HARVARD DATA SCIENCE REVIEW

Volume 4, Issue 2
DOI: /10.1162/99608f92.16c71dad
ISSN: 2644-2353

Private Prediction Sets

Anastasios N. Angelopoulos[*], Stephen Bates[*], Tijana Zrnic[*], Michael I. Jordan

University of California, Berkeley

April 15, 2022

Abstract

In real-world settings involving consequential decision-making, the deployment of machine learning systems generally requires both reliable uncertainty quantification and protection of individuals' privacy. We present a framework that treats these two desiderata jointly. Our framework is based on conformal prediction, a methodology that augments predictive models to return *prediction sets* that provide uncertainty quantification—they provably cover the true response with a user-specified probability, such as 90%. One might hope that when used with privately trained models, conformal prediction would yield privacy guarantees for the resulting prediction sets; unfortunately this is *not* the case. To remedy this key problem, we develop a method that takes any pretrained predictive model and outputs differentially private prediction sets. Our method follows the general approach of split conformal prediction; we use holdout data to calibrate the size of the prediction sets but preserve privacy by using a privatized quantile subroutine. This subroutine compensates for the noise introduced to preserve privacy in order to guarantee correct coverage. We evaluate the method on large-scale computer vision data sets.

1 Introduction

The impressive predictive accuracies of black-box machine learning algorithms on tightly controlled test beds do not sanctify their use in consequential applications. For example, given the gravity of medical decision-making, automated diagnostic predictions must come with rigorous instance-wise uncertainty to avoid silent, high-consequence failures. Furthermore, medical data science requires privacy guarantees, since individuals would suffer material harm were their data to be accessed or reconstructed by a nefarious actor. While uncertainty quantification and privacy are generally dealt with in isolation, they arise together in many real-world predictive systems, and, as we discuss, they interact. Accordingly, the work that we present here involves a framework that addresses uncertainty and privacy jointly. Specifically, we develop a differentially private version of conformal prediction that results in private, rigorous, finite-sample uncertainty quantification for any model and any data set at little computational cost.

This work takes a *modular* viewpoint on data science: we seek to give practitioners the flexibility to train whatever underlying private model gives the best performance (e.g., via deep learning), then later endow the model with rigorous statistical properties, without modifying that underlying model. See Figure 1. Zooming out, perhaps the most consistent trend in the history of engineering is that modular engineering building blocks—like integrated circuits or reaction chemistry—are the key to scalable and deployable systems where each part can be improved and debugged separately. *Private prediction sets* allow simple, private uncertainty quantification for any system and thus provide conceptual building blocks for data scientists constructing such systems.

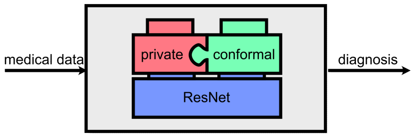

Figure 1: Modular data science.

[*] equal contribution

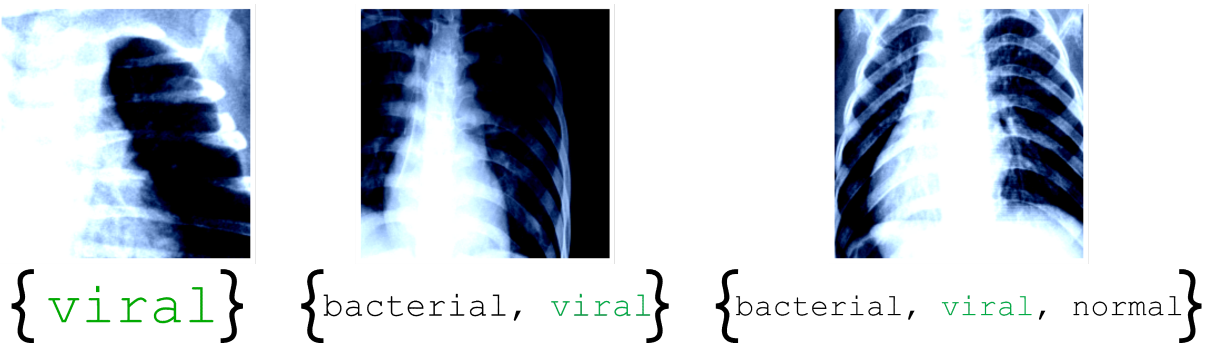

Figure 2: **Examples of private conformal prediction sets on COVID-19 data.** We show three examples of lung X-rays taken from the CoronaHack data set (Perez et al., 2020) with their corresponding private prediction sets at $\alpha = 10\%$ from a ResNet-18. All three patients had `viral pneumonia` (likely COVID-19). The classes in the prediction sets appear in ranked order according to the softmax score of the model; the center and right images are incorrectly classified if the predictor returns only the most likely class, but are correctly covered by the private prediction sets. See Experiment 4.4 for details.

Turning to the details, our approach builds on the notion of *prediction sets*—subsets of the response space that provably cover the true response variable with prespecified probability (e.g., 90%). Formally, for a test point with feature vector $X \in \mathcal{X}$ and response $Y \in \mathcal{Y}$, we compute an uncertainty set function, $\mathcal{C}(\cdot)$, mapping a feature vector to a subset of $\mathcal{Y}$ such that

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} \geq 1 - \alpha, \tag{1}$$

for a user-specified confidence level $1 - \alpha$, where $\alpha \in (0, 1)$. We use the output of an underlying predictive model (e.g., a pretrained, privatized neural network) along with a held-out *calibration data set*, $\{(X_i, Y_i)\}_{i=1}^n$, from the same distribution as (X, Y) to fit the set-valued function $\mathcal{C}(\cdot)$. The probability in expression (1) is therefore taken over both the randomness in (X, Y) and $\{(X_i, Y_i)\}_{i=1}^n$. If the underlying model expresses uncertainty, $\mathcal{C}$ will be large, signaling skepticism regarding the model's prediction.

Moreover, we introduce a *differentially private* mechanism for fitting $\mathcal{C}$, such that the sets that we compute have low sensitivity to the removal of any calibration point. This will allow an individual to contribute a calibration data point without fear that the prediction sets will reveal their sensitive information. Note that *even if the underlying model is trained in a privacy-preserving fashion, this provides no privacy guarantee for the calibration data.* Therefore, we will provide an adjustment that masks the calibration data set with additional randomness, addressing both privacy and uncertainty simultaneously.

See Figure 2 for a concrete example of private prediction sets applied to the automated diagnosis of COVID-19. In this setting, the prediction sets represent a set of plausible diagnoses based on an X-ray image—either `viral pneumonia` (presumed COVID-19), `bacterial pneumonia`, or `normal`. We guarantee that the true diagnosis is contained in the prediction set with high probability, while simultaneously ensuring that an adversary cannot detect the presence of any one of the X-ray images used to train the predictive system.

1.1 Our contribution

Our main contribution is a privacy-preserving algorithm that takes as input any predictive model together with a calibration data set, and outputs a set-valued function $\mathcal{C}(\cdot)$ that maps any input feature vector X to a set of labels such that the true label Y is contained in the predicted set with probability at least $1 - \alpha$, as per Equation (1). In order to generate prediction sets satisfying this property, we use ideas from split conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005; J. Lei et al., 2018), modifying this approach to ensure privacy. Importantly, if the provided predictive model is also trained in a differentially private way, then the whole pipeline that maps data to a prediction set function $\mathcal{C}(\cdot)$ is differentially private as well.

In Algorithm 1, we sketch our main procedure.

