Treatment of Statistical Estimation Problems in Randomized Smoothing for Adversarial Robustness

Václav Voráček
Tübingen AI center, University of Tübingen
vaclav.voracek@uni-tuebingen.de

Abstract

Randomized smoothing is a popular certified defense against adversarial attacks. In its essence, we need to solve a problem of statistical estimation which is usually very time-consuming since we need to perform numerous (usually 10^5) forward passes of the classifier for every point to be certified. In this paper, we review the statistical estimation problems for randomized smoothing to find out if the computational burden is necessary. In particular, we consider the (standard) task of adversarial robustness where we need to decide if a point is robust at a certain radius or not using as few samples as possible while maintaining statistical guarantees. We present estimation procedures employing confidence sequences enjoying the same statistical guarantees as the standard methods, with the optimal sample complexities for the estimation task and empirically demonstrate their good performance. Additionally, we provide a randomized version of Clopper-Pearson confidence intervals resulting in strictly stronger certificates. The code can be found at `https://github.com/vvoracek/RS_conf_seq`.

We encourage the reader only interested in statistics to start at subsection 2.1.

1 Introduction

Adversarial robustness: It is well known that a tiny, adversarial, perturbation of the input can change the output of basically any undefended machine learning model (Biggio et al., 2013; Szegedy et al., 2014); this is unpleasant and we continue in the mitigation of the problem. There are two main lines of work tackling this problem: (1) Empirical: the standard approach here is to use adversarial training (Madry et al., 2018; Goodfellow et al., 2014) where the model is trained on adversarial examples. This approach does not provide guarantees, only empirical evidence suggesting that the model may be robust. With stronger attacks, we might (yet again) realize it is not the case. (2) Certified: with formal robustness guarantees for the model. We will focus on this, and in particular on *randomized smoothing* (Lecuyer et al., 2019) which is currently the strongest certification method[1]. We will not cover other certification methods and we refer the reader to the survey Li et al. (2023) instead. We consider the standard task of certified robustness; the goal is to decide if the decision of a classifier F at a particular input x is robust against additive perturbations δ such that $\|\delta\| \leq r$ for some norm $\|\cdot\|$. Formally, we ask if $F(x) = F(x')$ whenever $\|x - x'\| \leq r$.

Randomized smoothing is a framework providing state of the art formal guarantees on the adversarial robustness for many datasets. One of its benefits lies in the fact that there are no assumptions on the model, making it possible to readily transfer the methods from defending image classifiers against sparse pixel changes to different modalities; e.g., defending large language models against change of

[1] see leaderboard `https://sokcertifiedrobustness.github.io/leaderboard/`

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

some letters/words/tokens. Randomized smoothing transforms any undefended classifier $F : \mathbb{R}^d \to \mathcal{Y}$ by a smoothing distribution φ into a smoothed classifier $H_\varphi(x) = \arg\max_{y\in\mathcal{Y}} \mathbb{P}_{\delta\sim\varphi}[\![F(x+\delta) = y]\!]$ for which robustness guarantees exist. We postpone the details for later. During the certification process, we need to estimate the maximum probability of a multinomial distribution from samples as the exact computation is intractable. This statistical estimation problem is the focus of this paper.

Speed issues: The main weakness of randomized smoothing is the extensive time required for both prediction and certification, making it troublesome for real-world applications. There is an inherent trade-off between the allowed probability of incorrectly claiming robustness[2] (type-1 error, α), the probability of incorrectly claiming non-robustness (type-2 error, β), and the number of samples used n. The standard practice is to set $\alpha = 0.001$, $n = 100\ 000$ and the value of β is then implicit. it might not be the most practically relevant setting since the implicitly set value of β is exponentially small in n when the sample is not close to the threshold. The claim is made precise in Example A.1.

The arguably more relevant setting is to set the values of α and β and leave n implicit. This is much more challenging since it is no longer possible to draw the predetermined number of samples and use a favourite concentration inequality. We propose a new certification procedure using confidence sequences to adaptively (and optimally) deciding how many samples to draw addressing the problem.

Contributions:

- We introduce a new, strictly better version of Clopper-Pearson confidence intervals for estimating the class probabilities in Subsection 2.1. The presented interval is optimal, and thus is the ultimate solution to the canonical statistical estimation of randomized smoothing.
- We propose new methods for the certification utilizing confidence sequences (instead of confidence intervals) in Subsection 2.2. This allows us to draw *just enough* samples to certify robustness of a point; greatly decreasing the number of samples needed.
- We provide a complete theoretical analysis of the proposed certification procedures. In particular, we provide matching (up to a constant factor) lower-bounds and upper-bounds for the width of the respective confidence intervals. We invert the bound and show that the certification procedure has the optimal sample-complexity in an adaptive estimation task.
- We provide empirical validation of the proposed methods confirming the theory.

Notation: Bernoulli random variable with mean p is denoted as $\mathcal{B}(p)$ and binomial random variable is $\mathcal{B}(n, p)$. Random variables are in capitals (X) and the realizations are lowercase (x). We type sequences in bold and denote $\mathbf{x}_{:t}$ first t elements of $\mathbf{x}$. We write $a \lesssim b$ if there exists a universal constant $C > 0$ such that $a \leq Cb$. If $a \lesssim b$ and $b \lesssim a$, then we write $a \asymp b$. Iverson bracket $[\![\Phi]\!]$ evaluates to 1 if Φ is true and to 0 otherwise.

1.1 Paper organization

First, in Section 2 we introduce randomized smoothing, then, in Subsection 2.1, we introduce Clopper-Pearson confidence intervals, show that they are conservative and propose their improved (optimal) randomized version. In Subsection 2.2 we discuss shortcomings of confidence intervals and introduce confidence sequences and provide lower and upper bounds for their performance. We use the confidence sequences in Section 3 where we consider a sequential estimation task.

2 Randomized Smoothing

As outlined in Introduction, consider a classifier $F : \mathbb{R}^d \to \mathcal{Y}$ and let the class probabilities under additive noise φ be $h_\varphi(x)_y = \mathbb{P}_{\delta\sim\varphi}[F(x + \delta) = y]$. Denote the highest probability (breaking ties arbitrarily) class in the original point $A = \arg\max_{y\in\mathcal{Y}} h_\varphi(x)_y$ and the second-highest probability class $B = \arg\max_{y\in\mathcal{Y}\setminus A} h_\varphi(x)_y$. Let the corresponding probabilities be p_A and p_B respectively. Recalling that $H_\varphi(x) = \arg\max_{y\in\mathcal{Y}} h_\varphi(x)_y$, then for a certain function $r : [0, 1]^2 \to \mathbb{R}_+$ we have

$$\|x - x'\| \leq r(p_A, p_B) \implies H_\varphi(x) = H_\varphi(x').$$

[2] of an input for a model at a certain radius

This r depends on the smoothing distribution φ and the considered norm. For example, if the considered norm is ℓ_2 and φ is isotropic Gaussian with standard deviation σ, then $r(p_A, p_B) = \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$ where Φ^{-1} is Gaussian quantile function Cohen et al. (2019). Note that in general, $r(\cdot, \cdot)$ is increasing in the first coordinate and decreasing in the second one. The intuition is that the larger the value of p_A at x, the larger it will be also in the neighborhood of x; similarly for p_B. It is common in the literature to use the bound $p_B \leq 1 - p_A$ and thus certify $r(p_A, 1 - p_A)$. We stick to the convention in the paper and discuss the topic in more details in Appendix B.

Statistical estimation: The crux of the paper lies in the statistical estimation problems for randomized smoothing. We consider the abstract framework for randomized smoothing, so the proposed techniques can be used as a drop-in replacement in all randomized smoothing works with a statistical-estimation component (i.e., not in the de-randomized ones such as Levine & Feizi (2021)). We do not only propose methods that work good empirically, we also provide theoretical analysis suggesting that we solve the problems optimally in certain strong sense. The main focus is on the following two constructs.

