A FRAMEWORK FOR ROBUSTNESS CERTIFICATION OF SMOOTHED CLASSIFIERS USING F-DIVERGENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Formal verification techniques that compute provable guarantees on properties of machine learning models, like robustness to norm-bounded adversarial perturbations, have yielded impressive results. Although most techniques developed so far requires knowledge of the architecture of the machine learning model and remains hard to scale to complex prediction pipelines, the method of *randomized smoothing* has been shown to overcome many of these obstacles. By requiring only black-box access to the underlying model, randomized smoothing scales to large architectures and is agnostic to the internals of the network. However, past work on randomized smoothing has focused on restricted classes of smoothing measures or perturbations (like Gaussian or discrete) and has only been able to prove robustness with respect to simple norm bounds. In this paper we introduce a general framework for proving robustness properties of smoothed machine learning models in the black-box setting. Specifically, we extend randomized smoothing procedures to handle *arbitrary* smoothing measures and prove robustness of the smoothed classifier by using f-divergences. Our methodology achieves *state-of-the-art* certified robustness on MNIST, CIFAR-10 and ImageNet and also audio classification task, Librispeech, with respect to several classes of adversarial perturbations.

1 INTRODUCTION

Predictors obtained from machine learning algorithms have been shown to be vulnerable to making errors when the inputs are perturbed by carefully chosen small but imperceptible amounts (Szegedy et al., 2014; Biggio et al., 2013). This has motivated significant amount of research in improving adversarial robustness of a machine learning model such as Madry et al. (2018); Goodfellow et al. (2015). While significant advances have been made, it has been shown that models that were estimated to be robust have later been broken by stronger attacks (Athalye et al., 2018; Uesato et al., 2018). This has led to the need for methods that offer provable guarantees that the predictor cannot be forced to misclassify an example by *any attack algorithm* restricted to produce perturbations within a certain set (for example, within an ℓ_p norm ball). While progress has been made leading to methods that are able to compute provable guarantees for several image and text classification tasks (Wong & Kolter, 2018; Wong et al., 2018; Raghunathan et al., 2018; Dvijotham et al., 2018; Katz et al., 2017; Huang et al., 2019; Jia et al., 2019), these methods require extensive knowledge of the architecture of the predictor and are not easy to extend to new models or architectures, requiring specialized algorithms for each new class of models. Further, the computational complexity of these methods grows significantly with input dimension and model size.

Consequently, to deal with these obstacles, recent work has proposed the *randomized smoothing* strategy for verifying the robustness of classifiers. Specifically, Lecuyer et al. (2018); Cohen et al. (2019) have shown that robustness properties can be more easily verified for the *smoothed* version of a base classifier h:

$$h_s(x) = \arg\max_{y \in \mathcal{Y}} \mathop{\mathbb{P}}_{X \sim \mu(x)}[h(X) = y] \ , \tag{1}$$

where the labels returned by the smoothed classifier h_s are obtained by taking a "majority vote" over the predictions of the original classifier h on random inputs drawn from a probability distribution $\mu(x)$, called the *smoothing measure* (here $\mathcal{Y}$ denotes the set of classes in the problem). Lecuyer et al. (2018) showed that verifying the robustness of this smoothed classifier is significantly simpler than verifying the original classifier h and only requires estimating the distribution of outputs of the

classifier under random perturbations of the input, but does not require access to the internals of the classifier h. We refer to this as *black-box verification.*

In this work, we develop a general framework for black-box verification that recovers prior work as special cases, and improves upon previous results in various ways.

Contributions Our contributions are summarized as follows:

1. We formulate the general problem of black-box verification via a generalized randomized smoothing procedure, which extends past approaches to allow for arbitrary smoothing measures. Specifically, we show that robustness certificates for smoothed classifiers can be obtained by solving a small convex optimization problem when adversarial perturbations can be characterized via divergence-based bounds on the smoothing measure. Our certificates generalize previous results obtained in related work (Lecuyer et al., 2018; Cohen et al., 2019), and vastly extend the class of perturbations that randomized smoothing procedures can certify.
2. We evaluate our framework experimentally for several audio and image classification tasks, obtaining robustness certificates that improve upon other black-box methods. In particular, we get state-of-the-art results on robustness to ℓ_0 and ℓ_1 norm perturbations on CIFAR-10 and ImageNet. We also obtain the first, to the best of knowledge, certifiably robust model for an audio classification task.

2 BLACK-BOX VERIFICATION FOR SMOOTHED CLASSIFIERS

Consider a binary classifier $h : \mathcal{X} \to \{\pm 1\}$ given to us as a black box, so we can only access the inputs and outputs of h but not its internals. We are interested in investigating the robustness of the smoothed classifier h_s (defined in Eq. 1) against adversarial perturbations of size at most ϵ with respect to a given norm $\|\cdot\|$. To determine whether a norm bounded adversarial attack on a fixed input $x \in \mathcal{X}$ with $h_s(x) = +1$ could be successful, we can solve the optimization problem

$$\min_{\|x'-x\| \leq \epsilon} \mathop{\mathbb{P}}_{X' \sim \mu(x')} [h(X') = +1] \ , \tag{2}$$

and check whether the minimum value can be smaller than $\frac{1}{2}$. This is a non-convex optimization problem for which we may not even be able to compute gradients since we only have black-box access to h. While techniques have been developed to address this problem, obtaining provable guarantees on whether these algorithms actually find the worst-case adversarial perturbation is difficult since we do not know anything about the nature of h.

Motivated by this difficulty, we take a different approach: Rather than studying the adversarial attack in the input space $\mathcal{X}$, we study it in the space of probability measures over inputs, denoted by $\mathcal{P}(\mathcal{X})$. Formally, this amounts to rewriting Eq. 2 as

$$\min_{\nu \in \{\mu(x') : \|x'-x\| \leq \epsilon\}} \mathop{\mathbb{P}}_{X' \sim \nu} [h(X') = +1] \ . \tag{3}$$

This is an infinite dimensional optimization problem over the space of probability measures $\nu \in \mathcal{P}(\mathcal{X})$ subject to the constraint $\nu \in \mathcal{D} = \{\mu(x') : \|x' - x\| \leq \epsilon\}$. While this set is still intractable to deal with, we can consider relaxations of this set defined by divergence constraints between ν and $\rho = \mu(x)$, i.e., $\mathcal{D} \subseteq \{\nu : D(\nu \| \rho) \leq \epsilon_D\}$ where D denotes some divergence between probability distributions. We will show in Section 3 that for several commonly used divergences (in fact, for any f-divergence cf. Ali & Silvey, 1966), the relaxed problem can be solved efficiently.

2.1 A GENERAL FRAMEWORK FOR ROBUST VERIFICATION

In order to make the above setting more general, instead of speaking only about binary classifiers, we focus on a *specification* $\phi : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}$: a generic function over the input space that we want to verify has certain properties. Unless otherwise specified, we will assume that $\mathcal{X} \subseteq \mathbb{R}^d$ (we work in a d dimensional input space).

In the binary classification setting, ϕ could simply be the binary classifier itself, but it can also cover other settings such as probabilistic classifiers, the multi-class setting, etc. Our framework also

involves a reference measure ρ (in the above example we would take $\rho = \mu(x)$) and a collection of perturbed distributions $\mathcal{D}$ (in the above example we would take $\mathcal{D} = \{\mu(x') : \|x' - x\| \leq \epsilon\}$).

Using these ingredients we postulate two closely related certification problems: *robust certification* and *information-limited robust certification*. In the former case, we are given black-box access to a specification ϕ and are asked to verify that it obeys some property for all probability distributions ν within some class. In the latter case, we are given only knowledge of certain expectations of ϕ under the reference distribution ρ, and are asked to verify that *any* specification ϕ that agrees with these expectations is robustly certified (in this case we restrict the definition to ternary-valued specifications for technical convenience). Although the information-limited case may seem more challenging because we need to provide guarantees that hold simultaneously over a whole class of specifications, it turns out that, for perturbation sets $\mathcal{D}$ specified by an f-divergence bound, both certification tasks can be solved efficiently using convex optimization.

Definition 2.1 (Robust certification)**.** Let $\phi : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}$ be a *specification* and $\rho \in \mathcal{P}(\mathcal{X})$ a *reference distribution* such that $\mathrm{E}_{X\sim\rho}[\phi(X)] \geq 0$. Given a collection of *perturbed distributions* $\mathcal{D} \subset \mathcal{P}(\mathcal{X})$ containing ρ, we say that ϕ is *robustly certified* at ρ with respect to $\mathcal{D}$ if for any $\nu \in \mathcal{D}$ we have $\mathrm{E}_{X\sim\nu}[\phi(X)] \geq 0$.

Verifying that a given specification ϕ is robustly certified is equivalent to checking whether the optimal value of the optimization problem

$$\mathrm{OPT}(\phi, \rho, \mathcal{D}) := \min_{\nu\in\mathcal{D}} \mathop{\mathrm{E}}_{X\sim\nu}[\phi(X)] \ , \tag{4}$$

satisfies $\mathrm{OPT}(\phi, \rho, \mathcal{D}) \geq 0$. Solving problems of this form is the key workhorse of our general framework for black-box certification of adversarial robustness for smoothed classifiers.

When faced with Eq. 4, a natural question to ask is how much information about ϕ, in the sense of black-box queries, is required to solve the verification problem. This motivates us to consider an information-limited scenario where a potential verification algorithm only has access to the distribution of $\phi(X)$ under the reference distribution $X \sim \rho$. Specifically, in many cases of practical interest (including verification of smoothed classifiers, cf. Section 2.2) it suffices to verify ternary-valued specifications taking values in $\{-1, 0, +1\}$, in which case we hope to find solutions to the verification problem that only depend on the probabilities $\theta_a = \mathbb{P}_{X\sim\rho}[\phi(X) = +1]$ and $\theta_b = \mathbb{P}_{X\sim\rho}[\phi(X) = -1]$. This motivates the following definition.

Definition 2.2 (Information-limited robust certification)**.** Given a *reference distribution* $\rho \in \mathcal{P}(\mathcal{X})$ and probabilities $0 \leq \theta_b \leq \theta_a \leq 1$ (with $\theta_a + \theta_b \leq 1$), define the class of specifications[1]

$$S = \left\{ \phi : \mathcal{X} \to \{-1, 0, +1\} \text{ such that } \mathop{\mathbb{P}}_{X\sim\rho}[\phi(X) = +1] \geq \theta_a, \mathop{\mathbb{P}}_{X\sim\rho}[\phi(X) = -1] \leq \theta_b \right\} \ . \tag{5}$$

For any collection of *perturbed distributions* $\mathcal{D} \subset \mathcal{P}(\mathcal{X})$ containing ρ we say that S is *information-limited robustly certified at ρ with respect to $\mathcal{D}$* if the following conditions hold:

$$\mathop{\mathrm{E}}_{X\sim\nu}[\phi(X)] \geq 0 \quad \forall \nu \in \mathcal{D}, \phi \in S \ .$$

2.2 Adversarial specification for smoothed classifiers

We first note that the definitions above are sufficient to capture the standard usage of randomized smoothing as it has been used in past work (e.g. Lecuyer et al., 2018; Cohen et al., 2019) to verify the robustness of smoothed multi-class classifiers. Specifically, consider smoothing a classifier $h : \mathcal{X} \to \mathcal{Y}$ with a finite set of labels $\mathcal{Y}$ using a smoothing measure $\mu : \mathcal{X} \mapsto \mathcal{P}(\mathcal{X})$. The resulting randomly smoothed classifier h_s is defined in Eq. 1. Our goal is to certify that the prediction $h_s(x)$ is robust to perturbations of size at most ϵ measured by distance function[2] $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$, i.e.,

$$h_s(x') = h_s(x) \quad \forall x' \text{ such that } d(x, x') \leq \epsilon \ . \tag{6}$$

To pose this question within our framework, we choose the reference distribution $\rho = \mu(x)$, the set of perturbed distributions $\mathcal{D}_{x,\epsilon} = \{\mu(x') : d(x, x') \leq \epsilon\}$, and the following specifications. Let

[1] Note that by construction we have $\mathrm{E}_{X\sim\rho}[\phi(X)] \geq 0$ for all $\phi \in S$.

