Invariance-Aware Randomized Smoothing Certificates

Jan Schuchardt
Technical University of Munich
j.schuchardt@tum.de

Stephan Günnemann
Technical University of Munich
s.guennemann@tum.de

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 gray-box certificates based on group orbits, which 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 orbit-based 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 [10–15].

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. [16–23]) 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 [24–26] – 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. By combining white-box knowledge about invariances with black-box knowledge about prediction probabilities we obtain *gray-box* certificates. We first derive a gray-box certificate that follows a post-processing paradigm: It takes an existing black-box certificate and augments the certified region using information about the model's invariances. This *orbit-based* certificate serves as a baseline for the provably tight gray-box certificates we derive in subsequent sections.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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 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 [27–34], LiDAR classification [35–38], drug discovery [39–41], particle physics [42–44] and protein folding [45–49].

Our main contributions are

- the first study on the interplay of invariance and certifiable robustness,
- a principled method for deriving tight invariance-aware randomized smoothing certificates,
- tight certificates for models invariant to translations and rotations.

We further demonstrate experimentally that the orbit-based certificates offer 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 [50] for a principled, high-level introduction into the realm of learning with invariances and equivariances.

Invariance and robustness. Two recent empirical studies [51, 52] 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 study models that are invariant by design. In [53], the negative effect of translation-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 certificates. While we propose the first gray-box certificate for invariant models, there exists prior work on combining white-box knowledge with black-box certification. In [54] and [55], knowledge about a classifier's gradients is used to derive tighter randomized smoothing certificates. In [56], knowledge about a graph neural network's receptive fields is used to derive collective gray-box certificates for multiple predictions. In [57], the message-passing scheme of graph neural networks is used for gray-box certification against adversaries that control all features of multiple nodes.

Adversarial attacks on spatial data. Adversarial attacks on spatial data either modify [58–66], insert [65, 67, 68], or delete [69–71] points in space and have been particularly actively studied for point cloud-classification. While the earliest work simply adopted gradient-based attacks from the image domain [58] more recent work has developed a rich assortment of domain-specific methods, for example to preserve object smoothness [64, 72], leverage critical points [69, 70] or craft physically realizable attacks (e.g. to attack models through LiDAR sensors) [61, 73–75]. Of particular note are attacks via isometries (e.g. rotations and translations) [76, 77], whose effectiveness motivates the use of invariant models. Note that we certify the robustness of invariant models to arbitrary point modification attacks – not just isometry attacks (see also "orthogonal research directions" below). Like in other domains, empirical defenses have been proposed [59, 78, 79] and subsequently broken [80–82], motivating the development of black-box [83–85] and white-box [86] 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 [83] or as a special case of the respective certificate [85].

Orthogonal research directions. One related but orthogonal research direction is transformation-specific certification [23, 85–93]. There, a model is assumed to potentially change its prediction under adversarial parametric transformations (e.g. rotations) and one certifies robustness for specific parameter ranges. White-box methods [23, 86–89] over-approximate the set of inputs reachable by a transformation and then propagate it through a model using existing white-box techniques. Black-box methods – namely transformation-specific randomized smoothing [90, 94] – randomize the transformation parameters to provide robustness guarantees for arbitrary models via statistical methods. Later work generalized this principle to spatial data [85], vector-field deformations [92] and multiplicative parameters [93]. Different from these methods, we assume our model to be invariant under a set of transformations, i.e. never change its prediction. We use this property as a tool for certification against arbitrary perturbations. Aside from that, there exist white-box certification techniques for specific *operations with invariances* (e.g. global max-pooling [86], message passing [95, 96] and batch normalization [97]). But prior work does not use or even discuss this property – it treats these invariant

operations as coincidental building blocks of the models it is trying to certify. Our work is the first to study *invariance itself* in the context of provable robustness and how to leverage it for certification.

3 Background

3.1 Randomized smoothing

Randomized smoothing is a black-box certification technique that can be adapted to various data types, tasks and threat models [98–103, 91, 104–106]. 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. [26] and its generalization to matrix data [83, 85].

