2' \circ W_2', h_2') &= \min_{f \in \mathcal{C}} \left\{ \mathrm{dis}_{(T_2')_{\mathcal{X}}}(h_2', f) + \mathrm{err}_{W_2'}(f) \right\} \\
&\le \mathrm{dis}_{(T_2')_{\mathcal{X}}}(h_2', f_1) + \mathrm{err}_{W_2'}(f_1) \\
&\le \mathrm{dis}_{(T_1)_{\mathcal{X}}}(h_1, f_1) + \frac{1}{|T_1|} + \mathrm{err}_{W_1}(f_1) + \frac{1}{|W_1|} \\
&= q(T_1 \circ W_1, h_1) + \frac{1}{|T_1|} + \frac{1}{|W_1|}.
\end{aligned}$$

Since $|T_1| \ge n\varepsilon$ and $|W_1| \le n - n\varepsilon$, we have $\frac{|W_1|}{|T_1|} \le \frac{1-\varepsilon}{\varepsilon}$. Thus,

$$\begin{aligned}
\exp(-\varepsilon \cdot q(T_2' \circ W_2', h_2')/2\Delta) &\ge \exp\left(-\varepsilon \cdot \left(q(T_1 \circ W_1, h_1) + \frac{1}{|T_1|} + \frac{1}{|W_1|}\right)/2\Delta\right) \\
&= \exp(-\varepsilon \cdot q(T_1 \circ W_1, h_1)/2\Delta) \cdot \exp\left(-\frac{\varepsilon}{2} \cdot \left(\frac{|W_1|}{|T_1|} + \frac{|W_1|}{|W_1|}\right)\right) \\
&\ge \exp(-\varepsilon \cdot q(T_1 \circ W_1, h_1)/2\Delta) \cdot \exp\left(-\frac{\varepsilon}{2} \cdot \left(\frac{1-\varepsilon}{\varepsilon} + 1\right)\right) \\
&= \exp(-\varepsilon \cdot q(T_1 \circ W_1, h_1)/2\Delta) \cdot \exp(-1/2).
\end{aligned}$$

By symmetry (because the above analysis only relies on the facts that W_1 and W_2' are neighboring and h_1 and h_2' agree on $\hat{T}_{\mathcal{X}}$), we have

$$\exp(-\varepsilon \cdot q(T_1 \circ W_1, h_1)/2\Delta) \ge \exp(-\varepsilon \cdot q(T_2' \circ W_2', h_2')/2\Delta) \cdot \exp(-1/2).$$

Then, the fact that $1 \le |P_1(c)|, |P_2'(c)| \le 2$ gives

$$\sum_{h_1 \in P_1(c)} \exp(-\varepsilon \cdot q(T_1 \circ W_1, h_1)/2\Delta) \ge \frac{1}{2} \sum_{h_2' \in P_2'(c)} \exp(-\varepsilon \cdot q(T_2' \circ W_2', h_2')/2\Delta) \cdot \exp(-1/2).$$

Summing over all hypotheses in H_1, we get

$$\begin{aligned}
\sum_{f \in H_1} \exp(-\varepsilon \cdot q(T_1 \circ W_1, f)/2\Delta) &= \sum_{\hat{T}^c} \sum_{h_1 \in P_1(c)} \exp(-\varepsilon \cdot q(T_1 \circ W_1, h_1)/2\Delta) \\
&\ge \sum_{\hat{T}^c} \frac{1}{2} \sum_{h_2' \in P_2'(c)} \exp(-\varepsilon \cdot q(T_2' \circ W_2', h_2')/2\Delta) \cdot \exp(-1/2) \\
&= \frac{1}{2\sqrt{e}} \sum_{f \in H_2'} \exp(-\varepsilon \cdot q(T_2' \circ W_2', f)/2\Delta).
\end{aligned}$$