[2] d is an arbitrary distance function (not necessarily a metric e.g. ℓ_0).

$c = h_s(x)$. For every $c' \in \mathcal{Y} \setminus \{c\}$, we define the specification $\phi_{c,c'} : \mathcal{X} \mapsto \{-1, 0, +1\}$ as follows:

$$\phi_{c,c'}(x) = \begin{cases} +1 & \text{if } h(x) = c \ , \\ -1 & \text{if } h(x) = c' \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Then, Eq. 6 holds if and only if every $\phi_{c,c'}$ is robustly certified at $\mu(x)$ with respect to $\mathcal{D}_{x,\epsilon}$ (see Appendix A.1). This can be extended to soft or probabilistic classifiers as well, as shown in Appendix A.2.

2.3 Constraint sets from f-divergences

Dealing with the set $\mathcal{D}_{x,\epsilon}$ directly is difficult due to its possibly non-convex geometry. In this section, we discuss specific relaxations of this set, i.e., choices for sets $\mathcal{D}$ such that $\mathcal{D}_{x,\epsilon} \subseteq \mathcal{D}$ that are easier to optimize over. In particular, we focus on a general family of constraint sets defined in terms of f-divergences. These divergences satisfy a number of useful properties and include many well-known instances (e.g. relative entropy, total variation); see Appendix A.3 for details.

Definition 2.3. (f-divergence constraint set). Given $\rho, \nu \in \mathcal{P}(\mathcal{X})$, their f-divergence is defined[3] as

$$D_f(\nu \| \rho) = \mathop{\mathrm{E}}_{X \sim \rho}\left[f\left(\frac{\nu(X)}{\rho(X)} \right) \right] \ ,$$

where $f : \mathbb{R}_+ \mapsto \mathbb{R}$ is a convex function with $f(1) = 0$. Given a reference distribution ρ, an f-divergence D_f and a bound $\epsilon_f \geq 0$, we define the *f-divergence constraint set* to be:

$$\mathcal{D}_f = \{\nu \in \mathcal{P}(\mathcal{X}) : D_f(\nu \| \rho) \leq \epsilon_f\} \ .$$

Relaxations using f-divergence This construction immediately allows us to obtain relaxations of $\mathcal{D}_{x,\epsilon}$. For example, by choosing $f(u) = u \log(u)$, we have the KL divergence. Using KL-divergence yields the following relaxation between norm-based and divergence-based constraint sets for Gaussian smoothing measures, i.e. $\mu(x) = \mathcal{N}(x, \sigma^2 I)$:

$$\mathcal{D}_{x,\epsilon} = \{\mu(x') : \|x - x'\|_2 \leq \epsilon\} \subseteq \left\{\nu : \mathrm{KL}(\nu \| \mu(x)) \leq \epsilon^2/(2\sigma^2)\right\} \ .$$

Tighter relaxations can be constructed by combining multiple divergence-based constraints. In particular, suppose $\mathcal{F}$ is a collection of convex functions each defining an f-divergence, and assume each $f \in \mathcal{F}$ has a bound ϵ_f associated with it. Then we can define the constraint set containing perturbed distributions where all the bounds hold simultaneously (Fig. 1):

$$\mathcal{D}_{\mathcal{F}} := \bigcap_{f \in \mathcal{F}} \mathcal{D}_f = \{\nu : \forall\, f \in \mathcal{F} \ \ D_f(\nu \| \rho) \leq \epsilon_f\} \ .$$

Concrete verification algorithms based on these strategies will be discussed in Section 3.

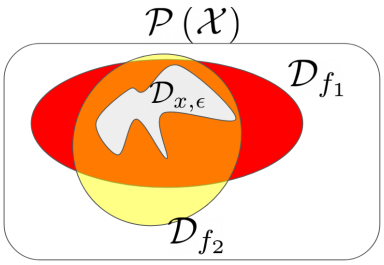


Figure 1: Intersecting f-divergence constraints to obtain better relaxations $\mathcal{D}_{\mathcal{F}}$ (depicted by the orange region) of $\mathcal{D}_{x,\epsilon}$.

3 Efficient black-box verification with f-divergences

We now show how the black-box verification problem (cf. Eq. 4) can be solved efficiently for the case of f-divergences. This allows us, by extension, to solve the problem for related divergences like Rényi divergences. Before we state the result, we introduce the notion of a convex conjugate: for any function $f : \mathbb{R}_+ \mapsto \mathbb{R}$, its *convex conjugate* is defined as

$$f^*(u) = \max_{v \geq 0} \left(uv - f(v)\right) \ .$$

[3]This definition should technically use the Radon-Nikodym derivative of the measure ν with respect to ρ, but we ignore measure-theoretic issues in this paper for simplicity of exposition. For continues distributions, ν and ρ should be treated as densities, and for discrete distributions as probability mass functions.

The following two theorems provide the main foundation for the verification procedures in the paper. They show that in the case of a constraint set of the form $\mathcal{D}_{\mathcal{F}}$ with $\mathcal{F} = \{f_1, \ldots, f_M\}$ and $\epsilon_{f_i} = \epsilon_i$, we can verify robust certification and information-limited robust certification, respectively, using a simple optimization procedure.

Theorem 1 (Verifying robust certification)**.** Define $f_\lambda(u) = \sum_{i=1}^M \lambda_i f_i(u)$ and denote its convex conjugate by f_λ^*. The specification ϕ is robustly certified at ρ with respect to $\mathcal{D}_{\mathcal{F}}$ *if and only if* the optimal value of the following optimization problem is non-negative:

$$\max_{\lambda \geq 0, \kappa} \kappa - \sum_{i=1}^M \lambda_i \epsilon_i - \mathop{\mathrm{E}}_{X \sim \rho} [f_\lambda^*(\kappa - \phi(X))] \ . \tag{7}$$

We note that the special case where $M = 1$ reduces to Proposition 1 of Duchi & Namkoong (2018), although the result is used in a completely different context in that work.

Our next main theorem concerns the extension of this verification procedure to the information-limited setting. Although it may seem more challenging to verify robustness over both all $\nu \in \mathcal{D}_{\mathcal{F}}$ *and* all specifications ϕ that obey $\theta_a \leq \mathbb{P}_{X \sim \rho}[\phi(X) = +1]$ and $\theta_b \geq \mathbb{P}_{X \sim \rho}[\phi(X) = -1]$, it turns out that the latter can also be accomplished via a convex optimization problem:

Theorem 2. The class of specifications S in Definition 2.2 is information-limited robustly certified at ρ with respect to $\mathcal{D}_{\mathcal{F}}$ *if and only if* the optimal value of the following three-variable convex optimization problem is non-negative:

$$\min_{\zeta_a, \zeta_b, \zeta_c \geq 0} \quad \zeta_a - \zeta_b \tag{8a}$$

$$\text{Subject to } \zeta_a + \zeta_b + \zeta_c = 1 \ , \quad D_{f_i}(\zeta \| \theta) \leq \epsilon_i \quad i = 1, \ldots, M \ , \tag{8b}$$

where $\theta = (\theta_a, \theta_b, 1 - \theta_a - \theta_b)$.

Theorem 2 has an intuitive interpretation: We consider the space of distributions over the possible values of the specification $\{-1, 0, +1\}$ and search for one which is within a distance of ϵ_i from θ with respect to the f_i-divergence looking for one that has the smallest expected value. If there exists one that has a negative expected value, robust certification does not hold. This is a rather surprising result, since it says that f-divergence constraints over the input space $\mathcal{P}(\mathcal{X})$ can simply be translated to f-divergence constraints over the outputs $\mathcal{P}(\{-1, 0, +1\})$.

The proof of Theorem 1 uses standard duality results to show that the dual of the verification optimization problem has the desired form. The proof of Theorem 2 rests on the fact that for this form of optimal solution, it is possible to directly compute the expectation in Eq. 7, and in fact this expectation only depends on ϕ via the probabilities θ_a and θ_b. Full proofs are given in Appendices A.4 and A.5.

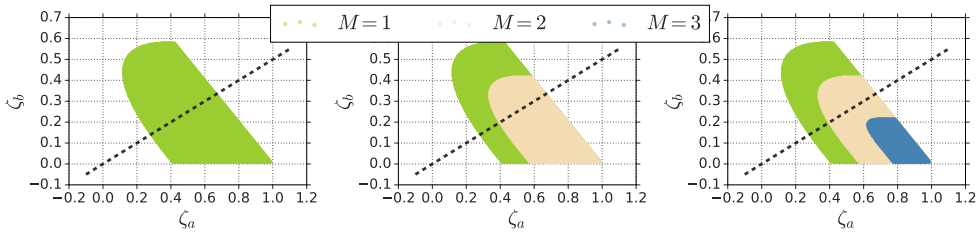


Figure 2: Geometric interpretation of Eq. 8: We show the *intersection* of constraints of the form $\{R_\alpha(\nu \| \rho) \leq \epsilon_\alpha\}$ using several Rényi divergences. For each α, we compute the worst case Rényi divergence over the set $\mathcal{D} = \{\mu(x') : \|x' - x\|_0 \leq \epsilon\}$ (i.e., the set of smoothing measures from ℓ_0 perturbations of a nominal point) and choose ϵ_α to be this value (we choose μ to be discrete smoothing measure described in appendix A.10). We then show the set $\{R_\alpha(\nu \| \rho) \leq \epsilon_\alpha\}$ ($\alpha = 2.5, M = 1$) depicted as the green region, the set set $\cap_{\alpha \in \{2.5, 3.0\}} \{R_\alpha(\nu \| \rho) \leq \epsilon_\alpha\}$ ($M = 2$, brown region in the middle figure) and $\cap_{\alpha \in \{2.5, 3.0, 5.0\}} \{R_\alpha(\nu \| \rho) \leq \epsilon_\alpha\}$($M = 3$, blue region on the right figure). The dashed line shows $\zeta_a - \zeta_b = 0$, so that the first two sets are not certified but the third set is.

Divergence constraint	$f(u)$	Certificate
KL divergence $\mathrm{KL}(\nu\|\rho) \leq \epsilon_{KL}$	$u\log(u)$	$\epsilon_{KL} \leq -\log\Big(1 - \big(\sqrt{\theta_a} - \sqrt{\theta_b}\big)^2\Big)$
Rényi divergences ($\alpha \geq 0$) $R_\alpha(\nu\|\rho) \leq \epsilon_{R,\alpha}$	$\mathrm{sign}(\alpha-1)(u^\alpha - 1)$	$\epsilon_{R,\alpha} \leq -\log(1 - \theta_a - \theta_b + 2\eta)$ $\eta = \left(\frac{{\theta_a}^{(1-\alpha)} + {\theta_b}^{(1-\alpha)}}{2}\right)^{\left(\frac{1}{1-\alpha}\right)}$
Infinite Rényi divergence $R_\infty(\nu\|\rho) \leq \epsilon_{R,\infty}$	N/A	$\epsilon_{R,\infty} \leq -\log(1 - (\theta_a - \theta_b))$
Hockey-stick divergences ($\beta \geq 0$) $D_{\mathrm{HS},\beta}(\nu\|\rho) \leq \epsilon_{\mathrm{HS},\beta}$	$[u-\beta]_+ - [1-\beta]_+$	$\epsilon_{\mathrm{HS},\beta} \leq \left[\frac{\beta(\theta_a - \theta_b) - \lvert\beta - 1\rvert}{2}\right]_+$

Table 1: Certificates for various f-divergences. Note that the Rényi divergences are not proper f-divergences, but are defined as $R_\alpha(\nu\|\rho) = \frac{1}{\alpha-1}\log(1 + D_f(\nu\|\rho))$. The infinite Rényi divergence, defined as $\sup_x \log(\nu(x)/\mu(x))$, is obtained by taking the limit $\alpha \to \infty$. All certificates depend on the gap between θ_a and θ_b. Notation: $[u]_+ = \max(u, 0)$.