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 $p_{\boldsymbol{X},y} = \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[g(\boldsymbol{Z}) = y]$ 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 and $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$ a perturbed input. One can show $f(\boldsymbol{X}') = y^*$ by proving that, for perturbed input $\boldsymbol{X}'$, y^* is more likely than all other classes combined. That is, $p_{\boldsymbol{X}',y^*} > 0.5$. A tighter certificate can be obtained by proving that y^* is more likely than the second most likely class, i.e. $p_{\boldsymbol{X}',y^*} > \max_{y'\neq y^*} p_{\boldsymbol{X}',y'}$. For the sake of exposition, we use the first approach throughout the main text and generalize all results to the second one in Appendix H. One can lower-bound $p_{\boldsymbol{X}',y^*}$ by finding the *worst-case classifier* from a set of functions $\mathbb{H}$ with $g \in \mathbb{H}$:

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

For $\mathbb{H} = \left\{h : \mathbb{R}^{N\times D} \to \mathbb{Y} \mid \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[h(\boldsymbol{Z}) = y^*] \geq p_{\boldsymbol{X},y^*}\right\}$, the classifiers that are at least as likely as g to classify $\boldsymbol{X}$ as y^*, the exact solution is given by the Neyman-Pearson lemma [107] (see Appendix F.2). 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 $p_{\boldsymbol{X},y^*}$.

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.

3.2 Group invariance

Let $f : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be a classifier and $\mathbb{T}$ a group, i.e. a set and associated operator $\cdot : \mathbb{T}\times\mathbb{T} \to \mathbb{T}$ that is closed and associative under the operation, has an inverse t^{-1} for each element $t \in \mathbb{T}$ and features an identity element e. Further let $\mathbb{T}$ act on the input space via a *group action* $\circ : \mathbb{T}\times\mathbb{R}^{N\times D} \to \mathbb{R}^{N\times D}$ that preserves the group structure, i.e. $(t \cdot t') \circ \boldsymbol{X} = t \circ (t' \circ \boldsymbol{X})$. Group actions naturally partition the domain into *orbits*, i.e. sets that can be reached by applying group actions to inputs:

Definition 1 (Orbits)**.** *The orbit of an input $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ w.r.t. a group $\mathbb{T}$ is $[\boldsymbol{X}]_{\mathbb{T}} = \{t \circ \boldsymbol{X} \mid t \in \mathbb{T}\}$.*

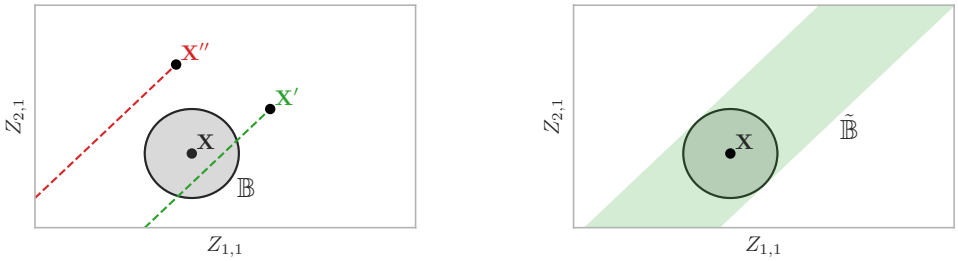
Classifier f is said to be invariant under group $\mathbb{T}$ if $\forall \boldsymbol{X} \in \mathbb{R}^{N\times D}, \forall \boldsymbol{X}' \in [\boldsymbol{X}]_{\mathbb{T}} : f(\boldsymbol{X}) = f(\boldsymbol{X}')$.

3.3 Haar measures

Haar measures [108] are a generalization of Lebesgue measures for integration over groups. Their key property is invariance, meaning the group operator does not affect the measure.

Definition 2 (Right Haar measure)**.** *Let η be a finite, regular measure on the Borel subsets of $\mathbb{T}$. If $\eta(\mathbb{S}) = \eta(\{s \cdot t \mid s \in \mathbb{S}\})$ for all $t \in \mathbb{T}$ and Borel subsets $\mathbb{S} \subseteq \mathbb{T}$, then η is a right Haar measure.*

For the groups we consider, the Haar measure is unique up to a multiplicative constant [109, 110]. In Section 6.1 we use Haar measures as a notational tool to summarize certificates for different groups. However, an understanding of measure theory is not required to follow any part of the derivations.



(a) Robustness for individual points

(b) Augmented certified region

Figure 1: Orbit-based gray-box certificate 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 augmented region $\tilde{\mathbb{B}}$, the union over all translations of $\mathbb{B}$.