Algorithm 1 Private prediction sets (informal)

```
input: predictor $\hat{f}(\cdot)$, calibration data $\{(X_i, Y_i)\}_{i=1}^n$, privacy level $\epsilon > 0$, confidence level $\alpha \in (0, 1)$
For $1 \leq i \leq n$, compute conformity score $s_i = S_{\hat{f}}(X_i, Y_i)$
Compute $\epsilon$-differentially private $(1 - \alpha + O((n\epsilon)^{-1}))$-quantile of $\{s_i\}_{i=1}^n$, denoted $\hat{s}$
output: $\mathcal{C}(\cdot) = \{y : S_{\hat{f}}(\cdot, y) \leq \hat{s}\}$
```

Algorithm 1 first computes the conformity scores for all training samples. Informally, these scores indicate how well a feature–label pair 'conforms' to the provided model $\hat{f}$, a low score implying high conformity and a high score being indicative of an atypical point from the perspective of $\hat{f}$. Then, the algorithm generates a certain carefully chosen private quantile of the scores. Finally, it returns a prediction set function $\mathcal{C}(\cdot)$ which, for a given input feature vector, returns all labels that result in a conformity score below the critical threshold $\hat{s}$.

Our main theoretical result asserts that Algorithm 1 has strict coverage guarantees and is differentially private. In addition, we show that the coverage is almost *tight*, that is, not much higher than $1 - \alpha$.

Theorem 1 (Informal preview)**.** *The prediction set function $\mathcal{C}(\cdot)$ returned by Algorithm 1 is ϵ-differentially private and satisfies*

$$1 - \alpha \leq \mathbb{P}\{Y \in \mathcal{C}(X)\} \leq 1 - \alpha + O((n\epsilon)^{-1}).$$

We obtain a gap between the lower and upper bound on the probability of coverage to be roughly of the order $O((n\epsilon)^{-1})$, similar to the standard gap $O(n^{-1})$ without the privacy requirement. With this, we provide the first theoretical insight into the cost of privacy in conformal prediction. To shed further light on the properties of our procedure, we perform an extensive empirical study where we evaluate the tradeoff between the level of privacy on one hand, and the coverage and size of prediction sets on the other.

1.2 Related work

Differential privacy (Dwork et al., 2006) has become the de facto standard for privacy-preserving data analysis, as witnessed by its widespread adoption in large-scale systems such as those by Google (Erlingsson et al., 2014; Bittau et al., 2017), Apple (Differential Privacy Team Apple, 2017), Microsoft (Ding et al., 2017), and the US Census Bureau (Abowd, 2018; Dwork, 2019). This increasing adoption of differential privacy goes hand in hand with steady progress in differentially private model training, ranging across both convex (Chaudhuri et al., 2011; Bassily et al., 2014) and nonconvex (Abadi et al., 2016; Neel et al., 2020) settings. Our work complements these works by proposing a procedure that can be combined with any differentially private model training algorithm to account for the uncertainty of the resulting predictive model by producing a prediction set function with formal guarantees. At a technical level, closest to our algorithm on the privacy side are existing methods for reporting histograms and quantiles in a privacy-preserving fashion (Dwork et al., 2006; Xu et al., 2013; J. Lei, 2011; Smith, 2011; Feldman & Steinke, 2017). Finally, there have also been significant efforts to quantify uncertainty with formal privacy guarantees through various types of private confidence intervals (Karwa & Vadhan, 2017; Sheffet, 2017; Gaboardi et al., 2019; Wang et al., 2019). While prediction sets resemble confidence intervals, they are fundamentally different objects as they do not aim to cover a fixed parameter of the population distribution, but rather a randomly sampled outcome. As a result, existing methods for differentially private confidence intervals do not generalize to our problem setting.

Prediction sets as a way to represent uncertainty are a classical idea, going back at least to tolerance regions in the 1940s (Wilks, 1941, 1942; Wald, 1943; Tukey, 1947). See Krishnamoorthy & Mathew (2009) for an overview of tolerance regions and Park et al. (2020) for a recent application to deep learning models. Conformal prediction (Vovk et al., 1999, 2005; Shafer & Vovk, 2008) is a related way of producing predictive sets with finite-sample guarantees. Most relevant to the present work, *split conformal prediction* (Papadopoulos et al., 2002; J. Lei et al., 2015, 2018) is a convenient version that uses data splitting to give prediction sets in a computationally efficient way. Vovk (2015) and Barber et al. (2021) refine this approach to reuse data for both training and calibration, improving statistical efficiency. Recent work has targeted desiderata such as small set sizes (Sadinle et al., 2019; Angelopoulos et al., 2020), coverage that is approximately balanced across feature space (Vovk, 2012; Foygel Barber et al., 2019; Romano et al., 2019;

Izbicki et al., 2019; Romano et al., 2020; Guan, 2020; Cauchois, Gupta, & Duchi, 2020), and coverage that is balanced across classes (J. Lei, 2014; Sadinle et al., 2019; Hechtlinger et al., 2018; Guan & Tibshirani, 2019). Further extensions address problems in distribution estimation (Vovk et al., 2017, 2020), handling or testing distribution shift (Tibshirani et al., 2019; Cauchois, Gupta, Ali, & Duchi, 2020; Hu & Lei, 2020), causal inference (L. Lei & Candès, 2020), and controlling other notions of statistical error (Bates et al., 2021). We suggest (Angelopoulos & Bates, 2021) and (Shafer & Vovk, 2008) as introductory tutorials on conformal prediction for the unfamiliar reader. Lastly, we highlight two alternative approaches with a similar goal to conformal prediction. First, the calibration technique in Jung et al. (2020) and Gupta et al. (2021) generates prediction sets via the estimation of higher moments across many overlapping subpopulations. Second, there is a family of techniques that define a utility function balancing set-size and coverage and then search for set-valued predictors to maximize this utility (Grycko, 1993; del Coz et al., 2009; Mortier et al., 2020). The present work builds on split conformal prediction, but modifies the calibration step to preserve privacy.

2 Preliminaries

In this section, we formally introduce the main concepts in our problem setting. Split conformal prediction assumes access to a predictive model, $\hat{f}$, and aims to output *prediction sets* that achieve coverage by quantifying the uncertainty of $\hat{f}$ and the intrinsic randomness in X and Y. It quantifies this uncertainty using a *calibration data set* consisting of n i.i.d. samples, $\{(X_i, Y_i)\}_{i=1}^n$, that were not used to train $\hat{f}$. The calibration proceeds by defining a *score function* $S_{\hat{f}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Without loss of generality we take the range of this function to be the unit interval $[0, 1]$. The reader should think of the score as measuring the degree of consistency of the response Y with the features X based on the predictive model $\hat{f}$ (e.g., the size of the residual in a regression model), but any score function would lead to correct coverage. To simplify notation we will write $S(\cdot, \cdot)$ to denote the score, where we implicitly assume an underlying model $\hat{f}$. From this score function, one forms prediction sets as follows:

$$\mathcal{C}(x) = \{y : S(x, y) \leq \hat{s}\}, \tag{2}$$

for a choice of $\hat{s}$ based on the calibration dataset. In particular, $\hat{s}$ is taken to be a quantile of the calibration scores $s_i = S(X_i, Y_i)$ for $i = 1, \ldots, n$. In nonprivate conformal prediction, one simply takes $\hat{s}$ to be the $\big((n+1)(1-\alpha)\big)/n$ quantile, and then a standard argument shows that the coverage property in (1) holds. In this work we show how to take a modified private quantile that maintains this coverage guarantee.

As a concrete example of standard split conformal prediction, consider classifying an image in $\mathcal{X} = \mathbb{R}^{m \times d}$ into one of a thousand classes, $\mathcal{Y} = \{1, ..., 1000\}$. Given a standard classifier outputting a probability distribution over the classes, $\hat{f} : \mathcal{X} \to [0, 1]^{1000}$ (e.g., the output of a softmax layer), we can define a natural score function based on the activation of the correct class, $S(x, y) = 1 - \hat{f}(x)_y$. Then we take $\hat{s}$ as the upper $\lceil 0.9(n+1) \rceil / n$ quantile of the calibration scores $s_1, \ldots, s_n$ and define $\mathcal{C}$ as in equation (2). That is, we take as the cutoff $\hat{s}$ the value such that if we include all classes with estimated probability greater than $1 - \hat{s}$, our sets have (only slightly more than) 90% coverage on the calibration data. The result $\mathcal{C}(x)$ on a test point is then a set of plausible classes guaranteed to contain the true class with probability 90%. Our proposed method will follow a similar workflow, but with a slightly different choice of $\hat{s}$ to guarantee both coverage and privacy.