3.1 Applications

Closed-form certificates for hard classifiers To verify the ternary-valued specifications $\phi_{c,c'}$ corresponding to the robustness of μ-smoothing a classifier h (cf. Section 2.2) it suffices to consider the top two labels c and c' with:

$$c = \arg\max_y \mathop{\mathbb{P}}_{X\sim\mu(x)}[h(X) = y] \;, \quad \theta_a = \mathop{\mathbb{P}}_{X\sim\mu(x)}[\phi_{c,c'}(X) = +1] = \mathop{\mathbb{P}}_{X\sim\mu(x)}[h(X) = c] \;,$$
$$c' = \arg\max_{y\neq c} \mathop{\mathbb{P}}_{X\sim\mu(x)}[h(X) = y] \;, \quad \theta_b = \mathop{\mathbb{P}}_{X\sim\mu(x)}[\phi_{c,c'}(X) = -1] = \mathop{\mathbb{P}}_{X\sim\mu(x)}[h(X) = c'] \;.$$

We are thus in the limited information setting and the verification reduces to a problem of the form Eq. 8. For this case, Table 1 provides simple closed-form certificates for several commonly used f-divergences (i.e. for constraint sets $\mathcal{F}$ with $M = |\mathcal{F}| = 1$).

Norm-based smoothing measures Our framework can be used whenever the smoothing measures are such that: (1) $X \sim \mu(x)$ can be sampled efficiently, and (2) $\max_{\{x':d(x,x')\leq\epsilon\}} D_f(\mu(x')\|\mu(x))$ can be computed or bounded efficiently for one or more f-divergences. In the case where the distance function is induced by some norm $d(x, x') = \|x - x'\|$, a natural choice is to take smoothing measures with density $\mu(x)[z] \propto \exp(-\|z - x\|)$. Such smoothing measures are additive in the sense that $X \sim \mu(x)$ satisfies $X = x + Z$ with $Z \sim \mu(0)$; this reduces the sampling question to sampling from $\mu(0)$. Furthermore, it can be shown using the triangle inequality that for any such distribution, we have $R_\infty(\mu(x')\|\mu(x)) = \|x' - x\|$. It is thus straightforward to compute certificates under infinite Rényi divergences for any such measure.

Lemma 6 in Appendix A.8 shows that we can also efficiently compute bounds for a large family of f-divergences under several norms of interest (including $\ell_1, \ell_2, \ell_\infty$ vector norms as well as the matrix nuclear and spectral norms). The same appendix also outlines efficient procedures to sample from the corresponding smoothing measures. A similar approach can also deal with discrete perturbations (i.e. induced by the ℓ_0 pseudo-norm) as shown in Appendix A.10.

3.2 Information-limited robust certification and tight relaxations

Ideally, we would like to certify robustness of specifications (e.g. $\phi_{c,c'}$ for smoothed classifiers) with respect to sets of the form $\mathcal{D}_{x,\epsilon} = \{\mu(x') : d(x, x') \leq \epsilon\}$. The following result shows that the gap between the ideal $\mathcal{D}_{x,\epsilon}$ and the tractable constraint sets $\mathcal{D}_\mathcal{F}$ can be closed *in the context of information-limited robust certification* provided that we can measure hockey-stick divergences of *every* non-negative order $\beta \geq 0$. The proof is given in Appendix A.7.

Theorem 3. Let $\theta_a, \theta_b, \rho, \mathcal{D}, S$ be as in Definition 2.2. Let $\epsilon_\beta = \max_{\nu\in\mathcal{D}} D_{\mathrm{HS},\beta}(\nu\|\rho)$ for $\beta \geq 0$ and define the constraint set $\mathcal{D}_{\mathrm{HS}} = \cap_{\beta\geq 0}\{\nu \in \mathcal{P}(\mathcal{X}) : D_{\mathrm{HS},\beta}(\nu\|\rho) \leq \epsilon_\beta\}$. Then, S is information-limited robustly certified at ρ with respect to $\mathcal{D}$ *if and only if* S is information-limited robustly certified at ρ with respect to $\mathcal{D}_{\mathrm{HS}}$.

4 CONNECTIONS WITH PRIOR WORK

Connection with prior work on information-limited black-box verification Cohen et al. (2019) study the problem of verifying hard classifiers smoothed by Gaussian noise, and derive optimal certificates with respect to ℓ_2 perturbations of the input. Their results can be recovered as a special case of our framework when applied to sets defined via constraints on hockey-stick divergences (which can be computed efficiently for Gaussian measures). More concretely, in Corollary 5 (Appendix A.7), we show that the results of Cohen et al. (2019) can be recovered as a special case of our framework by applying the theorem 3 to a Gaussian smoothing measure. Lee et al. (2019) study the problem of strict black-box verification with general smoothing measures. However, they are only able to derive certificates under the assumption that the likelihood ratio between measures $\frac{\nu(X)}{\rho(X)}$ only takes a finite set of values. This is a restrictive assumption that prevents the authors from accommodating natural smoothing measures like Gaussian or Laplacian measures. Further, the complexity of computing the certificates in their framework is often significant – they require an $O(d^3)$ computation (where d is the input dimension) to certify smoothness to ℓ_0 perturbations (in addition to the cost of estimating θ_a, θ_b by sampling). On the other hand, our formulation provides a straightforward method to compute certificates for ℓ_0 perturbations using Table 1. Finally, these works are both restricted to the information-limited black-box verification setting where only θ_a, θ_b are known. On the other hand, our formulation can compute improved certificates when more information about ϕ is available (see Section 5.1).

Connections with pixel differential privacy Lecuyer et al. (2018) introduced the notion of *pixel differential privacy* (pixelDP) to study robustness of smoothed classifiers: a distribution-valued function $G : \mathbb{R}^d \to \mathcal{P}(\mathcal{Z})$ satisfies (ε, τ)-*pixelDP with respect to* ℓ_p *perturbations* if for any $\|x - x'\|_p \leq 1$ it holds that $D_{\mathrm{DP},e^\varepsilon}(G(x)\|G(x')) \leq \tau$, where

$$D_{\mathrm{DP},e^\varepsilon}(G(x)\|G(x')) = \sup_E \left(\mathop{\mathbb{P}}_{X\sim G(x)}[X \in E] - e^\varepsilon \mathop{\mathbb{P}}_{X'\sim G(x')}[X' \in E] \right) \tag{9}$$

and the supremum is over all (measurable) subsets E of $\mathcal{Z}$. In particular, Lecuyer et al. show that using a smoothing measure μ satisfying pixelDP with respect to a certain type of perturbations leads to adversarially robust classifiers. Since $D_{\mathrm{HS},e^\varepsilon}$ is in fact equal to the hockey-stick divergence of order $\beta = e^\varepsilon$ (Barthe & Olmedo, 2013), their results can be directly expressed in our framework as follows. Take $\rho = \mu(x)$ and for fixed $\varepsilon \geq 0$ and $\tau \in [0, 1]$ define the set of perturbed distributions[4]

$$\mathcal{D}_{\varepsilon,\tau} = \{\nu : D_{\mathrm{DP},e^\varepsilon}(\nu\|\rho) \leq \tau \text{ and } D_{\mathrm{DP},e^\varepsilon}(\rho\|\nu) \leq \tau\} \ . \tag{10}$$

It immediately follows that if μ satisfies (ε, τ)-pixelDP with respect to ℓ_p perturbations, then we have the relaxation condition $\{\mu(x') : \|x - x'\|_p \leq 1\} \subseteq \mathcal{D}_{\varepsilon,\tau}$. From this point of view, the main result from (Lecuyer et al., 2018) applied to smoothed hard classifiers[5] yields the information-limited black-box certificate

$$\tau \leq \frac{\theta_a - e^{2\varepsilon}\theta_b}{e^\varepsilon + 1} \ . \tag{11}$$

For comparison, the certificate obtained by our method (HS certificate in Table 1) for the relaxation $\{\nu : D_{\mathrm{DP},e^\varepsilon}(\nu\|\rho) \leq \tau\}$ of $\mathcal{D}_{\varepsilon,\tau}$ already improves on the certificate by Lecuyer et al. whenever $\theta_a - \theta_b \geq (e^\varepsilon - 1)(1 - \theta_a - \theta_b)$, which, for example, is always true in the binary classification case. Since Theorem 2 provides optimal certificates for the full constraint set $\mathcal{D}_{\varepsilon,\tau}$, we have the following.

Corollary 4. The optimal certificates for the constraint set $\mathcal{D}_{\varepsilon,\tau}$ (cf. Eq. 10) obtained from Theorem 2 are stronger than those obtained from Eq. 11.

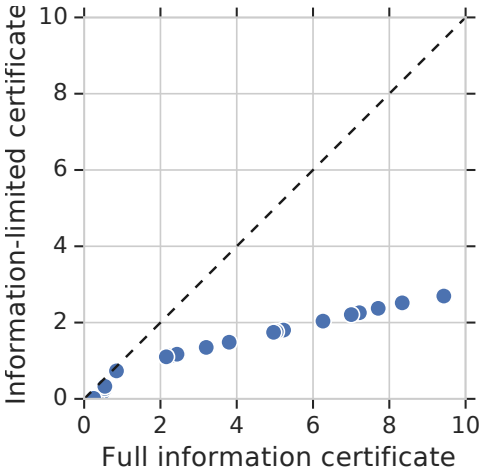
5 EXPERIMENTS

5.1 INFORMATION-LIMITED VS FULL-INFORMATION

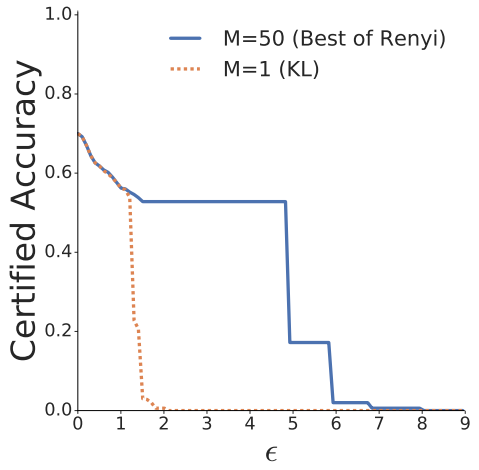
For ImageNet we trained a classifier using ResNet-152 (He et al., 2016) and tested our full-information certificate methodology on 50 randomly selected examples from the test set. We compare using our

[4]Closure of f-divergences under reversal implies that $\mathcal{D}_{\varepsilon,\tau}$ can be written in the form $\mathcal{D}_\mathcal{F}$ (cf. Section 2.3).

[5]Although Lecuyer et al. (2018) state their main result for soft classifiers, the extension to hard classifiers is straightforward.



(a) Plot showing the maximum perturbation radius we can provide a certificatre for ℓ_2 perturbations.