Note that $(T_1)^{h_1}$ and $(T_2')^{h_2'}$ are neighboring datasets (since h_1 and h_2' agree on $\hat{T}_{\mathcal{X}}$). Then by the fact that $\mathcal{A}$ is $(1, \delta)$-differentially private, we have

$$\begin{aligned}
&\Pr[\mathcal{A}_{\text{Relabel}}(T_1, W_1) = (T_1)^{h_1}] \cdot \Pr[\mathcal{A}((T_1)^{h_1}) \in O] \\
&= \frac{\exp(-\varepsilon \cdot q(T_1 \circ W_1, h_1)/2\Delta)}{\sum_{f \in H_1} \exp(-\varepsilon \cdot q(T_1 \circ W_1, f)/2\Delta)} \cdot \Pr[\mathcal{A}((T_1)^{h_1}) \in O] \\
&\leq 2e \cdot \frac{\exp(-\varepsilon \cdot q(T_2' \circ W_2', h_2')/2\Delta)}{\sum_{f \in H_2'} \exp(-\varepsilon \cdot q(T_2' \circ W_2', f)/2\Delta)} \cdot (e \cdot \Pr[\mathcal{A}((T_2')^{h_2'}) \in O] + \delta) \\
&= 2e \cdot \Pr[\mathcal{A}_{\text{Relabel}}(T_2', W_2') = (T_2')^{h_2'}] \cdot (e \cdot \Pr[\mathcal{A}((T_2')^{h_2'}) \in O] + \delta).
\end{aligned}$$

We can then bound $p_1(I)$ as follows:

$$\begin{aligned}
p_1(I) &= \Pr[\mathcal{A}(\mathcal{A}_{\text{Relabel}}(T_1, W_1)) \in O] \\
&= \sum_{\hat{T}^c} \sum_{h_1 \in P_1(c)} \Pr[\mathcal{A}_{\text{Relabel}}(T_1, W_1) = (T_1)^{h_1}] \cdot \Pr[\mathcal{A}((T_1)^{h_1}) \in O] \\
&\leq \sum_{\hat{T}^c} 2 \sum_{h_2' \in P_2'(c)} 2e \cdot \Pr[\mathcal{A}_{\text{Relabel}}(T_2', W_2') = (T_2')^{h_2'}] \cdot (e \cdot \Pr[\mathcal{A}((T_2')^{h_2'}) \in O] + \delta) \\
&= 4e \cdot \big(e \cdot \Pr[\mathcal{A}(\mathcal{A}_{\text{Relabel}}(T_2', W_2')) \in O] + \delta\big) \\
&= e^{2+2\ln 2} p_2((I \setminus \{1\}) \cup \{i\}) + 4e\delta.
\end{aligned}$$

Note that the summation $\sum_{I:1 \in I} \sum_{i \in [n] \setminus I} p_2((I \setminus \{1\}) \cup \{i\})$ actually counts every $p_2(I)$ (where $1 \notin I$) exactly $|I|$ times. Thus,

$$\begin{aligned}
\sum_{I:1 \in I} p_1(I) &= \frac{1}{n - |I|} \sum_{I:1 \in I} \sum_{i \in [n] \setminus I} p_1(I) \\
&\leq \frac{1}{n - |I|} \sum_{I:1 \in I} \sum_{i \in [n] \setminus I} \Big[e^{2+2\ln 2} p_2((I \setminus \{1\}) \cup \{i\}) + 4e\delta \Big] \\
&= \frac{|I|}{n - |I|} \sum_{I:1 \notin I} e^{2+2\ln 2} p_2(I) + 4e\delta \cdot \binom{n-1}{|I|-1} \\
&= O(\varepsilon) \sum_{I:1 \notin I} p_2(I) + O(\delta) \cdot \binom{n-1}{|I|-1},
\end{aligned}$$

where in the last line we use the fact that

$$\frac{|I|}{n - |I|} = \frac{\lceil \varepsilon n \rceil}{n - \lceil \varepsilon n \rceil} \leq \frac{2\varepsilon n}{n - 2\varepsilon n} \leq 6\varepsilon$$

assuming $\varepsilon n \geq 1$ and $\varepsilon \leq 1/3$. This implies that

$$\begin{aligned}
\Pr[\mathcal{A}_{\text{Agnostic}}(S_1) \in O] &= \frac{1}{\binom{n}{|I|}} \left(\sum_{I:1 \notin I} p_1(I) + \sum_{I:1 \in I} p_1(I) \right) \\
&\leq \frac{1}{\binom{n}{|I|}} \left(e^{\varepsilon} \sum_{I:1 \notin I} p_2(I) + O(\varepsilon) \sum_{I:1 \notin I} p_2(I) + O(\delta) \cdot \binom{n-1}{|I|-1} \right) \\
&\leq (e^{\varepsilon} + O(\varepsilon)) \cdot \frac{1}{\binom{n}{|I|}} \sum_{I} p_2(I) + O(\delta) \cdot \frac{|I|}{n} \\
&\leq e^{O(\varepsilon)} \Pr[\mathcal{A}_{\text{Agnostic}}(S_2) \in O] + O(\varepsilon\delta).
\end{aligned}$$