We next formally define differential privacy. We say that two data sets $\mathcal{D}, \mathcal{D}'$ are *neighboring* if they differ in a single element, that is, either data set can be obtained from the other by removing a single entry. For example, $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^n$ and $\mathcal{D}' = \mathcal{D} \setminus \{(X_0, Y_0)\}$, for some $(X_0, Y_0) \in \mathcal{D}$. Differential privacy then requires that two neighboring data sets produce similar distributions on the output.

Definition 1 (Differential privacy (Dwork et al., 2006))**.** A randomized algorithm $\mathcal{A}$ is ϵ-*differentially private* if for all neighboring data sets $\mathcal{D}$ and $\mathcal{D}'$, it holds that:

$$\mathbb{P}\{\mathcal{A}(\mathcal{D}) \in \mathcal{O}\} \leq e^{\epsilon} \mathbb{P}\{\mathcal{A}(\mathcal{D}') \in \mathcal{O}\},$$

for all measurable sets $\mathcal{O}$.

In short, if no adversary observing the algorithm's output can distinguish between $\mathcal{D}$ and a data set $\mathcal{D}'$ with the i-th entry removed, the presence of individual i in the analysis cannot be detected and hence their privacy is not compromised.

A key ingredient to our procedure is a privatized quantile of the conformity scores. We obtain this private quantile by discretizing the scores into bins and applying the exponential mechanism (McSherry & Talwar, 2007), one of the most ubiquitous tools in differential privacy. Our private quantile routine is then an extension of the private median routine proposed by Feldman and Steinke (Feldman & Steinke, 2017) to handle arbitrary quantiles. Specifically, let us fix a number of bins $m \in \mathbb{N}$, as well as edges $0 \equiv e_0 < e_1 < ... < e_{m-1} < e_m \equiv 1$. The edges define the bins $I_j = (e_{j-1}, e_j]$, $j = 1, ..., m$. We use Algorithm 2 with appropriately chosen quantile level q as a subroutine of our main conformal procedure.

Algorithm 2 Differentially private quantile

```
input: calibration scores $\{s_1, \dots, s_n\}$, bins $\{I_1, \dots, I_m\}$, privacy level $\epsilon$, level $q \in [0,1]$
For all $1 \le i \le n$, compute discretized score $[s_i] = e_j$, where $s_i \in I_j$
For all $1 \le j \le m$, compute $w_j = \max\left\{\frac{\#\{i:[s_i]<e_j\}}{q}, \frac{\#\{i:[s_i]>e_j\}}{1-q}\right\}$
Let $\hat{s} = e_j$ with probability $e^{-\frac{\epsilon}{2}w_j} / \sum_{j'=1}^{m} e^{-\frac{\epsilon}{2}w_{j'}}$
output: private quantile $\hat{s}$
```

3 Algorithm and Guarantees

We next precisely state our main algorithm and its formal guarantees. First, our algorithm has a calibration step, Algorithm 3, carried out one time using the calibration scores $s_1, \dots, s_n$ as input; this is the heart of our proposed procedure. The output of this step is a cutoff $\hat{s}$ learned from the calibration data. With this in hand, one forms the prediction set for a test point x as in equation (2), which for completeness we state in Algorithm 4.

Algorithm 3 Differentially private calibration

```
input: calibration scores $\{s_1, \dots, s_n\}$, privacy parameter $\epsilon$, coverage level $\alpha$, bins $\{I_1, \dots, I_m\}$
Compute $\tilde{q}$-quantile of $\{s_1, \dots, s_n\}$ via Algorithm 2, where $\tilde{q}$ is defined in (3), denoted $\hat{s}$
output: calibrated score cutoff $\hat{s}$
```

Algorithm 4 Differentially private prediction set

```
input: test point $x$, calibrated score cutoff $\hat{s}$
output: prediction set as in (2): $\mathcal{C}(x) = \{y : S(x,y) \le \hat{s}\}$.
```

This algorithm both satisfies differential privacy and guarantees correct coverage, as stated next in Proposition 1 and Theorem 2, respectively. The privacy property is a straightforward consequence of the privacy guarantees on the exponential mechanism (McSherry & Talwar, 2007).

Proposition 1 (Privacy guarantee)**.** *Algorithm 3 is ϵ-differentially private.*

Therefore, the main challenge for theory lies in understanding how to compensate for the added differentially private noise in order to get strict, distribution-free coverage guarantees.

Theorem 2 (Coverage guarantee)**.** *Fix the differential privacy level $\epsilon > 0$ and miscoverage level $\alpha \in (0.5, 1)$, as well as a free parameter $\gamma \in (0,1)$. Let*

$$\tilde{q} = \frac{(n+1)(1-\alpha)}{n(1-\gamma\alpha)} + \frac{2}{\epsilon n}\log\left(\frac{m}{\gamma\alpha}\right), \tag{3}$$

and let $\hat{s}$ be the output of Algorithm 2 at level $\min\{\tilde{q}, 1\}$. Then, the prediction sets in (2) with cutoff $\hat{s}$ satisfy the coverage property in (1).

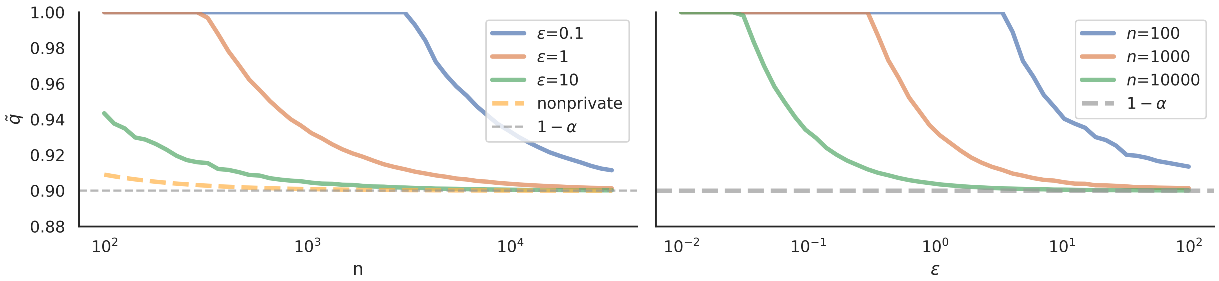

Figure 3: **The private quantile $\tilde{q}$ as n and ϵ grow.** We demonstrate the adjusted quantile from (3) as n and ϵ increase, with automatically chosen values for m and γ described in Appendix C. As the number of samples grows and the privacy constraint relaxes, the procedure chooses a less conservative quantile, eventually approaching the limiting value $1 - \alpha$. The mild fluctuations in the curves are due to differing choices of m^* and γ.

Remark 1. We can choose γ to minimize $\tilde{q}$, which leads to smallest prediction sets. The optimal value γ^* depends only on n, m, and α, and can be found by taking a derivative of (2); see Appendix C.

Note that the significance level $\tilde{q}$ in (3) is just a slightly inflated version of the nonprivate conformal quantile: $\tilde{q} \geq \frac{(n+1)(1-\alpha)}{n} \geq 1 - \alpha$. Indeed, taking $\epsilon \to \infty, \gamma \to 0$ in (2) recovers the nonprivate quantile. Intuitively, we must raise the significance level to compensate for the noise introduced to preserve privacy. We note that the additive factor of order $\frac{1}{n\epsilon}$ is in fact necessary to compute an approximate quantile with ϵ-differential privacy (Bun et al., 2017).

We informally sketch the main ideas in the proof, deferring the details to the Appendix.

