Private Prediction Sets

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

University of California, Berkeley

February 12, 2021

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 pre-trained 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 with experiments on the CIFAR-10, ImageNet, and CoronaHack datasets.

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 dataset at little computational cost.

Our approach builds on the notion of *prediction sets*—subsets of the response space that provably cover the true response variable with pre-specified 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 $1-\alpha \in (0,1)$. We use the output of an underlying predictive model (e.g., a pre-trained, privatized neural network) along with a held-out *calibration dataset*, $\{(X_i, Y_i)\}_{i=1}^n$, 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

[*] equal contribution

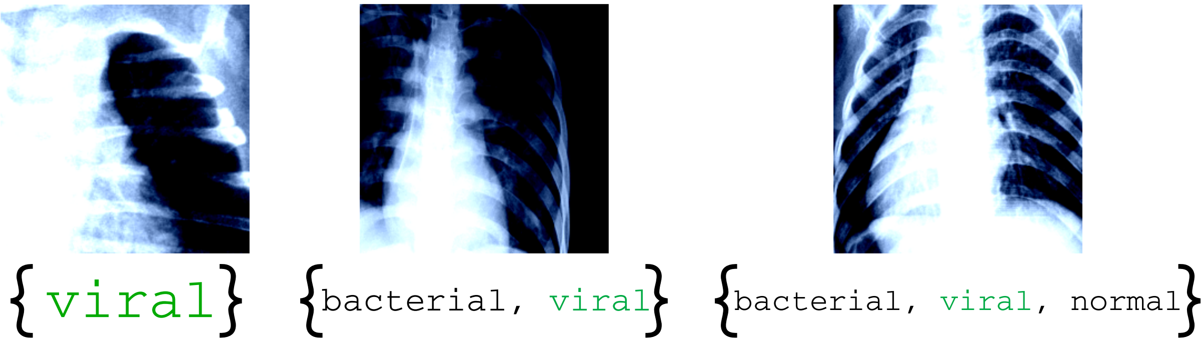

Figure 1: **Examples of private conformal prediction sets on COVID-19 data.** We show three examples of lung X-rays taken from the CoronaHack dataset [1] with their corresponding private prediction sets at $\alpha = 10\%$ from a ResNet-18. All three patients had `viral pneumonia` (likely COVID-19). The 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.

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 dataset with additional randomness, addressing both privacy and uncertainty simultaneously.

See Figure 1 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 which takes as input any predictive model together with a calibration dataset, 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 Eq. (1). In order to generate prediction sets satisfying this property, we will use ideas from split conformal prediction [2, 3, 4], 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 empirical CDF of $\{s_i\}_{i=1}^n$, $\tilde{F}(s)$
Compute $\hat{s} = \inf \left\{ s : \tilde{F}(s) \geq 1 - \alpha + O\left(\frac{\sqrt{\log(1/\alpha)}}{(n\epsilon)^{2/3}}\right) \right\}$
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 differentially private empirical CDF of the computed scores, and finds a critical threshold $\hat{s}$ which roughly corresponds to the $1-\alpha$ quantile of the private CDF, corrected for noise due to privacy. 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*

$$1-\alpha \leq \mathbb{P}\{Y \in \mathcal{C}(X)\} \leq 1-\alpha+O((n\epsilon)^{-2/3}).$$

We obtain a gap between the lower and upper bound on the probability of coverage to be roughly of the order $O((n\epsilon)^{-2/3})$, in contrast with 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 [5] 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 [6, 7], Apple [8], Microsoft [9], and the US Census Bureau [10, 11]. This increasing adoption of differential privacy goes hand in hand with steady progress in differentially private model training, ranging across both convex [12, 13] and non-convex [14, 15] 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 [5, 16, 17, 18]. Indeed, our work builds on work on private histogram computation by Dwork et al. [5]. Finally, there have also been significant efforts to quantify uncertainty with formal privacy guarantees through various types of private confidence intervals [19, 20, 21, 22]. 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 [23, 24, 25, 26]. See Krishnamoorthy & Mathew [27] for an overview of tolerance regions and Park et al. [28] for a recent application to deep learning models. Conformal prediction [29, 3, 30] is a related way of producing predictive sets with finite-sample guarantees. Most relevant to the present work, *split conformal prediction* [2, 31, 4] is a convenient version that uses data splitting to give prediction sets in a computationally efficient way. Vovk [32] and Barber et al. [33] refine this approach to re-use data for both training and calibration, improving statistical efficiency. Recent work has targeted desiderata such as small set sizes [34, 35], coverage that is approximately balanced across feature space [36, 37, 38, 39, 40, 41, 42], and coverage that is balanced across classes [43, 34, 44, 45]. Further extensions address problems in distribution estimation [46, 47], handling or testing distribution shift [48, 49, 50], causal inference [51], and controlling other notions of statistical error [52]. Lastly, we highlight two alternative approaches with a similar goal to conformal prediction. First, the calibration technique in Jung et al. [53] and Gupta et al. [54] generates prediction sets via the estimation of higher moments across many overlapping sub-populations. 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 [55, 56, 57]. 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 dataset* 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 non-private conformal prediction, one simply takes $\hat{s}$ to be the $\lceil (n+1)(1-\alpha) \rceil / n$ 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 Eq. (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 datasets $\mathcal{D}, \mathcal{D}' \in (\mathcal{X} \times \mathcal{Y})^n$ are *neighboring* if they differ in a single element, i.e., either dataset can be obtained from the other by replacing a single entry. Differential privacy then requires that two neighboring datasets produce similar distributions on the output.