■

Appendix C. Proof of Claim 16

We use the following technical lemma (Anthony and Bartlett, 1999) in bounding the sample complexity incurred by the exponential mechanism.

Lemma 22 *Let $d \geq 1$ and $\alpha, \beta \in (0,1)$. Then if $n \geq \frac{2d\ln(2/\alpha)+2\ln(1/\beta)}{\alpha}$, we have*

$$n\alpha \geq d \ln\left(\frac{en}{d}\right) + \ln\left(\frac{1}{\beta}\right).$$

Proof [Proof of Claim 16] Define the following three events:

- E_1: For every $c \in \mathcal{C}$, it holds that $|\text{err}_{\mathcal{D}}(c) - \text{err}_W(c)| \leq \alpha/9$.
- E_2: The exponential mechanism chooses an $h \in H$ such that
$$q(T \circ W, h) \leq \min_{f \in H} q(T \circ W, f) + \alpha/9.$$
- E_3: For any $h_1, h_2 \in \mathcal{C}$ such that $\text{dis}_{T_{\mathcal{X}}}(h_1, h_2) \leq \alpha/3$, it holds that $\text{dis}_{\mathcal{D}_{\mathcal{X}}}(h_1, h_2) \leq 2\alpha/3$.

We first show that T will be relabeled by some h such that $\text{err}_{\mathcal{D}}(h) \leq \inf_{c \in \mathcal{C}} \text{err}_{\mathcal{D}}(c) + \alpha$ if the above events happen. Let $\eta = \inf_{c \in \mathcal{C}} \text{err}_{\mathcal{D}}(c)$. Let $f_0 \in \mathcal{C}$ be some concept that minimizes the empirical error on W, i.e., $\text{err}_W(f_0) = \min_{c \in \mathcal{C}} \text{err}_W(c)$. Then E_1 implies that $\text{err}_W(f_0) \leq \inf_{c \in \mathcal{C}} \text{err}_W(c) \leq \inf_{c \in \mathcal{C}} \text{err}_{\mathcal{D}}(c) + \alpha/9 = \eta + \alpha/9$. Since every labeling in $\Pi_{\mathcal{C}}(T_{\mathcal{X}})$ is labeled by some $h \in H$, there exists some $h_0 \in H$ that agrees with f_0 on $T_{\mathcal{X}}$. Thus,

$$\begin{aligned}
q(T \circ W, h_0) &= \min_{f \in \mathcal{C}} \{\text{dis}_{T_{\mathcal{X}}}(h_0, f) + \text{err}_W(f)\} \\
&\leq \text{dis}_{T_{\mathcal{X}}}(h_0, f_0) + \text{err}_W(f_0) \\
&\leq \eta + \alpha/9.
\end{aligned}$$

Then, event E_2 ensures that the exponential mechanism outputs some $h \in H$ such that

$$\begin{aligned} q(T \circ W, h) &\le \min_{f \in H} q(T \circ W, f) + \alpha/9 \\ &\le q(T \circ W, h_0) + \alpha/9 \\ &\le \eta + 2\alpha/9. \end{aligned}$$

Suppose $q(T \circ W, h) = \mathrm{dis}_{T_{\mathcal{X}}}(h, f) + \mathrm{err}_W(f)$ for some $f \in \mathcal{C}$. Event E_1 ensures that

$$\mathrm{err}_{\mathcal{D}}(f) \le \mathrm{err}_W(f) + \alpha/9 \le q(T \circ W, h) + \alpha/9 \le \eta + \alpha/3.$$