4 Problem setting

We consider a similar setup to that described in Section 3.1, 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 is invariant under a group $\mathbb{T}$. This does not necessarily mean that the *smoothed classifier* f shares the same invariances. But, we can use the isotropy of smoothing distribution $\mu_{\boldsymbol{X}}$ to prove (see Appendix D) that randomized smoothing preserves invariance under Euclidean isotropies (rotation, reflection and translation) and permutation, which we shall leverage in Section 5.

Theorem 1. *Let base classifier $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under group $\mathbb{T}$ with $\mathbb{T} \subseteq E(D)$ or $\mathbb{T} \subseteq S(N)$, where $E(D)$ is the Euclidean group and $S(N)$ is the permutation group. Then the isotropically smoothed classifier f, as defined in Section 3.1, is also invariant under $\mathbb{T}$.*

Note that the above result also holds for subgroups of $E(D)$, such as the translation group $T(D)$, the rotation group $SO(D)$ and the roto-translation group $SE(D)$. We define these groups and their actions on $\mathbb{R}^{N\times D}$ more formally in Appendix C.

5 Orbit-based gray-box certificates

Since we are the first to consider robustness certification for invariant models, we begin by defining a baseline that other certificates can be benchmarked against: It 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 group $\mathbb{T}$ corresponds to sets of *transformations* that preserve predictions. By transitivity, a perturbed input $\boldsymbol{X}'$ that can reach $\mathbb{B}$ via a transformation from $\mathbb{T}$, i.e. $\exists t \in \mathbb{T} : t \circ \boldsymbol{X}' \in \mathbb{B}$, must fulfill $f(\boldsymbol{X}') = y^*$ and cannot be an adversarial example. Combining all such inputs, i.e. combining the orbits of elements of $\mathbb{B}$ (see Definition 1), yields an augmented certified region $\tilde{\mathbb{B}} \supseteq \mathbb{B}$ (proof in Appendix E.1):

Theorem 2 (Orbit-based certificates)**.** *Let $f \in \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under a group $\mathbb{T}$. Let $y^* = f(\boldsymbol{X})$ be a prediction that is certifiably robust to a set of perturbed inputs $\mathbb{B} \subseteq \mathbb{R}^{N\times D}$, i.e. $\forall \boldsymbol{Z} \in \mathbb{B} : f(\boldsymbol{Z}) = y^*$. Let $\tilde{\mathbb{B}} = \cup_{\boldsymbol{Z}\in\mathbb{B}}[\boldsymbol{Z}]_{\mathbb{T}}$. Then $\forall \boldsymbol{X}' \in \tilde{\mathbb{B}} : f(\boldsymbol{X}') = y^*$.*

This orbit-based approach is illustrated in Fig. 1. While we focus on randomized smoothing, Theorem 2 holds for arbitrary models and certified regions $\mathbb{B}$. However, obtaining $\mathbb{B}$ via other means may not be possible. For instance, there exist no white-box certificates for rotation-invariant models.

Because randomized smoothing preserves invariance to Euclidean isometries and permutation (see Theorem 1), we may apply the orbit-based approach with $\mathbb{T} \subseteq E(D)$ or $\mathbb{T} \subseteq S(N)$ to the certified region $\mathbb{B} = \{\boldsymbol{X} + \boldsymbol{\Delta} \mid ||\boldsymbol{\Delta}||_2 < r\}$ with $r = \sigma\Phi^{-1}(p_{\boldsymbol{X},y^*})$ that we derived in Section 3.1.

While $\tilde{\mathbb{B}} = \cup_{\boldsymbol{Z}\in\mathbb{B}}[\boldsymbol{Z}]_{\mathbb{T}}$, is a valid certificate, one may desire an equivalent, but more explicit characterization of the certified region $\tilde{\mathbb{B}}$ to determine whether a specific perturbed input preserves the prediction y^*. We discuss such explicit characterizations in Appendix E. In particular, translation invariance (i.e. $\mathbb{T} = T(D)$) leads to certified region $\tilde{\mathbb{B}} = \left\{\boldsymbol{X} + \boldsymbol{\Delta} \mid ||\boldsymbol{\Delta} - \mathbf{1}_N\overline{\boldsymbol{\Delta}}||_2 < r\right\}$, where

$\overline{\boldsymbol{\Delta}} \in \mathbb{R}^{1\times D}$ are column-wise averages and r is defined as above. Rotation invariance (i.e. $\mathbb{T} = SO(D)$) leads to certified region $\tilde{\mathbb{B}} = \left\{\boldsymbol{X}' \mid ||\boldsymbol{X}'\boldsymbol{R}^T - \boldsymbol{X}||_2 < r\right\}$, where $\boldsymbol{R}$ is an optimal rotation matrix defined by the singular value decomposition of $\boldsymbol{X}^T\boldsymbol{X}'$ [111, 112] (see Appendix E.2.2).