1. Confidence intervals: A standard component of randomized smoothing pipelines is the Clopper-Pearsons confidence interval. It is known to be conservative[3]; thus, the certification procedures are unnecessarily underestimating the certified robustness. We provide the *optimal* confidence interval for binomial random variables, resolving this issue completely.
2. Confidence sequences: In the standard randomized smoothing practice, we draw a certain, predetermined, number of samples and then we compute the certified radius on a confidence level $1 - \alpha$. We improve on this by allowing for adaptive estimation procedures employing confidence sequences; We demonstrate the performance in the standard task of adversarial robustness, where we want to decide if a point is robust at radius r with type-1 (resp. 2) error rates α (resp. β) using as few samples as possible.

Overview of randomized smoothing: The most relevant related works are Horváth et al. (2022); Chen et al. (2022) and they are discussed in Section 3. Here we briefly summarize literature relevant to randomized smoothing in general. The choice of the smoothing distribution φ is a crucial decision determined mainly by the threat model with respect to which we want to be robust. For example, if we are after certifying ℓ_1 robustness, we choose a uniform distribution in a $d-$dimensional ℓ_∞ ball (Lee et al., 2019; Yang et al., 2020), or better, splitting noise (Levine & Feizi, 2021), but we do not go into details here. Alternatively, for $p-$norms, $p \geq 2$ one would usually use a $d-$dimensional normal distribution Lecuyer et al. (2019); Cohen et al. (2019). The variance of the distribution based on how large perturbations do we allow in our threat model. We refer the reader to Yang et al. (2020) for a broader discussion on the smoothing distributions. It is possible to use methods in the spirit of randomized smoothing to certify other threat models, such as patch attacks Levine & Feizi (2020), sparse attack Bojchevski et al. (2020) which can be readily extended to other modalities. Sometimes, the "smoothing" distribution can be made supported on a small, discrete set and then we can evaluate the expectation exactly, yielding deterministic certification (often called de-randomized smoothing (Levine & Feizi, 2021)). See also Kumari et al. (2023) for a survey on randomized smoothing containing examples of when the certification is beyond additive ℓ_p-norm threat model; even such techniques use the Bernoulli estimation subroutine.

2.1 Confidence Intervals

We do not have access to the class A probability p_A and only have to estimate it from binomial samples; hence, the name *randomized* smoothing. Because of the randomness, we can only provide probabilistic statements about the robustness of a classifier in the following spirit "with probability at least $1 - \alpha$, robust radius is at least r" for a small α, usually 0.001. This failure probability corresponds to the event of overestimating p_A and we control it with the help of confidence intervals.

The standard choice for calculating the upper confidence interval is the Clopper-Pearson interval, sometimes called *exact*. Regardless of this pseudonym, it is in reality conservative. In this subsection, we introduce the Clopper-Pearson confidence interval for the mean of binomial random variables, demonstrate its limitations and introduce its (better) randomized version.

[3] See `https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval`.

Definition 2.1 (Confidence interval for binomials)**.** Let u, v map sample to a real number. They form a (possibly randomized) confidence interval $I(x) = [u(x), v(x)]$ with coverage $1 - \alpha$ if for any $p \in [0, 1]$ it holds that

$$\mathbb{P}_{X \sim \mathcal{B}(n,p), I} \left(p \in I(X) \right) \geq 1 - \alpha.$$

We will mainly use one-sided confidence intervals; that is, $u(\cdot) = 0$ (lower confidence interval) or $v(\cdot) = 1$ (upper confidence interval). When we will talk about probability of confidence interval containing the parameter, it will be in the sense of its definition, keeping in mind that confidence intervals provide no guarantees post-hoc for any individual estimation.

Clearly, if $I(x) = [0, 1]$ regardless of x, it will be a valid confidence interval but rather useless; thus we aim for short intervals. Ideally it would hold for every p that $\mathbb{P}_X(p \notin I(X)) = \alpha$, otherwise, some values are included in the confidence intervals unnecessarily often they can be shortened. In the following we introduce the standard Clopper-Pearson confidence intervals Clopper & Pearson (1934).

Definition 2.2 (Clopper-Pearson intervals)**.** One sided upper interval is defined as $v(x) = 1$ and

$$u(x) = \inf\{p \,|\, \mathbb{P}(\mathcal{B}(n, p) \geq x) > \alpha\}.$$

The lower one is defined as $u(x) = 0$ and

$$v(x) = \sup\{p \,|\, \mathbb{P}(\mathcal{B}(n, p) \leq x) > \alpha\}.$$

Amongst the deterministic confidence intervals, they are the shortest possible; however, they are in general conservative. In the binomial case $\mathcal{B}(n, p)$, there are only $n + 1$ possible outcomes; and thus only $n + 1$ possible confidence intervals suggesting that the actual coverage can be $1 - \alpha$ only for at most $n + 1$ values of p. The problem strikingly arises for upper confidence interval for large values of p. When we sample from $\mathcal{B}(n, p)$, regardless of the outcome, all values larger than $\sqrt[n]{\alpha}$ are contained in the confidence interval. This is a usual problem in the context of randomized smoothing, leading to sharp drops towards the end of robustness curves. We demonstrate this sub-optimality in the first part of Example A.2. We mitigate this problem by introducing randomness into the confidence intervals. They will still have the desired coverage level $1 - \alpha$, but will be shorter. Intuitively, we do so by "interpolating" between the deterministic confidence intervals in the spirit of Stevens (1950).

Definition 2.3 (Randomized Clopper-Pearson intervals)**.** Let W be uniform on the interval $[0, 1]$. The randomized one sided upper interval is defined as $v_r(x) = 1$ $u_r(x) = u'_r(x, W)$ where

$$u'_r(x, w) = \inf\{p \,|\, \mathbb{P}(\mathcal{B}(n, p) > x) + w\mathbb{P}(\mathcal{B}(n, p) = x) > \alpha\}.$$

The lower one is defined as $u_r(x) = 0$ and $v_r(x) = v'_r(x, W)$ where

$$v'_r(x, w) = \sup\{p \,|\, \mathbb{P}(\mathcal{B}(n, p) < x) + w\mathbb{P}(\mathcal{B}(n, p) = x) > \alpha\}.$$

Proposition 2.4. *Randomized Clopper-Pearson interval (I_{rCP}) have coverage exactly $1 - \alpha$. Furthermore, for any confidence interval I at level $1 - \alpha$, and any $p \geq q \in [0, 1]$ it holds that*

$$\mathbb{P}_{X \sim \mathcal{B}(n,p)}(q \in I(X)) \geq \mathbb{P}_{X \sim \mathcal{B}(n,p)}(q \in I_{rCP}(X)).$$

The proof is in the Appendix D and we remark that the interval can be efficiently found by binary search. Proposition 2.4 implies that the randomized Clopper-Pearson bounds are optimal and all the other confidence intervals for binomial random variables are more conservative. It remains to demonstrate the advantage of the randomized confidence intervals. We refer to Figure 1 for the comparison of the randomized and deterministic Clopper-Pearson confidence intervals and how they affect the robustness. We note that the most significant difference is towards the high values of p and for certification functions r such that $\lim_{p \to 1} r(p) = \infty$, such as when smoothing with normal distribution. This explains the common sharp drop by the end of the robustness curve.

Width of confidence intervals: For the simplicity of exposition, let the width of a confidence interval at level $1 - \alpha$ with n samples be $\asymp \sqrt{\log(1/\alpha)/n}$. This way, we hide the dependency on p into $\asymp$. In the full generality, the width of the confidence intervals exhibits many decay regimes between the rates $\sqrt{p(1-p)\log(1/\alpha)/n}$ (when $np \gg 1$) and $\log(1/\alpha)/n$ (when $np \asymp 1$). Our algorithms capture the correct scaling of the confidence intervals. See Boucheron et al. (2013) and the discussion on Bennett's inequality which captures the correct rates for Bernoulli mean estimation.