Proof sketch. We can write the probability of coverage as:

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} = \mathbb{E}\left[F(\hat{s})\right],$$

where F is the distribution of appropriately discretized empirical scores. We observe that for all $\tilde{q}$, the exponential mechanism with input $\tilde{q}$ and $s_1, \dots, s_n$ returns an empirical quantile no smaller than the $\tilde{q} - O(1/(n\epsilon))$ empirical quantile. This allows us to write

$$\mathbb{E}\left[F(\hat{s})\right] \geq (1 - \gamma\alpha)\mathbb{E}\left[F(\hat{F}^{-1}(\tilde{q} - O(1/(n\epsilon)))\right],$$

where $\hat{F}$ denotes the empirical distribution of the discretized scores. For any q, the random variable $F(\hat{F}^{-1}(q))$ is distributed as the $\lceil nq \rceil$-th order statistic of a super-uniform distribution, which implies that it can be stochastically lower bounded by the $\lceil nq \rceil$-th order statistic of a uniform distribution. This order statistic follows a beta distribution with known parameters, whose expectation can hence be evaluated analytically. Carefully choosing $\tilde{q}$ as a function of this expectation completes the proof of the theorem. □

With the validity of Algorithm 3 established, we next prove that the algorithm is not too conservative in the sense that the coverage is not far above $1 - \alpha$. A key quantity in our upper bound is

$$p_{\max}^m := \max_{1 \leq j \leq m} \mathbb{P}\{s_1 \in I_j\}.$$

This quantity captures the impact of the score discretization. Smaller $p_{\max}^m$ corresponds to mass being spread more evenly throughout the bins. For well-behaved score functions, we expect $p_{\max}^m$ to scale as $O(m^{-1})$. Indeed, if the scores have any continuous density on $[0, 1]$ bounded above and we take uniformly spaced bins, then $p_{\max}^m = O(m^{-1})$. In terms of $p_{\max}^m$, we have the following upper bound.

Theorem 3 (Coverage upper bound)**.** *The prediction sets in (2) with $\hat{s}$ is as in Theorem 2, satisfy the following coverage upper bound:*

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} \leq 1 - (1-\gamma)\alpha + (1-\gamma\alpha)\left(2p_{\max}^m + \frac{2\left(1 + \max\left\{\frac{\tilde{q}}{1-\tilde{q}}, 1\right\}\right)\log(m/(\gamma\alpha))}{(n+1)\epsilon}\right),$$

where $\tilde{q}$ is defined in (3).

If we further assume a weak regularity condition on the scores, then by balancing the rates in the expression above we arrive at an explicit upper bound.

Corollary 1 (Coverage upper bound, simplified form)**.** *Suppose that the input scores follow a continuous distribution on* $[0, 1]$ *with a density that is bounded above. Take* $m \propto n\epsilon$ *and* $\gamma = 1/m$*. Then, the prediction sets in* (2)*, with* $\hat{s}$ *as in Theorem 2, satisfy the following upper bound:*

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} \leq 1 - \alpha + O\left(\frac{\log\left(n\epsilon/\alpha\right)}{n\epsilon}\right).$$

We emphasize that the assumptions on the score distribution are only needed to prove the upper bound; the coverage lower bound holds for any distribution. In any case, these assumptions are very weak, essentially requiring only that the score distribution contains no point masses. In fact, this requirement could even be enforced ex post facto by adding a small amount of tiebreaking noise, in which case we would need no restrictions on the input distribution of scores whatsoever.

The upper bound answers an important practical question: how many bins should we take? If m is too small, then the histogram only coarsely approximates the empirical distribution of the scores. On the other hand, if m is too large, then the histogram is accurate, but the private quantile in 3 can grow as well. This tension can be observed in the terms in Theorem 3 that have a dependence on m. Corollary 1 suggests that the correct balance—which leads to minimal excess coverage—is to take $m \propto n\epsilon$. In practice, because the dependence of $\tilde{q}$ on m is only logarithmic, m is often very large.

This upper bound also gives insight to an important theoretical question: what is the cost of privacy in conformal prediction? In nonprivate conformal prediction, the upper bound is $1 - \alpha + O(n^{-1})$ (J. Lei et al., 2018). In private conformal prediction, we achieve an upper bound of $1 - \alpha + \tilde{O}((n\epsilon)^{-1})$, a relatively modest cost incurred by privacy-preserving calibration.

4 Experiments

We now turn to an empirical evaluation of differentially private conformal prediction for image classification problems. In this setting, each image X_i has a single unique class label $Y_i \in \{1, ..., K\}$ estimated by a predictive model $\hat{f} : \mathcal{X} \to [0, 1]^K$. We seek to create private prediction sets, $\mathcal{C}(X_i) \subseteq \{1, ..., K\}$, achieving coverage as in equation (1), using the following score function:

$$S(x, y) = 1 - \hat{f}(x)_y,$$

as in Sadinle et al. (2019). This section evaluates the prediction sets generated by Algorithm 3 by quantifying the cost of privacy and the effects of the model, number of calibration points, and number of bins used in our procedure. We use the CIFAR-10 data set (Krizhevsky & Hinton, 2009) wherever we require a privately trained neural network. Otherwise, we use a nonprivate model on the ImageNet data set (Deng et al., 2009), to investigate the performance of our procedure in a more challenging setting with a large number of possible labels. Except where otherwise mentioned, we use an automated number of uniformly spaced bins m^* to construct the privatized CDF. Appendix C describes the algorithm for choosing an approximately optimal value of m^* when the conformal scores are roughly uniform based on fixed values of n, ϵ, and α. We finish the section by providing private prediction sets for diagnosing viral pneumonia on the CoronaHack data set (Perez et al., 2020). The reader can reproduce the experiments exactly using our public GitHub repository.

4.1 Isolating the effects of private model training and private conformal prediction

We would like to disentangle the effects of private conformal prediction from those of private model training. To that end, we report the coverage and set sizes of the following four procedures: private conformal prediction with a private model, nonprivate conformal prediction with a private model, private conformal

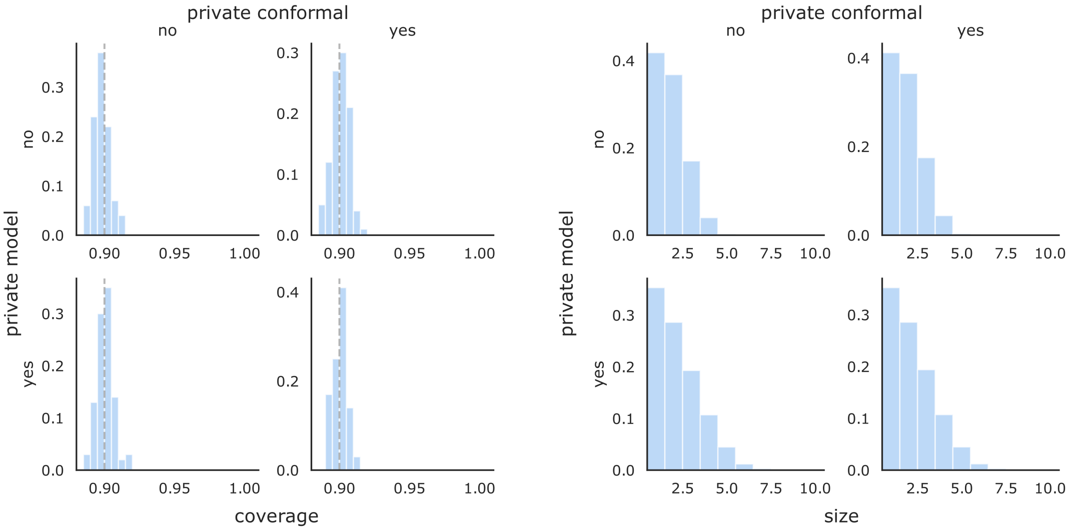

Figure 4: **Coverage and set size with private/nonprivate models and private/nonprivate conformal prediction.** We demonstrate histograms of coverage and set size of nonprivate/private models and nonprivate/private conformal prediction at the level $\alpha = 0.1$, with $\epsilon = 8$, $\delta = 1e - 5$, and $n = 5000$.

