Invariance-Aware Randomized Smoothing Certificates

Anonymous Author(s)
Affiliation
Address
email

Abstract

Building models that comply with the invariances inherent to different domains, such as invariance under translation or rotation, is a key aspect of applying machine learning to real world problems like molecular property prediction, medical imaging, protein folding or LiDAR classification. For the first time, we study how the invariances of a model can be leveraged to provably guarantee the robustness of its predictions. We propose a gray-box approach, enhancing the powerful black-box randomized smoothing technique with white-box knowledge about invariances. First, we develop a post-processing-based gray-box certification procedure that can be applied to arbitrary models with invariance under permutation and Euclidean isometries. Then, we derive provably tight gray-box certificates. We experimentally demonstrate that the provably tight certificates can offer much stronger guarantees, but that in practical scenarios the post-processing method is a good approximation.

1 Introduction

It is well-established that machine learning models are susceptible to adversarial attacks [1–4]. Even without malevolent actors, adversarial attacks can be considered worst case scenarios in environments with noisy, erroneous or otherwise corrupted data, thus necessitating robust machine learning methods.

Invariance is a central design principle that has so far received little dedicated attention in the realm of adversarially robust machine learning. Over the past decades, there has been ongoing research into developing machine learning models that comply with the invariances inherent to different data types and tasks. Prominent recent examples include Deep Sets [5], PointNet [6], Group Equivariant CNNs [7], Spherical CNNs [8] and Graph Convolutional Networks [9], but the study of invariant models significantly precedes the surge in popularity of deep learning methods [10–16].

For the first time, we explore the following question: *Can a-priori knowledge about invariances be leveraged in deriving provable guarantees for a model's robustness to adversarial attacks?*

Going by a loose categorization of prior work, we could adopt one of two possible approaches for our exploration: A white-box or black-box one. White-box certificates (e.g. [17–24]) analyze a model's internals, such at its weights and non-linearities, to provably guarantee that a prediction does not change under adversarial attack. Black box certificates – specifically randomized smoothing [25–27] – use statistical methods to provide provable guarantees that hold for all models sharing the same prediction probabilities under a random input distribution, irrespective of their internals.

We opt for a randomized smoothing approach, as it allows us to focus on the interplay between invariances and robustness, rather than the specific means of implementing these invariances. Based on this decision, we can define a more specific goal: Combining white-box knowledge about invariances with black-box knowledge about prediction probabilities to derive *gray-box* certificates.

For our exploration, we focus on models operating on spatial data, rather than structured data (e.g. images and sequences) – both because spatial symmetries can be elegantly formalized using algebraic

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

concepts and because there is an ongoing trend towards using machine learning for real-world applications with inherent spatial invariances, e.g. molecular property prediction [28–32], LiDAR classification [33–36], drug discovery [37–39], particle physics [40–42] and protein folding [43–47].

Our main contributions are

- a principled method for deriving tight invariance-aware randomized smoothing certificates,
- a tight certificate for models invariant to translations in arbitrary dimensions,
- tight certificates for models invariant to rotations in 2D and elemental rotations in 3D.

We further prove that smoothing with isotropic Gaussian noise preserves invariance to permutation and Euclidian isometries, allowing the classic randomized smoothing certificate of Cohen et al. [27] to be augmented via a post-processing approach. We experimentally demonstrate that it offers a good approximation of our tight certificates, if the variance of the smoothing distribution is small.

2 Related work

Invariant machine learning. Given the diversity of approaches to invariant machine learning, and the fact that our approach is model-agnostic, we refrain from attempting a survey and instead refer to [48] for a principled, high-level introduction into the realm of learning with invariances and equivariances.

Invariance and robustness. Two recent empirical studies [49, 50] demonstrate that data augmentation meant to increase robustness to l_p-norm adversarial attacks reduces robustness to semantically meaningful transformations (e.g. rotation) and vice-versa, suggesting an inherent invariance-robustness trade-off. However, neither studies models that are invariant by design. In [51], the negative effect of shift-invariance on the robustness of image classifiers is investigated. Note that our work is not meant to resolve potential trade-offs, but to tightly bound the actual robustness of models.

Gray-box robustness certificates. While we propose the first gray-box certificate for invariant models, there exists prior work on combining other white-box knowledge with black-box certificates. In [52] and [53], knowledge about a classifier's gradients is used to derive tighter randomized smoothing certificates. In [54], knowledge about a graph neural network's receptive fields is combined with randomized smoothing to derive collective robustness certificates for multiple predictions.

Adversarial attacks on spatial data. Adversarial attacks on spatial data have been extensively studied, in particular for point cloud classification. One can differentiate between attacks that modify [55–63], insert [62, 64, 65] or delete [66–68] coordinates. Like in other domains, empirical defenses have been proposed [56, 69, 70] and subsequently broken [71–73], motivating the development of black-box [74–76] and white-box [77] robustness certificates for spatial data. It should be noted that Gaussian randomized smoothing, without invariance information, has already been used in prior work – either as a baseline [74] or as a special case of the respective certificate [76].

Orthogonal research directions. One related but orthogonal research direction is transformation-specific certification [24, 76–82]. There, a model is assumed to potentially change its prediction under adversarial parametric transformations (e.g. rotations) and one derives provable robustness guarantees for specific parameter ranges. Here, on the other hand, we assume our model to be invariant under a set of transformations, i.e. never change its prediction, and use this property for certification. Aside from that, there exist white-box certification techniques for specific *operations with invariances* (global max-pooling [77], message passing [83, 84] and batch normalization [85]). Prior work treats these operations as coincidental building blocks of the models it is trying to certify. It does not study *invariance itself* and how to leverage it for robustness certification.

3 Background on randomized smoothing

Randomized smoothing is a black-box certification technique that can be adapted to various data types, tasks and threat models [86–95]. Instead of directly certifying a classifier g, it constructs a smoothed classifier f that returns the most likely prediction under random perturbations of its input. It then certifies the robustness of this smoothed classifier. We present the tight Gaussian smoothing certificate derived by Cohen et al. [27]. and its generalization to matrices.[1] [74, 76]

[1] This generalization is a direct consequence of vectorizing all matrices and following the derivations in [27].

Assume a continuous $(N \times D)$-dimensional input space $\mathbb{R}^{N \times D}$, label set $\mathbb{Y}$ and *base classifier* $g : \mathbb{R}^{N \times D} \to \mathbb{Y}$. Let $\mu_{\boldsymbol{X}}(\boldsymbol{Z}) = \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{Z}_{:,d} \mid \boldsymbol{X}_{:,d}, \sigma^2 \mathbf{I}_N\right)$ be the isotropic matrix normal distribution with mean $\boldsymbol{X}$ and standard deviation σ. Let $p_{\boldsymbol{X},y} = \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}}\left[g(\boldsymbol{Z}) = y\right]$ be the probability of g predicting class y under this *smoothing distribution*. One can then define a *smoothed classifier* $f(\boldsymbol{X}) = \operatorname{argmax}_{y \in \mathbb{Y}} p_{\boldsymbol{X},y}$ that returns the most likely prediction of g under $\mu_{\boldsymbol{X}}$.