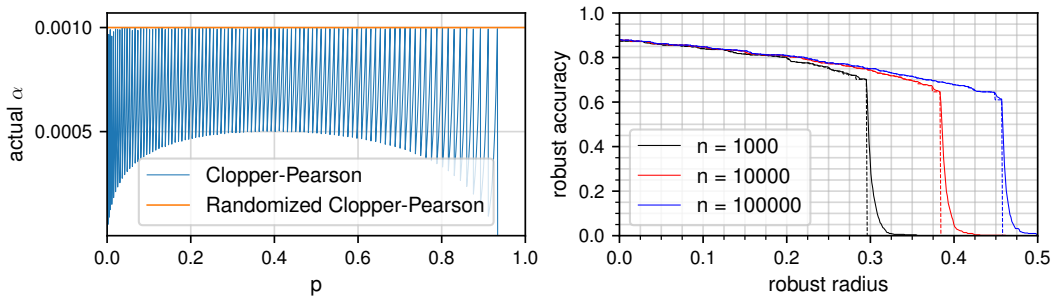

Figure 1: **left**: Comparison of coverages of confidence intervals for the mean estimation of $\mathcal{B}(100, p)$ when $\alpha := 0.001$. Note that for $p > \sqrt[n]{\alpha} \sim 0.93$, the coverage is 1. **right**: Comparison of ℓ_2-robustness curves with the standard (dashed) or the randomized (solid) Clopper-Pearson bounds on a CIFAR-10 dataset under the standard setting. The experimental details are in Appendix C.

2.2 Confidence Sequences

Limitations of confidence intervals: In order to compute the confidence intervals presented in the previous subsection, we need to collect samples and then run an estimation procedure once which brings certain limitations. Consider the following two scenarios: (1) It might be the case that we do not need all 100 000 samples and after only 10 it would be enough for our purposes because we could already conclude that the point cannot be certified here; thus, we wasted 99 990 samples. (2) Alternatively, we could see that even 10^5 samples are not enough, and we need to draw more samples. However, we have already spent our failure budget α, so we cannot even carry another test at all.

This motivates the introduction of confidence sequences. They generalize confidence intervals in the way that they provide a confidence interval after every received sample such that we control the probability that the true parameter is contained in *all* the confidence intervals *simultaneously*.

Definition 2.5 (Confidence sequence)**.** Let $\{u_t, v_t\}_{t=1}^{\infty}$ be mappings from a sequence of observations to a real number. They form a confidence sequence $I_t(\mathbf{x}_{:t}) = [u_t(\mathbf{x}_{:t}), v_t(\mathbf{x}_{:t})]$ for all $t \geq 1$ with confidence level $1 - \alpha$ if

$$\mathbb{P}\left(p \in I_t(\mathbf{X}_{:t}), \forall t > 1\right) \geq 1 - \alpha$$

for any $p \in [0, 1]$, where $\mathbf{X}$ is an infinite sequence of Bernoulli random variables $\mathcal{B}(p)$.

Remark 2.6. Since we want the estimated parameter to be contained in all the confidence intervals simultaneously, we will have by convention that $I_{t+1}(x_{:t+1}) \subseteq I_t(x_{:t})$.

To simplify presentation, we would consider the symmetric ones; i.e., those where we consider the two possible failures - when we overestimate or underestimate the mean - to be equal. However, they will be constructed from two one-sided bounds, so the generalization is straightforward and will not be discussed.

Related work: While we are not aware of any work that contains precisely the result developed in the paper, the used techniques in the betting part are in the spirit of Orabona & Jun (2023); Orabona (2019); Howard et al. (2020); Waudby-Smith & Ramdas (2023). The differences mainly lies in the fact that we are interested in Bernoulli random variables, which allows us to use specialized tools at place, as opposed to the referred works which usually consider bounded random variables. As an analogy to confidence intervals, the previous works construct Bernstein-type inequalities, while we construct a Clopper-Pearson-like bounds. We refer specially to Orabona & Jun (2023) containing more general result which was a great inspiration for us and it also contains a proper literature review on the topic. The construction of confidence sequences based on union bound employs the doubling trick which is widely used in online learning to convert fixed-horizon algorithms to anytime algorithms. In this direction, we refer to Mnih et al. (2008) as the direct predecessor of this work.

2.3 Union bound confidence sequence

A natural way how to extend the confidence intervals to confidence sequences is to construct a confidence interval at every time step and use a union bound to control the total failure probability. In

the following, we first show that a naive application of this approach is asymptotically suboptimal, and then we provide a way how to construct optimal confidence intervals in a certain strong sense.

Intuition on the width of confidence sequences: For any random variable with finite variance, the optimal width of the confidence interval for the mean parameter scales as $\sqrt{\log(1/\alpha)/t}$ with the increasing number of samples t at confidence level $1-\alpha$ Lugosi & Mendelson (2019). On the other hand, it is well known that the width of the optimal confidence sequence scales as $\sqrt{(\log(1/\alpha)+\log\log t)/t}$ as t increases due to the law of iterated logarithms Ledoux & Talagrand (1991). A naive use of union bound, computing a confidence interval using failure probability at time step t, $\alpha_t = \frac{\alpha c}{t^\gamma}$ for some c and $\gamma > 1$ such that $\sum_{i=1}^{\infty} \alpha_t = \alpha$ yields a confidence sequence whose width scales as $\sqrt{\log(1/\alpha_t)/t} \approx \sqrt{(\log(t)+\log(1/\alpha))/t}$. We cannot choose any monotonous α_t schedule decaying slower because even for $\gamma = 1$ we still keep the $\log$ factor while the sum $\sum_{i=1}^{\infty} \delta_t$ diverges.

Now consider non-monotonous schedules of α_t, two key ideas follows. (1) In order to have the optimal rate $\log(1/\alpha_t) \approx \log(1/\alpha) + \log\log t$, we need $\alpha_t \asymp \alpha/\log t$. Clearly, if this holds for all t, then $\sum_{t=1}^{\infty} \alpha_t$ diverges. (2) A confidence interval at time t is also a valid confidence interval for all $t' > t$. Furthermore, if t' is not much larger than t, then it may still asymptotically have the optimal width up to a multiplicative constant. Thus, updating the confidence sequence when t is a power of (say) 2 result in the optimal width. This reasoning is formalized in the following theorem.

Theorem 2.7. *Fix $\alpha > 0$. Consider a sequence*

$$\alpha_t = \begin{cases} \frac{\alpha}{k(k+1)} & \text{if } t = 2^k \text{ for integer } k, \\ 0 & \text{otherwise.} \end{cases}$$

Then Algorithm 1 produces a confidence sequence at level $1-\alpha$ of the following width which is attained in the worst case

$$\varepsilon_t \lesssim \sqrt{\frac{\log(1/\alpha)+\log\log(t)}{t}}$$

where ε_t is $U - L$ at time t, and the confidence intervals are randomized Clopper-Pearson intervals.

The proof is in Appendix E. We remark that the statement of Theorem 2.7 prioritizes simplicity over its full generality. The generalization to other schedules of α_t from the second bullet point as described in the previous paragraph is routine and described in detail in Appendix C.2.

Corollary 2.8. *The asymptotic rate of 2.7 is optimal due to law-of-iterated-logarithm (Ledoux & Talagrand, 1991) and even in the finite sample regime due to Balsubramani (2014)[Theorem 2].*

2.4 Confidence sequences based on betting

A recent alternative approach to confidence sequences is based on a hypothetical betting game. For the illustration, consider a fair sequential game; e.g., sequentially betting on outcomes of a coin. If we guess the outcome correctly, we win the staked amount, otherwise we lose it. If the coin is fair, in expectation, our wealth stays the same. On the other hand, if the game is not fair and the coin is biased, we can win money. For example, if the true head-probability is 0.51, we start increasing our wealth in an exponential fashion, see Example A.3; thus, if we win lots of money, we can conclude that the game is not fair. We instantiate a betting game for every possible mean $0 \le p \le 1$ that would be fair if the true mean is p. Then we observe samples of the random variable and as soon as we win enough money, we drop that particular p from the confidence sequence. To make things formal, we introduce the necessary concepts from probability theory. The evolution of our wealth throughout a fair game is modeled by martingales[4], sequences of random variables for which, independently of the past, the expected value stays the same.

Definition 2.9 (Martingale)**.** A sequence of random variables $W_1, W_2, \ldots$ is called a *martingale* if for any integer $n > 0$, we have $\mathbb{E}(|W_n|) < \infty$ and $\mathbb{E}(W_{n+1}|W_1, \ldots, W_n) = W_n$. If we instead have $\mathbb{E}(W_{n+1}|W_1, \ldots, W_n) \le W_n$, then the sequence is called a *supermartingale*.