6 Tight gray-box certificates

Now that we have established a baseline for gray-box certification, one may naturally wonder about its 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 $\mu_{\boldsymbol{X}}$. Similar to Section 3.1, we do so by finding a worst-case classifier. In addition to constraining its clean prediction probability, we constrain the classifier to be invariant under $\mathbb{T}$, 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 orbit of $\boldsymbol{Z}$ w.r.t. $\mathbb{T}$ (see Definition 1). In the following, we show how to solve the above optimization problem and then apply our certification methodology to specific invariances.

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 orbit 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, the constraint can be enforced by replacing all these variables with a single variable. We propose to formalize this idea by using canonical maps, which map all inputs from an orbit to a single, distinct representative:

Definition 3 (Canonical map)**.** *A canonical map 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*

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

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

In Appendix F.1, we prove that canonical maps let us discard the invariance constraints:

Lemma 1. *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 map for invariance under $\mathbb{T}$, then*

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

Like in Section 3.1, the optimization problem without invariance constraints from Lemma 1 could now be solved exactly using the Neyman-Pearson lemma – if we could find the distribution of $\gamma(\boldsymbol{Z})$, i.e. the distribution of representatives[1]. This can be achieved for groups $T(D)$, $SO(D)$ and $SE(D)$ via careful change of variables into an alternative parameterization of $\mathbb{R}^{N\times D}$ (see Appendix F.3).

This leads us to our main result, which we derive more formally in Appendix F.3. 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 recall from Section 3.3 that a right Haar measure is a measure for integration over a group $\mathbb{T}$.

Theorem 3. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under $\mathbb{T}$ with $\mathbb{T}$ chosen from $\{T(D), SO(D), SE(D)\}$. For $SO(D)$ and $SE(D)$, let $D \in \{2, 3\}$. Let $\mathbb{H}_\mathbb{T}$ be defined as in Eq.* (2) *and η be a right Haar measure on $\mathbb{T}$. Define the indicator function $h^* : \mathbb{R}^{N\times D} \to \{0, 1\}$ with*

$$h^*(\boldsymbol{Z}) = \mathbb{1}\left[\frac{\beta_{\boldsymbol{X}'}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}(\boldsymbol{Z})} \leq \kappa\right], \text{ where } \beta_{\boldsymbol{X}}(\boldsymbol{Z}) = \int_{t\in\mathbb{T}} \exp\left(\langle t\circ\boldsymbol{Z}, \boldsymbol{X}\rangle_\mathrm{F} / \sigma^2\right) d\eta(t) \quad (5)$$

$$\text{and } \kappa \in \mathbb{R} \text{ such that } \mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[h^*(\boldsymbol{Z})] = p_{\boldsymbol{X},y^*}. \quad (6)$$

Then

$$\min_{h\in\mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}[h(\boldsymbol{Z}) = y^*] = \mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}[h^*(\boldsymbol{Z})]. \quad (7)$$

[1] which, by Definition 3, is equivalent to a distribution over orbits.

The indicator function h^* in Eq. (5) corresponds to the worst-case invariant classifier for clean prediction probability $p_{\boldsymbol{X},y^*}$. To classify an input sample $\boldsymbol{Z}$, it integrates Gaussian kernels of $(t \circ \boldsymbol{Z}, \boldsymbol{X}')$ and $(t \circ \boldsymbol{Z}, \boldsymbol{X})$ over group $\mathbb{T}$ (e.g. over all possible rotations). If the ratio of these integrals is below a threshold κ, it classifies $\boldsymbol{Z}$ as y^*. The constraint in Eq. (6) ensures that the probability of predicting y^* under the clean smoothing distribution $\mu_{\boldsymbol{X}}$ matches that of the actual base classifier g. The expected value with respect to perturbed smoothing distribution $\mu_{\boldsymbol{X}'}$ in Eq. (7) is the optimal value of our optimization problem. As discussed in Section 3.1, the prediction is certifiably robust if this optimal value is greater than $1/2$.

