Towards Optimal Randomized Smoothing: A Semi-Infinite Linear Programming Approach

Brendon G. Anderson [1] **Samuel Pfrommer** [1] **Somayeh Sojoudi** [1]

Abstract

Randomized smoothing is a leading approach to producing certifiably robust classifiers. The goal of optimal randomized smoothing is to maximize the average certified radius over the space of smoothing distributions. We theoretically study this problem through the lens of infinite-dimensional optimization over measure spaces, and prove that the nonconvex infinite program is lower-bounded by a conic linear program wherein the classifier's confidence acts as a surrogate objective to optimize. A semi-infinite linear programming approximation to the problem is presented, whose sub-problems are proven to attain nontrivial strong duality. A proof-of-concept experiment demonstrates the effectiveness of the proposed approach.

1. Introduction

Neural networks have proliferated across a range of safety-critical applications, from autonomous driving (Bojarski et al., 2016; Wu et al., 2017) to power systems control (Kong et al., 2017) and medical diagnostics (Kang et al., 2021). It is therefore especially concerning that these models are vulnerable to *adversarial inputs*: human-imperceptible perturbations that can cause failures such as misclassification (Biggio et al., 2013; Szegedy et al., 2014). Heuristic defenses against such inputs have often been subsequently defeated by stronger attacks (Carlini & Wagner, 2017; Kurakin et al., 2017; Athalye et al., 2018), motivating classifiers with provable robustness properties.

Randomized smoothing, popularized by Lecuyer et al. (2019); Li et al. (2019); Cohen et al. (2019), remains one of the state-of-the-art methods to provably robustify classifiers. The idea of randomized smoothing is as follows: instead of directly classifying an input at test time, random perturbations of the input are classified, and the most probable class amongst the perturbed inputs is the class assigned to the input. Intuitively, this procedure acts to average out any problematic but unlikely perturbations near the input. This is illustrated in Figure 1, where the classifier decision boundaries are smoothed out, resulting in a larger robust region around the test input.

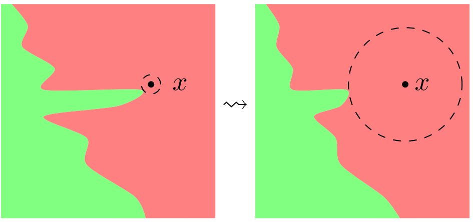

Figure 1. Randomized smoothing removes jagged regions where adversarial inputs reside and provably robustifies classifiers.

A line of work has been developed to prove certified safe radii for various input noise distributions. For example, Cohen et al. (2019) gives ℓ_2-norm safe radii using Gaussian smoothing and Teng et al. (2020) gives ℓ_1-norm safe radii using Laplacian smoothing. Norm ball robustness is further generalized by Yang et al. (2020a) under a range of smoothing distributions. Most works have only considered smoothing distributions that are fixed and unbiased, meaning the distribution of random noise is independent of the test input and has zero mean. These restrictions lead to conservative robustness guarantees, since a good distribution for an input far from the decision boundary is one with high variance (so as to maximize the certified radius), whereas a good distribution for an input close to the decision boundary is one with low variance and a potential bias (so as to maintain accuracy).

Recent works, such as Alfarra et al. (2020); Wang et al. (2021); Eiras et al. (2021); Chen et al. (2021), attempt to develop input-dependent smoothing schemes, but most restrict themselves to Gaussian noise, and their robustness certificates often do not hold for the actual classifier used at test time (Anderson & Sojoudi, 2022). In this work, we consider generalizing randomized smoothing to allow for arbitrary input-dependent smoothing schemes, and we theoretically

[1] University of California, Berkeley, USA. Correspondence to: Brendon G. Anderson <bganderson@berkeley.edu>.

1st Workshop on Formal Verification of Machine Learning, Baltimore, Maryland, USA. Colocated with ICML 2022. Copyright 2022 by the author(s).

study the problem of finding such a scheme that optimizes the average certified radius over the data distribution.

1.1. Preliminaries

Throughout, we allow $\|\cdot\|$ to denote an arbitrary norm. We use the notation $\|\cdot\|_\infty$ to represent both the element-wise maximum norm for finite-dimensional vectors, as well as the supremum norm for real-valued bounded functions. The domain of a norm will always be clear from context. The interior of a set X is denoted by $\operatorname{int} X$. The set of nonnegative real numbers is denoted by $\mathbb{R}_+$. The vector in $\mathbb{R}^d$ with all elements equal to 1 is written $\mathbf{1}_d$. For a Lipschitz continuous map φ between metric spaces, we denote its Lipschitz constant by $\operatorname{Lip}(\varphi)$. Given a function h on $\mathbb{R}^d$ and a point $y \in \mathbb{R}^d$, we define the function $\tau_y h$ on $\mathbb{R}^d$ by $\tau_y h(x) = h(x+y)$.

Let X be a subset of $\mathbb{R}^d$. Define $\mathcal{B}(X)$ to be the Borel σ-algebra on X and $(\mathcal{M}(X), \|\cdot\|_{\text{TV}})$ to be the Banach space of finite signed measures on the measurable space $(X, \mathcal{B}(X))$ equipped with the total variation norm. The Lebesgue measure of $B \in \mathcal{B}(X)$ is denoted $m(B)$. Define $\mathcal{M}(X)_+ = \{\mu \in \mathcal{M}(X) : \mu \geq 0\}$ to be the convex cone of finite positive measures and $\mathcal{P}(X) = \{\mu \in \mathcal{M}(X)_+ : \mu(X) = 1\}$ to be the set of probability measures. If X is compact, then denote by $(\mathcal{C}(X), \|\cdot\|_\infty)$ the Banach space of real-valued continuous functions on X equipped with the supremum norm. Define $\mathcal{C}(X)_+ = \{h \in \mathcal{C}(X) : h \geq 0\}$ to be the convex cone of nonnegative continuous functions on X. Recall the bilinear form $\langle \cdot, \cdot \rangle : \mathcal{C}(X) \times \mathcal{M}(X) \to \mathbb{R}$ given by $\langle h, \mu \rangle = \int_X h(x) d\mu(x)$ for all $(h, \mu) \in \mathcal{C}(X) \times \mathcal{M}(X)$. When h is a continuous function defined on all of $\mathbb{R}^d$, we use the notation $\langle h, \mu \rangle$ to mean $\langle h|_X, \mu \rangle$, i.e., the integration is restricted to X. For a vector-valued $h \colon \mathbb{R}^d \to \mathbb{R}^n$, we use $\langle h, \mu \rangle$ to denote component-wise integration; $\langle h, \mu \rangle = (\langle h_1, \mu \rangle, \langle h_2, \mu \rangle, \dots, \langle h_n, \mu \rangle)$. If μ is a probability measure, we may sometimes write its bilinear evaluation on h as an expectation: $E_{\epsilon \sim \mu} h(\epsilon) = \langle h, \mu \rangle$. The set of all real-valued continuously differentiable functions on $\mathbb{R}^d$ is denoted by $\mathcal{C}^1(\mathbb{R}^d)$. We denote the normal distribution on $\mathbb{R}^d$ with mean $\overline{x}$ and covariance Σ by $N(\overline{x}, \Sigma)$. The distribution function of $N(0,1)$ on $\mathbb{R}$ is denoted by Φ, which we recall has a well-defined inverse.

Proofs are deferred to the appendices.