[4]It would be historically accurate to say that martingales actually model fair games.

Algorithm 1 Union-Bound Confidence Sequence

$T, H, K, L, U \leftarrow 0, 0, 0, 0, 1$
loop
  Obtain random x
  $H \leftarrow H + x$
  $T \leftarrow T + 1$
  if $T = 2^K$ **then**
    $K \leftarrow K + 1$
    $\alpha_T \leftarrow \alpha/(K(K+1))$
    $L \leftarrow \max\{L, \text{LowConfInt}(H, T, \alpha_T)\}$
    $U \leftarrow \min\{U, \text{UppConfInt}(H, T, \alpha_T)\}$
  end if
end loop

Algorithm 2 Betting Confidence Sequence

$\text{LOGQ}, T, H, L, U \leftarrow 0, 0, 0, 0, 1$
loop
  $\hat{q} \leftarrow (H + 1/2)/(T + 1)$
  Obtain random x
  $H \leftarrow H + x$
  $T \leftarrow T + 1$
  $\text{LOGQ} \leftarrow \text{LOGQ} + x \log(\hat{q}) + (1-x)\log(1-\hat{q})$
  $\text{LOGP}(p) := H \log(p) + (T - H)\log(1 - p)$
  $I_p \leftarrow \{p | \text{LOGQ} - \text{LOGP}(p) \leq \log(1/\alpha)\}$
  $L \leftarrow \max\{L, \min I_p\}$
  $U \leftarrow \min\{U, \max I_p\}$
end loop

In the coin-betting example, $W_1, W_2, \ldots$ is a martingale where W_n represents our wealth after playing the game for n rounds. We stress that $W_t \geq 0$ for all $t > 0$. By convention, we will also have $W_1 = 1$. We further need a time-uniform generalization of Markov's inequality.

Proposition 2.10 (Ville's inequality Durrett (2010))**.** *Let $W_1, W_2, \ldots$ be a non-negative supermartingale. then for any real $a > 0$*

$$\mathrm{P}\left[\sup_{n \geq 1} W_n \geq a\right] \leq \frac{\mathbb{E}\left[W_1\right]}{a}.$$

Thus, whenever we play a game and earn a lot, we can — with high probability — rule out the possibility that the game is fair. So far, this is still an abstract framework. We have yet to design the betting game and the betting strategy and describe how to run the infinite number of games.

Betting game: Let[5] $0 < p < 1$. Consider a coin-betting game where we win $1/p$ (resp. $1/(1-p)$) multiple of the staked amount if we correctly predicted heads (resp. tails). If the underlying heads probability is p, then regardless of our bet - in expectation - we still have the same amount of money; thus, this game is fair. We identify heads and tails with outcomes $1, 0$ respectively.

Betting strategy: We deconstruct the betting strategy into the two sub-tasks: (1) If we know the underlying heads probability, we can design the optimal betting strategy for any criterion. (2) Estimate the heads probability. **First sub-task:** Let p define the betting game from the previous paragraph and q be the true heads probability; Optimally, bet q-fraction of wealth to heads and $1 - q$ fraction to tails. It maximizes the expected log-wealth, or equivalently, the expected growth-rate of our wealth and is also known as the Kelly Criterion. This is known to be optimal for the adaptive estimation task, see Wald (1947) under the name sequential-probability-ratio-test (SPRT). One might have expected the optimal criterion to optimize to be the expected wealth; however, the betting strategy maximizing the expected wealth suggest to bet all the money on one of the outcomes. This strategy, however, leads to an eventual bankruptcy almost surely and is not recommended in this context. **Second sub-task:** A natural choice is to use the running sample mean of the observations as the estimator of q. Unfortunately, this estimator would be either 0 or 1 after the first observations, so we go bankrupt whenever the observed sequence contains both outcomes. Thus, we use a "regularized" sample mean and after observing H times heads in a sequence of length T, we estimate $\hat{q} = (H+0.5)/(T+1)$. This is the MAP estimate of the mean with Beta$(1/2, 1/2)$ prior and is known as Krichevsky–Trofimov estimator (Cesa-Bianchi & Lugosi (2006) Section 9.7), Krichevsky & Trofimov (1981) and is proven to be successful for building confidence sequences beyond the Bernoulli case Orabona & Jun (2023).

Parallel betting games: We have described a betting game for a certain p and a betting strategy. Employing Ville's inequality we can possibly reject the hypothesis that the true sampling distribution follows p. However, we need to run the game for all values of $p \in [0, 1]$ and after every observation report the smallest interval containing all the values of p that were not rejected so far. This is clearly impossible to do explicitly, but it turns out that the non-rejected values of p form an interval and we can use binary search twice to find the end-points of the interval. This is a non-trivial result

[5]For simplicity of exposition, similar arguments hold when $p \in \{0, 1\}$.

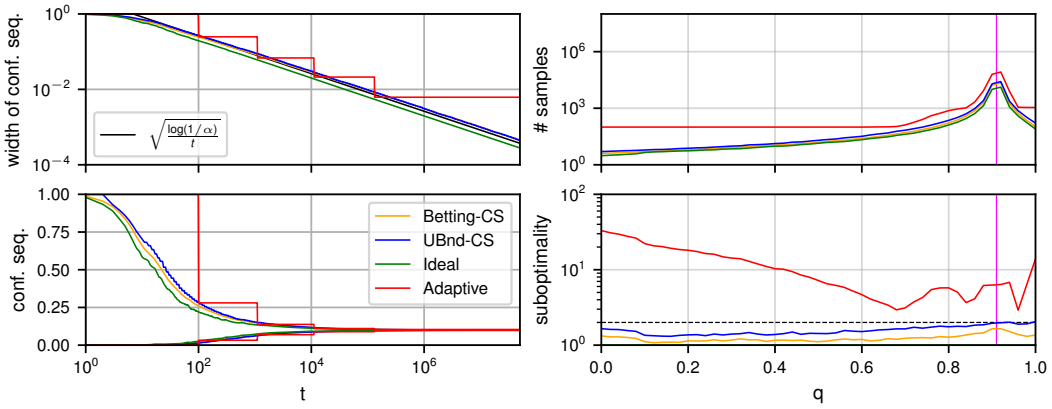

Figure 2: **left:** Comparison of widths of confidence sequences for the mean of Bernoulli $\mathcal{B}(0.1)$ with $\alpha = 0.001$. The width is on top and the actual confidence sequence on the bottom. In the notation of Algorithms 1 and 2, the sequence of $U - L$ is in the top figure, while both sequences U and L are in the bottom figure. Note the log-scale for t (and width on top). **right** : Instantiation of Task 3.1. The goal is to decide if $p = 0.91$ (vertical magenta line) or not with $\alpha = 0.001$. On top are the numbers of samples requested for the individual methods averaged over 1000 trials for 51 equally spaced values of $p \in [0, 1]$; on the bottom is the relative suboptimality of the individual methods; i.e., how many times more samples did they request compared to the ideal method. Note log scales on the $y-$axis. **methods:** UBnd-CS and Betting-CS are from Algorithm 1 and 2 respectively. Adaptive is from Horváth et al. (2022). The ideal is the unattainable lower-bound for the two tasks. On the LHS, it is a confidence interval on level $1 - \alpha$ computed independently at every time step. On the RHS, it is SPRT knowing both p, q which is optimal due to Wald (1947).

and generally does not need to hold for confidence intervals constructed by betting. The first key observations is that the betting strategy does not depend on p, so we can "play the game" just once. The second observation is that the resulting wealth is convex in p. To see why, let $\hat{q}_t$ (resp. x_t) be our estimate of q (resp. the coin-toss outcome, for brevity $0/1$ corresponds to heads/tails respectively and $H = \sum_{t=1}^{T} x_t$), then our log-wealth at time T can be written as a function of p.

$$\begin{aligned}\log W_T(p) &= \log \prod_{t=1}^{T} \left(\left(\frac{\hat{q}_t}{p} \right)^{x_t} \left(\frac{1-\hat{q}_t}{1-p} \right)^{1-x_t} \right) \\ &= \underbrace{\sum_{t=1}^{T} x_t \log(\hat{q}_t) + (1-x_t)\log(1-\hat{q}_t)}_{\text{LOGQ}} \underbrace{-H\log(p) - (T-H)\log(1-p)}_{\text{LOGP}(p)}.\end{aligned}$$

Therefore, at every time-step, we can compute the interval of values of p for which the betting game has not concluded yet and thus form the current confidence interval. The proof is in Appendix F.