Definition 1 (Differential privacy [5])**.** A randomized algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Z}$ is ϵ-*differentially private* if for all neighboring datasets $\mathcal{D}$ 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 dataset $\mathcal{D}'$ with the i-th entry replaced, 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 empirical distribution of the conformity scores. We obtain this from a privatized histogram where we bin the scores and add Laplace noise to each bin [5]. Specifically, let us fix a number of histogram bins $m \in \mathbb{N}$, as well as the histogram edges $0 \equiv e_0 < e_1 < ... < e_{m-1} < e_m \equiv 1$. The edges define the histogram bins $I_j = (e_{j-1}, e_j]$, $j = 1, ..., m$. With these bins, let $n_j = \big|\{i : s_i \in I_j\}\big|$ be the number of scores observed in bin j. To form a private empirical CDF, we return noisy versions of n_j. Algorithm 2 explicitly states a differentially private CDF algorithm due to Dwork et al. [5]. We use Algorithm 2 as a subroutine of our main conformal procedure.

Algorithm 2 Private empirical CDF [5]

```
input: calibration scores $\{s_1, \ldots, s_n\}$, bins $\{I_1, \ldots, I_m\}$, privacy level $\epsilon$
For all $1 \leq j \leq m$, compute bin count $n_j = |\{i : s_i \in I_j\}|$
For all $1 \leq j \leq m$, compute noisy bin count $\tilde{n}_j = n_j + \zeta_j$, $\zeta_j \sim$ Laplace $\left(\frac{2}{\epsilon}\right)$
output: private empirical CDF $\tilde{F}(s) = 1 - \frac{1}{n} \sum_{j=1}^m \tilde{n}_j \mathbf{1}\{s < e_j\}$
```

We will refer to the function $\tilde{F}$ output from Algorithm 2 as the *private empirical CDF*, but note that it is not exactly a CDF for two reasons. First, the added Laplace noise could make a bin count negative, in which case $\tilde{F}$ would fail to be non-decreasing. Second, $\tilde{F}(0)$ may not be zero (it could be either greater than or less than zero) due to the added noise. Nonetheless, this nuance will not impact our development, so the reader can safely think of this as a private CDF throughout this work.

3 Main 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 Eq. (2), which for completeness we state in Algorithm 4.

Algorithm 3 Differentially private calibration

```
input: calibration scores $\{s_1, \dots, s_n\}$, privacy parameter $\epsilon$, number of bins $m$, tuning parameter $\gamma$
Compute private histogram $\tilde{F}(\cdot)$ via Algorithm 2
Compute privacy noise adjustment $\zeta_{\max}^{n,m}(\gamma\alpha, \epsilon)$ in (3) by simulation  (or use the analytic upper bound in (6))
Compute the score cutoff $\hat{s}$ as in (4)
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): $\mathcal{C}(x) = \{y : S(x, y) \leq \hat{s}\}$.
```

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 histogram algorithm due to [5], followed by the closure-under-post-processing property.

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 α, as well as a free parameter $\gamma \in (0, 1)$. For any $\nu \in (0, 1)$, define*

$$\zeta_{\max}^{n,m}(\nu, \epsilon) := \min\left\{ t > 0 : \mathbb{P}\left\{ \max_{k \leq m} \left| \frac{1}{n} \sum_{j=1}^{k} \zeta_j \right| \geq t \right\} \leq \nu \right\}, \text{ where } \zeta_j \overset{i.i.d.}{\sim} \mathrm{Lap}\left(\frac{2}{\epsilon}\right). \tag{3}$$

Let

$$\hat{s} = \inf\left\{ z : \tilde{F}(z) \geq \frac{(n+1)(1-\alpha)}{n(1-\gamma\alpha)} + \zeta_{\max}^{n,m}(\gamma\alpha, \epsilon) \right\}, \tag{4}$$

Then, the prediction sets in (2) *with this choice of $\hat{s}$ satisfy the coverage property in* (1)*.*

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

Proof sketch. Suppose that $\hat{s} = \inf\{z : \tilde{F}(z) \geq \hat{q}\}$ for some $\hat{q}$. Then, we can write the probability of coverage as:

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} = \mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q})) \right], \tag{5}$$

where F is the distribution of the scores, appropriately discretized according to the histogram bins. We observe that uniformly across the domain, the empirical distribution of the discretized scores $\hat{F}$ is close to

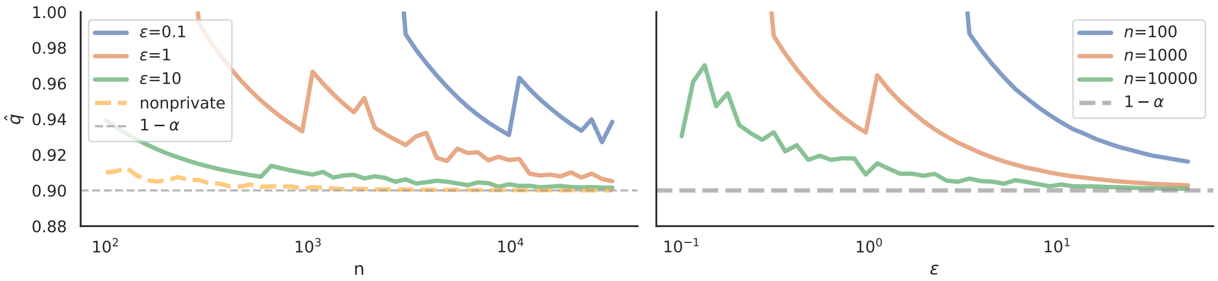

Figure 2: **The quantile of $\tilde{F}$ as n and ϵ grow.** We demonstrate the adjusted quantile, $\hat{q} = \frac{(n+1)(1-\alpha)}{n(1-\gamma\alpha)} + \zeta_{\max}^{n,m}(\gamma\alpha, \epsilon)$, as n and ϵ increase, with an automatically chosen values for m and γ described in Appendix C.1. As the number of samples grows and the privacy constraint relaxes, the procedure chooses a less conservative quantile of $\tilde{F}$, eventually approaching the limiting value $1-\alpha$. The non-monotonic fluctuations in the curves are due to the changing choice of the number of bins, m, in the discretization.