2. Problem Formulation

2.1. Conventional Smoothing

Let $d, n \in \mathbb{N}$ and consider the n-class classifier $f \colon \mathbb{R}^d \to \mathcal{Y}$, where $\mathcal{Y} = \{1, 2, \dots, n\}$, defined by $f(x) \in \arg\max_{i \in \mathcal{Y}} g_i(x)$ with associated soft classifier $g \colon \mathbb{R}^d \to [0,1]^n$. Conventional randomized smoothing performs a Gaussian averaging operation on g that smooths the decision boundaries of f so as to eradicate adversarial inputs which exist near "sharp" regions of the decision boundary. Formally, the soft classification rule is replaced by the following smoothed soft classifier:

$$g^\sigma(x) = E_{\epsilon \sim N(0, \sigma^2 I_d)} g(x + \epsilon) = \int_{\mathbb{R}^d} g(x+\epsilon) \phi^\sigma(\epsilon) d\epsilon,$$

where $\phi^\sigma \colon \mathbb{R}^d \to \mathbb{R}$ is the density function of $N(0, \sigma^2 I_d)$. The prediction is then assigned according to $f^\sigma(x) \in \arg\max_{i \in \mathcal{Y}} g_i^\sigma(x)$. Gaussian randomized smoothing has been shown to yield certified radii within which the (smoothed) prediction rule remains constant:

Theorem 1 (Cohen et al., 2019; Zhai et al., 2020)**.** *Let $\sigma^2 > 0$. Consider a point $x \in \mathbb{R}^d$ and let $y = f^\sigma(x) \in \arg\max_{i \in \mathcal{Y}} g_i^\sigma(x)$ and $y' \in \arg\max_{i \in \mathcal{Y} \setminus \{y\}} g_i^\sigma(x)$. Then $f^\sigma(x + \delta) = y$ for all $\delta \in \mathbb{R}^d$ such that*

$$\|\delta\|_2 \leq r^\sigma(x) := \frac{\sigma}{2} \left(\Phi^{-1}(g_y^\sigma(x)) - \Phi^{-1}(g_{y'}^\sigma(x)) \right).$$

Theorem 1 asserts that an adversarial perturbation δ of ℓ_2-norm radius less than $r^\sigma(x)$ cannot alter the prediction of the nominal input x upon using the smoothed classification scheme. Recent works have attempted optimizing over the smoothing variance σ^2 to maximize the certified radius $r^\sigma(x)$. However, some inputs may require a small σ^2 to maintain a correct prediction, whereas other inputs may permit a large σ^2 that yields a large certified radius, and therefore keeping σ^2 uniform with respect to x is overly conservative in general. To overcome this limitation, some works have allowed σ^2 to vary with x, but their proposed robustness certificates do not reflect the actual classification scheme being used at test time (Anderson & Sojoudi, 2022). To make their certificates valid, these works restrict themselves to using locally constant variances.

2.2. Generalized Smoothing

In this paper, we consider the generalized setting where the smoothing measure is allowed to vary with the input and is not necessarily Gaussian, and we seek to choose such a measure so as to maximize the average certified radius with respect to the data distribution. To this end, let X be a fixed subset of $\mathbb{R}^d$. For a measure-valued map $\mu \colon \mathbb{R}^d \to \mathcal{P}(X)$, we consider the smoothing scheme given by

$$g^\mu(x) = E_{\epsilon \sim \mu(x)} g(x+\epsilon) = \int_X g(x+\epsilon) d(\mu(x))(\epsilon),$$

$$f^\mu(x) \in \arg\max_{i \in \mathcal{Y}} g_i^\mu(x).$$

Of course, the set X over which we smooth the classifier is our choice. We will keep X arbitrary and fixed, albeit we impose the following assumption:

Assumption 1. The set X is compact, contains $0 \in \mathbb{R}^d$, and has positive Lebesgue measure.

Assumption 1 is natural with respect to our goal; such a set implies that the smoothing operation is non-negligible and being done locally around the input x of interest. Notice that, even with Gaussian smoothing, the smoothing is effectively performed on a compact set, since the Gaussian density function is nearly zero outside of some compact set. Using our notations, we see that $g^{\mu}(x) = \langle \tau_x g, \mu(x) \rangle$, which explicitly shows the linearty of $g^{\mu}(x)$ in μ.

It is not only important to maximize the size of the neighborhood around an input on which the classifier is constant (for robustness purposes), but also it is imperative to ensure that the classifier prediction is actually correct. We encode the correctness as follows: given a particular input-label pair $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$, the certified radius is

$$r^{\mu}(x, y) = \inf \left\{ \|x' - x\| : f^{\mu}(x') \neq y,\ x' \in \mathbb{R}^d \right\}. \quad (1)$$

Under this definition, if $f^{\mu}(x) \neq y$, then $r^{\mu}(x, y) = 0$; the radius of robustness is zero whenever the smoothed classifier does not predict the label y correctly. We define our metric of interest to be the *average certified radius* $E_{(x,y)\sim D} r^{\mu}(x, y)$, where D is the data distribution. Our problem now consists of solving the optimization

$$p^{\star} := \sup_{\mu \in \mathcal{F}} E_{(x,y)\sim D} r^{\mu}(x, y), \quad (2)$$

where $\mathcal{F}$ is the feasible set, which we leave as an arbitrary subset of all maps $\mu \colon \mathbb{R}^d \to \mathcal{P}(X)$ for the time being. Not only is the outer maximization in (2) a challenging infinite-dimensional problem in general, but also the inner minimization (1) is nonconvex. Thus, finding μ to attain the maximum $p^{\star}$ is intractable, and instead we propose a more amenable lower bound on the average certified radius, and then maximize this lower bound over μ.

3. Lower-Bounding the Certified Radius

In this section, we incorporate appropriate Lipschitz continuity conditions into the base soft classifier g and feasible set $\mathcal{F}$ to develop a lower bound on the certified radius that is linear in the variable μ. This will allow us to lower-bound $p^{\star}$ by an infinite linear program, as we will soon see. To this end, let ρ denote the metric induced by $\|\cdot\|$; $\rho(x', x) = \|x' - x\|$ for all $x', x \in \mathbb{R}^d$. Then the first such Lipschitz condition is formalized in the following assumption.

Assumption 2. Every component function g_i, as a map from $\left(\mathbb{R}^d, \rho\right)$ to $[0, 1]$ equipped with the metric induced by the absolute value $|\cdot| \colon \mathbb{R} \to [0, \infty)$, is Lipschitz continuous.

There exist many methods in the literature to numerically solve for upper bounds on the Lipschitz constants in Assumption 2—see, e.g., Weng et al. (2018); Fazlyab et al. (2019); Zhang et al. (2019); Jordan & Dimakis (2020). Alternatively, it suffices to replace the base classifier with one that has been preemptively smoothed using conventional randomized smoothing and to then employ the methods we present, as the Lipschitz constant of every g_i would then be given in closed-form by Salman et al. (2019).

We now make concrete the feasible set of measure-valued maps that we consider. In particular, we fix a constant $K \geq 0$ and let

$$\mathcal{F} = \{\mu \colon \mathbb{R}^d \to \mathcal{P}(X) : \mathrm{Lip}(\mu) \leq K\}. \quad (3)$$

Notice that the metric on $\mathcal{P}(X)$ is that induced by the total variation norm. The feasible set $\mathcal{F}$ restricts the smoothing scheme to those that do not yield drastically different smoothing measures for two distinct nearby inputs. This is precisely the condition that is missing from many works on input-dependent randomized smoothing, and without this condition, the resulting robustness certificates fail to hold in practice since no relation is granted between the smoothing distribution used at the nominal input and that used at a perturbed input, which is the one applied in reality. In general, directly enforcing the constraint in (3) is intractable, so we will later derive a tractable inner-approximation of the constraint set $\mathcal{F}$ that maintains certified radii in Section 4.1.

We are now in a position to lower-bound the certified radius, which depends on μ in a "nasty" way, by a quantity that is linear in μ.

Lemma 1. *Let $\mu \in \mathcal{F}$, and let $L^{(i)} \in \mathbb{R}_+$ be such that $\mathrm{Lip}(g_i) \leq L^{(i)}$ for all $i \in \mathcal{Y}$. Then, for all $i \in \mathcal{Y}$, it holds that g_i^{μ} is Lipschitz continuous with constant $C := \max\{L^{(i)} : i \in \mathcal{Y}\} + K$.*

Lemma 2. *Let $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$ be a fixed input-label pair, and let μ be a measure-valued map from $\mathbb{R}^d$ into $\mathcal{P}(X)$. Let $y' \in \arg\max_{i \in \mathcal{Y}\setminus\{y\}} g_i^{\mu}(x)$. If g_i^{μ} is Lipschitz continuous with constant $C \in \mathbb{R}_+$ for all $i \in \mathcal{Y}$, then*

$$\frac{1}{2C}\left(g_y^{\mu}(x) - g_{y'}^{\mu}(x)\right) \leq r^{\mu}(x, y).$$

Proposition 1. *Let $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$ be a fixed input-label pair and let $\mu \in \mathcal{F}$. Let $y' \in \arg\max_{i \in \mathcal{Y}\setminus\{y\}} g_i^{\mu}(x)$. Then*

$$\frac{1}{2C}\left(g_y^{\mu}(x) - g_{y'}^{\mu}(x)\right) \leq r^{\mu}(x, y),$$

where C is as in Lemma 1.