Applying the certificate to a group $\mathbb{T}$ requires three steps: 1.) Calculating the Haar integrals in Eq. (5). 2.) Solving Eq. (6) for classification threshold κ. 3.) Evaluating the expected value in Eq. (7).

Connection to prior work. This result differs from black-box randomized smoothing with Gaussian noise, where the worst-case classifier is a linear model [26]. The *group averaging* performed by the worst-case invariant classifier, i.e. integrating a function over group $\mathbb{T}$, is a key technique for building invariant models [7, 113–117]. Group-averaged kernels have been proposed in [15]. It is fascinating to see them naturally materialize from nothing but an invariance constraint.

6.2 Translation invariance

In the case of translation invariance (i.e. $\mathbb{T} = T(D)$), we can evaluate the worst-case classifier (Eq. (5), solve for threshold κ (Eq. (6)) and evaluate the perturbed prediction probability (Eq. (7)) analytically. This leads to the following result (proof in Appendix F.4.1):

Theorem 4. *Let $g : \mathbb{R}^{N \times D} \to \mathbb{Y}$ be invariant under $\mathbb{T} = T(D)$ and $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. Then*

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

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

Certificate parameters. For a fixed smoothing standard deviation σ, the certificate depends on a single parameter: The norm $\left|\left|\boldsymbol{\Delta} - \mathbf{1}_N \overline{\boldsymbol{\Delta}}\right|\right|_2$ of the mean-centered perturbation matrix $\boldsymbol{\Delta}$.

Comparison to the orbit-based certificate. Substituting into robustness condition $\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] > \frac{1}{2}$ shows that $f(\boldsymbol{X}') = y^*$ if $\left|\left|\boldsymbol{\Delta} - \mathbf{1}_N \overline{\boldsymbol{\Delta}}\right|\right|_2 < \sigma \Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right)$. This result, obtained via our tight certification methodology, is identical to the orbit-based certificate for translation invariance from the end of Section 5. In other words: Despite its simplicity, the orbit-based certificate is the best possible gray-box certificate for translation invariance.

6.3 Rotation invariance in 2D

Considering the previous result, one may suspect the orbit-based certificate to be tight for arbitrary invariances. This is not the case. For instance, it is not tight for rotation invariance ($\mathbb{T} = SO(D)$):

Theorem 5. *Let $g : \mathbb{R}^{N \times D} \to \mathbb{Y}$ be invariant under $\mathbb{T} = SO(D)$ and $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. Assume that perturbed input $\boldsymbol{X}'$ is not obtained via rotation of $\boldsymbol{X}$, i.e. $\nexists \boldsymbol{R} \in SO(D) : \boldsymbol{X}' = \boldsymbol{X}\boldsymbol{R}^T$. Further assume that $p_{\boldsymbol{X},y^*} \in (0, 1)$. Then, for all $\boldsymbol{R} \in SO(D)$:*

$$\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] > \Phi\left(\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right) - \frac{1}{\sigma}\left|\left|\boldsymbol{X}'\boldsymbol{R}^T - \boldsymbol{X}\right|\right|_2\right). \tag{8}$$

In other words: The tight certificate is strictly stronger. Proof in Appendix G.

Next, we apply Theorem 3 to obtain the strictly stronger, tight certificate for rotation invariance in 2D. In the following, let $\boldsymbol{R}(\theta) \in SO(2)$ be the matrix that rotates counter-clockwise by angle θ and $\mathcal{I}_0$ be the modified Bessel function of the first kind and order 0. The Haar integral from Eq. (5) that defines the worst-case classifier can be evaluated analytically (see Appendix F.4.2)):

$$\beta_{\boldsymbol{X}}(\boldsymbol{Z}) = \mathcal{I}_0\left(\frac{1}{\sigma^2}\sqrt{\langle \boldsymbol{Z}, \boldsymbol{X}\rangle_{\mathrm{F}}^2 + \langle \boldsymbol{Z}, \boldsymbol{X}\boldsymbol{R}(-\pi/2)^T\rangle_{\mathrm{F}}^2}\right), \tag{9}$$

We can substitute this into Theorem 3 and use the fact that Eq. (9) depends on linear transformations of the matrix normal random variable $\boldsymbol{Z}$ to obtain the following certificate (proof in Appendix F.4.2):