the privately computed empirical CDF. Specifically, $|\tilde{F}(s) - \hat{F}(s)| \leq \zeta_{\max}^{n,m}(\gamma\alpha, \epsilon)$ holds for all $s \in (0,1)$ with probability at least $1-\gamma\alpha$. This allows us to replace $\tilde{F}$ with $\hat{F}$ in Eq. (5) by writing

$$\mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q}))\right] \geq (1-\gamma\alpha)\mathbb{E}\left[F(\hat{F}^{-1}(\hat{q} - \zeta_{\max}^{n,m}(\gamma\alpha, \epsilon)))\right].$$

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 $\hat{q}$ as a function of this expectation completes the proof of the theorem. □

In Theorem 2 we present a general statement of our coverage result—without an explicit bound on $\zeta_{\max}^{n,m}(\nu, \epsilon)$—since in practice the way to get the best performance is to evaluate this term via simulation. This is computationally inexpensive, so the simulation error can be made negligible. See Figure 2 for a numerical evaluation of this choice of adjusted quantile. If we wish instead to obtain an expression that lends itself to theoretical understanding, we can do so by incorporating an explicit upper bound on $\zeta_{\max}^{n,m}(\nu, \epsilon)$.

Corollary 1 (Coverage guarantee, simplified form)**.** *Fix the differential privacy level $\epsilon > 0$ and miscoverage level α such that $\alpha > 4\exp(-m)$. Fix also a free parameter $\gamma \in [4\exp(-m)/\alpha, 1)$. Let*

$$\hat{s} = \inf\left\{z : \tilde{F}(z) \geq \frac{(n+1)(1-\alpha)}{n(1-\gamma\alpha)} + \frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon}\right\}. \tag{6}$$

Then, the prediction sets in (2) *with this choice of $\hat{s}$ satisfy the coverage property in* (1)*.*

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 $1-\alpha$. As with our lower bound on coverage, we proceed by proving an abstract result followed by an explicit special case.

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 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:*

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} \leq 1 - \alpha + \gamma\alpha + (1 - \gamma\alpha)\left(\frac{1}{n+1} + p^m_{\max} + 2\zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)\right).$$

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 2 (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* $m \propto (n\epsilon)^{2/3}$ *and* $\gamma = 1/m$*. Then, the prediction sets in* (2)*, with* $\hat{s}$ *as in Theorem 2, satisfy the following upper bound:*

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} \leq 1 - \alpha + O\left(\frac{\sqrt{\log\left(n\epsilon/\alpha\right)}}{(n\epsilon)^{2/3}}\right).$$

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 there is little noise addition due to privacy, but the histogram is an overly coarse approximation of the empirical distribution of the scores. On the other hand, if m is too large, then the histogram is accurate, but there is a lot of additive noise implied by the requirement of differential privacy. This tension can be observed in the terms in Theorem 3 that have a dependence on m, namely $p^m_{\max}$ and $\zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)$. Corollary 2 suggests that the correct balance—which leads to minimal excess coverage—is to take $m \propto (n\epsilon)^{-2/3}$.

This upper bound also gives insight to an important theoretical question: what is the cost of privacy in conformal prediction? In non-private conformal prediction, the upper bound is $1-\alpha+O(n^{-1})$ [4]. In private conformal prediction, we achieve an upper bound of $1 - \alpha + \tilde{O}((n\epsilon)^{-2/3})$, 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 Eq. (1), using the following score function:

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

as in Sadinle et al. [34]. 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 dataset [58] wherever we require a privately trained neural network. Otherwise, we use a non-private model on the ImageNet dataset [59], 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.1 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 dataset [1]. The reader can reproduce the experiments exactly using our public GitHub repository.

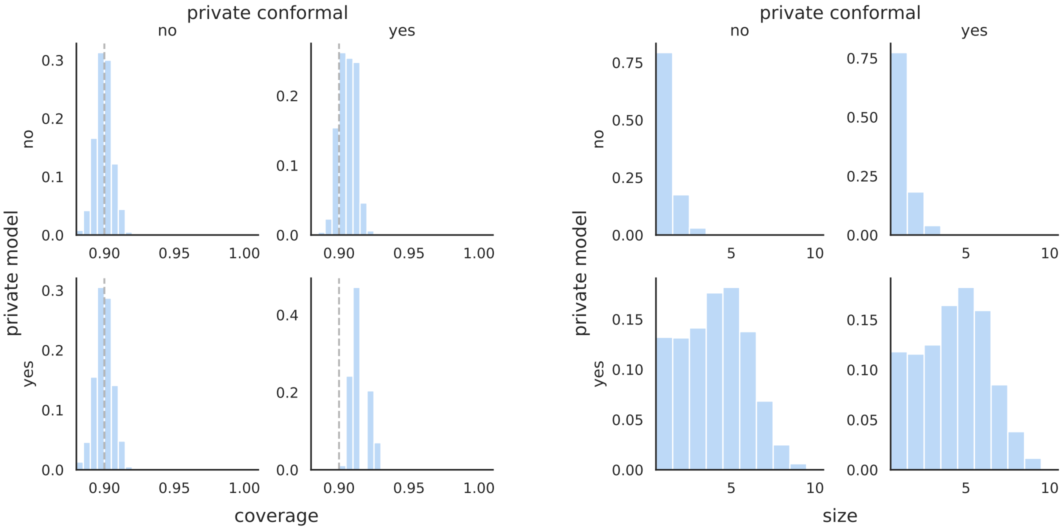

Figure 3: **Coverage and set size with private/non-private models and private/non-private conformal prediction.** We demonstrate histograms of coverage and set size of non-private/private models and non-private/private conformal prediction at the level $\alpha = 0.1$, with $\epsilon = 8$, $\delta = 1e - 5$, and $n = 5000$.