Let $y^* = f(\boldsymbol{X})$ be a smoothed prediction. Its robustness to adversarially perturbed inputs $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$ can be certified as follows: First, one can use the fact that $f(\boldsymbol{X}') = y^*$ if y^* is more likely than all other classes combined [2] , i.e. $p_{\boldsymbol{X}',y^*} > 0.5$. The probability $p_{\boldsymbol{X}',y^*}$ can then be lower-bounded by finding the *worst-case classifier* from a set of functions $\mathbb{H}$ with $g \in \mathbb{H}$.

$$p_{\boldsymbol{X}',y^*} \geq \min_{h \in \mathbb{H}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right]. \tag{1}$$

For $\mathbb{H} = \left\{h : \mathbb{R}^{N \times D} \to \mathbb{Y} \mid \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}}\left[h(\boldsymbol{Z}) = y^*\right] \geq p_{\boldsymbol{X},y^*}\right\}$, the set of classifiers that are at least as likely as g to predict y^*, the exact solution is given by the Neyman-Pearson lemma [96] from statistical testing. The optimal value is $\Phi\left(\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right) - \frac{||\Delta||_2}{\sigma}\right)$, where Φ is the standard-normal CDF and $|| \cdot ||_2$ is the Frobenius norm. If $||\boldsymbol{\Delta}||_2 < \sigma \Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right)$, then $p_{\boldsymbol{X}',y^*} > 0.5$ and the prediction is provably robust. Because Eq. (1) was solved exactly, this is a *tight certificate*, i.e. the best possible certificate that can be obtained by only using black-box knowledge about prediction probability $p_{\boldsymbol{X},y^*}$ of base classifier g under smoothing distribution $\mu_{\boldsymbol{X}}$.

Probabilistic certificates. For neural networks, the prediction probability $p_{\boldsymbol{X},y^*}$ can usually not be computed analytically. Instead, one has to use Monte Carlo sampling to compute a lower confidence bound $\underline{p_{\boldsymbol{x}',y^*}}$ that holds with high probability $1 - \alpha$. The resulting certificate is a probabilistic one.

4 Problem setting

We consider a similar setup to that described in the previous section, i.e. we have a smoothed classifier $f : \mathbb{R}^{N \times D} \to \mathbb{Y}$ that is the result of randomly smoothing a base classifier g with an isotropic matrix normal distribution $\mu_{\boldsymbol{X}}$. Given a clean prediction $y^* = f(\boldsymbol{X})$ with clean prediction probability $p_{\boldsymbol{X},y^*}$, we want to determine whether $f(\boldsymbol{X}') = y^*$ for an adversarially perturbed input $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$. Different from all prior work, we additionally assume that the *base classifier* g (not f) is invariant under a group of transformations $\mathbb{T}$, i.e. a set of functions that contains the identitiy function as well as the inverse of each element, and is associative and closed under function composition.

Definition 1 (Group-invariance)**.** *A function $h : \mathbb{A} \to \mathbb{B}$ is invariant under a group of transformations $\mathbb{T}$ if $\forall x \in \mathbb{A}, \forall \tau \in \mathbb{T} : h(x) = h(\tau(x))$.*

Translation, rotation and permutation invariance in $\mathbb{R}^{N \times D}$ correspond to the groups $\left\{\boldsymbol{X} \mapsto \boldsymbol{X} + \mathbf{1}_N \boldsymbol{c}^T \mid \boldsymbol{c} \in \mathbb{R}^D\right\}$, $\left\{\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(D)\right\}$, and $\left\{\boldsymbol{X} \mapsto \boldsymbol{P}\boldsymbol{X} \mid \boldsymbol{P} \in S_N\right\}$, where $SO(D)$ is the special orthogonal group, i.e. the set of all $D \times D$ rotation matrices, and S_N is the symmetric group, i.e. the set of all $N \times N$ permutation matrices.

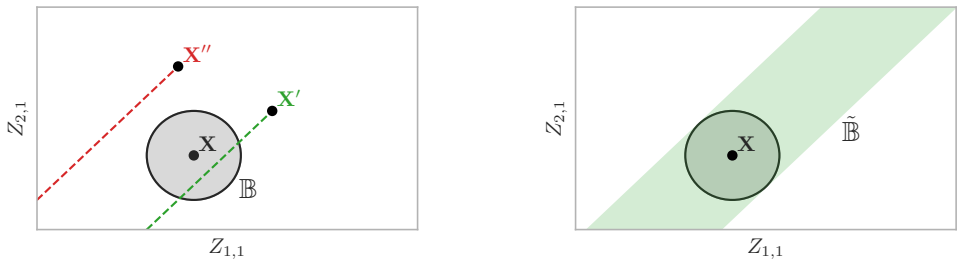
5 Post-processing-based gray-box certificates

Our first method for leveraging invariances for certification is based on the insight that certificates correspond to sets of *inputs* $\mathbb{B} \subseteq \mathbb{R}^{N \times D}$ that preserve clean prediction y^* and invariances correspond to sets of *transformations* that preserve predictions. Consequently, post-processing $\mathbb{B}$ by applying the transformations from $\mathbb{T}$ yields a new set of inputs $\tilde{\mathbb{B}} \supseteq \mathbb{B}$ that provably preserve the clean prediction.

Theorem 1. *Let $h \in \mathbb{R}^{N \times D} \to \mathbb{Y}$ be invariant under a group of transformations $\mathbb{T}$. Let $y^* = h(\boldsymbol{X})$ be a prediction that is certifiably robust to a set of perturbed inputs $\mathbb{B} \subseteq \mathbb{R}^{N \times D}$, i.e. $\forall \boldsymbol{Z} \in \mathbb{B} : h(\boldsymbol{Z}) = y^*$. Let $\tilde{\mathbb{B}} = \{\tau(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathbb{B}, \tau \in \mathbb{T}\}$. Then $\forall \boldsymbol{X}' \in \tilde{\mathbb{B}} : h(\boldsymbol{X}') = y^*$.*

To simplify the comparison with the tight gray-box certificates we shall derive shortly, we adopt an equivalent, pre-processing-based perspective for the special case of Frobenius norm certificates: Input $\boldsymbol{X}'$ is provably not an adversarial example, if it can be mapped into $\mathbb{B}$ via a transformation from $\mathbb{T}$.

[2] One can also combine two of the *binary* certificates discussed here into a *multi-class certificate* that guarantee $p_{\boldsymbol{X}',y^*} > \max_{y' \neq y^*} p_{\boldsymbol{X}',y'}$. We discuss multi-class invariance-aware certificates in Appendix F.



(a) Pre-processing perspective

(b) Post-processing perspective

Figure 1: Pre-processing and post-processing perspective for translation invariance and $N = 2, D = 1$. a.) Input $\boldsymbol{X}'$ can be translated into certified region $\mathbb{B}$. It is not an adversarial example. Input $\boldsymbol{X}''$ can not be translated into $\mathbb{B}$. It might be an adversarial example. b.) The prediction $y^* = f(\boldsymbol{X})$ is certifiably robust to all perturbed inputs from region $\tilde{\mathbb{B}}$, the union over all translations of $\mathbb{B}$.