Intuitively, Proposition 1 shows that, for sufficiently continuous classifiers, the relative confidence between the two top-ranking classes provides an estimate of the certified radius. Lemma 1 reveals that the Lipschitz constant of the smoothed classifier may actually increase relative to that of the base classifier. We emphasize that this is *not* at odds with the goal of robustification; smoothness of decision boundaries is not always a good surrogate notion for robustness, and indeed decreasing the Lipschitz constant too much may cause a drop in classification accuracy (Yang et al., 2020b; Anderson & Sojoudi, 2022). Figure 2 illustrates this phenomenon, where a linear classifier with a low Lipschitz constant is less robust than a nonlinear classifier with a higher Lipschitz constant—the extent of nonlinearity should be determined based on the data distribution, not by the assumption that smoothness yields robustness.

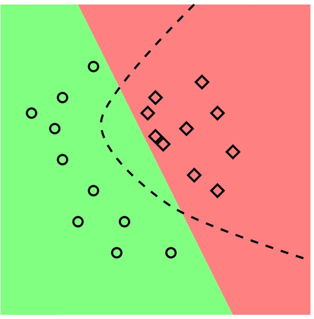

Figure 2. The nonlinear classifier (high Lipschitz constant) is more robust than the linear one (low Lipschitz constant).

The lower bound in Propostion 1 depends nonlinearly on μ since the runner-up class y' may depend on μ. To overcome this hurdle, we introduce the following standard assumption:

Assumption 3. The image of g is a probability simplex; $\sum_{i=1}^n g_i(x) = 1$ for all $x \in \mathbb{R}^d$.

Under Assumption 3, we obtain our final lower bound on the certified radius that is linear in μ:

Proposition 2. *Let (x, y), μ, and C be as in Proposition 1. Then*

$$\frac{1}{2C}\left(2g_y^\mu(x) - 1\right) \leq r^\mu(x, y). \tag{4}$$

4. Surrogate Optimization

With our lower bound (4) on the certified radius in place, we may formulate the following optimization as a surrogate for $p^\star$:

$$\sup_{\mu \in \mathcal{F}} E_{(x,y)\sim D} \frac{1}{2C}\left(2g_y^\mu(x) - 1\right) \leq p^\star, \tag{5}$$

where we recall that $C = \max\{L^{(i)} : i \in \mathcal{Y}\} + K$ and $L^{(i)} \in \mathbb{R}_+$ are such that $\text{Lip}(g_i) \leq L^{(i)}$ for all $i \in \mathcal{Y}$.

Now, although the feasible set $\mathcal{F}$ in (5) is convex, it is defined by an intractable constraint. We propose to overcome this challenge by relaxing the constraint to ones that maintain a lower bound on the certified radius and permit a tractable inner-approximation. This inner-approximation, defined by conic linear constraints, is derived in the next section.

4.1. Inner-Approximating the Lipschitz Constraint

We start by relaxing (5). In particular, notice that $\mathcal{F}$ is a subset of

$$\mathcal{F}' \coloneqq \{\mu \colon \mathbb{R}^d \to \mathcal{P}(X) : \text{Lip}(g_i^\mu) \leq C \text{ for all } i \in \mathcal{Y}\}, \tag{6}$$

as proven in Lemma 1. But also, by Lemma 2, we know that $\mu \in \mathcal{F}'$ maintains a lower bound on the certified radius, and hence we find that

$$\sup_{\mu \in \mathcal{F}} E_{(x,y)\sim D} \frac{1}{2C}\left(2g_y^\mu(x) - 1\right) \leq \sup_{\mu \in \mathcal{F}'} E_{(x,y)\sim D} \frac{1}{2C}\left(2g_y^\mu(x) - 1\right) \leq p^\star.$$

Hence, we may focus our attention on the feasible set $\mathcal{F}'$, and in particular, we seek to derive a tractable inner-approximation of this set. Let us first recall a few basic facts from analysis.

Lemma 3. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and let $\|\cdot\|_*$ be its corresponding dual norm. Assume that $h \colon \mathbb{R}^d \to \mathbb{R}$ is differentiable. Then $\text{Lip}(h) \leq \sup_{x \in \mathbb{R}^d} \|\nabla h(x)\|_*$.*

Lemma 4 (Folland, 1999, Chapter 5)**.** *All norms on $\mathbb{R}^d$ are equivalent. In particular, there exists $C_{*,\infty} \in \mathbb{R}_+$ such that $\|x\|_* \leq C_{*,\infty}\|x\|_\infty$ for all $x \in \mathbb{R}^d$.*

It is well-known for the common cases of $\|\cdot\| = \|\cdot\|_1$, $\|\cdot\| = \|\cdot\|_2$, and $\|\cdot\| = \|\cdot\|_\infty$ that the equivalence constant in Lemma 4 can be taken as $C_{*,\infty} = 1$, $C_{*,\infty} = \sqrt{d}$, and $C_{*,\infty} = d$, respectively.

Going forward, we make the following assumption.[1]

Assumption 4. The base classifier g is continuously differentiable; i.e., $g_i \in \mathcal{C}^1(\mathbb{R}^d)$ for all $i \in \mathcal{Y}$.

Now, let $\mathcal{F}_0$ be the class of maps $\mu \colon \mathbb{R}^d \to \mathcal{M}(X)_+$ with the following properties:

1. $\mu(x)$ has nonnegative density $\phi_x \in \mathcal{C}^1(\text{int}\, X)$ with respect to Lebesgue measure for all $x \in \mathbb{R}^d$, and
2. the density gradients are uniformly bounded with constant $C_0 \leq \frac{C}{m(X)}$, i.e., $\sup_{\epsilon \in \text{int}\, X} \|\nabla_x \phi_x(\epsilon)\|_* \leq C_0$ for all $x \in \mathbb{R}^d$, and
3. $(\mu(x))(X \setminus \text{int}\, X) = 0$ for all $x \in \mathbb{R}^d$.

The first property merely restricts attention to "nice" smoothing distributions with well-defined densities. The second property restricts those densities to ones that do not vary too

[1] It is easily seen that Assumption 4 can be relaxed to the assumption that g is continuously differentiable almost everywhere with respect to Lebesgue measure, making our results hold in most practical settings, including for ReLU neural networks.

rapidly with respect to the base classifier—this is natural, for if $\sup_{\epsilon\in\text{int } X}\|\nabla\phi_x(\epsilon)\|_*$ were too large, the distribution would nearly resemble a Dirac measure, resulting in a trivial shift of the classifier input upon smoothing. The third property is a benign technical condition that requires the measures to nullify the boundary of the set X. This is satisfied when the measure $\mu(x)$ is absolutely continuous with respect to Lebesgue measure. We now present our inner-approximation to the feasible set $\mathcal{F}'$.

Proposition 3. *For all $i\in\mathcal{Y}$ and $x\in\mathbb{R}^d$, let $A_i(x,\cdot)\colon\mathcal{M}(X)\to\mathbb{R}^{2d+1}$ be the linear operator defined by*

$$A_i(x,\eta)=\langle(\tau_x\nabla g_i,-\tau_x\nabla g_i,1),\eta\rangle\,,$$

and let $b_i\in\mathbb{R}^{2d+1}$ be the vector $b_i=(C_b\mathbf{1}_d,C_b\mathbf{1}_d,-1)$, where $C_b:=\frac{C-C_0m(X)}{C_{*,\infty}}\geq 0$, C is as in Lemma 1, and $C_{*,\infty}$ is as in Lemma 4. Define $\mathcal{K}\subseteq\mathbb{R}^{2d+1}$ to be the convex cone $\mathcal{K}=\mathbb{R}^d_+\times\mathbb{R}^d_+\times\{0\}$. Then, it holds that

$$\hat{\mathcal{F}}:=\bigcap_{i=1}^{n}\left\{\mu\in\mathcal{F}_0: A_i(x,\mu(x))+b_i\in\mathcal{K}\text{ for all } x\in\mathbb{R}^d\right\}$$

is a subset of the feasible set (6).

The inner-approximation $\hat{\mathcal{F}}$ is defined by affine conic constraints on μ, and is therefore much more amenable than directly constraining Lipschitz constants. If a map $\mu\colon\mathbb{R}^d\to\mathcal{M}(X)_+$ satisfies these conic constraints and the regularity properties defining $\mathcal{F}_0$, then Proposition 3 shows that the confidence $g_y^\mu(x)$ can be used to certifiably lower-bound the radius of robustness $r^\mu(x,y)$.

