Certifying Ensembles: A General Certification Theory with $\mathcal{S}$-Lipschitzness

Aleksandar Petrov [1 2 *] **Francisco Eiras** [2] **Amartya Sanyal** [3] **Philip H.S. Torr** [2] **Adel Bibi** [2 *]

Abstract

Improving and guaranteeing the robustness of deep learning models has been a topic of intense research. Ensembling, which combines several classifiers to provide a better model, has been shown to be beneficial for generalisation, uncertainty estimation, calibration, and mitigating the effects of concept drift. However, the impact of ensembling on certified robustness is less well understood. In this work, we generalise Lipschitz continuity by introducing $\mathcal{S}$-Lipschitz classifiers, which we use to analyse the theoretical robustness of ensembles. Our results are precise conditions when ensembles of robust classifiers are more robust than any constituent classifier, as well as conditions when they are less robust.

1. Introduction

Ensembling consists in combining several classifiers to obtain a better-performing one (Hansen & Salamon, 1990; Sagi & Rokach, 2018). While it was originally proposed to improve the accuracy of weak classifiers (Rokach, 2016; Allen-Zhu & Li, 2023), it is also beneficial for improving uncertainty estimation and calibration (Lakshminarayanan et al., 2017; Zhang et al., 2020), as well as mitigating the effects of concept drift (Sagi & Rokach, 2018). These benefits of ensembling have inspired research into studying its effect on robustness. For example, recent empirical works have shown that encouraging diversity in the non-maximal predictions (Pang et al., 2019), or in the gradient directions (Kariyappa & Qureshi, 2019) of individual classifiers results in ensembles with improved robustness.

However, the degree of improved performance depends on the ensembled classifiers. When the constituent classifiers are all highly accurate, there is little room for improvement after ensembling; the gains are most pronounced with weak classifiers. Possibly, a similar limitation holds for robustness: perhaps ensembles of robust classifiers enjoy lower robustness improvements than ensembles of non-robust classifiers. Pang et al. (2019), Horváth et al. (2021), Yang et al. (2022) and Puigcerver et al. (2022) propose theoretical justifications for why ensembles boost robustness but stop short of quantifying the improvement, especially when the individual classifiers are already robust. This raises the following questions on the robustness limitations of ensembles:

i. *For a collection of robust classifiers, can their ensemble be more robust than its constituents? If so, what is the maximum achievable improvement, and under which conditions?*

ii. Conversely: *Is it possible for an ensemble of robust classifiers to be less robust than its constituents? If so, what is the worst possible drop in robustness, and under which conditions?*

We tackle these questions by introducing $\mathcal{S}$-Lipschitzness: a generalization of Lipschitz continuity that enables tight analysis of the theoretical robustness of ensembles. $\mathcal{S}$-Lipschitzness gives rise to certificates which need not be symmetric and are guaranteed to certify regions at least as large as the classical Lipschitz ones. Building on the $\mathcal{S}$-Lipschitzness framework, we answer the above questions:

i. It is possible for ensembles to certify every perturbation that any of the individual classifiers can certify, and even a *superset of their union*. However, the gain is most pronounced when the individual classifiers are not robust; as the robustness of the classifiers improves, the ensembling gain becomes more limited.

ii. It is possible for ensembles to fail to certify perturbations that every individual classifiers certifies, *e.g.* the ensemble certificate can be a proper *subset of the intersection* of the constituent certificates. In the worst case, ensembles of robust classifiers *do not certify any perturbation at all*. However, we show that as long as all classifiers have the same prediction, the ensemble certificate will never be a *subset of the intersection*.

[*] Equal contribution [1] Department of Computer Science, University of Oxford, Oxford, UK [2] Department of Engineering Science, University of Oxford, Oxford, UK [3] ETH AI Center, ETH Zürich, Zürich, Switzerland. Correspondence to: A. Petrov <aleks@robots.ox.ac.uk>, A. Bibi <adel.bibi@eng.ox.ac.uk>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Related work

Certified Adversarial Robustness. Deep neural networks are vulnerable to adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2015). The emergence of empirical defences to these mechanisms (Papernot et al., 2017; Madry et al., 2017; de Jorge et al., 2022), has motivated the need for methods that achieve *certified* robustness. Those methods can be classified into *exact*, *i.e.*, complete (Katz et al., 2017; Ehlers, 2017; Huang et al., 2017; Lomuscio & Maganti, 2017; Bunel et al., 2018), or *conservative*, *i.e.*, sound but incomplete (Gowal et al., 2018; Mirman et al., 2018; Wang et al., 2018; Ayers et al., 2020). Probabilistic methods, mostly based on *randomized smoothing* (Lecuyer et al., 2019; Cohen et al., 2019), have been shown to scale to large networks but have high inference time complexity.

Robustness of Ensembles. While ensembles have long been used to boost the accuracy of classifiers, interest in their robustness properties is more recent. Pang et al. (2019) propose diversifying the non-maximal predictions of individual classifiers which leads to empirically better robustness. Kariyappa & Qureshi (2019) recommend a different type of regularisation: encouraging misaligned gradients. Moreover, Horváth et al. (2021) and Yang et al. (2022) observe that applying randomized smoothing after ensembling results in more certifiably robust models than applying it to the individual classifiers. Xu et al. (2021) proposed using a mixture of clean and robust experts, while Puigcerver et al. (2022) studied the Lipschitz continuity of ensembles.

3. $\mathcal{S}$-Certificates with $\mathcal{S}$-Lipschitzness

We start by introducing the definition of point-wise adversarial robustness of a classifier[1].

Definition 1 (Robustness)**.** *Given a classifier* $f\colon \mathbb{R}^d \to \mathbb{R}^K$, *an* $x \in \mathbb{R}^d$ *and a set* $Q \subset \mathbb{R}^d$, f *is said to be* robust *at* x *if* $\arg\max_{i\in 1,\dots,K} f_i(x) = \arg\max_{i\in 1,\dots,K} f_i(x+\delta)$, $\forall \delta \in Q$, *where* f_i *is the prediction for the* i*-th class. We will call* Q *a* certificate *at* x.

3.1. Lipschitz Certificates

The Lipschitz constant of a function is closely related to its gradients. The larger the norm of the gradients, the more sensitive the function is to perturbations and the larger its Lipschitz constant becomes. Furthermore, given a Lipschitz classifier with a Lipschitz constant L, the *prediction gaps*, *i.e.*, the differences between the confidence of the top prediction and the other classes, fully determine the certificate Q. As such, we have the following proposition.

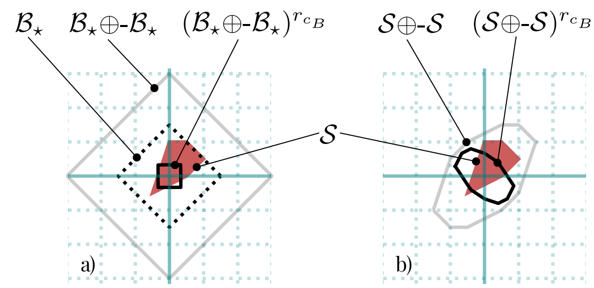

Figure 1. Lipschitz certificate for the ℓ_∞ norm (a) and $\mathcal{S}$-certificate (b). These are shown with . The $\mathcal{S}$-certificate is a superset of the Lipschitz certificate. Both certificates are in the uniform setting (U) for a classifier $f : \mathbb{R}^d \to \mathbb{R}^K$ with range space of gradients $\mathcal{S} = \{\nabla f_i(x) : x \in \mathbb{R}^d, i{=}1, \dots, K\}$ (shown in). We assume $r_{c_B} = 1$. $\mathcal{B}_\star$ is the smallest ℓ_1 ball containing $\mathcal{S}$.

Proposition 1 (Certification of Lipschitz classifiers)**.** *Take a differentiable*[2] *classifier* $f : \mathbb{R}^d \to \mathbb{R}^K$ *such that* $\sup_x \|\nabla f_i(x)\|_\star \le L_i$, $\forall i$. *Then* f_i *is* L_i*-Lipschitz with respect to* $\|\cdot\|$. *Moreover,* f *has a certificate*

$$Q = \left\{ \delta \in \mathbb{R}^d : \|\delta\| \le \min_{i \ne c_A} \frac{f_{c_A}(x) - f_i(x)}{L_i + L_{c_A}} = \min_{i \ne c_A} \frac{r_i}{L_i + L_{c_A}} \right\}. \tag{1}$$

Here, $\|\cdot\|_\star$ *is the dual norm to* $\|\cdot\|$ *and* c_A *is* $\arg\max_i f_i(x)$. *If all classes have the same Lipschitz constant* L, *i.e.,* $L_i \le L, \forall i$, *the certificate simplifies to*

$$Q = \left\{ \delta \in \mathbb{R}^d : \|\delta\| \le \frac{f_{c_A}(x) - f_{c_B}(x)}{2L} = \frac{r_{c_B}}{2L} \right\}, \tag{2}$$

where $c_B = \arg\max_{i \ne c_A} f_i(x)$. (Proof on p. 16)

We refer to the formulation in Equation (1) as *class-wise Lipschitz continuity* (CW) since it accounts for the classes potentially having different Lipschitz constants. Often, however, in prior art, all classes are considered to have the same Lipschitz constant L set such that $L \ge \max_i L_i$. We refer to this setting captured by Equation (2) as *uniform Lipschitz continuity* (U). Moreover, the Lipschitz certificates apply to any choice of norm; the main text considers only ℓ_p norms but we give further examples in Appendix C.1.

Figure 1a shows the relationship between the norm of the gradients of a classifier (its Lipschitzness) and the resulting certificates from Proposition 1. Take a classifier $f : \mathbb{R}^d \to \mathbb{R}^K$ and the set of all its gradients $\mathcal{S} = \{\nabla f_i(x) : x \in \mathbb{R}^d, i{=}1, \dots, K\}$ shown in . For simplicity, assume also that $r_{c_B}{=}1$. As $\sup_{s\in\mathcal{S}} \|s\|_1 \le 1.5$, the f_i are 1.5-Lipschitz with respect to the ℓ_∞ norm. Therefore, from Equation (2) the certificate Q is the ℓ_∞ ball of radius $1/3$ shown with . Taking the supremum of the ℓ_1 norm intro-

[1] A list of symbols is provided in Appendix A.

[2] For simplicity, we work with differentiable classifiers, even though our results are also valid for continuous classifiers that are not differentiable at finite number of points.

duces overapproximation of the true set of gradients. The region has the same supremum ℓ_1 norm as $\mathcal{S}$ and hence has the same certificate. However, is a superset of $\mathcal{S}$ and must correspond to a more sensitive classifier. This is due to the overapproximating action of the supremum of the gradient norms. To rectify this, we offer a generalization of Lipschitzness working directly with the gradients $\mathcal{S}$.

3.2. $\mathcal{S}$-Certificates

We observed that Lipschitzness induces a larger gradient overapproximation to the set of gradients set $\mathcal{S}$. This begs the question: *Can we enlarge the certificates by avoiding the dual norm ball overapproximation of the gradients and work directly with the exact gradient set $\mathcal{S}$?* To this end, we first generalize the definition of a Lipschitz function which allows the use of the exact range space of the gradient as opposed to any overapproximation.

Definition 2 ($\mathcal{S}$-Lipschitz function). *A function $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{S}$-Lipschitz for a bounded set $\mathcal{S} \subset \mathbb{R}^d$ if it holds that:*

$$-\rho_{\mathcal{S}}(x-y) \leq f(y) - f(x) \leq \rho_{\mathcal{S}}(y-x),\ \forall x, y \in \mathbb{R}^d,$$

with $\rho_{\mathcal{S}}(\delta) = \sup_{c\in\mathcal{S}} c^\top \delta$. If $\mathcal{S}$ is convex, then $\rho_{\mathcal{S}}$ corresponds to its support function.

In contrast to the classical Lipschitzness, $\mathcal{S}$-Lipschitzness accounts not only for the magnitude of the gradients but also for their direction. We also can generalize the notion of dual norms to sets that are not norm balls:

Definition 3 (Polar set). *For a set $\mathcal{S} \subset \mathbb{R}^d$, the polar set*[3] *to $\mathcal{S}$ of radius $r > 0$ is defined as:*

$$(\mathcal{S})^r = \left\{\delta \in \mathbb{R}^d\ :\ \rho_{\mathcal{S}}(\delta) = \sup\nolimits_{x\in\mathcal{S}} x^\top \delta \leq r\right\}.$$

Let's generalize Proposition 1 with $\mathcal{S}$-Lipschitzness:

Theorem 1 ($\mathcal{S}$-certificates). *Let $f : \mathbb{R}^d \to \mathbb{R}^K$ be a classifier with f_i being differentiable and $\nabla f_i : \mathbb{R}^d \to \mathcal{S}_i$ for all $i = 1, \ldots, K$. Then, each f_i is $\mathcal{S}_i$-Lipschitz. Furthermore, for a fixed x, f is robust at x against all δ in*

$$Q = \bigcap\nolimits_{i\neq c_A} (\mathcal{S}_i \oplus -\mathcal{S}_{c_A})^{r_i}. \tag{3}$$

Here, $c_A = \arg\max_c f_c(x)$, $r_i = f_{c_A}(x) - f_i(x)$, and $\oplus$ is the Minkowski sum. If $\mathcal{S} \supseteq \mathcal{S}_i, \forall i$, then we have the simplified certificate

$$Q = (\mathcal{S} \oplus -\mathcal{S})^{r_{c_B}}, \tag{4}$$

where $c_B = \arg\max_{c\neq c_A} f_c(x)$. (Proof on p. 17)

[3] We are extending the standard notion of a polar set (Rockafellar, 1970) to encompass radii different from 1.

We show Theorem 1 is tight in an example in Proposition 7.

The certificate in Theorem 1 is a polar set (or intersection of polar sets), hence, it has a natural dependence on the gradient sets $\mathcal{S}$ and the prediction gap r:

Proposition 2 (Polar set dependence on $\mathcal{S}$ and r). *Let $\mathcal{S}, \mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4 \subset \mathbb{R}^d$ be bounded and $r, r_1, r_2 > 0$:*

i. $\mathcal{S}_1 \subseteq \mathcal{S}_2 \Rightarrow (\mathcal{S}_1 \oplus -\mathcal{S}_1) \subseteq (\mathcal{S}_2 \oplus -\mathcal{S}_2)$;
ii. $\mathcal{S}_1 \subseteq \mathcal{S}_2 \Rightarrow (\mathcal{S}_1)^r \supseteq (\mathcal{S}_2)^r$;
iii. $r_1 \leq r_2 \Rightarrow (\mathcal{S})^{r_1} \subseteq (\mathcal{S})^{r_2}$;
iv. $((\mathcal{S}_1 \subseteq \mathcal{S}_3) \wedge (\mathcal{S}_2 \subseteq \mathcal{S}_4)) \Rightarrow (\mathcal{S}_3 \oplus -\mathcal{S}_4)^r \subseteq (\mathcal{S}_1 \oplus -\mathcal{S}_2)^r$.

where $\oplus$ is the Minkowski sum operator. (Proof on p. 18)

3.3. $\mathcal{S}$-Certificates Subsume Lipschitz Certificates

In Section 3.1 we showed that the classifier in Figure 1 is 1.5-Lipschitz with respect to ℓ_∞ norm and that its Lipschitz certificate is therefore the ℓ_∞ ball of radius $1/3$. The same result can be viewed as a special case of $\mathcal{S}$-certification when we observe that the classifier is $\mathcal{B}_\star$-Lipschitz with $\mathcal{B}_\star = \{x \in \mathbb{R}^d : \|x\|_1 \leq 1.5\}$. Hence, for $r_{c_B} = 1$, from Equation (4) we get the same certificate $(\mathcal{B}_\star \oplus -\mathcal{B}_\star)^1 = (2\mathcal{B}_\star)^1 = \{\delta \in \mathbb{R}^d : \|\delta\|_\infty \leq 1/3\}$ (in Figure 1a). However, if we do not overapproximate $\mathcal{S}$ with $\mathcal{B}_\star$, then Equation (4) gives us the $\mathcal{S}$-certificate $(\mathcal{S} \oplus -\mathcal{S})^1$ (in Figure 1b). Clearly, the $\mathcal{S}$-certificate is larger than the Lipschitz one. Proposition 8 in the appendix shows that this is always the case.

Could it be that the $\mathcal{S}$-certificate in Figure 1 is larger than the Lipschitz certificate because of a suboptimal choice of norm? No, because whenever the set of gradients is not centrally symmetric, *i.e.*, $\mathcal{S} \neq -\mathcal{S}$, then no matter what norm we choose, we have $\mathcal{B}_\star \supset \mathcal{S}$ and thus an $\mathcal{S}$-certificate larger than the Lipschitz certificate. This is because norms are centrally symmetric by definition.