Theorem 2.11. *Algorithm 2 produces a valid confidence sequence at confidence level $1 - \alpha$, where we interpret the interval $[L, U]$ at iteration T of the algorithm as the confidence interval at time T with width ε at most (which is attained in the worst case):*

$$\varepsilon_t \lesssim \sqrt{\frac{\log(1/\alpha) + \log(t)}{t}}.$$

This is not asymptotically optimal; still, empirically it performs well, and the same techniques can be used to obtain a confidence sequence that follows law-of-iterated-logarithms Orabona & Jun (2023). We conclude this subsection by comparing the presented confidence sequences in Figure 2.

3 Experiments

Now we shall provide experimental evidence for the performance of the proposed methods. We benchmark the confidence sequences on the Sequential decision making task, where we try to certify

	r=0.5	r=1.25 ,	r=2
Adaptive Horváth et al. (2022)	1976 ± 41	3593 ± 574	4623 ± 47
Betting CS 2	531 ± 157	2169 ± 257	2130 ± 339
Union bound CS 1	635 ± 157	2557 ± 234	2670 ± 271
Adaptive Horváth et al. (2022)	0.13 ± 0.006s	0.23 ± 0.036s	0.3 ± 0.003s
Betting CS 2	0.05 ± 0.012s	0.17 ± 0.019s	0.17 ± 0.02s
Union bound CS 1	0.05 ± 0.006s	0.19 ± 0.018s	0.21 ± 0.02s

Table 1: Comparison of the average number of samples (resp. time) needed to decide if a point is certifiably robust with given radius. Cifar10, ℓ_2, details are in Appendix C.2.1

a certain radius at given confidence level with as few samples as possible; the definition that follows is general beyond randomized smoothing. We emphasize that this setting of certifying a certain radius is by far the most common one in the robustness literature. We stress that the comparison of the robustness curves (e.g., as in Figure 1) is vacuous, since in the adaptive task, we do not spend samples to *improve* the robustness curves, beyond the certified level.

Definition 3.1 (Sequential decision making task)**.** Let $\frac{1}{2} \leq p, q \leq 1$ and only p being known. Receive samples from $\mathcal{B}(q)$. After every sample, either halt and declare that $p > q$, $p < q$, or request another sample. The task is to minimize the number of samples while being wrong with frequency at-most α.

3.1 Related work

We identified Horváth et al. (2022) as the most relevant work. They distinguish between samples for which a predetermined radius r can be certified, and the samples for which it cannot. They use s values $n_1 < \cdots < n_s$ ($[10^2, 10^3, 10^4, 1.2 \cdot 10^5]$) sequentially as the number of samples. They try to certify radius r with n_1 samples; if it fails, then they try n_2 samples etc. They employ Bonferroni correction (union bound) and every sub-certification is allowed to fail with probability only $\frac{\alpha}{s}$. The key differences (details in Appendix C.2.1) to our method are that (1) It always abstains for hard tasks. (2) Splitting the α budget evenly degrades performance for small n. (3) method is only a heuristics. See Figure 5 and Tables 1, 3, 4. For the empirical comparison.

Another relevant work is Chen et al. (2022). Here, the certification is split in two phases. (1) Mean is crudely estimated. (2) The crude estimate selects the number of samples drawn so that the decrease (either multiplicative or absolute) in the certified radius is heuristically approximately at most a predetermined constant. We note that this heuristic for distributing samples can be made rigorous in a certain sense (see Dagum et al. (1995)). This is trivial for confidence sequences, as one can stop the estimation only as soon as they short enough and solve the task of Chen et al. (2022) with guarantees (instead of just heuristic). In this sense, we see our methods to be more general. We benchmark this in Table 2.

The works Seferis et al. (2023); Ugare et al. (2024) also address the speed issues of randomized smoothing, however, they are orthogonal to our directions. In particular, Ugare et al. (2024) uses an auxiliary network for which the certification is faster and transfer the certificates to the original model. Seferis et al. (2023) observes that few samples are sufficient for non-trivial certificates.

ε	0.01	0.02	0.03
UB-CS	197 628	49 198	21 513
Betting-CS	199 771	47 215	20 918
Horvath	768 560	94 900	81 080

Table 2: We run the confidence sequences until the width is smaller than ε on a (both sided) confidence level 0.999. That way we can certify certain radius knowing that the true probability is at most ε larger. We used the same network as for the ℓ_2 experiment in Table 1 (WideResnet-40 on CIFAR10, $\sigma = 1$). We report the average number of samples required over 500 images.

4 Conclusion

In this paper, we investigated the statistical estimation procedures related to randomized smoothing and improved them in the following two ways: (1) We have provided a strictly stronger version of confidence intervals than the Clopper-Pearson confidence interval. (2) We have developed confidence sequences for sequential estimation in the framework of randomized smoothing, which will greatly reduce the number of samples needed for adaptive estimation tasks. Additionally, we provided matching algorithmic upper bounds with problem lower bounds for the relevant statistical estimation task.

5 Broader Impact Statement

We hope that this paper enlarges the interest in statistical estimation within the ML community.

Acknowledgments

The author was supported by the DFG Cluster of Excellence "Machine Learning – New Perspectives for Science", EXC 2064/1, project number 390727645 and is thankful for the support of Open Philanthropy.

References

Balsubramani, A. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*, 2014.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *ECML PKDD*, 2013.

Bojchevski, A., Gasteiger, J., and Günnemann, S. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *ICML*, 2020.

Boucheron, S., Lugosi, G., and Massart, P. Concentration inequalities. 2013.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chen, R., Li, J., Yan, J., Li, P., and Sheng, B. Input-specific robustness certification for randomized smoothing. In *AAAI*, 2022.

Clopper, C. J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 1934.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.

Dagum, P., Karp, R., Luby, M., and Ross, S. An optimal algorithm for monte carlo estimation. In *FOCS*, 1995.

Durrett, R. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics, 2010.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *ICLR*, 2014.

Horváth, M. Z., Müller, M. N., Fischer, M., and Vechev, M. Boosting randomized smoothing with variance reduced classifiers. *ICLR*, 2022.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 2020.

Krichevsky, R. and Trofimov, V. The performance of universal encoding. *IEEE Transactions on Information Theory*, 1981.

Kumari, A., Bhardwaj, D., Jindal, S., and Gupta, S. Trust, but verify: A survey of randomized smoothing techniques. *arXiv preprint arXiv:2312.12608*, 2023.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *S&P*, 2019.

Ledoux, M. and Talagrand, M. *The Law of the Iterated Logarithm*, pp. 196–234. 1991.

Lee, G.-H., Yuan, Y., Chang, S., and Jaakkola, T. Tight certificates of adversarial robustness for randomly smoothed classifiers. *NeurIPS*, 2019.

Levine, A. and Feizi, S. (de) randomized smoothing for certifiable defense against patch attacks. *NeurIPS*, 2020.

Levine, A. J. and Feizi, S. Improved, deterministic smoothing for ℓ_1 certified robustness. In *ICML*, 2021.

Li, L., Xie, T., and Li, B. Sok: Certified robustness for deep neural networks. In *S&P*. IEEE, 2023.

Lugosi, G. and Mendelson, S. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Mnih, V., Szepesvári, C., and Audibert, J.-Y. Empirical bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pp. 672–679, 2008.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Orabona, F. and Jun, K.-S. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 2023.

Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. In *NeurIPS*, 2019.

Seferis, E., Burton, S., and Kollias, S. Randomized smoothing (almost) in real time? In *ICMLw The Many Facets of Preference-Based Learning*, 2023.

Stevens, W. L. Fiducial limits of the parameter of a discontinuous distribution. *Biometrika*, 1950.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *ICLR*, 2014.

Ugare, S., Suresh, T., Banerjee, D., Singh, G., and Misailovic, S. Incremental randomized smoothing certification. In *ICLR*, 2024.