Moreover, since $\mathrm{err}_{\mathcal{D}}(f) \ge \eta$, we have

$$\begin{aligned} \mathrm{dis}_{T_{\mathcal{X}}}(h, f) &= q(T \circ W, h) - \mathrm{err}_W(f) \\ &\le q(T \circ W, h) - (\mathrm{err}_{\mathcal{D}}(f) - \alpha/9) \\ &\le q(T \circ W, h) - \eta + \alpha/9 \\ &\le \eta + 2\alpha/9 - \eta + \alpha/9 \\ &= \alpha/3. \end{aligned}$$

Event E_3 then ensures that $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(h, f) \le 2\alpha/3$. By the triangle inequality, we obtain

$$\mathrm{err}_{\mathcal{D}}(h) \le \mathrm{err}_{\mathcal{D}}(f) + \mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(h, f) \le \eta + \alpha/3 + 2\alpha/3 = \eta + \alpha.$$

To complete the proof, we now show $E_1 \cap E_2 \cap E_3$ happens with probability $1-\beta$. By Lemma 8, E_1 happens with probability $1 - \beta/3$ given that $|W| \ge C \cdot \frac{\mathtt{VC}(\mathcal{C}) + \ln(1/\beta)}{\alpha^2}$ for some constant C.

We then consider E_2. By Sauer's Lemma (Lemma 6), we have $|H| \le \left(\frac{e|T|}{\mathtt{VC}(\mathcal{C})}\right)^{\mathtt{VC}(\mathcal{C})}$. Then by Lemma 11, with probability $1 - \beta/3$, the exponential mechanism selects some h such that

$$\begin{aligned} q(T \circ W, h) &\le \min_{f \in H} q(T \circ W, f) + \frac{2}{|W|\varepsilon} \ln(|H|/\beta) \\ &\le \min_{f \in H} q(T \circ W, f) + \frac{12}{|T|} \left(\mathtt{VC}(\mathcal{C}) \ln\left(\frac{e|T|}{\mathtt{VC}(\mathcal{C})}\right) + \ln(1/\beta) \right) \\ &\le \min_{f \in H} q(T \circ W, f) + \alpha/9, \end{aligned}$$

where in the second inequality we use $|W| \ge \frac{|T|}{6\varepsilon}$ and in the last inequality we apply Lemma 22, which requires $|T| \ge C' \cdot \frac{\mathtt{VC}(\mathcal{C}) \ln(1/\alpha) + \ln(1/\beta)}{\alpha}$. This means E_2 happens with probability $1 - \beta/3$.

Finally, Lemma 7 implies that event E_3 happens with probability $1 - \beta/3$ given that $|T| \ge C'' \cdot \frac{\mathtt{VC}(\mathcal{C}) \ln(1/\alpha) + \ln(1/\beta)}{\alpha}$. Therefore, we have $\Pr[E_1 \cap E_2 \cap E_3] \ge 1 - \beta$ by the union bound. ■

Appendix D. Proof of Lemma 17

To prove Lemma 17, we follow the proof strategy of Alon et al. (2020). We first construct an auxiliary algorithm ($\mathcal{A}_{\text{Auxiliary}}$) and prove the following claim, which allows us to employ the generalization property of DP. The proof is analogous to part of the proof of Lemma 15.

Claim 23 *Suppose $\mathcal{A}$ is $(1,\delta)$-differentially private. For any public V and W, $\mathcal{A}_{\text{Auxiliary}}(U,V,W)$ is $(2+2\ln 2, 4e\delta)$-differentially private with respect to U.*

Proof Let U_1 and U_2 be two neighboring datasets and O be any set of the outputs of $\mathcal{A}_{\text{Auxiliary}}$. Then $T_1 = U_1 \circ V$ and $T_2 = U_2 \circ V$ are also neighboring datasets. Therefore, there is some $\hat{T}$ of size $|T_1| - 1$ such that $T_1 \setminus \{(x_1, y_1)\} = T_2 \setminus \{(x_1', y_1')\}$ for data points $(x_1, y_1) \in T_1$ and $(x_1', y_1') \in T_2$. Let H_t be the candidate set constructed during the execution of $\mathcal{A}_{\text{Relabel}}(T_t, W)$, where $t \in \{1, 2\}$. For each possible labeling $\hat{T}^c$ of $\hat{T}_{\mathcal{X}}$, let $P_t(c)$ be the set of hypotheses added to H_t in the execution of $\mathcal{A}_{\text{Relabel}}(T_t, W)$. Pick arbitrary $h_1 \in P_1(c)$ and $h_2 \in P_2(c)$ and suppose $q(T_1 \circ W, h_1)$ is minimized by f_1, we have