Are CW certificates always supersets to the U certificates? The CW and U $\mathcal{S}$-certificates are larger than any Lipschitz certificate (Proposition 8). As CW generalizes U, its certificates are supersets to the ones of U. This follows from CW reducing to U by taking $\mathcal{S} \supseteq \cup\, \mathcal{S}_i$, *i.e.*, overapproximating some of the classes with a larger $\mathcal{S}$. This is analogous to setting $L \geq \max L_i$ in the Lipschitz case. Then, from Proposition 2*iv*, it directly follows that CW certificates are always supersets of U certificates. Another view is that U certificates are restricted to only symmetric sets since $\mathcal{S} \oplus - \mathcal{S}$ is symmetric (Aux. Lemma 7), while CW certificates, *i.e.*, $\bigcap_{i\neq c_A} (\mathcal{S}_i \oplus - \mathcal{S}_{c_A})^{r_i}$, can be asymmetric.

The example in Figure 2 (with detailed calculations in Appendix C.3) shows how the certified regions can vary depending on whether we use $\mathcal{S}$-Lipschitz or Lipschitz certificates and on the CW or U modes.

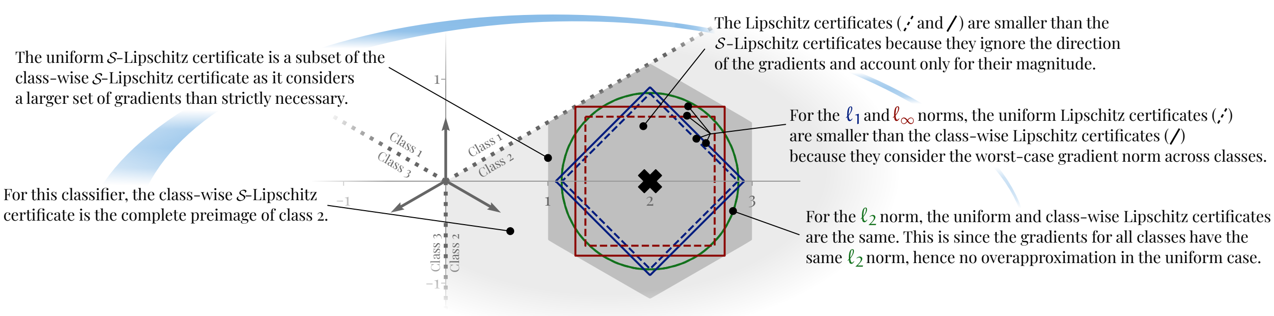

Figure 2. Lipschitz and $\mathcal{S}$-Lipschitz certificates at $x = [2, 0]^\top$ for a linear classifier that splits the domain into three equal sectors. Step-by-step explanation of the construction of the certificates is provided in Appendix C.3.

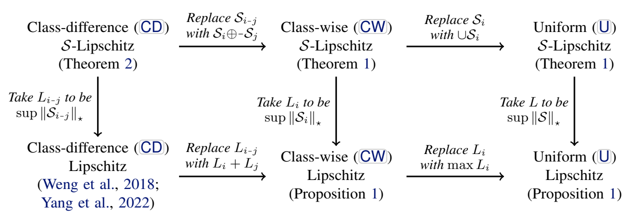

Figure 3. The lattice of continuity certificates. $A \to B$ means that the certificate provided by B is a subset of the certificate of A. Therefore, class-difference $\mathcal{S}$-certificates are the largest, while uniform Lipschitz certificates are the smallest.

3.4. Tightening Certificates via Class Differences

We conclude this section by showing how to further enlarge the certificates by directly targeting the $\mathcal{S}$-Lipschitzness of the class difference. Recall the $\mathcal{S}$-certificate $Q = \bigcap_{i \neq c_A} (\mathcal{S}_i \oplus -\mathcal{S}_{c_A})^{r_i}$ for the CW mode from Theorem 1. The role of the $\mathcal{S}_i \oplus -\mathcal{S}_{c_A}$ term is to measure the $\mathcal{S}$-Lipschitz continuity of $h_{i\text{-}c_A} = f_i - f_{c_A}$. It is straightforward to see that $h_{i\text{-}c_A}$ is indeed $(\mathcal{S}_i \oplus -\mathcal{S}_{c_A})$-Lipschitz. However, it is not necessarily the tightest $\mathcal{S}$ for $h_{i\text{-}c_A}$. Intuitively, $\mathcal{S}_i \oplus -\mathcal{S}_{c_A}$ takes the differences of the gradients of f_i and f_{c_A}, *regardless* of the input x. However, the set of gradients of $h_{i\text{-}c_A}$ are the difference of gradients of f_i and f_{c_A} *at the same* x. If all classes are similarly sensitive at a given x but their sensitivity varies *jointly* across the domain, the difference between $\mathcal{S}_i \oplus -\mathcal{S}_{c_A}$ and the gradients of $h_{i\text{-}c_A}$ can be significant. Using this, we can tighten Theorem 1 with class-difference (CD) certificates.

Theorem 2. *Let $f : \mathbb{R}^d \to \mathbb{R}^K$ be a classifier such that $h_{i\text{-}j} = f_i - f_j$ is $\mathcal{S}_{i\text{-}j}$-Lipschitz, $\forall i, j \in 1, \dots, K, i \neq j$. Then, given an input $x \in \mathbb{R}^d$, f is robust at x against all δ in $Q = \bigcap_{i \neq c_A} (\mathcal{S}_{i\text{-}c_A})^{r_i}$.* (Proof on p. 19)

This approach generalizes the CW $\mathcal{S}$-certificates from Theorem 1 and provides the tightest certificates. Hence, throughout the rest of the paper, we will use class difference unless stated otherwise. Prior work looked at the Lipschitz CD certificates (Weng et al., 2018) and regularization (Yang et al., 2022). Still, we believe to be the first to offer a theoretical justification of why it enlarges the certificates through the new lens of $\mathcal{S}$-Lipschitzness. Figure 3 summarizes the big picture relating the certificates with function continuity and positions our new results with respect to prior art.

4. Robustness of Ensembles of Classifiers

We can use $\mathcal{S}$-Lipschitzness to study how the robustness properties of individual classifiers affect the robustness of an ensemble of them. Given N classifiers $f^j : \mathbb{R}^d \to \mathbb{R}^K$, consider their weighted ensemble:

$$g(x) = \textstyle\sum_{j=1}^{N} \alpha_j f^j(x), \;\; \alpha_j \geq 0, \;\; \sum_{j=1}^{N} \alpha_j = 1. \tag{5}$$

We will indicate the prediction gaps of f^j as r^j. We can use the $\mathcal{S}$-certificates from Theorem 2 in order to relate the ensemble robustness to that of the individual classifiers.

Theorem 3 (Addition of $\mathcal{S}$-Lipschitz classifiers)**.** *Take an ensemble as in Equation (5) with $N = 2$ and the CD setting, i.e., $h^j_{i\text{-}k} = f^j_i - f^j_k$ is $\mathcal{S}^j_{i\text{-}k}$-Lipschitz. Then, at a fixed $x \in \mathbb{R}^d$, it holds that g is robust against all δ in*

$$Q_g = \bigcap_{i \neq c^g_A} \left(\alpha_1 \mathcal{S}^1_{i\text{-}c^g_A} \oplus \alpha_2 \mathcal{S}^2_{i\text{-}c^g_A} \right)^{r^g_i},$$

with $c^g_A = \arg\max_i g_i$ and $r^g_i = g_{c^g_A} - g_i$. The case for $N > 2$ follows by induction. (Proof on p. 19)

We study whether ensembling two classifiers f_1 and f_2 results in better robustness by comparing the ensemble certificate Q_g with the individual certificates Q_1 and Q_2. We identify three regimes:

$Q_g \supset Q_1 \cup Q_2$	uniform improvement,	❶
$Q_1 \cap Q_2 \subseteq Q_g \subseteq Q_1 \cup Q_2$	inconclusive,	❷
$Q_1 \cap Q_2 \supset Q_g$	uniform reduction.	❸

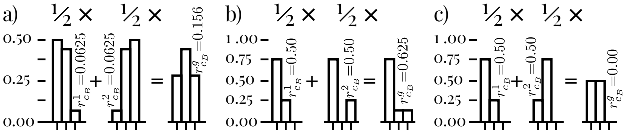

Figure 4. Ensembles of two classifiers (N=2, K=3) in regime ❶ (**a** and **b**), and regime ❸ with $r^g = 0$ and hence $Q_g = \{0\}$ (**c**).

Ideally, we wish to construct ensembles that are in regime ❶. We may tolerate ensembles in ❷. But most importantly, we want to avoid ensembles in regime ❸ at all costs.

The certification regime depends on whether we are in the $\boxed{\mathsf{U}}$ or $\boxed{\mathsf{CD}}$ mode. It also depends on the ensemble agreement on the top predictions, *i.e.*, which of the following holds:

$$c_A = c_A^j = \arg\max_i f_i^j(x), \quad \text{for all } j \in 1, \dots, N \quad \boxed{c_A^=}$$

$$c_A^j \neq c_A^{j'}, \quad \text{for } j \neq j' \quad \boxed{c_A^{\neq}}$$

$$c_B = c_B^j = \arg\max_{i \neq c_A^j} f_i^j(x), \quad \text{for all } j \in 1, \dots, N \quad \boxed{c_B^=}$$

The rest of this section outlines the conditions leading to each one of the ❶,❷ and ❸ certification regimes.

Let us first examine a common scenario: the constituent classifiers agreeing on the top two predictions ($\boxed{c_A^=}$ and $\boxed{c_B^=}$). Under the common $\boxed{\mathsf{U}}$ mode where all classes are similarly Lipschitz, one might guess that ensembling such agreeing classifiers must boost robustness. However, the above conditions put the ensemble solidly in regime ❷:

Theorem 4. *Consider an ensemble of* $\boxed{\mathsf{U}}$ *classifiers and a fixed x for which* $\boxed{c_A^=}$ *and* $\boxed{c_B^=}$ *hold. Then, for any choice of weights α_j in Equation (5), the $\mathcal{S}$-certificate of the ensemble satisfies* ❷. (Proof on p. 19)

Theorem 4 is particularly concerning when $\mathcal{S}^1$ and $\mathcal{S}^2$ are balls with the same norm but different radii, as we show with an example in Appendix C.4. This section shows how relaxing the conditions in Theorem 4 enables all regimes.

Regime ❶ is possible. If all classifiers are $\boxed{\mathsf{U}}$ and $\mathcal{S}^j = \mathcal{S}$, then the differences between Q_1, Q_2 and Q_g are fully determined by r^1, r^2 and r^g. We will refer to this setting as $\boxed{\mathsf{U}'}$. This restriction is not uncommon: if randomized smoothing is used, then $\mathcal{S}$ is uniquely defined by the smoothing distribution (Yang et al., 2020; Eiras et al., 2022; Rumezhak et al., 2023), which is the same for all constituents. For a $\boxed{\mathsf{U}'}$ ensemble, prediction gaps satisfying $r^g_{c_B} > \overline{r} = \max_j r^j_{c_B}$ imply ❶. We show two examples of such predictions gaps in Figure 4a and b.

The margin of improvement is limited. Although the feasibility of regime ❶ is noteworthy, the improvement of of the ensemble over the most robust constituent classifier is limited. To simplify the analysis, we assume $\boxed{c_A^=}$ holds. This is reasonable as $\boxed{c_A^=}$ prevents ❸ (as we will show in Proposition 5). We will also assume that all $\mathcal{S}$ are of the same shape, *e.g.*, norms, though not necessarily of the same size[4]. This allows us to work with scalar radii instead of sets.

Proposition 3. *Take two classifiers $f^1, f^2 : \mathbb{R}^d \to \mathbb{R}^K$ satisfying* $\boxed{c_A^=}$*. Further, assume that all $h^j_{i\text{-}k} = f^j_i - f^j_k$ are $\epsilon_{j,i\text{-}k}\mathcal{B}_\star$-Lipschitz for some closed convex symmetric set $\mathcal{B}_\star$. Then, the maximum improvement in the certified radius R^g of g relative to the larger one of R^1 and R^2 is*

$$R^g - \max\{R^1, R^2\} \leq \frac{1}{\min\{M^1, M^2\}} - \frac{\min\{r^1_{c^1_B}, r^2_{c^2_B}\}}{\min\{M^1, M^2\} + \Delta},$$

where we have defined M^k as $\min_{i \neq c_A} \epsilon_{k,i\text{-}c_A}$ and Δ as $\max_{k=1,2} \max_{i \neq c_A} (\epsilon_{k,i\text{-}c_A} - M^k)$. (Proof on p. 20)

In the above proposition, $\min\{M^1, M^2\}$ refers to the radius of the least sensitive classifier, *i.e.*, the one with smallest Lipschitz constant or $\mathcal{S}$-Lipschitzness. Δ measures how the Lipschitzness ranges amongst the classes and classifiers. $\Delta = 0$ implies that all $\epsilon_{k,i-c_A}$ are the same and therefore, all classifiers have the same Lipschitzness for all class pairs. On the other hand, large Δ means that some classifiers are more robust for some class pairs while others are very sensitive for particular class pairs.

Proposition 3 is more restrictive when the individual classifiers have large predictions gaps ($r^1_{c^1_B}, r^2_{c^2_B}$) and/or similar Lipschitzness (small Δ). Both factors likely hold for robust classifiers: the large prediction gap is necessary for a large certificate and the similar Lipschitzness ensures similarly sized certificates for the different classes. Therefore, the ensembling improvement is only significant when the individual classifiers are not robust.

Regime ❸ is also possible. Similarly to the ❶ case, for $\boxed{\mathsf{U}'}$ ensembles prediction gaps satisfying $r^g_{c_B} < \underline{r} = \min_j r^j_{c_B}$ imply ❸. An example of such predictions gaps is in Figure 4c. This regime unfortunately applies to the real-world classifiers in Appendix B: for all of them the constituent models are on average more robust than the ensemble.

Ensembles can result in zero robustness. To make matters worse, not only is it possible that $r^g_{c_B}$ is smaller than all individual gaps, but it can even be 0, resulting in $Q_g = \{0\}$. This is the very case illustrated in Figure 4c. Figures B.3 and B.4 show examples of this scenario occurring in practice.

Proposition 4. *For any set of $N \geq 2$ classifiers satisfying* $\boxed{c_A^{\neq}}$*, there exist weights α_j for which the resulting ensemble has $r^g_{c_B} = 0$ and a certified perturbation set $Q_g = \{0\}$.* (Proof on p. 22)

Same top predictions prevent regime ❸. The possibility of

[4]This is more general than the $\boxed{\mathsf{U}'}$ condition which restricted the sizes to also be the same.

❸ and the complete loss of robustness is disappointing but, as long as all classifiers have the same top prediction, the ensemble cannot have a decision boundary passing through x. Not only that, but also it will never be in regime ❸:

Proposition 5. *No ensemble of classifiers as in Theorem 3 satisfying* $\boxed{c_A^=}$ *can be in regime* ❸. (Proof on p. 22)

The sufficient conditions for improved certification are restrictive. Focusing again on the setting of Proposition 3, we can provide sufficient conditions for regime ❶:

Proposition 6. *Take an ensemble as in Proposition 3. Assume two different second top predictions and that classes that are not in the top two predictions of any individual classifier have low confidences*[5]*. Then* ❶ *occurs when:*

$$f^1_{c_A} > f^1_{c_B^2} + r^2_{c_B^2} \frac{\epsilon_{1,c_B^2-c_A}}{\epsilon_{2,c_B^2-c_A}} \quad \textit{and} \quad f^2_{c_A} > f^2_{c_B^1} + r^1_{c_B^1} \frac{\epsilon_{2,c_B^1-c_A}}{\epsilon_{1,c_B^1-c_A}}.$$

(Proof on p. 23)

The conditions in Proposition 6 are rather limiting: the second class predicted by f^2 should have low enough confidence by f^1 and vice versa. This means that ensembling ends up being beneficial at a fixed x if each classifier has a different second prediction and all other predictions are very close to 0. Therefore, regime ❶ is unlikely to occur unless the classifiers are carefully regularized. Pang et al. (2019) suggest encouraging diversity among the non-maximal predictions. Proposition 6 theoretically justifies this approach.

5. Discussion

In this section, we provide some comments on the implications and limitations of our theoretical analysis.