Voracek, V. and Hein, M. Improving l1-certified robustness via randomized smoothing by leveraging box constraints. In *International Conference on Machine Learning*, 2023.

Wald, A. *Sequential analysis*. 1947.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B Methodological*, 2023.

Yang, G., Duan, T., Hu, J. E., Salman, H., Razenshteyn, I., and Li, J. Randomized smoothing of all shapes and sizes. In *ICML*, 2020.

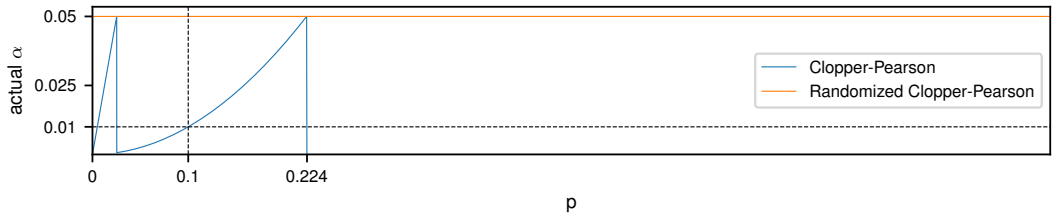

Figure 3: Actual coverages for (randomized) Clopper-Pearson confidence intervals for $\mathcal{B}(2, p)$.

A Deferred examples

Example A.1. [implicit type-2 error is exponentially small] Let us sample from $X \sim \mathcal{B}(p)$ to decide if the mean is smaller or larger than $p - \varepsilon$ using $n = 100\,000$ samples. We use Hoeffdings' inequality to bound the probability that p is incorrectly estimated to be lower than $p - \varepsilon$, i.e.,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \leq \mathbb{E}[X] - \varepsilon\right] \leq e^{-2n\varepsilon^2}.$$

Considering ε to be a constant, we see that this probability scales as e^{-n}. For example, when the true probability is 0.5 and we want to decide if it is smaller or larger than 0.4, already 1000 samples make the probability of incorrectly decision to be roughly $2 \cdot 10^{-9}$.

Example A.2. [suboptimality of Clopper-Pearson confidence interval and optimality of the randomized one] Recall that the coverage for p is the probability that p is included in the confidence interval when it is the true parameter and α is the allowed type-1 error and $1 - \alpha$ should be the coverage. Consider samples from $X \sim \mathcal{B}(2, p)$ and $\alpha = 0.05$. By definition, the Clopper-Pearson upper intervals are $[0, 1]$, $[0.025, 1]$, $[0.224, 1]$ for observations $x = 0, 1, 2$ respectively. Coverage for $p = 0.224$ is 0.95 because the event that p is outside of the confidence interval is $\mathbb{P}(X \in \{0, 1\}) = 1 - p^2 \approx 0.95$. On the other hand, coverage for $p = 0.1$ is $\mathbb{P}(X \in \{0, 1\}) = 1 - p^2 = 0.99$ and for all $p > 0.224$ it is 1; see Figure A for the coverages. Now we turn on to the randomized Clopper-Pearson interval for $p = 0.5$. Recall the definition of the upper interval,

$$u'_r(x, w) = \inf\{p \,|\, \mathbb{P}(\mathcal{B}(n, p) > x) + w\mathbb{P}(\mathcal{B}(n, p) = x) > \alpha\}.$$

The randomized confidence interval for some value x interpolates between the confidence intervals for x and $x + 1$ when $x < n$. Thus, when $x \neq 2$, $p > 0.224$ is always in the confidence interval (this happens with probability $1 - p^2 = 0.75$. Otherwise, we solve the following for w (because we set $n = 2$, $x = 2$):

$$\begin{aligned}\mathbb{P}(\mathcal{B}(2, p) > 2) + w\mathbb{P}(\mathcal{B}(2, p) = 2) &= \alpha,\\ 0 + p^2 w &= \alpha,\\ v &= \alpha/p^2.\end{aligned}$$

Thus, $p > 0.224$ is not contained in the randomized confidence interval iff $W \leq \alpha/p^2$ and $X = 2$. Since these random variables are independent, the resulting probability is $\mathbb{P}(W \leq 0.2)\mathbb{P}(X = 2) = \alpha/p^2 \cdot p^2 = \alpha$, as desired.

Example A.3 (Exponential increase of wealth for a biased coin)**.** For simplicity of exposition we assume that the coin falls on head 51 times from 100 tosses. To get a high probability statement is straightforward. Let us always bet 0.51 fraction of our money to to heads and 0.49 to tails (equivalently, just bet 0.02 of the money to the heads). If we win, we win 2% of our money, otherwise we lose 2%. Then, our wealth after 100 tosses will be $1.02^{51} \cdot 1.02^{-49} \sim 1.04$. Thus, every 100 tosses we multiply our wealth by the factor of 1.04 which is the desired exponential function.

B Binary or multiclass certification

Although all the standard benchmarking datasets for randomized smoothing are multiclass (cifar10 and imagenet), the commonly used randomized smoothing certification protocol is for the binary

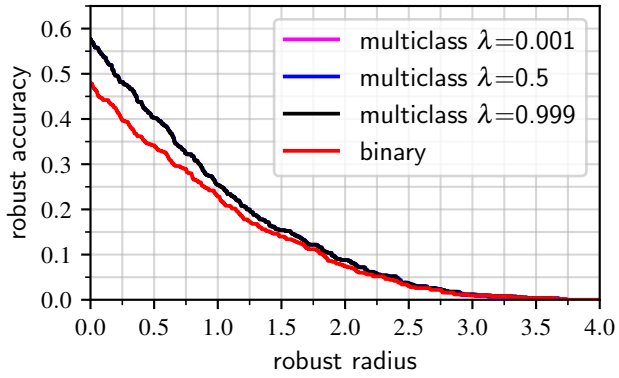

Figure 4: Comparison of the robustness curves for binary and multiclass certification. In the binary case, all the failure budget $\alpha = 0.001$ was spent on controlling the top-1 class probability. In the multiclass setting, we spend λ fraction of the budget in bounding p_A and the remaining $1 - \lambda$ part on bounding p_B. Note that this has no significant effect. The average certified radius for binary certification is 0.50, while for the multiclass it is 0.61. The experimental details are in Appendix C. The only difference is that now $\sigma = 1$.

setting, where we certify class A against all the other classes merged in a super class. In that case, the certification is done using the formula $r(\hat{p}_A, 1 - \hat{p}_A)$, where we have to guarantee that $p_A \geq \hat{p}_A$ at confidence level at least $1 - \alpha$. The alternative is to use multiclass certification; here, the certification is done via formula $r(\hat{p}_A, \hat{p}_B)$ ensuring that $p_A \geq \hat{p}_A$ and $p_B \leq \hat{p}_B$ at the same time at confidence level at least $1 - \alpha$. The difference between these two bounds naturally manifests in the regime when p_A is small. Strikingly, when $p_A < 0.5$, the binary certification approach cannot certify any robustness, while the multiclass one possible can. The cost for the multiclass procedure is only that we have to divide the failure budget between the the two estimation procedures. This is usually insignificant. In the ℓ_2 (and thus ℓ_∞ case), the role of α is rather minor, see Cohen et al. (2019) Figure 8. This is even more pronounced in the ℓ_1 case. Here α plays an absolutely negligible role in the resulting certified radius, see Voracek & Hein (2023), Subsection 2.7 for the discussion and Figures 4, 6. While this might be known to many, we believe that some readers may benefit from reading this argument. We demonstrate the difference in certification power in Figure 4. This multiclass certification fits in our setting effortlessly. We can run one confidence sequence for p_A, and another for p_B. We do not even need to know p_B and we can run it for all of them at the same time. This means, that we run it only for the second most observed class. This second most observed class does not need to be the actual runner-up class, but since it was possibly observed more times than the actual runner-up class, it will also provide a wider confidence interval, so the statistical estimation is still correct.