(b) Comparison of certified accuracy using $M = 50$ (50 Rényi divergences) vs. $M = 1$ for ℓ_0 perturbations.

Figure 3: Comparison between full-information vs. information-limited certificates, and one vs. many f-divergences in $\mathcal{D}_{\mathcal{F}}$ for ImageNet.

full-information methodology vs. limited-information for ℓ_2 perturbations. The results are shown in Figure 3a. Our full-information certificate can be provided for significantly larger perturbation radii compared to using the information-limited certificate from (Cohen et al., 2019). In particular, we are able to provide a certificate for perturbation radius $\epsilon = 9.42$ in the full-information case whilst the limited-information certificate can only be provided for perturbation radius $\epsilon = 2.69$. This is a substantial difference demonstrating the significance of using full-information.

5.2 $M = 1$ vs. $M > 1$

Another advantage of our framework is the ability to consider several f-divergences simultaneously. We plot the results comparing the certificate derived from just the KL divergence ($\alpha = 1$) vs. using several Rényi divergences ($\alpha \in [1.1, 80]$) in Figure 3b. Here we show that when we increase the number of Rényi divergences to $M = 50$, we can maintain a certifiable accuracy of 52.8%[6] for an ℓ_0 radius of $\epsilon = 5.0$. In contrast, when we use only KL divergence the certified accuracy quickly drops to 0% around $\epsilon = 2.0$.

5.3 Experiments on Certifiable Accuracy

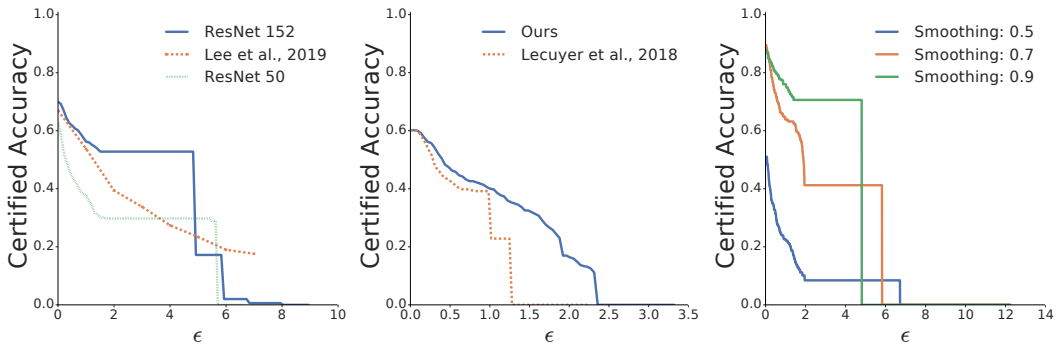
Our experimental protocol closely follows (Cohen et al., 2019). First, a classifier is trained using data augmentation by including samples from the smoothing measure on each original training point. Then, at prediction time, we use the same smoothing measure to create a smoothed classifier for which we compute robustness certificates. We then compute the fraction of test examples that are correctly classified and certified to be robust to a given value of ϵ. We show results for several values of the smoothing parameter controlling the amount of noise added while smoothing; this induces a natural trade-off between accuracy and robustness. For the ℓ_0 we use the discrete smoothing measure from appendix A.10 and for ℓ_1, we use the Laplacian smoothing measure.

[6]The constant line at a certified accuracy of 52.8% is an artifact of the design of the certification procedure. During certification, our certificate is specified by only two numbers, (1) the number of times our guess for the true label appeared as the top-class, (2) the number of times our guess for the second-most likely class appeared as the top-class. If, given two inputs, these numbers are identical, they will produce the same certificate. A large number of inputs during certification have identical values for (1) and (2). For example, at small smoothing values it is the case that the guess for the true label was predicted as the top-class in all queries used for certification.

We report the certificates computed using the "Best of Rényi" approach: We compute the certificate from Table 1 for Rényi divergences over a range[7] of values of α and pick the best one. The probabilities θ_a and θ_b are estimated using $100K$ samples from the smoothing measure for LibriSpeech and ImageNet, 1 million for CIFAR10, and 10 million for MNIST. The confidence bound, i.e. the probability the certificate does not fail, is set to 0.99 (Lecuyer et al. (2018) use a value of 0.95). For MNIST, the results are summarized in Table 2, where our results significantly outperform those reported in (Lee et al., 2019).

Table 2: MNIST

Certificate	Norm	Certified Accuracy						
		$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$	$\epsilon = 4$	$\epsilon = 5$	$\epsilon = 6$	$\epsilon = 7$
Lee et al. (2019)	ℓ_0	0.921	0.774	0.539	0.524	0.357	0.202	0.097
Ours	ℓ_0	0.954	0.905	0.856	0.808	0.772	0.738	0.699
Lecuyer et al. (2018)	ℓ_1	0.772	0.548	0.424	0.061	0	0	0
Ours	ℓ_1	0.860	0.716	0.584	0.447	0.325	0.201	0.017



(a) ImageNet: ℓ_0 robustness. (b) ImageNet: ℓ_1 robustness (c) Librispeech: ℓ_0 robustness

Figure 4: Certified accuracy on ImageNet & Librispeech. ℓ_0 ImageNet results from (Lee et al., 2019) are taken from their paper.

The results for ImageNet are plotted in Figure 4a. For the ℓ_0 norm, our results are not as strong as the ones from (Lee et al., 2019). However, the certification algorithm used there has a significant computational cost that scales cubically in the input dimension, while our certification procedure is constant time (after the values θ_a and θ_b have been estimated).

We also obtain results on an audio classification task, Librispeech (Panayotov et al., 2015), where the task is to learn identify the speaker given a dataset of recordings from ten speakers. We created ℓ_0 perturbations by replacing part of the audio signal with silence (with the ℓ_0 perturbation controlling the fraction of time-stamps at which the signal is zeroed out). The results are plotted in Figure 4c.

6 Conclusion

We have introduced a general framework for black-box verification that utilizes multiple f-divergences and provides state-of-the-art results on both image classification and audio tasks by a significant margin. We theoretically demonstrate that with a sufficient number of f-divergences we can obtain tight relaxation for arbitrary perturbation sets in an information-limited setting. Empirically, we have also show the advantages of being able to bound multiple f-divergences and using full-information certificate.

[7]We use 50 values $\alpha \in [1.1, 80]$ for ℓ_0 robustness and 25 values $\alpha \in [2, 10]$ for ℓ_1 robustness.

REFERENCES

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28 (1):131–142, 1966.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pp. 403–412, 2018.

Gilles Barthe and Federico Olmedo. Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs. In *International Colloquium on Automata, Languages, and Programming*, pp. 49–60. Springer, 2013.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 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.

Sebastian Bubeck. Orf523 (advanced optimization): Introduction. *https://blogs.princeton.edu/imabandit/2013/02/05/orf523-advanced-optimization-introduction/*, 2013.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Krishnamuthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy Mann, and Pushmeet Kohli. Towards scalable verification of neural networks: A dual approach. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Machine Learning*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. *arXiv preprint arXiv:1909.01492*, 2019.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986*, 2019.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pp. 97–117. Springer, 2017.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*, 2018.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi S Jaakkola. A stratified approach to robustness for randomly smoothed classifiers. *arXiv preprint arXiv:1906.04948*, 2019.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, April 2015. doi: 10.1109/ICASSP.2015.7178964.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Bys4ob-Rb`.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2016.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, 2018.

Eric Wong and J Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. *International Conference on Machine Learning*, 2018.

Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, pp. 8400–8409, 2018.

A APPENDIX

A.1 ADVERSARIAL SPECIFICATION FOR SMOOTHED CLASSIFIERS

Note that for any $\nu \in \mathcal{D}_{x,\epsilon}$ we have

$$\mathop{\mathrm{E}}_{X\sim\nu}[\phi_{c,c'}(X)] = \mathop{\mathbb{P}}_{X\sim\nu}[h(X) = c] - \mathop{\mathbb{P}}_{X\sim\nu}[h(X) = c'] \ .$$

Therefore, $\mathrm{E}_{X\sim\nu}[\phi_{c,c'}(X)] \geq 0$ for all $c' \in \mathcal{Y} \setminus \{c\}$ is equivalent to $c \in \arg\max_{y\in\mathcal{Y}} \mathbb{P}_{X\sim\nu}[h(X) = y]$. For $\nu = \mu(x')$, this means that $h_s(x') = c$ (assuming the argmax is unique). In other words, $\mathrm{E}_{X\sim\nu}[\phi_{c,c'}(X)] \geq 0$ for all $c' \in \mathcal{Y} \setminus \{c\}$ and all $\mu(x') \in \mathcal{D}_{x,\epsilon}$ if and only if $h_s(x') = c$ for all x' such that $d(x, x') \leq \epsilon$, proving the required robustness certificate.

A.2 SPECIFICATIONS FOR ROBUSTNESS OF SOFT CLASSIFIERS

Consider a soft classifier $H : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ that for each input x returns a probability distribution $H(x)$ over the set of potential labels $\mathcal{Y}$ (e.g. H might represent the outputs of the soft-max layer of a neural network). As in the case of hard classifiers, our methodology can be used to provide robustness guarantees for smoothed soft classifiers obtained by applying a smoothing measure $\mu(x)$ to the input. In this case, the smoothed classifier is again a soft classifier given by $H_s(x) = \mathrm{E}_{X\sim\mu(x)}[H(X)]$.

Let x be a fixed input point and write $p = H_s(x) \in \mathcal{P}(\mathcal{Y})$ to denote the distribution over labels. A number of robustness properties about the soft classifier $\tilde{H}$ at x can be phrased in terms of Definition 2.1. For example, suppose $\mathcal{Y} = \{1, \ldots, K\}$ and assume, without loss of generality, that $p_1 \geq p_2 \geq \cdots \geq p_K$ so that $\{1, \ldots, k\}$ are the top k labels at x. Then we can verify that the set of top k labels will not change when moving the input from x to x' with $\|x - x'\| \leq \epsilon$ by defining the specifications $\phi_{i,j}(z) = H(z)_i - H(z)_j$ for $i \in [1, k]$ and $j \in [k + 1, K]$, and showing that all of these $\phi_{i,j}$ are robustly certified at $\mu(x)$ with respect to the set $\mathcal{D}_{x,\epsilon}$ defined above. Note that the case $k = 1$ corresponds to robustness of the standard classification rule which outputs the label with the largest score. Another example is robustness of classifiers which are allowed to abstain. For example, suppose we build a hard classifier $\tilde{h}$ out of H_s which returns the label with the maximum score as long as the gap between this score and the score of any other label is at least γ; otherwise it produces no output. Then we can certify that $\tilde{h}$ will not abstain and return the label $c = \arg\max_{y\in\mathcal{Y}} p_y$ at any point close to x by showing that every $\phi_{c'}(z) = H(z)_c - H(z)_{c'} - \gamma$, $c' \neq c$, is robustly certified at $\mu(x)$ with respect to $\mathcal{D}_{x,\epsilon}$.

A.3 BACKGROUND ON f-DIVERGENCES

A number of well-known properties about f-divergences are used throughout the paper, both explicitly and implicitly. Here we review such properties for the readers' convenience. Proofs and further details can be found in, e.g., (Csiszár et al., 2004; Liese & Vajda, 2006).