The conditions preventing regime ❸ also prevent accuracy gain for the ensemble. Proposition 5 showed that $\boxed{c_A^=}$ prevents regime ❸. However, ensembling cannot boost accuracy when in the $\boxed{c_A^=}$ regime. Hence robustness seems to be at odds with accuracy, in line with the robustness-accuracy trade-off (Zhang et al., 2019; Tsipras et al., 2019).

Ensembling can generate directionally-balanced certificates. When we have different shapes for $\mathcal{S}^1$ and $\mathcal{S}^2$, an ensemble can be used to trade-off classifiers that specialize in robustness in particular directions. As shown in Figure 5, this technique can be used to construct more directionally-balanced certificates. Therefore, depending on the notion of robustness, ❷ can be desirable when proper care is taken.

The prediction gap and $\mathcal{S}$-Lipschitzness are not independent. Throughout this paper, we treated the $\mathcal{S}$-Lipschitzness and the prediction gaps as two independent tools for boosting robustness. Intuitively, one would like to have as much as possible from both: smooth classifiers with high prediction gaps. However, this is not possible. The smoother a classifier is, the lower its prediction gaps are likely to be. Therefore, the robustness gains from ensembling are likely even smaller than the already conservative bounds we have. Appendix B offers experiments demonstrating this effect.

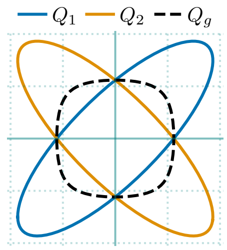

Figure 5. Highly directional certificates can be ensembled to obtain directionally balanced certificates. Q_g is constructed for $\alpha_j = 1/2$ and $r^1_{c_B} = r^2_{c_B} = 1$.

Robustness over distributions rather than single points. Section 4 focused on point-wise robustness: all the results presented there are for a fixed x. In reality, we are usually interested in the expected robustness over a distribution of inputs. Even if the ensemble performs worse than the best individual classifier (*e.g.*, ❷) at all x, it might still be overall more robust than any individual classifier. Furthermore, the unfavourable conditions in Proposition 4 might exist for some x, but it is likely that they are rare for real classifiers and distributions. We provide experimental observations to this effect in Appendix B. The highlight is that for all ensembles considered, the ensemble certificates are smaller than these of the individual classifiers for more than 50% of the inputs. Hence, real world ensembles seem to worsen robustness across distributions of inputs.

Limitations of the $\mathcal{S}$-Lipschitzness analysis. Most of the results in this paper are valid within the context of $\mathcal{S}$-certificates: inferring certificates for ensembles from the $\mathcal{S}$-Lipschitzness properties of the individual classifiers. While this framework was necessary for the theoretical analysis, it might be conservative. Methods that construct certificates without direct reliance on ($\mathcal{S}$)-Lipschitzness properties, *e.g.*, abstract interpretation (Gehr et al., 2018) or SMT solvers (Huang et al., 2017), might be able to provide larger certificates than what our theory predicts. However, these methods cannot provide general theoretical analysis of the type we offer in this work.

6. Conclusion

We propose $\mathcal{S}$-Lipschitzness which offers tighter robustness certificates. We use it to analyse the robustness properties of ensembles of classifiers. Our results show that ensembling can improve the certification over the most robust individual classifier only under very strict conditions. Moreover, even when improvements are possible, they are theoretically very small. We prove that an ensemble can be worse than the least robust constituent classifier. Even worse, it may result in a classifier with zero robustness.

[5] "Low confidences" is formally defined in the proof.

Acknowledgements

AP has received funding from Toyota Motor Europe. FE and PT have received funding from FiveAI. AS acknowledges partial support from the ETH AI Center fellowship. AB has received funding from the Amazon Research Awards. This work is supported by a UKRI grant Turing AI Fellowship (EP/W002981/1) and the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (EP/S024050/1). AB would like to thank Konstantin Mishchenko for the early insightful discussions. We also thank the Royal Academy of Engineering and FiveAI.

References

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *International Conference on Learning Representations*, 2023.

Ayers, E. W., Eiras, F., Hawasly, M., and Whiteside, I. PaRoT: A practical framework for robust deep neural network training. In *NASA Formal Methods Symposium*, 2020.

Bunel, R. R., Turkaslan, I., Torr, P., Kohli, P., and Mudigonda, P. K. A unified view of piecewise linear neural network verification. In *Advances in Neural Information Processing Systems*, 2018.

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

de Jorge, P., Bibi, A., Volpi, R., Sanyal, A., Torr, P. H., Rogez, G., and Dokania, P. K. Make some noise: Reliable and efficient single-step adversarial training. In *AdvML Frontiers Workshop at the International Conference on Machine Learning*, 2022.

Ehlers, R. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, 2017.

Eiras, F., Alfarra, M., Torr, P., Kumar, M. P., Dokania, P. K., Ghanem, B., and Bibi, A. ANCER: Anisotropic certification via sample-wise volume maximization. *Transactions on Machine Learning Research*, 2022.

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., and Vechev, M. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *IEEE Symposium on Security and Privacy*, 2018.

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

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Hansen, L. K. and Salamon, P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Horváth, M. Z., Mueller, M. N., Fischer, M., and Vechev, M. Boosting randomized smoothing with variance reduced classifiers. In *International Conference on Learning Representations*, 2021.

Huang, X., Kwiatkowska, M., Wang, S., and Wu, M. Safety verification of deep neural networks. In *International Conference on Computer Aided Verification*, 2017.

Kariyappa, S. and Qureshi, M. K. Improving adversarial robustness of ensembles with diversity training. *Preprint arXiv:1901.09981*, 2019.

Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. Reluplex: An efficient SMT solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, 2017.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

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

Lomuscio, A. and Maganti, L. An approach to reachability analysis for feed-forward ReLU neural networks. *arXiv preprint arXiv:1706.07351*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mirman, M., Gehr, T., and Vechev, M. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, 2018.

Pang, T., Xu, K., Du, C., Chen, N., and Zhu, J. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, 2019.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017.

Puigcerver, J., Jenatton, R., Ruiz, C. R., Awasthi, P., and Bhojanapalli, S. On the adversarial robustness of mixture of experts. In *Advances in Neural Information Processing Systems*, 2022.

Rockafellar, R. T. *Convex analysis*. Princeton University Press, 1970.

Rokach, L. Decision forest: Twenty years of research. *Information Fusion*, 2016.

Rumezhak, T., Eiras, F. G., Torr, P. H., and Bibi, A. RANCER: Non-axis aligned anisotropic certification with randomized smoothing. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

Sagi, O. and Rokach, L. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 2018.

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.

Schechter, E. *Handbook of Analysis and Its Foundations*. Academic Press, 1997.

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

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Wang, S., Chen, Y., Abdou, A., and Jana, S. Mixtrain: Scalable training of verifiably robust neural networks. *arXiv preprint arXiv:1811.02625*, 2018.

Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J., and Daniel, L. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018.

Xu, K., Wang, C., Cheng, H., Kailkhura, B., Lin, X., and Goldhahn, R. Mixture of robust experts (MoRE): A robust denoising method towards multiple perturbations. *Preprint arXiv:2104.10586*, 2021.

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

Yang, Z., Li, L., Xu, X., Kailkhura, B., Xie, T., and Li, B. On the certified robustness for ensemble models and beyond. In *International Conference on Learning Representations*, 2022.

Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L. E., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Zhang, J., Kailkhura, B., and Han, T. Y.-J. Mix-n-Match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, 2020.

A. List of symbols

For the ease of the reader, we have summarized the notation used in the paper in the following table:

α_j	The weight of the j-th classifier in the ensemble
$\mathcal{B}$	A norm ball
$\mathcal{B}_\star$	A dual norm ball
c_A^j	The class predicted by the j-th classifier with the highest confidence
c_B^j	The class predicted by the j-th classifier with the second highest confidence
c_A^g	The class predicted by the ensemble with the highest confidence
c_B^g	The class predicted by the ensemble with the second highest confidence
$\boxed{c_A^=}$	All top predictions in the ensemble are the same
$\boxed{c_A^{\neq}}$	At least two classifiers in the ensemble differ in their top prediction
$\boxed{c_B^=}$	All second highest predictions in the ensemble are the same
f	A classifier
f_i	The confidence for the i-th class of the classifier f
f^j	The j-th classifier in the ensemble of classifiers
g	An ensemble of classifiers $f_1, \ldots, f_N$
$h_{i\text{-}k}$	The difference of the confidence of classes i and k
i	Class index
j	Classifier index in an ensemble
K	Number of classes
L_i	The Lipschitz constant for the i-th class
N	Number of classifiers in the ensemble
Q	Certificate
Q_j	Certificate for the j-th classifier in the ensemble
Q_g	Certificate for the ensemble
r_i^j	The confidence gap between the top class and the i-th class for the j-th classifier in the ensemble
r_i^g	The confidence gap between the top class and the i-th class for the ensemble
$\overline{r}$	The maximum confidence gap in the ensemble ($\max_j r_{c_B}^j$)
$\underline{r}$	The minimum confidence gap in the ensemble ($\min_j r_{c_B}^j$)
R^j	Certified radius for the j-th classifier in the ensemble when Q_j is a norm ball
R^g	Certified radius for the ensemble when Q_g is a norm ball
$\rho_{\mathcal{S}}$	Support function
$\mathcal{S}$	Range space of gradients
$\mathcal{S}_i$	Range space of gradients for the i-th class
$\mathcal{S}^j$	Range space of gradients for the j-th classifier in the ensemble
$\mathcal{S}_{i\text{-}k}$	Range space of gradients for the difference of the confidence of classes i and k ($h_{i\text{-}k}$)
$(\mathcal{S})^r$	Polar set of $\mathcal{S}$ with radius r
σ	Smoothing Gaussian noise for randomised smoothing
$\boxed{\mathsf{U}}$	Uniform continuity regime
$\boxed{\mathsf{U}'}$	Uniform continuity regime with all classifiers having the same $\mathcal{S}$-Lipschitz for all classes
$\boxed{\mathsf{CW}}$	Class-wise continuity regime
$\boxed{\mathsf{CD}}$	Class-difference continuity regime

B. Experiments

In this appendix we describe several experiments that validate and illustrate the observations in the main body of the paper.

Experimental setup We use the ensembles trained by Horváth et al. (2021) that they have released publicly[6]. The classifiers are based on the ResNet20 and ResNet50 architectures (He et al., 2016) and are trained respectively on CIFAR10 (Krizhevsky, 2009) and ImageNet (Russakovsky et al., 2015). We use randomized smoothing (Lecuyer et al., 2019; Cohen et al., 2019) to obtain individual classifiers with known continuity properties ($\mathcal{S}$). Concretely, a model smoothed with independent Gaussian noise with variance σ^2 is $\sqrt{2/\pi\sigma^2}$-Lipschitz for the ℓ_2 norm (Salman et al., 2019). As standard with randomized smoothing, each classifier is trained with Gaussian noise with variance matching the smoothing variance (Lecuyer et al., 2019).

We consider the following ensembles:

i. Ensemble of N=6 ResNet20 classifiers trained on CIFAR10 (K=10), trained and smoothed with σ=0.25.
ii. Ensemble of N=6 ResNet20 classifiers trained on CIFAR10 (K=10), trained and smoothed with σ=0.50.
iii. Ensemble of N=6 ResNet20 classifiers trained on CIFAR10 (K=10), trained and smoothed with σ=1.00.
iv. Ensemble of N=3 ResNet50 classifiers trained on ImageNet (K=1000), trained and smoothed with σ=1.00.

We construct each ensemble with uniform weights $\alpha_j = 1/N$. As all classifiers comprising an ensemble have the same $\mathcal{S}$ and are in the uniform continuity regime (U), they are also in the U′ regime. Hence, as discussed in Section 4, we can directly infer the robustness certificates from the prediction gaps alone.

Note that for the experiments in this appendix, we *first smoothen the individual classifiers and then ensemble them*. This is as to make sure that the individual classifiers are smooth. This is opposite to the procedure suggested by Horváth et al. (2021) and Yang et al. (2022). They *ensemble first and smoothen the ensemble second*.

Regime ❶ is possible but occurs rarely in practice. From the 1000 CIFAR10 inputs at which we evaluated the three ResNet20 ensembles not a single one had an ensemble gap $r^g_{c_B}$ larger than the best individual classifier gap $\overline{r}$. This is shown in the left-most column in Figure B.1 that shows $r^g_{c_B}$ against $\overline{r}$: there is no points over the diagonal. The ResNet50 ensemble, though, has 7 samples out of 500 in regime ❶, *i.e.*, for which the ensemble has a larger certified radius than the best individual classifier (left plot in Figure B.2). However, this amounts to only 1.4% of the inputs being in regime ❶.

Regime ❸ occurs in practice but is also rare. Regime ❸, in which the ensemble fails to certify perturbations that every one of the individual classifiers certifies, does occur in practice as well. This is evident from the points under the line in the middle plots in Figures B.1 and B.2 which show $r^g_{c_B}$ against $\underline{r}$. For all four ensembles, there are inputs in regime ❸. For the ResNet20 ensembles evaluated on CIFAR10, this occurs in respectively 3.9%, 4.1% and 3.3% of the cases. The ResNet50 ensemble has 10.2% of its ImageNet samples in regime ❸. Still, all four ensembles have much larger rates of regime ❸ compared to regime ❶. Therefore, this indicates that for real-world ensembles, most inputs are likely in regime ❷, with some in regime ❸, and very few, if any, in regime ❶.

Overall, the ensembles have smaller certificates than the individual classifiers. Most inputs of real-world ensembles seem to be in regime ❷. This means that the ensemble prediction gap for an input x (and hence certified radius) is between the smallest and the largest individual classifier gaps at x. However, this does not tell us much about how the ensemble compares with a *single* individual classifier, which is what one needs in order to decide whether it is better to use the ensemble or a single model.

We can make this comparison with the help of the leftmost and rightmost plots in Figures B.1 and B.2 which show $r^g_{c_B}$ against respectively the best individual classifier gap $\overline{r}$ and the gap of one of the classifiers in the ensemble $r^1_{c_B}$. The plots also show the average ensemble gap $r^g_{c_B}$ and average individual gap $r^1_{c_B}$ across all samples. We can see that for all four ensembles, the average ensemble gap is smaller than the average gap of the individual classifier. Therefore, as far as the average certified radius is concerned, the ensembles have *lower* robustness than the individual classifier. Furthermore, only between 35% and 48% of the inputs have an ensemble gap that is larger than the individual gap. Hence, it appears that if one cares about robustness, they would be better off selecting one of the individual classifiers rather than the ensemble, for all four of these examples.

Ensembles of robust predictions can be non-robust in practice. Proposition 4 showed that it is possible that ensembles which, at a given x, all have $r^j_{c_B} > 0$, when ensembled can have $r^g_{c_B} = 0$ and hence a certificate $Q_g = \{0\}$, regardless of the continuity properties of the classifiers. One would hope that this is a purely theoretical curiosity and such situations do not occur in practice. However, as all of the centre plots in Figures B.1 and B.2 show, for every ensemble, there are inputs for which the worst individual classifier has gap well above 0, while the ensemble gap is practically 0. These are the points close to the horizontal axis. We discuss two examples in more details.