$$\begin{aligned}
q(T_2 \circ W, h_2) &= \min_{f \in \mathcal{C}} \left\{ \text{dis}_{(T_2)_{\mathcal{X}}}(h_2, f) + \text{err}_W(f) \right\} \\
&\le \text{dis}_{(T_2)_{\mathcal{X}}}(h_2, f_1) + \text{err}_W(f_1) \\
&\le \text{dis}_{(T_1)_{\mathcal{X}}}(h_1, f_1) + \frac{1}{|T_1|} + \text{err}_W(f_1) \\
&= q(T_1 \circ W, h_1) + \frac{1}{|T_1|}.
\end{aligned}$$

Since $|T_1| \ge \varepsilon n$ and $|W| \le n - \varepsilon n$, we have $\frac{|W|}{|T_1|} \le \frac{1-\varepsilon}{\varepsilon}$. Thus,

$$\begin{aligned}
\exp(-\varepsilon \cdot q(T_2 \circ W, h_2)/2\Delta) &\ge \exp\left(-\varepsilon \cdot \left(q(T_1 \circ W, h_1) + \frac{1}{|T_1|}\right)/2\Delta\right) \\
&= \exp(-\varepsilon \cdot q(T_1 \circ W, h_1)/2\Delta) \cdot \exp\left(-\frac{\varepsilon |W|}{2|T_1|}\right) \\
&\ge \exp(-\varepsilon \cdot q(T_1 \circ W, h_1)/2\Delta) \cdot \exp\left(-\frac{1-\varepsilon}{2}\right) \\
&\ge \frac{1}{\sqrt{e}} \cdot \exp(-\varepsilon \cdot q(T_1 \circ W, h_1)/2\Delta).
\end{aligned}$$

By symmetry, we also have

$$\exp(-\varepsilon \cdot q(T_1 \circ W, h_1)/2\Delta) \ge \frac{1}{\sqrt{e}} \cdot \exp(-\varepsilon \cdot q(T_2 \circ W, h_2)/2\Delta).$$

The fact that $1 \le |P_1(c)|, |P_2(c)| \le 2$ gives

$$\sum_{h_1 \in P_1(c)} \exp(-\varepsilon \cdot q(T_1 \circ W, h_1)/2\Delta) \ge \frac{1}{2\sqrt{e}} \sum_{h_2 \in P_2(c)} \exp(-\varepsilon \cdot q(T_2 \circ W, h_2)/2\Delta).$$

Therefore,

$$\begin{aligned}
\sum_{f \in H_1} \exp(-\varepsilon \cdot q(T_1 \circ W, f)/2\Delta) &= \sum_{\hat{T}^c} \sum_{h_1 \in P_1(c)} \exp(-\varepsilon \cdot q(T_1 \circ W, h_1)/2\Delta) \\
&\ge \sum_{\hat{T}^c} \frac{1}{2} \sum_{h_2 \in P_2(c)} \exp(-\varepsilon \cdot q(T_2 \circ W, h_2)/2\Delta) \cdot \exp(-1/2) \\
&= \frac{1}{2\sqrt{e}} \sum_{f \in H_2} \exp(-\varepsilon \cdot q(T_2 \circ W, f)/2\Delta).
\end{aligned}$$