Recall that the f-divergences can be defined for any convex function $f : \mathbb{R}_+ \to \mathbb{R}$ such that $f(1) = 0$. We note that this requirement holds without loss of generality as the map $x \mapsto f(x) - f(1)$ is convex whenever f is convex. Any f-divergence D_f satisfies the following:

1. $D_f(\nu\|\rho) \geq 0$.
2. $D_f(\rho\|\rho) = 0$, and $D_f(\nu\|\rho) = 0$ implies $\nu = \rho$ whenever f is strictly convex at 1.
3. $D_f(F_*(\nu)\|F_*(\rho)) \leq D_f(\nu\|\rho)$ for any function F, where $F_*(\rho)$ is the push-forward of ρ.
4. $D_f(\nu\|\rho) = D_{\bar{f}}(\rho\|\nu)$ where $\bar{f}(u) = uf\left(\frac{1}{u}\right)$ is again convex with $\bar{f}(1) = 0$.

A.4 PROOF OF THEOREM 1

For simplicity of exposition (and to avoid measure theoretic issues), we focus on the case where ν, ρ have well defined densities $\nu(x), \rho(x)$ such that $\rho(x) > 0$ whenever $\nu(x) > 0$.

We begin by rewriting the optimization problem in terms of the likelihood ratio $r(X) = \frac{\nu(X)}{\rho(X)}$: We have

$$\mathop{\mathrm{E}}_{X\sim\nu}[\phi(X)] = \mathop{\mathrm{E}}_{X\sim\rho}[r(X)\phi(X)] \ , \quad D_{f_i}(\rho\|\nu) = \mathop{\mathrm{E}}_{X\sim\rho}[f_i(r(X))] \ , \quad \mathop{\mathrm{E}}_{\mathcal{X}\sim\rho}[r(X)] = 1 \ ,$$

where the first two equalities follow directly by plugging in $\nu(X) = \rho(X)r(X)$ and the third is obtained using the fact that ν is a probability measure. Using these relations, the optimization over ν can be rewritten as

$$\min_{r\geq 0} \mathop{\mathrm{E}}_{X\sim\rho}[r(X)\phi(X)] \tag{12a}$$

$$\text{Subject to } \mathop{\mathrm{E}}_{X\sim\rho}[f_i(r(X))] \leq \epsilon_i, \quad \mathop{\mathrm{E}}_{X\sim\rho}[r(X)] = 1 \ , \tag{12b}$$

where $r \geq 0$ denotes that $r(x) \geq 0 \quad \forall x \in \mathcal{X}$. The optimization over r is a convex optimization problem and can be solved using Lagrangian duality as follows – we first dualize the constraints on r to obtain

$$\begin{aligned}
&\min_{r\geq 0} \mathop{\mathrm{E}}_{X\sim\rho}[r(X)\phi(X)] + \sum_i \lambda_i \left(\mathop{\mathrm{E}}_{X\sim\rho}[f_i(r(X))] - \epsilon \right) + \kappa \left(1 - \mathop{\mathrm{E}}_{X\sim\rho}[r(X)] \right) \\
&= \min_{r\geq 0} \mathop{\mathrm{E}}_{X\sim\rho} \left[r(X)\phi(X) + \sum_i \lambda_i f_i(r(X)) - \kappa r(X) \right] + \kappa - \sum_i \lambda_i \epsilon \\
&= \kappa - \sum_i \lambda_i \epsilon_i - \mathop{\mathrm{E}}_{X\sim\rho} \left[\max_{r\geq 0} \kappa r - r\phi(X) - \sum_i \lambda_i f_i(r) \right] \\
&= \kappa - \sum_i \lambda_i \epsilon_i - \mathop{\mathrm{E}}_{X\sim\rho} \left[\max_{r\geq 0} \left(r(\kappa - \phi(X)) - f_\lambda(r) \right) \right] \\
&= \kappa - \sum_i \lambda_i \epsilon_i - \mathop{\mathrm{E}}_{X\sim\rho}[f_\lambda^*(\kappa - \phi(X))] \ .
\end{aligned}$$

By strong duality, it holds that maximizing the final expression with respect to $\lambda \geq 0, \kappa$ achieves the optimal value in Eq. 12a. Thus, if the optimal value is smaller than 0, the specification is not robustly certified and if it is larger than 0, the specification is robustly certified. This concludes the proof of correctness of the certificate Eq. 7.

A.5 Proof of theorem 2

For the next result, we observe that when ϕ is ternary valued, the optimization over κ, λ above can be written as

$$\max_{\kappa,\lambda\geq 0} \kappa - \sum_i \lambda_i \epsilon_i - \theta_a f_\lambda^*(\kappa - 1) - \theta_b f_\lambda^*(\kappa + 1) - \theta_c f_\lambda^*(\kappa) \ ,$$

where $\theta_a = \mathbb{P}_{X\sim\rho}[\phi(X) = +1], \theta_b = \mathbb{P}_{X\sim\rho}[\phi(X) = -1], \theta_c = \mathbb{P}_{X\sim\rho}[\phi(X) = 0]$.

Writing out the expression for f^*, we obtain

$$\begin{aligned}
\max_{\lambda\geq 0,\kappa} \min_{\gamma\geq 0} \kappa - \sum_i \lambda_i \epsilon_i - \theta_a \left((\kappa - 1)\gamma_a - \sum_i \lambda_i f_i(\gamma_a) \right) - \theta_b \left((\kappa + 1)\gamma_b - \sum_i \lambda_i f_i(\gamma_b) \right) \\
- \theta_c \left(\kappa\gamma_c - \sum_i \lambda_i f_i(\gamma_a) \right) \\
= \min_{\gamma\geq 0} \max_{\lambda\geq 0,\kappa} \kappa(1 - \theta_a\gamma_a - \theta_b\gamma_b - \theta_c\gamma_c) + \sum_i \lambda_i \left(\sum_{y\in\{a,b,c\}} \theta_y f_i(\gamma_y) - \epsilon_i \right) + \theta_a\gamma_a - \theta_b\gamma_b \ ,
\end{aligned}$$

where the second inequality follows from strong duality. The inner maximization is unbounded unless

$$\sum_{y\in\{a,b,c\}} \gamma_y\theta_y = 1 \ , \quad \sum_{y\in\{a,b,c\}} \theta_y f_i(\gamma_y) \leq \epsilon_i \ .$$

One thing to note is that, we can rewrite these constraints in terms of $\zeta = \theta \odot \gamma$, i.e. $\zeta_y = \theta_y \gamma_y$ for $y \in \{a, b, c\}$. These constraints ensure that ζ is a probability distribution over $\{+1, 0, -1\}$ and furthermore

$$\sum_{y \in \{a,b,c\}} \theta_y f_i(\gamma_y) = D_{f_i}(\zeta \| \theta) \ .$$

Thus, the second constraint above is equivalent to $D_{f_i}(\zeta \| \theta) \leq \epsilon_i$. Writing the optimization problem in terms of ζ, we obtain

$$\begin{aligned} \min_{\zeta_a, \zeta_b, \zeta_c \geq 0} \quad & \zeta_a - \zeta_b \\ \text{Subject to } & D_{f_i}(\zeta \| \theta) \leq \epsilon_i \quad i = 1, \ldots, M \ , \\ & \zeta_a + \zeta_b + \zeta_c = 1 \ . \end{aligned}$$

A.6 Calculation of closed-form certificates in table 1

We present the derivation of certificates for Hockey-Stick and Rényi divergences. The certificates for the KL and infinite Rényi divergence can be derived by taking limits of the Rényi certificate (as $\alpha \to 1, \infty$ respectively).

A.6.1 Calculation of certificate for hockey-stick divergence

The function $f(u) = \max(u - \beta, 0) - \max(1 - \beta, 0)$ is a convex function with $f(1) = 0$. Then, we have

$$\begin{aligned} f^*_\lambda(u) &= \max_{v \geq 0} \left(uv - \lambda \max(v - \beta, 0) \right) + \lambda \max(1 - \beta, 0) \\ &= \begin{cases} \max(\beta u, 0) + \lambda \max(1 - \beta, 0) & \text{if } u \leq \lambda \ , \\ \infty & \text{if } u > \lambda \ . \end{cases} \end{aligned}$$

The certificate given by Eq. 7 in Theorem 1 for this divergence in the case of a smoothed hard classifier takes the form

$$\max_{\kappa \in \mathbb{R}, \lambda \geq 0} \left(\kappa - \mathop{\mathrm{E}}_{X \sim \rho} [f^*_\lambda(\kappa - \phi(X))] \right) - \lambda \epsilon \geq 0 \ ,$$

where the specification takes the values

$$\phi(X) = \begin{cases} +1 & \text{w.p. } \theta_a \ , \\ -1 & \text{w.p. } \theta_b \ , \\ 0 & \text{w.p. } 1 - \theta_a - \theta_b \ . \end{cases}$$

Plugging in the expression for f^* the objective function above takes the form

$$\kappa - \beta \big(\theta_a [\kappa - 1]_+ + \theta_b [\kappa + 1]_+ + (1 - \theta_a - \theta_b)[\kappa]_+ \big) - \lambda(\epsilon + \max(1 - \beta, 0)) \ ,$$

where we use the notation $[u]_+ = \max(u, 0)$ and assumed the constraints $\kappa \leq \lambda - 1$ since the objective is $-\infty$ otherwise. If $\beta \leq 1$, the objective is increasing monotonically in κ, so the optimal value is to set κ to its upper bound $\lambda - 1$. Plugging this in, the possible values of the derivative with respect to λ are

$$\begin{cases} \beta(1 - \theta_b) - \epsilon & \text{if } 0 \leq \lambda < 1 \ , \\ \beta \theta_a - \epsilon & \text{if } 1 < \lambda < 2 \ , \\ -\epsilon & \text{if } \lambda > 2 \ . \end{cases}$$

Thus, if $\epsilon \leq \beta \theta_a$, the maximum is attained at 2, if $\beta \theta_a \leq \epsilon \leq \beta(1 - \theta_b)$, the maximum is attained at 1, else the maximum is attained at 0, leading to the certificate:

$$\begin{cases} -1 & \text{if } \epsilon \geq \beta(1 - \theta_b) \ , \\ \beta(1 - \theta_b) - \epsilon - 1 & \text{if } \beta \theta_a \leq \epsilon \leq \beta(1 - \theta_b) \ , \\ \beta(1 + (\theta_a - \theta_b)) - 2\epsilon - 1 & \text{if } \epsilon \leq \beta \theta_a \ . \end{cases}$$

Thus, the certificate is non-negative only if

$$\epsilon \leq \max\left(\frac{\beta(1+(\theta_a-\theta_b))-1}{2}, 0\right) \ .$$

The case $\beta \geq 1$ can be worked out similarly, leading to

$$\epsilon \leq \max\left(\frac{\beta(-1+(\theta_a-\theta_b))+1}{2}, 0\right) \ .$$

The two cases can be combined as

$$\epsilon \leq \max\left(\frac{\beta(\theta_a-\theta_b)-|\beta-1|}{2}, 0\right) \ .$$

A.6.2 Calculation of certificate for Rényi divergence

We consider the cases $\alpha \geq 1$ and $\alpha \leq 1$ separately.