[6]Trained models are available at https://github.com/eth-sri/smoothing-ensembles

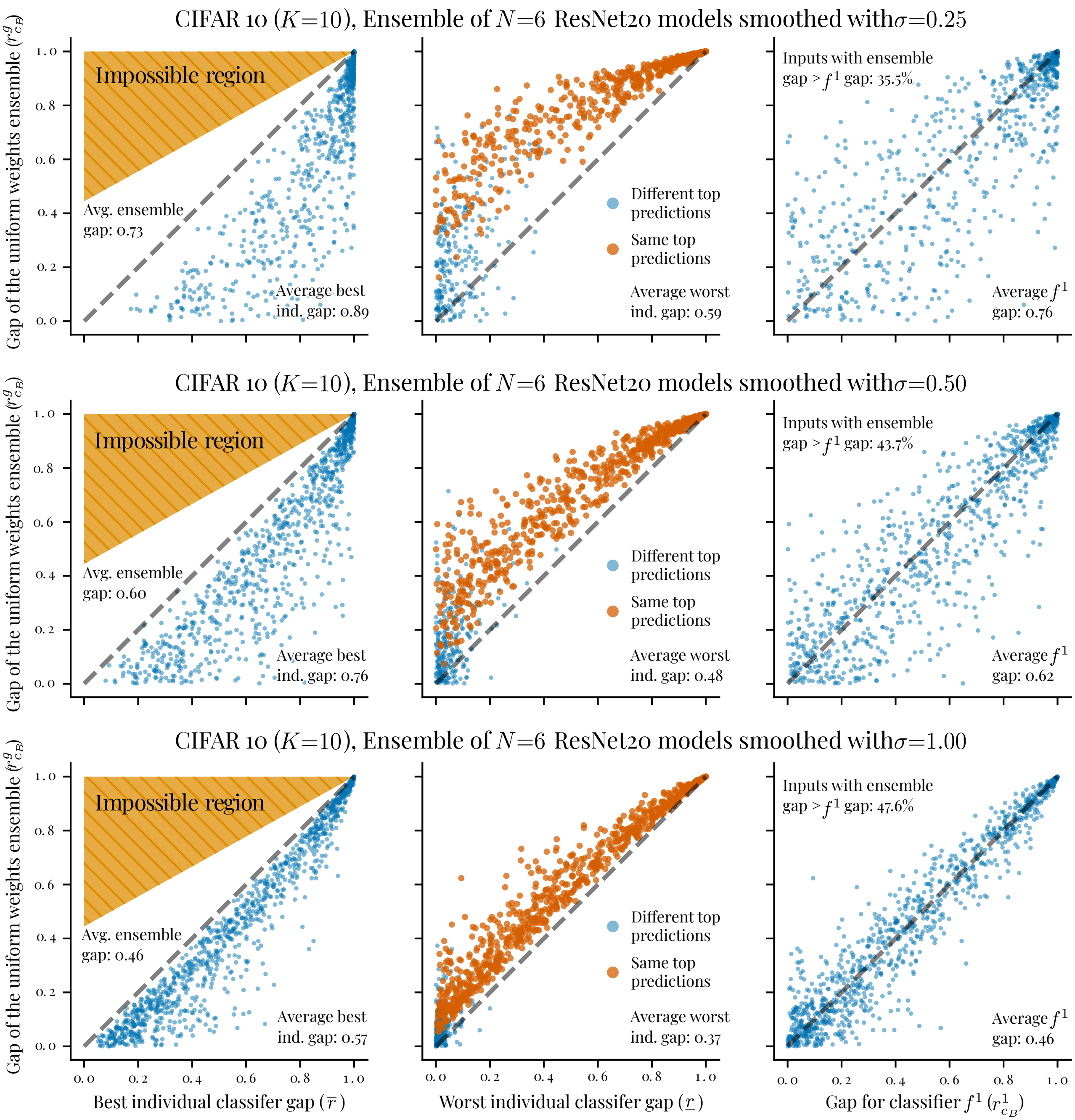

Figure B.1. Gap of the uniform weights ensemble plotted against the best individual gap (left), the worst individual gap (center) and against the gap of one of the constituent classifiers (right). The plots against the other classifiers are similar and are hence omitted. Each row shows one ensemble of 6 classifiers. Each individual classifier is a smoothed ResNet20 classifier trained by Horváth et al. (2021) using the train split of CIFAR10 and a different random seed. For these plots, we evaluate all classifiers at the same 1000 inputs from the CIFAR10 test split, each corresponding to a single point in the plots. We have reported the average value for the horizontal and vertical axis for each plot. The percentage of inputs for which the ensemble has a larger gap than the individual classifier, is also shown in the rightmost plots.

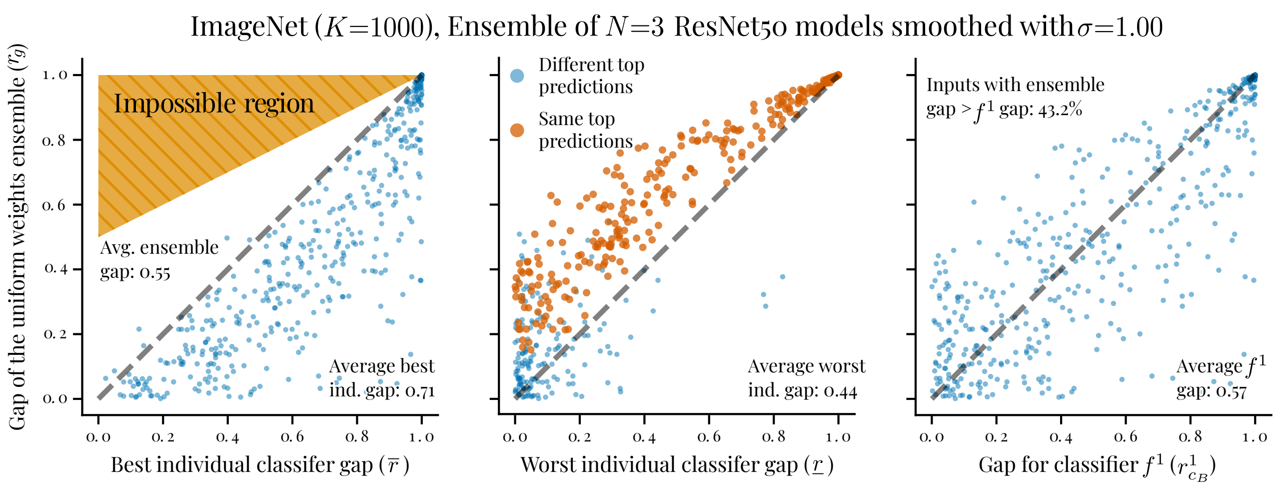

Figure B.2. Gap of the uniform weights ensemble plotted against the best individual gap (left), the worst individual gap (center) and against the gap of one of the constituent classifiers (right). The plots against the other classifiers are similar and are hence omitted. Each individual classifier is a smoothed ResNet50 classifier trained by Horváth et al. (2021) using the train split of ImageNet and a different random seed. For these plots, we evaluate all classifiers at the same 500 inputs from the ImageNet test split, each corresponding to a single point in the plots. We have reported the average value for the horizontal and vertical axis for each plot. The percentage of inputs for which the ensemble has a larger gap than the individual classifier, is also shown in the rightmost plot.

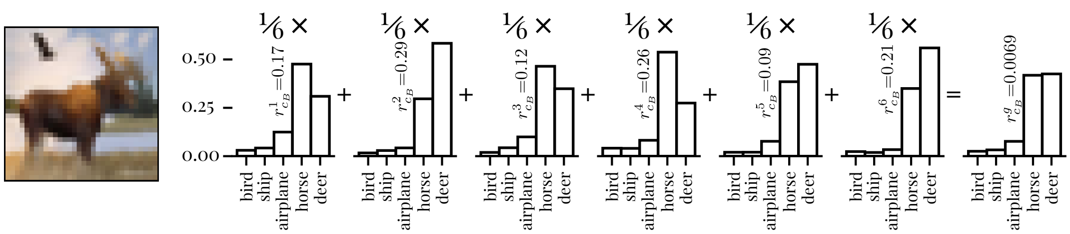

Figure B.3. A CIFAR10 sample for which the ResNet20 ($\sigma = 1.00$) ensemble is in regime ❸ and has a certificate Q_g barely larger than $\{0\}$. For clarity, only the 5 classes with the highest confidences are shown.

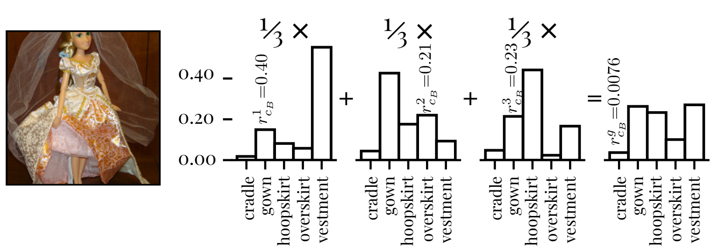

Figure B.4. An ImageNet sample for which the ResNet50 ensemble is in regime ❸ and has a certificate Q_g barely larger than $\{0\}$. For clarity, only the 5 classes with the highest confidences are shown.

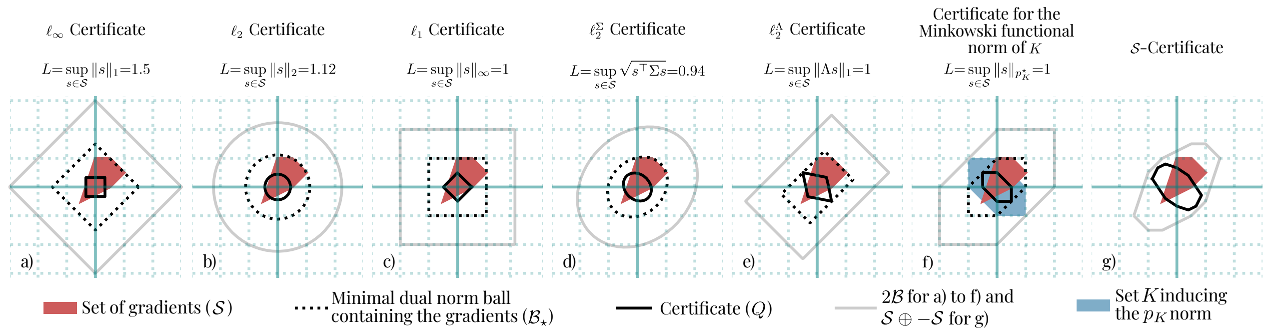

Figure B.5. (**a**-**f**) are Lipschitz certificates for the set of gradients $\mathcal{S}=\{\nabla f_i(x) : x\in\mathbb{R}^d, i=1,\dots,K\}$. We assume the U mode and $r_{c_B}=1$. $\mathcal{B}_\star$, the minimum dual norm ball containing $\mathcal{S}$, is shown. The certificate Q is the polar set $(2\mathcal{B}_\star)^1$. For (**d**) and (**e**) we have $\Sigma=\Lambda=\begin{bmatrix} 5/4 & 1/4 \\ 1/4 & 5/4 \end{bmatrix}$. (**f**) is the certificate constructed using the Minkowski functional norm (gauge) of K, the closed convex symmetric set marked in blue. (**g**) is the $\mathcal{S}$-certificate for the same $\mathcal{S}$. As there is no overapproximation of $\mathcal{S}$, the certificate is directly $Q=(\mathcal{S}\oplus-\mathcal{S})^1$, the largest of them all. Note that (**a**) and (**g**) are the same as (**a**) and (**b**) in Figure 1.

Figure B.3 shows one CIFAR10 sample and its predictions by all 6 ResNet20 ($\sigma=1.00$) models and the ensemble prediction. On average, the 6 classifiers have prediction gap 0.19, with the smallest one being $\underline{r}=r^5_{c_B}=0.09$. However, the ensemble gap is $r^g_{c_B}=0.0069$, more than an order of magnitude smaller than the smallest individual gap. Hence, the ensemble certificate would too be more than an order of magnitude smaller than the smallest individual certificate. This situation occurs as the 6 classifiers are split between classifying the input as a horse or a deer, resulting in very close predictions for the ensemble.

Similarly, the three ResNet50 classifiers have three different predictions for the input in Figure B.4, none of which is the correct class (`overskirt`). With $\underline{r}=0.21$ and $r^g_{c_B}=0.0076$, this leads to almost 30 times smaller certified radius of the ensemble compared with the least robust individual classifier.

In both of these examples, people would also likely be confused and would make mistakes. Perturbing just a couple of pixels in the CIFAR10 input would likely be sufficient to nudge one in classifying the input as `horse` or as `deer`. Therefore, lack of robustness in the ensemble might not be a bug, but in fact be a feature: a sign of better calibration.

Different top prediction is sufficient to ensure an ensemble is not in regime ❷. From Proposition 5 we know that inputs for which all individual classifiers agree ($c_A^=$) must be in regimes ❶ or ❷. From the center plots in Figures B.1 and B.2 one can observe that all inputs corresponding to this regime (in orange) are above the diagonal. Therefore, our experimental results support Proposition 5.

C. Additional examples

C.1. Examples of Lipschitz certificates for different norms

In the main text, we gave an illustration with an ℓ_∞ Lipschitz certificate in Figure 1. We offer some further examples here that we illustrate in Figure B.5 using the same classifier as in Figure 1.

Other ℓ_p certificates. Let's take a look at the other two commonly used ℓ_p certificates. First, there is the ℓ_2 certificate. From the Hölder inequality we have that the dual norm of ℓ_2 is again ℓ_2. Hence, the certificate can be computed by finding the radius of the smallest ℓ_2 ball that contains the gradients $\mathcal{S}$. In the case illustrated in Figure B.5b we have $\sup_{s\in\mathcal{S}}\|s\|_2=1.12$. Hence, f is 1.12-Lipschitz with respect to the ℓ_2 norm, and from Proposition 1 we have that the certificate Q is $\{\delta\in\mathbb{R}^2 : \|\delta\|_2\leq {}^1/_{2.24}\}$ which corresponds to the circle marked with ⁄ in Figure B.5b.

Similarly, the dual norm for ℓ_1 is ℓ_∞. Hence, we observe that f is 1-Lipschitz with respect to the ℓ_1 norm, that is $\sup_{s\in\mathcal{S}}\|s\|_\infty=1$. Therefore, the ℓ_1 certificate is $Q=\{\delta\in\mathbb{R}^2 : \|\delta\|_1\leq {}^1/_2\}$, the rhombus marked with ⁄ in Figure B.5c.

Anisotropic certificates. Proposition 1 is not limited to ℓ_p norms. Anisotropic certificates can be larger in some directions and smaller in others. This is in contrast with the ℓ_p certificates which have the same radius in all directions. This allows anisotropic certificates, in either of the CD, CW or U modes, to be tighter in directions with smaller gradients. For example, ellipsoidal certificates —certificates with the ℓ_2^Σ norm defined as $\|\delta\|_2^\Sigma=\sqrt{\delta^\top\Sigma^{-1}\delta}$— can be constructed by bounding the gradients with its dual norm $\ell_2^{\Sigma^{-1}}$. Similarly, generalized cross-polytopes can be con-

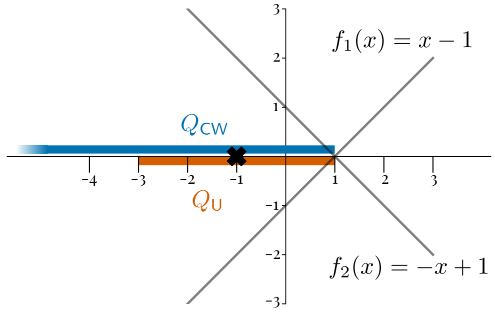

Figure C.1. Illustration of the one-dimensional binary classifier example in Appendix C.2.

structed with the ℓ_1^{Λ} norm defined as $\|\delta\|_1^{\Lambda} = \left\|\Lambda^{-1}\delta\right\|_1$ by bounding gradients with its dual norm $\ell_\infty^{\Lambda^{-1}}$. The smallest norm balls (⋰) for $\Sigma = \Lambda = \left[\begin{smallmatrix} 5/4 & 1/4 \\ 1/4 & 5/4 \end{smallmatrix}\right]$ and the corresponding certificates (╱) are shown in Figure B.5d and e. Refer to Eiras et al. (2022) for further details.

Arbitrary norms defined as Minkowski functionals. Any closed convex symmetric set $K \subset \mathbb{R}^d$ containing the origin gives rise to a norm on $\mathbb{R}^d$ defined as $p_K(x) := \inf\{a \in \mathbb{R} : a > 0 \text{ and } x \in aK\}$. This is called *Minkowski functional* or *gauge* of K (Schechter, 1997). Intuitively, $p_K(x)$ measures how much we need to scale K in order to have x barely fitting in it, *i.e.*, x being on the border of the scaled K. Figure 1f illustrates such a closed convex symmetric set K in ● and the minimum dual $p_K^\star$ norm containing $\mathcal{S}$ (⋰) with a radius $\sup_{s \in \mathcal{S}} \|s\|_{p_K^\star} = 1$. Therefore, the certificate is the p_K norm ball of radius $1/2$, shown in ╱.

Comparison with the $\mathcal{S}$-certificate. The $\mathcal{S}$-certificate shown with ╱ in Figure 1g is the largest of all seven certificates. Proposition 8 shows that this must always be the case: there is no norm for which the Lipschitz certificate will be a strict superset of the $\mathcal{S}$-certificate. More detailed explanation is offered in Section 3.3 in the main text.