C Experimental details

C.1 Figure 1

The model in Figure 1 is the pretrained cifar10 model (Exactly the same model/setting from the example in README) of Salman et al. (2019), `https://github.com/Hadisalman/smoothing-adversarial`; in particular, it was ResNet-110 smoothed with Gaussian noise $\sigma = 0.12$ for ℓ_2 robustness. We set $\alpha = 0.001$ as usual and skip every 20 images of the test dataset (using 500 images, as is the standard practice).

C.2 Parameters of union bound confidence sequences

We were enlarging the sample size by a factor of $\beta = 1.1$ between estimations (that is, the condition $T = 2^K$ is replaced by $T > \beta^K = 1.1^K$ and our schedule is $\alpha_k = 5\alpha/((t+4)(t+5))$ The method is not sensitive to the choice of hyperparameters, see 5. In fact, the hyperparameters β, γ should be selected based on what is the "interesting" regime. There is an inherent tradeoff between a good asymptotical performance ($\beta, \gamma \sim 1$) and low-sample performance ($\beta, \gamma > 1$). This is confirmed in

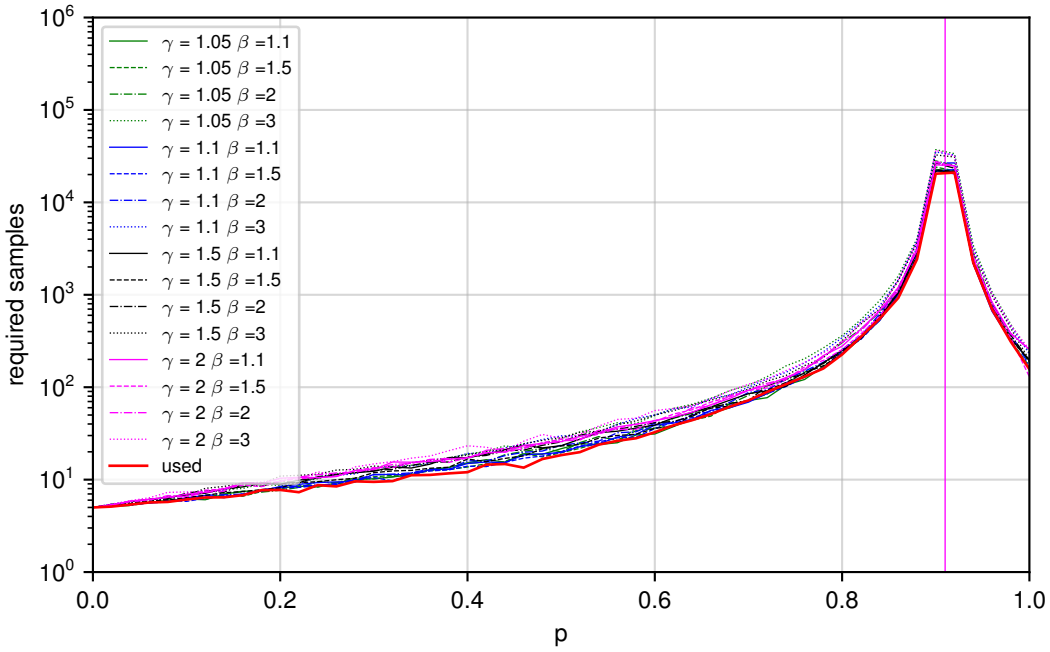

Figure 5: Samples needed for the adaptive estimation task as in Figure 2 for different hyperparameters. β is the factor by which we enlarge the sample size before computing new confidence interval, γ is the scaling of α as described in the main text. I.e., $k-$th estimation will have $\alpha_k = \frac{\alpha c}{k^\gamma}$ where c is the normalization constant such that $\sum_{k=1}^{\infty} \alpha_k = \alpha$.

Figure 5. The width of the confidence interval scales as:

$$\sqrt{\frac{\beta(\log(1/(\gamma-1)) + \log(1/\alpha) + \gamma \log\log_\beta t)}{t}}.$$

This is because $\sum_{t=1}^{\infty} \frac{1}{t^\gamma} \asymp \frac{1}{\gamma-1}$, $\alpha_k \asymp \frac{\alpha(\gamma-1)}{k^\gamma}$ and $t \asymp \beta^k$, then $k \asymp \log_\beta t$. Plugging in these identites in $\sqrt{\frac{\log(\alpha_t)}{t}}$, and remembering that the confidence interval is recomputed only after elarging the sample size by β factor, then for time t, the actual t we use in the formula can be as small as t/β.

C.2.1 Comparison with Horváth et al. (2022)

The method from Horváth et al. (2022) uses a finite collection of values of $n = n_1 < \cdots < n_s$ for which the confidence intervals are computed. A direct consequence is that the hard examples for which more than n_s samples are needed cannot be certified. Additionally, due to Bonferroni correction, with large s, the term $\log(s/\alpha)$ appearing in the confidence interval becomes large compared to t for small values of t (compare with the polynomial scaling that we propose where this is not the case). Consider $n_i = n_1^i$, then the width of the confidence interval scales as (not considering the regime when $t > n_s$ where the width is constant):

$$\sqrt{\frac{n_1(\log s + \log 1/\alpha)}{t}}.$$

In particular, when we want to have confidence sequence up $n_s = N$ samples, then the width becomes

$$\sqrt{\frac{\sqrt[s]{N}(\log s + \log 1/\alpha)}{t}}$$

and we either pay for the fact that we have large differences between the steps ($\sqrt[s]{N}$), or for the fact that we have lot of steps ($\log(s)$) which is detrimental for small values of t.

C.3 Details for 1, 3, 4

We had WideResNet-40-2 for CIFAR-10 trained for 120 epochs with SGD and learning rate 0.1, Nesterov momentum 0.9, weight decay 0.0001 and cosine annealing. batch size 64. The loss was the standard, noise-augmented training using the same noise as for the certification. We either used Gaussian smoothing for ℓ_2 robustness of uniform in ℓ_∞ box for ℓ_1 robustness.

For the certification we used batch size 100 (natural for Horváth et al. (2022)). For our method, we had a mixed batch of data points so the data points for which we have the least amount of samples and are not decided yet are put in the batch.

D Proof of Proposition 2.4

Proof. We show it for the upper interval, the lower is analogical. First, we note that $f(p) = \mathbb{P}(\mathcal{B}(n,p) \geq a)$ is non-decreasing in p for any n, a; Thus, $u_r(X) \leq p$ if and only if $\mathbb{P}(\mathcal{B}(n,p) > x) + w\mathbb{P}(\mathcal{B}(n,p) = x) > \alpha$. Now, we show that $\mathbb{P}(u_r(X) \leq p) = 1 - \alpha$ for $X \sim \mathcal{B}(n,p)$. To shorten the notation, let $\alpha_1 = \mathbb{P}(X \geq a)$ and $\alpha_2 = \mathbb{P}(X > a)$ for such a that $\alpha_1 \leq \alpha \leq \alpha_2$. We have

$$\begin{aligned}\mathbb{P}(u_r(X) \leq p) = &\mathbb{P}(u_r(X) \leq p \mid X < a)\mathbb{P}(X < a)\\ &+ \mathbb{P}(u_r(X) \leq p \mid X = a)\mathbb{P}(X = a)\\ &+ \mathbb{P}(u_r(X) \leq p \mid X > a)\mathbb{P}(X > a),\end{aligned}$$

which we evaluate to $\mathbb{P}(u_r(X) \leq p) = (1 - \alpha_1) + (\alpha_1 - \alpha_2)\mathbb{P}(u_r(X) \leq p \mid X = a) + 0$. We note that given the event $X = a$, it holds $u_r(X) \leq p \iff \alpha_2 + W(\alpha_1 - \alpha_2) > \alpha$, so $\mathbb{P}(u_r(X) \leq p \mid X = a) = \mathbb{P}(\alpha_2 + W(\alpha_1 - \alpha_2) > \alpha) = \mathbb{P}\left(W > \frac{\alpha - \alpha_2}{\alpha_1 - \alpha_2}\right) = \frac{\alpha_1 - \alpha}{\alpha_1 - \alpha_2}$. Overall, $\mathbb{P}(u_r(X) \leq p) = (1 - \alpha_1) + (\alpha_1 - \alpha_2)\frac{\alpha_1 - \alpha}{\alpha_1 - \alpha_2} = 1 - \alpha$.

