Randomized smoothing and information geometry

For the moment, I will abandon randomized smoothing. Here are the reasons.

- Randomized smoothing is a statistical "Monte-Carlo" method. Thus, the certification is only "with high probability". Moreover, the MC sampling may suffer from the curse of dimensionality.
- Randomized smoothing is a post-hoc method (by opposition to by-design methods such as Lipschtz neural networks). It is a strength, because it doesn't depend on the specificities of the base model. But it is also a weakness, because it doesn't help to design robust base model. It won't help us answer the question of how to design models that "understand the task". Randomized smoothing is a certification method, not a robustification method.
- One motivation to study randomized smoothing was that it might be reformulated using information geometry. However, I have recently discover that information geometry doesn't bring much to the statistical theory (only the e and m-connections, which are only useful for providing new intuitions or reformulations, as far as I know). Thus, using information geometry to reformulate randomized smoothing seems useless.
- I believe that certified robustness (in the sense of stability to small perturbations, generally using L_p norms) is not a useful goal. Stability is not a good property. A stable model doesn't understand the task more than an unstable one (as shown by invariance attacks). Using empirical or certified methods to measure stability is not a good evaluation of robustness.

1 Literature review about randomized smoothing

1.1 Cohen et al.

[1]

As far as I know, randomized smoothing is the only certified defense that is scalable to large models. The idea is very simple. Suppose you have a base classifier $F : \mathbb{R}^n \to \mathcal{Y}$ for a classification task. Let $x \in \mathbb{R}^n$. Choose a parameter $\sigma > 0$ and consider the probability space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_x)$ where the distribution of $\mathbb{P}_x$ is $\mathcal{N}(x, \sigma^2 I)$.

We can define a new smoothed classifier $G : \mathbb{R}^n \to \mathcal{Y}$ (that is *not* a neural network) by:

$$G(x) = \arg\max_c \mathbb{P}_x(F^{-1}(c)),$$

It can be shown that, if there is a class $c_A \in \mathcal{Y}$ and two probabilities $\underline{p_A}, \overline{p_B}$ such that:

$$\mathbb{P}_x(F^{-1}(c_A)) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}_x(F^{-1}(c)), \tag{1}$$

then we have $G(x + \delta) = c_A$ for all $\|\delta\|_2 \leq R$ where:

$$R = \frac{\sigma}{2} \left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}) \right), \tag{2}$$

with Φ^{-1} the quantile function of the standard normal distribution.

Moreover, Cohen et al. [1] showed that this bound is tight, i.e., if $\underline{p_A} + \overline{p_B} \leq 1$, then for any perturbation δ such that $\|\delta\|_2 \geq R$, there exists a classifier F verifying equation 1 for which $G(x + \delta) \neq c_A$. Hence, the set of perturbations to which the smoothed classifier is provably robust is precisely an l_2 ball.

Thus, smoothing with other distributions (other than Gaussian) may lead to certified guarantees for other l_p norm, or even other type of attacks. Hao et al. may have done this for spatial attacks [2].

The radius depends linearly on σ. But most importantly, the radius is larger when $\underline{p_A}$ is close to 1 and $\overline{p_B}$ is close to 0. The higher is σ, the less likely this will be true, thus there is a trade-off. This trade-off depends on the ability of the model F to be robust to Gaussian noise. Hence, this result shows a direct link between robustness to Gaussian noise and certified robustness to adversarial attacks.

In practice, it is impossible to compute $\mathbb{P}_x(F^{-1}(c))$. There are two problems:

1. Prediction $\rightarrow$ estimate what class is predicted by G.
2. Certification $\rightarrow$ estimate the radius R

For prediction, the authors of [1] propose a method to estimate $\mathbb{P}_x(F^{-1}(c))$ with Monte Carlo sampling. The method guarantees that the predicted class for $G(x)$ is not incorrect with probability $1-\alpha$ where α is a chosen risk. If the model is unsure, it will abstain to make any prediction. For a fixed α, in order to avoid abstaining, we must sample a large number of points. The prediction methods is the following (I'm not sure if these algorithms are for binary classification or multiclass classification, I guess they are for binary classification):

1. Sample N points from $f(\mathcal{N}(x, \sigma^2 I))$.
2. Let N_A and N_B be the number of points in N associated with largest and second largest classes.
3. For c_A to be the predicted class of G, the probability of obtained c_A from the $N_A + N_B$ points of the first two classes must be greater than $1/2$. So we compute the probability (p-value) that there are N_A successes among $N_A + N_B$ trials of a Binomial distribution with parameter $1/2$. If this p-value is smaller than α, return c_A, else, abstain.

The certification is a little bit more complicated. To compute equation 2, we need to evaluate a lower bound $\underline{p_A}$. Then, we set $\overline{p_B} = 1 - \underline{p_A}$.

1. Sample a small number N_0 of points from $f(\mathcal{N}(x, \sigma^2 I))$.
2. Take a guess for c_A as the largest class among the N_0 points.
3. Sample a larger number N of points from $f(\mathcal{N}(x, \sigma^4 I))$ (I don't know if there is an error here in [1] and it should be σ^2 instead of σ^4). Let N_A be the number of c_A among the N points.
4. Compute the confidence interval $[p, +\infty)$ of the Binomial distribution with N trials, N_A successes, and risk α. We set $\underline{p_A} = p$.
5. If $\underline{p_A} > 1/2$, return the prediction c_A and the radius $\sigma \Phi^{-1}(\underline{p_A})$, else, abstain.