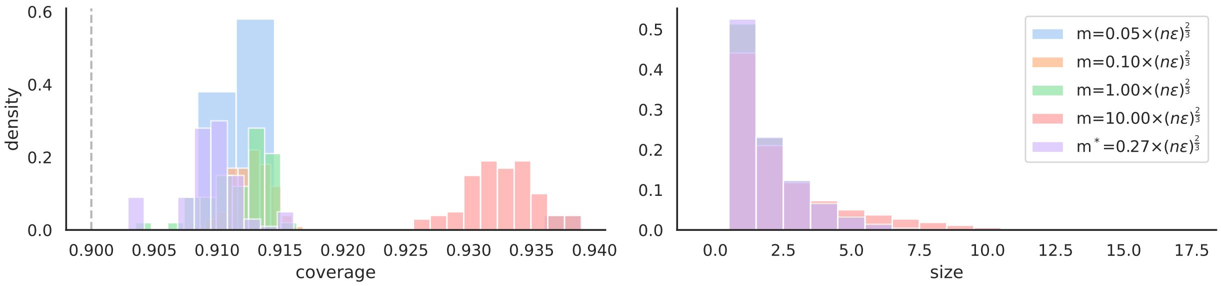

Figure 4: **Coverage and set size for different values of m.** We demonstrate the performance on Imagenet of private conformal prediction using a non-private ResNet-152 as the base model at $\alpha = 0.1$ and $\epsilon = 1$. The coverage improves as m approaches $\approx (n\epsilon)^{\frac{2}{3}}$, then degrades. See Section 4.2 for details.

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, non-private conformal prediction with a private model, private conformal prediction with a non-private model, and non-private conformal prediction with a non-private model. The non-private model and private model are both ResNet-18s [60]. The private model is trained with private SGD [14], as implemented in the `Opacus` library, with privacy parameters $\epsilon = 8$ and $\delta = 1e - 5$. The non-private model's accuracy (83%) was significantly higher than that of the private model (49%). We used the suggested private model training parameters from the `Opacus` library with minor adjustments (see Appendix C.2), but did not optimize hyperparameters as our work does not aim to improve private model training.

Figure 3 shows histograms of the coverages and set sizes of these procedures over 1000 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 much 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

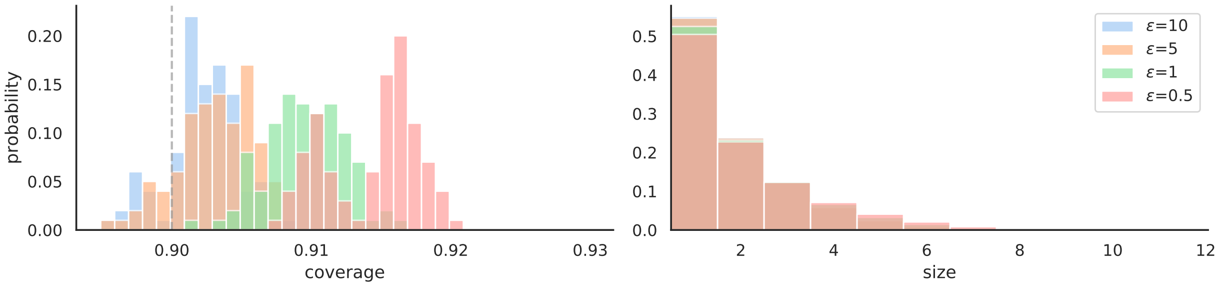

Figure 5: **Coverage and set size for different values of ϵ.** We demonstrate the performance on ImageNet of private conformal prediction using a non-private ResNet-152 as the base model with $\alpha = 0.1$. The coverage improves for liberal (large) ϵ. See Section 4.3 for details.

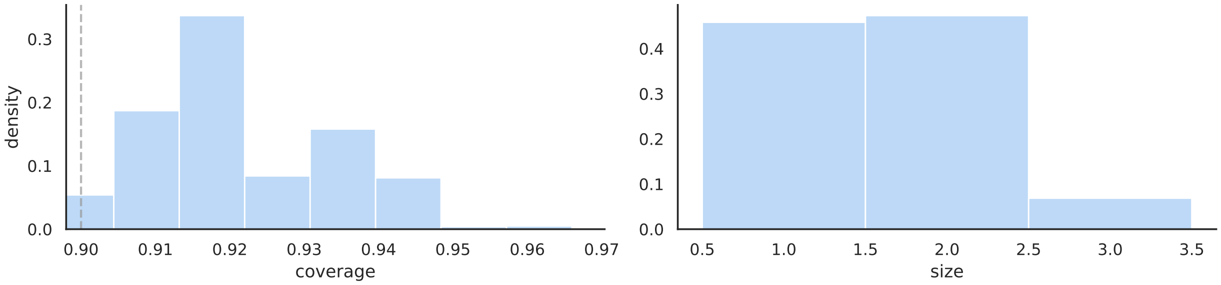

Figure 6: **Coverage and set size on the CoronaHack dataset.** We demonstrate the performance on the CoronaHack dataset of private conformal prediction using a non-private ResNet-18 as the base model with $\alpha = 0.1$. The average coverage was 92%. See Section 4.4 for details.

right quadrant of the plot.