Theorem 2. *Let $h \in \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under a group of transformations $\mathbb{T}$. Let $y^* = h(\boldsymbol{X})$ be a prediction that is certifiably robust to a set of perturbed inputs $\mathbb{B} = \{\boldsymbol{Z} \mid ||\boldsymbol{Z} - \boldsymbol{X}||_2 < r\}$, i.e. $\forall \boldsymbol{Z} \in \mathbb{B} : h(\boldsymbol{Z}) = y^*$. If $\min_{\tau\in\mathbb{T}} ||\tau(\boldsymbol{X}') - \boldsymbol{X}||_2 < r$, then $h(\boldsymbol{X}') = y^*$.*

Fig. 1 compares the two perspectives. Note that Theorem 2 is only a way of determining whether $\boldsymbol{X}' \in \tilde{\mathbb{B}}$ for a specific $\boldsymbol{X}'$. It is not necessary for specifying the post-processed certified region. Applying it to translation invariance shows that $h(\boldsymbol{X}') = y^*$ if $\left|\left|\boldsymbol{\Delta} - \boldsymbol{1}_N\overline{\boldsymbol{\Delta}}\right|\right|_2 < r$, where $\overline{\boldsymbol{X}}, \overline{\boldsymbol{\Delta}} \in \mathbb{R}^{1\times D}$ are column-wise averages. For rotation invariance, Theorem 2 guarantees robustness if $\left|\left|\boldsymbol{X}'\boldsymbol{R}^T - \boldsymbol{X}\right|\right|_2 < r$, where $\boldsymbol{R}$ is an optimal rotation matrix defined by the singular value decomposition of $\boldsymbol{X}^T\boldsymbol{X}'$ [97, 98]. In Appendix C, we prove these results and discuss other invariances.

5.1 Application to randomized smoothing certificates

While the post-processing approach may seem straightforward, it has two practical limitations that we shall resolve shortly. Firstly, one may not be able to obtain a certified region $\mathbb{B}$. For example, no white-box robustness certificate for rotation-invariant models has been proposed thus far. Secondly, randomized smoothing, which can be applied to arbitrary models, only certifies the predictions of *smoothed classifier* f. Recall from Section 4, that we only know *base classifier* g to be invariant under group $\mathbb{T}$. There is no guarantee that f shares the same invariances, i.e. postprocessing $\mathbb{B}$ with $\mathbb{T}$ may yield invalid certificates. We address this issue by proving the following result in Appendix C:

Theorem 3. *Let base classifier $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under a group of transformations $\mathbb{T} \subseteq \{\boldsymbol{X} \mapsto \boldsymbol{P}\eta(\boldsymbol{X}) \mid \boldsymbol{P} \in S(N), \eta \in E(D)\}$, where $S(N)$ is the symmetric group in N dimensions, $E(D)$ is the Euclidean group in D dimensions and $\eta : \mathbb{R}^D \to \mathbb{R}^D$ is applied row-wise. Then the isotropically smoothed classifier f, as defined in Section 3, is also invariant under $\mathbb{T}$.*

In other words: Isotropic Gaussian smoothing preserves invariance to any group of transformations composed of permutation and distance-preserving functions in Euclidean space, i.e. rotation, reflection and translation. We may thus instantiate Theorems 1 and 2 with $h = f$ and $r = \sigma\Phi^{-1}(p_{\boldsymbol{X},y^*})$.

6 Tight gray-box certificates

Now that we have derived the first gray-box certificates for invariant models, one may naturally wonder about their optimality. We answer this by deriving tight gray-box certificates, i.e. the best certificates that can be obtained for prediction y^* using only the invariances of base classifier g under group $\mathbb{T}$ and its prediction probability $p_{\boldsymbol{X},y^*}$ under clean smoothing distribution $\mu_{\boldsymbol{X}}$. To do so, we find the worst-case invariant classifier, i.e. solve $\min_{h\in\mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}[h(\boldsymbol{Z}) = y^*]$ with

$$\mathbb{H}_\mathbb{T} = \left\{ h : \mathbb{R}^{N\times D} \to \mathbb{Y} \mid \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[h(\boldsymbol{Z}) = y^*] \geq p_{\boldsymbol{X},y^*}, \forall \boldsymbol{Z}, \forall \boldsymbol{Z}' \in [\boldsymbol{Z}]_\mathbb{T} : h(\boldsymbol{Z}) = h(\boldsymbol{Z}') \right\}, \quad (2)$$

where $[\boldsymbol{Z}]_\mathbb{T}$ is the *equivalence class* of $\boldsymbol{Z}$ with respect to invariance under group $\mathbb{T}$:

Definition 2 (Equivalence classes)**.** *The equivalence class of an input $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ w.r.t. invariance under a group of transformations $\mathbb{T}$ is $[x]_\mathbb{T} = \{\tau(x) \mid \tau \in \mathbb{T}\}$.*

6.1 Certification methodology

To work with invariance constraints, it is convenient to not think of h as a function, but a family of variables $(h_{\boldsymbol{Z}}) \in \mathbb{Y}$ indexed by $\mathbb{R}^{N\times D}$. The invariance constraint states that all variables from an equivalence class should have the same value, i.e. $\forall \boldsymbol{Z}, \forall \boldsymbol{Z}' \in [\boldsymbol{Z}]_{\mathbb{T}} : h_{\boldsymbol{Z}} = h_{\boldsymbol{Z}'}$. Intuitively, such constraints can be enforced by replacing each occurrence of a variable with a distinct representative of its equivalence class. We propose to formalize this idea by using canonical representations:

Definition 3 (Canonical representation)**.** *A canonical representation for invariance under a group of transformations $\mathbb{T}$ is a function $\gamma : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times D}$ with*

$$\forall \boldsymbol{Z} \in \mathbb{R}^{N\times D} : \gamma(\boldsymbol{Z}) \in [\boldsymbol{Z}]_{\mathbb{T}}, \tag{3}$$

$$\forall \boldsymbol{Z} \in \mathbb{R}^{N\times D}, \forall \boldsymbol{Z}' \in [\boldsymbol{Z}]_{\mathbb{T}} : \gamma(\boldsymbol{Z}) = \gamma(\boldsymbol{Z}'). \tag{4}$$

In Appendix D.3, we prove that canonical representations let us discard the invariance constraints:

Theorem 4. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under group $\mathbb{T}$ and let $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. If $\gamma : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times D}$ is a canonical representation for invariance under $\mathbb{T}$, then*

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \min_{h:\mathbb{R}^{N\times D}\to\mathbb{Y}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\gamma(\boldsymbol{Z})) = y^*] \; s.t. \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}} [h(\gamma(\boldsymbol{Z})) = y^*] \geq p_{\boldsymbol{X},y^*}.$$

After that, we can perform a carefully chosen substitution $\xi : \mathbb{R}^{N\times D} \to \mathbb{R}^K$ for some $K \in \mathbb{N}$ to modify the probability integrals and show that applying the canonical representation is equivalent to mapping $\mathbb{R}^{N\times D}$ onto a lower-dimensional subspace of $\xi\left(\mathbb{R}^{N\times D}\right)$ that is occupied by the representatives of the different equivalence classes. We can then find the marginal distribution over this subspace and apply the Neyman-Pearson lemma [96] from statistical testing to solve our problem exactly, i.e. find the worst-case classifier. We present a more formal discussion of our method in Appendix D.2.

6.2 Translation invariance

Applying our procedure to translation invariance (see Appendix D.4) shows the following:

Theorem 5. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be translation invariant and $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. Then*

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \Phi\left(\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right) - \frac{\left|\left|\boldsymbol{\Delta} - \boldsymbol{1}_N\overline{\boldsymbol{\Delta}}\right|\right|_2}{\sigma}\right),$$

where $\overline{\boldsymbol{\Delta}} \in \mathbb{R}^{1\times D}$ are the column-wise average of $\Delta = \boldsymbol{X}' - \boldsymbol{X}$ and σ is the standard deviation of the isotropic matrix normal smoothing distribution $\mu_{\boldsymbol{X}}$.

