Towards Evading the Theoretical Limitations of Randomized Smoothing

Anonymous Authors[1]

Abstract

Randomized smoothing is the dominant standard for designing provable defenses against adversarial examples. Nevertheless, it has recently been proven that current certificates suffer from information theoretic limitations, seeming to doom the method. In this paper, we investigate how to develop new approaches to circumvent these limitations. To do so, we first show that existing methods use too little information about the classifier, which results in severely sub-optimal robustness certificates when the decision boundary has a steep local curvature. Moreover, we show that this underestimation can become arbitrarily bad as the dimension grows, thus explaining the information theoretic limits of existing certificates.

Our main contribution lies in showing that we can circumvent this shortcoming by combining several noise-based queries on the classifier; essentially showing that we can gather perfect information on the decision boundary when the number of noises is large enough. Building upon this result, we introduce a new framework, based on the generalized Neyman-Pearson lemma, which allows to collect information from any number (and types) of noise as needed. This opens the door to designing methods that can provide asymptotically optimal certificates, thus circumventing the information theoretic limitations of Random smoothing at the cost of high computational cost. Finally, we provide some insights into methods for computing these certificates and how to overcome the computational difficulties they represent.

1. Introduction

Modern day machine learning models are vulnerable to adversarial examples, *i.e.*, small perturbations to their inputs that force misclassification (Szegedy et al., 2014; Goodfellow et al., 2015). Although a considerable amount of work has been devoted to mitigating their impact, defending a model against such malicious inputs remains genuinely difficult. *Randomized smoothing* has recently emerged as a dominant standard among existing defenses. It converts any base classifier into a more robust one through a convolution with a noise distribution, and has the advantage of providing provable, classifier agnostic guarantees of robustness in the neighborhood of each point. (Lecuyer et al., 2018; Li et al., 2019a; Cohen et al., 2019; Salman et al., 2019). It is also easy to implement and interpret, computationally efficient, and offers state-of-the-art provable robustness for multiple classification tasks.

However, current certification methods for Randomised smoothing suffer from significant impossibility results: the maximum radius that can be certified for a given smoothing distribution vanishes as the dimension increases (Hayes, 2020; Yang et al., 2020; Kumar et al., 2020; Mohapatra et al., 2021). Some recent works actually set out to to get around that impossibility, by collecting more information about the classifier. Specifically, Mohapatra et al. (2020) pioneered this approach, by leveraging higher-order derivatives to build a better certificate.

Our Contributions. We propose a theoretical certification framework to circumvent the current shortcoming of Randomized Smoothing. In this paper:

1. We quantify the gap between current certificates and the optimal bound for two classes of decision boundaries, and show that it can become arbitrarily bad as the input dimension increases.
2. We show that by collecting more information using noise-based queries only, it is possible to approximate the perfect certificate with arbitrary precision, and therefore bypass the current impossibility results.
3. We introduce a method, based on the generalized *Neyman-Pearson* lemma, to derive these certificates, and provide some techniques to make them more computationally viable.
4. Finally, we show that by introducing randomization in the certification process, it is possible to compute bounds that do not depend on the input dimension.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Background & Related Work

In this section, we review previous works on adversarial attacks and certified defenses. We focus on Randomized Smoothing, its strengths, weaknesses and recent work that attempts to limit its range of use.

Adversarial Attacks & Certified Defenses. Since the discovery of adversarial attacks (Globerson et al., 2006; Szegedy et al., 2014; Goodfellow et al., 2015), *i.e.*, small crafted perturbations to the input of neural networks that remain imperceptible to humans but lead to misclassification, multiple defenses strategies have been developed. At first, state-of-the-art empirical defenses have been devised (Goodfellow et al., 2015; Madry et al., 2018) leading to the development of stronger attacks (Carlini et al., 2019; Madry et al., 2018; Athalye et al., 2018) breaking these defenses. In order to break this cat-and-mouse game of attacks and defenses, provable methods are necessary (Pinot et al., 2020). Several works have proposed to design 1-Lipschitz neural network providing guarantees on the output of the classifier (Li et al., 2019b; Trockman et al., 2021; Singla & Feizi, 2021; Meunier et al., 2021; Singla et al., 2021). Indeed, provided that we know the Lipschitz constant of the networks and the margin of the prediction, it is easy to compute a certificate given a specific perturbation radius. Another popular approach is to compute a lower bound on the certificate via semidefinite relaxations (Raghunathan et al., 2018a; Wong & Kolter, 2018; Weng et al., 2018; Raghunathan et al., 2018b). Unfortunately, these types of defenses do not scale to large use cases (*i.e.*, ImageNet Dataset (Deng et al., 2009)). On the other hand, *Randomized Smoothing* has been developed as a noise-based certified defense and has the advantage to be architecture-independent and, therefore, can scale to large neural networks.

Randomized Smoothing. In optimization, several authors (Yousefian et al., 2012; Duchi et al., 2012) have noted that convolving the input with a predefined probability distribution would transform the initial function into a smoothed one. The idea behind this approach is that convolving two functions yields a resulting function which is at least as smooth as the smoothest of the two original functions. In the context of devising defense against adversarial examples, it is a natural idea to *smooth* the classifier making it less sensitive to input perturbations. At first, several noise-based defenses mechanisms have been proposed (Cao & Gong, 2017; Liu et al., 2018; Pinot et al., 2019) to defend against adversarial examples without the current theoretical guarantees of Randomized Smoothing. The first noise-based certified defenses was proposed by Lecuyer et al. (2018) (based on the differential privacy framework) and Li et al. (2019a). However, the first tight certificate based was proposed by Cohen et al. (2019). Building upon Cohen et al.'s work, Salman et al. (2019) combined Randomized Smoothing with Adversarial Training (Madry et al., 2018) in order to achieve better certified accuracy on well-known datasets. Finally, Yang et al. (2020) extend the work of Cohen et al.'s extensively by generalizing their approach to a lot of different smoothing distributions.

Limitations of Randomized Smoothing. However, in its current form, Randomized smoothing suffers from significant drawbacks. Several works have started to demonstrate the limitations of the approach with "impossibility" results for specific norms or settings. The first results on the hardness of Randomized Smoothing was proposed by (Blum et al., 2020) and (Kumar et al., 2020). Blum et al. (2020) demonstrate that any smoothing distribution for ℓ_p with $p > 2$ must have a "large component-wise magnitude" essentially saying that in order to certify high-dimensional images, the intensity of the noise would eventually dominate the useful information in the images. Hayes (2020) demonstrated a certified radius bound for $p = 1, 2$ and "location-scale distribution" (which includes the generalized Gaussian distribution) but their use of the KL divergence makes the bound loose. Kumar et al. (2020) and Yang et al. (2020) demonstrate a tight bound on the certified radius. More specifically, Kumar et al. (2020) demonstrate that the largest ℓ_p-radius of certification for $p > 2$ for some specific class of smoothing distributions decreases with rate $\mathcal{O}(1/d^{\frac{1}{2}-\frac{1}{p}})$ where d is the dimension of the images. Yang et al. (2020) find a similar result bound for a larger class of smoothing distribution. In the same vein, Wu et al. (2021) demonstrate a similar result for $p > 2$ for spherical symmetric distributions and a tight bound for ℓ_2-radius.

Towards Full-Information Certificates. All of these impossibility results (direct and indirect) are obtained in the minimal-information setting, where we only have access to the probability of the dominant class under the smoothing noise. To bypass these limitations, the natural way is to consider more information about the local shape of the decision boundary, and so to provide a bound over a smaller set of classifiers.

Building on that idea, several papers have proposed way to gather more information: Dvijotham et al. (2020) introduces a "full-information" certificate, whose name may be misleading : they do not use full information on the classifier, but only on the probability distribution at the point of interest after smoothing. This means knowing the probability for each class $c_1, \ldots, c_m$ instead of only the dominant one. They also introduce an interesting alternative to the Neyman-Pearson certification by relaxing the attack constraint using f-divergences. The additional information their method gives on the classifier itself is marginal, but their relaxation process can be used complementary of any

information-gathering method.

Mohapatra et al. (2020) shows that in the case of Gaussian smoothing, it is possible to reconstruct the smoothed classifier everywhere with arbitrary precision using the information on all its successive derivatives at the point of interest, which can be obtained using Monte-Carlo sampling. They also provide a certification algorithm that successfully improves upon the certificate proposed by Cohen et al. (2019) using first-order information, although marginally. There are two limitations to this method: first of all, derivatives are exponentially expensive to compute as the order increases, which makes all orders greater than the first currently intractable. Secondly, this only provides information on the Gaussian smoothed classifier, and cannot be easily used with any other smoothing scheme. However, their information-gathering process is orthogonal to ours and could be directly combined with our work, since it also relies on the generalized Neyman-Pearson lemma.

In this paper, we will provide a more general information-gathering framework, that extends those result by providing full information on the base classifier, allowing certificates to bypass information theoretic limitations with any smoothing scheme.

3. General framework for Randomized Smoothing certification

Let $\mathcal{X} = \mathbb{R}^d$ be our input space and $\mathcal{Y} = \{0, 1\}$ our label space. Let $\mathcal{H}$ be the class of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$, and $h \in \mathcal{H}$ be a base classifier. Randomised smoothing consists in smoothing h by averaging it over some probability density function q_0 over $\mathcal{X}$. When receiving an input $x \in \mathcal{X}$, we compute the probability that h takes value 1 for a point drawn from $q_0(\,\cdot\, - x)$:

$$p(x, h, q_0) = \int h(z) q_0(z - x)\ \mathrm{d}z$$

The smoothed classifier then returns the most probable class:

Definition 1. *The q_0-randomized smoothing of h is the classifier:*

$$h_{q_0} : x \mapsto \mathbb{1}\left\{p(x, h, q_0) > \frac{1}{2}\right\}$$

In the rest of the paper, we will consider the points x such that $p(x, h, q_0) \geq \frac{1}{2}$, so that the smoothed classifier returns 1. The other case is exactly symmetrical. An adversarial attack $\delta \in \mathcal{X}$ is a small crafted perturbation, such that $\|\delta\| \leq \epsilon$ where ϵ is a small constant and $\|\cdot\|$ is the Euclidean norm. An adversarial attack is engineered such that:

$$h_{q_0}(x + \delta) \neq y$$

meaning $p(x + \delta, h, q_0) < \frac{1}{2}$. In the following, we will define the robustness guarantee provided by Randomized Smoothing.

Definition 2. *An ϵ-certificate for the q_0-randomized smoothing of h at the point x is a value $v \in \mathbb{R}$ such that:*

$$v \leq \inf_{\delta \in B(0,\epsilon)} p(x + \delta, h, q_0)$$

A certificate v is said to be successful if $v > \frac{1}{2}$.

A successful ϵ-certificate means that no attack of norm at most ϵ can fool the classifier. This definition allows us to compare different certificates for the same smoothed classifier.

Certificates for randomised smoothing are usually "black-box", *i.e.* we can only access the classifier h through limited queries. This means giving a bound on the worst-case scenario for some class of functions $\mathcal{G}$ that contains h.

Definition 3 (Partial information certificate). *Let $\mathcal{G} \subset \mathcal{H}$ be a class of functions. A partial information certificate w.r.t. $\mathcal{G}$ is a value $v \in \mathbb{R}$ such that*

$$v \leq \inf_{g \in \mathcal{G}} \inf_{\delta \in B(0,\epsilon)} p(x + \delta, g, q_0)$$

In particular, $\mathcal{G}$ can be defined by constraints over the average response to noise distributions.

Definition 4 (Noised-based certificate). *Let $\mathcal{Q}$ be a finite family of probability density functions. Let $q_0 \in \mathcal{Q}$, a noise-based ϵ-certificate for the q_0-randomized smoothing of h at point x is:*

$$\mathrm{NC}(h, q_0, x, \epsilon, \mathcal{Q}) = \inf_{g \in \mathcal{G}_\mathcal{Q}} \inf_{\delta \in B(0,\epsilon)} p(x + \delta, g, q_0)$$

where :

$$\mathcal{G}_\mathcal{Q} = \{g \in \mathcal{H} : \forall q \in \mathcal{Q}, p(x, g, q) = p(x, h, q)\}$$

It is a partial information certificate for $\mathcal{G}_\mathcal{Q}$. When the infimum over $\mathcal{G}_\mathcal{Q}$ is attained by some g, we call g a $\mathcal{Q}$-worst case classifier.

The certificate from Cohen et al. (2019) is a particular type of partial information certificate, where the class $\mathcal{G}_{\{q_0\}}$ contains all functions that have the same average response as h to the smoothing distribution. In the rest of the paper, we will denote that single-noised-based ϵ-certificate as ϵ-SNC.