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 $(n\epsilon)^{\frac{2}{3}}$, 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 results in excessive Laplacian noise added to the privatized empirical CDF. The optimal choice of m balances these two factors.

To demonstrate this tradeoff, we performed experiments on ImageNet. We used a non-private, pre-trained ResNet-152 from the `torchvision` repository as the base model. Figure 4 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.

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 5 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 dataset, a public chest X-ray dataset containing 5908 X-rays labeled as `normal`, `viral pneumonia` (primarily COVID-19), or `bacterial pneumonia`. Using 4408 training pairs over 14 epochs, we (non-privately) 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 6 plots the coverage and set size of this procedure over 1000 different train/calibration/validation splits of the dataset, and Figure 1 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 CDF estimation. 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 $\tilde{O}((n\epsilon)^{-2/3})$ compared to $O(n^{-1})$ for non-private conformal prediction, a mild decrease in efficiency. This is confirmed in our experiments, where we show that there is little difference between private and non-private 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 non-private 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

[1] J. C. Perez, C. de Blas Perez, F. L. Alvarez, and J. M. C. Contreras, "Databiology Lab CORONAHACK: Collection of public COVID-19 data," *bioRxiv*, 2020.

[2] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, "Inductive confidence machines for regression," in *Machine Learning: European Conference on Machine Learning*, pp. 345–356, 2002.

[3] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[4] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*, pp. 265–284, Springer, 2006.

[6] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054–1067, 2014.

[7] A. Bittau, Ú. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld, "Prochlo: Strong privacy for analytics in the crowd," in *Proceedings of the 26th Symposium on Operating Systems Principles*, pp. 441–459, 2017.

[8] Differential Privacy Team Apple, "Learning with privacy at scale," in *Apple Machine Learning Research*, 2017.

[9] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," in *Advances in Neural Information Processing Systems*, pp. 3571–3580, 2017.

[10] J. M. Abowd, "The US census bureau adopts differential privacy," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867, 2018.

[11] C. Dwork, "Differential privacy and the US census," in *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 1–1, 2019.

[12] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization.," *Journal of Machine Learning Research*, vol. 12, no. 3, 2011.

[13] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473, IEEE, 2014.

[14] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

[15] S. Neel, A. Roth, G. Vietri, and S. Wu, "Oracle efficient private non-convex optimization," in *International Conference on Machine Learning*, pp. 7243–7252, PMLR, 2020.

[16] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, "Differentially private histogram publication," *The VLDB Journal*, vol. 22, no. 6, pp. 797–822, 2013.

[17] J. Lei, "Differentially private m-estimators," *Advances in Neural Information Processing Systems*, vol. 24, pp. 361–369, 2011.

[18] A. Smith, "Privacy-preserving statistical estimation with optimal convergence rates," in *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 813–822, 2011.

[19] V. Karwa and S. Vadhan, "Finite sample differentially private confidence intervals," *arXiv preprint arXiv:1711.03908*, 2017.

[20] O. Sheffet, "Differentially private ordinary least squares," in *International Conference on Machine Learning*, pp. 3105–3114, PMLR, 2017.

[21] M. Gaboardi, R. Rogers, and O. Sheffet, "Locally private mean estimation: z-test and tight confidence intervals," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2545–2554, PMLR, 2019.

[22] Y. Wang, D. Kifer, and J. Lee, "Differentially private confidence intervals for empirical risk minimization," *Journal of Privacy and Confidentiality*, vol. 9, no. 1, 2019.

[23] S. S. Wilks, "Determination of sample sizes for setting tolerance limits," *Annals of Mathematical Statistics*, vol. 12, no. 1, pp. 91–96, 1941.

[24] S. S. Wilks, "Statistical prediction with special reference to the problem of tolerance limits," *Annals of Mathematical Statistics*, vol. 13, no. 4, pp. 400–409, 1942.

[25] A. Wald, "An extension of Wilks' method for setting tolerance limits," *Annals of Mathematical Statistics*, vol. 14, no. 1, pp. 45–55, 1943.

[26] J. W. Tukey, "Non-parametric estimation II. statistically equivalent blocks and tolerance regions—the continuous case," *Annals of Mathematical Statistics*, vol. 18, no. 4, pp. 529–539, 1947.

[27] K. Krishnamoorthy and T. Mathew, *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley, 2009.

[28] S. Park, O. Bastani, N. Matni, and I. Lee, "PAC confidence sets for deep neural networks via calibrated prediction," in *International Conference on Learning Representations*, 2020.

[29] V. Vovk, A. Gammerman, and C. Saunders, "Machine-learning applications of algorithmic randomness," in *International Conference on Machine Learning*, pp. 444–453, 1999.

[30] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, no. Mar, pp. 371–421, 2008.

[31] J. Lei, A. Rinaldo, and L. Wasserman, "A conformal prediction approach to explore functional data," *Annals of Mathematics and Artificial Intelligence*, vol. 74, pp. 29–43, 2015.

[32] V. Vovk, "Cross-conformal predictors," *Annals of Mathematics and Artificial Intelligence*, vol. 74, no. 1-2, pp. 9–28, 2015.

[33] R. F. Barber, E. J. Candes, A. Ramdas, R. J. Tibshirani, *et al.*, "Predictive inference with the jackknife+," *Annals of Statistics*, vol. 49, no. 1, pp. 486–507, 2021.

[34] M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *Journal of the American Statistical Association*, vol. 114, pp. 223 – 234, 2019.

[35] A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," *arXiv:2009.14193*, 2020.

[36] V. Vovk, "Conditional validity of inductive conformal predictors," in *Proceedings of the Asian Conference on Machine Learning*, vol. 25, pp. 475–490, 2012.