Substituting into the robustness condition $\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] > \frac{1}{2}$ and solving for $\left|\left|\boldsymbol{\Delta} - \boldsymbol{1}_N\overline{\boldsymbol{\Delta}}\right|\right|_2$ shows that $f(\boldsymbol{X}') = y^*$ if $\left|\left|\boldsymbol{\Delta} - \boldsymbol{1}_N\overline{\boldsymbol{\Delta}}\right|\right|_2 < \sigma\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right)$. We see that this certificate is identical to its post-processing-based counterpart from Section 5. In other words: The post-processing-based certificate for translation invariance is provably tight.

6.3 Rotation invariance in 2D

Considering the previous result, one may also suspect the post-processing-based certificate for rotation invariance, i.e. invariance under $\left\{\boldsymbol{X} \to \boldsymbol{X}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(D)\right\}$, to be tight. This is not the case:

Theorem 6. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be translation invariant and $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. Further assume that $\nexists \boldsymbol{R} \in SO(D) : \boldsymbol{X}'\boldsymbol{R}^T = \boldsymbol{X}$ and that $p_{\boldsymbol{X},y^*} \in (0,1)$. Then, for all $\boldsymbol{R} \in SO(D)$:*

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] \gneq \Phi\left(\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right) - \frac{\left|\left|\boldsymbol{X}'\boldsymbol{R}^T - \boldsymbol{X}\right|\right|_2}{\sigma}\right). \tag{5}$$

We present the proof in Appendix E. Our certification procedure based on canonical representations, however, lets us derive the tight, strictly stronger certificates for rotation invariance in 2D (see Appendix D.5). In the following, let $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_{\mathrm{F}}$ be the Frobenius inner product $\sum_{n=1}^{N} \boldsymbol{A}_n^T \boldsymbol{B}_n$ and let $\boldsymbol{R}(\theta) \in SO(2)$ be the matrix that rotates counter-clockwise by angle θ.

Theorem 7. *Let $g : \mathbb{R}^{N\times 2} \to \mathbb{Y}$ be rotation invariant and $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. Then*

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \Pr_{\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q}) \leq t] \tag{6}$$

$$\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(2)},\boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q}) \leq t] = p_{\boldsymbol{X},y^*} \tag{7}$$

$$\text{and } \rho(\boldsymbol{q}) = \mathcal{I}_0\left(\sqrt{q_1^2+q_2^2}\right) / \mathcal{I}_0\left(\sqrt{q_3^2+q_4^2}\right), \tag{8}$$

where $\mathcal{I}_0$ is the modified Bessel function of the first kind and order 0*, and the entries of $\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)} \in \mathbb{R}^4$, $\boldsymbol{\Sigma} \in \mathbb{R}^{4\times 4}$ are linear combinations (see Appendix D.5) of clean data norm $||\boldsymbol{X}||_2$, perturbation norm $||\boldsymbol{\Delta}||_2$ and parameters $\epsilon_1 = \langle \boldsymbol{X}, \boldsymbol{\Delta}\rangle_{\mathrm{F}}$, $\epsilon_2 = \left\langle \boldsymbol{X}\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T, \boldsymbol{\Delta}\right\rangle_{\mathrm{F}}$.*

Because this result is somewhat more involved than the previous ones, let us first examine its parameters, then explain how to evaluate it and finally discuss its connection to prior work.

Certificate parameters. Different from the black-box certificate, which only depends on $||\boldsymbol{\Delta}||_2$, this certificate also depends on the clean data norm $||\boldsymbol{X}||_2$. It further depends on ϵ_1 and ϵ_2, Frobenius inner products between $\boldsymbol{\Delta}$, and the clean input $\boldsymbol{X}$ before and after a rotation by $-\frac{\pi}{2}$. These parameters capture the orientation of $\boldsymbol{X}'$ relative to $\boldsymbol{X}$, as one would expect from a rotation-invariance aware certificate. Note that ϵ_1 and ϵ_2 always fulfill $\sqrt{\epsilon_1^2+\epsilon_2^2} \leq ||\boldsymbol{X}||_2||\boldsymbol{\Delta}||_2$ (see Appendix H).

Certificate evaluation. Evidently, we do not have a closed-form analytical expression for the probabilities in Eqs. (6) and (7). However, recall from Section 3 that randomized smoothing already involves an intractable probability, namely $p_{\boldsymbol{X},y^*}$. One has to use Monte Carlo sampling to compute a lower confidence bound that hold with high probability $1-\alpha$. We propose to adopt the same approach: First, we bound threshold t from Eq. (7). Then, we bound probability $\Pr_{\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q}) \leq t]$ from Eq. (6). We discuss the full algorithm and how to ensure that all bounds simultaneously hold in Appendix D.9. Because the probabilities only involve 4-dimensional normal distributions and do not depend on base classifier g, one can use a large number of samples to obtain narrow bounds at little computational cost (e.g. 0.59s for 100000 samples per confidence bound on an Intel Xeon E5-2630 CPU).

Connection to prior work. During our derivation of Theorem 7 (see Appendix D.5), we show that the worst-case invariant classifier, i.e. the minimizer of our optimization problem, is a function that predicts y^* if and only if $\frac{\beta(\boldsymbol{X}',\boldsymbol{Z})}{\beta(\boldsymbol{X},\boldsymbol{Z})} \leq t$ with $\beta(\boldsymbol{X},\boldsymbol{Z}) = \int_0^{2\pi} \exp\left(\frac{1}{\sigma^2}\left\langle \boldsymbol{X}, \boldsymbol{Z}\boldsymbol{R}(\theta)^T\right\rangle_{\mathrm{F}}\right) d\theta$. Different from black-box randomized smoothing, where the worst-case classifier is a linear model [27], this classifier evaluates Gaussian kernels of $(\boldsymbol{X},\boldsymbol{Z})$ and $(\boldsymbol{X}',\boldsymbol{Z})$ and averages them over all possible rotations. This *group averaging* is a key technique for building invariant models [7, 99–103] Group-averaged kernels have been proposed in [16] as "Haar integration kernels". It is fascinating to see them naturally materialize from nothing but an invariance constraint.

6.4 Generalizations

In Appendix D.6 we generalize our result to elemental rotations in 3D, i.e. rotations from $SO(3,d) := \{\boldsymbol{R} \mid \boldsymbol{R} \in SO(3), \boldsymbol{R}_d = \boldsymbol{e}_d\}$ with indicator vector e_d. The certificate differs from Theorem 7 in that it also depends on the amount of perturbation along rotation axis d, i.e. $||\boldsymbol{\Delta}_{:,d}||_2$. For $||\boldsymbol{\Delta}_{:,d}||_2 = 0$, both certificates are identical. Note that every model that is invariant to arbitrary rotations is also invariant to elemental rotations. In Appendices D.7 and D.8 we further prove that additionally enforcing translation invariance is equivalent to centering $\boldsymbol{X}$ and $\boldsymbol{\Delta}$ before evaluating the certificates for rotation invariance, which is consistent with our results from Section 6.2.

7 Limitations and broader impact

The main limitation of our work lies in its exploratory nature. In Section 5, we have derived the first certificates for models invariant under groups composed of permutation and arbitrary Euclidean isometries. However, there is a vast swath of invariances that we have not covered, such as spatio-temporal invariances [42], invariance under graph isometries [9] or invariances for planar images [7]. Furthermore, we have not yet derived tight certificates for all Euclidean isometries (though our approach based on canonical representations could conceivably be applied to them).