Case 1 ($\alpha \geq 1$) If $\alpha \geq 1$, the function $f(u) = (u^\alpha - 1)$ is a convex function with $f(1) = 0$. Then, we have

$$\begin{aligned} f^*_\lambda(u) &= \max_{v \geq 0} uv - \lambda(v^\alpha - 1) = \begin{cases} \lambda & \text{if } u \leq 0 \\ \lambda + \lambda(\alpha-1)\left(\frac{u}{\lambda\alpha}\right)^{\frac{\alpha}{\alpha-1}} & \text{if } u \geq 0 \end{cases} \\ &= \lambda + \lambda(\alpha-1)\left(\frac{\max(u,0)}{\lambda\alpha}\right)^{\frac{\alpha}{\alpha-1}} \ . \end{aligned}$$

Suppose we have a bound on the Rényi divergence $R_\alpha(\nu\|\rho) \leq \epsilon$. Then we know $D_f(\nu\|\rho) \leq \exp((\alpha-1)\epsilon) - 1$. Let $\beta = \frac{\alpha}{\alpha-1}$ and

$$B = \theta_a(\max(0, \kappa-1))^\beta + \theta_b(\max(0,\kappa+1))^\beta + (1-\theta_a-\theta_b)(\max(0,\kappa))^\beta \ .$$

Then the certificate Eq. 7 simplifies to (after some algebra)

$$\max_{\lambda \geq 0, \kappa} \kappa - \lambda \exp((\alpha-1)\epsilon) - B\lambda^{1-\beta}\frac{(\alpha-1)}{\alpha^\beta} \ .$$

Setting the derivative with respect to λ to 0 and solving for λ, we obtain

$$\lambda = \frac{1}{\alpha}\left(\left(\frac{B}{\exp((\alpha-1)\epsilon)}\right)\right)^{\left(\frac{1}{\beta}\right)} \ ,$$

and the optimal certificate reduces to

$$\max_\kappa \kappa - B^{\frac{1}{\beta}} \exp\left(\frac{\epsilon}{\beta}\right) \ .$$

For this number to be positive, we need that $\kappa \geq 0$ and

$$\frac{\kappa}{B^{\frac{1}{\beta}}} \geq \exp\left(\frac{\epsilon}{\beta}\right) \ .$$

The LHS above evaluates to

$$\left(\theta_a \max(0, 1-\gamma)^\beta + \theta_b \max(0, 1+\gamma)^\beta + 1 - \theta_a - \theta_b\right)^{-\frac{1}{\beta}} \ .$$

where $\gamma = \frac{1}{\kappa} \geq 0$. Maximizing this expression with respect to γ, we obtain

$$\gamma = \frac{\theta_a^{\alpha-1} - \theta_b^{\alpha-1}}{\theta_a^{\alpha-1} + \theta_b^{\alpha-1}} \ ,$$

so that the certificate reduces to

$$\left(2^\beta \theta_a \theta_b \left(\theta_a^{\alpha-1} + \theta_b^{\alpha-1}\right)^{\left(-\frac{1}{\alpha-1}\right)} + 1 - \theta_a - \theta_b\right)^{\left(\frac{1}{\beta}\right)} \geq \exp\left(\frac{\epsilon}{\beta}\right) \ .$$

Taking logarithms now gives the result.

Case 2 ($0 \leq \alpha \leq 1$) When $0 \leq \alpha \leq 1$, the function $f(u) = (1 - u^\alpha)$ is a convex function with $f(1) = 0$. Then, we have

$$f^*_\lambda(u) = \max_{v \geq 0} uv - \lambda(1 - v^\alpha) = \begin{cases} -\lambda + \lambda^{\frac{1}{1-\alpha}}(1-\alpha)\left(\frac{-u}{\alpha}\right)^{\left(-\frac{\alpha}{1-\alpha}\right)} & \text{if } u \leq 0 \ , \\ \infty & \text{otherwise} \ . \end{cases}$$

Further, a bound $R_\alpha(\nu \| \rho) \leq \epsilon$ implies

$$D_f(\nu \| \rho) \leq 1 - \exp((\alpha - 1)\epsilon) \ .$$

Then the certificate from Eq. 7 reduces to

$$\max_{\kappa, \lambda \geq 0} \kappa + \lambda \exp((\alpha - 1)\epsilon) - (1 - \alpha)\lambda^{\frac{1}{1-\alpha}} \alpha^{\frac{\alpha}{1-\alpha}} \left(\theta_a (1 - \kappa)^{-\frac{\alpha}{1-\alpha}} + \theta_b (-1 - \kappa)^{-\frac{\alpha}{1-\alpha}} + \theta_c (-\kappa)^{-\frac{\alpha}{1-\alpha}} \right)$$

with the constraint $\kappa \leq -1$ (otherwise the certificate is $-\infty$). Setting the derivative with respect to λ to 0 and solving for λ, we obtain

$$\lambda = \frac{\exp\left((\alpha - 1)\epsilon\left(\frac{1-\alpha}{\alpha}\right)\right)}{\alpha\omega} \ ,$$

where

$$\omega = \left(\theta_a (1 - \kappa)^{-\frac{\alpha}{1-\alpha}} + \theta_b (-1 - \kappa)^{-\frac{\alpha}{1-\alpha}} + \theta_c (-\kappa)^{-\frac{\alpha}{1-\alpha}} \right)^{\left(\frac{1-\alpha}{\alpha}\right)} \ .$$

Plugging this back into the certificate and setting $\beta = \frac{\alpha}{1-\alpha}$, we obtain

$$\kappa + \frac{\exp\left(-\frac{\epsilon}{\beta}\right)}{\omega} \ .$$

For this number to be positive, we require that

$$\frac{1}{-\kappa\omega} \geq \exp\left(\frac{\epsilon}{\beta}\right) \ .$$

The LHS of the above expression evaluates to

$$\left(\theta_a (1 + \gamma)^{(-\beta)} + \theta_b (1 - \gamma)^{(-\beta)} + 1 - \theta_a - \theta_b \right)^{\left(\frac{-1}{\beta}\right)} \ ,$$

where $\gamma = -\frac{1}{\kappa}$. Maximizing this expression over $\gamma \in [0, 1]$, we obtain the final certificate to be

$$\left(1 - \theta_a - \theta_b + 2\left(\frac{\theta_a^{1-\alpha} + \theta_b^{1-\alpha}}{2} \right)^{\left(\frac{1}{1-\alpha}\right)} \right)^{\left(-\frac{1}{\beta}\right)} \geq \exp\left(\frac{\epsilon}{\beta}\right) \ .$$

Taking logarithms, we obtain

$$\epsilon \leq -\log\left(1 - \theta_a - \theta_b + 2\left(\frac{\theta_a^{1-\alpha} + \theta_b^{1-\alpha}}{2} \right)^{\left(\frac{1}{1-\alpha}\right)} \right) \ .$$

A.7 Information-limited robust certification and tight relaxations

A.7.1 Proof of theorem 3

At a high level, the proof shows that, in the information-limited case, to achieve robust certification under an arbitrary set of constraints $\mathcal{D}$ it suffices to know the "envelope" of $\mathcal{D}$ with respect to all hockey-stick divergences of order $\beta \geq 0$, i.e. the function $\beta \mapsto \max_{\nu \in \mathcal{D}} D_{\mathrm{HS},\beta}(\nu \| \rho)$ captures all the necessary information to provide information-limited robust certification with respect to $\mathcal{D}$.

We start by considering the following optimization problem:

$$\min_{\Psi : \mathcal{X} \mapsto \{-1, 0, +1\}, \nu \in \mathcal{D}} \mathop{\mathrm{E}}_{X \sim \nu}[\Psi(X)] \tag{13a}$$

$$\text{Subject to } \mathop{\mathrm{E}}_{X \sim \rho}[1[\Psi(X) = +1]] \geq \theta_a \ , \tag{13b}$$

$$\mathop{\mathrm{E}}_{X \sim \rho}[1[\Psi(X) = -1]] \leq \theta_b \ . \tag{13c}$$

In the information-limited setting, this problem attains the minimum expected value over $\phi \in S$. Here $1[\phi(X) = 1]$ denotes the indicator function.

It will be convenient to write this in a slightly different form: Rather than looking at the outputs of Ψ as the $+1, 0, -1$, we look at them as vectors in $\mathbb{R}^3$:

$$\mathcal{Z} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

and define

$$\mathbf{a} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \quad , \quad \mathbf{a}_+ = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad , \quad \mathbf{a}_- = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad .$$

Then, we can write the optimization problem Eq. 13 equivalently as

$$\min_{\Psi: \mathcal{X} \mapsto \mathcal{Z}, \nu \in \mathcal{D}} \mathop{\mathrm{E}}_{X \sim \nu} \left[\mathbf{a}^T \Psi(X) \right] \tag{14a}$$

$$\text{Subject to } \mathop{\mathrm{E}}_{X \sim \rho} \left[{\mathbf{a}_+}^T \Psi(X) \right] \geq \theta_a \ , \tag{14b}$$

$$\mathop{\mathrm{E}}_{X \sim \rho} \left[{\mathbf{a}_-}^T \Psi(X) \right] \leq \theta_b \ . \tag{14c}$$

We first consider the minimization over Ψ for a fixed value of ν. We begin by observing that since the objective is linear, the optimization over Ψ can be replaced with the optimization over the convex hull of the set of Ψ that satisfy the constraints (Bubeck, 2013). Since each input $x \in \mathcal{X}$ can be mapped independently of the rest, the convex hull is simply the cross product of the convex hull at every x, to obtain the constraint set

$$\left\{ \Psi : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Z}) \text{ such that } \mathop{\mathrm{E}}_{X \sim \rho} \left[\mathbf{a}_+^T \Psi(X) \right] \geq \theta_a, \mathop{\mathrm{E}}_{X \sim \rho} \left[\mathbf{a}_-^T \Psi(X) \right] \leq \theta_b \right\} \ .$$

Therefore, the optimization problem reduces to

$$\min_{\Psi: \mathcal{X} \mapsto \mathcal{P}(\mathcal{Z})} \mathop{\mathrm{E}}_{X \sim \nu} \left[\mathbf{a}^T \Psi(X) \right] \tag{15a}$$

$$\text{Subject to } \mathop{\mathrm{E}}_{X \sim \rho} \left[{\mathbf{a}_+}^T \Psi(X) \right] \geq \theta_a \ , \tag{15b}$$

$$\mathop{\mathrm{E}}_{X \sim \rho} \left[{\mathbf{a}_-}^T \Psi(X) \right] \leq \theta_b \ . \tag{15c}$$

This is a convex optimization problem in Ψ. Denote

$$r(X) = \frac{\nu(X)}{\rho(X)} \ .$$

Considering the dual of this optimization problem with respect to the optimization variable Ψ, we obtain

$$\begin{aligned}
&\min_{\Psi} \mathop{\mathrm{E}}_{X \sim \rho} \left[\mathbf{a}^T \Psi(X) r(X) \right] - \lambda_a \left(\mathop{\mathrm{E}}_{X \sim \rho} \left[{\mathbf{a}_+}^T \Psi(X) \right] - \theta_a \right) + \lambda_b \left(\mathop{\mathrm{E}}_{X \sim \rho} \left[{\mathbf{a}_-}^T \Psi(X) \right] - \theta_b \right) \\
&= \min_{\Psi} \lambda_a \theta_a - \lambda_b \theta_b + \mathop{\mathrm{E}}_{X \sim \rho} \left[\left(r(X) \mathbf{a} - \lambda_a \mathbf{a}_+ + \lambda_b \mathbf{a}_- \right)^\top \Psi(X) \right] \\
&= \min_{\Psi} \lambda_a \theta_a - \lambda_b \theta_b + \mathop{\mathrm{E}}_{X \sim \rho} \left[\begin{pmatrix} r(X) - \lambda_a \\ 0 \\ -r(X) + \lambda_b \end{pmatrix}^T \Psi(X) \right] \ .
\end{aligned}$$

Since we can choose $\Psi(x)$ independently for each $x \in \mathcal{X}$, we can minimize each term in the expectation independently to obtain

$$\min_{\Psi(x) \in \mathcal{P}(\mathcal{Z})} \begin{pmatrix} r(x) - \lambda_a \\ 0 \\ r(x) + \lambda_b \end{pmatrix}^T \Psi(x) = \min(r(x) - \lambda_a, 0, -r(x) + \lambda_b) \ .$$