Lemma 1 (Neyman & Pearson (1933))**.** *Let q_0 be some probability density function, $\delta \in \mathcal{X}$, $k > 0$. Let:*

$$F_k : x \mapsto \mathbb{1}\{\mathcal{S}_k\}$$

Where $\mathcal{S}_k = \{q_0(x + \delta) \leq k q_0(x)\}$ is the Neyman-Pearson set. Then F_k minimizes $\int f(u) q_0(u + \delta)$ among all functions f such that $\int f(u) q_0(u)\ \mathrm{d}u \geq \int F_k(u) q_0(u)\ \mathrm{d}u$

The main result from Cohen et al. (2019) is that certificates can be computed using the Neyman-Pearson set:

Corollary 1. *Let $\mathcal{Q} = \{q_0\}$ and q_0 a probability density function. For any k such that $p(x, h, q_0) \geq p(x, F_k, q_0)$, then:*

$$\mathrm{NC}(h, q_0, x, \epsilon, \mathcal{Q}) \geq p(x + \delta, F_k, q_0) \tag{1}$$

Furthermore, if $p(x, h, q_0) = p(x, F_k, q_0)$, then Equation (1) is an equality, i.e. *F_k is the $\{q_0\}-$worst-case classifier.*

To evaluate the quality of this certificate, we now need a benchmark. For that, we will use the *perfect certificate*, *i.e.*, the tightest possible bound, that uses perfect information over the classifier h.

Definition 5 (Perfect certificate). *The perfect ϵ-certificate for the q_0-randomized smoothing of h at point x is:*

$$\mathrm{PC}(h, q_0, x, \epsilon) = \inf_{\delta \in B(0,\epsilon)} p(x + \delta, h, q_0)$$

The underestimation of a noise-based certificate can now be defined as the difference between both bounds.

Definition 6 (Underestimation of a noise-based certificate). *Let $\mathcal{Q}$ be a finite family of probability density functions and let $q_0 \in \mathcal{Q}$ and $\epsilon > 0$. We define the underestimation function ν as:*

$$\nu(h, q_0, x, \epsilon, \mathcal{Q}) = \mathrm{PC}(h, q_0, x, \epsilon) - \mathrm{NC}(h, q_0, x, \epsilon, \mathcal{Q})$$

The function ν computes the difference between the perfect ϵ-certificate and the noise-based ϵ-certificate for an classifier h with randomized smoothing q_0.

We will now show that the underestimation of single-noise certificates such as the ones from Cohen et al. (2019) depends on the shape of the decision boundary, and becomes large as the dimension increases.

4. Limitations of current certificates

In this section, we provide a deeper analysis of the underestimation function for the single noise certificate. Recall that it is $\mathrm{NC}(h, q_0, x, \epsilon, \mathcal{Q})$, where $\mathcal{Q}$ is the singleton $\{q_0\}$, *i.e.*, the same noise q_0 is used for the smoothing and the certification. Using information from a single noise presents two main weaknesses :

- Since $\mathcal{Q}$ is small, the certificate is obtained as a worst-case over a large set of functions $\mathcal{G}$, and will often be far worse than the optimal certificate PC for our specific classifier.
- With a classifier-agnostic certificate, we have limited possibilities of optimization for the choice of the base classifier h. In particular, current certificates are blind to the "local curvature" of the decision boundary, as will be illustrated shortly.

Since the curvature is a difficult notion to define for general decision boundaries, we consider the special cases of conic and piecewise-linear boundaries, where it can be characterized by a single real number, namely the angle of the curve.

In the following, we show, for uniform noises on the ℓ_2 ball and those two classes of decision boundaries, how much the single noise certificate underestimates the optimal one. We show that this underestimation is higher when the "local curvature" of the decision boundary is sharp, and that the single noise certificate can become arbitrary bad as dimension increases.

Let $(e_1, \ldots, e_d)$ be any orthonormal basis of $\mathbb{R}^d$.

Definition 7 (Cone of revolution). *Let $c \geq 0$. For any $x \in \mathbb{R}^d$, let $z, \rho, \phi_1, \ldots, \phi_{d-2}$ be the hyper-cylindrical coordinates of axis e_1 (see Definition 12). The cone of revolution of axis e_1, peaked at c and of angle $\theta \in \left[0, \frac{\pi}{2}\right]$ is the set $\mathcal{C}(c, \theta)$, defined by:*

$$\left\{ \begin{array}{l} z \in \mathbb{R} \\ \rho \in \mathbb{R}_+ \\ \phi_1, \ldots, \phi_{d-2} \in [0, \pi]\,. \end{array} \middle| \; z > c \text{ and } \rho \leq z \tan\theta \right\}$$

when $\theta \leq \frac{\pi}{2}$ (convex cone), and the set:

$$\mathcal{C}(c, \theta) = \{z \geq c \text{ or } \rho \geq -z \tan(\pi - \theta)\}\,.$$

for the concave cone ($\theta > \frac{\pi}{2}$).

We define a classifier with conical decision boundary as $h_\theta : x \mapsto \mathbb{1}\{x \notin \mathcal{C}(c, \theta)\}$.

Definition 8 (2-piecewise Linear set). *Let $c \geq 0$. Let $x_1, \ldots x_n$ be the euclidean coordinates in the base $(e_1, \ldots, e_n)$. The 2-piecewise linear decision region of axis e_1 and e_2, of distance c and angle $\theta \in \left[0, \frac{\pi}{2}\right]$ is the set :*

$$\left\{ x_1, \ldots, x_n \in \mathbb{R} \mid x_1 > c \text{ and } \arctan\left(\frac{x_2}{x_1}\right) \in [-\theta, \theta] \right\}$$

This is the set of all points contained between two intersecting hyperplanes. We now show that for both conic and 2-piecewise linear decision regions, the underestimation of the single noise certificate depends on the angle θ and grows with the dimension.

Theorem 1 (Underestimation of single noise-based certificates). *Let $\mathcal{Q} = \{q_0\}$ where q_0 is a uniform distribution over an ℓ_2 ball $B_2^d(0, r)$. Let $0 < \epsilon \leq r$, we denote $\theta_m = \arccos\left(\frac{\epsilon}{2r}\right)$. For any $\theta \in [0, \theta_m]$ we denote by h_θ the classifier whose decision boundary is a cone of revolution of peak 0, axis e_1 and angle θ. Then, $\nu(h_\theta, q_0, 0, \epsilon, \mathcal{Q})$ is a continuous and decreasing function of θ. Furthermore, we have*

- $\nu(h_0, q_0, 0, \epsilon, \mathcal{Q}) = 1 - I_{1-(\frac{\epsilon}{2r})^2}\left(\frac{d+1}{2}, \frac{1}{2}\right)$

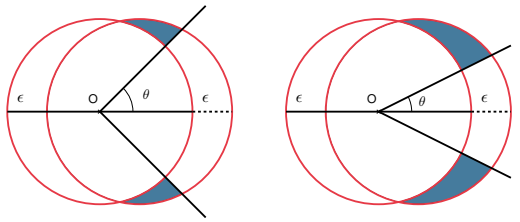

Figure 1. Illustration of theorem 1. The decision boundary is a cone of revolution, or two half-hyperplanes. Cohen et al.'s certificate underestimates the perfect one by the relative area of the blue zone, which increases with the curvature θ.

- $\nu(h_{\theta_m}, q_0, 0, \epsilon, \mathcal{Q}) = 0$

where $I_z(a, b)$ is the incomplete regularized beta function. Furthermore, for any ϵ, r, $\nu(h_0, q_0, 0, \epsilon, \mathcal{Q}) \xrightarrow[d\to\infty]{} 1$. The same result holds for 2-piecewise linear sets.

When using single noise certificate with uniform distributions, we assume the worst case, namely that all points lost after translating the ball (the left crescent zone outside the intersection in figure 1) are of the good class, and all points gained are of the bad class. Hence, the difference between the true certificate and the single noise certificate is the relative area of the blue zone.

For conical decision boundaries, the blue zone has a rotational symmetry around δ, whereas for $2-$piecewise linear sets it is invariant by translation along coordinates $x_3, \ldots, x_n$. In both cases the volume of the blue zone decreases when θ grows.

The last part of the result shows that as the dimension gets higher, the single noise certificate can become arbitrarily bad. In the extreme case ($\nu(\theta) = 1$), it returns 0 even though the classification task is trivial. This is due to the fact that the volume of balls tend to concentrate on their surface in high dimension, and so the relative weight of the crescent zone increases.

In the following, we show that we can identify the points where the single-noise classifier is optimal, by injecting successive noises. Namely, if the local decision boundary is such that the probability of returning the correct class increases with the radius of the noise, then single-noise certificates cannot be optimal. This process allows us to evaluate the proportion of points where better certificates may be obtained.

Proposition 1 (Identifying points of underestimation). *Let $q_\sigma \sim \mathcal{N}(0, \sigma I_d)$. For any $x \in \mathbb{R}^d$ and $\sigma_1 > 0$, let g_{σ_1} be a $\{q_{\sigma_1}\}$-worst-case classifier. Then $p(x, g_{\sigma_1}, q_\sigma)$ is a decreasing function of σ over $[\sigma_1, \infty)$.*

A similar result holds for uniform distributions $(q_r)_r$ over $B_2^d(0, r)$.

This means that for points where the single-noise certificate is optimal, the probability of the correct class should decrease as we increase the radius of the noise. Inversely, when that probability increases locally, we know that this certificate cannot be optimal.

Table 1. Proportion of CIFAR10 testset where p grows with σ, hinting at a locally highly convex decision region.

σ_1	σ_2	**Uniform noise**	**Gaussian noise**
0.25	0.30	40.9%	41.1%
0.50	0.55	41.9%	38.1%
0.75	0.80	41.3%	34.2%

Boundary Curvature of SOTA models on CIFAR10 We will use the result from proposition 1 to quantify the proportion of corrected classified points of CIFAR10 dataset where single-noise certificates are underestimating the real bounds. We considered uniform (first column) and Gaussian (second column) noises $q(\sigma)$ of increasing variance $\sigma > 0$, and measure the probability growth $\Delta p(\sigma_1, \sigma_2)$, where σ_1 is the same amplitude as the noise injected during the training of the model, and $\sigma_2 > \sigma_1$. Table 1 show the results for models by trained with uniform and Gaussian noises by Yang et al. (2020). We observe that nearly half of the corrected classified points have the probability growing with the variance suggesting that single-noise certificates would underestimating the real bounds.

5. Better certificates using noise-based information gathering

To solve this underestimation issue, we propose a new way of computing noise-based certificates. This will allow, for example, to gather more information on the classifier via noises of increasing variance (as in Figure 3), but is far more general and allows for any combination of noise distributions.

5.1. A framework for obtaining noise-based certificates

Lemma 2 (Generalized Neyman-Pearson lemma Chernoff & Scheffe (1952)**).** *Let $f, q_0, \ldots, q_n$ be probability density functions, and $\Phi_\mathcal{K}$ be any function of the form:*

$$\Phi_\mathcal{K} = \mathbb{1}\{\mathcal{S}_\mathcal{K}\}$$

Where $\mathcal{K} = \{k_1, \ldots, k_n\}$ with $k_1, \ldots, k_n > 0$, and $\mathcal{S}_\mathcal{K} = \{f(x) < \sum_{i=0}^n k_i q_i(x)\}$ is the generalized Neyman-Pearson set.

Then $\Phi_\mathcal{K}$ minimises $\int \Phi f d\mu$ over all functions $0 \le \Phi \le 1$ such that $\int \Phi q_i d\mu \ge \int \Phi_\mathcal{K} q_i d\mu$ for all i.

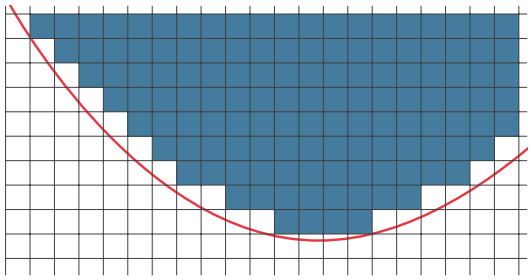

Figure 2. Illustration of the approximation theorem. By querying the classifier with uniform noises on the squares of the grid, we can compute an approximation of the true certificate using the blue squares. As we refine the grid with smaller squares, the approximation becomes increasingly good, and converges to the perfect certificate.

We now show that this can be used, with isotropic noises, to obtain certificates.

Corollary 2 (Computing noise-based certificates). *Let $\mathcal{Q} = \{q_0, \dots, q_n\}$ be a finite family of isotropic probability density functions, of same center. Let $\epsilon > 0$, and any δ of norm ϵ. If the k_i are such that $\forall i, p(x, \Phi_{\mathcal{K}}, q_i) \leq p(x, h, q_i)$, then we have:*

$$\mathrm{NC}(h, q_0, x, \epsilon, \mathcal{Q}) \geq p(x + \delta, \Phi_{\mathcal{K}}, q_0)$$

Furthermore, this becomes an equality if $\forall i, p(x, \Phi_{\mathcal{K}}, q_i) = p(x, h, q_i)$

This means that by choosing the k_i such that $p(x, \Phi_{\mathcal{K}}, q_i)$ is as close as possible from $p(x, h, q_i)$ while remaining lower, we can get arbitrary close to $\mathrm{NC}(h, q_0, x, \epsilon, \mathcal{Q})$ using the Neyman-Pearson classifier $\Phi_{\mathcal{K}}$.