Theorem 6. *Let $g : \mathbb{R}^{N\times 2} \to \mathbb{Y}$ be invariant under $\mathbb{T} = SO(2)$ and $\mathbb{H}_\mathbb{T}$ be defined as in Eq.* (2)*. Define the indicator function $\rho : \mathbb{R}^4 \to \{0, 1\}$ with*

$$\rho(\boldsymbol{q}) = \mathbb{1}\left[\mathcal{I}_0\left(\sqrt{q_1^2 + q_2^2}\right) / \mathcal{I}_0\left(\sqrt{q_3^2 + q_4^2}\right) \leq \kappa\right], \tag{10}$$

$$\text{with } \kappa \in \mathbb{R} \text{ such that } \mathop{\mathbf{E}}_{\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(2)},\boldsymbol{\Sigma}\right)}\left[\rho(\boldsymbol{q})\right] = p_{\boldsymbol{X},y^*}. \tag{11}$$

Then

$$\min_{h\in\mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right] = \mathop{\mathbf{E}}_{\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)}\left[\rho(\boldsymbol{q})\right]. \tag{12}$$

where $\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)} \in \mathbb{R}^4$, $\boldsymbol{\Sigma} \in \mathbb{R}^{4\times 4}$ are linear combinations (see Appendix F.4.2) of $||\boldsymbol{X}||_2^2 \,/\, \sigma^2$, $||\boldsymbol{\Delta}||_2^2 \,/\, \sigma^2$ and parameters $\epsilon_1 = \langle \boldsymbol{X}, \boldsymbol{\Delta}\rangle_\mathrm{F}$, $\epsilon_2 = \langle \boldsymbol{X}\boldsymbol{R}\left(-\pi/2\right)^T, \boldsymbol{\Delta}\rangle_\mathrm{F}$.

Certificate parameters. This certificate depends on the perturbation norm $||\boldsymbol{\Delta}||_2$ and clean data norm $||\boldsymbol{X}||_2$, relative to the smoothing standard deviation σ. It further depends on parameters ϵ_1 and ϵ_2, which are Frobenius inner products between $\boldsymbol{\Delta}$ and the clean input $\boldsymbol{X}$ before and after a rotation by $-\pi/2$. These parameters capture the orientation of $\boldsymbol{X'} = \boldsymbol{X} + \boldsymbol{\Delta}$ relative to $\boldsymbol{X}$, as one would expect from a rotation-invariance aware certificate.

Monte Carlo evaluation. Evidently, we do not have a closed-form expression for the expectations in Eqs. (11) and (12). However, recall from Section 3.1 that randomized smoothing already involves an intractable expectation, namely the prediction probability $p_{\boldsymbol{X},y^*}$. One has to use Monte Carlo sampling to compute a lower confidence bound that hold with high probability $1 - \alpha$. We adopt the same approach: First, we lower-bound threshold κ from Eq. (11), i.e. make the classifier slightly less likely to predict y^*. Then, we lower-bound bound expectation $\mathbf{E}_{\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)}\left[\rho(\boldsymbol{q})\right]$ from Eq. (12). Because the resulting value is slightly smaller than than optimal value of our optimization problem $\min_{h\in\mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right]$, this procedure yields a valid certificate. We discuss the full algorithm and how to ensure that all bounds simultaneously hold in Appendix F.5. Because the expectations only require sampling from four-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).

6.4 Rotation invariance in 3D

To evaluate the tight certificate for 3D rotation invariance (i.e. $\mathbb{T} = SO(3)$), we adopt the same Monte Carlo evaluation approach discussed in Section 6.3. Different from 2D rotation invariance, evaluating the worst-case classifier analytically is not tractable. It can however be evaluated by numerical integration over an Euler angle parameterization of rotation group $SO(3)$ (see Appendix F.4.3).

6.5 Roto-translation invariance in 2D and 3D

In Appendix F.4.4 we prove that additionally enforcing translation invariance (i.e. $\mathbb{T} = SE(D)$) is equivalent to centering $\boldsymbol{X}$ and $\boldsymbol{\Delta}$ before evaluating the certificates for rotation invariance. This is consistent with our result from Section 6.2, i.e. the orbit-based approach being optimal for translation.

7 Limitations and broader impact

The main limitation of our work lies in its exploratory nature. We have derived the first certificates for models invariant under arbitrary Euclidean isometries and/or permutation. While these invariances are of key importance to many practical applications , there is a vast swath of invariances that we have not covered, such as spatio-temporal invariances [44], invariance under graph isomorphisms [9] or invariances for planar images [7]. Furthermore, we have not yet derived tight certificates for reflections and permutations (though our certification methodology could conceivably be applied to them).