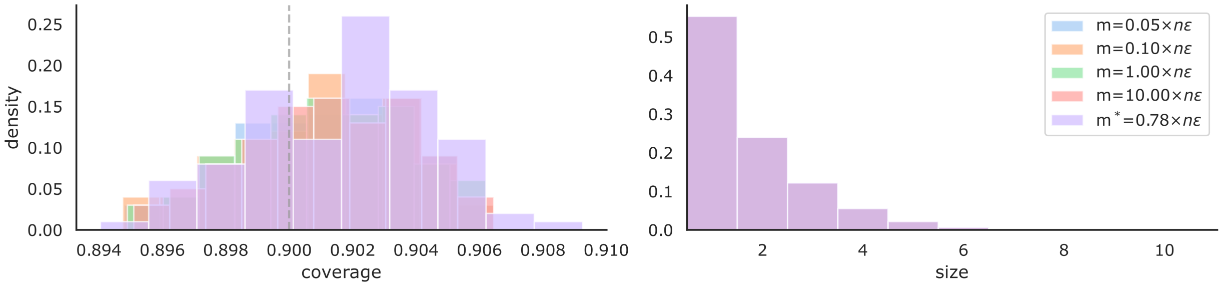

Figure 5: **Coverage and set size for different values of m.** We demonstrate the performance on Imagenet of private conformal prediction using a nonprivate ResNet-152 as the base model at $\alpha = 0.1$ and $\epsilon = 1$. The coverage is nearly constant over three orders of magnitude of bin numbers. All of the histograms on the right hand side are overlapping. See Section 4.2 for details.

prediction with a nonprivate model, and nonprivate conformal prediction with a nonprivate model. The nonprivate model and private model are both the same stock convolutional architecture from the `Opacus` library. The private model is trained with private SGD (Abadi et al., 2016), as implemented in the `Opacus` library, with privacy parameters $\epsilon = 8$ and $\delta = 1e - 5$. We used the suggested private model training parameters from the `Opacus` library (see Appendix C), as our work does not aim to improve private model training. The nonprivate model's accuracy (73%) was significantly higher than that of the private model (67%).

Figure 4 shows histograms of the coverages and set sizes of these procedures over 1,000 random splits of the CIFAR-10 validation set with $n = 5000$. Notably, the results show the price of private conformal prediction is very low, as evidenced by the minuscule increase in set size caused by private conformal prediction. However, the private model training causes a larger set size due to the private model's comparatively poor performance. Note that a user desiring a fully private pipeline will use the procedure in the bottom right quadrant of the plot.

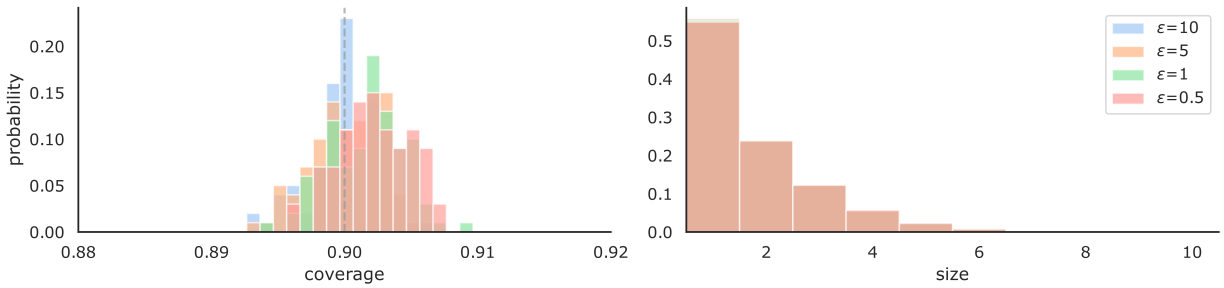

Figure 6: **Coverage and set size for different values of ϵ.** We demonstrate the performance on ImageNet of private conformal prediction using a nonprivate ResNet-152 as the base model with $\alpha = 0.1$. The coverage improves slightly for liberal (large) ϵ, although the cost of privacy is evidently very low. All of the histograms on the right hand side are overlapping. See Section 4.3 for details.

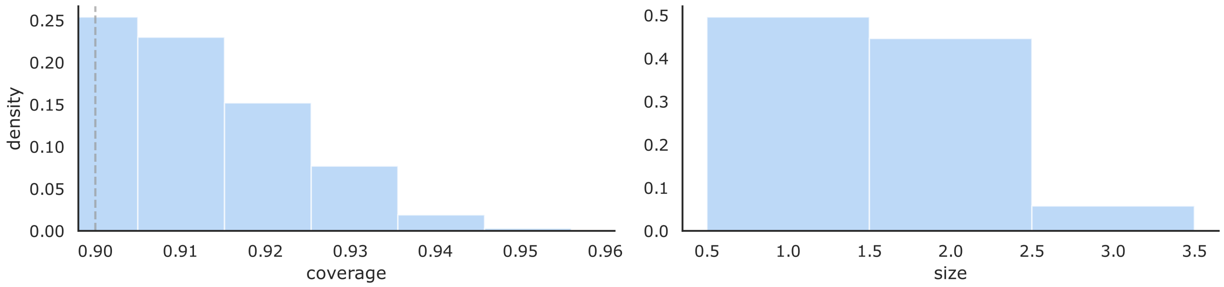

Figure 7: **Coverage and set size on the CoronaHack data set.** We demonstrate the performance on the CoronaHack data set of private conformal prediction using a nonprivate ResNet-18 as the base model with $\alpha = 0.1$. The median coverage was 90.4%. See Section 4.4 for details.

4.2 Varying number of bins m

Here we probe the performance of private prediction sets as the number of uniformly spaced bins m in our procedure changes. Based on our theoretical results, m should be on the order of $n\epsilon$, with the exact number dependent on the underlying model and the choices of α, n, and ϵ. A too-small choice of m coarsely quantizes the scores, so Algorithm 4 may be forced to round up to a very conservative private quantile. A too-large choice of m increases the logarithmic term in 3. The optimal choice of m balances these two factors.

To demonstrate this tradeoff, we performed experiments on ImageNet. We used a nonprivate, pretrained ResNet-152 from the `torchvision` repository as the base model. Figure 5 shows the coverage and set size of private prediction sets over 100 random splits of ImageNet's validation set for several choices of m; we used $n = 30000$ and evaluated on the remaining 20000 images. The experimental results suggest m^* works comparatively well, and that our method is relatively insensitive to the number of bins over several orders of magnitude.

4.3 Varying privacy level ϵ

Next we quantify how the coverage changes with the privacy parameter ϵ. We used $n = 30000$ calibration points and 20000 evaluation points as in Experiment 4.3. For each value of ϵ we choose a different value of m^*. Figure 6 shows the coverage and set size of private prediction sets over 100 splits of ImageNet's validation set for several choices of ϵ. As ϵ grows, the procedure becomes less conservative. Overall the procedure exhibits little sensitivity to ϵ.

4.4 COVID-19 diagnosis

Next we show results on the CoronaHack data set, a public chest X-ray data set containing 5908 X-rays labeled as `normal`, `viral pneumonia` (primarily COVID-19), or `bacterial pneumonia`. Using 4408 training pairs over 14 epochs, we (nonprivately) fine-tuned the last layer of a pretrained ResNet-18 from `torchvision` to predict one of the three diagnoses. The private conformal calibration procedure saw a further $n = 1000$ examples, and we used the remaining 500 for validation. The ResNet-18 had a final accuracy of 75% after fine-tuning. Figure 7 plots the coverage and set size of this procedure over 1000 different train/calibration/validation splits of the dataset, and Figure 2 shows selected examples of these sets.