Computing the certificates: This gives us the following algorithm to compute certificates using noises $q_0, \dots, q_n$:

1. Compute $p(x, h, q_i)$ for all noises q_i via Monte-Carlo sampling.
2. Compute the generalized Neyman-Pearson function $\Phi_{\mathcal{K}}$ by fitting the constants k_i such that $p(x, \Phi_{\mathcal{K}}, q_i) \leq p_i$ with the closest possible approximation.
3. Fix some δ of norm ϵ. If the noises used are all isotropic, we can directly compute the certificate $p(x+\delta, \Phi_{\mathcal{K}}, q_0)$ via sampling. For non-isotropic noises, different arguments must be used. We will give some examples later.

If the noise distributions are not isotropic or have different centers (breaking the isotropy of the problem), then we must take a lower bound over all $p(x + \delta, \Phi_{\mathcal{K}}, q_0)$ (where $\Phi_{\mathcal{K}}$ depends on δ), as we will do in the proof of theorem 2. We will also show later that introducing randomness in the certification process allows to obtain certificate even in the non-isotropic case.

The main advantage of this method is that since our class of function $\mathcal{G}$ shrinks with the number of noise used, a bound obtained with several noises will always be at least as good as the one proposed by Cohen et al. (2019).

5.2. Approximation results

In this part, we will show that noise-based information is enough to approximate any non-pathological classifier:

Theorem 2 (General approximation Theorem). *Let q_0 be the uniform noise on the ℓ_∞ ball $B_\infty(\cdot, r)$ for some $r > 0$. For any $\epsilon > 0$, $\xi > 0$ and $x \in \mathbb{R}^d$, there exists a finite family of probability density $\mathcal{Q}$ such that:*

$$\mathrm{NC}(h, q_0, x, \epsilon, \mathcal{Q}) \geq \mathrm{PC}(h, q_0, x, \epsilon) - \xi$$

Theorem 2 shows that it is possible to collect asymptotically perfect information on the decision boundary using only noise-based queries. The main improvement compared to the result of Mohapatra et al. (2020) is that we reconstruct the base classifier itself, and not just the gaussian smoothed version of it, hence it works for any smoothing scheme. This shows that the black-box approach to randomized smoothing certification is viable, and can bypass the theoretical limitations when using several noises instead of one.

Computational cost In the proof of Theorem 2, we use a very generic information-gathering method, which uses around $(\frac{r+\epsilon}{\xi})^d$ noises. In practice, we would refine the grid only for the squares that cover the decision boundary (*i.e.*, give a probability that is not close to 0 or 1), but that would still remain very costly.

However, if we have more prior information on the decision boundary, it will be possible to design much more efficient noise-based information gathering schemes. In the following, we present a result demonstrating that we can obtain full information in the case of conical or 2-piecewise linear decision boundaries by using only concentric uniform noises.

Definition 9 (Volume growth for a decision boundary). *Let $\mathcal{A}$ be a set, $r_1 > r_2 \geq 0$. We define the volume growth of $\mathcal{A}$ from r_1 to r_2 as:*

$$\Delta V(\mathcal{A}, r_1, r_2) = \mathrm{Vol}\left(\mathcal{A} \cap B_2^d(0, r_2)\right) - \mathrm{Vol}\left(\mathcal{A} \cap B_2^d(0, r_1)\right)$$

where B_2^d is the ℓ_2-ball in dimension d.

Definition 10 (Linear half-space). *Let $c \geq 0$. The half-space of translation c is, in hypercylindrical coordinates of axis e_1, the set :*

$$H(c) = \left\{ \begin{array}{l} z \in \mathbb{R} \\ \rho \in \mathbb{R}_+ \\ \phi_2, \dots, \phi_{d-1} \in [0, \pi]\,. \end{array} \middle| \; z > c \right\}$$

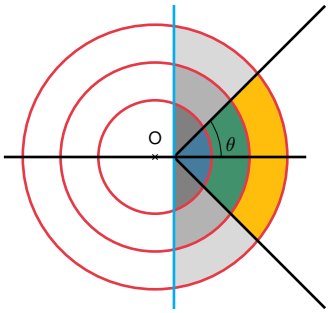

Figure 3. Illustration of Proposition 2. We see that the difference of volume captured between two balls (blue, green and yellow zones) grows with θ. For $\theta < \frac{\pi}{2}$, the volume growth is lower than for a hyperplan decision boundary (the cyan line) at the same distance. The difference is shown by the gray zones.

Proposition 2 (Growth function for concentric noises). *Let $c \geq 0$, $r_2 > r_1 > c$. Then $\Delta V(C(c,\theta), r_1, r_2)$ is a continuous, increasing function of θ, that is a bijection from $[0,\pi]$ to $[0, 2\Delta V(H(c), r_1, r_2)]$.*

The same result holds for 2-piecewise linear sets of parameter θ.

The Figure 3 is an intuitive illustration of Proposition 2. We can see that the volume of the cone captured by the balls grows with respect to θ.

Also, Proposition 2 means means that for these classes of decision boundaries, a small number of concentric noises is enough to obtain full information, in two steps :

- Evaluate the distance c to the decision boundary by finding the threshold such that $p(x, h, q(r)) \neq 1$;
- Use two noises of radius r_1 and r_2 to identify the angle θ as presented in Proposition 2.

This hints at a more general result for piecewise linear decision boundaries (which includes all neural networks with ReLus activations) : it may be possible to gather perfect information using only a bounded number of concentric noises to "map" the fractures.

However, the computation of noise-based certificates is a costly problem. In the next section, we will show some ways of reducing that complexity, and open a discussion on several choices of distributions that exhibit nice properties for generalized Neyman-Pearson certification.

6. Reducing the computational cost

In this section, we analyze the computational challenges of implementing noise-based certificates, as well as avenues to reduce them.

There are currently three main obstacles to computing noise-based certificates with this method :

1. Computing integrals via Monte-Carlo sampling in high-dimension can become very costly. Monte-Carlo and Markov-Chain Monte-Carlo are both methods that suffers the curse of dimensionality. While it could be possible (but difficult) to compute the integrals for a low dimensional datasets (*i.e.*, CIFAR10), it would be nearly impossibly to do that for a high-dimensional dataset (*i.e.*, ImageNet).
2. When computing the integrals in high dimension, numbers can become very small or very large, leading to computational instability.[1] Performing this type of computation without arbitrary precision library could be impossible with the dimension used in classical ML applications.
3. Finally, fitting the k_i to compute the generalized Neyman-Pearson set is a hard stochastic optimization problem.

We show that we can bypass problems 1. when using Gaussian noise, as well as problem 2. when also using a method that we call *high probability certification*. Finally, we show that uniform noise considerably reduces problem 3, although suffering from problems 1 and 2.

Theorem 3 (Sampling in low dimension for gaussian noise). *When $q_0, \ldots, q_n$ are normal distributions of center 0, we can express the certificate as $\mathbb{E}[f(V)]$, where V is a random variable in dimension at most n+1, and f is a function with output in $\{0,1\}$. V depends however on the input dimension d, and f on the Neyman-Pearson variables $k_1, \ldots, k_n$.*

Definition 11 (High Probability Certification). *An α-probable ϵ-certificate for the q_0 randomized smoothing is a value $v \in \mathbb{R}$ and a probability distribution Q such that, for any δ of norm at most ϵ:*

$$\mathbb{P}_Q\left[p(x+\delta, h, q_0) \geq v\right] \geq 1 - \alpha$$

Note that the randomness is not on the attack (we do not certify for "the majority of attacks", which would not work since attacks are engineered, thus worst-case guarantees are required).

Theorem 4 (Asymptotic dimension-independent certificate). *Let $q_0, \ldots, q_n$ are normal distribution of same variance σ, centered on points $x, z_1, \ldots, z_n$, where the z_i constitute a random family of orthonormal vectors, at constant distance from x. For any $\alpha \in (0,1), \epsilon > 0$, there exists an α-probable ϵ-certificate v_d and a value $v \in [0,1]$ such that $v_d \xrightarrow[d\to\infty]{} v$ and v can be written as $\mathbb{E}[f(V)]$, where V is a random variable in dimension at most n+1 that is independent from the dimension d.*

[1] For example, the volume of an ℓ_2 ball in dimension 784 (MNIST dimension) is approximately equal to $\exp(-1503.90)$.

Furthermore, that convergence is fast : for CIFAR10, with $\alpha = 0.05$, the approximation term is already of 0.99. This means that we can compute the value v (independent of the dimension) and use it as a high-probability certificate.

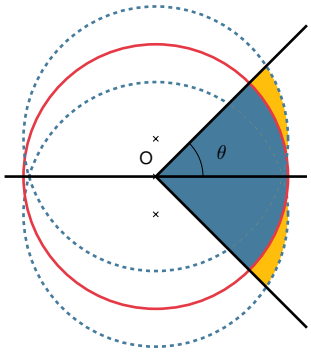

Figure 4. Illustration of Theorem 4. We can use translated noises to gather information (here identify the yellow zones), and in high dimensions the random translations will be almost orthogonal to the attack with high probability, whatever the direction of the attack.

The core argument of Theorem 4 is that in high dimension, for any attack vector δ, random directions will be almost orthogonal to δ with high probability. This allows us to obtain a certificate using the generalized Neyman-Pearson lemma, although our our noises do not have the same center so the problem is not isotropic.

This means that, by introducing some randomness during the information gathering, we can obtain a certificate using non-isotropic noise that is very easy to compute, since independent on the dimension. This removes one of the main barriers for our certificates to scale in very large dimension.

Proposition 3 (*Computing the k_i for uniform noises*)**.** *Let $q_0, \dots, q_n$ are uniform distributions, where $n \ll d$ there are only at most 2^n possible values for the generalized Neyman-Pearson set $\mathcal{S}$.*

This means that the exact values of the k_i do not matter, and the research of a set $\mathcal{S}$ shifts from a hard optimization problem to a combinatorial problem with only at most 2^n values to try where n correspond to the number of noise and is usually much lower than the input dimension. Also, we should remark that smart choices of noises can make that combinatorial problem easier in practice, since all of the sets do not intersect with each other.

Discussion on the choice of noise We see from Theorem 4 that by using normal distributions whose centers are drawn at random around point x, we can bypass problems 1 and 2, since the certificate can be computed as a random variable in low dimension. The technique of introducing randomness thus allows us to derive certificates even though the problem is not isotropic any more. This opens up a lot of possibilities for using translated noises to gather information. On the other hand, using uniform noises, fitting the k_i becomes a combinatorial task (Proposition 3), but the certificate $p(x + \delta, \Phi_0, q_i)$ remains hard to compute.

7. Future Works

Computing the k_i Theorem 3 and Theorem 4 successfully reduce the difficulty of the problem. However, even with those simplifications, fitting the k_i of the generalized Neyman-Pearson set remain a difficult problem. Indeed, fitting the k_i is an stochastic optimization problem where each step requires the computation of an integral via Monte-Carlo sampling. A potential direction would be to use the relaxation introduced by Dvijotham et al. (2020) for an easier to compute approximation of the Neyman-Pearson set. Both techniques from Yang et al. (2020) to compute ordinary Neyman-Pearson sets can also be extended to our general sets, for more computational efficiency.

Choosing the base classifier h Now that our certificates use more specific information on the classifier, it is possible to optimize the combination between the base classifier and the noise distributions used. For example, we may adjust our training to ensure that the decision boundary has the highest possible curvature, since it is where our new certificates will shine. The work from Salman et al. (2019) suggests that noise injection at the training time has a huge impact on the certification performance, and as we see in Table 1, this also impacts the curvature of the decision boundary. This could lead to a method for controlling the curvature at each point by adjusting the amount of noise injected.

8. Conclusion

State-of-the art certification methods for Randomized Smoothing suffer from severe underestimations, that increase with the dimension of the input space and are the source of the information-theoretical limitations to their performance. By collecting more information on the decision boundary, we can obtain tighter black-box certificates, which have the potential to scale as well as previous ones for larger network architectures, while not suffering from the same limitations.

Multiple-noise certification opens up a world of possibilities in terms of certification methods. Specific combination of noises can be chosen for each classifier we wish to certify on, with the guarantee of always performing better than current certificates. Specific base classifiers may be chosen to optimize the efficiency of the smoothing mechanism.

Much work remains however to be done, in order to actually implement certificates using this framework. The main diffi-

culty is to fit the constants $k_1, \ldots, k_n$, which are hard to do using Monte-Carlo sampling for most choices of noises. But the obstacle has now shifted from an impossibility result to computational challenges, restoring hope that Randomized Smoothing may someday be a definitive solution against adversarial example attacks.

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Blum, A., Dick, T., Manoj, N., and Zhang, H. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *Journal of Machine Learning Research*, 2020.

Blumenson, L. A derivation of n-dimensional spherical coordinates. *The American Mathematical Monthly*, 67, 1960.

Cao, X. and Gong, N. Z. Mitigating evasion attacks to deep neural networks via region-based classification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pp. 278–287, 2017.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., and Madry, A. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Chernoff, H. and Scheffe, H. A generalization of the neyman-pearson fundamental lemma. *The Annals of Mathematical Statistics*, 1952.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 2012.