[37] R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "The limits of distribution-free conditional predictive inference," *Information and Inference: A Journal of the IMA*, 2019.

[38] Y. Romano, E. Patterson, and E. Candès, "Conformalized quantile regression," in *Advances in Neural Information Processing Systems*, vol. 32, pp. 3543–3553, 2019.

[39] R. Izbicki, G. T. Shimizu, and R. B. Stern, "Flexible distribution-free conditional predictive bands using density estimators," *arXiv:1910.05575*, 2019.

[40] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," *arXiv:2006.02544*, 2020.

[41] L. Guan, "Conformal prediction with localization," *arXiv:1908.08558*, 2020.

[42] M. Cauchois, S. Gupta, and J. Duchi, "Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction," *arXiv:2004.10181*, 2020.

[43] J. Lei, "Classification with confidence," *Biometrika*, vol. 101, pp. 755–769, 10 2014.

[44] Y. Hechtlinger, B. Poczos, and L. Wasserman, "Cautious deep learning," *arXiv:1805.09460*, 2018.

[45] L. Guan and R. Tibshirani, "Prediction and outlier detection in classification problems," *arXiv:1905.04396*, 2019.

[46] V. Vovk, J. Shen, V. Manokhin, and M.-g. Xie, "Nonparametric predictive distributions based on conformal prediction," *Machine Learning*, pp. 1–30, 2017.

[47] V. Vovk, I. Petej, P. Toccaceli, A. Gammerman, E. Ahlberg, and L. Carlsson, "Conformal calibrators," in *Conformal and Probabilistic Prediction and Applications*, pp. 84–99, PMLR, 2020.

[48] R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," in *Advances in Neural Information Processing Systems 32*, pp. 2530–2540, 2019.

[49] M. Cauchois, S. Gupta, A. Ali, and J. C. Duchi, "Robust validation: Confident predictions even when distributions shift," *arXiv:2008.04267*, 2020.

[50] X. Hu and J. Lei, "A distribution-free test of covariate shift using conformal prediction," *arXiv:2010.07147*, 2020.

[51] L. Lei and E. J. Candès, "Conformal inference of counterfactuals and individual treatment effects," *arXiv:2006.06138*, 2020.

[52] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan, "Distribution-free, risk-controlling prediction sets," *arXiv:2101.02703*, 2021.

[53] C. Jung, C. Lee, M. M. Pai, A. Roth, and R. Vohra, "Moment multicalibration for uncertainty estimation," *arXiv:2008.08037*, 2020.

[54] V. Gupta, C. Jung, G. Noarov, M. M. Pai, and A. Roth, "Online multivalid learning: Means, moments, and prediction intervals," *arXiv:2101.01739*, 2021.

[55] E. Grycko, "Classification with set-valued decision functions," in *Information and Classification*, pp. 218–224, 1993.

[56] J. J. del Coz, J. Díez, and A. Bahamonde, "Learning nondeterministic classifiers," *Journal of Machine Learning Research*, vol. 10, no. 79, pp. 2273–2293, 2009.

[57] T. Mortier, M. Wydmuch, K. Dembczyński, E. Hüllermeier, and W. Waegeman, "Efficient set-valued prediction in multi-class classification," *arXiv:1906.08129*, 2020.

[58] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[61] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

A Auxiliary results

Lemma 1 (Lévy's maximal inequality)**.** *Let $X_1, ..., X_m$ be independent symmetric random variables. Define $S_k = \sum_{i \le k} X_i$ for $k = 1, ..., m$. Let $Z_k = |S_k|$. Then, for $t \ge 0$,*

$$\mathbb{P}\left\{\max_{k \le m} Z_k \ge t\right\} \le 2\mathbb{P}\{Z_n \ge t\}\,.$$

For a proof, see, for example, Lemma 11.12 in Boucheron et al. [61].

Lemma 2. *Let $\zeta \sim \text{Laplace}(b)$. Then, ζ is $(2b, \sqrt{2}b)$-subexponential:*

$$\mathbb{E}[e^{\lambda\zeta}] \le e^{2b^2\lambda^2} \text{ for all } |\lambda| < \frac{1}{\sqrt{2}b}.$$

Proof. The moment-generating function of ζ is $M(\lambda) = \frac{1}{1-b^2\lambda^2}$, for $|\lambda| < \frac{1}{b}$. Therefore, to complete the proof it suffices to show

$$\frac{1}{1-b^2\lambda^2} \le 1 + 2b^2\lambda^2 \tag{7}$$

for $|\lambda| < \frac{1}{\sqrt{2}b}$, since $1 + x \le e^x$. Rearranging the terms, we observe that condition (7) is equivalent to:

$$0 \le (2b^2 - b^2 - 2b^4\lambda^2)\lambda^2,$$

and this is clearly satisfied for $|\lambda| < \frac{1}{\sqrt{2}b}$. □

Lemma 3 (Bernstein's inequality)**.** *Let X be a (ν, b)-subexponential random variable, meaning that*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\nu^2\lambda^2/2} \text{ for all } |\lambda| < \frac{1}{b}.$$

Then,

$$\mathbb{P}\{|X| \geq t\} \leq \begin{cases} 2\exp\left(-\frac{t^2}{2\nu^2}\right), & 0 \leq t \leq \frac{\nu^2}{b}, \\ 2\exp\left(-\frac{t}{2b}\right), & t > \frac{\nu^2}{b}. \end{cases}$$

Lemma 4. *Let F be the CDF of a distribution supported on a finite set $\{a_1, \dots, a_m\}$. Let $Z_1, \dots, Z_n \overset{i.i.d.}{\sim} F$, and let $\hat{F}$ denote the empirical CDF corresponding to $Z_1, \dots, Z_n$. 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 non-decreasing order statistic of $Z_1, \dots, Z_n$. By monotonicity of F, we further have that $F(Z_{(\lceil nq\rceil)})$ is identical to the $\lceil nq\rceil$-th non-decreasing order statistic of $F(Z_1), \dots, F(Z_n)$. By a standard argument, the samples $F(Z_1), \dots, F(Z_n)$ are super-uniform, i.e. $\mathbb{P}\{F(Z_1) \leq u\} \leq u$ 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 $i \in \{1, \dots, n\}$,

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

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:

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

and $\tilde{F}$ denotes the private CDF returned by Algorithm 2.