Broader impact. With the growing prevalence of machine learning in safety-critical and sensitive domains like autonomous driving [104, 105] or healthcare [106, 107], models that are not only accurate, but trustworthy promise to become increasingly important. Robustness certification is one pillar of trustworthy machine learning, alongside other concepts like algorithmic fairness [108, 109] or differential privacy [110, 111]. Unique to our work is that we certify robustness for models with spatial invariances. These invariances naturally occur in the physical sciences, including those with a direct societal impact like pharmacology and biochemistry. Our work can be seen as a first step towards provable trustworthiness for tasks like drug discovery [37–39] or protein folding [43–47].

8 Experimental evaluation

We know that the post-processing-based certificate for translation invariance is provably tight and can certify robustness for a region with infinitely larger volume than black-box randomized smoothing (see Fig. 1). We therefore focus our experimental evaluation on comparing the post-processing-based certificates for rotation invariance from Section 5 to the provably tight ones from Section 6. Recall that the post-processed certificate for rotation invariance guarantees robustness for the set $\left\{\boldsymbol{Z}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(D), ||\boldsymbol{Z} - \boldsymbol{X}||_2 \leq r\right\}$ with $r = \sigma\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right)$. In other words: It certifies robustness for perturbations with rotational components that can be eliminated to bring $\boldsymbol{X}'$ into distance r of $\boldsymbol{X}$. We want to understand whether the tight certificates offer any benefit beyond that, or if the strict inequality in Theorem 6 is due to some negligible ϵ. To this end, we first thoroughly examine the four-dimensional parameter space of the tight certificate for 2D rotation invariance and then verify our findings by applying our certificates to rotation invariant point cloud classifiers.

All parameters and experimental details are specified in Appendix B. A reference implementation is made available with the supplementary material. We use 10000 samples per confidence bound and set $\alpha = 0.001$, i.e. all certificates hold with 99.9% probability.

8.1 Tight certificate parameter space

Recall that the tight certificate for rotation invariance depends on $||\boldsymbol{X}||_2$, $||\boldsymbol{\Delta}||_2$ and parameters ϵ_1 and ϵ_2, which capture the orientation of the perturbed point cloud and fulfill $\sqrt{\epsilon_1^2 + \epsilon_2^2} \leq ||\boldsymbol{X}||_2 \cdot ||\boldsymbol{\Delta}||_2$. To avoid clutter, we define $\tilde{\epsilon}_k := \epsilon_k \,/\, (||\boldsymbol{X}||_2 \cdot ||\boldsymbol{\Delta}||_2)$. As our metric for this section, we report $p_{\min}$, the smallest probability $p_{\boldsymbol{X},y^*}$ for which a prediction can still be certified [3] (recall that the base classifier g only affects the certificate through this prediction probability).

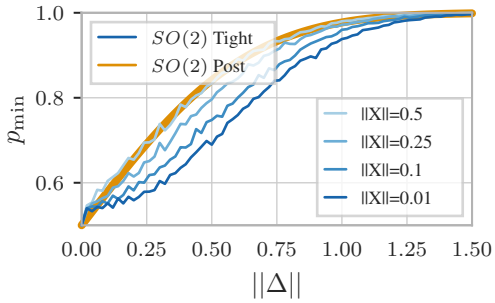


Figure 2: Comparison of tight and post-processing-based certificates applied to adversarial scaling for $\sigma = 0.5$ and varying $||\boldsymbol{\Delta}||$ and $||\boldsymbol{X}||$ (smaller $p_{\min}$ is better). As $||\boldsymbol{X}||$ increases, the difference between the certificates shrinks.

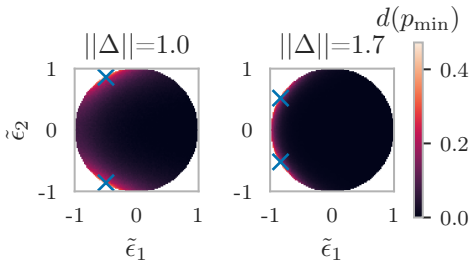


Figure 3: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for $\sigma = 0.5$, $||\boldsymbol{X}|| = 1.0$. Crosses correspond to adversarial rotations. Large $d(p_{\min})$ can be observed near adversarial rotations.

Adversarial scaling. First, we assume that $\boldsymbol{X}' = (1 + \beta)\boldsymbol{X}$, i.e. the input is adversarially scaled. In this case, we have $\tilde{\epsilon}_1 = 1$ and $\tilde{\epsilon}_2 = 0$. We then vary $||\boldsymbol{\Delta}||_2$ and $||\boldsymbol{X}||_2$ and evaluate our certificates. Note that such attacks have no rotational component, i.e. the post-processing-based certificate is identical to the black-box one. Fig. 2 shows that, even in the absence of rotations, the tight certificate

[3] We discuss how to compute these inverse certificates in Appendix G.

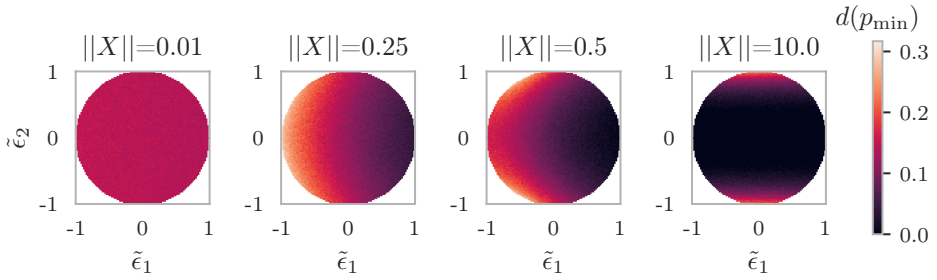


Figure 4: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for $\sigma = 0.5$, $||\boldsymbol{\Delta}|| = 0.5$ under varying $||\boldsymbol{X}||$, $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$. As $||\boldsymbol{X}||$ increases, the regions where the tight certificate outperforms the black-box baseline shrink.

can yield significantly stronger guarantees. For $\sigma = 0.5$ and $||\boldsymbol{X}||_2 = 0.01$, post-processing can only certify robustness for a prediction with $p_{\boldsymbol{X},y^*} = 0.8$ if $||\boldsymbol{\Delta}||_2 \leq 0.4$. The tight certificate can certify robustness up to $||\boldsymbol{\Delta}||_2 = 0.73$. However, the gap shrinks, as the norm of the clean data increases.

Effect of data norm. We would like to see if this is a pervasive pattern. To this end, we fix $||\boldsymbol{\Delta}||_2$, gradually increase $||\boldsymbol{X}||_2$, and evaluate the tight certificate for $\tilde{\epsilon}_1, \tilde{\epsilon}_2$ on a 100×100 rasterization of $[0, 1] \times [0, 1]$. We then measure the difference $d(p_{\min})$ to the black-box certificate (we will discuss how these results relate to the post-processing-based one shortly). Fig. 4 shows that for small $||\boldsymbol{X}||_2$, the tight certificate outperforms the black-box certificate for arbitrary perturbations of norm $||\boldsymbol{\Delta}||_2$. But, as $||\boldsymbol{X}||_2$ increases, the regions where it outperforms the black-box one shrink.