Dvijotham, K. D., Hayes, J., Balle, B., Kolter, Z., Qin, C., Gyorgy, A., Xiao, K., Gowal, S., and Kohli, P. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020.

Globerson, A. et al. Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Hayes, J. Extensions and limitations of randomized smoothing for robustness guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Kumar, A., Levine, A., Goldstein, T., and Feizi, S. Curse of dimensionality on randomized smoothing for certifiable robustness. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Lecuyer, M., Atlidakais, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, 2018.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, 2019a.

Li, Q., Haque, S., Anil, C., Lucas, J., Grosse, R. B., and Jacobsen, J.-H. Preventing gradient attenuation in lipschitz constrained convolutional networks. In *Advances in Neural Information Processing Systems*, 2019b.

Li, S. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

Liu, X., Cheng, M., Zhang, H., and Hsieh, C.-J. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Meunier, L., Delattre, B., Araujo, A., and Allauzen, A. Scalable lipschitz residual networks with convex potential flows. *arXiv preprint arXiv:2110.12690*, 2021.

Mohapatra, J., Ko, C.-Y., Weng, T.-W., Chen, P.-Y., Liu, S., and Daniel, L. Higher-order certification for randomized smoothing. In *Advances in Neural Information Processing Systems*, 2020.

Mohapatra, J., Ko, C.-Y., Weng, L., Chen, P.-Y., Liu, S., and Daniel, L. Hidden cost of randomized smoothing. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Neyman, J. and Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

Olver, F. W., Lozier, D. W., Boisvert, R. F., and Clark, C. W. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.

Pinot, R., Meunier, L., Araujo, A., Kashima, H., Yger, F., Gouy-Pailler, C., and Atif, J. Theoretical evidence for adversarial robustness through randomization. In *Advances in Neural Information Processing Systems*, 2019.

Pinot, R., Ettedgui, R., Rizk, G., Chevaleyre, Y., and Atif, J. Randomization matters how to defend against strong adversarial attacks. In *International Conference on Machine Learning*, pp. 7717–7727. PMLR, 2020.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018a.

Raghunathan, A., Steinhardt, J., and Liang, P. S. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, 2018b.

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 *Advances in Neural Information Processing Systems*, 2019.

Singla, S. and Feizi, S. Skew orthogonal convolutions. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Singla, S., Singla, S., and Feizi, S. Householder activations for provable robustness against adversarial attacks. *arXiv preprint arXiv:2108.04062*, 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Trockman, A. et al. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations*, 2021.

Weng, L., Zhang, H., Chen, H., Song, Z., Hsieh, C.-J., Daniel, L., Boning, D., and Dhillon, I. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, 2018.

Wong, E. and Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, 2018.

Wu, Y., Bojchevski, A., Kuvshinov, A., and Günnemann, S. Completing the picture: Randomized smoothing suffers from the curse of dimensionality for a large family of distributions. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Yang, G., Duan, T., Hu, J. E., Salman, H., Razenshteyn, I., and Li, J. Randomized smoothing of all shapes and sizes. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Yousefian, F., Nedić, A., and Shanbhag, U. V. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 2012.

A. Definitions

In the following, we will consider the dimension $d \geq 3$.

Definition 12 (**Hyper-cylindrical coordinates** This is an extension of the hyperspherical coordinates, defined in Blumenson (1960))**.** *Let $e_1, \ldots, e_d$ be an orthonormal base of $\mathbb{R}^d$, with corresponding Euclidean coordinates $(x_1, \ldots, x_d)$. The hyper-cylindrical coordinates of axis e_1 are the following change of variable:*

$$z = x_1 \tag{2}$$

$$\rho = \sqrt{x_2^2 + \cdots + x_d^2} \tag{3}$$

$$\phi_i = \operatorname{arccot}\left(\frac{x_i}{\sqrt{x_n^2 + \cdots + x_i^2}}\right) \tag{4}$$

$$\phi_{d-1} = 2\operatorname{arccot}\left(\frac{x_{d-1} + \sqrt{x_{d-1}^2 + x_d^2}}{x_d}\right) \tag{5}$$

with the following reverse transformation:

$$x_1 = z \tag{6}$$

$$x_2 = r\cos(\phi_1) \tag{7}$$

$$x_i = r\left(\prod_{i=1}^{i-2}\sin(\phi_i)\right)\cos(\phi_{i-1}) \tag{8}$$

$$x_d = r\left(\prod_{i=1}^{d-2}\sin(\phi_{i-2})\right) \tag{9}$$

where $i \in \{2, \ldots, d-1\}$. This is a bijection, where $\phi_i \in [0, \pi]$, $r \in \mathbb{R}_+$, and $\phi_{d-1} \in [0, 2\pi]$, with the convention that $\phi_k = 0$ when $x_k, \ldots, x_n = 0$. Note that it is simply a change of variables to hyperspherical coordinates on the $d-1$ last variables.

Definition 13 (Incomplete Regularized Beta)**.** *Let $z \in \mathbb{R}$, $a > 0$ and $b > 0$. The Incomplete Regularized Beta Function is the function defined as:*

$$I_z(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^z t^{a-1}(1-t)^{b-1}\,\mathrm{d}t \tag{10}$$

Definition 14 (ℓ_p-ball)**.** *The ℓ_p-ball of dimension d, radius $r > 0$ and center $c \in \mathbb{R}^d$ is the set:*

$$B_p^d(c, r) = \{\forall x \in \mathbb{R}^d \mid \|x\|_p \leq r\} \tag{11}$$

Definition 15 (Spherical Cap of an ℓ_p-ball (Li, 2011))**.** *A* spherical cap *is the portion of the sphere that is cut away by an hyperplane of distance $r - a$ from the origin. The formula for the volume of the spherical cap is given by:*

$$\operatorname{Vol}(\operatorname{Cap}(a, r, d)) = \frac{1}{2}\operatorname{Vol}(B_p^d(0, r))I_{\frac{2ra-h^2}{r^2}}\left(\frac{d+1}{2}, \frac{1}{2}\right) \tag{12}$$

B. Proofs of Section 4

B.1. Proof of Theorem 1

Theorem 1 (Underestimation of single noise-based certificates). *Let $\mathcal{Q} = \{q_0\}$ where q_0 is a uniform distribution over an ℓ_2 ball $B_2^d(0, r)$. Let $0 < \epsilon \leq r$, we denote $\theta_m = \arccos\left(\frac{\epsilon}{2r}\right)$. For any $\theta \in [0, \theta_m]$ we denote by h_θ the classifier whose decision boundary is a cone of revolution of peak 0, axis e_1 and angle θ. Then, $\nu(h_\theta, q_0, 0, \epsilon, \mathcal{Q})$ is a continuous and decreasing function of θ. Furthermore, we have*

- $\nu(h_0, q_0, 0, \epsilon, \mathcal{Q}) = 1 - I_{1-(\frac{\epsilon}{2r})^2}\left(\frac{d+1}{2}, \frac{1}{2}\right)$
- $\nu(h_{\theta_m}, q_0, 0, \epsilon, \mathcal{Q}) = 0$

where $I_z(a, b)$ is the incomplete regularized beta function. Furthermore, for any ϵ, r, $\nu(h_0, q_0, 0, \epsilon, \mathcal{Q}) \xrightarrow[d\to\infty]{} 1$. The same result holds for 2-piecewise linear sets.

Lemma 3 (Limit of the regularized incomplete beta function). *Let $z \leq 1$, $b = \frac{1}{2}$ fixed. Then $I_z(a, b) \xrightarrow[a\to\infty]{} 0$.*

Proof. For any $z < 1, a > 1, b = \frac{1}{2}$, we have:

$$I_z(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^z t^{a-1}(1-t)^{b-1}\, \mathrm{d}t \tag{13}$$

$$\leq \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} z^a (1-z)^{-\frac{1}{2}} \tag{14}$$

From Olver et al. (2010), Equation 5.11.12, we have the following approximation:

$$\frac{\Gamma(a+b)}{\Gamma(a)} \underset{a\to+\infty}{\sim} a^b \tag{15}$$

from Equation (15), we can show that:

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \underset{a\to+\infty}{\sim} \frac{a^b}{\Gamma(b)} \tag{16}$$

Finally, we have:

$$I_z(a,b) \leq \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} z^a (1-z)^{-\frac{1}{2}} \tag{17}$$

$$\underset{a\to+\infty}{\sim} \frac{a^b}{\Gamma(b)} z^a (1-z)^{-\frac{1}{2}} \tag{18}$$

$$\xrightarrow[a\to+\infty]{} 0 \tag{19}$$

which concludes the proof. □

In what follows, $V = \mathrm{Vol}(B_2^d(0, r))$

Lemma 4. *The optimal attack of size ϵ against $\mathcal{C}(0, \theta)$ is the translation fully along its axis e_1,* i.e. ϵe_1.

Proof. Let $A = \mathrm{Span}(e_1), B = \mathrm{Span}(e_2, \ldots, e_d)$. For any vector $u \in \mathbb{R}^d$, we write $u = u_A + u_B$ where u_A and u_B are the orthogonal projections of u on A and B respectively.

First we will show that since the cone is invariant by rotation around e_1, the orthogonal component of the attack is as well. Without any attack, the probability of returning 1 at point 0, is

$$\frac{1}{V} \int_{\mathbb{R}^d} \mathbb{1}\left\{\|x_A - 0\|^2 + \|x_B - 0\|^2 \leq r^2\right\} \mathbb{1}\left\{\|x_B\| \leq \|x_A\| \tan(\theta)\right\} \mathrm{d}x \tag{20}$$

where V is the volume of the ball of radius r and center 0.

Let δ be any attack vector, $\|\delta\| = \epsilon$. Attacking by δ amounts to shifting the center of the ball from $(0,0)$ to (δ_A, δ_B). Let

$$f(x,\delta) = \mathbb{1}\left\{\|x_A - \delta_A\|^2 + \|x_B - \delta_B\|^2 \leq r^2\right\} \mathbb{1}\left\{\|x_B\| \leq \|x_A\| \tan(\theta)\right\}, \qquad \forall x \in \mathbb{R}^d \tag{21}$$

Then the probability of returning 1 at point 0, under attack δ, is $p(\delta) = \frac{1}{V} \int\limits_{\mathbb{R}^d} f(x,\delta)\ \mathrm{d}x$ where V is independent of δ.

Now let g be any isometric mapping such that $g|_A = \mathrm{Id}_A$. Let $\tilde{\delta} = g(\delta)$. Recall that as g is an isometry hence g and g^{-1} are also affine, hence we have

$$\begin{aligned}
f(g^{-1}(x),\delta) &= \mathbb{1}\left\{\|g^{-1}(x_A) - \delta_A\|^2 + \|g^{-1}(x_B) - \delta_B\|^2 \leq r^2\right\} \mathbb{1}\left\{\|g^{-1}(x_B)\| \leq \|g^{-1}(x_A)\| \tan(\theta)\right\} && (22)\\
&= \mathbb{1}\left\{\|g^{-1}(x_A - \tilde{\delta}_A)\|^2 + \|g^{-1}(x_B - \tilde{\delta}_B)\|^2 \leq r^2\right\} \mathbb{1}\left\{\|g^{-1}(x_B)\| \leq \|g^{-1}(x_A)\| \tan(\theta)\right\} && (23)\\
&= \mathbb{1}\left\{\|x_A - \tilde{\delta}_A\|^2 + \|x_B - \tilde{\delta}_B\|^2 \leq r^2\right\} \mathbb{1}\left\{\|x_B\| \leq \|x_A\| \tan(\theta)\right\} && (24)\\
&= f(x,\tilde{\delta}) = f(x, g^{-1}(\delta)) && (25)
\end{aligned}$$

since g^{-1} is an isometry. It follows, by a change of variable in the integral:

$$\begin{aligned}
p(\delta) &= \frac{1}{V} \int\limits_{\mathbb{R}^d} f(x,\delta)\ \mathrm{d}x && (26)\\
&= \frac{1}{V} \int\limits_{\mathbb{R}^d} f(g^{-1}(u),\delta)\ \mathrm{d}u && (27)\\
&= \frac{1}{V} \int\limits_{\mathbb{R}^d} f(u, g^{-1}(\delta))\ \mathrm{d}u && (28)\\
&= p(\tilde{\delta}) && (29)
\end{aligned}$$

In particular, we can always choose g such that the $g(\delta_B) = \delta_2 e_2$. In what follows, we will consider δ of the form $\delta = \delta_1 e_1 + \delta_2 e_2$ and show that the attack is optimal when $\delta_2 = 0$. For this, we will compute the difference between the attack translated by $\delta_1 e_1$ and the one translated by $\delta_1 e_1 + \delta_2 e_2$, to show that the orthogonal component actually reduces the efficiency of the attack. Let us first recall that

$$\begin{aligned}
p(\delta) &= \frac{1}{V} \int\limits_{\mathbb{R}^d} \mathbb{1}\left\{\|x_1 - \delta_1\|^2 + \|x_2 - \delta_2\|^2 + \|x_3\|^2 + \cdots + \|x_d\|^2 \leq r^2\right\} \mathbb{1}\left\{\|x_B\| \leq \|x_A\| \tan(\theta)\right\}\ \mathrm{d}x_1 \ldots \mathrm{d}x_d && (30)\\
&= \frac{1}{V} \mathrm{Vol}(B(\delta, r) \cap C(0,\theta)). && (31)
\end{aligned}$$

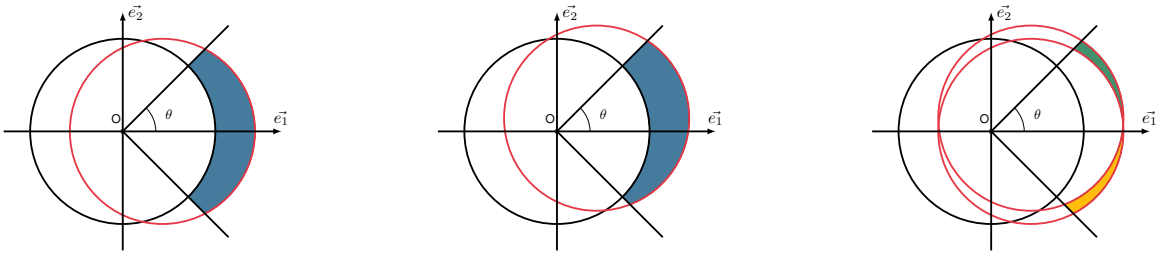

Figure 5. Illustration of the proof. The illustration on the left shows that there is always a gain by translating along e_1, the illustration in the middle shows the gain when translating along both e_1 and e_2, and finally, the illustration on the right shows the difference. The loss incurred by the second translation, visible in the yellow zone, is greater than the gain (blue zone). The argument of the proof is that the symmetric of the blue zone is contained in the yellow zone.