Broader impact. With the growing prevalence of machine learning in safety-critical and sensitive domains like autonomous driving [118, 119] or healthcare [120, 121], trustworthy models promise to become increasingly important. Certification is one pillar of trustworthiness, alongside concepts like fairness [122, 123] or differential privacy [124, 125]. Unique to our work is that we certify models

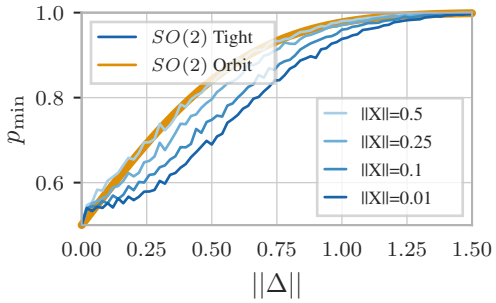


Figure 2: Comparison of tight and orbit-based certificates applied to adversarial scaling for $\sigma = 0.5$ and varying $||\boldsymbol{\Delta}||$ and $||\boldsymbol{X}||$ (smaller $p_{\min}$ is better). As $||\boldsymbol{X}||$ increases, the difference between the certificates shrinks.

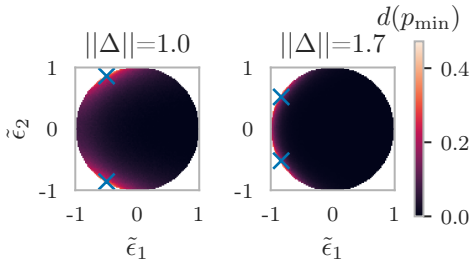


Figure 3: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for $\sigma = 0.5$, $||\boldsymbol{X}|| = 1.0$. Crosses correspond to adversarial rotations. Large $d(p_{\min})$ can be observed near adversarial rotations.

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 [39–41] or protein folding [45–49].

8 Experimental evaluation

We already know that the orbit-based certificate for translation invariance is tight and certifies robustness for an infinitely larger volume than black-box randomized smoothing (see Fig. 1). Therefore, we focus our experiments on the certificates for rotation invariance. Recall that the orbit-based certificate guarantees robustness for the set $\tilde{\mathbb{B}} = \left\{\boldsymbol{X}' \mid ||\boldsymbol{X}'\boldsymbol{R}^T - \boldsymbol{X}||_2 < r\right\}$ with $r = \sigma\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right)$ and optimal rotation matrix $\boldsymbol{R}$. 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 5 is due to some negligible ϵ. To this end, we first thoroughly examine the four-dimensional parameter space of the tight certificate for rotation invariance in 2D, before applying our certificates to rotation invariant point cloud classifiers.

All parameters and experimental details are specified in Appendix B. We use 10000 samples per confidence bound and set $\alpha = 0.001$, i.e. all certificates hold with 99.9% probability. A reference implementation will be made available at https://www.cs.cit.tum.de/daml/invariance-smoothing.

8.1 Tight certificate parameter space

The tight certificate for 2D rotation invariance depends on $||\boldsymbol{X}||_2 / \sigma$, $||\boldsymbol{\Delta}||_2 / \sigma$ and parameters ϵ_1 and ϵ_2, which capture the orientation of the perturbed point cloud and fulfill $\sqrt{\epsilon_1^2 + \epsilon_2^2} \leq ||\boldsymbol{X}||_2 \cdot ||\boldsymbol{\Delta}||_2$ (see Appendix J). 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 [2]

Adversarial scaling. First, we assume that $\boldsymbol{X}' = (1 + c)\boldsymbol{X}$, 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 orbit-based certificate is identical to the black-box one. Fig. 2 shows that, even in the absence of rotations, the tight certificate can yield significantly stronger guarantees. For $\sigma = 0.5$ and $||\boldsymbol{X}||_2 = 0.01$, the baseline 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

[2] We discuss how to compute these inverse certificates in Appendix I.