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 $\hat{F}^{-1}(\delta)$ and $\tilde{F}^{-1}(\delta)$. Finally, we denote $\bar{\zeta} = \frac{1}{n}\max_{1\leq j\leq m} |\sum_{i=1}^j \zeta_i|$.

Notice that we can write $\tilde{F}(s) = \hat{F}(s) + \xi(s)$, where $\xi(s) = \frac{1}{n}\sum_{i=1}^m \zeta_i \mathbf{1}\{e_i \leq s\}$. Thus, we have

$$\left\{\bar{\zeta} \leq \zeta_{\max}^{n,m}(\nu, \epsilon),\ \tilde{F}(z) \geq q\right\} \subseteq \left\{\bar{\zeta} \leq \zeta_{\max}^{n,m}(\nu, \epsilon),\ \hat{F}(z) \geq q - \zeta_{\max}^{n,m}(\nu, \epsilon)\right\},$$

and consequently

$$\left\{\bar{\zeta} \leq \zeta_{\max}^{n,m}(\nu, \epsilon),\ z \leq \tilde{F}^{-1}(q)\right\} \supseteq \left\{\bar{\zeta} \leq \zeta_{\max}^{n,m}(\nu, \epsilon),\ z \leq \hat{F}^{-1}(q - \zeta_{\max}^{n,m}(\nu, \epsilon))\right\}. \quad (8)$$

Denote $\hat{q} = \frac{(n+1)(1-\alpha)}{n(1-\gamma\alpha)} + \zeta_{\max}^{n,m}(\gamma\alpha, \epsilon)$, and notice that $\hat{s} = \tilde{F}^{-1}(\hat{q})$. With this, we can write

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} = \mathbb{P}\{S(X, Y) \leq \hat{s}\} = \mathbb{P}\left\{S(X, Y) \leq \tilde{F}^{-1}(\hat{q})\right\} = \mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q}))\right].$$

In the last step we use the fact that the event $\{S(X,Y) \leq \tilde{F}^{-1}(\hat{q})\}$ is equivalent to $\{[S(X,Y)] \leq \tilde{F}^{-1}(\hat{q})\}$, because $\tilde{F}^{-1}(\hat{q})$ is supported on $\{e_i\}_{i=0}^m$.

By splitting up the analysis depending on whether $\bar{\zeta} \leq \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon)$, we obtain the following:

$$\begin{aligned}\mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q}))\right] &= \mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q}))\mathbf{1}\{\bar{\zeta} > \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon)\}\right] + \mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q}))\mathbf{1}\{\bar{\zeta} \leq \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon)\}\right]\\ &\geq \gamma\alpha \cdot 0 + \mathbb{E}\left[\mathbf{1}\{[S(X,Y)] \leq \tilde{F}^{-1}(\hat{q}), \bar{\zeta} \leq \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon)\}\right]\\ &\geq \mathbb{E}\left[\mathbf{1}\{[S(X,Y)] \leq \hat{F}^{-1}(\hat{q} - \zeta_{\max}^{n,m}(\nu,\epsilon)), \bar{\zeta} \leq \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon)\}\right]\\ &= (1-\gamma\alpha)\mathbb{E}\left[F(\hat{F}^{-1}(\hat{q} - \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon)))\right],\end{aligned}$$

where in the third step we apply Eq. (8). Thus, it suffices to show that

$$\mathbb{E}\left[F(\hat{F}^{-1}(\hat{q} - \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon)))\right] \geq \frac{1-\alpha}{1-\gamma\alpha}. \tag{9}$$

Let $j^* = \lceil n(\hat{q} - \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon))\rceil$. Then, by Lemma 4,

$$F(\hat{F}^{-1}(\hat{q} - \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon))) \succeq \text{Beta}(j^*, n - j^* + 1),$$

so

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

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

$$\frac{\lceil n(\hat{q} - \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon))\rceil}{n+1} \geq \frac{1-\alpha}{1-\gamma\alpha},$$

holds, which implies Eq. (9) and thus completes the proof.

B.2 Proof of Corollary 1

It suffices to show that

$$\frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon} \geq \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon). \tag{10}$$

To prove so, we use Lévy's maximal inequality, stated in Lemma 1; we get

$$\mathbb{P}\left\{\frac{1}{n}\max_{1\leq j\leq m}\left|\sum_{i=1}^{j}\zeta_i\right| \geq \frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon}\right\} \leq 2\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{m}\zeta_i\right| \geq \frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon}\right\}.$$

The noise variables $\{\zeta_i\}_{i=1}^m$ are independent $(4/\epsilon, 2\sqrt{2}/\epsilon)$-subexponential random variables; see Lemma 2 for a proof. Therefore, their sum is $(4\sqrt{m}/\epsilon, 2\sqrt{2}/\epsilon)$-subexponential. Applying Bernstein's inequality (Lemma 3) with this choice of parameters gives