Let $A = B(\delta, r) \cap C(0, \theta) \setminus B(\delta_1 e_1, r)$ and $D = B(\delta_1 e_1, r) \cap C(0, \theta) \setminus B(\delta, r)$.

$$p(\delta_1 e_1) - p(\delta) = \frac{1}{V} \operatorname{Vol}(B(\delta, r) \cap C(0, \theta)) - \frac{1}{V} \operatorname{Vol}(B(\delta_1 e_1, r) \cap C(0, \theta)) \tag{32}$$

$$= \frac{1}{V} \left(\operatorname{Vol}(B(\delta, r) \cap C(0, \theta) \cap B(\delta_1 e_1, r)) + \operatorname{Vol}\left(B(\delta, r) \cap C(0, \theta) \setminus B(\delta_1 e_1, r)\right) \right. \tag{33}$$

$$\left. - \operatorname{Vol}(B(\delta_1 e_1, r) \cap C(0, \theta)) \cap B(\delta, r)) - \operatorname{Vol}\left(B(\delta_1 e_1, r) \cap C(0, \theta) \setminus B(\delta, r)\right) \right) \tag{34}$$

$$= \frac{1}{V} \left(\operatorname{Vol}(D) - \operatorname{Vol}(A) \right) \tag{35}$$

We have $p(\delta_1 e_1) - p(\delta) = \frac{1}{V} \left(\operatorname{Vol}(D) - \operatorname{Vol}(A) \right)$. To show that it is positive, we will show that there is an isometry v (preserving volumes) such that $v(A) \subset D$.

Let v be the reflection across the hyperplan $\left\{ x \in \mathbb{R}^d, x_2 = \frac{\delta_2}{2} \right\}$. We have $v(x_1, \dots, x_d) = (x_1, \delta_2 - x_2, x_3, \dots, x_d)$. For simplicity, for any $x \in \mathbb{R}^d$, we denote $v(x) = \tilde{x}$.

Let $x \in A$. We will show that $\tilde{x}$ is in D. As x is in A, we have

$$\begin{cases} (x_1 - \delta_1)^2 + (x_2 - \delta_2)^2 + x_3^2 + \dots + x_d^2 \leq r^2 & (36) \\ x_1 > 0 & (37) \\ x_2^2 + \dots + x_d^2 \leq x_1^2 \tan^2(\theta) & (38) \\ (x_1 - \delta_1)^2 + x_2^2 + x_3^2 + \dots + x_d^2 > r^2 & (39) \end{cases}$$

Equation (36) states that $x \in B(\delta, r)$, Equation (37) and Equation (38) state that it is in the cone, whereas Equation (39) says that $x \notin B(\delta_1 e_1, r)$.

Let us first show that $\tilde{x} \in C(0, \theta)$. $\tilde{x_1} = x_1 > 0$, and subtracting Equation (36) from Equation (39) gives us $x_2^2 > (x_2 - \delta_2)^2$. It follows:

$$\tilde{x_2}^2 + \dots + \tilde{x_d}^2 = (\delta_2 - x_2)^2 + x_3^2 + \dots + x_d^2 \tag{40}$$

$$< x_2^2 + \dots + x_d^2 \tag{41}$$

$$\leq x_1^2 \tan^2(\theta) \qquad \text{(from Equation (38))} \tag{42}$$

$$= \tilde{x_1} \tan^2(\theta) \tag{43}$$

Now we show that $\tilde{x} \in B(\delta_1 e_1, r)$.

$$(\tilde{x_1} - \delta_1)^2 + \tilde{x_2}^2 + \dots + \tilde{x_d}^2 = (x_1 - \delta_1)^2 + (x_2 - \delta_2)^2 + x_3^2 + \dots + x_d^2 \tag{44}$$

$$\leq r^2 \tag{45}$$

Finally we show $\tilde{x} \notin B(\delta, r)$.

$$(\tilde{x_1} - \delta_1)^2 + (\tilde{x_2} - \delta_2)^2 + \dots + \tilde{x_d}^2 = (x_1 - \delta_1)^2 + x_2^2 + x_3^2 + \dots + x_d^2 \tag{46}$$

$$> r^2 \qquad \text{(from Equation (39))} \tag{47}$$

Combining the above, we get $\tilde{x} \in D$. As x was chosen arbitrarily in A, we get $v(A) \subset D$. As v is isometric, we finally get $p(\delta_1 e_1) - p(\delta) = \frac{1}{V} \left(\operatorname{Vol}(A) - \operatorname{Vol}(D) \right) = \frac{1}{V} \left(\operatorname{Vol}(v(A)) - \operatorname{Vol}(D) \right) \leq 0$. The component orthogonal to the axis is detrimental to the attack.

We now only need to prove that $p(\delta)$ is strictly increasing with δ_1. For what follow, we consider an attack $\delta_1 e_1$, and another one $((\delta_1 + \Delta) e_1)$. We will use the same technique:

Let $A = B(\delta_1 e_1, r) \cap C(0, \theta) \setminus B((\delta_1 + \Delta) e_1, r)$, and $D = B((\delta_1 + \Delta) e_1, r) \cap C(0, \theta) \setminus B(\delta_1 e_1, r)$. We have $p((\delta_1 + \Delta) e_1) - p(\delta_1 e_1) = \frac{1}{V} \left(\operatorname{Vol}(D) - \operatorname{Vol}(A) \right)$, and we will show that there is an isometry v such that $v(A) \subset D$.

Let v be the reflection across the hyperplane $\left\{x \in \mathbb{R}^d \mid x_1 = \delta_1 + \frac{\Delta}{2}\right\}$. Let $x = (x_1, \dots, x_d) \in A$. It verifies the following equations:

$$\begin{cases} (x_1 - \delta_1)^2 + x_2^2 + x_3^2 + \cdots + x_d^2 \leq r^2 & (48) \\ x_1 > 0 & (49) \\ x_2^2 + \cdots + x_d^2 \leq x_1^2 \tan^2(\theta) & (50) \\ (x_1 - \delta_1 - \Delta)^2 + x_2^2 + x_3^2 + \cdots + x_d^2 > r^2 & (51) \end{cases}$$

$v(x_1, \dots, x_d) = (2\delta_1 + \Delta - x_1, x_2, \dots, x_d) = \tilde{x}$.

First of all, subtracting Equation (51) from Equation (48) gives:

$$(x_1 - \delta_1 - \Delta)^2 > (x_1 - \delta_1)^2 \Rightarrow \Delta^2 - 2\Delta(x_1 - \delta_1) > 0 \tag{52}$$

$$\Rightarrow x_1 < \delta_1 + \frac{\Delta}{2} \tag{53}$$

Let us show $\tilde{x} \in D$.

$$\tilde{x_1}^2 \tan^2(\theta) = (2\delta_1 + \Delta - x_1)^2 \tan^2(\theta) \tag{54}$$

$$\geq \left(2\delta_1 + \Delta - \delta_1 - \frac{\Delta}{2}\right)^2 \tan^2(\theta) \tag{55}$$

$$= \left(\delta_1 + \frac{\Delta}{2}\right)^2 \tan^2(\theta) \tag{56}$$

$$\geq x_1^2 \tan^2(\theta) \tag{57}$$

$$\geq x_2^2 + \cdots + x_d^2 \tag{58}$$

$$= \tilde{x_2}^2 + \cdots + \tilde{x_d}^2 \tag{59}$$

Hence $\tilde{x} \in C(0, \theta)$. Then:

$$(\tilde{x_1} - \delta_1 - \Delta)^2 + \tilde{x_2}^2 + \cdots + \tilde{x_d}^2 = (x_1 - \delta_1)^2 + x_2^2 + \cdots + x_d^2 \tag{60}$$

$$\leq r^2 \tag{61}$$

Hence $\tilde{x} \in B((\delta_1 + \Delta)e_1, r)$. Finally,

$$(\tilde{x_1} - \delta_1)^2 + \tilde{x_2}^2 + \cdots + \tilde{x_d}^2 = (x_1 - \delta_1 - \Delta)^2 + x_2^2 + \cdots + x_d^2 \tag{62}$$

$$> r^2 \tag{63}$$

and we have $\tilde{x} \notin B(\delta_1 e, r)$. We have thus shown that $\mathrm{Vol}(D) \geq \mathrm{Vol}(A)$, and so the attack is increasing in δ_1.

We have shown that any component of the attack that is orthogonal to e_1 is detrimental to the attack, and that an increase along e_1 benefits the attack. It follows that the optimal attack of size at most ϵ is ϵe_1. □

Proof of Theorem 1. Let $r > 0$, $0 < \epsilon \leq r$, $\delta = [\epsilon, 0, \dots, 0]^\top \in \mathbb{R}^d$, $\theta \in [0, \theta_m]$ with $\theta_m = \arccos\left(\frac{\epsilon}{2r}\right)$. Let $\mathcal{C}(0, \theta)$ be a cone of revolution of peak 0, axis e_1 and angle θ. Let us consider all functions h_θ whose decision boundary is the cone of revolution $\mathcal{C}(0, \theta)$. The probability $p(x, h_\theta, q_0)$ of returning class 1 at the point 0 for the classifier h_θ after smoothing by q_0 is:

$$p(x, h_\theta, q_0) = \int_{\mathbb{R}^d} \frac{\mathbb{1}\left\{x \in B_2^d(0, r)\right\}}{\mathrm{Vol}(B_2^d(0, r))} \mathbb{1}\left\{x \in \mathcal{C}(0, \theta)^c\right\} \, \mathrm{d}x \tag{64}$$

$$= \frac{\mathrm{Vol}(B_2^d(0, r) \cap \mathcal{C}(0, \theta)^c)}{\mathrm{Vol}(B_2^d(0, r))} \tag{65}$$

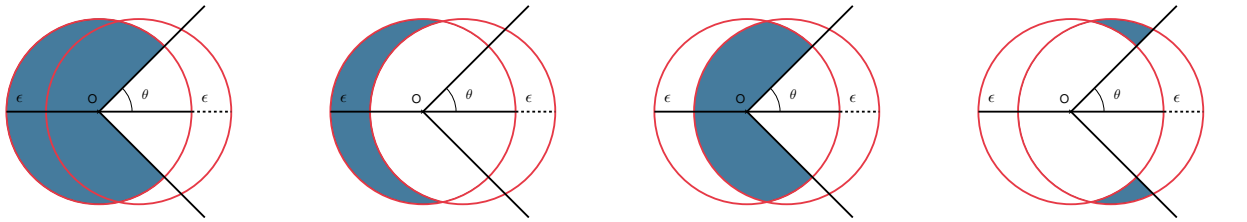

Figure 6. Illustration of proof of Theorem 1. The worst-case classifier using only the information $p(\theta)$ assumes that the zone in the second figureis entirely lost, whereas for the perfect certificate, the blue zone in the fourth figure is not lost. That zone grows as θ shrinks.