Since h_1 and h_2 agree on $\hat{T}_{\mathcal{X}}$, $(T_1)^{h_1}$ and $(T_2)^{h_2}$ are neighboring datasets. Moreover, note that h_1 and h_2 agree on $V_{\mathcal{X}}$ because $V_{\mathcal{X}}$ is just part of $\hat{T}_{\mathcal{X}}$. Therefore, $\mathcal{A}_{\text{Auxiliary}}(U_1, V, W)$ and $\mathcal{A}_{\text{Auxiliary}}(U_2, V, W)$ will select the same $\bar{h}$ from $V^{h_1} = V^{h_2}$. By the fact that $\mathcal{A}$ is $(1, \delta)$-differentially private, we obtain

$$\begin{aligned}
&\Pr[\mathcal{A}_{\text{Relabel}}(T_1, W) = (T_1)^{h_1}] \cdot \Pr[\mathcal{A}((T_1)^{h_1}) \oplus \bar{h} \in O] \\
&= \frac{\exp(-\varepsilon \cdot q(T_1 \circ W, h_1)/2\Delta)}{\sum_{f \in H_1} \exp(-\varepsilon \cdot q(T_1 \circ W, f)/2\Delta)} \cdot \Pr[\mathcal{A}((T_1)^{h_1}) \oplus \bar{h} \in O] \\
&\leq 2e \cdot \frac{\exp(-\varepsilon \cdot q(T_2 \circ W, h_2)/2\Delta)}{\sum_{f \in H_2} \exp(-\varepsilon \cdot q(T_2 \circ W, f)/2\Delta)} \cdot (e \cdot \Pr[\mathcal{A}((T_2)^{h_2}) \oplus \bar{h} \in O] + \delta) \\
&= 2e \cdot \Pr[\mathcal{A}_{\text{Relabel}}(T_2, W) = (T_2)^{h_2}] \cdot (e \cdot \Pr[\mathcal{A}((T_2)^{h_2}) \oplus \bar{h} \in O] + \delta).
\end{aligned}$$

Let $\bar{h} = \bar{h}(V^h)$ denote the selection rule of $\bar{h}$. Summing over all labelings gives

$$\begin{aligned}
&\Pr[\mathcal{A}_{\text{Auxiliary}}(U_1, V, W) \in O] \\
&= \sum_{\hat{T}^c} \sum_{h_1 \in P_1(c)} \Pr[\mathcal{A}_{\text{Relabel}}(T_1, W) = (T_1)^{h_1}] \cdot \Pr[\mathcal{A}((T_1)^{h_1}) \oplus \bar{h}(V^{h_1}) \in O] \\
&\leq \sum_{\hat{T}^c} 2 \sum_{h_2 \in P_2(c)} 2e \cdot \Pr[\mathcal{A}_{\text{Relabel}}(T_2, W) = (T_2)^{h_2}] \cdot (e \cdot \Pr[\mathcal{A}((T_2)^{h_2}) \oplus \bar{h}(V^{h_2}) \in O] + \delta) \\
&= 4e \cdot (e \cdot \Pr[\mathcal{A}_{\text{Auxiliary}}(U_2, V, W) \in O] + \delta) \\
&= e^{2 + 2\ln 2} \Pr[\mathcal{A}_{\text{Auxiliary}}(U_2, V, W) \in O] + 4e\delta.
\end{aligned}$$

■

Proof [Proof of Lemma 17] Recall that

$$n = O\left(\frac{m}{\varepsilon} + \frac{\texttt{VC}(\mathcal{C})\log(1/\alpha) + \log(1/\beta)}{\alpha\varepsilon} + \frac{\texttt{VC}(\mathcal{C}) + \log(1/\beta)}{\alpha^2}\right).$$

Moreover, assuming $\varepsilon \leq 1/3$ and $\varepsilon n \geq 1$, we have

$$\frac{|W|}{|T|} = \frac{n - \lceil \varepsilon n \rceil}{\lceil \varepsilon n \rceil} \geq \frac{n - 2\varepsilon n}{2\varepsilon n} \geq \frac{1}{6\varepsilon}.$$

Therefore, it is not hard to verify that the conditions in Claim 16 are fulfilled. Thus, with probability $1 - \beta$, T will be relabeled by some h with $\text{err}_{\mathcal{D}}(h) \leq \inf_{c \in \mathcal{C}} \text{err}_{\mathcal{D}}(c) + \alpha$.

