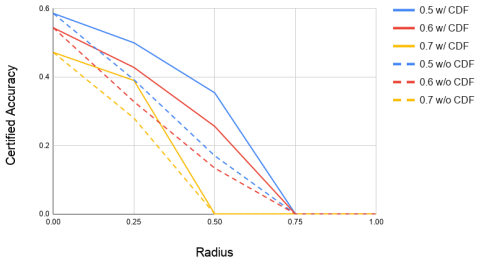
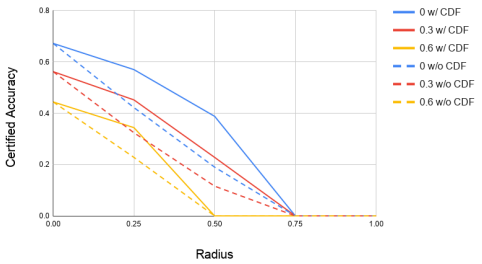
ImageNet results (R2, R3 & R4): As suggested by most reviewers, we perform our experiments on the ImageNet dataset by certifying a pre-trained ResNet-50 from Cohen et al. in [6] and produce the following plots similar to Figure 2 in our paper. Just like the CIFAR-10 results, the CDF-based method significantly outperforms the naive baseline for different levels of smoothing noise σ. Due to space constraints, we show plots for $\sigma = 0.25$ noting that plots for $\sigma = 0.5$ demonstrate similar trends.



(a) Average Prediction Score



(b) Margin

Figure 1: ImageNet Results ($\sigma = 0.25$).

Calibration of confidence information (R1): We agree that it is improper to interpret confidence scores as the "probability" of the classification being correct – indeed, this is the reason why we prefer the term "class score" over "class probability." However, we think it is widely accepted that scores are often strongly correlated with the chance of correct classification, even if this is not an exact or linear relationship because of poor calibration.

Correlation between prediction score and certified radius (R1): We do not mean to imply no correlation exists, but rather that the correlation in Fig 4 is *weak*; there are many images with large radius but low confidence, and visa versa.

Theorem 1 is an immediate application of Lemma 2 in Salman et al. (R1): While the two theorems might be related, the connection does not seem so obvious for us. Theorem 1 in our paper is regarding the *baseline* certificate against which we compare our CDF-based certificate (Theorem 2) which is our main technical contribution. We will make a reference to this lemma in the paper.

Parameters $s_1, \ldots, s_n$ are not discussed in text (R2): In our experiments, we set $s_1, \ldots, s_n$ such that the number of confidence score values in every interval $(a, s_1), (s_1, s_2), \ldots, (s_n, b)$ is equal. We chose this approach instead of setting the parameters at regular intervals in (a, b) because doing so would make a lot of intervals empty. It made sense to split the range in such a way that the intervals are well-balanced. There could be other reasonable ways to set these parameters which could have an impact on the performance. We will add this discussion to the paper.

In page 3, how to determine the parameter a, b and c? (R3): Parameters a and b depend on the confidence measure used and denote the range (a, b) in which the confidence scores lie. For instance, if you are using average prediction score, then $a = 0$ and $b = 1$ (see page 6 under section 4: Confidence measures) and if you are considering margin of average prediction score, then $a = -1$ and $b = 1$. Parameter c is the confidence threshold you would like to certify, i.e., the confidence score is guaranteed to be above c within the certified radius. This parameter can be set by the user to calculate the corresponding certified radius. In our experiments, to keep computing requirements within reach, we compute the threshold c for a chosen radius instead. But, the original task of computing the radius for a given c can be accomplished using a simple binary-search. (See page 5, lines 172-175).

The rationale behind confidence notions (R3):. The *average* score is a very conventional measure of classification confidence, and is the more direct analog of classical confidence scores in the setting of a smoothed classifier. The classifier *margin* is less conventional; it measures how much more certain the top class is over the second-place class. We find that the distribution of margins is often more concentrated than average scores, and we can therefore produce certificates with stronger bounds. There is a tradeoff here that the user can choose – the average is more interpretable but provides weaker bounds, the margin is less conventional (although large margin is still a good indicator of confidence) and provides tighter certificates.

Relationship between radii σ and confidence threshold (R3): If you consider two Guassians separated by ϵ units, their overlap will be large when σ is large relative to ϵ, and the overlap is very small when σ is small. For this reason, when σ is large, the confidence bound decreases more slowly with increasing ϵ, allowing us to certify larger radii. However, when ϵ is small (e.g., for certifying small radii), the overlap between Gaussians is large, even for small σ, and so we can produce certificates with small σ. In this case, we can take advantage of the fact that the sampled scores are more homogenous when σ is small to get tighter bounds. Our experiments show numerous values of σ to show that our results hold for a range of choices. However in practice the optimal σ is proportional to the radius you want to certify. We will include this discussion when we update our paper.

Related work (R2 & R3): We will be sure to add a section on related work in randomized smoothing and certifying confidences.