Single-noise certificates uses the fact that in the worst-case scenario, all the volume lost during the translation was in class 1, and all the volume gained is in the class 0 (see proof of Yang et al. (2020) [Theorem I.19]) This gives :

$$\begin{aligned}
\mathrm{NC}(h_\theta, q_0, x, \epsilon, \{q_0\}) &= p(x, h_\theta, q_0) - \frac{1}{V}\left(\mathrm{Vol}(B_2^d(0,r) \setminus B_2^d(\delta,r))\right) && (66)\\
&= \frac{1}{V}\left(\mathrm{Vol}(B_2^d(0,r) \cap \mathcal{C}(0,\theta)^c) - \left(\mathrm{Vol}(B_2^d(0,r) \setminus B_2^d(\delta,r))\right)\right) && (67)\\
&= \frac{1}{V}\mathrm{Vol}(B_2^d(0,r) \cap B_2^d(\delta,r) \cap \mathcal{C}(0,\theta)^c) && (68)
\end{aligned}$$

In Equation (66), we say that in the worst case scenario all the volume lost during the translation was in class 1, and all volume gained was in class 0, hence we loose everything outside the intersection.

Since by Lemma 4 the optimal attack against the cone is the translation along its axis, the perfect certificate for the probability p will be defined under the attack δ:

$$\mathrm{PC}(h_\theta, q_0, x, \epsilon) = \frac{1}{V}\mathrm{Vol}(B_2^d(\delta,r) \cap \mathcal{C}(0,\theta)^c) \tag{69}$$

The difference between the perfect certificate and the single-noise based certificate (as in Definition 6) is:

$$\begin{aligned}
\nu(h_\theta, q_0, x, \epsilon, \{q_0\}) &= \frac{1}{V}\left(\mathrm{Vol}(B_2^d(\delta,r) \cap \mathcal{C}(0,\theta)^c) - \mathrm{Vol}(B_2^d(0,r) \cap B_2^d(\delta,r) \cap \mathcal{C}(0,\theta)^c)\right) && (70)\\
&= \frac{1}{V}\mathrm{Vol}(B_2^d(\delta,r) \cap \mathcal{C}(0,\theta)^c \setminus \left(B_2^d(0,r) \cap B_2^d(\delta,r) \cap \mathcal{C}(0,\theta)^c\right)) && (71)\\
&= \frac{1}{V}\mathrm{Vol}(B_2^d(\delta,r) \cap \mathcal{C}(0,\theta)^c \cap \left(B_2^d(0,r) \cap B_2^d(\delta,r) \cap \mathcal{C}(0,\theta)^c\right)^c) && (72)\\
&= \frac{1}{V}\mathrm{Vol}(B_2^d(\delta,r) \cap \mathcal{C}(0,\theta)^c \cap \left(B_2^d(0,r)^c \cup B_2^d(\delta,r)^c \cup \mathcal{C}(0,\theta)\right)) && (73)\\
&= \frac{1}{V}\mathrm{Vol}(B_2^d(\delta,r) \cap B_2^d(0,r)^c \cap \mathcal{C}(0,\theta)^c) && (74)\\
&= \int && (75)
\end{aligned}$$

It follows that ν is a continuous function with respect to $\theta \in [0, \theta_m]$. It is decreasing, since $\mathcal{C}(0,\theta_1) \subset \mathcal{C}(0,\theta_2)$ when $\theta_1 < \theta_2$.

Furthermore, when $\theta = 0$:

$$\nu(h_0, q_0, x, \epsilon, \{q_0\}) = \frac{1}{V}\operatorname{Vol}(B_2^d(\delta, r) \cap B_2^d(0, r)^c) \tag{76}$$

$$= \frac{1}{V}\operatorname{Vol}(B_2^d(\delta, r) \setminus B_2^d(0, r)) \tag{77}$$

$$= \frac{1}{V}\operatorname{Vol}(B_2^d(\delta, r)) - \operatorname{Vol}(B_2^d(0, r) \cap B_2^d(\delta, r)) \tag{78}$$

$$= \frac{1}{V}\left(\operatorname{Vol}(B_2^d(\delta, r) - 2\operatorname{Vol}(\operatorname{Cap}(r - \frac{\epsilon}{2}, r, d))\right) \tag{79}$$

$$= 1 - I_{1-\left(\frac{\epsilon}{2r}\right)^2}\left(\frac{d+1}{2}, \frac{1}{2}\right) \tag{80}$$

where the step from Equation (78) to Equation (79) is due because the intersection of both spheres is the union of two spherical caps.

Moreover, from Lemma 3, we have $\nu(h_0, q_0, x, \epsilon, \{q_0\}) \xrightarrow[d\to\infty]{} 1$:

And, when $\theta = \theta_m$, we are going to prove that $\nu(h_{\theta_m}, q_0, x, \epsilon, \{q_0\}) = 0$. Equivalently, we want to show that the set defined by:

$$\left\{(x, \rho) \in \mathbb{R} \;\middle|\; x < \frac{\epsilon}{2} \;\middle|\; (x - \epsilon)^2 + \rho^2 \leq r^2 \;\middle|\; \rho > x \tan\theta_m\right\} \tag{81}$$

is an empty set. Let (x, ρ) in this set. We have:

$$\rho > x \tan\left(\arccos\left(\frac{\epsilon}{2r}\right)\right) \tag{82}$$

$$= \frac{2rx}{\epsilon}\sqrt{1 - \frac{\epsilon^2}{4r^2}} \tag{83}$$

due to the equality: $\tan(\arccos(x)) = \frac{\sqrt{1-x^2}}{x}$. Then, we have:

$$r^2 \geq (x - \epsilon)^2 + \rho^2 \tag{84}$$

$$\geq x^2 - 2x\epsilon + \epsilon^2 + \frac{4r^2x^2}{\epsilon^2}\left(1 - \frac{\epsilon^2}{4r^2}\right) \tag{85}$$

$$= x^2 - 2x\epsilon + \epsilon^2 + \frac{4r^2x^2}{\epsilon^2} - x^2 \tag{86}$$

$$= \frac{4r^2}{\epsilon^2}x^2 - 2x\epsilon + \epsilon^2 \tag{87}$$

Hence we have:

$$\frac{4r^2}{\epsilon^2}x^2 - 2x\epsilon + \epsilon^2 - r^2 \leq 0 \quad \text{and} \quad x \leq \frac{\epsilon}{2} \tag{88}$$

But the minimum of the right hand side is: $\frac{\epsilon^3}{4r^2} \leq \frac{\epsilon}{2}$ because $r \leq \epsilon$. Therefore, the r.h.s is increasing on the interval $[\frac{\epsilon}{2}, \infty]$ and is equal to 0 when $x = \frac{\epsilon}{2}$, which proves that no point verifies Equation (88). That allows us to conclude that: $\nu(h_{\theta_m}, q_0, x, \epsilon, \{q_0\}) = 0$.

□

B.2. Proof of proposition 1

Proposition 1 (Identifying points of underestimation). *Let $q_\sigma \sim \mathcal{N}(0, \sigma I_d)$. For any $x \in \mathbb{R}^d$ and $\sigma_1 > 0$, let g_{σ_1} be a $\{q_{\sigma_1}\}$-worst-case classifier. Then $p(x, g_{\sigma_1}, q_\sigma)$ is a decreasing function of σ over $[\sigma_1, \infty)$.*

A similar result holds for uniform distributions $(q_r)_r$ over $B_2^d(0, r)$.

Proof. Let h be the classifier of interest, $x \in \mathcal{X}, p_1 = p(x, h, q_{\sigma_1})$. We know from the proof of Theorem 1 in Cohen et al. (2019) that for any δ of norm ϵ, g_{σ_1} is a $\{q_{\sigma_1}\}$-worst classifier with optimal attack δ, where:

$$g_{\sigma_1} : z \mapsto \mathbb{1}\left\{\delta^T(z - x) \leq \sigma_1 \|\delta\| \Phi^{-1}(p_1)\right\} \tag{89}$$

where Φ is the cumulative distribution function of $\mathcal{N}(0, 1)$, *i.e.*, $\Phi(u) = \mathbb{P}[\mathcal{N}(0, 1) \leq u]$.

We see that the set $D_{\sigma_1} = \{z \in \mathcal{X} \mid g_{\sigma_1}(z) = 1\}$ is a half-space, whose supporting hyperplane is orthogonal to δ. Furthermore, let $d_1 = d(x, D_{\sigma_1})$. For any σ, we have:

$$\begin{aligned} p(x, g_{\sigma_1}, q_\sigma) &= \mathbb{P}\left[\mathcal{N}(x, \sigma I_d) \in D_{\sigma_1}\right] && (90)\\ &= \mathbb{P}\left[\mathcal{N}(0, \sigma) \leq d_1\right] && (91)\\ &= \mathbb{P}\left[\mathcal{N}(0, 1) \leq \frac{d_1}{\sigma}\right] && (92)\\ &= \Phi\left(\frac{d_1}{\sigma}\right) && (93) \end{aligned}$$

which is a decreasing function of σ, hence the result.

For uniform noises, recall that we denote q_r the uniform distribution of radius $r > 0$. As we saw in the proof of Theorem 1, for any δ of norm ϵ, the $\{q_{r_1}\}$-worst classifier g_{r_1} has a decision region D_{r_1} that is entirely contained in $B_2^d(x, r_1)$.

Hence for any $r > r_1$,

$$\begin{aligned} p(x, g_{r_1}, q_r) &= \mathbb{P}\left[U(x, r) \in D_{r_1}\right] && (94)\\ &= \frac{\mathrm{Vol}(B_2^d(0, r) \cap D_{r_1})}{\mathrm{Vol}(B_2^d(x, r))} && (95)\\ &= \frac{\mathrm{Vol}(B_2^d(x, r_1) \cap D_{r_1})}{\mathrm{Vol}(B_2^d(x, r))} && (96) \end{aligned}$$

because $B_2^d(x, r) \cap D_{r_1} \subset B_2^d(x, r_1)$. Since r_1 is constant, this is a decreasing function in r.

A similar proof works for 2-piecewise linear decision boundaries. □

B.3. Proof of Proposition 2

Proposition 2 (Growth function for concentric noises). *Let $c \geq 0$, $r_2 > r_1 > c$. Then $\Delta V(C(c, \theta), r_1, r_2)$ is a continuous, increasing function of θ, that is a bijection from $[0, \pi]$ to $[0, 2\Delta V(H(c), r_1, r_2)]$.*

The same result holds for 2-piecewise linear sets of parameter θ.

$\mathcal{A}(r_1, r_2) = H(c) \cap B_2^d(0, r_2) \setminus B_2^d(0, r_1)$ $\quad$ $\mathcal{B}_\theta(r_1, r_2) = \mathcal{C}(c, \theta) \cap B_2^d(0, r_2) \setminus B_2^d(0, r_1)$

Figure 7. Illustration of the proof of Proposition 2.

Proof of Proposition 2. Let $r_2 > r_1 > 0$, $c > 0$. Let $\mathcal{C}(c, \theta)$ be the cone of revolution of peak c and angle θ. Let $\mathcal{A}(r_1, r_2) = H(c) \cap B_2^d(0, r_2) \setminus B_2^d(0, r_1)$ and $\mathcal{B}_\theta(r_1, r_2) = \mathcal{C}(c, \theta) \cap B_2^d(0, r_2) \setminus B_2^d(0, r_1)$.

There exist a constant K such as:

$$\mathrm{Vol}(\mathcal{A}(r_1, r_2)) = \frac{1}{K} \mathrm{Vol}(H(c) \cap B_2^d(0, r_2) \setminus B_2^d(0, r_1)) \tag{97}$$

$$= \frac{1}{K} \int_{x \geq 0} \int_{r_1^2 - x^2}^{r_2^2 - x^2} \rho^{d-2} \, \mathrm{d}\rho \, \mathrm{d}x \tag{98}$$

and

$$\mathrm{Vol}(\mathcal{B}_\theta(r_1, r_2)) = \frac{1}{K} \mathrm{Vol}(\mathcal{C}(c, \theta) \cap B_2^d(0, r_2) \setminus B_2^d(0, r_1)) \tag{99}$$

$$= \frac{1}{K} \left[\int_{x = r_1 \cos\theta}^{r} \int_{r_1^2 - x^2}^{x \tan\theta} \rho^{d-2} \, \mathrm{d}\rho \, \mathrm{d}x + \right.$$

$$\left. \int_{r_1}^{r_2 \cos\theta} \int_{\rho=0}^{x \tan\theta} \rho^{d-2} \, \mathrm{d}\rho \, \mathrm{d}x + \int_{r_2 \cos\theta}^{r_2} \int_{\rho=0}^{r_2^2 - x^2} \rho^{d-2} \, \mathrm{d}\rho \, \mathrm{d}x \right] \tag{100}$$

Then:

$$\mathrm{Vol}(\mathcal{A}(r_1, r_2)) - \mathrm{Vol}(\mathcal{B}_\theta(r_1, r_2)) = \frac{1}{K} \int_{r_1 \cos\ \theta}^{r_2 \cos\theta} \int_{x \tan\theta}^{r_2^2 - x^2} \rho^{d-2} \, \mathrm{d}\rho \, \mathrm{d}x \tag{101}$$

Therefore the difference $\mathrm{Vol}(\mathcal{A}(r_1, r_2)) - \mathrm{Vol}(\mathcal{B}_\theta(r_1, r_2))$ is a continuous, decreasing function of θ over $[0, \pi]$, that is positive for $\theta \leq \frac{\pi}{2}$ and negative for $\theta \geq \frac{\pi}{2}$ by symmetry.