How can we extend randomized smoothing to regression (maybe it has already been done)? Can we interpret randomized smoothing with information geometry?

1.2 Ettedgui et al.

[3].

3 objectives when choosing the base classifier:

1. Choose base classifier that maximize accuracy of smoothed classifier $\rightarrow$ it may be possible to bypass the trade-off between variance of smoothing (to obtain large radii) and the accuracy
2. Choose base classifier that maximize the certified radii of the smoothed classifier (in some sense that must be spelled out since there are always points with radius 0) $\rightarrow$ or equivalently, maximize the perfect certificate (again in some sense)

3. Choose base classifier that minimize the computational cost of certification process → approximate the perfect certificate as best as possible, with as less as possible computation (trade-off) → base classifier with "good" local curvature

Problems:

- Generalize Proposition 2 for piecewise linear base classifier → how to choose the smoothing distributions to approximate the perfect certificate?
- Explicit what is "local curvature of decision boundary".

1.3 Sukenik et al.

[4].

The authors prove several theoretical results about input-dependent randomized smoothing (IDRS). We are in the framework of Cohen et al. [1] but with a function $\sigma(x)$ instead of a constant σ. The authors don't investigate much how to choose the function $\sigma(x)$. They assume that $\sigma(x)$ is given and investigate how to derive tight certified radii, both in theory and in practice.

Moreover, they show that IDRS suffer from curse of dimensionality in the sense that the certified radius for a design $\sigma(x)$ will shrink when the dimension increases, except if the function $\sigma(x)$ is r-semi-elastic[1] with very small r, i.e., the function $\sigma(x)$ is almost constant and thus we loose any benefit of using an input-dependent scheme.

I don't really follow their derivations, so I don't know if this is a definitive no-go for IDRS. Maybe, more investigation on how to choose $\sigma(x)$ may help. Intuitively, $\sigma(x)$ should be large when far from decision boundary and small when close to decision boundary.

Note that there are other works on IDRS but more empirical. I read this one: [5]. Here, the authors use backprop and gradient descent to directly optimize $\sigma(x)$ to obtain large certified radius. In order to achieve certification, they use a very, very heuristic method which is certainly not tight at all. But they claim that their method achieves larger radii than Cohen *et al.*[1] with constant σ.

Last paragraph of the paper: "*The most intriguing and promising direction for the future work lies in the development of new $\sigma(x)$ functions, which could treat the mentioned issues even more efficiently. This improvement is necessary to make IDRS able to significantly beat constant smoothing, as it happens in our toy example in Appendix A. The difference between CIFAR10 and MNIST and the toy dataset is that $\sigma(x)$ from equation 1 is well-suited for the toy dataset's geometry, but not to the same extent for the geometry of images. One possible reason is that this $\sigma(x)$ does not correspond to the distances from the decision boundary across the entire dataset, because the euclidean distances between images do not align with the "distances in images' content". We believe, however, that improvements in $\sigma(x)$ design are possible and we leave the investigation in this matter as an open and attractive research question. One way to go is to define a metric on the input space that would better reflect the geometry of images and convolutional neural networks.*"

1.4 Hao et al.

[2]

The authors extend randomized smoothing to semantic attacks (unrestricted attacks that are not l_p-based). They propose different results depending on the properties of the semantic attack.

1.5 Mohapatra et al. (a)

[6]

The authors study the randomized smoothing method from Cohen et al. [1]. They show that randomized smoothing tends to be biased against certain classes when the noise σ grows. The certified radius and the accuracy *of certain classes* shrink. This shrinking effect is highly dependent on the "geometry" of the dataset.

[1] I don't know this notion but it seems more or less related to Lipschitz, except using log, if that makes sense.

2 Randomized smoothing

The challenge will be to find the good formulation. It should be general enough to provide flexibility and solve all the limits of current randomized smoothing methods, but not too general in order for us to prove something valuable. We will take inspiration from existing generalizations such as: [3, 2, 7].

2.1 General formulation

Let $\mathcal{X}$ be a measurable subset of $\mathbb{R}^d$.
Let $\mathcal{Y} = \{-1, 1\}$ be a set of classes (for the moment, we only consider binary classification).
Let $\mathcal{H} = \{F : \mathcal{X} \to \mathcal{Y}\}$ be a family of *base classifiers*.
Define the *smoothing map* $\Phi : \mathcal{H} \to \mathcal{F}(\mathcal{S}(\mathcal{X}), \mathcal{P}(\mathcal{Y}))$ where:

- $\mathcal{S}(\mathcal{X})$ is a statistical manifold of distributions sampled from $\mathcal{X}$,
- $\mathcal{P}(\mathcal{Y})$ is the statistical manifold of distributions on $\mathcal{Y}$, i.e., the family of Bernoulli distributions.
- $\mathcal{F}(\mathcal{S}(\mathcal{X}), \mathcal{P}(\mathcal{Y}))$ is the set of maps from $\mathcal{S}(\mathcal{X})$ to $\mathcal{P}(\mathcal{Y})$.