Rotational components. Simple algebra (see Appendix H) shows that for any combination of $||\boldsymbol{X}||_2$ and $||\boldsymbol{\Delta}||_2$, adversarial rotations, i.e. $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{R}^T$ such that $||\boldsymbol{X}' - \boldsymbol{X}|| = ||\boldsymbol{\Delta}||$, correspond to two specific points with $\sqrt{\tilde{\epsilon}_1^2 + \tilde{\epsilon}_2^2} = 1$ in the certificate's parameter space. In Fig. 3 we fix $\sigma = 0.5$ and $||\boldsymbol{X}||_2 = 1$, vary $||\boldsymbol{\Delta}||$ and highlight these two points. We observe that the tight certificate only outperforms the black-box certificate for values of $\tilde{\epsilon}_1, \tilde{\epsilon}_2$ that are close to adversarial rotations. However, robustness to such attacks can also be readily certified via post-processing. Combined with our previous observations, we expect the tight and the post-processing-based certificate to perform similarly well, assuming that the smoothing standard deviation σ is small relative to $||\boldsymbol{X}||_2$.

8.2 Application to point cloud classification

To verify whether our observations hold in practice, we apply our certificates to point cloud classification. We consider two datasets: 3D point cloud representations of ModelNet40 [112], which consists of CAD models from 40 different categories, and 2D point cloud representations of MNIST [113]. We apply the same pre-processing steps as in [6]. As our base classifiers, we use rotation and translation (i.e. $SE(D)$) invariant versions of two well-established models: PointNet [6] and DGCNN [114]. To implement the invariances, we center the input data, perform principal component analysis, apply the

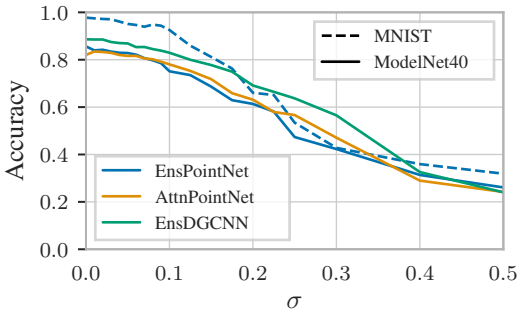


Figure 5: Test set accuracy of smoothed translation and rotation invariant point cloud classifiers on the ModelNet40 and MNIST pointcloud datasets, under varying standard deviation σ.

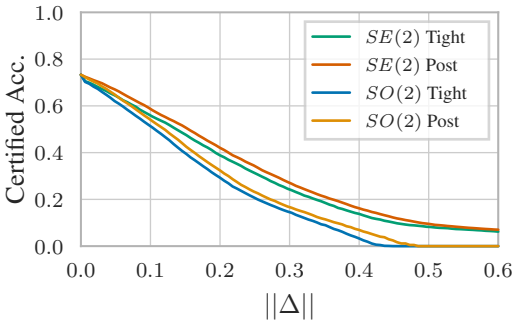
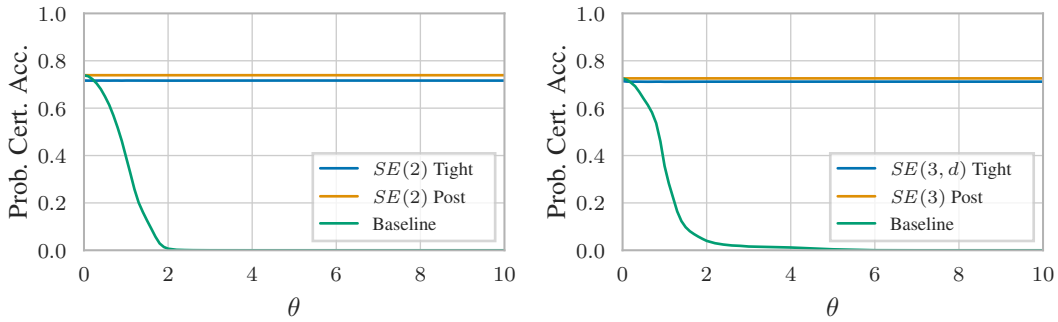


Figure 6: Comparison of certificates for adversarial scaling of MNIST with EnsPointNet and $\sigma = 0.15$. The tight and post-processing-based certificates yield similar certified accuracies.

model to all possible poses (see discussion in [115]) and average the output logits (EnsPointNet and EnsDGCNN). In addition, we consider a more refined model [116] that combines canonical poses via a self-attention mechanism (AttnPointNet). Certification is performed on the default test sets.

Practical smoothing parameters. Fig. 5 shows the test set accuracies of randomly smoothed models under varying standard deviations σ. Values of σ that preserve an accuracy above 50% are small, relative to the average norm of the test sets (10.67 for MNIST, 19.17 for ModelNet40). Going by our previous results, we expect the tight and post-processing based certificates to perform similarly well.

Adversarial scaling. In Fig. 6 we again consider adversarial scaling, i.e. attacks without rotational components, but applied to the MNIST point cloud dataset with $\sigma = 0.15$. We report the certified accuracy, i.e. the percentage of correct and provably robust predictions, for certification with ($SE(2)$) and without ($SO(2)$) translation invariance. The tight and post-processing-based certificate yield similar results. Note that the certified accuracies of the tight method are in fact marginally lower, because we bound the certificates using Monte Carlo sampling (see Section 6.3). We further observe that additionally enforcing translation invariance (i.e. $SE(2)$) strengthens both approaches.



(a) EnsPointNet on MNIST

(b) EnsDGCNN on ModelNet40

Figure 7: Comparison of tight, post-processing-based and black-box certificates for randomly perturbed inputs with $||\mathbf{\Delta}|| = \sigma = 0.1$, rotated by angle θ. The gray-box certificates effectively eliminate the induced rotation, while the black-box method cannot certify robustness for large θ.

Rotational components. Finally, we study perturbations with rotational components. We fix $||\mathbf{\Delta}||$, randomly sample perturbations of the specified norm and then rotate $\boldsymbol{Z} = X + \mathbf{\Delta}$ by a specified angle $\theta \in [0, 10°]$ (in the case of ModelNet40, around one randomly chosen coordinate axis). For each element of the test set and each θ, we generate 10 such samples. We then compute the percentage of samples $\boldsymbol{X}'$ for which $f(\boldsymbol{X})$ is correct and $f(\boldsymbol{X}') = f(\boldsymbol{X})$ is provably guaranteed ('probabilistic certified accuracy"). Fig. 7 shows results for MNIST and ModelNet40 evaluated with $||\mathbf{\Delta}|| = \sigma = 0.1$ ($SE(3, d)$ refers to elemental rotations). The black-box baseline's probabilistic certified accuracy drops close to 0 for $\theta = 2°$. The gray-box certificates are almost constant in θ, i.e. effectively eliminate any induced rotation. However, the tight certificate did not offer any benefit beyond that. We did not observe a single sample for which only the tight certificate could guarantee robustness.

In Appendix A we repeat the experiments from this and the previous section for various other combinations of parameter values. All results are consistent with the ones presented here, confirming that the post-processing approach offers a good approximation of the tight certificates in practice.

9 Conclusion

For the first time, we have studied the use of invariances for robustness certification. We proposed a gray-box approach, combining white-box knowledge about invariances with black-box randomized smoothing. We have derived a post-processing-based procedure for certification that can be applied to arbitrary models with invariance to permutations and Euclidean isometries. We have proven that the post-processed certificate for translation invariance is tight and derived strictly stronger certificates for rotation invariance. Our experiments are to be interpreted in two ways: The fact that it is possible to derive tight invariance-aware certificates and there exist scenarios such that they offer stronger guarantees for arbitrary perturbations should be an exciting inspiration for future work. The fact that post-processing yields semantically meaningful certificates that offer good approximations of our tight certificates should invite its application to real-world tasks with inherent invariances.