This implies that the Lagrangian evaluates to

$$\lambda_a \theta_a - \lambda_b \theta_b + \mathop{\mathrm{E}}_{X \sim \rho} \left[\min(r(X) - \lambda_a, 0, r(X) + \lambda_b) \right] \ .$$

We now consider two cases:

Case 1 ($\lambda_a \geq \lambda_b \geq 0$) In this case, we can see that

$$\begin{aligned}&\min(r(X) - \lambda_a, 0, -r(X) + \lambda_b) = \min(r(X) - \lambda_a, 0) + \min(-r(X) + \lambda_b, 0) \\ &= r(X) - \lambda_a - \max(r(X) - \lambda_a, 0) - \max(r(X) - \lambda_b, 0) \ .\end{aligned}$$

Then, the Lagrangian reduces to

$$\begin{aligned}&\lambda_a\theta_a - \lambda_b\theta_b + \mathop{\mathrm{E}}_{X\sim\rho}[r(X) - \lambda_a] - \mathop{\mathrm{E}}_{X\sim\rho}[\max(r(X) - \lambda_a, 0)] - \mathop{\mathrm{E}}_{X\sim\rho}[\max(r(X) - \lambda_b, 0)] \\ &= 1 - \lambda_a(1 - \theta_a) - \lambda_b\theta_b - (D_{\mathrm{HS},\lambda_a}(\nu\|\rho) + \max(1 - \lambda_a, 0)) - (D_{\mathrm{HS},\lambda_b}(\nu\|\rho) + \max(1 - \lambda_b, 0)) \ .\end{aligned}$$

Case 2 ($\lambda_b \geq \lambda_a \geq 0$) In this case, we can see that

$$\begin{aligned}&\min(r(X) - \lambda_a, 0, -r(X) + \lambda_b) = \min(r(X) - \lambda_a, -r(X) + \lambda_b) \\ &= r(X) - \lambda_a + 2\min\left(0, \frac{\lambda_a + \lambda_b}{2} - r(X)\right) \\ &= r(X) - \lambda_a - 2\max\left(r(X) - \frac{\lambda_a + \lambda_b}{2}, 0\right) \ .\end{aligned}$$

Then, the Lagrangian reduces to

$$\begin{aligned}&\lambda_a\theta_a - \lambda_b\theta_b + \mathop{\mathrm{E}}_{X\sim\rho}[r(X) - \lambda_a] - 2\mathop{\mathrm{E}}_{X\sim\rho}\left[\max\left(r(X) - \frac{\lambda_a + \lambda_b}{2}, 0\right)\right] \\ &= 1 - \lambda_a(1 - \theta_a) - \lambda_b\theta_b - 2\left(D_{\mathrm{HS},\frac{\lambda_a+\lambda_b}{2}}(\nu\|\rho) + \max\left(1 - \frac{\lambda_a + \lambda_b}{2}, 0\right)\right) \ .\end{aligned}$$

We know that $1 - \theta_a \geq \theta_b$ and $\lambda_b \geq \lambda_a$. If $\lambda_b > \lambda_a$, by choosing $\lambda'_a = \lambda_a + \kappa$ and $\lambda'_b = \lambda_b - \kappa$ for some small $\kappa > 0$, we know that the the sum of the first three terms would reduce while the final term would remain unchanged. Thus, at the the optimum in this case, we can assume $\lambda_a = \lambda_b$ and we obtain

$$1 - \lambda_a(1 - \theta_a) - \lambda_a\theta_b - 2(D_{\mathrm{HS},\lambda_a}(\nu\|\rho) + \max(1 - \lambda_a, 0)) \ .$$

Final analysis of Lagrangian Combining the two cases we can write the dual problem as

$$\begin{aligned}\max_{\lambda_a\geq\lambda_b\geq 0} \ & 1 - \lambda_a(1 - \theta_a) - \lambda_b\theta_b - (D_{\mathrm{HS},\lambda_a}(\nu\|\rho) + \max(1 - \lambda_a, 0)) && (16) \\ & - (D_{\mathrm{HS},\lambda_b}(\nu\|\rho) + \max(1 - \lambda_b, 0)) \ . && (17)\end{aligned}$$

By strong duality, the optimal value of the above problem precisely matches the optimal value of Eq. 15 (and hence Eq. 13). Thus, information limited robust certification with respect to $\mathcal{D}$ holds if and only if Eq. 17 has a non-negative optimal value for each $\nu \in \mathcal{D}$. Since we have that

$$\max_{\nu\in\mathcal{D}} D_{\mathrm{HS},\lambda_a}(\nu\|\rho) = \epsilon_{\lambda_a} \ , \quad \max_{\nu\in\mathcal{D}} D_{\mathrm{HS},\lambda_b}(\nu\|\rho) = \epsilon_{\lambda_b} \ ,$$

information-limited robust certification holds if and only if the optimal value of

$$\begin{aligned}\max_{\lambda_a\geq\lambda_b\geq 0} \ & 1 - \lambda_a(1 - \theta_a) - \lambda_b\theta_b - (\epsilon_{\lambda_a} + \max(1 - \lambda_a, 0)) && (18) \\ & - (\epsilon_{\lambda_b} + \max(1 - \lambda_b, 0)) && (19)\end{aligned}$$

is non-negative. Further, since the optimal value only depended on the value of $D_{\mathrm{HS},\beta}(\nu\|\rho)$ for $\beta \geq 0$, it is equivalent to information-limited robust certification with respect to $\mathcal{D}_{\mathrm{HS}}$.

The above argument also shows that in this case, information-limited robust certification with respect to $\mathcal{D}$ is equivalent to requiring that the following convex optimization problem has a non-negative optimal value:

$$\max_{\lambda_a\geq\lambda_b\geq 0} \quad 1 - \lambda_a(1 - \theta_a) - \lambda_b\theta_b - \left(\epsilon_{\lambda_a} + [1 - \lambda_a]_+\right) - \left(\epsilon_{\lambda_b} + [1 - \lambda_b]_+\right) \ . \tag{20}$$

A.7.2 GAUSSIAN SMOOTHING MEASURES

Consider the specification $\phi_{c,c'}$ applied to hard classifiers smoothed using a smoothing measure $\mu : \mathcal{X} \mapsto \mathcal{P}(\mathcal{X})$. The strict black-box verification problem is to certify that for all x' with $d(x, x') \leq \epsilon$ we have $\mathrm{E}_{X\sim\mu(x')}[\phi_{c,c'}(X)] \geq 0$ knowing only that $\mathrm{E}_{X\sim\mu(x)}[\phi_{c,c'}(X) = +1] \geq \theta_a$ and $\mathrm{E}_{X\sim\mu(x)}[\phi_{c,c'}(X) = -1] \leq \theta_b$. The above lemma gives us the exact solution to this problem provided that we can compute

$$\max_{x':d(x,x')\leq\epsilon} D_{\mathrm{HS},\beta}(\mu(x')\|\mu(x))$$

for each $\beta \geq 0$. In particular, when μ is a Gaussian measure $\mu(x) = \mathcal{N}\big(x, \sigma^2 I\big)$, it can be shown that (see, e.g., Balle & Wang (2018)):

$$\begin{aligned}&\max_{x':\|x-x'\|_2\leq\epsilon} D_{\mathrm{HS},\beta}(\mu(x')\|\mu(x)) \\ &= \Psi_g\bigg(\frac{\epsilon}{2\sigma} - \frac{\log(\beta)\sigma}{2\epsilon}\bigg) - \beta\Psi_g\bigg(-\frac{\epsilon}{2\sigma} - \frac{\log(\beta)\sigma}{2\epsilon}\bigg) - \max(1-\beta, 0) \ ,\end{aligned}$$

where Ψ_g is the CDF of a standard normal random variable $\mathcal{N}(0, 1)$. Applying Eq. 20 to this expression, we recover the main result from (Cohen et al., 2019):

Corollary 5. (θ_a, θ_b) are strictly robustly certified at $\rho = \mathcal{N}\big(x, \sigma^2 I\big)$ with respect to

$$\mathcal{D}_{x,\epsilon} = \{\mathcal{N}\big(x', \sigma^2 I\big) : \|x - x'\|_2 \leq \epsilon\}$$

if and only if $1 \geq \theta_a \geq \theta_b \geq 0$ and

$$\Psi_g\Big(\Psi_g^{-1}(\theta_a) - \frac{\epsilon}{\sigma}\Big) + \Psi_g\Big(\Psi_g^{-1}(1-\theta_b) - \frac{\epsilon}{\sigma}\Big) \geq 1 \ .$$

Proof. We have

$$\epsilon_\beta = \Psi_g\bigg(\frac{\epsilon}{2\sigma} - \frac{\log(\beta)\sigma}{2\epsilon}\bigg) - \beta\Psi_g\bigg(-\frac{\epsilon}{2\sigma} - \frac{\log(\beta)\sigma}{2\epsilon}\bigg) - \max(1-\beta, 0) \ ,$$

so that Eq. 20 reduces to

$$\begin{aligned}\max_{\lambda_a\geq\lambda_b\geq 0} &1 - \lambda_a(1-\theta_a) - \lambda_b\theta_b \\ &- \bigg(\Psi_g\bigg(\frac{\epsilon}{2\sigma} - \frac{\log(\lambda_a)\sigma}{2\epsilon}\bigg) - \lambda_a\Psi_g\bigg(-\frac{\epsilon}{2\sigma} - \frac{\log(\lambda_a)\sigma}{2\epsilon}\bigg)\bigg) \\ &- \bigg(\Psi_g\bigg(\frac{\epsilon}{2\sigma} - \frac{\log(\lambda_b)\sigma}{2\epsilon}\bigg) - \lambda_b\Psi_g\bigg(-\frac{\epsilon}{2\sigma} - \frac{\log(\lambda_b)\sigma}{2\epsilon}\bigg)\bigg) \ .\end{aligned}$$

The result then follows from setting the derivatives of this expresion to 0 with respect to λ_a, λ_b. □

A.8 EFFICIENT SAMPLING AND F-DIVERGENCE COMPUTATION FOR NORM-BASED SMOOTHING MEASURES

Lemma 6. The smoothing measure $\mu : \mathcal{X} \mapsto \mathcal{P}(\mathcal{X})$ with density $\mu(x)[z] \propto \exp(-\|z - x\|)$ satisfies

$$\max_{\|\delta\|\leq\epsilon} R_\infty(\mu(x+\delta)\|\mu(x)) \leq \epsilon \ .$$

if $\|x\|$ is any norm. Further, if f is convex function with $f(1) = 0$ such that $f\big(\frac{1}{u}\big)$ is convex and monotonically increasing in u, then

$$\max_{\|\delta\|\leq\epsilon} D_f(\mu(x+\delta)\|\mu(x)) \leq \max_{\|\delta\|=\epsilon} \mathrm{E}_{X\sim\mu(0)}[f(\exp(-\|X-\delta\| + \|X\|))] \ . \tag{21}$$

Proof. By the triangle inequality, we have

$$\frac{\mu(x')[z]}{\mu(x)[z]} = \exp(\|z-x\| - \|z-x'\|) \leq \exp(\|x-x'\|)$$

so that

$$R_\infty(\mu(x')\|\mu(x)) \leq \|x - x'\| \ .$$

Similarly, for f that satisfy the conditions of the theorem, it can be shown that $D_f(\mu(x')\|\mu(x))$ is convex in x' so that its maximum over the convex set $\|x' - x\| \leq \epsilon$ is attained on the boundary. □

For several norms, the optimization problem in Eq. 21 can be solved in closed form. These include $\ell_1, \ell_2, \ell_\infty$ norms and the matrix spectral norm and nuclear norm (the final two are relevant when $\mathcal{X}$ is a space of matrices). The results are documented in Table 3. Thus, every f-divergence that meets the conditions of Lemma 6 can be estimated efficiently for these norms. In particular, the divergences that are induced by the functions $\tilde{f}(u^{-\alpha})$ for any monotonic convex function $\tilde{f}$ and $\alpha \geq 0$ satisfy this constraint. This gives us a very flexible class of f-divergences that can be efficiently estimated for these norm-based smoothing measures.