5 Discussion

We introduce a method to produce differentially private prediction sets that contain the true response with a user-specified probability by blending split conformal prediction with differentially private quantile computation. The primary challenge we resolve in this work is simultaneously satisfying the coverage property and privacy property, which requires a careful choice of the conformal score threshold to account for the added privacy noise. Our corresponding upper bound shows that the coverage does not greatly exceed the nominal level $1 - \alpha$, meaning that our procedure is not too conservative. Moreover, our upper bound gives insight into the price of privacy in conformal prediction: the upper bound scales as $\tilde{O}((n\epsilon)^{-1})$ compared to $O(n^{-1})$ for nonprivate conformal prediction, a mild decrease in efficiency. This is confirmed in our experiments, where we show that there is little difference between private and nonprivate conformal prediction when using the same predictive model. We also observe the familiar phenomenon that there is a substantial decrease in accuracy for private model fitting compared to nonprivate model fitting. We conclude that the cost of privacy lies primarily in the model fitting—private calibration has a comparatively minor effect on performance. We also note that any improvement in private model training would immediately translate to smaller prediction sets returned by our method. In sum, we view private conformal prediction as an appealing method for uncertainty quantification with differentially private models.

References

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

Abowd, J. M. (2018). The US census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2867–2867).

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*.

Angelopoulos, A. N., Bates, S., Malik, J., & Jordan, M. I. (2020). Uncertainty sets for image classifiers using conformal prediction. *arXiv:2009.14193*.

Barber, R. F., Candes, E. J., Ramdas, A., Tibshirani, R. J., et al. (2021). Predictive inference with the jackknife+. *Annals of Statistics*, *49*(1), 486–507.

Bassily, R., Smith, A., & Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science* (pp. 464–473).

Bates, S., Angelopoulos, A., Lei, L., Malik, J., & Jordan, M. I. (2021). Distribution-free, risk-controlling prediction sets. *arXiv:2101.02703*.

Bittau, A., Erlingsson, Ú., Maniatis, P., Mironov, I., Raghunathan, A., Lie, D., ... Seefeld, B. (2017). Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th symposium on operating systems principles* (pp. 441–459).

Bun, M., Steinke, T., & Ullman, J. (2017). Make up your mind: The price of online queries in differential privacy. In *Proceedings of the twenty-eighth annual ACM-SIAM symposium on discrete algorithms* (pp. 1306–1325).

Cauchois, M., Gupta, S., Ali, A., & Duchi, J. C. (2020). Robust validation: Confident predictions even when distributions shift. *arXiv:2008.04267*.

Cauchois, M., Gupta, S., & Duchi, J. (2020). Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *arXiv:2004.10181*.

Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, *12*(3).

del Coz, J. J., Díez, J., & Bahamonde, A. (2009). Learning nondeterministic classifiers. *Journal of Machine Learning Research*, *10*(79), 2273-2293. Retrieved from http://jmlr.org/papers/v10/delcoz09a.html

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Differential Privacy Team Apple. (2017). Learning with privacy at scale. In *Apple machine learning research.*

Ding, B., Kulkarni, J., & Yekhanin, S. (2017). Collecting telemetry data privately. In *Advances in neural information processing systems* (pp. 3571–3580).

Dwork, C. (2019). Differential privacy and the US census. In *Proceedings of the 38th acm sigmod-sigact-sigai symposium on principles of database systems* (pp. 1–1).

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

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

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

Feldman, V., & Steinke, T. (2017). Generalization for adaptively-chosen estimators via stable median. In *Conference on learning theory* (pp. 728–757).

Foygel Barber, R., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, *10*, 455–482.

Gaboardi, M., Rogers, R., & Sheffet, O. (2019). Locally private mean estimation: z-test and tight confidence intervals. In *The 22nd international conference on artificial intelligence and statistics* (pp. 2545–2554).

Grycko, E. (1993). Classification with set-valued decision functions. In *Information and classification* (pp. 218–224).

Guan, L. (2020). Conformal prediction with localization. *arXiv:1908.08558*.

Guan, L., & Tibshirani, R. (2019). Prediction and outlier detection in classification problems. *arXiv:1905.04396*.

Gupta, V., Jung, C., Noarov, G., Pai, M. M., & Roth, A. (2021). Online multivalid learning: Means, moments, and prediction intervals. *arXiv:2101.01739*.

Hechtlinger, Y., Poczos, B., & Wasserman, L. (2018). Cautious deep learning. *arXiv:1805.09460*.

Hu, X., & Lei, J. (2020). A distribution-free test of covariate shift using conformal prediction. *arXiv:2010.07147*.

Izbicki, R., Shimizu, G. T., & Stern, R. B. (2019). Flexible distribution-free conditional predictive bands using density estimators. *arXiv:1910.05575*.

Jung, C., Lee, C., Pai, M. M., Roth, A., & Vohra, R. (2020). Moment multicalibration for uncertainty estimation. *arXiv:2008.08037*.

Karwa, V., & Vadhan, S. (2017). Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*.

Krishnamoorthy, K., & Mathew, T. (2009). *Statistical tolerance regions: Theory, applications, and computation*. Wiley. Retrieved from https://books.google.com/books?id=1jQh0miU6PQC

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Lei, J. (2011). Differentially private m-estimators. In *Advances in neural information processing systems* (pp. 361–369).

Lei, J. (2014, 10). Classification with confidence. *Biometrika*, *101*(4), 755-769.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111.

Lei, J., Rinaldo, A., & Wasserman, L. (2015). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, *74*, 29-43.

Lei, L., & Candès, E. J. (2020). Conformal inference of counterfactuals and individual treatment effects. *arXiv:2006.06138*.

McSherry, F., & Talwar, K. (2007). Mechanism design via differential privacy. In *48th annual IEEE symposium on foundations of computer science (FOCS)* (pp. 94–103).

Mortier, T., Wydmuch, M., Dembczyński, K., Hüllermeier, E., & Waegeman, W. (2020). Efficient set-valued prediction in multi-class classification. *arXiv:1906.08129*.

Neel, S., Roth, A., Vietri, G., & Wu, S. (2020). Oracle efficient private non-convex optimization. In *International conference on machine learning* (pp. 7243–7252).

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine Learning: European Conference on Machine Learning* (pp. 345–356).

Park, S., Bastani, O., Matni, N., & Lee, I. (2020). PAC confidence sets for deep neural networks via calibrated prediction. In *International conference on learning representations*. Retrieved from https://openreview.net/forum?id=BJxVI04YvB

Perez, J. C., de Blas Perez, C., Alvarez, F. L., & Contreras, J. M. C. (2020). Databiology Lab CORONAHACK: Collection of public COVID-19 data. *bioRxiv*.

Romano, Y., Patterson, E., & Candès, E. (2019). Conformalized quantile regression. In *Advances in neural information processing systems* (pp. 3543–3553).

Romano, Y., Sesia, M., & Candès, E. J. (2020). Classification with valid and adaptive coverage. *arXiv:2006.02544*.

Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, *114*, 223 - 234.

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, *9*(Mar), 371–421.

Sheffet, O. (2017). Differentially private ordinary least squares. In *International conference on machine learning* (pp. 3105–3114).

Smith, A. (2011). Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on theory of computing* (pp. 813–822).

Tibshirani, R. J., Foygel Barber, R., Candes, E., & Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in neural information processing systems* (pp. 2530–2540).

Tukey, J. W. (1947). Non-parametric estimation II. statistically equivalent blocks and tolerance regions—the continuous case. *Annals of Mathematical Statistics*, *18*(4), 529–539.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Proceedings of the asian conference on machine learning* (Vol. 25, pp. 475–490).

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, *74*(1-2), 9–28.

Vovk, V., Gammerman, A., & Saunders, C. (1999). Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning* (pp. 444–453).

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer. doi: 10.1007/b106715.