C.2. One-dimensional binary classifier example

Linear classifiers are easy to analyse as their $\mathcal{S}$ sets are singleton sets. Let's then see the difference between the Lipschitz and the $\mathcal{S}$-Lipschitz certificates for a one-dimensional linear binary classifier defined as

$$f_1(x) = x - 1, \qquad f_2(x) = -x + 1.$$

For this classifier we have $\mathcal{S}_1 = \{+1\}$ and $\mathcal{S}_2 = \{-1\}$. We want to compute certificates for the input $x = -1$. Hence $c_A = 2$ and $r_{c_B} = f_2(-1) - f_1(-1) = 4$. Let's first consider the $\boxed{\mathsf{CW}}$ certificate from Equation (3). We have $Q_{\mathsf{CW}} = (\mathcal{S}_1 \oplus -\mathcal{S}_2)^r = (\{1\} \oplus -\{-1\})^4 = \{2\}^4 = (-\infty, 2]$. This certificate is shown in blue in Figure C.1. If we instead construct the $\boxed{\mathsf{U}}$ certificate by taking the smallest $\mathcal{S}$ such that f_1 and f_2 are $\mathcal{S}$-Lipschitz, then we get $\mathcal{S} = \{-1, 1\}$. Certifying using this $\mathcal{S}$, Equation (4) gives us $Q_{\mathsf{U}} = (\mathcal{S} \oplus -\mathcal{S})^r = \{-2, 0, 2\}^4 = [-2, 2]$. This certificate is shown in orange in Figure C.1. f is 1-Lipschitz with respect to any ℓ_p norm and the Lipschitz certificate Proposition 1 results in the same certified perturbation set: $[-2, 2]$ for any ℓ_p. Therefore, even in this simple case, we see that the $\boxed{\mathsf{CW}}$ $\mathcal{S}$-certificate covers the whole domain in which f predicts 2 while the Lipschitz approach and the $\boxed{\mathsf{U}}$ $\mathcal{S}$-certificate are limited to the largest *symmetric* perturbation set.

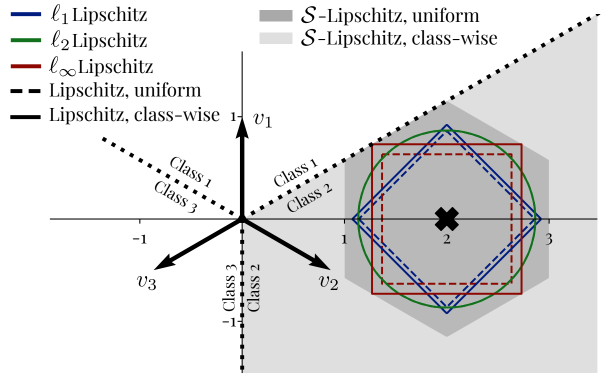

Figure C.2. Illustration of the two-dimensional three-way classifier example from Appendix C.3 and Figure 2.

C.3. Derivation of the certificates in Figure 2

This is an extended explanation of Figure 2 with all the intermediate steps and calculations.

Consider the 3-class two-dimensional linear classifier defined as:

$$\begin{aligned} f_1(x) &= x^\top v_1 = [0, 1] \cdot x \\ f_2(x) &= x^\top v_2 = [\sqrt{3}/2, -1/2] \cdot x \\ f_3(x) &= x^\top v_3 = [-\sqrt{3}/2, -1/2] \cdot x \end{aligned}$$

We want to construct a certificate for $x_0 = [2, 0]^\top$. We then have $f_1(x_0) = 0$, $f_2(x_0) = \sqrt{3}$, $f_3(x_0) = -\sqrt{3}$, $c_A = 2$, $c_B = 1$, $r_1 = \sqrt{3}$, $r_3 = 2\sqrt{3}$.

Let's first consider the $\boxed{\mathsf{U}}$ Lipschitz case using the observation that f_1, f_2, and f_3 are L^p-Lipschitz for the ℓ_p norm with $L^1 = L^2 = 1$, $L^\infty = (\sqrt{3}+1)/2$ (from Aux. Lemma 2). The respective certificates would be the ℓ_p ball with radius $r_1/2L^p$, as shown in Figure C.2. Now, let's compare with the

CW case.

$$\mathcal{S}_1 = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} \qquad L_1^1 = 1 \qquad L_1^2 = 1 \quad L_1^\infty = 1$$
$$\mathcal{S}_2 = \left\{ \begin{bmatrix} \sqrt{3}/2 \\ -1/2 \end{bmatrix} \right\} \qquad L_2^1 = \frac{\sqrt{3}}{2} \quad L_2^2 = 1 \quad L_2^\infty = \frac{\sqrt{3}+1}{2}$$
$$\mathcal{S}_3 = \left\{ \begin{bmatrix} -\sqrt{3}/2 \\ -1/2 \end{bmatrix} \right\} \quad L_3^1 = \frac{\sqrt{3}}{2} \quad L_3^2 = 1 \quad L_3^\infty = \frac{\sqrt{3}+1}{2}$$

The respective certificates would be the intersection of the ℓ_p balls with radius $\min\{r_1/(L_1^p + L_2^p), r_3/(L_3^p + L_2^p)\}$. For ℓ_1 and ℓ_∞ we observe increased certified radii when using CW Lipschitzness: respectively from $\sqrt{3}/2$ to $2\sqrt{3}/(2+\sqrt{3})$ and from $\sqrt{3}/(1+\sqrt{3})$ to $2\sqrt{3}/(3+\sqrt{3})$. The certified radius for ℓ_2 remained unchanged: $\sqrt{3}/2$: that's because $L_1^2 = L_2^2 = L_3^2$ and hence we don't overapproximate the true smoothness in the U case.

Next, let's do the same analysis using $\mathcal{S}$-Lipschitzness instead. In the U case, we have that f is $\mathcal{S}$-Lipschitz with $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$. Therefore, the certified set is the hexagon in Figure C.2 (via Aux. Lemma 8).

Finally, let's take a look at the CW $\mathcal{S}$-certificate: this should give us the largest certified region. Again using Aux. Lemma 8 we have

$$\begin{aligned} Q &= (\mathcal{S}_1 \oplus -\mathcal{S}_2)^{r_1} \cap (\mathcal{S}_3 \oplus -\mathcal{S}_2)^{r_2} \\ &= \{x \in \mathbb{R}^d : [-1/2, \sqrt{3}/2] \cdot x \leq 1 \ \wedge\ [-1/2, 0] \cdot x \leq 1\}. \end{aligned}$$

This is all of the domain that f classifies as class 2.

Hence, the CW $\mathcal{S}$-Lipschitz approach gives us the maximum possible certified domain: the whole preimage of the class 2 prediction. All CW Lipschitz certificates are smaller than the CW $\mathcal{S}$-certificate as they consider only the norm of the gradients and ignores their orientation. Similarly all U Lipschitz certificates are smaller than the U $\mathcal{S}$-certificate. The U $\mathcal{S}$-certificate is smaller than the CW $\mathcal{S}$-certificate as it ignores the class-wise differences, and similarly the U Lipschitz certificate is smaller than the CW Lipschitz certificate.

C.4. Example for Theorem 4

Take two classifiers $f^1, f^2 : \mathbb{R}^2 \to \mathbb{R}^K$ under the conditions in Theorem 4. Assume further that their $\mathcal{S}$-Lipschitz sets have the same shape but possibly different sizes. That is, $\mathcal{S}^1 = \epsilon_1 \mathcal{B}_\star, \mathcal{S}^2 = \epsilon_2 \mathcal{B}_\star$, $\epsilon_1, \epsilon_2 > 0$ where $\mathcal{B}_\star = \{x \in \mathbb{R}^d : \|x\|_\star \leq 1\}$ for some norm $\|\cdot\|_\star$. We use $\mathcal{B}$ to denote the unit ball defined by the dual norm $\|\cdot\|$. From Theorem 2, we have

$$Q_1 = \frac{r^1_{c_B}}{2\epsilon_1}\mathcal{B}, \quad Q_2 = \frac{r^2_{c_B}}{2\epsilon_2}\mathcal{B}, \quad Q_g = \frac{\alpha_1 r^1_{c_B} + \alpha_2 r^2_{c_B}}{2(\alpha_1\epsilon_1 + \alpha_2\epsilon_2)}\mathcal{B}.$$

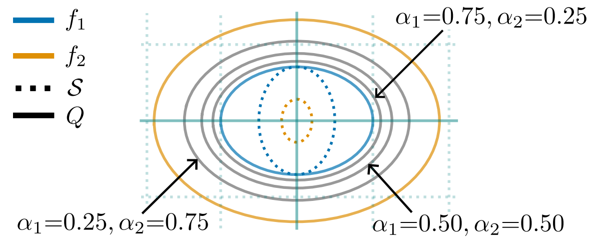

Figure C.3. Illustration for the example in Appendix C.4.

The radius of Q_g interpolates from $r^1_{c_B}/2\epsilon_1$ to $r^2_{c_B}/2\epsilon_2$ and can never be larger than $\max\{r^1_{c_B}/2\epsilon_1, r^2_{c_B}/2\epsilon_2\}$. Therefore, in this setting, ensembling will always result in a smaller certified radius than the most robust individual classifier.

We illustrate this phenomenon in Figure C.3. Consider the anisotropic ellipsoidal norm $\|x\| = \sqrt{x^\top \left[\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right] x}$ (see Appendix C.1 for further details on this norm). The radii of $\mathcal{S}^1$ and $\mathcal{S}^2$ are respectively $\epsilon_1{=}1/2$ and $\epsilon_2{=}1/5$ (shown in $\therefore$), and their prediction gaps are $r^1_{c_B}{=}1$ and $r^2_{c_B}{=}3/4$. We show the certificate Q_1 for f_1 as the smallest ellipse and the certificate Q_2 for f_2 as the largest one. We also show three sets of mixing coefficients α_1, α_2 in grey, which all fall between Q_1 and Q_2. This illustrates how in the U, $c_A^=$, $c_B^=$ and same shape of the $\mathcal{S}$-Lipschitzness regime, we will always have the largest certified radius by picking the best individual classifier (f_2 in this case), instead of ensembling.

D. Deferred Proofs

Aux. Lemma 1. *Consider a classifier $f : \mathbb{R}^d \to \mathbb{R}^K$ such that f_i is L_i-Lipschitz with respect to the norm $\|\cdot\|$, that is $|f_i(x) - f_i(x')| \leq L_i \|x - x'\|$, $\forall i, x, x'$. Then, at a fixed x, we have $\arg\max_i f_i(x+\delta) = c_A$ for all $\|\delta\| \leq \min_{i \neq c_A} (f_{c_A}(x) - f_i(x))/(L_{c_A} + L_i)$, where $c_A = \arg\max_i f_i(x)$.*

Proof of Aux. Lemma 1. From the definition of f_i being L_i-Lipschitz it follows that for all $i = 1, \ldots, K$:

$$f_i(x) + L_i \|\delta_i\| \geq f_i(x + \delta_i) \geq f_i(x) - L_i \|\delta_i\|$$

For $\arg\max_c f_c(x+\delta) = c_A$ it must be that $f_i(x+\delta) \leq f_{c_A}(x+\delta)$ for all $i \neq c_A$. By applying the above inequalities for every $i \neq c_A$ we obtain:

$$\begin{aligned} f_{c_A}(x) - L_{c_A}\|\delta_i\| - f_i(x) - L_i\|\delta_i\| \geq 0 \\ \|\delta_i\| \leq \frac{f_{c_A}(x) - f_i(x)}{L_{c_A} + L_i}. \end{aligned} \tag{9}$$

Equation (9) is an upper bound of the perturbation that will not change the prediction from c_A to i. Since this must hold for all $i \neq c_A$, it is only valid for the intersection of these perturbation sets, *i.e.*, $\|\delta\| \leq \min_{i \neq c_A} {}^{(f_{c_A}(x) - f_i(x))}/_{(L_{c_A} + L_i)}$. □

Aux. Lemma 2. *Consider a differentiable $h : \mathbb{R}^d \to \mathbb{R}$, such that $\sup_x \|\nabla h(x)\|_\star \leq L$, where $\|\cdot\|_\star$ is the dual norm of $\|\cdot\|$. Then h is L-Lipschitz with respect to $\|\cdot\|$.*

Proof of Aux. Lemma 2. See proof of Proposition 1 from (Eiras et al., 2022). □

Proposition 1 (Certification of Lipschitz classifiers)**.** *Take a differentiable[5] classifier $f : \mathbb{R}^d \to \mathbb{R}^K$ such that $\sup_x \|\nabla f_i(x)\|_\star \leq L_i$, $\forall i$. Then f_i is L_i-Lipschitz with respect to $\|\cdot\|$. Moreover, f has a certificate*

$$Q = \left\{ \delta \in \mathbb{R}^d : \|\delta\| \leq \min_{i \neq c_A} \frac{f_{c_A}(x) - f_i(x)}{L_i + L_{c_A}} = \min_{i \neq c_A} \frac{r_i}{L_i + L_{c_A}} \right\}. \tag{1}$$

Here, $\|\cdot\|_\star$ is the dual norm to $\|\cdot\|$ and c_A is $\arg\max_i f_i(x)$. If all classes have the same Lipschitz constant L, i.e., $L_i \leq L, \forall i$, the certificate simplifies to

$$Q = \left\{ \delta \in \mathbb{R}^d : \|\delta\| \leq \frac{f_{c_A}(x) - f_{c_B}(x)}{2L} = \frac{r_{c_B}}{2L} \right\}, \tag{2}$$

where $c_B = \arg\max_{i \neq c_A} f_i(x)$.

Proof of Proposition 1. Follows directly from Aux. Lemmas 1 and 2. □

Aux. Lemma 3. *For a bounded set $\mathcal{S} \subseteq \mathbb{R}^d$, it holds that $\rho_{\mathrm{hull}\,\mathcal{S}}(\delta) = \rho_{\mathcal{S}}(\delta)$, $\forall \delta \in \mathbb{R}^d$. In other words, f being $\mathcal{S}$-Lipschitz is the same as it being $(\mathrm{hull}\,\mathcal{S})$-Lipschitz.*

Aux. Lemma 4 ($\mathcal{S}$-Lipschitz function and gradients)**.** *Consider a differentiable $f : \mathbb{R}^d \to \mathbb{R}$. If $\nabla f : \mathbb{R}^d \to \mathcal{S}$, then f is $\mathcal{S}$-Lipschitz. The reverse also holds: if f is $\mathcal{S}$-Lipschitz, then, $\nabla f(x) \in \mathrm{hull}\,\mathcal{S}$, $\forall x \in \mathbb{R}^d$.*

Proof of Aux. Lemma 4. Let's first start by showing that a function with bounded gradients is $\mathcal{S}$-Lipschitz. Consider any $x, y \in \mathbb{R}^d$ and define $\gamma : [0, 1] \to \mathbb{R}^d$ where $\gamma(t) = (1-t)x + ty$. Then we have that:

$$\begin{aligned}
f(y) - f(x) &= f(\gamma(1)) - f(\gamma(0)) \\
&= \int_0^1 \frac{df(\gamma(t))}{dt} dt \\
&= \int_0^1 \frac{df(\gamma(t))}{d\gamma(t)} \frac{d\gamma(t)}{dt} dt \\
&= \int_0^1 \nabla_x f(\gamma(t))^\top \nabla_t \gamma(t) dt \\
&= \int_0^1 \nabla_x f((1-t)x + ty)^\top (y - x) dt \\
&\leq \int_0^1 \max_{t \in [0,1]} \left\{ \nabla_x f((1-t)x + ty)^\top (y - x) \right\} dt \\
&= \max_{t \in [0,1]} \left\{ \nabla_x f((1-t)x + ty)^\top (y - x) \right\} \\
&\leq \sup_{\nabla f \in \mathcal{S}} \nabla f^\top (y - x) \\
&= \rho_{\mathcal{S}}(y - x),
\end{aligned}$$

where we used the fundamental theorem of calculus, the fact that f is continuous, and Definition 2. Similarly,

$$\begin{aligned}
f(y) - f(x) &\geq \int_0^1 \min_{t \in [0,1]} \left\{ \nabla_x f((1-t)x + ty)^\top (y - x) \right\} dt \\
&= \min_{t \in [0,1]} \left\{ \nabla_x f((1-t)x + ty)^\top (y - x) \right\} \\
&= - \max_{t \in [0,1]} \left\{ \nabla_x f((1-t)x + ty)^\top (x - y) \right\} \\
&\geq - \sup_{\nabla f \in \mathcal{S}} \nabla f^\top (x - y) \\
&= -\rho_{\mathcal{S}}(x - y).
\end{aligned}$$