Let $F \in \mathcal{H}$ be a base classifier.
Define the *smoothed model* ΦF of F:
for $p \in \mathcal{S}(\mathcal{X})$, $\Phi F(p)$ is a Bernoulli distribution defined by:

$$\Phi F(p) = \mathbb{E}_p[\mathbb{1}\{F = 1\}] = \int_{\mathcal{X}} \mathbb{1}\{F(x) = 1\} \mathrm{d}p(x). \tag{3}$$

Define a *smoothing manifold* $\mathcal{M} = \{p_x, x \in \mathcal{X}\} \subseteq \mathcal{S}(\mathcal{X})$ parameterized by $\mathcal{X}$.
$\mathcal{M}$ is an embedded submanifold of $\mathcal{S}(\mathcal{X})$.
Define the *smoothing classifier* $\tilde{F} : \mathcal{X} \to \mathcal{Y}$:
for all $x \in \mathcal{X}$, $\tilde{F}$ predicts:

$$\tilde{F}(x) = \begin{cases} 1 & \text{if } \Phi F(p_x) > \dfrac{1}{2}, \\ -1 & \text{otherwise.} \end{cases} \tag{4}$$

For example, the smoothing scheme presented in [1] consists in fixing $\sigma > 0$ and choosing $p_x = \mathcal{N}_d(x, \sigma^2 I_d)$.
Define a *certification scheme* for a fixed smoothing manifold $\mathcal{M}$.
Let $x \in \mathcal{X}$.
Let $\mathcal{B}(x, r)$ be a set of allowed perturbations of x.
r is the maximum distance between x and any perturbation, using some distance function.
For example, $\mathcal{B}(x, r)$ can be the l_2 ball of radius r.
The *certified radius* R is defined by:

$$R = \sup\{r : \forall z \in \mathcal{B}(x, r), \forall H \in \mathcal{E}(F), \tilde{H}(z) = \tilde{H}(x)\}. \tag{5}$$

The certified radius is obtained with the Neyman-Pearson lemma.
The certified radius relies on the *certification family* $\mathcal{E}(F) \subseteq \mathcal{H}$ at x.
Note that we must have $F \in \mathcal{E}(F)$.
We aim to study:

1. What properties of $\mathcal{E}(F)$ impact the certified radius.
2. How to achieve these properties.
3. What is the computational cost of the certification scheme.

For example, in [8], $\mathcal{E}(F) = \{H \in \mathcal{H} : \Phi H(p_x) = \Phi F(p_x)\}$.
In [3], $\mathcal{E}(F) = \cap_{q \in \mathcal{Q}} \{H \in \mathcal{H} : \Phi H(q) = \Phi F(q)\}$, where $\mathcal{Q}$ is a well-chosen family of distributions.
Here, there is a trade-off between finding the smallest family $\mathcal{Q}$ (to keep the computation cost reasonable) and achieving the smallest certification family $\mathcal{E}(F)$.
Informally, smaller certification families $\mathcal{E}(F)$ lead to larger certified radii.
Note that input-dependent (IDRS) or anisotropic (ARS) randomized smoothing are consequences of the choice of the smoothing manifold $\mathcal{M}$.

2.2 The Neyman-Pearson lemma

Even if other methods have been proposed to derive certified radius with randomized smoothing, it seems that all recent papers rely on the Neyman-Pearson lemma (NPL) or some generalization of it to derive the certified radius. The reason is that the NPL leads to tight radius. So, in a sense, the NPL is the heart of randomized smoothing.

Unfortunately, I'm still very confused on how exactly the NPL is used to derive certified radii in a general context. So, we will start by stating the NLP, and see how it is applied in Cohen et al. [1] which is the first paper to apply the NPL to randomized smoothing. Then, we will see how it is extended to more general framework, in particular in Yang et al. [7], Hao et al. [2], Ettedgui et al. [3].

We will also look for a geometric interpretation of the NPL. Fortunately, some clues are already available in Eguchi & Copas [9].

2.2.1 The original Neyman-Pearson lemma

The original NPL is about hypothesis testing, so I will present it in this context. But note that this is only an interpretation of the NPL, and it can be used in other context.

Let $\mathcal{X}$ be a sample space and consider two hypothesis:

- H is the **null hypothesis** with distribution $\mathbb{P}_H$ and density p_H.
- A is the **alternative hypothesis** with distribution $\mathbb{P}_A$ and density p_A.

We assume that both distributions have support equal to $\mathcal{X}$.
Let X be a random variable which probability distribution is either $\mathbb{P}_H$ or $\mathbb{P}_A$.
A **test** $T : \mathcal{X} \to \{0, 1\}$ is a measurable function that takes in input a realization $z \in \mathcal{X}$ of X and returns 1 if it decides that z has been sampled from $\mathbb{P}_A$, and 0 if it decides that z has been sampled from $\mathbb{P}_H$.
The **risk** of a test T is defined as $\mathbb{P}_H(T(X) = 1)$, i.e., the probability of rejecting H while H is true. This is a Type I error.
The **power** of a test T is defined as $\mathbb{P}_A(T(X) = 1)$, i.e., the probability of rejecting H while A is true. This is the opposite of a Type II error.
It is commonly assumed that Type I error are worse than Type II error. Thus, we will fix an acceptable risk and then look for the test T which maximizes the power for the given risk.

Theorem 2.1 (Neyman-Pearson lemma)**.**
Let $\alpha \in (0, 1)$. The most powerful test T such that $\mathbb{P}_H(T(X) = 1) = \alpha$ is

$$T(z) = \mathbb{1}\left\{\frac{p_A(z)}{p_H(z)} > t\right\}, \tag{6}$$

for some t chosen to respect the risk constraint.