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{m}\zeta_i\right| \geq \frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon}\right\} \leq 2\exp\left(-\frac{n^2\epsilon^2}{32m}\left(\frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon}\right)^2\right) = \frac{\gamma\alpha}{2}.$$

Note that we are in the regime of Bernstein's inequality with faster decay because $\alpha \geq \frac{4}{\gamma}\exp(-m)$ implies $\frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon} \leq \frac{8m}{\sqrt{2}n\epsilon}$. Putting everything together, we get the following bound on the second term:

$$\mathbb{P}\left\{\frac{1}{n}\max_{1\leq j\leq m}\left|\sum_{i=1}^{j}\zeta_i\right| \geq \frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon}\right\} \leq 2\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{m}\zeta_i\right| \geq \frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon}\right\} \leq \gamma\alpha.$$

With this, we have proved $\frac{4\sqrt{2m\log(4/(\gamma\alpha))}}{n\epsilon} \geq \zeta_{\max}^{n,m}(\gamma\alpha,\epsilon)$, as desired.

B.3 Proof of Theorem 3

We adopt the definitions of $F, \hat{F}, \tilde{F}$, and $\bar{\zeta}$ from the proof of Theorem 2. By a similar reasoning as in Theorem 2, we have

$$\left\{\bar{\zeta} \leq \zeta^{n,m}_{\max}(\nu, \epsilon),\ \tilde{F}(z) \geq q\right\} \supseteq \left\{\bar{\zeta} \leq \zeta^{n,m}_{\max}(\nu, \epsilon),\ \hat{F}(z) \geq q + \zeta^{n,m}_{\max}(\nu, \epsilon)\right\},$$

and consequently

$$\left\{\bar{\zeta} \leq \zeta^{n,m}_{\max}(\nu, \epsilon),\ z \leq \tilde{F}^{-1}(q)\right\} \subseteq \left\{\bar{\zeta} \leq \zeta^{n,m}_{\max}(\nu, \epsilon),\ z \leq \hat{F}^{-1}(q + \zeta^{n,m}_{\max}(\nu, \epsilon))\right\}. \tag{11}$$

We can again write

$$\mathbb{P}\{Y \in \mathcal{C}(X)\} = \mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q}))\right],$$

where $\hat{q} = \frac{(n+1)(1-\alpha)}{n(1-\gamma\alpha)} + \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)$, and thus $\hat{s} = \tilde{F}^{-1}(\hat{q})$.

By splitting up the analysis depending on whether $\bar{\zeta} \leq \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)$, we obtain the following:

$$\begin{aligned}
\mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q}))\right] &= \mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q}))\mathbf{1}\{\bar{\zeta} > \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)\}\right] + \mathbb{E}\left[F(\tilde{F}^{-1}(\hat{q}))\mathbf{1}\{\bar{\zeta} \leq \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)\}\right] \\
&\leq \gamma\alpha \cdot 1 + \mathbb{E}\left[\mathbf{1}\{[S(X,Y)] \leq \tilde{F}^{-1}(\hat{q}), \bar{\zeta} \leq \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)\}\right] \\
&\leq \gamma\alpha + \mathbb{E}\left[\mathbf{1}\{[S(X,Y)] \leq \hat{F}^{-1}(\hat{q} + \zeta^{n,m}_{\max}(\nu, \epsilon)), \bar{\zeta} \leq \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)\}\right] \\
&= \gamma\alpha + (1-\gamma\alpha)\mathbb{E}\left[F(\hat{F}^{-1}(\hat{q} + \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)))\right],
\end{aligned} \tag{12}$$

where in the third step we apply Eq. (11). Let $j^* = \lceil n(\hat{q} + \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon))\rceil$. By Lemma 4, we have

$$F(\hat{F}^{-1}(\hat{q} + \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon))) \preceq \textsc{Beta}(j^*, n - j^* + 1) + p^m_{\max},$$

so

$$\mathbb{E}\left[F(\hat{F}^{-1}(\hat{q} + \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)))\right] \leq \frac{j^*}{n+1} + p^m_{\max} = \frac{\lceil n(\hat{q} + \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon))\rceil}{n+1} + p^m_{\max}. \tag{13}$$

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

$$\frac{\lceil n(\hat{q} + \zeta^{n,m}_{\max}(\gamma\alpha, \epsilon))\rceil}{n+1} \leq \frac{\frac{1-\alpha}{1-\gamma\alpha}(n+1) + 2\zeta^{n,m}_{\max}(\gamma\alpha, \epsilon)n + 1}{n+1} = \frac{1-\alpha}{1-\gamma\alpha} + 2\zeta^{n,m}_{\max}(\gamma\alpha, \epsilon) + \frac{1}{n+1}. \tag{14}$$

Putting together equations (12), (13), and (14) completes the proof.

C Experimental details

C.1 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. The algorithm simply entails simulating uniformly distributed scores and then choosing the (m^*, γ) 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 values around $(n\epsilon)^{\frac{2}{3}}$ and γ can be chosen from coarsely spaced values from $1e-4$ to 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 $s_i$ for the optimal $\gamma$ chosen by grid search.
  output: $m^*$, $\gamma$
```

C.2 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 non-private model training is the use of the `--disable-dp` flag, turning off the added noise but preserving all other settings. In the private model training, we make a minor modification to the noise scaling due to the fact that we are working under the replacement definition of differential privacy and the `Opacus` privacy accounting is done assuming the removal definition. More precisely, if C is the clipping value of the algorithm, then the ℓ_2-sensitivity to the removal of a data point is C, while the ℓ_2-sensitivity to the replacement of a data point is $2C$. For this reason, instead of adding noise with level σC, we add noise with level $2\sigma C$. We run the private training procedure for 470 epochs to achieve $\epsilon = 8$.