The second part of the statement follows from Neymann-Pearson lemma. Concretely, we consider the following binary hypothesis testing problem from sample $\mathcal{B}(n,\theta)$, and we decide if $\theta = p$ or $\theta = q$. Both confidence intervals has size α and can be interpreted as binary tests – just return the indicator function of $q \in I(x)$. Neymann-Pearson lemma states that the (unique) uniformly most powerful test is the likelihood ratio test, which is implemented by the randomized Clopper-Pearson interval.

□

E Proof of Theorem 2.7

Proof of Theorem 2.7. First, $\sum_{i=1}^{\infty} \alpha_t = \sum_{k=1}^{\infty} \frac{\alpha}{k(k+1)} = \alpha$; thus, by union bound, all the computed confidence intervals are simultaneously correct at confidence level $1 - \alpha$. Next we show that the width is as claimed. When $t = 2^k$, we directly have from Hoeffding's inequality

$$\varepsilon \lesssim \sqrt{\frac{\log \frac{1}{\alpha_t}}{t}} \asymp \sqrt{\frac{\log \frac{1}{\alpha} + \log\log t}{t}}.$$

Otherwise, we would use a confidence interval of some previous t' such that $t' < t < 2t'$ with width

$$\varepsilon \lesssim \sqrt{\frac{\log \frac{1}{\alpha} + \log\log t'}{t'}} \asymp \sqrt{\frac{\log \frac{1}{\alpha} + \log\log t}{t}},$$

Noting that Clopper-Pearson's confidence intreval is shorter than Hoeffding's finishes the proof. □

F Proof of Theorem 2.11

Proof. First we verify that everything in the algorithm is well defined and the logarithms take positive inputs. Now let $W_t = \exp\left(\frac{\text{LOGQ}_t}{\text{LOGP}_t}\right)$ where subscript t denotes iteration of the algorithm. We show that it is a martingale when $X \sim \mathcal{B}(p)$ for $0 < p < 1$. In that case,

$$\mathbb{E}_X[W_t] = W_{t-1}\mathbb{E}_X\left[\left(\frac{\hat{q}_t}{p}\right)^X \left(\frac{1-\hat{q}_t}{1-p}\right)^{1-X}\right] = W_{t-1}\left(p\frac{\hat{q}}{p} + (1-p)\frac{1-\hat{q}}{1-p}\right) = W_{t-1}.$$

	r=0.5	r=1.25 ,	r=2
Adaptive Horváth et al. (2022)	1976 ± 41	3593 ± 574	4623 ± 47
Betting CS 2	531 ± 157	2169 ± 257	2130 ± 339
Union bound CS 1	635 ± 157	2557 ± 234	2670 ± 271
Adaptive Horváth et al. (2022)	0.13 ± 0.006s	0.23 ± 0.036s	0.3 ± 0.003s
Betting CS 2	0.05 ± 0.012s	0.17 ± 0.019s	0.17 ± 0.02s
Union bound CS 1	0.05 ± 0.006s	0.19 ± 0.018s	0.21 ± 0.02s
	r=0.5	r=1.25 ,	r=2
Adaptive Horváth et al. (2022)	4150 ± 523	2266 ± 640	4760 ± 131
Betting CS 2	2206 ± 150	1665 ± 84	3932 ± 199
Union bound CS 1	2665 ± 78	1674 ± 37	3717 ± 361
Adaptive Horváth et al. (2022)	0.27 ± 0.04s	0.15 ± 0.04s	0.3 ± 0.009s
Betting CS 2	0.17 ± 0.011s	0.13 ± 0.006s	0.3 ± 0.014s
Union bound CS 1	0.2 ± 0.005s	0.13 ± 0.002s	0.28 ± 0.02s

Table 3: Comparison of the average number of sample needed to decide if the point is certifiably robust with given radius in the top. The time needed is on the bottom. The upper table was in the main paper, the bottom one is the exact same experiment but with a retrained model for ℓ_2 robustness robustness with $\sigma = 1$.

If $p \in \{0, 1\}$, then we would have a deterministic sequence and $W_t \leq W_{t-1}$. Thus, W_t is a supermartingale. It is also output of the exponential function and so is non-negative. Therefore, the assumptions of Ville's inequality are satisfied and can be applied. Whenever p is excluded from the confidence interval, it happened that $\text{LOGQ}_t - \text{LOGP}_t \geq \log(1/\alpha)$, or equivalently, $W_t \geq 1/\alpha$ which can only happen with probability α and it is thus a valid confidence sequence. We also recall that in the main text we have shown that I_p is a sub-level set of a convex function and is thus convex and can be efficiently found by binary search. The width of the confidence interval follows from the standard regret bounds for the Krichevsky–Trofimov estimator Cesa-Bianchi & Lugosi (2006) Section 9.7; Krichevsky & Trofimov (1981). The result then follows from Orabona (2019), Subsection 12.7.

□

We remark that $W(\mathbf{x}_{:t})$ does not depend on the order of $\mathbf{x}_{:t}$, so we write it as $W(h, t)$; thus, for every T, we can compute what is the minimal number $H(T)$ needed of observed 1 so that p is outside of the lower-confidence interval. $H(T)$ is clearly non-decreasing in T; also, $W(h, t)$ can be easily computed from $W(h - 1, t)$ and also from $W(h, t - 1)$ using few flops. The whole dynamic programming approach can be summarized in the following scheme:

- If $W(h, t) \geq \frac{1}{\alpha}$: $H(t) := h$, $t := t + 1$, compute $W(h, t + 1)$
- Else: $h := h + 1$, compute $W(h + 1, t)$.

Both lines are executed in constant time. Also, we start from $W(0, 0)$ and $h, t < N + 1$. Executing a line, either h or t increases and thus to compute the thresholds for sequence of length up to N, we need to execute at most $2N$ lines of the scheme. We note that for the simplicity, we did not handle the case when $W(h, t) < 1/\alpha$. In that case we would just set $H(t) = t + 1$ and increase t.

	r=0.5	r=1 ,	r=1.5
Adaptive Horváth et al. (2022)	423 ± 38	3520 ± 82	3823 ± 127
Betting CS 2	138 ± 28	2680 ± 221	3007 ± 122
Union bound CS 1	150 ± 3	2926 ± 18	3144 ± 347
Adaptive Horváth et al. (2022)	0.04 ± 0.006s	0.23 ± 0.006s	0.25 ± 0.009s
Betting CS 2	0.19 ± 0.002s	0.21 ± 0.017s	0.23 ± 0.01s
Union bound CS 1	0.016 ± 0.006s	0.22 ± 0.009s	0.25 ± 0.03s
	r=0.5	r=1 ,	r=1.5
Adaptive Horváth et al. (2022)	1370 ± 596	4790 ± 50	8463 ± 714
Betting CS 2	592 ± 94	4055 ± 459	5822 ± 81
Union bound CS 1	806 ± 113	4327 ± 106	5795 ± 321
Adaptive Horváth et al. (2022)	0.1 ± 0.04s	0.30 ± 0.003s	0.53 ± 0.05s
Betting CS 2	0.05 ± 0.007s	0.31 ± 0.03s	0.44 ± 0.005s
Union bound CS 1	0.06 ± 0.008s	0.33 ± 0.008s	0.44 ± 0.02s

Table 4: Comparison of the average number of samples needed to decide if the point is certifiably robust with given radius in the top. The time needed is on the bottom. top and bottom are again the exact same experiment but with a retrained model for ℓ_1 robustness with $\sigma = 1$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We promised treatment of statistical estimation problems for randomized smoothing and we delivered them.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We think we provided fair discussion of the results. For the provided theorems, the assumptions are reasonable. The experiments seem convincing to us.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide proofs in the appendices although the main paper should be enough to understand them.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide details in C. However, some of the experiments are heavily dependent on random seed so they are unlikely to be reproduced. The important thing there are not the absolute results, but the relative performance of the methods which should be robust.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: In linked repo.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In C, also in the repository.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provided averages over runs and reported standard deviations where appropriate.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The total compute is under a day on a single GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: just yes

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: yes

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: no need.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the code we used, but not the datasets as it is standard (cifar)

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: na

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: na

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: na

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.