Proof. Let $W = \left\{z \in \mathcal{X} : \frac{p_A(z)}{p_H(z)} > t\right\}$ where t is chosen[2] such that $\mathbb{P}_H(W) \geq \alpha$. Now, let W^* be any other measurable subset of $\mathcal{X}$ such that $\mathbb{P}_H(W^*) \leq \alpha$. We have $\mathbb{P}_A(W) = \mathbb{P}_A(W \setminus W^*) + \mathbb{P}_A(W \cap W^*)$ and $\mathbb{P}_A(W^*) = \mathbb{P}_A(W^* \setminus W) + \mathbb{P}_A(W \cap W^*)$, thus:

$$\begin{aligned}\mathbb{P}_A(W) - \mathbb{P}_A(W^*) = \mathbb{P}_A(W \setminus W^*) - \mathbb{P}_A(W^* \setminus W) &= \int_{W \setminus W^*} p_A(z)dz - \int_{W^* \setminus W} p_A(z)dz, \\ &\geq t\left(\int_{W \setminus W^*} p_H(z)dz - \int_{W^* \setminus W} p_H(z)dz\right) = t(\mathbb{P}_H(W) - \mathbb{P}_H(W^*)) \geq 0.\end{aligned}$$

□

[2]Note that such a t necessarily exists in $(0, +\infty)$ because $\lim_{t \to 0} \mathbb{P}_H(W) = 1$ and $\lim_{t \to +\infty} \mathbb{P}_H(W) = 0$.

2.3 Certified radius against Euclidean adversary using Gaussian noise

Cohen et al. [1] provide a slightly more general version of the NPL. Here, I will purposefully formulate it without mention to hypothesis testing.

Theorem 2.2 (Slightly generalized Neyman-Pearson lemma)**.**
Let X and Y be two random variables with support on $\mathcal{X}$, with densities p_X and p_Y, and distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$.

1. *Let $W_1 = \left\{z \in \mathcal{X} : \frac{p_Y(z)}{p_X(z)} > t\right\}$ with $t > 0$. Let W_1^* be such that $\mathbb{P}_X(W_1^*) \leq \mathbb{P}_X(W_1)$. Then $\mathbb{P}_Y(W_1^*) \leq \mathbb{P}_Y(W_1)$.*
2. *Let $W_2 = \left\{x \in \mathcal{X} : \frac{p_Y(z)}{p_X(z)} \leq t\right\}$ with $t > 0$. Let W_2^* be such that $\mathbb{P}_X(W_2^*) \geq \mathbb{P}_X(W_2)$. Then $\mathbb{P}_Y(W_2^*) \geq \mathbb{P}_Y(W_2)$.*

The proof is identical to the proof of Theorem 2.1.

Example 2.3. Let us apply Theorem 2.2 to $X \sim \mathcal{N}(x, \sigma^2 I)$ and $Y \sim \mathcal{N}(x+\delta, \sigma^2 I)$ with $\mathcal{X} = \mathbb{R}^d$. We have:

$$\begin{aligned}\frac{p_Y(z)}{p_X(z)} &= \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^d (z_i - (x_i+\delta_i))^2\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^d (z_i - x_i)^2\right)} \\ &= \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^d 2z_i\delta_i - \delta_i^2 - 2x_i\delta_i\right) \\ &= \exp(a\delta^T z + b)\end{aligned}$$

where $a = \frac{1}{\sigma^2}$ and $b = -\frac{2\delta^T x + \|\delta\|^2}{2\sigma^2}$ are constants for x. Then:

$$\frac{p_Y(z)}{p_X(z)} > t \Leftrightarrow \exp(a\delta^T z + b) > t \Leftrightarrow \delta^T z > \frac{\ln(t) - b}{a}.$$

Thus, for $\beta = \frac{\ln(t)-b}{a}$, we can apply Theorem 2.2 with $W_1 = \{z : \delta^T z > \beta\}$ and $W_2 = \{z : \delta^T z \leq \beta\}$.

2.3.1 Certified radius for multiclass classification

Now, we can derive the certified radius.

Theorem 2.4 (Certified radius with Gaussian noise for multiclass classification)**.**
Let $F : \mathbb{R}^d \to \mathcal{Y}$ be a base classifier. Let $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let $x \in \mathbb{R}^d$.
Define the smoothed classifier by $G(x) = \arg\max_k \mathbb{P}(F(x+\epsilon) = k)$.
Assume that there exist $k_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ such that:

$$\mathbb{P}(F(x+\epsilon) = k_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{k \neq k_A} \mathbb{P}(F(x+\epsilon) = k).$$

Then, for all $\|\delta\|_2 < R$, we have $G(x+\delta) = k_A$ where:

$$R = \frac{\sigma}{2}\left(\phi^{-1}(\underline{p_A}) - \phi^{-1}(\overline{p_B})\right).$$