Next, let's show the reverse: that if a function is $\mathcal{S}$-Lipschitz, then its gradients must be in $\mathrm{hull}\,\mathcal{S}$. If f is $\mathcal{S}$-Lipschitz, then $f(y) - f(x) \leq \sup_{c \in \mathcal{S}} c^\top (y - x)$. Consider the directional derivative of f in direction $v \in \mathbb{R}^d$ at x, and taking $y = x + hv$:

$$\begin{aligned}
\nabla_v f(x) &= \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h} \\
&\leq \lim_{h \to 0} \frac{\sup_{c \in S} c^\top (hv)}{h} \\
&= \lim_{h \to 0} \frac{h \sup_{c \in S} c^\top v}{h} \\
&\overset{*}{=} \sup_{c \in S} c^\top v \\
&= \rho_{\mathcal{S}}(v),
\end{aligned}$$

with $*$ following from the L'Hôpital's rule. Similarly, $\nabla_v f(x) \geq -\rho_{\mathcal{S}}(-v)$, $\forall x \in \mathbb{R}^d$. Hence, we have:

$$-\rho_{\mathcal{S}}(-v) = -\sup_{c \in S} c^\top (-v) \leq \nabla f(x)^\top v \leq \sup_{c \in S} c^\top v. \tag{10}$$

Now we need to show that Equation (10) implies that $\nabla f(x) \in \operatorname{hull}\mathcal{S}$. By the properties of support functions of convex sets we have that $\nabla f(x) \in \operatorname{hull}\mathcal{S}$ iff

$$\sup_{c' \in \operatorname{hull}\mathcal{S}} \left\{c'^\top \nabla f(x) - \rho_{\mathcal{S}}(c')\right\} = 0.$$

In the above, we use that $\rho_{\mathcal{S}}(\delta) = \rho_{\operatorname{hull}\mathcal{S}}(\delta)$ (Aux. Lemma 3). Substituting from Equation (10) we get:

$$\begin{aligned}&\sup_{c' \in \operatorname{hull}\mathcal{S}} \left\{c'^\top \nabla f(x) - \rho_{\mathcal{S}}(c')\right\} \\ &\quad \le \sup_{c' \in \operatorname{hull}\mathcal{S}} \left\{\sup_{c\in\mathcal{S}} c^\top c' - \sup_{c\in\mathcal{S}} c^\top c'\right\} = 0,\end{aligned}$$

hence, $\nabla f(x) \in \operatorname{hull}\mathcal{S},\ \forall x \in \mathbb{R}^d$. ☐

Aux. Lemma 5. *Given $\mathcal{S} \subseteq \mathbb{R}^n$, it holds that $\rho_{\mathcal{S}}(\delta) = \rho_{-\mathcal{S}}(-\delta),\ \forall \delta \in \mathbb{R}^d$.*

Proof of Aux. Lemma 5. $\rho_{-\mathcal{S}}(-\delta) = \sup_{c\in-\mathcal{S}} c^\top(-\delta) = \sup_{c\in\mathcal{S}} (-c)^\top(-\delta) = \sup_{c\in\mathcal{S}} c^\top\delta = \rho_{\mathcal{S}}(\delta)$. ☐

Aux. Lemma 6. *Given $\mathcal{S}, \mathcal{S}' \subseteq \mathbb{R}^n$, for all $\delta \in \mathbb{R}^d$ it holds that $\rho_{\mathcal{S}}(\delta) + \rho_{-\mathcal{S}'}(\delta) = \rho_{\mathcal{S}\oplus-\mathcal{S}'}$, where $\oplus$ is the Minkowski sum operator.*

Proof of Aux. Lemma 6.

$$\begin{aligned}\sup_{c\in\mathcal{S}}(c^\top\delta) + \sup_{c'\in-\mathcal{S}'}(c'^\top\delta) &= \sup_{c\in\mathcal{S}}(c^\top\delta) + \sup_{c'\in\mathcal{S}'}(-c'^\top\delta) \\ &= \sup_{c\in\mathcal{S}, c'\in\mathcal{S}'} (c-c')^\top\delta.\end{aligned}$$

At the same time, by definition of the Minkowski sum:

$$\rho_{\mathcal{S}\oplus-\mathcal{S}'}(\delta) = \sup_{c\in\mathcal{S}\oplus-\mathcal{S}'} c^\top\delta = \sup_{c\in\mathcal{S}, c'\in-\mathcal{S}'} (c+c')^\top\delta.$$

☐

Theorem 1 ($\mathcal{S}$-certificates)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^K$ be a classifier with f_i being differentiable and $\nabla f_i : \mathbb{R}^d \to \mathcal{S}_i$ for all $i = 1, \dots, K$. Then, each f_i is $\mathcal{S}_i$-Lipschitz. Furthermore, for a fixed x, f is robust at x against all δ in*

$$Q = \bigcap\nolimits_{i\neq c_A} \left(\mathcal{S}_i \oplus -\mathcal{S}_{c_A}\right)^{r_i}. \tag{3}$$

Here, $c_A = \arg\max_c f_c(x)$, $r_i = f_{c_A}(x) - f_i(x)$, and $\oplus$ is the Minkowski sum. If $\mathcal{S} \supseteq \mathcal{S}_i, \forall i$, then we have the simplified certificate

$$Q = (\mathcal{S} \oplus -\mathcal{S})^{r_{c_B}}, \tag{4}$$

where $c_B = \arg\max_{c\neq c_A} f_c(x)$.

Proof of Theorem 1. The connection between gradients and $\mathcal{S}$-Lipschitzness comes from Aux. Lemma 4.

Following Definition 2, we have:

$$\begin{aligned}f_{c_A} - \rho_{\mathcal{S}_{c_A}}(x-y) &\le f_{c_A}(y) \\ f_i + \rho_{\mathcal{S}_i}(y-x) &\ge f_i(y),\ \forall i \neq c_A.\end{aligned}$$

We want $f_{c_A}(y) > f_i(y),\ \forall i \neq c_A$, hence, a sufficient condition following the two inequalities above is

$$\begin{aligned}&f_i + \rho_{\mathcal{S}_i}(y-x) < f_{c_A} - \rho_{\mathcal{S}_{c_A}}(x-y) \iff \\ &\qquad \rho_{\mathcal{S}_i}(y-x) + \rho_{\mathcal{S}_{c_A}}(x-y) < f_{c_A} - f_i.\end{aligned}$$

Using Aux. Lemmas 5 and 6 and setting $\delta = y - x$ we get:

$$\rho_{\mathcal{S}_i\oplus-\mathcal{S}_{c_A}}(\delta) < f_{c_A} - f_i,$$

which is the definition of $(\mathcal{S}_i \oplus -\mathcal{S}_{c_A})^{r_i}$. As this needs to hold for all $i \neq c_A$ we take the intersection.

To show Equation (4) observe that $r_{c_B} \le r_i, \forall i \neq c_A$. Hence, by Proposition 2iii, $(\mathcal{S}_i \oplus -\mathcal{S}_{c_A})^{r_i} \supseteq (\mathcal{S}_i \oplus -\mathcal{S}_{c_A})^{r_{c_B}}$. The rest follows from $\mathcal{S}_i \subseteq \mathcal{S}$ and Proposition 2iv. ☐

Proposition 7 (Tightness of $\mathcal{S}$-certificates)**.** *For any $\delta \notin (\mathcal{S} \oplus -\mathcal{S})^r$, there exists an $f : \mathbb{R}^d \to \mathbb{R}^K$ with f_i $\mathcal{S}$-Lipschitz for all i and $r_{c_B} = f_{c_A}(x) - f_{c_B}(x)$ such that $\arg\max_i f_i(x+\delta) \neq c_A$.*

Proof of Proposition 7. Let's take a constructive approach and provide a classifier that classifies x and $x+\delta$ differently. For simplicity, we will consider a binary classifier. Let's fix $x \in \mathbb{R}^d$, $\delta \notin (\mathcal{S} \oplus -\mathcal{S})^{r_{c_B}}$ and construct a classifier that reduces the gap between c_A and c_B as much as possible, while still being $\mathcal{S}$-Lipschitz. Take $c, c' \in \mathcal{S}$ that attain the supremum

$$\rho_{\mathcal{S}\oplus-\mathcal{S}}(\delta) = \sup_{c\in\mathcal{S},\ c'\in\mathcal{S}} (c-c')^\top\delta > r_{c_B}.$$

Note that c, c' depend only on $\mathcal{S}$ and δ but not on the classifier f. Now, let's define the classifier f as:

$$\begin{aligned}f_{c_A}(y) &= (y-x)^\top c' + r_{c_B} \\ f_{c_B}(y) &= (y-x)^\top c.\end{aligned}$$

We can verify that $f_{c_A}(x) > f_{c_B}(x)$ and that $f_{c_A}(x) - f_{c_B}(x) = r_{c_B}$, as well as that f_{c_A} and f_{c_B} are $\mathcal{S}$-Lipschitz, hence satisfying all requirements. However, we also have:

$$\begin{aligned}&f_{c_A}(x+\delta) - f_{c_B}(x+\delta) \\ &= \delta^\top c' + r_{c_B} - \delta^\top c = r_{c_B} - (c-c')^\top\delta \\ &< 0,\end{aligned}$$

hence $\arg\max_i f_i(x+\delta) = c_B$. ☐

Proposition 2 (Polar set dependence on $\mathcal{S}$ and r)**.** *Let $\mathcal{S}, \mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4 \subset \mathbb{R}^d$ be bounded and $r, r_1, r_2 > 0$:*

i. $\mathcal{S}_1 \subseteq \mathcal{S}_2 \Rightarrow (\mathcal{S}_1 \oplus \text{-}\mathcal{S}_1) \subseteq (\mathcal{S}_2 \oplus \text{-}\mathcal{S}_2)$*;*
ii. $\mathcal{S}_1 \subseteq \mathcal{S}_2 \Rightarrow (\mathcal{S}_1)^r \supseteq (\mathcal{S}_2)^r$*;*
iii. $r_1 \leq r_2 \Rightarrow (\mathcal{S})^{r_1} \subseteq (\mathcal{S})^{r_2}$*;*
iv. $((\mathcal{S}_1 \subseteq \mathcal{S}_3) \wedge (\mathcal{S}_2 \subseteq \mathcal{S}_4)) \Rightarrow (\mathcal{S}_3 \oplus \text{-}\mathcal{S}_4)^r \subseteq (\mathcal{S}_1 \oplus \text{-}\mathcal{S}_2)^r$.

where $\oplus$ is the Minkowski sum operator.

Proof of Proposition 2.
Proof of i.: For all $s \in (\mathcal{S}_1 \oplus -\mathcal{S}_1)$ there must be some $s', s'' \in \mathcal{S}_1$ such that $s' - s'' = s$. But $s', s'' \in \mathcal{S}_2$ and hence $s' - s''$ must also be in $\mathcal{S}_2 \oplus -\mathcal{S}_2$.
Proof of ii.: We have to show that $\forall y \in \mathbb{R}^d$ we have $\sup_{x \in \mathcal{S}_2} x^\top y \leq r$ implying $\sup_{x \in \mathcal{S}_1} x^\top y \leq r$. This is equivalent to showing that

$$\sup_{x \in \mathcal{S}_1} x^\top y \leq \sup_{x \in \mathcal{S}_2} x^\top y, \ \forall y \in \mathbb{R}^d. \tag{11}$$

We can rewrite the right-hand side as

$$\sup_{x \in \mathcal{S}_2} x^\top y = \max \left\{ \sup_{x \in \mathcal{S}_1} x^\top y, \sup_{x \in \mathcal{S}_2 \setminus \mathcal{S}_1} x^\top y \right\},$$

for all $y \in \mathbb{R}^d$, hence Equation (11) is always true.
Proof of iii.: If $y \in (S)^{r_1}$ then:

$$\sup_{x \in \mathcal{S}} x^\top y \leq r_1.$$

But then it also holds that $\sup_{x \in \mathcal{S}2} x^\top y \leq r_2$ as $r_2 \geq r_1$ and hence $y \in (\mathcal{S})^{r_2}$.
Proof of iv.: If $y \in (\mathcal{S}_3 \oplus -\mathcal{S}_4)^r$ then for all $s_3 \in \mathcal{S}_3, s_4 \in \mathcal{S}_4$ it holds that $(s_3 - s_4)^\top y \leq r$. However, as $\mathcal{S}_1$ and $\mathcal{S}_2$ are subsets of respectively $\mathcal{S}_3$ and $\mathcal{S}_4$ it must then also hold that $\forall s_1 \in \mathcal{S}_1, \forall s_2 \in \mathcal{S}_2$ we have $(s_1 - s_2)^\top y \leq r$. This implies that $y \in (\mathcal{S}_1 \oplus -\mathcal{S}_2)^r$. $\square$

Proposition 8 (The $\mathcal{S}$-certificate subsumes any Lipschitz certificate)**.** *Take $f : \mathbb{R}^d \to \mathbb{R}^K$ to be a classifier that such that:*

i. *f_i is $\mathcal{S}_i$-Lipschitz for every $i = 1, \ldots, K$ and $\mathcal{S}_i$ is the smallest such set (CW case); or*
ii. *f_i is $\mathcal{S}$-Lipschitz for all $i = 1, \ldots, K$ (U case) and $\mathcal{S}$ is the smallest such set.*

Consider a fixed input $x \in \mathbb{R}^d$. Then, the corresponding $\mathcal{S}$-certificate from Theorem 1 at x is always a superset of the Lipschitz certificate for any norm $\|\cdot\|$.

Proof of Proposition 8. We will only consider the CW case as the U follows trivially from it. As discussed in the main text, if f_i is L_i-Lipschitz with respect to the norm $\|\cdot\|$, then the Lipschitz certificate at x is equal to the $\mathcal{B}_{i,\star}$-Lipschitz certificate, where $\mathcal{B}_{i,\star} = \{y \in \mathbb{R}^d : \|y\|_\star \leq L_i\}$. Formally:

$$\begin{aligned} Q_{\text{Lip}} &= \left\{ \delta \in \mathbb{R}^d : \|\delta\| \leq \min_{i \neq c_A} \frac{r_i}{L_i + L_{c_A}} \right\} \\ &= \bigcap_{i \neq c_A} (\mathcal{B}_{i,\star} \oplus \text{-}\mathcal{B}_{c_A,\star})^{r_i}. \end{aligned}$$

At the same time, the $\mathcal{S}$-certificate is:

$$Q_{\mathcal{S}} = \bigcap_{i \neq c_A} (\mathcal{S}_i \oplus -\mathcal{S}_{c_A})^{r_i}.$$

Next, note that $\mathcal{S}_i \subseteq \mathcal{B}_{i,\star}$, regardless of the choice of the norm $\|\cdot\|$. This follows from the definitions of $\mathcal{S}_i = \{\nabla f_i(z) : z \in \mathbb{R}^d\}$ and $\mathcal{B}_{i,\star} = \{y \in \mathbb{R}^d : \|y\|_\star \leq \sup_z \|\nabla f_i(z)\|_\star\}$.
As $\mathcal{S}_i \subseteq \mathcal{B}_{i,\star}$, from Proposition 2iv we have:

$$(\mathcal{S}_i \oplus -\mathcal{S}_{c_A})^{r_i} \supseteq (\mathcal{B}_{i,\star} \oplus -\mathcal{B}_{i,\star})^{r_i}.$$

Finally, as set intersection preserves the superset relation, we have that

$$Q_{\mathcal{S}} = \bigcap_{i \neq c_A} (\mathcal{S}_i \oplus \text{-}\mathcal{S}_{c_A})^{r_i} \supseteq \bigcap_{i \neq c_A} (\mathcal{B}_{i,\star} \oplus \text{-}\mathcal{B}_{c_A,\star})^{r_i} = Q_{\text{Lip}}.$$

$\square$

Aux. Lemma 7. *For any bounded set $\mathcal{S} \subset \mathbb{R}^d$ it holds that $\mathcal{S} \oplus -\mathcal{S}$ is symmetric, i.e.*

$$x \in (\mathcal{S} \oplus -\mathcal{S}) \Rightarrow -x \in (\mathcal{S} \oplus -\mathcal{S}).$$

Furthermore, for any $r > 0$ and any symmetric $\mathcal{S} \subset \mathbb{R}^d$, it holds that $\mathcal{S}^r$ is also symmetric. Finally, if $\mathcal{S}$ is symmetric and convex, then

$$\mathcal{S} \oplus -\mathcal{S} = 2\mathcal{S}.$$

Proof of Aux. Lemma 7. If $x \in (\mathcal{S} \oplus -\mathcal{S})$ then $\exists s_1, s_2 \in \mathcal{S}$ such that $x = s_1 - s_2$. However, then it also must hold that $s_2 - s_1 = -x$ is in $\mathcal{S} \oplus -\mathcal{S}$.