Since $\mathcal{A}$ is an (α, β)-PAC empirical learner, the final output hypothesis $g = \mathcal{A}(T^h)$ satisfies $\text{err}_{T^h}(g) \leq \alpha$ with probability $1 - \beta$. This is equivalent to $\text{dis}_{T_{\mathcal{X}}}(h, g) \leq \alpha$. Suppose we choose $|U| = |V| = |T|/2$ in $\mathcal{A}_{\text{Auxiliary}}$. Note that h and $\bar{h}$ agree on $V_{\mathcal{X}}$, the realizable generalization property (Lemma 7) implies that $\text{dis}_{\mathcal{D}_{\mathcal{X}}}(h, \bar{h}) \leq \alpha$ with probability $1 - \beta$. Applying Lemma 7 again (over $T_{\mathcal{X}}$), we have $\text{dis}_{T_{\mathcal{X}}}(h, \bar{h}) \leq 2\alpha$ with probability $1 - \beta$. Therefore,

$$\text{dis}_{U_{\mathcal{X}}}(g, \bar{h}) \leq \text{dis}_{T_{\mathcal{X}}}(g, \bar{h}) \leq \text{dis}_{T_{\mathcal{X}}}(g, h) + \text{dis}_{T_{\mathcal{X}}}(h, \bar{h}) \leq 3\alpha.$$

We then bound the generalization disagreement between g and $\bar{h}$ using the generalization property of DP. By Claim 23, $\mathcal{A}_{\text{Auxiliary}}$ is $(O(1), O(\delta))$-differentially private with respect to U. Therefore, it is also $(O(1), O(\delta + \beta/|U|))$-differentially private. Then by Lemma 12 and the fact that

$|U| = |T|/2 \geq C\ln(1/\beta)/\alpha$ for some large constant C, we have

$$\begin{aligned}\mathrm{dis}_{\mathcal{D}_\mathcal{X}}(g,\bar{h}) &\leq O\left(\mathrm{dis}_{U_\mathcal{X}}(g,\bar{h}) + \frac{1}{|U|}\log\left(\frac{1}{\delta|U|+\beta}\right)\right)\\ &\leq O\left(\alpha + \frac{1}{|U|}\log\left(\frac{1}{\beta}\right)\right)\\ &\leq O(\alpha)\end{aligned}$$

with probability $1 - O(\delta|U| + \beta)$.

Putting all things together, the union bound ensures that with probability $1 - O(\delta|U| + \beta) = 1 - O(\varepsilon\delta n + \beta)$, we have

$$\begin{aligned}\mathrm{err}_{\mathcal{D}}(g) &\leq \mathrm{err}_{\mathcal{D}}(h) + \mathrm{dis}_{\mathcal{D}_\mathcal{X}}(h,g)\\ &\leq \inf_{c\in\mathcal{C}} \mathrm{err}_{\mathcal{D}}(c) + \alpha + \mathrm{dis}_{\mathcal{D}_\mathcal{X}}(h,\bar{h}) + \mathrm{dis}_{\mathcal{D}_\mathcal{X}}(g,\bar{h})\\ &\leq \inf_{c\in\mathcal{C}} \mathrm{err}_{\mathcal{D}}(c) + O(\alpha).\end{aligned}$$

■

Appendix E. Discussion on the Computational Complexity

The relabeling procedure in our transformation has two main steps:

1. Constructing a candidate set that contains all the labelings.
2. Running the exponential mechanism.

By Sauer's Lemma, the size of the candidate set is $O(n^{\mathtt{VC}(\mathcal{C})})$, where n is the sample size. If we do not require the time complexity to be polynomial in $\mathtt{VC}(\mathcal{C})$ (i.e., we treat $\mathtt{VC}(\mathcal{C})$ as a fixed constant), step 2 can be done with a polynomial number of calls of an ERM oracle. Although step 1 requires enumerating 2^n labelings for general classes, it can be done efficiently for classes with certain structures (e.g., point functions, thresholds, and axis-aligned rectangles). In these cases, the reduction is efficient. In summary, the transformation is inefficient in general, but can be made efficient for special classes.

We remark that there are classes (e.g., halfspaces) that are (computationally) easy to learn in the realizable setting but hard to learn in the agnostic setting under computational or cryptographic assumptions (Feldman et al., 2009). Since our algorithm is a reduction from agnostic learning to realizable learning, we cannot hope that it will be generally efficient given these hardness results. Hence, we focus on information-theoretic bounds in this work.