In practice, one may enforce that $\mu\in\mathcal{F}_0$ by instantiating a particular parameterization of the smoothing measures. For example, one may consider thresholded Gaussian densities with a fixed lower bound on the variance parameter, and consider the set of all maps corresponding to summations of k such densities for some fixed $k\in\mathbb{N}$. Alternatively, one may choose an adequately smooth class of neural networks to generate the parameterization.

4.2. Semi-Infinite Approximation

Let $(x_1,y_1),\dots,(x_N,y_N)\sim D$ be data to be used in the empirical risk optimization. The optimization, using the inner-approximation of Proposition 3, becomes

$$\hat{p}:=\sup\left\{\frac{1}{CN}\sum_{j=1}^{N}\left(\langle\tau_{x_j}g_{y_j},\mu(x_j)\rangle-1/2\right):\right.$$
$$\left.A(x,\mu(x))+b\in\mathcal{K}^n\text{ for all } x\in\mathbb{R}^d,\ \mu\colon\mathbb{R}^d\to\mathcal{M}(X)_+\right\}.$$

where $A(x,\cdot)=(A_1(x,\cdot),\dots,A_n(x,\cdot))$ for all x, $b=(b_1,\dots,b_n)$, and $\mathcal{K}^n=\mathcal{K}\times\cdots\times\mathcal{K}$. The "for all $x\in\mathbb{R}^d$" makes the problem infinite; the search space is infinite-dimensional and there are infinitely many constraints. To make the problem semi-infinite, we take the common discretization approach for infinite programs and consider

$$\hat{p}':=\frac{1}{CN}\sum_{j=1}^{N}\sup\left\{\langle\tau_{x_j}g_{y_j},\mu_j\rangle-1/2:\right.$$
$$\left.A(x_j,\mu_j)+b\in\mathcal{K}^n,\ \mu_j\in\mathcal{M}(X)_+\right\}.$$

Theoretical guarantees for the convergence of such discretization approaches to solutions of the original infinite program can be found in the the literature on solving general capacity problems, e.g., Lai & Wu (1992); Wu et al. (2001).

Now, for all $j\in\{1,2,\dots,N\}$, denote the jth subproblem in the above summation by

$$\hat{p}_j':=\sup\left\{\langle\tau_{x_j}g_{y_j},\mu_j\rangle-1/2:\right.$$
$$\left.A(x_j,\mu_j)+b\in\mathcal{K}^n,\ \mu_j\in\mathcal{M}(X)_+\right\}.\quad(7)$$

Every subproblem (7) is a canonical semi-infinite linear program; the objective is affine, the search space is infinite-dimensional as the decision variable is a measure, and the constraints are affine conic ones.

To generate an approximate solution for $\hat{p}$, we may solve (7) for $\mu_j^\star$ for all j, and then extend the measure-valued map to $\mathbb{R}^d$ by means of inverse-distance weighted interpolation, e.g., using Shepard's method (Shepard, 1968).

4.3. Solving the Subproblems

The subproblem (7) has an infinite-dimensional variable with finitely many constraints. More commonly considered is robust linear programming in finite-dimensional search spaces; problems with a finite-dimensional variable but infinitely many constraints. Indeed, the latter has readily available solution methods, e.g., discretizations and penalty methods, with associated convergence guarantees (Fang & Wu, 1994; Reemtsen & Görner, 1998; Fang, 2018). Fortunately, these types of semi-infinite linear programs are duals of one another. Following Shapiro (2001), the dual of (7) is

$$\hat{d}_j':=\inf\left\{-b^\top\lambda-1/2:\right.$$
$$\left.A^*(x_j,\lambda)-\tau_{x_j}g_{y_j}\in\mathcal{C}(X)_+,\ \lambda\in-\mathcal{K}^{n*}\right\},\quad(8)$$

where $A^*(x,\cdot)$ is the adjoint of the linear operator $A(x,\cdot)$, and where $\mathcal{K}^{n*}$ is the dual cone of $\mathcal{K}^n$. The dual problem (8) is a linear program with a finite-dimensional search space, but infinitely many constraints in the form of a nonnegative functional constraint. Now, a key challenge is that, in general, *strong duality does not hold for semi-infinite linear programs* (Shapiro, 2001). If this were the case for our problem, then $\hat{p}_j'<\hat{d}_j'$, which would void the lower bound on the certified radius. However, we can show that for the specific problem at hand, strong duality does indeed hold:

Theorem 2. *For all $j \in \{1, 2, \dots, N\}$, if $\hat{p}'_j$ is feasible, then strong duality holds; $\hat{p}'_j = \hat{d}'_j$.*

We emphasize that the duality theory given in Theorem 2 is nontrivial, and relies intimately on the topology of the feasible set of our particular semi-infinite program. The optimal value of the dual consequently yields an immediate certified radii around the given data in the case that the smoothing scheme satisfies our inner-approximation:

Corollary 1. *Let $\mu_j^\star$ solve $\hat{p}'_j$ for all $j \in \{1, 2, \dots, N\}$, and let $\mu \in \hat{\mathcal{F}}$ be such that $\mu(x_j) = \mu_j^\star$ for all such j. Then $\frac{1}{C}\hat{d}'_j = \frac{1}{C}\hat{p}'_j \leq r^\mu(x_j, y_j)$.*

Corollary 1 gives a simple way of estimating how suboptimal the certificate of a (possibly naively designed) smoothing scheme may be around various data points. In particular, if $\nu \colon \mathbb{R}^d \to \mathcal{P}(X)$ is a smoothing scheme with known certified radius $r^\nu(x_j, y_j) \ll \hat{d}'_j$, then Corollary 1 shows the existence of an alternative smoothing scheme μ with a larger certified radius attainable by solving (7).

5. Experiment

Consider a 3-layer ReLU neural network g with 10 neurons per hidden layer, 2 outputs, and 2 inputs for the purposes of visualization. The weights and biases are randomly initialized using normally distributed entries. We consider $N = 2$ data points $x_1 = (3, 0), x_2 = (-3, 0)$, shown in red in Figure 3, with associated classes $y_1 = y_2 = 2$, which corresponds to the blue decision regions. Clearly, g misclassifies x_1 and x_2. We perform conventional Gaussian smoothing with $\sigma = 0.5$ to arrive at the modified classifier g^σ in the middle of Figure 3. Since conventional smoothing is data-blind, it still misclassifies x_1 and x_2, albeit the decision boundaries have been smoothed. Finally, we apply our approach by solving (7) with μ_j parameterized by truncated Gaussians with support $X = [-5, 5]^2$, and then interpolate between these measures for general $x \in \mathbb{R}^2$ using the inverse-distance weighted Shepard's method (Shepard, 1968) with respect to x_1, x_2 to construct g^μ. As shown at the bottom of Figure 3, our method is able to correctly manipulate the decision boundary locally around the given data without compromising the general global behavior of the base classifier. In Appendix C, we give additional details and plots, including the recovery of the density of the optimal smoothing measure via the dual (8), and the use of Corollary 1 to quantify the suboptimality of conventional randomized smoothing.

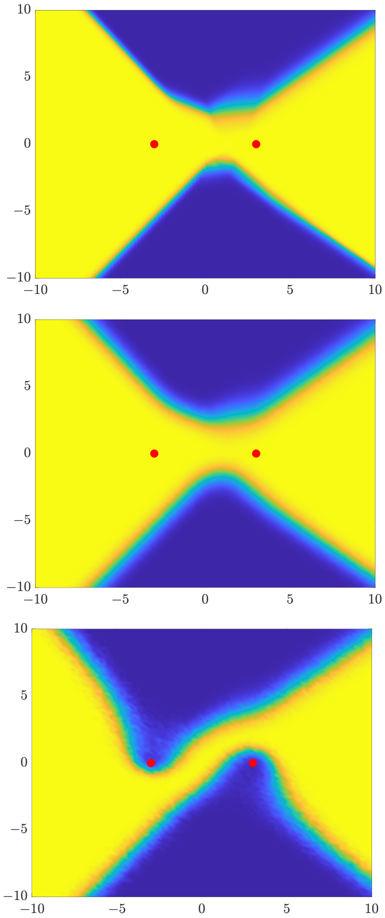

Figure 3. Plots of g (top), g^σ (middle), and g^μ (bottom). Optimal g^μ correctly learns how to manipulate decision boundary.