Let's now prove the second part. The condition for $-y$ to be in $\mathcal{S}^r$ when $\mathcal{S}$ is symmetric is $\sup_{x \in \mathcal{S}} x^\top(-y) \leq r$. The left side can be rewritten as:

$$\sup_{x \in \mathcal{S}} x^\top(-y) = \sup_{x \in \mathcal{S}} (-x)^\top y = \sup_{x \in -\mathcal{S}} x^\top y = \sup_{x \in \mathcal{S}} x^\top y,$$

which is the same as the condition for y to be in $\mathcal{S}^r$. In the above, we use the fact that $\mathcal{S} = -\mathcal{S}$, the definition of $\mathcal{S}$ being symmetric.

For the last part we have

$$\mathcal{S} \oplus -\mathcal{S} = S \oplus S = 2\mathcal{S}.$$

The last equality follows from convexity: for any $s_1, s_2 \in \mathcal{S}$ it holds that $(s_1 + s_2)/2 \in \mathcal{S}$ and hence $s_1 + s_2 \in 2\mathcal{S}$. $\square$

Aux. Lemma 8. *For a $\mathcal{S} \subset \mathbb{R}^d$ its polar set of radius r is the intersection of $|\mathcal{S}|$ half-spaces:*

$$(\mathcal{S})^r = \bigcap_{s \in \mathcal{S}} \left\{ x \in \mathbb{R}^d : \frac{1}{r} s^\top x \leq 1 \right\}.$$

Theorem 2. *Let $f : \mathbb{R}^d \to \mathbb{R}^K$ be a classifier such that $h_{i\text{-}j} = f_i - f_j$ is $\mathcal{S}_{i\text{-}j}$-Lipschitz, $\forall i, j \in 1, \ldots, K, i \neq j$. Then, given an input $x \in \mathbb{R}^d$, f is robust at x against all δ in $Q = \bigcap_{i \neq c_A} (\mathcal{S}_{i\text{-}c_A})^{r_i}$.*

Proof of Theorem 2. For a fixed class $i \neq c_A$ we have that the following must hold from Definition 2:

$$\begin{aligned} h_{i\text{-}c_A}(y) - h_{i\text{-}c_A}(x) &\leq \rho_{\mathcal{S}_{i\text{-}j}}(y - x) \\ f_i(y) - f_{c_A}(y) - f_i(x) + f_{c_A}(x) &\leq \rho_{\mathcal{S}_{i\text{-}j}}(y - x). \end{aligned}$$

Rearranging the terms gives:

$$f_i(y) - f_{c_A}(y) \leq \rho_{\mathcal{S}_{i\text{-}j}}(y - x) + \underbrace{f_i(x) - f_{c_A}(x)}_{-r_i}.$$

We are interested in the values of y for which the left-hand side is nonpositive as these are inputs for which the confidence is higher for c_A than for i. Hence, we restrict the right-hand side to be upper-bounded by zero:

$$\rho_{\mathcal{S}_{i\text{-}j}}(\underbrace{y - x}_{\delta}) \leq r_i.$$

The values of δ satisfying this inequality are exactly the polar set $(\mathcal{S}_{i\text{-}j})^{r_i}$ (Definition 3).

Finally, as we need that the confidence for class c_A is larger than the confidences for any other class, we need to take the intersection over $i \neq c_A$ resulting in the certificate $Q = \bigcap_{i \neq c_A} (\mathcal{S}_{i\text{-}c_A})^{r_i}$. □

Aux. Lemma 9 (Scaling of $\mathcal{S}$-Lipschitz Classifiers)**.** *Consider a constant $\alpha > 0$ and a classifier $f : \mathbb{R}^d \to \mathbb{R}^K$ such that $h_{i\text{-}j} = f_i - f_j$ is $\mathcal{S}_{i\text{-}j}$-Lipschitz for all $i \neq j$. Then $\alpha h_{i\text{-}j} = \alpha f_i - \alpha f_j$ is $\alpha\mathcal{S}_{i\text{-}j}$-Lipschitz but f and αf have the same certificates:*

$$Q_f = \bigcap_{i \neq c_A^g} (\mathcal{S}_{i\text{-}c_A^g})^{r_i} = \bigcap_{i \neq c_A^g} (\alpha\mathcal{S}_{i\text{-}c_A^g})^{\alpha r_i} = Q_{\alpha f}. \quad (12)$$

Proof of Aux. Lemma 9. Note that scaling with a positive constant α does not change the top class c_A^g and also scales the prediction gaps proportionally: $\alpha f_{c_A^g}(x) - \alpha f_i(x) = \alpha r_i$.

Next, let's show that αf_i is $\alpha\mathcal{S}_i$-Lipschitz. If $g : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{S}$-Lipschitz, then we have

$$-\sup_{c \in \mathcal{S}} c^\top (x - y) \leq g(y) - g(x) \leq \sup_{c \in \mathcal{S}} c^\top (y - x)$$

for all $x, y \in \mathbb{R}^d$. As $\alpha > 0$, multiplying everything by α results in:

$$\begin{aligned} -\sup_{c \in \mathcal{S}} \alpha c^\top (x - y) \leq \alpha g(y) - \alpha g(x) &\leq \sup_{c \in \mathcal{S}} \alpha c^\top (y - x) \\ -\sup_{c \in \alpha\mathcal{S}} c^\top (x - y) \leq \alpha g(y) - \alpha g(x) &\leq \sup_{c \in \alpha\mathcal{S}} c^\top (y - x), \end{aligned}$$

which is the condition for αg being $(\alpha\mathcal{S})$-Lipschitz.

Now, we can show that scaling the $\mathcal{S}$ set and the polar set radius with the same constant does not change the polar set:

$$\begin{aligned} (\alpha\mathcal{S})^{\alpha r} &= \left\{ y \in \mathbb{R}^d \ : \ \sup_{x \in \alpha\mathcal{S}} x^\top y \leq \alpha r \right\} \\ &= \left\{ y \in \mathbb{R}^d \ : \ \sup_{x \in \mathcal{S}} \alpha x^\top y \leq \alpha r \right\} \\ &= \left\{ y \in \mathbb{R}^d \ : \ \sup_{x \in \mathcal{S}} x^\top y \leq r \right\} \\ &= (\mathcal{S})^r. \end{aligned}$$

Equation (12) directly follows. □

Theorem 3 (Addition of $\mathcal{S}$-Lipschitz classifiers)**.** *Take an ensemble as in Equation (5) with $N = 2$ and the* CD *setting, i.e., $h_{i\text{-}k}^j = f_i^j - f_k^j$ is $\mathcal{S}_{i\text{-}k}^j$-Lipschitz. Then, at a fixed $x \in \mathbb{R}^d$, it holds that g is robust against all δ in*

$$Q_g = \bigcap\nolimits_{i \neq c_A^g} \left(\alpha_1 \mathcal{S}_{i\text{-}c_A^g}^1 \oplus \alpha_2 \mathcal{S}_{i\text{-}c_A^g}^2 \right)^{r_i^g},$$

with $c_A^g = \arg\max_i g_i$ and $r_i^g = g_{c_A^g} - g_i$. The case for $N > 2$ follows by induction.

Proof of Theorem 3. From Aux. Lemma 9 we know that $\alpha_1 h_1^{i\text{-}j}$ and $\alpha_2 h_2^{i\text{-}j}$ are $\alpha_1 \mathcal{S}_{i\text{-}j}^1$- and $\alpha_2 \mathcal{S}_{i\text{-}j}^2$-Lipschitz. Then, following Aux. Lemma 4, we have that $\nabla \alpha_1 h_1^{i\text{-}j} \in \operatorname{hull} \alpha_1 \mathcal{S}_{i\text{-}j}^1$ and $\nabla \alpha_2 h_2^{i\text{-}j} \in \operatorname{hull} \alpha_2 \mathcal{S}_{i\text{-}j}^2$. Since $\nabla h_{i\text{-}j}^g = \nabla (g_i - g_j) = \nabla \alpha_1 h_1^{i\text{-}j} + \nabla \alpha_2 h_2^{i\text{-}j}$, we have $\nabla h_{i\text{-}j}^g \in \alpha_1 \operatorname{hull} \mathcal{S}_{i\text{-}j}^1 \oplus \alpha_2 \operatorname{hull} \mathcal{S}_{i\text{-}j}^2 = \operatorname{hull}(\alpha_1 \mathcal{S}_{i\text{-}j}^1 \oplus \alpha_2 \mathcal{S}_{i\text{-}j}^2)$ as constructing the convex hull and taking the Minkowski sum commute. By Theorem 1, $h_{i\text{-}j}^g$ is $(\operatorname{hull}(\alpha_1 \mathcal{S}_{i\text{-}j}^1 \oplus \alpha_2 \mathcal{S}_{i\text{-}j}^2))$-Lipschitz which by Aux. Lemma 3 is the same as being $(\alpha_1 \mathcal{S}_{i\text{-}j}^1 \oplus \alpha_2 \mathcal{S}_{i\text{-}j}^2)$-Lipschitz. The rest follows from Theorem 2. □

Theorem 4. *Consider an ensemble of* U *classifiers and a fixed x for which $c_A^=$ and $c_B^=$ hold. Then, for any choice of weights α_j in Equation (5), the $\mathcal{S}$-certificate of the ensemble satisfies* ❷.

Proof of Theorem 4. We will prove only the case for $N = 2$. $N > 2$ follows by induction. Furthermore, we assume $\alpha_j \geq 0, \forall j$ as in Equation (5).

$$Q_g = \left\{ y \in \mathbb{R}^n : \sup_{\substack{x_1, x_2 \in \mathcal{S}^1 \\ x_3, x_4 \in \mathcal{S}^2}} \left\{ \alpha_1 (x_1 - x_2)^\top y + \alpha_2 (x_3 - x_4)^\top y \right\} \leq \alpha_1 r^1 + \alpha_2 r^2 \right\} \tag{13}$$

We will denote the individual classifier gaps and the ensemble gap as $r^1 = f^1_{c_A}(x) - f^1_{c_B}(x)$, $r^2 = f^2_{c_A}(x) - f^2_{c_B}(x)$, $r^g = g_{c_A}(x) - g_{c_B}(x)$. First, from Theorem 1 we have

$$\begin{aligned} Q_1 &= (\mathcal{S}^1 \oplus -\mathcal{S}^1)^{r^1}, \\ Q_2 &= (\mathcal{S}^2 \oplus -\mathcal{S}^2)^{r^2}, \\ Q_g &= \left((\alpha_1 \mathcal{S}^1 \oplus \alpha_2 \mathcal{S}^2) \oplus -(\alpha_1 \mathcal{S}^1 \oplus \alpha_2 \mathcal{S}^2) \right)^{\alpha_1 r^1 + \alpha_2 r^2} \\ &= (\alpha_1 (\mathcal{S}^1 \oplus -\mathcal{S}^1) \oplus \alpha_2 (\mathcal{S}^2 \oplus -\mathcal{S}^2))^{\alpha_1 r^1 + \alpha_2 r^2}. \end{aligned}$$

Q_g can also be expanded as Equation (13). Consider the two inequalities that define Q_1 and Q_2:

$$\sup_{x_1, x_2 \in \mathcal{S}^1} (x_1 - x_2)^\top y \leq r^1, \sup_{x_3, x_4 \in \mathcal{S}^2} (x_3 - x_4)^\top y \leq r^2.$$

If for a given y, both of these hold, then the inequality in Equation (13) also must hold. Hence, the intersection of Q_1 and Q_2 must be a subset of Q_g. Similarly, it is necessary for at least one of them to hold, hence every element of Q_g must be an element of the union of Q_1 and Q_2. □

Aux. Lemma 10. *Let $S \subseteq \mathbb{R}^d$ be a convex set, $\alpha, \beta \geq 0$, and $a, b \in \mathbb{R}^d$. Then it holds that*

$$(\alpha S + a) \oplus (\beta S + b) = (\alpha + \beta) S + (a + b).$$

A special case is the summing of ℓ_p norm balls ($p \geq 1$):

$$\mathcal{B}_p[\mu_1, \epsilon_1] \oplus \mathcal{B}_p[\mu_1, \epsilon_1] = \mathcal{B}_p[\mu_1 + \mu_2, \epsilon_1 + \epsilon_2].$$

Proof of Aux. Lemma 10. It is trivial to see that

$$(\alpha S + a) \oplus (\beta S + b) = \alpha S \oplus \beta S + (a + b).$$

Hence, we only need to show if $\alpha S \oplus \beta S \stackrel{?}{=} (\alpha + \beta) S$ which is the same as:

$$S_L = \{\alpha x + \beta y : x, y \in S\} \stackrel{?}{=} \{(\alpha + \beta) x' : x' \in S\} = S_R.$$

It is obvious that $S_R \subseteq S_L$. So we only need to show that $S_L \subseteq S_R$. Take a $z \in S_L$. Then there must be $x, y \in S$ such that $\alpha x + \beta y = z$. Now, take $x' = z/(\alpha+\beta)$:

$$x' = \frac{z}{\alpha + \beta} = \frac{\alpha x + \beta y}{\alpha + \beta} = \frac{\alpha}{\alpha + \beta} x + \frac{\beta}{\alpha + \beta} y.$$

x' is a linear combination of elements of the convex S, hence x' is also in S. Therefore, for every $z \in S_L$ we can construct an $x' \in S$ such that $(\alpha + \beta) x' = z \in S_R$. This concludes our proof that $S_L = S_R$.

The ℓ_p norm ball special case follows directly when we note that $\mathcal{B}_p[\mu, \epsilon]$ can be represented as

$$\epsilon \cdot \{x \in \mathbb{R}^d : \|x\|_p \leq 1\} + \mu.$$

□

Aux. Lemma 11. *Take to be $\mathcal{B}_\star \subset \mathbb{R}^d$ a closed convex symmetric set. Define $\mathcal{B}$ to be the norm ball of its dual norm, i.e.:*

$$\mathcal{B} = \left\{ y \in \mathbb{R}^d : \sup_{x \in \mathcal{B}_\star} x^\top y \leq 1 \right\}.$$

Then, the polar set of $\epsilon \mathcal{B}$ with radius r is:

$$(\epsilon \mathcal{B}_\star)^r = \frac{r}{\epsilon} \mathcal{B}.$$

Proposition 3. *Take two classifiers $f^1, f^2 : \mathbb{R}^d \to \mathbb{R}^K$ satisfying $c_A^=$. Further, assume that all $h^j_{i\text{-}k} = f^j_i - f^j_k$ are $\epsilon_{j,i\text{-}k} \mathcal{B}_\star$-Lipschitz for some closed convex symmetric set $\mathcal{B}_\star$. Then, the maximum improvement in the certified radius R^g of g relative to the larger one of R^1 and R^2 is*

$$R^g - \max\{R^1, R^2\} \leq \frac{1}{\min\{M^1, M^2\}} - \frac{\min\{r^1_{c^1_B}, r^2_{c^2_B}\}}{\min\{M^1, M^2\} + \Delta},$$

where we have defined M^k as $\min_{i \neq c_A} \epsilon_{k,i\text{-}c_A}$ and Δ as $\max_{k=1,2} \max_{i \neq c_A} (\epsilon_{k,i\text{-}c_A} - M^k)$.

Proof of Proposition 3. We assume that the predictions of each classifier are normalized, *i.e.*, $\sum_i f^j_i = 1, \ \forall j = 1, \ldots, N$. Because all difference smoothness sets $\mathcal{S}$ have the same shape $\mathcal{B}_\star$ and the shape is closed under scaling and Minkowski sum (Aux. Lemma 10), we can simply work with a certified radius rather than a certified set. Note, however, that the shape of the certified sets would be the dual of the shape of the smoothness, *i.e.*, $\mathcal{B}$. Therefore, from Aux.