Proof. Let $k_B \neq k_A$. Let $X = x + \epsilon \sim \mathcal{N}(x, \sigma^2 I)$ and $Y = x + \delta + \epsilon \sim \mathcal{N}(x+\delta, \sigma^2 I)$.
Let $W_1^* = \{z : F(z) = k_B\}$ and $W_2^* = \{z : F(z) = k_A\}$.
It suffices to show that $\mathbb{P}(F(Y) = k_B) = \mathbb{P}_Y(W_1^*) < \mathbb{P}_Y(W_2^*) = \mathbb{P}(F(Y) = k_A)$.
The idea is to find two sets W_1 and W_2 as in Example 2.3 such that:

$$\begin{aligned}\mathbb{P}_X(W_1^*) &\leq \mathbb{P}_X(W_1) \\ \mathbb{P}_X(W_2) &\leq \mathbb{P}_X(W_2^*)\end{aligned}$$

Then, using Theorem 2.2, we can conclude that:

$$\mathbb{P}_Y(W_1^*) \leq \mathbb{P}_Y(W_1)$$
$$\mathbb{P}_Y(W_2) \leq \mathbb{P}_Y(W_2^*)$$

Then, we will see under what condition does the inequality $\mathbb{P}_Y(W_1) < \mathbb{P}_Y(W_2)$ hold, such that:

$$\mathbb{P}_Y(W_1^*) \leq \mathbb{P}_Y(W_1) < \mathbb{P}_Y(W_2) \leq \mathbb{P}_Y(W_2^*).$$

By assumption, we have $\mathbb{P}_X(W_2^*) \geq \underline{p_A}$ and $\mathbb{P}_X(W_1^*) \leq \overline{p_B}$. So we will look for W_1 and W_2 such that $\mathbb{P}_X(W_2) = \underline{p_A}$ and $\mathbb{P}_X(W_1) = \overline{p_B}$. Let us start with W_2. X is a Gaussian vector with independent components, hence $\delta^T X$ is a univariate Gaussian distribution with $\mathbb{E}[\delta^T X] = \delta^T x$ and $\text{Var}[\delta^T X] = \sigma^2 \|\delta\|_2^2$. Let $U = \frac{\delta^T X - \delta^T x}{\sigma \|\delta\|_2} \sim \mathcal{N}(0,1)$. We are looking for β such that:

$$\begin{aligned}
\underline{p_A} &= \mathbb{P}_X(\{z : \delta^T z \leq \beta\}) \\
&= \mathbb{P}(\delta^T X \leq \beta) \\
&= \mathbb{P}\left(U \leq \frac{\beta - \delta^T x}{\sigma \|\delta\|_2}\right) \\
&= \phi\left(\frac{\beta - \delta^T x}{\sigma \|\delta\|_2}\right) \\
&\Rightarrow \beta = \sigma \|\delta\|_2 \phi^{-1}(\underline{p_A}) + \delta^T x.
\end{aligned}$$

Thus:

$$W_2 = \{z : \delta^T (z - x) \leq \sigma \|\delta\|_2 \phi^{-1}(\underline{p_A})\}.$$

Similarly, we can show that:

$$W_1 = \{z : \delta^T (z - x) > \sigma \|\delta\|_2 \phi^{-1}(1 - \overline{p_B})\}.$$

To complete the proof, we look for a condition such that $\mathbb{P}_Y(W_1) < \mathbb{P}_Y(W_2)$. We have $\mathbb{E}[\delta^T Y] = \delta^T x + \|\delta\|_2^2$ and $\text{Var}[\delta^T Y] = \sigma^2 \|\delta\|_2^2$. Then:

$$\begin{aligned}
\mathbb{P}_Y(W_2) &= \mathbb{P}(\delta^T Y \leq \sigma \|\delta\|_2 \phi^{-1}(\underline{p_A}) + \delta^T x) \\
&= \mathbb{P}\left(U \leq \phi^{-1}(\underline{p_A}) - \frac{\|\delta\|}{\sigma}\right) \\
&= \phi\left(\phi^{-1}(\underline{p_A}) - \frac{\|\delta\|}{\sigma}\right).
\end{aligned}$$

Similarly, we get:

$$\mathbb{P}_Y(W_1) = \phi\left(\phi^{-1}(\overline{p_B}) + \frac{\|\delta\|}{\sigma}\right).$$

Now, we can see that $\mathbb{P}_Y(W_1) < \mathbb{P}_Y(W_2)$ if and only if:

$$\|\delta\|_2 < \frac{\sigma}{2}\left(\phi^{-1}(\underline{p_A}) - \phi^{-1}(\overline{p_B})\right).$$

□

The tightness of this certified radius can be proved by building, for each $\|\delta\|_2 > R$, a base classifier F^* such that $\mathbb{P}(F^*(x + \epsilon) = k_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{k \neq k_A} \mathbb{P}(F^*(x + \epsilon) = k)$, yet its smoothed classifier verifies $G^*(x + \delta) \neq k_A$. The base classifier F^* can be defined as follows:

$$F^*(x) = \begin{cases} k_A & \text{if } x \in W_2 \\ k_B & \text{if } x \in W_1 \\ \text{other classes} & \text{otherwise} \end{cases}$$

Note that this classifier is well defined if $\underline{p_A} + \overline{p_B} \leq 1$. This is always possible since we can set $\overline{p_B} = 1 - \underline{p_A}$.

2.3.2 Certified radius for binary classification

In order to focus on the important ideas of the proof (and ignore needless details), we reformulate the result for binary classification. Moreover, we change the notations in preparation of the more general framework that will be introduced later. We also apply the NPL with the log ratio of densities, in preparation of the connections with KL-divergence and information geometry that will be uncovered later.