6. Conclusions and Future Work

In this paper, we theoretically study optimal randomized smoothing using the framework of infinite-dimensional optimization. We derive a lower bound on the certified radius of the smoothed classifier that is linear in the smoothing measure being optimized over. We prove that the bound is maintained upon replacing the problem's intractable Lipschitz constraints with more amenable conic linear constraints. A discretization approach is proposed as a means for numerically approximating the resulting problem by semi-infinite linear programs, which are then proven to enjoy nontrivial strong duality. An exploratory case study demonstrates that the theory developed offers a viable approach for solving general optimal randomized smoothing problems. Three avenues of future research are of primary interest: 1) further developing numerical tools for efficiently solving for and sampling from optimal smoothing distributions, 2) establishing a relationship between the dual (8) and the similar-looking problem of attacking a base classifier, and 3) connecting our semi-infinite programming approach to the related, yet more established topics of generalized moment problems and discounted optimal control using occupation measures.

References

Alfarra, M., Bibi, A., Torr, P. H., and Ghanem, B. Data dependent randomized smoothing. *arXiv preprint arXiv:2012.04351*, 2020.

Anderson, B. G. and Sojoudi, S. Certified robustness via locally biased randomized smoothing. In *Learning for Dynamics and Control*. PMLR, 2022.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 3–14, 2017.

Chen, C., Kong, K., Yu, P., Luque, J., Goldstein, T., and Huang, F. Insta-RS: Instance-wise randomized smoothing for improved robustness and accuracy. *arXiv preprint arXiv:2103.04436*, 2021.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Eiras, F., Alfarra, M., Kumar, M. P., Torr, P. H., Dokania, P. K., Ghanem, B., and Bibi, A. Ancer: Anisotropic certification via sample-wise volume maximization. *arXiv preprint arXiv:2107.04570*, 2021.

Fang, S.-C. LP with infinitely many constraints, August 2018. Lecture notes.

Fang, S.-C. and Wu, S.-y. An inexact approach to solving linear semi-infinite programming problems. *Optimization*, 28(3-4):291–299, 1994.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. J. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11427–11438, 2019.

Folland, G. B. *Real Analysis: Modern Techniques and Their Applications*, volume 40. John Wiley & Sons, 1999.

Hernández-Lerma, O. and Lasserre, J. B. *Markov Chains and Invariant Probabilities*, volume 211. Birkhäuser, 2012.

Jordan, M. and Dimakis, A. G. Exactly computing the local Lipschitz constant of ReLU networks. *Advances in Neural Information Processing Systems*, 33:7344–7353, 2020.

Kang, J., Ullah, Z., and Gwak, J. MRI-based brain tumor classification using ensemble of deep features and machine learning classifiers. *Sensors (Basel, Switzerland)*, 21, 2021.

Klerk, E. d. and Laurent, M. A survey of semidefinite programming approaches to the generalized problem of moments and their error analysis. In *World Women in Mathematics 2018*, pp. 17–56. Springer, 2019.

Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., and Zhang, Y. Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid*, 10(1):841–851, 2017.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

Lai, H. and Wu, S.-Y. Extremal points and optimal solutions for general capacity problems. *Mathematical Programming*, 54(1):87–113, 1992.

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*, pp. 656–672. IEEE, 2019.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Li, J. Certified defenses iii: Randomized smoothing, November 2019. Lecture notes.

Reemtsen, R. and Görner, S. Numerical methods for semi-infinite programming: a survey. In *Semi-infinite programming*, pp. 195–275. Springer, 1998.

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*, volume 32, 2019.

Shapiro, A. On duality theory of conic linear problems. In *Semi-Infinite Programming*, pp. 135–165. Springer, 2001.

Shepard, D. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, pp. 517–524, 1968.

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

Teng, J., Lee, G.-H., and Yuan, Y. ℓ_1 adversarial robustness certificates: A randomized smoothing approach. *Preprint*, 2020. URL https://openreview.net/forum?id=H1lQIgrFDS.

Wang, L., Zhai, R., He, D., Wang, L., and Jian, L. Pretrain-to-finetune adversarial training via sample-wise randomized smoothing. *Preprint*, 2021. URL https://openreview.net/pdf?id=Te1aZ2myPIu.

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.

Wu, B., Iandola, F., Jin, P. H., and Keutzer, K. SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 129–137, 2017.

Wu, S.-Y., Fang, S.-C., and Lin, C.-J. Solving general capacity problem by relaxed cutting plane approach. *Annals of Operations Research*, 103(1):193–211, 2001.

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*, pp. 10693–10705. PMLR, 2020a.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8588–8601, 2020b.

Zhai, R., Dan, C., He, D., Zhang, H., Gong, B., Ravikumar, P., Hsieh, C.-J., and Wang, L. Macer: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020.

Zhang, H., Zhang, P., and Hsieh, C.-J. RecurJac: An efficient recursive algorithm for bounding Jacobian matrix of neural networks and its applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5757–5764, 2019.

A. Supplementary Materials for Section 3 (Lower-Bounding the Certified Radius)

In this section, we prove the results given in Section 3.

Lemma 1. *Let $\mu \in \mathcal{F}$, and let $L^{(i)} \in \mathbb{R}_+$ be such that $\operatorname{Lip}(g_i) \leq L^{(i)}$ for all $i \in \mathcal{Y}$. Then, for all $i \in \mathcal{Y}$, it holds that g_i^μ is Lipschitz continuous with constant $C \coloneqq \max\{L^{(i)} : i \in \mathcal{Y}\} + K$.*

Proof. Let $i \in \mathcal{Y}$. By Lipschitz continuity of g_i, it holds that $|g_i(x' + \epsilon) - g_i(x + \epsilon)| \leq L^{(i)}\|x' - x\|$ for all $x', x \in \mathbb{R}^d$. Furthermore, by Lipschitz continuity of μ, it holds that $\|\mu(x') - \mu(x)\|_{\mathrm{TV}} \leq K\|x' - x\|$ for all $x', x \in \mathbb{R}^d$. Therefore, for all $x', x \in \mathbb{R}^d$, it holds that

$$
\begin{aligned}
|g_i^\mu(x') - g_i^\mu(x)| &= |\langle \tau_{x'} g_i, \mu(x')\rangle - \langle \tau_x g_i, \mu(x)\rangle| \\
&\leq |\langle \tau_{x'} g_i - \tau_x g_i, \mu(x')\rangle| + |\langle \tau_x g_i, \mu(x') - \mu(x)\rangle| \\
&\leq \int_X |g_i(x' + \epsilon) - g_i(x + \epsilon)|\, d(\mu(x'))(\epsilon) + \|\tau_x g_i\|_\infty \|\mu(x') - \mu(x)\|_{\mathrm{TV}} \\
&\leq L^{(i)}\|x' - x\| + \|g_i\|_\infty \|\mu(x') - \mu(x)\|_{\mathrm{TV}} \\
&\leq \left(L^{(i)} + K\right)\|x' - x\| \\
&\leq \left(\max\{L^{(i)} : i \in \mathcal{Y}\} + K\right)\|x' - x\|,
\end{aligned}
$$

where we used the fact that $\|\tau_x g_i\|_\infty = \|g_i\|_\infty$ and that $|g_i| \leq 1$. Thus, g_i^μ is indeed Lipschitz continuous with the claimed constant. □

Lemma 2. *Let $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$ be a fixed input-label pair, and let μ be a measure-valued map from $\mathbb{R}^d$ into $\mathcal{P}(X)$. Let $y' \in \arg\max_{i \in \mathcal{Y} \setminus \{y\}} g_i^\mu(x)$. If g_i^μ is Lipschitz continuous with constant $C \in \mathbb{R}_+$ for all $i \in \mathcal{Y}$, then*

$$
\frac{1}{2C}\left(g_y^\mu(x) - g_{y'}^\mu(x)\right) \leq r^\mu(x, y).
$$

Proof. First, note that, if $y \notin \arg\max_{i \in \mathcal{Y}} g_i^\mu(x)$, then $g_y^\mu(x) - g_{y'}^\mu(x) \leq 0$, so the result trivially holds due to the nonnegativity of $r^\mu(x, y)$. Therefore, we assume without loss of generality that $y \in \arg\max_{i \in \mathcal{Y}} g_i^\mu(x)$.