References

[1] 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.

[2] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[3] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

[4] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

[5] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

[6] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[7] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

[8] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations*, 2018.

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[10] Kunihiko Fukushima. Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. *IEICE Technical Report, A*, 62(10):658–665, 1979.

[11] K FUKUSHIMA. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol, Cybern*, 36:193–202, 1980.

[12] C Lee Giles and Tom Maxwell. Learning, invariance, and generalization in high-order neural networks. *Applied optics*, 26(23):4972–4978, 1987.

[13] Christopher J Burges and Bernhard Schölkopf. Improving the accuracy and speed of support vector machines. *Advances in neural information processing systems*, 9, 1996.

[14] Olivier Chapelle and Bernhard Schölkopf. Incorporating invariances in non-linear support vector machines. *Advances in neural information processing systems*, 14, 2001.

[15] Dennis DeCoste and Bernhard Schölkopf. Training invariant support vector machines. *Machine learning*, 46(1):161–190, 2002.

[16] Bernard Haasdonk, A Vossen, and Hans Burkhardt. Invariance in kernel methods by haar-integration kernels. In *Scandinavian Conference on Image Analysis*, pages 841–851. Springer, 2005.

[17] Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess. Maximum resilience of artificial neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 251–268. Springer, 2017.

[18] 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*, pages 97–117. Springer, 2017.

[19] Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 269–286. Springer, 2017.

[20] Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pages 5276–5285. PMLR, 2018.

[21] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

[22] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2018.

[23] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *Advances in neural information processing systems*, 31, 2018.

[24] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL): 1–30, 2019.

[25] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Computer Vision – ECCV 2018*, pages 381–397. Springer International Publishing, 2018. doi: 10.1007/978-3-030-01234-2_23.

[26] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 656–672. IEEE, 2019.

[27] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019.

[28] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

[29] Connor W Coley, Regina Barzilay, William H Green, Tommi S Jaakkola, and Klavs F Jensen. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling*, 57(8):1757–1772, 2017.

[30] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[31] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

[32] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020.

[33] Joachim Niemeyer, Franz Rottensteiner, and Uwe Soergel. Contextual classification of lidar data and building object detection in urban areas. *ISPRS journal of photogrammetry and remote sensing*, 87:152–165, 2014.

[34] Aili Wang, Xin He, Pedram Ghamisi, and Yushi Chen. Lidar data classification using morphological profiles and convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(5):774–778, 2018.

[35] Xin He, Aili Wang, Pedram Ghamisi, Guoyu Li, and Yushi Chen. Lidar data classification using spatial transformation and cnn. *IEEE Geoscience and Remote Sensing Letters*, 16(1): 125–129, 2018.

[36] Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A Chapman, Dongpu Cao, and Jonathan Li. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3412–3432, 2020.

[37] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

[38] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, et al. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2020.

[39] Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei. Transfer learning for drug discovery. *Journal of Medicinal Chemistry*, 63(16): 8683–8694, 2020.

[40] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):1–9, 2014.

[41] Dan Guest, Kyle Cranmer, and Daniel Whiteson. Deep learning and its application to lhc physics. *Annual Review of Nuclear and Particle Science*, 68:161–181, 2018.

[42] Alexander Bogatskiy, Brandon Anderson, Jan Offermann, Marwah Roussi, David Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics. In *International Conference on Machine Learning*, pages 992–1002. PMLR, 2020.

[43] Ning Qian and Terrence J Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology*, 202(4):865–884, 1988.

[44] Piero Fariselli, Osvaldo Olmea, Alfonso Valencia, and Rita Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein engineering*, 14(11):835–843, 2001.

[45] Jinbo Xu. Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, 116(34):16856–16865, 2019.

[46] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell systems*, 8(4):292–301, 2019.

[47] Kiersten M Ruff and Rohit V Pappu. Alphafold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167208, 2021.

[48] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[49] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International Conference on Machine Learning*, pages 9561–9571. PMLR, 2020.

[50] Sandesh Kamath, Amit Deshpande, Subrahmanyam Kambhampati Venkata, and Vineeth N Balasubramanian. Can we have it all? on the trade-off between spatial and adversarial robustness of neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[51] Vasu Singla, Songwei Ge, Basri Ronen, and David Jacobs. Shift invariance can reduce adversarial robustness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1858–1871. Curran Associates, Inc., 2021.

[52] Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Higher-order certification for randomized smoothing. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4501–4511. Curran Associates, Inc., 2020.

[53] Alexander Levine, Aounon Kumar, Thomas Goldstein, and Soheil Feizi. Tight second-order certificates for randomized smoothing. *arXiv preprint arXiv:2010.10549*, 2020.

[54] Jan Schuchardt, Aleksandar Bojchevski, Johannes Klicpera, and Stephan Günnemann. Collective robustness certificates. In *9th International Conference on Learning Representations, ICLR*, pages 4–7, 2021.

[55] Jong-Chyi Su, Matheus Gadelha, Rui Wang, and Subhransu Maji. A deeper look at 3d shape classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[56] Daniel Liu, Ronald Yu, and Hao Su. Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2279–2283. IEEE, 2019.

[57] Hang Zhou, Dongdong Chen, Jing Liao, Kejiang Chen, Xiaoyi Dong, Kunlin Liu, Weiming Zhang, Gang Hua, and Nenghai Yu. Lg-gan: Label guided adversarial network for flexible targeted attack of point cloud based deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10356–10365, 2020.

[58] Tzungyu Tsai, Kaichen Yang, Tsung-Yi Ho, and Yier Jin. Robust adversarial objects against deep learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 954–962, 2020.

[59] Kibok Lee, Zhuoyuan Chen, Xinchen Yan, Raquel Urtasun, and Ersin Yumer. Shapeadv: Generating shape-aware adversarial 3d point clouds. *arXiv preprint arXiv:2005.11626*, 2020.

[60] Yiren Zhao, Ilia Shumailov, Robert Mullins, and Ross Anderson. Nudge attacks on point-cloud dnns. *arXiv preprint arXiv:2011.11637*, 2020.

[61] Yuxin Wen, Jiehong Lin, Ke Chen, CL Philip Chen, and Kui Jia. Geometry-aware generation of adversarial point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[62] Jaeyeon Kim, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Minimal adversarial examples for deep learning on 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7797–7806, 2021.

[63] Yiming Sun, Feng Chen, Zhiyu Chen, and Mingjie Wang. Local aggressive adversarial attacks on 3d point cloud. In *Asian Conference on Machine Learning*, pages 65–80. PMLR, 2021.

[64] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9136–9144, 2019.

[65] Atrin Arya, Hanieh Naderi, and Shohreh Kasaei. Adversarial attack by limited point cloud surface modifications. *arXiv preprint arXiv:2110.03745*, 2021.

[66] Jiancheng Yang, Qiang Zhang, Rongyao Fang, Bingbing Ni, Jinxian Liu, and Qi Tian. Adversarial attack and defense on point sets. *arXiv preprint arXiv:1902.10899*, 2019.

[67] Matthew Wicker and Marta Kwiatkowska. Robustness of 3d deep learning in an adversarial setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11767–11775, 2019.