Lemma 11 we have:

$$
\begin{aligned}
Q_j &= \bigcap_{i \neq c_A} (\epsilon_{j,i\text{-}c_A} \mathcal{B}_\star)^{r_i^j} \\
&= \min_{i \neq c_A} \left\{ \frac{r_i^j}{\epsilon_{j,i\text{-}c_A}} \right\} \mathcal{B} \qquad \text{for } j = 1, 2 \qquad (14)\\
&= \min_{i \neq c_A} \left\{ R_i^j \right\} \mathcal{B}, \\
&= R^j \\
Q_\cup &= Q_1 \cup Q_2 \\
&= \max\{\min_{i \neq c_A} R_i^1, \min_{i \neq c_A} R_i^2\} \mathcal{B} \qquad (15)\\
&= \max\{R^1, R^2\} \mathcal{B} \\
&= R^\cup \mathcal{B}_\star \\
Q_\cap &= Q_1 \cap Q_2 \\
&= \min\{\min_{i \neq c_A} R_i^1, \min_{i \neq c_A} R_i^2\} \mathcal{B} \qquad (16)\\
&= \min\{R^1, R^2\} \mathcal{B} \\
&= R^\cap \mathcal{B}
\end{aligned}
$$

Next, note that as we have the same top predictions c_A, according to Aux. Lemma 12 it holds that $r_i^g = \alpha_1 r_i^1 + \alpha_2 r_i^2$. Therefore, from Aux. Lemmas 10 and 11 we have

$$
\begin{aligned}
Q_g &= \bigcap_{i \neq c_A} (\alpha_1 \epsilon_{1,i\text{-}c_A} \mathcal{B}_\star \oplus \alpha_2 \epsilon_{2,i\text{-}c_A} \mathcal{B}_\star)^{\alpha_1 r_i^1 + \alpha_2 r_i^2} \\
&= \bigcap_{i \neq c_A} ((\alpha_1 \epsilon_{1,i\text{-}c_A} + \alpha_2 \epsilon_{2,i\text{-}c_A}) \mathcal{B}_\star)^{\alpha_1 r_i^1 + \alpha_2 r_i^2} \\
&= \bigcap_{i \neq c_A} \frac{\alpha_1 r_i^1 + \alpha_2 r_i^2}{\alpha_1 \epsilon_{1,i\text{-}c_A} + \alpha_2 \epsilon_{2,i\text{-}c_A}} \mathcal{B} \qquad (17)\\
&= \bigcap_{i \neq c_A} R_i^g \mathcal{B} \\
&= \min_{i \neq c_A} \{R_i^g\} \mathcal{B} \\
&= R^g \mathcal{B}.
\end{aligned}
$$

From Equation (17) it follows that the absolute values of α_1 and α_2 do not matter, only their relative size. This follows from the fact that R_i^g is unchanged if we normalize α_1 and α_2 by diving by $\alpha_1 + \alpha_2$. Therefore, with no loss of generality we will assume that $\alpha_1 = \alpha$ and $\alpha_2 = 1 - \alpha$ for the rest of the proof. This also follows from Aux. Lemma 9.

While the claims of this proposition can be proven algebraically, we opt for a more intuitive graphical approach. First, notice that R_i^g is a linear-fractional function in α and is monotonically increasing or decreasing from R_i^2 (for $\alpha = 0$) to R_i^1 (for $\alpha = 1$). Therefore, we can plot the R_i^j and R_i^g (as functions of α) as in Figure D.1.

From Equation (15) we know that $R^\cup$ equals the larger one between the smallest radius on the left and the smallest

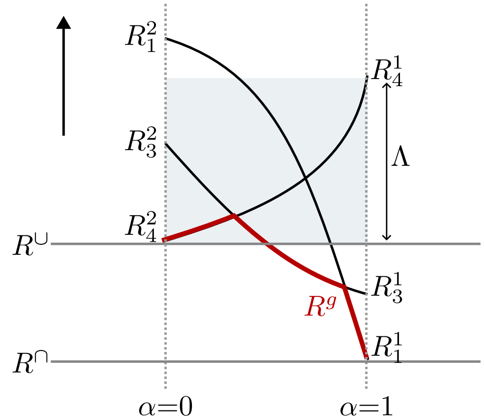

Figure D.1. Illustration of the certified radii in Proposition 3. The specific parameters used are $K = 4, c_A = 2$

radius on the right. Similarly, from Equation (16) we have that $R^\cap$ is the smallest radius on either side. Finally, R^g is the minimum of R_i^g across $\alpha \in [0, 1]$, or the thick red line in Figure D.1.

Note the peak of R^g cannot be larger than the smaller of the largest R_i^1 or the largest R_i^2 due to the monotonicity of R_i^g. Recall also that $\max\{R^1, R^2\} = R^\cup$ from Equation (15). Hence, $R^g - \max\{R^1, R^2\}$ is upper bounded by Λ, the height of the shaded region in Figure D.1:

$$
\Lambda = \min\{\max_{i \neq c_A} R_i^1, \max_{i \neq c_A} R_i^2\} - \max\{\min_{i \neq c_A} R_i^1, \min_{i \neq c_A} R_i^2\} \qquad (18)
$$

As $r_{c_B^j}^j$ is the smallest gap for f_j and as no gap can be larger than 1 we have:

$$
\frac{r_{c_B^j}^j}{\epsilon_{j,i\text{-}c_A}} \leq R_i^j = \frac{r_i^j}{\epsilon_{j,i\text{-}c_A}} \leq \frac{1}{\epsilon_{j,i\text{-}c_A}}. \qquad (19)
$$

Using this we can upper-bound the first term in Equation (18):

$$
\begin{aligned}
&\min\left\{\max_{i \neq c_A} R_i^1, \max_{i \neq c_A} R_i^2\right\} \\
&\leq \min\left\{\max_{i \neq c_A} \frac{1}{\epsilon_{1,i\text{-}c_A}}, \max_{i \neq c_A} \frac{1}{\epsilon_{2,i\text{-}c_A}}\right\} \text{ (from Eq. 19)} \qquad (20)\\
&= \min\left\{\frac{1}{M^1}, \frac{1}{M^2}\right\} \\
&= \frac{1}{\max\{M^1, M^2\}}.
\end{aligned}
$$

Using Equation (19) we can also lower-bound the second term in Equation (18):

$$\begin{aligned}
&\max\left\{\min_{i\neq c_A} R_i^1, \min_{i\neq c_A} R_i^2\right\}\\
\geq&\max\left\{\min_{i\neq c_A}\frac{r^1_{c_B^1}}{\epsilon_{1,i\text{-}c_A}}, \min_{i\neq c_A}\frac{r^2_{c_B^2}}{\epsilon_{2,i\text{-}c_A}}\right\}\\
=&\max\left\{\frac{r^1_{c_B^1}}{\max_{i\neq c_A}\epsilon_{1,i\text{-}c_A}}, \frac{r^2_{c_B^2}}{\max_{i\neq c_A}\epsilon_{2,i\text{-}c_A}}\right\}\\
\geq&\max\left\{\frac{\min\{r^1_{c_B^1}, r^2_{c_B^2}\}}{\max_{i\neq c_A}\epsilon_{1,i\text{-}c_A}}, \frac{\min\{r^1_{c_B^1}, r^2_{c_B^2}\}}{\max_{i\neq c_A}\epsilon_{2,i\text{-}c_A}}\right\}\\
=&\frac{\min\{r^1_{c_B^1}, r^2_{c_B^2}\}}{\min\{\max_{i\neq c_A}\epsilon_{1,i\text{-}c_A}, \max_{i\neq c_A}\epsilon_{2,i\text{-}c_A}\}}\\
\geq&\frac{\min\{r^1_{c_B^1}, r^2_{c_B^2}\}}{\max\{M^1, M^2\}+\Delta}.
\end{aligned} \tag{21}$$

Finally, substituting Equations (20) and (21) into Equation (18), we get the upper bound for $R^g - \max\{R^1, R^2\}$:

$$\begin{aligned}
&R^g - \max\{R^1, R^2\}\\
\leq&\Lambda\\
\leq&\frac{1}{\max\{M^1, M^2\}} - \frac{\min\{r^1_{c_B^1}, r^2_{c_B^2}\}}{\max\{M^1, M^2\}+\Delta}.
\end{aligned}$$

□

Proposition 4. *For any set of $N \geq 2$ classifiers satisfying $c_A^{\neq}$, there exist weights α_j for which the resulting ensemble has $r^g_{c_B} = 0$ and a certified perturbation set $Q_g = \{0\}$.*

Proof of Proposition 4. First, note that $Q_{g(\alpha)} = \{0\}$ if $r^g_{c_B} = 0$, that is if the top two classes of g have the same confidence. In other words, if x is on the decision boundary for the ensemble g. Therefore, we want to show that it is possible to construct an ensemble for which the decision boundary passes through x.

Let's first consider the case with two classifiers ($N = 2$) and when the top prediction of g is one of the top predictions of the individual classifiers for any α: $c_A^g \in \{c_A^1, c_A^2\}, \forall\alpha \in [0, 1]$. We denote by $g(\alpha)$ the ensemble $g(\alpha) = \alpha f_1 + (1-\alpha)f_2$. Therefore, we have $c_A^{g(\alpha)} = c_A^2$ when α is close to 0 and $c_A^{g(\alpha)} = c_A^1$ when α is close to 1. The switch between the two values happens at

$$\alpha^\star = \frac{f^2_{c_A^2} - f^2_{c_A^1}}{f^1_{c_A^1} - f^1_{c_A^2} + f^2_{c_A^2} - f^2_{c_A^1}}, \tag{22}$$

if the denominator is not 0. It follows that when $\alpha = \alpha^\star$ we have that $g_{c_A^g} = g_{c_B^g}$, hence $r^g_{c_B} = 0$ and $Q_{g(\alpha)} = \{0\}$. Note that if the denominator in Equation (22) is 0, then $Q_{g(\alpha)} = \{0\}$ for all α.

Now consider the case when for some α the top prediction of g is not one of the top predictions of the individual classifiers. Then we can split the domain $[0, 1]$ for α into a subset that has only two top predictions and apply the above analysis to this subset. Therefore, when $N = 2$ the proposition holds.

To see that it holds for $N > 2$, note that we can always fix $N - 2$ of the α_j weights to 0. As long as we select two individual classifiers with different top predictions to have non-zero weights, we can apply the $N = 2$ analysis to them. Therefore, the proposition holds for all N. □

Aux. Lemma 12. *For any ensemble of N normalized K-class classifiers satisfying $c_A^{=}$ it holds that the i-th class prediction gap r_i^g of the ensemble is the weighted sum of the gaps r_i^j of the individual classifiers:*

$$r_i^g = \sum_{j=1}^N \alpha_j r_i^j,\ \forall i = 1, \ldots, K.$$

Proof of Aux. Lemma 12.

$$\begin{aligned}
r_i^g &= \sum_{j=1}^N \alpha_j f_j^{c_A^g} - \sum_{j=1}^N \alpha_j f_i^j\\
&= \sum_{j=1}^N \alpha_j f^j_{c_A^j} - \sum_{j=1}^N \alpha_j f_i^j\\
&= \sum_{j=1}^N \alpha_j \left(f^j_{c_A^j} - f_i^j\right)\\
&= \sum_{j=1}^N \alpha_j r_i^j.
\end{aligned}$$

□

Proposition 5. *No ensemble of classifiers as in Theorem 3 satisfying $c_A^{=}$ can be in regime ❸.*

Proof of Proposition 5. We will deal only with the $N = 2$ case as $N \geq 2$ follows by induction. Furthermore, we will assume that $\alpha_2 = 1 - \alpha_1$ for simplicity. This doesn't affect the proof as the $\alpha_1 + \alpha_2$ scaling does not affect the certificate (Aux. Lemma 9).

We prove by contradiction. We restrict to same c_A, and assume that we have α_1, classifier outputs and $\mathcal{S}$ sets such that:

$$Q_g \subset Q_1 \cap Q_2, \tag{23}$$

where

$$Q_1 = \bigcap_{i \neq c_A} \left(\mathcal{S}^1_{i\text{-}c_A}\right)^{r_i^1}$$
$$Q_2 = \bigcap_{i \neq c_A} \left(\mathcal{S}^2_{i\text{-}c_A}\right)^{r_i^2}$$
$$Q_g = \bigcap_{i \neq c_A} \left(\alpha_1 \mathcal{S}^1_{i\text{-}c_A} \oplus (1-\alpha_1)\mathcal{S}^2_{i\text{-}c_A}\right)^{r_i^g}.$$

Hence, Equation (23) becomes

$$\bigcap_{i \neq c_A} \left(\alpha_1 \mathcal{S}^1_{i\text{-}c_A} \oplus (1\text{-}\alpha_1)\mathcal{S}^2_{i\text{-}c_A}\right)^{r_i^g} \subset \bigcap_{j=1,2} \bigcap_{i \neq c_A} \left(\mathcal{S}^j_{i\text{-}c_A}\right)^{r_i^j}. \tag{24}$$

This implies that there must be a point x in the right-hand side of Equation (24) that is not in the left-hand side. This x must satisfy:

$$\sup_{t \in \mathcal{S}^j_{i\text{-}c_A}} t^\top x \leq r_i^j \text{ for all } j = 1, 2,\ i \neq c_A.$$

For the left-hand side of Equation (24), using $c_A^=$ and Aux. Lemma 12 we have:

$$\begin{aligned} &\left(\alpha_1 \mathcal{S}^1_{i\text{-}c_A} \oplus (1-\alpha_1)\mathcal{S}^2_{i\text{-}c_A}\right)^{r_i^g} \\ = &\left(\alpha_1 \mathcal{S}^1_{i\text{-}c_A} \oplus (1-\alpha_1)\mathcal{S}^2_{i\text{-}c_A}\right)^{\alpha_1 r_i^1 + (1-\alpha_1) r_i^2}. \end{aligned}$$

We can see that x must be in this polar set:

$$\begin{aligned} &\sup_{\substack{t_1 \in \mathcal{S}^1_{i\text{-}c_A} \\ t_2 \in \mathcal{S}^2_{i\text{-}c_A}}} \left(\alpha_1 t_1^\top x + (1-\alpha_1) t_2^\top x\right) \\ = &\ \alpha_1 \sup_{t_1 \in \mathcal{S}^1_{i\text{-}c_A}} t_1^\top x + (1-\alpha_1) \sup_{t_2 \in \mathcal{S}^2_{i\text{-}c_A}} t_2^\top x \\ \leq &\ \alpha_1 r_i^1 + (1-\alpha_1) r_i^2. \end{aligned}$$

As this holds for all $i \neq c_A$, x must also be in the intersection and hence in Q_g. This is a contradiction of the assumption that x is not in Q_g. □

Proposition 6. *Take an ensemble as in Proposition 3. Assume two different second top predictions and that classes that are not in the top two predictions of any individual classifier have low confidences*[5]*. Then* ❶ *occurs when:*

$$f^1_{c_A} > f^1_{c_B^2} + r^2_{c_B^2} \frac{\epsilon_{1,c_B^2\text{-}c_A}}{\epsilon_{2,c_B^2\text{-}c_A}} \ \textit{ and } \ f^2_{c_A} > f^2_{c_B^1} + r^1_{c_B^1} \frac{\epsilon_{2,c_B^1\text{-}c_A}}{\epsilon_{1,c_B^1\text{-}c_A}}.$$

Proof of Proposition 6. We restrict ourselves to the $c_A^=$ and different c_B setting as this is the regime that prevents ❸ and allows for ❶ (Theorem 4 and Proposition 5). We ask the classes that are not in the top two predictions of any individual classifier to have low predictions in order for them to not compete for the top ensemble prediction. Formally:

$$f^i_c < f^j_{c_B^k}, \forall j, k \in \{1, 2\}, \forall c \notin \{c_A\} \cup \{c_B^l : l = 1, 2\}.$$

From the proof of Proposition 3 and our small third predictions assumptions we have:

$$R^\cup = \max\left\{\frac{r^1_{c_B^1}}{\epsilon_{1,c_B^1\text{-}c_A}}, \frac{r^2_{c_B^2}}{\epsilon_{2,c_B^2\text{-}c_A}}\right\},$$

and

$$\begin{aligned} &R^g \\ &= \min_{i \neq c_A}\left\{\frac{\alpha_1 r_i^1 + \alpha_2 r_i^2}{\alpha_1 \epsilon_{1,i\text{-}c_A} + \alpha_2 \epsilon_{2,i\text{-}c_A}}\right\} \\ &= \min\left\{\frac{\alpha_1 r^1_{c_B^1} + \alpha_2 r^2_{c_B^1}}{\alpha_1 \epsilon_{1,c_B^1\text{-}c_A} + \alpha_2 \epsilon_{2,c_B^1\text{-}c_A}}, \frac{\alpha_1 r^1_{c_B^2} + \alpha_2 r^2_{c_B^2}}{\alpha_1 \epsilon_{1,c_B^2\text{-}c_A} + \alpha_2 \epsilon_{2,c_B^2\text{-}c_A}}\right\}. \end{aligned}$$

Both terms in the above minimum are monotonic. Therefore, the only way that $R^g > R^\cup$ for some α_1, α_2 is that the first term is decreasing while the second is increasing. This happens when

$$\frac{r^1_{c_B^2}}{\epsilon_{1,c_B^2 - c_A}} > \frac{r^2_{c_B^2}}{\epsilon_{2,c_B^2 - c_A}}$$
$$\frac{r^2_{c_B^1}}{\epsilon_{2,c_B^1 - c_A}} > \frac{r^1_{c_B^1}}{\epsilon_{1,c_B^1 - c_A}},$$

which, when rearranged, results in the conditions in the proposition. □