Now, let $x' \in \mathbb{R}^d$ be such that $\|x' - x\| \leq \frac{1}{2C}\left(g_y^\mu(x) - g_{y'}^\mu(x)\right)$. Such x' exists since $y \in \arg\max_{i \in \mathcal{Y}} g_i^\mu(x)$. This, together with Lipschitz continuity of g_y^μ, gives that

$$
|g_y^\mu(x') - g_y^\mu(x)| \leq \frac{1}{2}\left(g_y^\mu(x) - g_{y'}^\mu(x)\right),
$$

which implies that

$$
g_y^\mu(x') \geq \frac{1}{2}\left(g_y^\mu(x) + g_{y'}^\mu(x)\right).
$$

On the other hand, for $i \neq y$, Lipschitz continuity of g_i^μ gives that

$$
|g_i^\mu(x') - g_i^\mu(x)| \leq \frac{1}{2}\left(g_y^\mu(x) - g_{y'}^\mu(x)\right),
$$

which implies that

$$
g_i^\mu(x') \leq g_i^\mu(x) + \frac{1}{2}\left(g_y^\mu(x) - g_{y'}^\mu(x)\right) \leq \frac{1}{2}\left(g_y^\mu(x) + g_{y'}^\mu(x)\right) \leq g_y^\mu(x').
$$

Therefore, $f^\mu(x') = y$ for all $x' \in \mathbb{R}^d$ such that $\|x' - x\| \leq \frac{1}{2C}\left(g_y^\mu(x) - g_{y'}^\mu(x)\right)$, which implies that

$$
\frac{1}{2C}\left(g_y^\mu(x) - g_{y'}^\mu(x)\right) \leq \inf\{\|x' - x\| : f^\mu(x') \neq y,\ x' \in \mathbb{R}^d\} = r^\mu(x, y),
$$

which proves the result. □

Proposition 1. *Let $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$ be a fixed input-label pair and let $\mu \in \mathcal{F}$. Let $y' \in \arg\max_{i\in\mathcal{Y}\setminus\{y\}} g_i^\mu(x)$. Then*

$$\frac{1}{2C}\left(g_y^\mu(x) - g_{y'}^\mu(x)\right) \le r^\mu(x, y),$$

where C is as in Lemma 1.

Proof. This follows directly from Lemmas 1 and 2. □

Proposition 2. *Let (x, y), μ, and C be as in Proposition 1. Then*

$$\frac{1}{2C}\left(2g_y^\mu(x) - 1\right) \le r^\mu(x, y). \tag{4}$$

Proof. By Assumption 3 and linearity of the Lebesgue integral, we have that

$$\sum_{i=1}^n g_i^\mu(x) = \sum_{i=1}^n E_{\epsilon\sim\mu(x)} g_i(x+\epsilon) = E_{\epsilon\sim\mu(x)} \sum_{i=1}^n g_i(x+\epsilon) = 1.$$

Hence, for $y' \in \arg\max_{i\in\mathcal{Y}\setminus\{y\}} g_i^\mu(x)$, it holds that

$$g_{y'}^\mu(x) = \max_{i\in\mathcal{Y}\setminus\{y\}} g_i^\mu(x) \le \sum_{i\ne y} g_i^\mu(x) = 1 - g_y^\mu(x).$$

Therefore, by Proposition 1

$$\frac{1}{2C}\left(2g_y^\mu(x) - 1\right) \le \frac{1}{2C}\left(g_y^\mu(x) - g_{y'}^\mu(x)\right) \le r^\mu(x, y).$$

□

B. Supplementary Materials for Section 4 (Surrogate Optimization)

In this section, we prove the results given in Section 4.

Lemma 3. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and let $\|\cdot\|_*$ be its corresponding dual norm. Assume that $h\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable. Then $\mathrm{Lip}(h) \le \sup_{x\in\mathbb{R}^d} \|\nabla h(x)\|_*$.*

Proof. Let $x', x \in \mathbb{R}^d$. By the mean-value theorem, we have that $h(x') = h(x) + \nabla h(\overline{x})^\top(x' - x)$ for some $\overline{x} \in \mathbb{R}^d$ on the line segment between x and x'. Therefore,

$$|h(x') - h(x)| = |\nabla h(\overline{x})^\top(x' - x)| \le \|\nabla h(\overline{x})\|_* \|x' - x\| \le \sup_{y\in\mathbb{R}^d} \|\nabla h(y)\|_* \|x' - x\|.$$

Since $x', x \in \mathbb{R}^d$ are arbitrary, we conclude that h is Lipschitz continuous with constant $\sup_{x\in\mathbb{R}^d} \|\nabla h(x)\|_*$. □

Proposition 3. *For all $i \in \mathcal{Y}$ and $x \in \mathbb{R}^d$, let $A_i(x, \cdot)\colon \mathcal{M}(X) \to \mathbb{R}^{2d+1}$ be the linear operator defined by*

$$A_i(x, \eta) = \langle(\tau_x\nabla g_i, -\tau_x\nabla g_i, 1), \eta\rangle,$$

and let $b_i \in \mathbb{R}^{2d+1}$ be the vector $b_i = (C_b\mathbf{1}_d, C_b\mathbf{1}_d, -1)$, where $C_b \coloneqq \frac{C - C_0 m(X)}{C_{*,\infty}} \ge 0$, C is as in Lemma 1, and $C_{*,\infty}$ is as in Lemma 4. Define $\mathcal{K} \subseteq \mathbb{R}^{2d+1}$ to be the convex cone $\mathcal{K} = \mathbb{R}_+^d \times \mathbb{R}_+^d \times \{0\}$. Then, it holds that

$$\hat{\mathcal{F}} \coloneqq \bigcap_{i=1}^n \left\{\mu \in \mathcal{F}_0 : A_i(x, \mu(x)) + b_i \in \mathcal{K} \text{ for all } x \in \mathbb{R}^d\right\}$$

is a subset of the feasible set (6).

Proof. Let $\mu \in \hat{\mathcal{F}}$. Let $x \in \mathbb{R}^d$ and consider the density $\phi_x \in \mathcal{C}^1(\operatorname{int} X)$ of $\mu(x)$. Let $i \in \mathcal{Y}$. Then

$$g_i^\mu(x) = \int_X g_i(x+\epsilon) d(\mu(x))(\epsilon) = \int_{\operatorname{int} X} g_i(x+\epsilon)\phi_x(\epsilon) d\epsilon.$$

Now, by the definition of $\mathcal{F}_0$ and a standard application of the dominated convergence theorem together with the mean-value theorem (see, e.g., Folland (1999, Theorem 2.27) and the discussion thereafter), we may interchange differentiation with Lebesgue integration:

$$\nabla_x \int_{\operatorname{int} X} g_i(x+\epsilon)\phi_x(\epsilon) d\epsilon = \int_{\operatorname{int} X} \nabla_x \left(g_i(x+\epsilon)\phi_x(\epsilon)\right) d\epsilon.$$

Hence, by the product rule for differentiation,

$$\begin{aligned}
\nabla g_i^\mu(x) &= \int_{\operatorname{int} X} \left((\nabla_x g_i(x+\epsilon))\phi_x(\epsilon) + g_i(x+\epsilon)\nabla_x \phi_x(\epsilon)\right) d\epsilon \\
&= \int_X \nabla g_i(x+\epsilon) d(\mu(x))(\epsilon) + \int_{\operatorname{int} X} g_i(x+\epsilon)\nabla_x\phi_x(\epsilon) d\epsilon \\
&= \langle \tau_x \nabla g_i, \mu(x)\rangle + \int_{\operatorname{int} X} \tau_x g_i(\epsilon) \nabla_x \phi_x(\epsilon) d\epsilon.
\end{aligned}$$

Therefore, by Lemma 4 and the fact that $|g_i| \le 1$,

$$\begin{aligned}
\|\nabla g_i^\mu(x)\|_* &\le \| \langle \tau_x \nabla g_i, \mu(x)\rangle \|_* + \int_{\operatorname{int} X} |\tau_x g_i(\epsilon)| \|\nabla_x \phi_x(\epsilon)\|_* d\epsilon \\
&\le C_{*,\infty} \| \langle \tau_x \nabla g_i, \mu(x)\rangle \|_\infty + \int_{\operatorname{int} X} \sup_{\epsilon \in \operatorname{int} X} \|\nabla_x \phi_x(\epsilon)\|_* d\epsilon \\
&\le C_{*,\infty} \| \langle \tau_x \nabla g_i, \mu(x)\rangle \|_\infty + C_0 m(X).
\end{aligned}$$