Vovk, V., Petej, I., Toccaceli, P., Gammerman, A., Ahlberg, E., & Carlsson, L. (2020). Conformal calibrators. In *Conformal and probabilistic prediction and applications* (pp. 84–99).

Vovk, V., Shen, J., Manokhin, V., & Xie, M.-g. (2017). Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 1–30.

Wald, A. (1943). An extension of Wilks' method for setting tolerance limits. *Annals of Mathematical Statistics*, *14*(1), 45–55.

Wang, Y., Kifer, D., & Lee, J. (2019). Differentially private confidence intervals for empirical risk minimization. *Journal of Privacy and Confidentiality*, *9*(1).

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics*, *12*(1), 91–96.

Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Annals of Mathematical Statistics*, *13*(4), 400–409. Retrieved from https://doi.org/10.1214/aoms/1177731537 doi: 10.1214/aoms/1177731537

Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., & Winslett, M. (2013). Differentially private histogram publication. *The VLDB Journal*, *22*(6), 797–822.

A Auxiliary results

We start with a result about the error of the private quantile mechanism, stated in Algorithm 2. The following is an extension of the the analogous result for the private median due to Feldman & Steinke (2017).

Lemma 1. *For any $\delta \in (0,1)$, the differentially private quantile algorithm (Algorithm 2) satisfies:*

$$\mathbb{P}\left\{\frac{1}{n}\#\{i : [s_i] \leq \hat{s}\} \geq q - \frac{2\max\left\{\frac{1-q}{q}, 1\right\}\log(m/\delta)}{n\epsilon}\right\} \geq 1 - \delta$$

and

$$\mathbb{P}\left\{\frac{1}{n}\#\{i : [s_i] < \hat{s}\} \leq q + \frac{2\max\left\{\frac{q}{1-q}, 1\right\}\log(m/\delta)}{n\epsilon}\right\} \geq 1 - \delta.$$

Proof. By the standard utility guarantee for the exponential mechanism (McSherry & Talwar, 2007) (e.g., Corollary 3.12 in (Dwork & Roth, 2014)), we have

$$\mathbb{P}\left\{\max\left\{\frac{\#\{j : [s_j] < \hat{s}\}}{q}, \frac{\#\{j : [s_j] > \hat{s}\}}{1-q}\right\} < \min_i w_i + \frac{2\max\{1/q, 1/(1-q)\}\log(m/\delta)}{\epsilon}\right\} \geq 1 - \delta. \tag{4}$$

First we argue that $\min_i w_i \leq n$. Let $s^* = \min\{s \in \{e_0, \ldots, e_m\} : \#\{i : [s_i] \leq s\} > qn\}$. Then, $\#\{i : [s_i] > s^*\} < (1-q)n$ trivially. Furthermore, $\#\{i : [s_i] < s^*\} \leq qn$ by the definition of s^*, since s^* is the *first* point at which the cumulative fraction of scores less than or equal to s^* exceeds q. Since we have identified a bin where $\max\left\{\frac{\#\{j:[s_j]<s^*\}}{q}, \frac{\#\{j:[s_j]>s^*\}}{1-q}\right\} \leq n$, we can conclude that $\min_i w_i \leq n$.

Going back to Equation (4), we have that with probability at least $1-\delta$,

$$\begin{aligned}\frac{1}{n}\#\{i : s_i \leq \hat{s}\} &= 1 - \frac{1}{n}\#\{i : s_i > s\} \\ &\geq 1 - (1-q)\frac{\min_i w_i}{n} - \frac{2\max\{(1-q)/q, 1\}\log(m/\delta)}{n\epsilon} \\ &\geq q - \frac{2\max\{(1-q)/q, 1\}\log(m/\delta)}{n\epsilon}.\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{1}{n}\#\{i : s_i < \hat{s}\} &\leq q\frac{\min_i w_i}{n} + \frac{2\max\{1, q/(1-q)\}\log(m/\delta)}{n\epsilon} \\ &\leq q + \frac{2\max\{1, q/(1-q)\}\log(m/\delta)}{n\epsilon}.\end{aligned}$$

□

Next, we package some classical facts about the distribution of order statistics in a form helpful for analyzing conformal prediction.

Lemma 2. *Let F be the CDF of a distribution supported on a finite set $\{a_1, \ldots, a_m\}$. Let $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} F$, and let $\hat{F}$ denote the empirical CDF corresponding to $Z_1, \ldots, Z_n$. Denote also $p_{\max}^m = \max_{1 \leq i \leq m} \mathbb{P}\{Z_1 = a_i\}$. Then,*

$$Z_{\text{Beta}} + p_{\max}^m \succeq F(\hat{F}^{-1}(q)) \succeq Z_{\text{Beta}},$$

where Z_{Beta} follows the beta distribution $\text{Beta}(\lceil nq \rceil, n - \lceil nq \rceil + 1)$ and $\succeq$ denotes first-order stochastic dominance.

Proof. Since we take $\hat{F}^{-1}(q) = \inf\{z : \hat{F}(z) \geq q\}$ by definition, then that implies $\hat{F}^{-1}(q) = Z_{(\lceil nq \rceil)}$, where $Z_{(i)}$ denotes the i-th nondecreasing order statistic of $Z_1, \ldots, Z_n$. By monotonicity of F, we further have that $F(Z_{(\lceil nq \rceil)})$ is identical to the $\lceil nq \rceil$-th nondecreasing order statistic of $F(Z_1), \ldots, F(Z_n)$. By a standard argument, the samples $F(Z_1), \ldots, F(Z_n)$ are super-uniform, that is, $\mathbb{P}\{F(Z_1) \leq u\} \leq u$ for all $u \in [0,1]$. In other words, they are stochastically larger than a uniform distribution on $[0,1]$, and thus their $\lceil nq \rceil$-th order statistic is stochastically lower bounded by the $\lceil nq \rceil$-th order statistic of a uniform distribution, which follows the $\text{Beta}(\lceil n\alpha \rceil, n - \lceil n\alpha \rceil + 1)$ distribution. This completes the proof of the lower bound. For the upper bound, we use the fact that $\mathbb{P}\{F(Z_1) \leq u\} \geq u - p_{\max}^m$, and so $F(Z_i)$ are stochastically dominated by $U_i + p_{\max}^m$, where $\{U_i\}_{i=1}^n$ are i.i.d. uniform on $[0,1]$. Their $\lceil nq \rceil$-th order statistic is distributed as $Z_{\text{Beta}} + p_{\max}^m$, which completes the proof. □

B Proofs

B.1 Proof of Theorem 2

First we introduce some notation. By F we will denote the discretized CDF of the scores; in particular, for any $i \in \{1, \ldots, n\}$,

$$F(s) = \mathbb{P}\{[s_i] \leq s\}.$$

Here, by $[s_i]$ we denote a *discretized* version of s_i where we set $[s_i] = e_j$ if $s_i \in I_j$. We also let $\hat{F}$ denote the empirical distribution of the discretized scores:

$$\hat{F}(s) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{[s_i] \leq s\}.$$

By convention, we let $F^{-1}(\delta)$ denote the left-continuous inverse of F, i.e. $F^{-1}(\delta) := \inf\{s : F(s) \geq \delta\}$, and we similarly define $\hat{F}^{-1}(\delta)$.