□

C. Proofs of Section 5

C.1. Proof of Lemma 2

Lemma 2 (Generalized Neyman-Pearson lemma Chernoff & Scheffe (1952)**).** *Let $f, q_0, \dots, q_n$ be probability density functions, and $\Phi_{\mathcal{K}}$ be any function of the form:*

$$\Phi_{\mathcal{K}} = \mathbb{1}\{\mathcal{S}_{\mathcal{K}}\}$$

Where $\mathcal{K} = \{k_1, \dots, k_n\}$ with $k_1, \dots, k_n > 0$, and $\mathcal{S}_{\mathcal{K}} = \{f(x) < \sum_{i=0}^{n} k_i q_i(x)\}$ is the generalized Neyman-Pearson set. Then $\Phi_{\mathcal{K}}$ minimises $\int \Phi f d\mu$ over all functions $0 \leq \Phi \leq 1$ such that $\int \Phi q_i d\mu \geq \int \Phi_{\mathcal{K}} q_i d\mu$ for all i.

Proof of Lemma 2. By definition of Φ_0, we have :

$$\int (\Phi - \Phi_0)(f_0 - \sum_{k=1}^{m} k_i q_i) d\mu \geq 0 \tag{102}$$

since the integrand is always positive. Hence :

$$\int (\Phi - \Phi_0) q_0 d\mu \geq \sum_{k=1}^{m} k_i \int (\Phi - \Phi_0) q_i d\mu \tag{103}$$

Since $\int (\Phi - \Phi_0) q_i d\mu \geq 0$, we have:

$$\int (\Phi - \Phi_0) q_0 d\mu \geq 0 \tag{104}$$

which is the desired result. □

C.2. Proof of Theorem 2

Theorem 2 (General approximation Theorem). *Let q_0 be the uniform noise on the ℓ_∞ ball $B_\infty(\cdot, r)$ for some $r > 0$. For any $\epsilon > 0$, $\xi > 0$ and $x \in \mathbb{R}^d$, there exists a finite family of probability density $\mathcal{Q}$ such that:*

$$\mathrm{NC}(h, q_0, x, \epsilon, \mathcal{Q}) \geq \mathrm{PC}(h, q_0, x, \epsilon) - \xi$$

Proof of Theorem 2. Let $n > 0$, and some $x \in \mathcal{X}$. We can construct a grid of $(n(r+\epsilon))^d$ squares of side size $\frac{1}{n}$, of the form $\left[\frac{a_1}{n}, \frac{a_1+1}{n}\right] \times \cdots \times \left[\frac{a_d}{n}, \frac{a_d+1}{n}\right]$ that will cover the ball $B_\infty^d(x, r)$, as well as its translation by ϵ in any direction.

Let us call the squares in this grid A_j for $j = 1 \dots m$ and $m = (\frac{d+\epsilon}{n})^d$. They all have the same volume $V_n = (\frac{1}{n})^d$.

Let q_j denote the probability density function of the uniform noise over A_j :

$$\forall j \in \{1, \dots, m\}, z \in \mathbb{R}^d, q_j(z) = \frac{1}{V_n} \mathbb{1}_{z \in A_j} \tag{105}$$

For $j \in \{1, \dots, m\}$, let $p_i = \int h(z) q_i(z) dz$ be our available information on the classifier h, and the coefficients k_j such that :

$$k_j = \begin{cases} \frac{V_n}{V} & \text{if } p_j = 1, \textit{i.e.}, h = 1 \text{ almost surely on } A_j \\ 0 & \text{otherwise.} \end{cases} \tag{106}$$

Let δ be any attack vector of norm ϵ, and $\tilde{q_0} = q_0(\cdot + \delta)$ be the distribution after attack by δ. The support of $\tilde{q_0}$ is $B_\infty^d(x + \delta, r)$ which is fully contained in $\bigcup_{i=1}^{m} A_i$.

Let Φ_0 be the Neyman-Pearson function defined by the k_i:

$$\Phi_0(x) = \begin{cases} 1 & \text{if } \tilde{q_0}(x) < \sum_{i=1}^{n} k_i q_i(x) \\ 0 & \text{otherwise} \end{cases} \tag{107}$$

We know that $\Phi_0 = 1$ outside of $B^d_\infty(x+\delta, r)$ since $\tilde{p_0} = 0$ there. Let $x \in B^d_\infty(x+\delta, r)$. The A_i are disjoint and cover the ball, so there is exactly one j such that $x \in A_j$. We then have:

$$\Phi_0(x) = \begin{cases} 1 & \text{if } h = 1 \text{ almost surely on } A_j \\ 0 & \text{otherwise} \end{cases} \tag{108}$$

Hence $\Phi_0|_{A_j} = \operatorname*{ess\,inf}_{A_j}(h)$, since h has values in $\{0, 1\}$. It follows that :

$$\int \Phi_0(z)\tilde{q_0}(z)dz = \sum_{i=1}^{m} (\operatorname*{ess\,inf}_{A_j}(h)) \operatorname{Vol}(A_i \cap B^d_\infty(x-\delta, r)) \tag{109}$$

That is a lower Riemann sum for the integral $\int\limits_{B^d_\infty(x-\delta,r)} h$, and so converges to it when $m \to \infty$ as h is Riemann integrable. Hence we can choose n such that, for any δ,

$$\int \Phi_0(z)\tilde{q_0}(z)dz \leq \int h(z)\tilde{q_0}(z)dz + \xi \tag{110}$$

which gives us the desired result, since this is true for any δ. □

D. Proofs of Section 6

D.1. Proof of Theorem 3

Theorem 3 (Sampling in low dimension for gaussian noise). *When $q_0, \dots, q_n$ are normal distributions of center 0, we can express the certificate as $\mathbb{E}[f(V)]$, where V is a random variable in dimension at most n+1, and f is a function with output in $\{0, 1\}$. V depends however on the input dimension d, and f on the Neyman-Pearson variables $k_1, \dots, k_n$.*

Proof of Theorem 3. Let $x \in \mathbb{R}^d$, $\delta = [\epsilon, 0, \dots, 0]^\top$ with $\epsilon > 0$. Let $\mathcal{Q} = \{q_0, \dots, q_n\}$ be a finite family of Gaussian probability density functions centered at 0 with variance σ_i. The noise based certificate can be written as follows:

$$\mathrm{NC}(h, q_0, x, \epsilon, \mathcal{Q}) = \mathbb{P}\left[\mathcal{N}(x + \delta, \sigma_0^2) \in \mathcal{S}\right] \tag{111}$$

where $\mathcal{S}$ is the Neyman-Person set defined as:

$$\mathcal{S} = \left\{ x \in \mathbb{R}^d \;\middle|\; \exp\left(-\frac{\|x - \delta\|^2}{2\sigma_0^2}\right) \leq \sum_{i=0}^{n} k_i \exp\left(-\frac{\|x\|^2}{2\sigma_i^2}\right) \right\} \tag{112}$$

The certificate can be expressed as:

$$\mathbb{P}\left[\mathcal{N}(x + \delta, \sigma_0^2) \in \mathcal{S}\right] = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma_0^d} \int_{\mathbb{R}^d} \exp\left(-\frac{\|x - \delta\|^2}{2\sigma_0^2}\right) \mathbb{1}_{x \in \mathcal{S}} \, \mathrm{d}x \tag{113}$$

By expressing Equation (113) with cylindrical coordinates, we have:

$$\mathbb{P}\left[\mathcal{N}(x + \delta, \sigma_0^2) \in \mathcal{S}\right] = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma_0^d} \int_{\mathbb{R}^d} \exp\left(-\frac{\|x - \delta\|^2}{2\sigma_0^2}\right) \mathbb{1}_{x \in \mathcal{S}} \, \mathrm{d}x \tag{114}$$

$$= \frac{1}{(2\pi)^{\frac{d}{2}}\sigma_0^d} \int_{\substack{\mu \in \mathbb{R} \\ r \in \mathbb{R}^+ \\ \phi_1, \dots, \phi_{d-3} \in [-\pi, \pi] \\ \phi_{d-2} \in [0, 2\pi]}} \exp\left(-\frac{r^2 + (\mu - \epsilon)^2}{2\sigma^2}\right) r^{d-2} \mathbb{1}_{(r,\mu) \in \tilde{\mathcal{S}}} \tilde{J}(\phi_1, \dots, \phi_{d-2}) \, \mathrm{d}r \, \mathrm{d}\mu \, \mathrm{d}\phi_1 \dots \mathrm{d}\phi_{d-2} \tag{115}$$

where from Blumenson (1960), the Jacobian J of the change of variables is:

$$J = r^{d-2}\tilde{J}(\phi_1, \dots, \phi_{d-2}) = r^{d-2} \prod_{k=1}^{d-2} \sin^k \phi_{d-1-k} \tag{116}$$

and where $\tilde{\mathcal{S}}$ is the updated Neyman-Person set:

$$\tilde{\mathcal{S}} = \left\{ r, \mu \in \mathbb{R} \;\middle|\; \exp\left(-\frac{r^2 + (\mu - \epsilon)^2}{2\sigma_0^2}\right) \leq \sum_{i=0}^{n} k_i \exp\left(-\frac{r^2 + \mu^2}{2\sigma_i^2}\right) \right\} \tag{117}$$

Given that the indicator function is independent of the $\phi_1, \dots, \phi_{d-2}$, we can rearrange the above equation as follows:

$$\mathbb{P}\left[\mathcal{N}(x + \delta, \sigma_0^2) \in \mathcal{S}\right] = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma_0^d} \left(\int_{\substack{\mu \in \mathbb{R} \\ r \in \mathbb{R}^+}} \exp\left(-\frac{r^2 + (\mu - \epsilon)^2}{2\sigma^2}\right) r^{d-2} \mathbb{1}_{(r,\mu) \in \tilde{\mathcal{S}}} \, \mathrm{d}r \, \mathrm{d}\mu \right) \left(\int_{\substack{\phi_1, \dots, \phi_{d-3} \in [-\pi, \pi] \\ \phi_{d-2} \in [0, 2\pi]}} \tilde{J}(\phi_1, \dots, \phi_{d-2}) \, \mathrm{d}\phi_1 \dots \mathrm{d}\phi_{d-2} \right) \tag{118}$$

By setting A as:

$$A = \int_{\substack{\phi_1,\dots,\phi_{d-3}\in[-\pi,\pi]\\ \phi_{d-2}\in[0,2\pi]}} \tilde{J}(\phi_1,\dots,\phi_{d-2})\ \mathrm{d}\phi_1 \dots \mathrm{d}\phi_{d-2} \tag{119}$$

we have:

$$\mathbb{P}\left[\mathcal{N}(x+\delta,\sigma_0^2)\in\mathcal{S}\right] = \frac{A}{(2\pi)^{\frac{d}{2}}\sigma^d}\left(\int_{\mu\in\mathbb{R}}\int_{r\in\mathbb{R}^+} \exp\left(-\frac{r^2}{2\sigma^2}\right)\exp\left(-\frac{(\mu-\epsilon)^2}{2\sigma^2}\right) r^{d-2}\mathbb{1}_{(r,\mu)\in\mathcal{S}}\ \mathrm{d}r\ \mathrm{d}\mu\right) \tag{120}$$

Finally, we can express this probability with an expected value over a Gaussian and Chi distribution:

$$\mathbb{P}\left[\mathcal{N}(x+\delta,\sigma_0^2)\in\mathcal{S}\right] = \mathbb{E}_{\substack{\mu\sim\mathcal{N}(\epsilon,\sigma^2)\\ r\sim\chi(d-1,0,\sigma_0^2)}}\left[\mathbb{1}_{(r,\mu)\in\tilde{\mathcal{S}}}\right] \tag{121}$$

which concludes the proof.