Theorem 2.5 (Certified radius with Gaussian noise for binary classification)**.**
Let $F : \mathbb{R}^d \to \{0, 1\}$ be a base classifier. Let q_0 be the density of $\mathcal{N}(0, \sigma^2 I)$. Let $x \in \mathbb{R}^d$.
Define the smoothed classifier by:

$$\Phi F(x) = \mathbb{1}\left\{\int_{\mathbb{R}^d} F(z)q_0(z-x)dz > \frac{1}{2}\right\}.$$

Assume that there exists α such that:

$$\frac{1}{2} < \alpha \leq \int_{\mathbb{R}^d} F(z)q_0(z-x)dz.$$

Then, for all $\|\delta\|_2 < R$, we have $\Phi F(x+\delta) = 1$ where:

$$R = \sigma\phi^{-1}(\alpha).$$

Proof. We are looking for a set $W(\delta)$ of the form:

$$W(\delta) = \left\{z \in \mathbb{R}^d : \ln \frac{q_0(z-x-\delta)}{q_0(z-x)} \leq u(\delta)\right\},$$

with $u(\delta) \in \mathbb{R}$, and such that[3]:

$$\int_{W(\delta)} q_0(z-x)dz = \alpha.$$

Since $q_0(\cdot - x) \sim \mathcal{N}(x, \sigma^2 I)$ and $q_0(\cdot - x - \delta) \sim \mathcal{N}(x+\delta, \sigma^2 I)$, we have:

$$\ln \frac{q_0(z-x-\delta)}{q_0(z-x)} = \frac{\delta^T\left(z - x - \frac{\delta}{2}\right)}{\sigma^2}.$$

For $z \sim q_0(\cdot - x)$, we have that $\ln \frac{q_0(z-x-\delta)}{q_0(z-x)}$ is a univariate Gaussian distribution with $\mathbb{E}\left[\ln \frac{q_0(z-x-\delta)}{q_0(z-x)}\right] = -\frac{\|\delta\|_2^2}{2\sigma^2}$ and $\text{Var}\left[\ln \frac{q_0(z-x-\delta)}{q_0(z-x)}\right] = \frac{\|\delta\|_2^2}{\sigma^2}$. Hence $\int_{W(\delta)} q_0(z-x)dz = \alpha$ is equivalent to:

$$\phi\left(\frac{\sigma}{\|\delta\|_2}\left(u(\delta) + \frac{\|\delta\|_2^2}{2\sigma^2}\right)\right) = \alpha.$$

Thus:

$$u(\delta) = \frac{\|\delta\|_2}{\sigma}\phi^{-1}(\alpha) - \frac{\|\delta\|_2^2}{2\sigma^2}.$$

Now, the NPL (Theorem 2.2) yields:

$$\int_{W(\delta)} q_0(z-x-\delta)dz \leq \int_{\mathbb{R}^d} F(z)q_0(z-x-\delta)dz.$$

For $z \sim q_0(\cdot - x - \delta)$, we have $\mathbb{E}\left[\ln \frac{q_0(z-x-\delta)}{q_0(z-x)}\right] = \frac{\|\delta\|_2^2}{2\sigma^2}$ and $\text{Var}\left[\ln \frac{q_0(z-x-\delta)}{q_0(z-x)}\right] = \frac{\|\delta\|_2^2}{\sigma^2}$. Hence:

$$\begin{aligned}\int_{W(\delta)} q_0(z-x-\delta)dz &= \phi\left(\frac{\sigma}{\|\delta\|_2}\left(u(\delta) - \frac{\|\delta\|_2^2}{2\sigma^2}\right)\right) \\ &= \phi\left(\phi^{-1}(\alpha) - \frac{\|\delta\|_2}{\sigma}\right)\end{aligned}$$

[3]In fact, we want $W(\delta)$ such that: $\frac{1}{2} < \int_{W(\delta)} q_0(z-x)dz \leq \int_{\mathbb{R}^d} F(z)q_0(z-x)dz$. Then we can define: $\alpha = \int_{W(\delta)} q_0(z-x)dz$.

Finally, we have $\Phi F(x+\delta) > 1/2$ if:

$$\begin{aligned}
&\int_{W(\delta)} q_0(z-x-\delta)dz > \frac{1}{2} \\
\Leftrightarrow \quad & \phi\left(\phi^{-1}(\alpha) - \frac{\|\delta\|_2}{\sigma}\right) > \frac{1}{2} \\
\Leftrightarrow \quad & \phi^{-1}(\alpha) - \frac{\|\delta\|_2}{\sigma} > 0 \\
\Leftrightarrow \quad & \sigma\phi^{-1}(\alpha) > \|\delta\|_2.
\end{aligned}$$

□

Let $\mathcal{H}$ be the set of measurable functions from $\mathbb{R}^d$ to $\{0,1\}$. For $\alpha > 1/2$, define the family of base classifiers:

$$\mathcal{F}(\alpha) = \left\{ G \in \mathcal{H} : \int_{\mathbb{R}^d} G(z) q_0(z-x)dz \geq \alpha \right\}.$$

Theorem 2.5 can be extended into the following equivalence:

$$\inf_{G \in \mathcal{F}(\alpha)} \int_{\mathbb{R}^d} G(z) q_0(z-x-\delta)dz > \frac{1}{2} \Leftrightarrow \|\delta\|_2 < \sigma\phi^{-1}(\alpha).$$