[68] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Pointcloud saliency maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1598–1606, 2019.

[69] Hang Zhou, Kejiang Chen, Weiming Zhang, Han Fang, Wenbo Zhou, and Nenghai Yu. Dup-net: Denoiser and upsampler network for 3d adversarial point clouds defense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1961–1970, 2019.

[70] Xiaoyi Dong, Dongdong Chen, Hang Zhou, Gang Hua, Weiming Zhang, and Nenghai Yu. Self-robust 3d point recognition via gather-vector guidance. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11513–11521. IEEE, 2020.

[71] Daniel Liu, Ronald Yu, and Hao Su. Adversarial shape perturbations on 3d point clouds. In *European Conference on Computer Vision*, pages 88–104. Springer, 2020.

[72] Chengcheng Ma, Weiliang Meng, Baoyuan Wu, Shibiao Xu, and Xiaopeng Zhang. Efficient joint gradient based attack against sor defense for 3d point cloud classification. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1819–1827, 2020.

[73] Jiachen Sun, Karl Koenig, Yulong Cao, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of 3d point cloud classification under adaptive attacks. *arXiv preprint arXiv:2011.11922*, 2020.

[74] Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Pointguard: Provably robust 3d point cloud classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6186–6195, 2021.

[75] Dishanika Dewani Denipitiyage, Thalaiyasingam Ajanthan, Parameswaran Kamalaruban, and Adrian Weller. Provable defense against clustering attacks on 3d point clouds. In *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*, 2021.

[76] Wenda Chu, Linyi Li, and Bo Li. Tpc: Transformation-specific smoothing for point cloud models. *arXiv preprint arXiv:2201.12733*, 2022.

[77] Tobias Lorenz, Anian Ruoss, Mislav Balunović, Gagandeep Singh, and Martin Vechev. Robustness certification for point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7608–7618, 2021.

[78] Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. Certifying geometric robustness of neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[79] Marc Fischer, Maximilian Baader, and Martin Vechev. Certified defense to image transformations via randomized smoothing. *Advances in Neural Information Processing Systems*, 33: 8404–8417, 2020.

[80] Anian Ruoss, Maximilian Baader, Mislav Balunović, and Martin Vechev. Efficient certification of spatial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2504–2513, 2021.

[81] Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. Tss: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 535–557, 2021.

[82] Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Towards verifying robustness of neural networks against a family of semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[83] Daniel Zügner and Stephan Günnemann. Certifiable robustness and robust training for graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 246–256, 2019.

[84] Daniel Zügner and Stephan Günnemann. Certifiable robustness of graph convolutional networks under structure perturbations. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1656–1665, 2020.

[85] Akhilan Boopathy, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Cnn-cert: An efficient framework for certifying robustness of convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3240–3247, 2019.

[86] Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

[87] Aleksandar Bojchevski, Johannes Klicpera, and Stephan Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *International Conference on Machine Learning*, pages 1003–1013. PMLR, 2020.

[88] Alexander J Levine and Soheil Feizi. Improved, deterministic smoothing for l_1 certified robustness. In *International Conference on Machine Learning*, pages 6254–6264. PMLR, 2021.

[89] Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. *arXiv preprint arXiv:1912.09899*, 2019.

[90] Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection with median smoothing. *Advances in Neural Information Processing Systems*, 33:1275–1286, 2020.

[91] Aounon Kumar and Tom Goldstein. Center smoothing: Certified robustness for networks with structured outputs. *Advances in Neural Information Processing Systems*, 34, 2021.

[92] Marc Fischer, Maximilian Baader, and Martin Vechev. Scalable certified segmentation via randomized smoothing. In *International Conference on Machine Learning*, pages 3340–3351. PMLR, 2021.

[93] Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Tao Xie, Ce Zhang, and Bo Li. Provable robust learning based on transformation-specific smoothing. *arXiv preprint arXiv:2002.12398*, 4, 2020.

[94] Binghui Wang, Xiaoyu Cao, Neil Zhenqiang Gong, et al. On certifying robustness against backdoor attacks via randomized smoothing. *arXiv preprint arXiv:2002.11750*, 2020.

[95] Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pages 8230–8241. PMLR, 2020.

[96] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[97] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[98] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[99] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[100] Ryan L Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. *arXiv preprint arXiv:1811.01900*, 2018.

[101] Ryan Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Relational pooling for graph representations. In *International Conference on Machine Learning*, pages 4663–4673. PMLR, 2019.

[102] Omri Puny, Matan Atzmon, Heli Ben-Hamu, Edward J Smith, Ishan Misra, Aditya Grover, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *arXiv preprint arXiv:2110.03336*, 2021.

[103] Dmitry Yarotsky. Universal approximations of invariant maps by neural networks. *Constructive Approximation*, 55(1):407–474, 2022.

[104] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.

[105] Khan Muhammad, Amin Ullah, Jaime Lloret, Javier Del Ser, and Victor Hugo C de Albuquerque. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4316–4336, 2020.

[106] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6): 1236–1246, 2018.

[107] Beau Norgeot, Benjamin S Glicksberg, and Atul J Butte. A call for deep-learning healthcare. *Nature medicine*, 25(1):14–15, 2019.

[108] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[109] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35, 2021.

[110] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[111] Zhanglong Ji, Zachary C Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*, 2014.

[112] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[113] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[114] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[115] Zelin Xiao, Hongxin Lin, Renjie Li, Lishuai Geng, Hongyang Chao, and Shengyong Ding. Endowing deep 3d models with rotation invariance based on principal component analysis. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[116] Feiran Li, Kent Fujiwara, Fumio Okura, and Yasuyuki Matsushita. A closer look at rotation-invariant deep point cloud analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16218–16227, 2021.

[117] Xu Yan. Pointnet/pointnet++ pytorch, 2019. URL `https://github.com/yanx27/Pointnet_Pointnet2_pytorch`.

[118] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *International Conference on Computer Vision (ICCV)*, 2019.

[119] William Jake Johnson. *Comparing Variations of the Neyman-Pearson Lemma*. PhD thesis, Montana State University, 2016.

[120] Francisco Eiras, Motasem Alfarra, M Pawan Kumar, Philip HS Torr, Puneet K Dokania, Bernard Ghanem, and Adel Bibi. Ancer: Anisotropic certification via sample-wise volume maximization. *arXiv preprint arXiv:2107.04570*, 2021.

[121] Emily A Cooper and Hany Farid. A toolbox for the radial and angular marginalization of bivariate normal distributions. *arXiv preprint arXiv:2005.09696*, 2020.

[122] DLMF. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.5 of 2022-03-15. URL `http://dlmf.nist.gov/`. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[123] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[124] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469.

[125] George B Dantzig and Abraham Wald. On the fundamental lemma of neyman and pearson. *The Annals of Mathematical Statistics*, 22(1):87–93, 1951.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We derive the described certificates and our experimental results are consistent with the claims in the abstract and introduction.
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix C for post-processing-based certificates, Appendix D for tight certificates and Appendix E for the non-tightness of post-processing-based certificates for rotation.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code and instructions are made available with the supplemental material. Links to the datasets are provided with the readme file.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] However, each individual robustness certificate is based on confidence intervals, i.e. holds with 99.9% probability w.r.t. randomness of the smoothing distribution.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix B.4.
   (b) Did you mention the license of the assets? [Yes] See Appendix B.4.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Yes, our code is made available with the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We only experimented on synthetic data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We only experimented on synthetic data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]