Constraint on δ	Bound on Eq. 21	Sampling from $X \sim \mu(0)$
$\|\delta\|_1 \leq \epsilon$	$\mathop{\mathrm{E}}_{X\sim\mu_1(0)}[f(\exp(\|X - \epsilon e_0\|_1 - \|X\|_1))]$	$X_i \sim \mathrm{Lap}(0,1)$ iid
$\|\delta\|_2 \leq \epsilon$	$\mathop{\mathrm{E}}_{X\sim\mu_2(0)}[f(\exp(\|X - \epsilon e_0\|_2 - \|X\|_2))]$	$X = Ru$ $R \sim \Gamma(d,1)$ $u \sim \mathcal{U}(\partial\mathcal{B}_2)$
$\|\delta\|_\infty \leq \epsilon$	$\mathop{\mathrm{E}}_{X\sim\mu_\infty(0)}[f(\exp(\|X - \epsilon\mathbf{1}\|_\infty - \|X\|_\infty))]$	$X = Ru$ $R \sim \Gamma(d+1,1)$ $u \in \mathcal{U}(\mathcal{B}_\infty)$
$\|\delta\|_{nuc} \leq \epsilon$	$\mathop{\mathrm{E}}_{\substack{s\sim\mu_1(0)\\U,V\sim\mathcal{U}(\mathcal{O})}}\left[f\left(\exp\left(\left\|U\,[[s]]V^T - \epsilon\,[[e_0]]\right\|_{nuc} - \|s\|_1\right)\right)\right]$	$X = U\,[[s]]V^T$ $s \sim \mu_1, U, V \sim \mathcal{U}(\mathcal{O})$
$\|\delta\|_\star \leq \epsilon$	$\mathop{\mathrm{E}}_{\substack{s\sim\mu_\infty(x)\\U,V\sim\mathcal{U}(\mathcal{O})}}\left[f\left(\exp\left(\left\|U\,[[s]]V^T - \epsilon\,[[e_0]]\right\|_\star - \|s\|_\infty\right)\right)\right]$	$X = U\,[[s]]V^T$ $s \sim \mu_\infty, U, V \sim \mathcal{U}(\mathcal{O})$

Table 3: Bounds on f-divergences: e_0 is the vector with 1 in the first coordinate and zeros in all other coordinates and $\mathbf{1}$ is the vector with all coordinates equal to 1. μ_p refers to the smoothing measure induced by the ℓ_p norm, $\mathcal{U}(S)$ refers to the uniform measure over the set S, $\mathcal{O}$ is the set of orthogonal matrices and $\mathcal{B}_p = \{\|z\|_p \leq 1\}$ is the unit ball in the ℓ_p norm.

Efficient sampling The only other requirement for obtaining a certificate computationally is to be able to sample from $\mu(x)$ to estimate θ_a, θ_b. Since $\mu(x)$ is log-concave, there are general purpose polynomial time algorithms for sampling from this measure. However, for most norms, more efficient methods exist, as outlined below.

The random variable $X \sim \mu(x)$ can be obtained as $X = x + Z$ with $Z \sim \mu(0)$. Thus, to sample from $\mu(x)$ for any x it is enough to be able to sample from $\mu(0)$. For $\|\cdot\|_1$, this reduces to sampling from a Laplace distribution which can be done easily. For $\|\cdot\|_\infty$, (Steinke & Ullman, 2016) give the following efficient sampling procedure: first sample r from a Gamma distribution with shape $d+1$ and mean $d+1$, i.e. $r \sim \Gamma(d+1,1)$, and then sample each Z_i, $i \in [d]$, uniformly from $[-r, r]$. Theorem 7 gives a short proof of correctness for this procedure. Theorem 8 also has a similar result for the case of $\|\cdot\|_2$ and Table 3 lists the sampling procedures for several norms.

Theorem 7. The random variable $Z \in \mathbb{R}^d$ obtained by first sampling $R \sim \Gamma(d+1,1)$ and then sampling each Z_i, $i \in [d]$, uniformly from $[-R, R]$ has density $\propto e^{-\|z\|_\infty}$.

Proof. We first compute the normalization constant for a density of the form $\propto e^{-\|z\|_\infty}$ as follows:

$$\int_{\mathbb{R}^d} e^{-\|z\|_\infty} dz = \int_0^\infty \left(\int_{\mathbb{R}^d} 1[\|z\|_\infty = t] dz\right) e^{-t} dt = \int_0^\infty 2^d dt^{d-1} e^{-t} dt = d!2^d \ .$$

Next we show the density of Z satisfies $p_Z(z) = e^{-\|z\|_\infty}/(d!2^d)$ by noting that conditioned on $R = r$ we have $p_{Z|R=r}(z) = 1[\|z\|_\infty \leq r]/(2r)^d$ because of the uniform sampling used in each

coordinate, and integrating over R sampled from a Gamma distribution with shape $d+1$ and mean $d+1$ yields

$$p_Z(z) = \int_0^\infty p_{Z|R=r}(z)p_R(r)dr = \int_0^\infty \frac{1[\|z\|_\infty \le r]}{(2r)^d}\frac{r^d e^{-r}}{d!}dr = \frac{1}{d!2^d}\int_{\|z\|_\infty}^\infty e^{-r}dr = \frac{e^{-\|z\|_\infty}}{d!2^d} \ .$$

□

Theorem 8. The random variable $Z \in \mathbb{R}^d$ obtained by first sampling $Z' \sim \mathcal{N}(0, I)$ and $R \sim \Gamma_p(d, 1)$ and then taking $Z = R\frac{Z'}{\|Z'\|_2}$ has density $\propto e^{-\|z\|_2^p}$. Here $\Gamma_p(d, a)$ denotes the generalized Gamma distribution of order $p > 0$ with shape d and scale a.

Proof. First note that $W = \frac{Z'}{\|Z'\|_2} \sim \mathcal{U}(\mathcal{B}_2)$; i.e. it is uniform on the ℓ_2 ball of radius 1. Therefore, RW is uniform on the ℓ_2 ball of radius R and the conditional density of Z given R is given by $p_{Z|R=r}(z) = \frac{1[\|z\|_2=r]\Gamma(d/2)}{2\pi^{d/2}r^{d-1}}$. Since R has density $p_R(r) \propto r^{d-1}e^{-r^p}$, we get

$$\begin{aligned} p_Z(z) &= \int_0^\infty p_{Z|R=r}(z)p_R(r)dr \\ &\propto \int_0^\infty \frac{1[\|z\|_2 = r]\Gamma(d/2)}{2\pi^{d/2}r^{d-1}} r^{d-1}e^{-r^p}dr \\ &\propto e^{-\|z\|_2^p} \ . \end{aligned}$$

□

A.9 High-confidence estimates of expected values

Let $Z_1, \ldots, Z_t$ be independent, identically distributed random variables with range R and mean m. Let the empirical mean be $\bar{Z} = \frac{1}{t}\sum_{i=1}^t Z_i$ and the empirical variance be $\bar{\sigma}^2 = \frac{1}{t}\sum_{i=1}^t (Z_i - \bar{Z})^2$. Applying Bernstein's inequality to the sum and the sum of the squares of these random variables, we get the empirical Bernstein bound (Audibert et al., 2009), which states that with probability at least $1-\zeta$,

$$|\bar{Z} - m| \le \sqrt{\frac{2\bar{\sigma}^2 \log(3/\zeta)}{t}} + \frac{3R\log(3/\zeta)}{t}. \tag{22}$$

The main benefit of the above inequality is that as long as the variance of the sample $Z_1, \ldots, Z_t$ is small, the convergence rate becomes essentially $O(1/t)$ instead of the standard $O(1/\sqrt{t})$. Also, since Eq. 22 only contains empirical quantities apart from the range R, it can be used to obtain computable bounds for the expectation μ: with probability at least $1-\zeta$,

$$\bar{Z} - \sqrt{\frac{2\bar{\sigma}^2 \log(3/\zeta)}{t}} - \frac{3R\log(3/\zeta)}{t} \le m \le \bar{Z} + \sqrt{\frac{2\bar{\sigma}^2 \log(3/\zeta)}{t}} + \frac{3R\log(3/\zeta)}{t} \ . \tag{23}$$

This bound can be applied to approximate the expectation in Eq. 7 with high probability for given values of λ and κ. More specifically, if the function $f_\lambda^*(\kappa - \phi(\cdot))$ is bounded with range R, then taking t samples $X_1, \ldots, X_t$ independently from ρ, and defining $Z_i = f_\lambda^*(\kappa - \phi(X_i))$, and $\bar{Z}$ and $\bar{\sigma}^2$ as above, Eq. 23 implies that with probability at least $1-\zeta$,

$$\mathop{\mathrm{E}}_{X\sim\rho}[f_\lambda^*(\kappa - \phi(X_i))] \le \bar{Z} + \sqrt{\frac{2\bar{\sigma}^2 \log(3/\zeta)}{t}} + \frac{3R\log(3/\zeta)}{t} \ .$$

Plugging in this bound to Eq. 7 gives a high-probability lower bound for the function to be maximized for any given λ and κ.

In practice, the way we apply this bound is to use a stochastic gradient method to optimize the values of λ, κ, using samples of $X \sim \rho$ to get an unbiased estimate of the gradient of the objective. We then freeze the values of λ, κ (after a fixed number of optimization steps) and then use the above procedure to get a high confidence lower bound on the objective.

A.10 DISCRETE PERTURBATIONS

We can also handle discrete perturbations in our framework. A natural case to consider is ℓ_0 perturbations. In this case, we assume that $\mathcal{X} = A^d$ where

$$A = \{1, \dots, K\}$$

is a discrete set. Then, we can choose

$$\mu(x)[z] = \prod_{i=1}^{d} p^{1[z_i = x_i]} \left(\frac{q}{K-1} \right)^{1[z_i \neq x_i]} ,$$

where $p + q = 1$, $p \geq q \geq 0$, and p denotes the probability that the measure retains the value of x and $\frac{q}{K-1}$ denotes a uniform probability of switching it to a different value. In this case, it can be shown that for every $\alpha > 0$ that

$$R_\alpha(\mu(x') \| \mu(x)) = \|x - x'\|_0 \left(\frac{\log\left(p\left(\frac{q}{(K-1)p} \right)^{(\alpha)} + q\left(\frac{(K-1)p}{q} \right)^{(\alpha)} \right)}{\alpha - 1} \right)$$

so that we can derive a certificate with respect to ℓ_0 perturbations using any set of Rényi divergences (or combinations of theses).

This can be extended to structured discrete perturbations by introducing coupling terms between the perturbations:

$$\mu(x)[z] \propto \prod_{i=1}^{d} p^{1[z_i = x_i]} \left(\frac{q}{K-1} \right)^{1[z_i \neq x_i]} \exp\left(\sum_{i=1}^{d-1} \eta 1[z_i = x_i] 1[z_{i+1} = x_{i+1}] \right) .$$

This would correlate perturbations between adjacent features (which for example may be useful to model correlated perturbations for time series data). Since this can be viewed as a Markov Chain, Rényi divergences between $\mu(x), \mu(x')$ are still easy to compute.