It is an equivalence because, given δ such that $\|\delta\|_2 \geq \sigma\phi^{-1}(\alpha)$, we can define the linear classifier $F^*(z) = \mathbb{1}\{z \in W(\delta)\}$. By definition of $W(\delta)$, we have:

$$\begin{aligned}
\Phi F^*(x) &= \mathbb{1}\left\{ \int_{\mathbb{R}^d} F^*(z) q_0(z-x)dz > \frac{1}{2} \right\} \\
&= \mathbb{1}\left\{ \int_{W(\delta)} q_0(z-x)dz > \frac{1}{2} \right\} \\
&= \mathbb{1}\left\{ \alpha > \frac{1}{2} \right\} \\
&= 1,
\end{aligned}$$

but:

$$\begin{aligned}
\Phi F^*(x+\delta) &= \mathbb{1}\left\{ \int_{\mathbb{R}^d} F^*(z) q_0(z-x-\delta)dz > \frac{1}{2} \right\} \\
&= \mathbb{1}\left\{ \int_{W(\delta)} q_0(z-x-\delta)dz > \frac{1}{2} \right\} \\
&= 0,
\end{aligned}$$

because $\|\delta\|_2 \geq \sigma\phi^{-1}(\alpha)$.

The strength of randomized smoothing is to derive a link between the **margin** α of the smoothed classifier ΦF at x, and the minimum perturbation size R required to change the prediction of ΦF.

2.4 Certified radius with multiple distributions

Ettedgui et al. [3] rely on a generalized version of the NPL for an arbitrary number of densities, first proved by Chernoff and Scheffe.

Theorem 2.6 (Generalized Neyman-Pearson lemma)**.**
Let $q_0, \ldots, q_n$ be probability density functions with associated distributions $\mathbb{P}_0, \ldots, \mathbb{P}_n$. Let $t_1, \ldots, t_n > 0$.
Define the Neyman-Pearson set $W = \{z : q_0(z) \leq \sum_{i=1}^n t_i q_i(z)\}$.
Let W^* be such that, for all $i \in \{1, \ldots, n\}$, $\mathbb{P}_i(W) \leq \mathbb{P}_i(W^*)$.
Then, $\mathbb{P}_0(W) \leq \mathbb{P}_0(W^*)$.

Proof.

$$\mathbb{P}_0(W) - \mathbb{P}_0(W^*) = \mathbb{P}_0(W \setminus W^*) - \mathbb{P}_0(W^* \setminus W) = \int_{W \setminus W^*} q_0(z)dz - \int_{W^* \setminus W} q_0(z)dz$$
$$\leq \int_{W \setminus W^*} \sum_{i=1}^{n} t_i q_i(z)dz - \int_{W^* \setminus W} \sum_{i=1}^{n} t_i q_i(z)dz = \sum_{i=1}^{n} t_i(\mathbb{P}_i(W) - \mathbb{P}_i(W^*)) \leq 0.$$

□

This result can be used to derive *noise-based certificates*.

Theorem 2.7 (Certified radius with multiple distributions)**.**
Let $F : \mathbb{R}^d \to \{0, 1\}$ be a base classifier.
Let $q_1, \ldots, q_n$ be probability density functions.
Assume that for all $i \in \{1, \ldots n\}$, q_i is isotropic[4]*.*
Let $x \in \mathbb{R}^d$ and $\epsilon > 0$. Let $\delta \in \mathbb{R}^d$ such that $\|\delta\|_2 = \epsilon$.
Let $t_1(\delta), \ldots, t_n(\delta) > 0$ and define:

$$W(\delta) = \left\{ z \in \mathbb{R}^d : q_0(z - x - \delta) \leq \sum_{i=1}^{n} t_i(\delta) q_i(z - x) \right\}.$$

Assume that the $t_i(\delta)$ are chosen such that, for all $i \in \{1, \ldots, n\}$:

$$\int_{W(\delta)} q_i(z - x)dz \leq \int_{R^d} F(z) q_i(z - x)dz.$$

Let $\mathcal{H}$ be the set of measurable functions from $\mathbb{R}^d$ to $\{0, 1\}$. Define the family of functions:

$$\mathcal{F} = \left\{ G \in \mathcal{H} : \forall i \in \{1, \ldots, n\}, \int_{\mathbb{R}^d} G(z) q_i(z - x)dz = \int_{\mathbb{R}^d} F(z) q_i(z - x)dz \right\}.$$

Then:

$$\int_{W(\delta)} q_0(z - x - \delta)dz \leq \inf_{G \in \mathcal{F}} \inf_{\|\delta\|_2 \leq \epsilon} \int_{\mathbb{R}^d} G(z) q_0(z - x - \delta)dz,$$

with equality if $\forall i \in \{1, \ldots, n\}$, $\int_{W(\delta)} q_i(z - x)dz = \int_{R^d} F(z) q_i(z - x)dz$.