By the definitions of A_i and b_i, together with the fact that $A_i(x, \mu(x)) + b_i \in \mathcal{K}$, it holds that

$$\| \langle \tau_x \nabla g_i, \mu(x)\rangle \|_\infty \le C_b = \frac{C - C_0 m(X)}{C_{*,\infty}},$$

which therefore implies that

$$\|\nabla g_i^\mu(x)\|_* \le C.$$

Hence, by Lemma 3, we have that $\operatorname{Lip}(g_i^\mu) \le C$, which proves that $\hat{\mathcal{F}}$ is a subset of (6) as desired. □

Theorem 2. *For all $j \in \{1, 2, \dots, N\}$, if $\hat{p}'_j$ is feasible, then strong duality holds; $\hat{p}'_j = \hat{d}'_j$.*

Proof. Throughout the proof, we use the following notations for a normed vector space $(\mathcal{X}, \|\cdot\|)$. The closed unit ball of radius $r \ge 0$ centered at $y \in \mathcal{X}$ is denoted by $B_{\mathcal{X}}(y, r) = \{x \in \mathcal{X} : \|x - y\| \le r\}$. The dual space of $\mathcal{X}$ is denoted by $\mathcal{X}^*$, and the operator norm on $\mathcal{X}^*$ is denoted by $\|\cdot\|_{\mathcal{X}^*}$.

Let $j \in \{1, 2, \dots, N\}$ and recall the primal and dual of interest:

$$\begin{aligned}
\hat{p}'_j &= \sup \left\{ \left\langle \tau_{x_j} g_{y_j}, \mu_j \right\rangle - 1/2 : A(x_j, \mu_j) + b \in \mathcal{K}^n,\ \mu_j \in \mathcal{M}(X)_+ \right\}, \\
\hat{d}'_j &= \inf \left\{ -b^\top \lambda - 1/2 : A^*(x_j, \lambda) - \tau_{x_j} g_{y_j} \in \mathcal{C}(X)_+,\ \lambda \in -\mathcal{K}^{n*} \right\}.
\end{aligned}$$

Assume that $\hat{p}'_j$ is feasible. For every μ_j feasible for $\hat{p}'_j$, it must be that $| \left\langle \tau_{x_j}, g_{y_j}, \mu_j \right\rangle | \le \|\tau_{x_j} g_{y_j}\|_\infty \|\mu_j\|_{\mathrm{TV}} \le 1$ since $|g_i| \le 1$ for all i and $\mu_j \in \mathcal{P}(X)$. Therefore, since $\hat{p}'_j$ is feasible, the optimal value of the primal maximization is finite. Hence, by Shapiro (2001, Proposition 2.6), it suffices to prove that the convex cone

$$\begin{aligned}
M &:= \left\{ (\lambda, \alpha) \in \mathbb{R}^{n(2d+1)} \times \mathbb{R} : \lambda = \nu - A(x_j, \mu_j),\ \alpha = \left\langle -\tau_{x_j} g_{y_j}, \mu_j \right\rangle,\ \mu_j \in \mathcal{M}(X)_+,\ \nu \in \mathcal{K}^n \right\} \\
&= \left\{ \left(\nu - A(x_j, \mu_j), \left\langle -\tau_{x_j} g_{y_j}, \mu_j \right\rangle\right) \in \mathbb{R}^{n(2d+1)} \times \mathbb{R} : \mu_j \in \mathcal{M}(X)_+,\ \nu \in \mathcal{K}^n \right\}
\end{aligned}$$

is closed with respect to the product topology on $\mathbb{R}^{n(2d+1)} \times \mathbb{R}$. To this end, let $\{(\lambda^{(k)}, \alpha^{(k)}) \in M : k \in \mathbb{N}\}$ be a sequence in M such that $(\lambda^{(k)}, \alpha^{(k)}) \to (\overline{\lambda}, \overline{\alpha})$ as $k \to \infty$ for some $(\overline{\lambda}, \overline{\alpha}) \in \mathbb{R}^{n(2d+1)} \times \mathbb{R}$. We will prove that M is closed by proving that $(\overline{\lambda}, \overline{\alpha}) \in M$.

First, note by definition of M that there exists $\{(\mu_j^{(k)}, \nu^{(k)}) \in \mathcal{M}(X)_+ \times \mathcal{K}^n : k \in \mathbb{N}\}$ such that $\lambda^{(k)} = \nu^{(k)} - A(x_j, \mu_j^{(k)})$ and $\alpha^{(k)} = \big\langle -\tau_{x_j} g_{y_j}, \mu_j^{(k)} \big\rangle$ for all $k \in \mathbb{N}$. Since $\nu^{(k)} \in \mathcal{K}^n$ for all $k \in \mathbb{N}$, it holds that $\nu_{2d+1}^{(k)} = 0$ for all such k and therefore

$$\lambda_{2d+1}^{(k)} = \nu_{2d+1}^{(k)} - A(x_j, \mu_j^{(k)})_{2d+1} = -\langle 1, \mu_j^{(k)} \rangle \to \overline{\lambda}_{2d+1}$$

as $k \to \infty$. Hence,

$$\lim_{k\to\infty} \|\mu_j^{(k)}\|_{\mathrm{TV}} = \lim_{k\to\infty} \mu_j^{(k)}(X) = \lim_{k\to\infty} \langle 1, \mu_j^{(k)} \rangle = -\overline{\lambda}_{2d+1}.$$

Let $\epsilon > 0$ be fixed. Then the above sequential convergence implies that there exists $k^\star \in \mathbb{N}$ such that $\|\mu_j^{(k)}\|_{\mathrm{TV}} \leq |\overline{\lambda}_{2d+1}| + \epsilon$ for all $k \geq k^\star$. Thus,

$$\mu_j^{(k)} \in B_{\mathcal{M}(X)}(0, r) = \{\mu_j \in \mathcal{M}(X) : \|\mu_j\|_{\mathrm{TV}} \leq r\}$$

for all $k \geq k^\star$, where $r = |\overline{\lambda}_{2d+1}| + \epsilon$. In other words, the scaled measure $\mu_j^{(k)}/r$ is in the closed unit ball $B_{\mathcal{M}(X)}(0,1) = \{\mu_j \in \mathcal{M}(X) : \|\mu_j\|_{\mathrm{TV}} \leq 1\}$ for all $k \geq k^\star$. Since X is a compact subset of the Hausdorff space $\mathbb{R}^d$, we have that the map $\varphi \colon \mathcal{M}(X) \to \mathcal{C}(X)^*$ defined by $\varphi(\mu_j)(\ell) = \int_X \ell d\mu_j$ is an isometric isomorphism from $\mathcal{M}(X)$ to $\mathcal{C}(X)^*$ by the Riesz representation theorem (Folland, 1999, Theorem 7.17). Therefore,

$$\varphi(\mu_j^{(k)}/r) \in B_{\mathcal{C}(X)^*}(0,1) = \{I \in \mathcal{C}(X)^* : \|I\|_{\mathcal{C}(X)^*} \leq 1\}$$

for all $k \geq k^\star$. By the Banach-Alaoglu theorem (Folland, 1999, Theorem 5.18), the ball $B_{\mathcal{C}(X)^*}(0,1)$ is compact in the weak-$*$ topology on $\mathcal{C}(X)^*$. Moreover, since X is compact, the space $\mathcal{C}(X)$ is separable, so the ball $B_{\mathcal{C}(X)^*}(0,1)$ is metrizable (Hernández-Lerma & Lasserre, 2012, Lemma 1.3.2), and therefore $B_{\mathcal{C}(X)^*}(0,1)$ is sequentially compact in the weak-$*$ topology. Therefore, the sequence $\{\varphi(\mu_j^{(k)}/r) : k \geq k^\star\} \subseteq B_{\mathcal{C}(X)^*}(0,1)$ has a convergent subsequence, i.e., there exists $\{k_l \geq k^\star : l \in \mathbb{N}\}$ and $\overline{I} \in B_{\mathcal{C}(X)^*}(0,1)$ such that

$$\varphi(\mu_j^{(k_l)}/r) \to \overline{I}$$

in the weak-$*$ topology as $l \to \infty$. Since φ is an isometric isomorphism from $\mathcal{M}(X)$ to $\mathcal{C}(X)^*$, this implies that

$$\mu_j^{(k_l)} \to \overline{\mu}_j \coloneqq r\varphi^{-1}(\overline{I}) \in B_{\mathcal{M}(X)}(0, r)$$

in the weak-$*$ topology on $\mathcal{M}(X)$ (induced by φ^{-1} on $\mathcal{C}(X)^*$) as $l \to \infty$, i.e., $\big\langle \ell, \mu_j^{(k_l)} \big\rangle \to \langle \ell, \overline{\mu} \rangle$ for all $\ell \in \mathcal{C}(X)$. In particular, $\big\langle -\tau_{x_j} g_{y_j}, \mu_j^{(k_l)} \big\rangle \to \big\langle -\tau_{x_j} g_{y_j}, \overline{\mu} \big\rangle$ as $\ell \to \infty$. Since $\big\langle -\tau_{x_j} g_{y_j}, \mu_j^{(k)} \big\rangle = \alpha^{(k)} \to \overline{\alpha}$ as $k \to \infty$, it must be that

$$\overline{\alpha} = \big\langle -\tau_{x_j} g_{y_j}, \overline{\mu} \big\rangle.$$