We can write

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} = \mathbb{P}\{S(X,Y) \leq \hat{s}\} = \mathbb{E}\left[F(\hat{s})\right].$$

Denote the event $E = \left\{\frac{1}{n}\#\{i : [s_i] \leq \hat{s}\} \geq \tilde{q} - \frac{2}{\epsilon n}\log(m/(\gamma\alpha))\right\}$, and note that by Lemma 1 and the fact that $\tilde{q} \geq 0.5$, $\mathbb{P}\{E\} \geq 1 - \gamma\alpha$. By splitting up the analysis depending on E, we obtain the following:

$$\begin{aligned}
\mathbb{E}\left[F(\hat{s})\right] &= \mathbb{E}\left[F(\hat{s})\mathbf{1}\{E^c\}\right] + \mathbb{E}\left[F(\hat{s})\mathbf{1}\{E\}\right] \\
&\geq \gamma\alpha \cdot 0 + \mathbb{E}\left[F(\hat{s})\mathbf{1}\{E\}\right] \\
&\geq (1-\gamma\alpha)\mathbb{E}\left[F\left(\hat{F}^{-1}\left(\tilde{q} - \frac{2}{\epsilon n}\log\left(m/(\gamma\alpha)\right)\right)\right)\right],
\end{aligned}$$

where the final inequality follows by the definition of E. Thus, it suffices to show that

$$\mathbb{E}\left[F\left(\hat{F}^{-1}\left(\tilde{q} - \frac{2}{\epsilon n}\log\left(m/(\gamma\alpha)\right)\right)\right)\right] \geq \frac{1-\alpha}{1-\gamma\alpha}. \tag{5}$$

Let $j^* = \left\lceil n\left(\tilde{q} - \frac{2}{\epsilon n}\log\left(m/(\gamma\alpha)\right)\right)\right\rceil$. Then, by Lemma 2,

$$F\left(\hat{F}^{-1}\left(\tilde{q} - \frac{2}{\epsilon n}\log\left(m/(\gamma\alpha)\right)\right)\right) \succeq \textsc{Beta}(j^*, n - j^* + 1),$$

so

$$\mathbb{E}\left[F\left(\hat{F}^{-1}\left(\tilde{q} - \frac{2}{\epsilon n}\log\left(m/(\gamma\alpha)\right)\right)\right)\right] \geq \frac{j^*}{n+1} = \frac{\lceil n(\tilde{q} - \frac{2}{\epsilon n}\log\left(m/(\gamma\alpha)\right))\rceil}{n+1}.$$

By the definition of $\tilde{q}$, we see that

$$\frac{\lceil n(\tilde{q} - \frac{2}{\epsilon n}\log\left(m/(\gamma\alpha)\right))\rceil}{n+1} \geq \frac{1-\alpha}{1-\gamma\alpha},$$

holds, which implies Equation (5) and thus completes the proof.

B.2 Proof of Theorem 3

We adopt the definitions of F, $\hat{F}$ from Theorem 2, and define E as the event

$$\left\{\frac{1}{n}\#\{i : [s_i] < \hat{s}\} \leq \tilde{q} + \frac{2\max\{\frac{\tilde{q}}{1-\tilde{q}}, 1\}\log(m/\gamma\alpha)}{n\epsilon}\right\},$$

which by Lemma 1 has probability at least $1 - \gamma\alpha$. By a similar reasoning as in Theorem 2, we obtain the following:

$$\begin{aligned}
\mathbb{E}\left[F(\hat{s})\right] &= \mathbb{E}\left[F(\hat{s})\mathbf{1}\{E^c\}\right] + \mathbb{E}\left[F(\hat{s})\mathbf{1}\{E\}\right] \\
&\leq \gamma\alpha \cdot 1 + \mathbb{E}\left[F(\hat{s})\mathbf{1}\{E\}\right] \\
&\leq \gamma\alpha + (1-\gamma\alpha)\left(\mathbb{E}\left[F\left(\hat{F}^{-1}\left(\tilde{q} + \frac{2\max\{\frac{\tilde{q}}{1-\tilde{q}}, 1\}}{\epsilon n}\log\left(m/(\gamma\alpha)\right)\right)\right)\right] + p^m_{\max}\right),
\end{aligned} \tag{6}$$

where the final inequality follows by the definition of E.

Let $j^* = \lceil n \left(\tilde{q} + \frac{2\max\{\frac{\tilde{q}}{1-\tilde{q}},1\}}{\epsilon n} \log\left(m/(\gamma\alpha)\right) \right) \rceil$. By Lemma 2, we have

$$F\left(\hat{F}^{-1}\left(\tilde{q} + \frac{2\max\{\frac{\tilde{q}}{1-\tilde{q}},1\}}{\epsilon n} \log\left(m/(\gamma\alpha)\right)\right)\right) \preceq \textsc{Beta}(j^*, n - j^* + 1) + p^m_{\max},$$

so

$$\mathbb{E}\left[F\left(\hat{F}^{-1}\left(\tilde{q} + \frac{2\max\{\frac{\tilde{q}}{1-\tilde{q}},1\}}{\epsilon n} \log\left(m/(\gamma\alpha)\right)\right)\right)\right] \leq \frac{j^*}{n+1} + p^m_{\max} = \frac{\lceil n \left(\tilde{q} + \frac{2\max\{\frac{\tilde{q}}{1-\tilde{q}},1\}}{\epsilon n} \log\left(m/(\gamma\alpha)\right) \right) \rceil}{n+1} + p^m_{\max}. \tag{7}$$

By the definition of $\tilde{q}$, we see that

$$\begin{aligned}\frac{\lceil n \left(\tilde{q} + \frac{2\max\{\frac{\tilde{q}}{1-\tilde{q}},1\}}{\epsilon n} \log\left(m/(\gamma\alpha)\right) \right) \rceil}{n+1} &\leq \frac{\frac{1-\alpha}{1-\gamma\alpha}(n+1) + \frac{2(1+\max\{\frac{\tilde{q}}{1-\tilde{q}},1\})}{\epsilon} \log\left(m/(\gamma\alpha)\right)}{n+1} \\ &= \frac{1-\alpha}{1-\gamma\alpha} + \frac{2(1+\max\{\frac{\tilde{q}}{1-\tilde{q}},1\})\log(m/(\gamma\alpha))}{(n+1)\epsilon}.\end{aligned} \tag{8}$$

Putting together Equations (6), (7), and (8) completes the proof.

C Experimental details

Choosing m^* and γ. Algorithm 5 gives automatic choices of the optimal number of uniformly spaced bins, m^*, and the tuning parameter γ that work well for approximately uniformly distributed scores. In a moment, we will show how to find the optimal value γ^* for a fixed value of m. Once γ^* gets chosen, we will simulate uniformly distributed scores to choose the value m^* that results in the best quantile for specific, pre-determined values of α, ϵ, and n. In practice, m^* can be chosen from a relatively coarse grid of, say, 50 values logarithmically spaced from 10^2 to 10^6.

We start choosing the optimal value γ^* by solving for the zeros of the derivative $\frac{\delta\tilde{q}}{\delta\gamma}$, leading to the quadratic expression,

$$\frac{\delta\tilde{q}}{\delta\gamma} = 0 \iff \alpha^2\gamma^2 - \left(\frac{\alpha(1-\alpha)\epsilon(n+1)}{2} + 2\alpha\right)\gamma + 1 = 0. \tag{9}$$

Letting Γ be the roots of (9), we can then choose the optimal value γ^* as

$$\gamma^* = \underset{\gamma\in\Gamma\cap(0,1)\cup\{1e-12\}}{\arg\min} \left[\frac{(n+1)(1-\alpha)}{n(1-\gamma\alpha)} + \frac{2}{\epsilon n}\log\left(\frac{m}{\gamma\alpha}\right)\right], \tag{10}$$

where the number 1e-12 takes care of the case that both roots lie outside the interval $(0,1)$.

Algorithm 5 Get optimal number of bins and γ

```
input: number of calibration points $n$, privacy level $\epsilon > 0$, confidence level $\alpha \in (0,1)$
  Simulate $n$ uniform conformity scores $s_i \sim \mathrm{Unif}(0,1), i = 1,...,n$
Choose $m^*$ to be the value of $m$ minimizing the output of Algorithm 3 on the $s_i$ with the optimal $\gamma^*$ chosen by (10).
output: $m^*$, $\gamma^*$
```

Private training procedure. We used the `Opacus` library with the default parameter choices included in the CIFAR-10 example code. The only difference in the nonprivate model training is the use of the `--disable-dp` flag, turning off the added noise but preserving all other settings.