□

D.2. Proof of Theorem 4

The main idea of this proof is to bound the dot product between δ and the x_i

Lemma 5. *Let u be a random vector drawn uniformly on the unit sphere of $\mathbb{R}^d$, $\delta\in\mathbb{R}^d$ of norm ϵ. Then :*

$$\mathbb{P}\left[|\langle u,\delta\rangle| \geq \frac{\epsilon}{\sqrt{(d+2)\alpha}}\right] \leq \alpha \tag{122}$$

Proof. This is the direct consequence of Chebyshev's inequality, since $\langle u,\delta\rangle$ is a random vector of expectation 0 (by symmetry), and :

$$\begin{aligned} V(u,\delta) &= \sum_{i=1}^{d} V(\delta_i u_i) \\ &= \sum_{i=1}^{d} \delta_i^2 V(u_i) \\ &= \frac{\|\delta\|^2}{d+2} = \frac{\epsilon^2}{d+2} \end{aligned}$$

□

Lemma 6. *Let $v_1,\dots,v_n$ be random, orthogonal vectors, $e_i = \frac{v_i}{\|v_i\|}$ δ a vector of norm ϵ, and δ_P the orthogonal projection of δ on $P = \mathrm{Span}(e_1,\dots,e_n)$. Then:*

$$\mathbb{P}\left[\|\delta_P\|^2 \leq \frac{n\epsilon^2}{(d+2)\alpha}\right] \geq (1-\alpha)^n \tag{123}$$

Proof. $\|\delta_P\|^2 = \sum_{i=1}^{n}\langle\delta,v_i\rangle^2$, so we have :

$$\mathbb{P}\left[\|\delta_P\|^2 \leq \frac{n\epsilon^2}{(d+2)\alpha}\right] \geq \mathbb{P}\left[\forall i, |\langle v_i,\delta\rangle| \leq \frac{\epsilon}{\sqrt{(d+2)\alpha}}\right] \tag{124}$$

$$\geq (1-\alpha)^n \tag{125}$$

□

Theorem 4 (Asymptotic dimension-independent certificate). *Let $q_0, \dots, q_n$ are normal distribution of same variance σ, centered on points $x, z_1, \dots, z_n$, where the z_i constitute a random family of orthonormal vectors, at constant distance from x. For any $\alpha \in (0,1), \epsilon > 0$, there exists an α-probable ϵ-certificate v_d and a value $v \in [0,1]$ such that $v_d \xrightarrow[d\to\infty]{} v$ and v can be written as $\mathbb{E}[f(V)]$, where V is a random variable in dimension at most n+1 that is independent from the dimension d.*

Proof of Theorem 4. Let z_0 be the center of the Gaussian noise used at train and test time. Let z_i $(i = 1 \dots, n)$ be the centers of new Gaussian noises used to gather information. We choose the z_i as random vectors on the unit sphere, and orthonormalize them such that $(z_0 - z_i)$ are an orthogonal family.

Let $P = \mathrm{Span}(z_1, \dots, z_n)$, and $\delta = \delta_P + \delta_{P^\perp}$ where $\delta_P \in P$ and $\delta_{P^\perp} \in P^\perp$ and let $\tilde{\epsilon} = \|\delta_{P^\perp}\|$.

We use a coordinate system centered on z_0, with $\mu_i = \left\langle \frac{z_0 - z_i}{\|z_0 - z_i\|}, z - z_0 \right\rangle$ $(i = 1 \dots m)$, $t = \left\langle \frac{\delta_{P^\perp}}{\epsilon}, x - z_0 \right\rangle$, and r, ϕ_i $(i = 1 \dots d - m - 2)$ the hypercylindrical coordinates of axis $\delta_{P^\perp}$ in $P^\perp$. Let $d_i = \|z_0 - z_i\|$. Let π_P be the orthogonal projection on P, $\pi_{P^\perp}$ on $P^\perp$. Then:

$$\|z - \delta\|^2 = \|\pi_P(z - \delta)\|^2 + \|\pi_{P^\perp}(z - \delta)\|^2 \tag{126}$$
$$= \|\pi_P(z) - \delta_P\|^2 + \|\pi_{P^\perp}(z) - \delta_{P^\perp}\|^2 \tag{127}$$
$$= \|\pi_P(z)\|^2 + \|\delta_P\|^2 - 2\langle \pi_P(z), \delta_P \rangle + r^2 + (t - \tilde{\epsilon})^2 \tag{128}$$
$$= \sum \mu_j^2 + r^2 + (t - \tilde{\epsilon})^2 + \|\delta_P\|^2 - 2\langle \pi_P(z), \delta_P \rangle \tag{129}$$

We can also write :

$$\forall i = 1 \dots m, \tag{130}$$
$$\|z - z_i\|^2 = \|z - z_0\|^2 + \|z_0 - z_i\|^2 - 2\langle z - z_0, z_0 - z_i \rangle \tag{131}$$
$$= r^2 + \sum_{j=1}^{m} \mu_j^2 + t^2 + d_i^2 - 2\mu_i d_i \tag{132}$$

The certificate given by the generalized Neyman-Pearson lemma for a given δ is:

$$\tilde{p} = \mathbb{P}\left[\mathcal{N}(z_0 + \delta, \sigma^2) \in \mathcal{S}\right] \tag{133}$$
$$= K_1 \int \exp\left(-\|z - \delta\|^2\right) \mathbb{1}_{\mathcal{S}} \, \mathrm{d}z \tag{134}$$
$$= K_1 \int \exp\left(-\frac{r^2}{2\sigma^2}\right) \exp\left(-\frac{\sum_{j=1}^n \mu_j^2}{2\sigma^2}\right) \exp\left(-\frac{(t - \tilde{\epsilon})^2}{2\sigma^2}\right) \exp\left(-\frac{\|\delta_P\|^2}{2\sigma^2}\right) \exp\left(\frac{2\langle \pi_P(z), \delta_P \rangle}{2\sigma^2}\right) \mathbb{1}_{\mathcal{S}} r^{d-n-2} \, \mathrm{d}r \, \mathrm{d}t \, \mathrm{d}\mu_i \tag{135}$$
$$\geq K_1 \int a(r, \mu_j, t) \exp\left(-\frac{n\epsilon^2}{2\sigma^2(d+2)\alpha}\right) \exp\left(-\sum_j \mu_j \frac{\epsilon}{2\sigma^2\sqrt{(d+2)\alpha}}\right) \mathbb{1}_{\mathcal{S}} r^{d-n-2} \, \mathrm{d}r \, \mathrm{d}t \, \mathrm{d}\mu_i \tag{136}$$

With probability at least $(1 - \alpha)^{n+1}$, by Lemma 5.

Where:

$$a(r, \mu_j, t) = \exp\left(-\frac{r^2}{2\sigma^2}\right) \exp\left(-\frac{\sum_{j=1}^n \mu_j^2}{2\sigma^2}\right) \exp\left(-\frac{(t - \tilde{\epsilon})^2}{2\sigma^2}\right) \tag{137}$$

and

$$b(r, \mu_i, t) = \sum_{i=1}^{m} k_i \exp\left(-\frac{r^2 + \sum_{i=1}^m \mu_j^2 + t^2 + d_i^2 - 2\mu_i d_i}{2\sigma^2}\right) \tag{138}$$

Then, we need to obtain a lower subset of $\mathcal{S}$. with high probability.

$$\mathcal{S} = \left\{ r > 0, \mu_i, t \in \mathbb{R}^d \;\middle|\; a(r, \mu_j, t) \leq \exp\left(-\frac{\|\delta_P\|^2}{2\sigma^2}\right) \exp\left(\frac{2\langle z - z_0, z_0 - x_i\rangle}{2\sigma^2}\right) b(r, \mu_j, t) \right\} \tag{139}$$

$$\supset \left\{ r > 0, \mu_i, t \in \mathbb{R}^d \;\middle|\; a(r, \mu_j, t) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2(d+2)\alpha}\right) \exp\left(-\sum_j \mu_j \frac{\epsilon}{2\sigma^2\sqrt{(d+2)\alpha}}\right) b(r, \mu_j, t) \right\} \tag{140}$$

We can further simplify $\tilde{\mathcal{S}}$ by posing $c_0(n, d, \alpha) = \exp\left(-\frac{n\epsilon^2}{2\sigma^2(d+2)\alpha}\right)$, $a(d, \alpha) = \frac{\epsilon}{2\sigma^2\sqrt{(d+2)\alpha}}$ and remarking that :

$$\tilde{\mathcal{S}} = \left\{ r > 0, \mu_i, t \in \mathbb{R}^d \;\middle|\; \exp\left(-\frac{r^2}{2\sigma^2}\right) \exp\left(-\frac{\sum_{j=1}^n (\mu_j - a(d,\alpha))^2}{2\sigma^2}\right) \exp\left(-\frac{(t-\tilde{\epsilon})^2}{2\sigma^2}\right) \right.$$
$$\left. \leq c_0 \sum_{i=1}^m k_i \exp\left(-\frac{r^2 + \sum_{i=1}^m \mu_j^2 + t^2 + d_i^2 - 2\mu_i d_i}{2\sigma^2}\right) \right\} \tag{141}$$

$$= \left\{ r > 0, \mu_i, t \in \mathbb{R}^d \;\middle|\; \exp\left(-\frac{(t-\tilde{\epsilon})^2}{2\sigma^2}\right) \leq c_0 \sum_{i=1}^m k_i \exp\left(-\frac{t^2 + d_i^2 - 2\mu_i d_i}{2\sigma^2}\right) \right\} \tag{142}$$

which only depends on variables $t, \mu_1, \ldots, \mu_n$, so we will write it $\tilde{\mathcal{S}}(t, \mu_1, \ldots, \mu_n)$. Hence:

$$\tilde{p} \geq K_1 \int \exp\left(-\frac{r^2}{2\sigma^2}\right) \exp\left(-\frac{\sum_{j=1}^n \mu_j^2}{2\sigma^2}\right) \exp\left(-\frac{(t-\tilde{\epsilon})^2}{2\sigma^2}\right) \exp\left(-\frac{n\epsilon^2}{2\sigma^2 d_0 \alpha(1-\zeta)}\right) r^{d-n-2} \mathbb{1}\left\{\tilde{\mathcal{S}}(t, \mu_1, \ldots, \mu_n)\right\} \mathrm{d}r \; \mathrm{d}t \; \mathrm{d}\mu_i \tag{143}$$

$$\geq K_2 \exp\left(-\frac{n\epsilon^2}{2\sigma^2 d_0 \alpha(1-\zeta)}\right) \int \exp\left(-\frac{\sum_{j=1}^n (\mu_j - a(d,\alpha))^2}{2\sigma^2}\right) \exp\left(-\frac{(t-\tilde{\epsilon})^2}{2\sigma^2}\right) \exp\left(-\frac{n\epsilon^2}{2\sigma^2 d_0 \alpha(1-\zeta(\alpha))}\right) \mathbb{1}_{\tilde{\mathcal{S}}} \; \mathrm{d}t \; \mathrm{d}\mu_i \tag{144}$$

$$\geq \mathbb{E}_{\substack{\mu_j \sim \mathcal{N}(a(d,\alpha), \sigma^2) \\ t \sim \mathcal{N}(\epsilon, \sigma^2)}} \left[\mathbb{1}\left\{\tilde{\mathcal{S}}(t, \mu_1, \ldots, \mu_n)\right\}\right] \tag{145}$$

Where $K_2 = \frac{K_1}{\int \exp\left(-\frac{r^2}{2\sigma^2}\right) r^{d-2} \; \mathrm{d}r}$, which corresponds to removing the normalization of the chi law.

When $d \to \infty$, we have $c_0(n, d, \alpha) \to 1$ and $a(d, \alpha) \to 0$. Let :

$$\mathcal{S}_0 = \left\{ r > 0, \mu_i, t \in \mathbb{R}^d \;\middle|\; \exp\left(-\frac{(t-\tilde{\epsilon})^2}{2\sigma^2}\right) \leq \sum_{i=1}^m k_i \exp\left(-\frac{t^2 + d_i^2 - 2\mu_i d_i}{2\sigma^2}\right) \right\} \tag{146}$$

Then

$$\mathbb{E}_{\substack{\mu_j \sim \mathcal{N}(a(d,\alpha), \sigma^2) \\ t \sim \mathcal{N}(\epsilon, \sigma^2)}} \left[\mathbb{1}\left\{\tilde{\mathcal{S}}(t, \mu_1, \ldots, \mu_n)\right\}\right] \xrightarrow[d \to \infty]{} \mathbb{E}_{\substack{\mu_j \sim \mathcal{N}(0, \sigma^2) \\ t \sim \mathcal{N}(\epsilon, \sigma^2)}} \left[\mathbb{1}\left\{\mathcal{S}_0(t, \mu_1, \ldots, \mu_n)\right\}\right] \tag{147}$$

which can be computed independently of the dimension. □

D.3. Proof of proposition 3

Proposition 3 (*Computing the k_i for uniform noises*). *Let $q_0, \ldots, q_n$ are uniform distributions, where $n \ll d$ there are only at most 2^n possible values for the generalized Neyman-Pearson set $\mathcal{S}$.*

Proof. Let $A_0, \ldots, A_n$ be the sets such that $q_i = \frac{\mathbb{1}_{A_i}}{\mathrm{Vol}(A_i)}$. The Neyman-Pearson set becomes :

$$S = \left\{ z \in \mathbb{R}^d \mid \frac{\mathbb{1}_{A_0}}{\mathrm{Vol}(A_0)} \leq \sum_{i=1}^{n} k_i \frac{\mathbb{1}_{A_i}}{\mathrm{Vol}(A_i)} \right\}$$
$$= \left\{ z \in \mathbb{R}^d \mid \mathbb{1}_{A_0} \leq \sum_{i=1}^{n} K_i \mathbb{1}_{A_i} \right\}$$

Where $K_i = k_i \frac{\mathrm{Vol}(A_0)}{\mathrm{Vol}(A_i)}$. We always have $A_0^c \subset S$, since the left-hand side is 0 there.

The right-hand site is piecewise constant, and can only take the value 1 or 0 on each intersection of A_i, for a maximum of 2^n possibilities. □