Before proving Theorem 2.7, let us emphasize some remarks:

- Theorem 2.7 can be used to derive certified radius against l_2 attacks by finding conditions on $\|\delta\|_2$ such that[5] $\int_{W(\delta)} q_0(z - x - \delta)dz > \frac{1}{2}$.
- The use of several *certification distributions* $q_1, \ldots, q_n$ besides the smoothing distribution q_0 allows to reduce the size of $\mathcal{F}$, thus to achieve higher certificate $\int_{W(\delta)} q_0(z - x - \delta)dz$ and large certified radii. Note that we do *not* need to know $\int_{W(\delta)} q_0(z - x)dz$ to obtain a certified radius.
- Theorem 2.7 requires the distributions q_i to be isotropic, which may lead to small certified radii.
- Theorem 2.7 requires that the smoothing distribution be of the form $q_0(\cdot - x)$, i.e., translation of a single distribution, which prevents the use of input-dependent randomized smoothing.
- Theorem 2.7 certifies only against l_2 attacks, while true robustness is measured by semantic attacks.

In order to obtain an explicit certified radius, we need:

[4]There is a function $\tilde{q}_i$ such that for all $z \in \mathbb{R}^d$, $q_i(z) = \tilde{q}_i(\|z\|)$.
[5]Here, we assume that $\Phi F(x) = 1$.

1. To find the $t_i(\delta)$ in order to define $W(\delta)$. We want the $\int_{W(\delta)} q_i(z-x)dz$ to be as high as possible without exceeding $\int_{R^d} F(z)q_i(z-x)dz$.
2. To compute $\int_{W(\delta)} q_0(z-x-\delta)dz$ explicitly and find conditions on $\|\delta\|_2$.

Both steps depend on the distributions $q_0, q_1, \ldots, q_n$.

Proof of Theorem 2.7. Let $G \in \mathcal{F}$. By assumption, for all i:

$$\int_{W(\delta)} q_i(z-x)dz \leq \int_{R^d} F(z)q_i(z-x)dz = \int_{R^d} G(z)q_i(z-x)dz.$$

Using Theorem 2.6:

$$\int_{W(\delta)} q_0(z-x-\delta)dz \leq \int_{R^d} G(z)q_0(z-x-\delta)dz.$$

Thus, since G was arbitrary:

$$\int_{W(\delta)} q_0(z-x-\delta)dz \leq \inf_{G\in\mathcal{F}} \int_{R^d} G(z)q_0(z-x-\delta)dz.$$

□

In fact, Theorem 2.7 is false. Roman provided a simple counter-example. Maybe, the Theorem was poorly formulated in [3]?

2.5 Geometric interpretation

Interestingly, the NPL is equivalent to the fact that the Kullback-Leibler divergence is non-negative. Indeed, we can rewrite $W = \left\{z \in \mathcal{X} : \ln \frac{p_A(z)}{p_H(z)} > u\right\}$ with $t = e^u$. Now, consider another test defined by the set $W^* = \left\{z \in \mathcal{X} : \ln \frac{p_Q(z)}{p_H(z)} > u\right\}$ where p_Q is another density. Then, the loss of power when using W^* instead of W is:

$$\mathbb{P}_A(W) - \mathbb{P}_A(W^*) \geq e^u \left(\mathbb{P}_H(W) - \mathbb{P}_H(W^*)\right).$$

The overall loss of power for different thresholds u is:

$$\int_{-\infty}^{+\infty} \left(\mathbb{P}_A(W) - \mathbb{P}_A(W^*)\right) du = \int_{-\infty}^{+\infty} \left(\mathbb{P}_A\left(\ln \frac{p_A(z)}{p_H(z)} > u\right) - \mathbb{P}_A\left(\ln \frac{p_Q(z)}{p_H(z)} > u\right)\right) du.$$

Somehow, using integration by parts (don't understand how), we get:

$$\int_{-\infty}^{+\infty} \left(\mathbb{P}_A(W) - \mathbb{P}_A(W^*)\right) du = D_{KL}(p_A||p_Q),$$

hence the non-negativity of D_{KL} implies that the test built with p_A has the highest power.

References

[1] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified Adversarial Robustness via Randomized Smoothing," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[2] Z. Hao, C. Ying, Y. Dong, H. Su, J. Song, and J. Zhu, "GSmooth: Certified Robustness against Semantic Transformations via Generalized Randomized Smoothing," in *Proceedings of the 39th International Conference on Machine Learning*, pp. 8465–8483, 2022.

[3] R. Ettedgui, A. Araujo, R. Pinot, Y. Chevaleyre, and J. Atif, "Towards Evading the Limits of Randomized Smoothing: A Theoretical Analysis," 2022.

[4] P. Súkeník, A. Kuvshinov, and S. Günnemann, "Intriguing Properties of Input-dependent Randomized Smoothing," 2022.

[5] M. Alfarra, A. Bibi, P. H. S. Torr, and B. Ghanem, "Data dependent randomized smoothing," in *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2022.

[6] J. Mohapatra, C.-Y. Ko, L. Weng, P.-Y. Chen, S. Liu, and L. Daniel, "Hidden Cost of Randomized Smoothing," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

[7] G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li, "Randomized Smoothing of All Shapes and Sizes," in *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[8] K. Dvijotham, J. Hayes, B. Balle, J. Z. Kolter, C. Qin, A. Gyorgy, K. Xiao, S. Gowal, and P. Kohli, "A Framework for Robustness Certification of Smoothed Classifiers using f-divergences," 2020.

[9] S. Eguchi and J. Copas, "Interpreting Kullback–Leibler divergence with the Neyman–Pearson lemma," *Journal of Multivariate Analysis*, vol. 97, pp. 2034–2040, Oct. 2006.