Similarly, we have that $A(x_j, \mu_j^{(k_l)}) \to A(x_j, \overline{\mu})$ as $l \to \infty$. Since $\nu^{(k)} - A(x_j, \mu_j^{(k)}) = \lambda^{(k)} \to \overline{\lambda}$, as $k \to \infty$, it must be that $\nu_{k_l} \to \overline{\lambda} + A(x_j, \overline{\mu}) =: \overline{\nu}$ as $l \to \infty$. That is,

$$\overline{\lambda} = \lim_{l\to\infty} (\nu^{(k_l)} - A(x_j, \mu_j^{(k_l)})) = \overline{\nu} - A(x_j, \overline{\mu}).$$

Since $\mathcal{M}(X)_+$ is closed in the weak-$*$ topology (Klerk & Laurent, 2019), we have that $\overline{\mu} \in \mathcal{M}(X)_+$. Furthermore, since $\nu^{(k)} \in \mathcal{K}^n$ for all $k \in \mathbb{N}$ and $\mathcal{K}^n$ is closed in $\mathbb{R}^{n(2d+1)}$, it holds that $\overline{\nu} \in \mathcal{K}^n$. Thus, we have proved that $(\overline{\lambda}, \overline{\alpha}) \in M$, so M is closed, which concludes the proof. □

C. Supplementary Materials for Section 5 (Experiment)

In this section, we present additional results from the experimental setup of Section 5.

We consider solving the dual problem (8) using a discretization of X with 100^2 uniformly spaced points. For guarantees on the convergence of such discretizations to the true solution of this robust linear program, we refer the interested reader to the literature on algorithms for semi-infinite linear programming, e.g., Fang & Wu (1994); Reemtsen & Görner (1998); Fang (2018). The resulting problem is a finite-dimensional linear program with $n(2d+1) = 10$ variables and $2dn + 100^2 = 10008$ constraints. We solve this problem and return an optimal solution for the associated double-dual, i.e., the dual of the discretization of (8). We denote the optimal double-dual variables associated to the discretization of the constraint $A^*(x_j, \lambda) - \tau_{x_j} g_{y_j} \in \mathcal{C}(X)_+$ by $\hat{\phi}_j^\star \in \mathbb{R}^{100^2}$. Now, the ith element of $\hat{\phi}_j^\star$ is the value of the density function corresponding to $\mu_j^\star$ at the ith point of the discretization of X, and thus we may visualize (a discretization-based approximation of) the optimal measure $\mu_j^\star$ that solves (7) by plotting a properly reshaped version of $\hat{\phi}_j^\star$. This is done for $x_j = (0, 0)$ with $y_j = 1$ and shown in Figure 4. As expected, the optimization learns to place the measure's density in the region associated with class $y_j = 1$. We remark that the shape of the density $\hat{\phi}_j^\star$ is non-Gaussian and nontrivial, indicating that Gaussian smoothing is likely to be suboptimal around this input. We quantitatively characterize this suboptimality next.

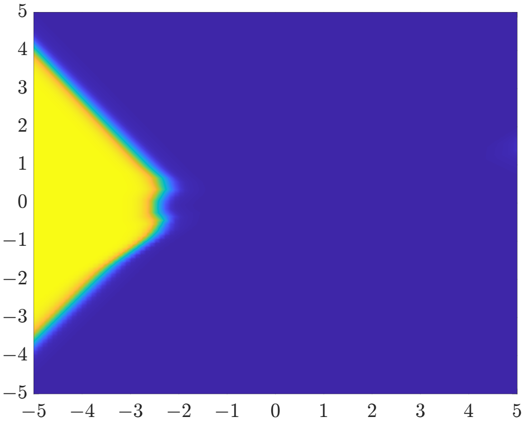

Figure 4. Density of the optimal parameterized measure μ_1 associated to $x_1 = (3, 0)$ with class $y_1 = 2$.

Our goal is to demonstrate the potential suboptimality of conventional Gaussian randomized smoothing by considering locally optimized smoothing measures that satisfy Corollary 1 for their respective data points. We then compare our lower bound $\frac{1}{C}\hat{d}'_j$ on the certified radius at x_j, computed via the dual (8), to that of Gaussian randomized smoothing (denoted $r^\sigma(x_j)$) over a range of variances. One technical nuance to be noted is that Corollary 1 requires the existence of a measure-valued map $\mu \in \hat{\mathcal{F}}$ such that $\mu(x_j) = \mu_j^\star$. Such μ always exists by simply taking $\mu(x) = \mu_j^\star$ for all $x \in \mathbb{R}^d$, and in this case, our comparisons amount to quantifying how suboptimal Gaussian smoothing is at the point x_j relative to the optimal smoothing measure at x_j, in a pointwise sense. To extend this assessment to multiple inputs, we sample inputs x_j uniformly at random over the domain of Figure 3 all with class $y_j = 1$, and compute the average of the difference $\hat{d}'_j/C - r^\sigma(x_j)$ over these samples. We remark that the Lipschitz constant of the Gaussian-smoothed classifier g_i^σ using variance σ is given in closed-form as $\sqrt{\frac{2}{\pi\sigma^2}}$ (Salman et al., 2019; Li, 2019), and therefore in what follows, we take $C = \sqrt{\frac{2}{\pi\sigma^2}}$ as the maximum allowable Lipschitz constant of g_i^μ so as to ensure a fair comparison between the certified radius using our method and that using Gaussian smoothing with variance σ.

Figure 5 demonstrates an interesting trend: at low values of σ, we find that, on average, the Gaussian smoothing certified radius is better than the lower bound given by the dual optimal value $\hat{d}'_j$, and that this is likely an artifact of the lower bound in Corollary 1 not being tight due to the reliance on the global Lipschitz constant C (which is large since σ is small). This points to a possible direction for future work in which the lower bounds utilize local Lipschitz constants, in effect tightening the gap between the semi-infinite linear programming bounds and the true certified radius $r^\mu(x_j, y_j)$. On the other hand, as σ increases, we see that our lower bound on the certified radius becomes increasingly large relative to that of Gaussian smoothing, indicating that Gaussian smoothing becomes increasingly suboptimal (in a pointwise sense) as the variance being used becomes larger. This provides quantitative evidence that, unlike Gaussian smoothing that flattens the confidence of the classifier towards a constant function in an uninformed manner as σ increases, an optimal (data-informed) smoothing scheme has the ability to learn where to allocate density in order to maintain large certified radii even as the Lipschitz constant of the smoothed classifier is required to decrease.

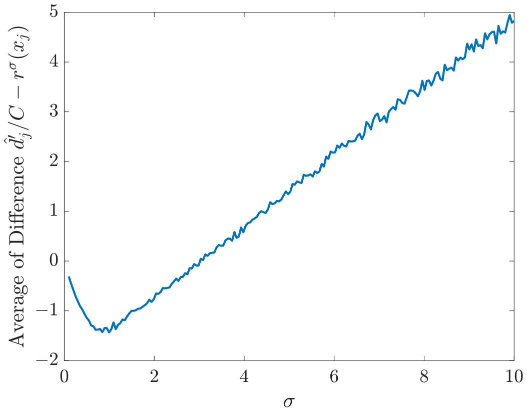

Figure 5. Gap between optimized certified radius via dual (8) and certified radius of Gaussian smoothing.