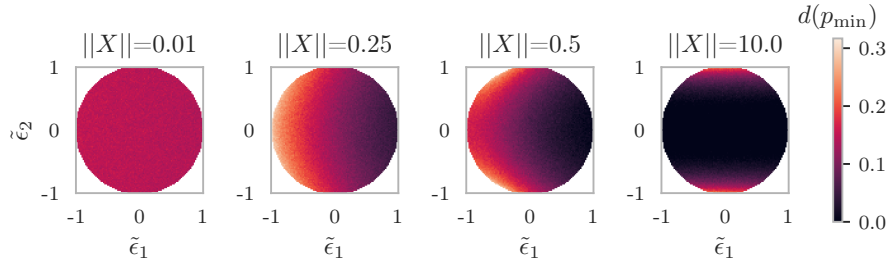


Figure 4: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for $\sigma = 0.5$, $||\boldsymbol{\Delta}|| = 0.5$ under varying $||\boldsymbol{X}||$, $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$. As $||\boldsymbol{X}||$ increases, the regions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

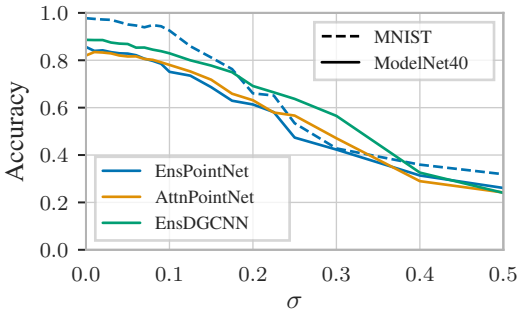


Figure 5: Test set accuracy of smoothed translation and rotation invariant point cloud classifiers on the ModelNet40 and MNIST pointcloud datasets, under varying standard deviation σ.

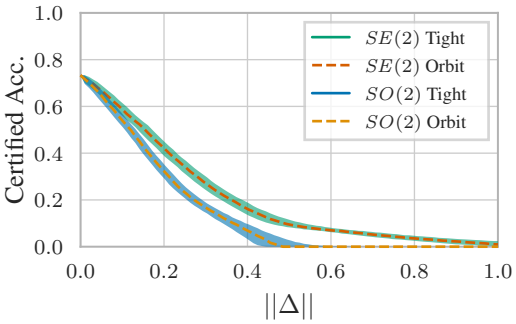


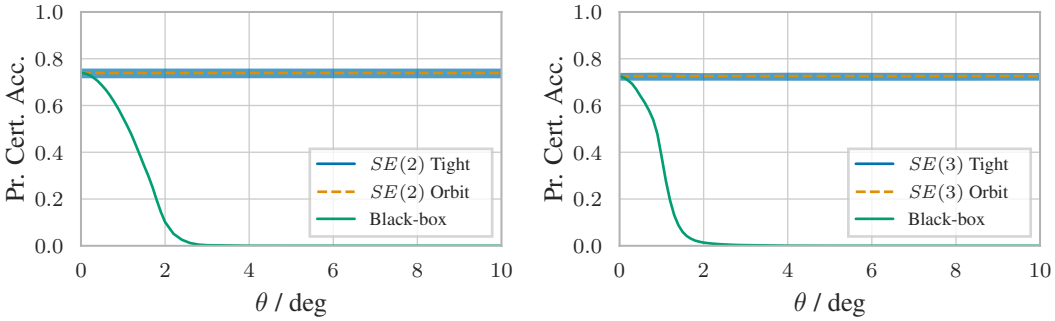
Figure 6: Comparison of certificates for adversarial scaling of MNIST with EnsPointNet and $\sigma = 0.15$. The tight and orbit-based certificates yield similar certified accuracies.

how these results relate to the orbit-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 J) 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 by the orbit-based approach. Combined with our previous observations, and because the certificate only depends on norms relative to smoothing standard deviation σ, i.e. $||\boldsymbol{X}||_2/\sigma$ and $||\boldsymbol{\Delta}||_2/\sigma$, we expect the tight and the orbit-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 [126], which consists of CAD models from 40 different categories, and 2D point cloud representations of MNIST [127]. We apply the same pre-processing steps as in [6]. Certification is performed on the default test sets. As our base classifiers, we use rotation and translation (i.e. $SE(D)$) invariant versions of two well-established models that are used in prior work on robustness certification for point clouds: PointNet [6] and DGCNN [128]. To implement the invariances, we center the input data, perform principal component analysis, apply the model to all possible poses (see discussion in [129, 130]) and average the output logits (EnsPointNet and EnsDGCNN). In addition, we consider a more refined model [129] that combines canonical poses via a self-attention mechanism (AttnPointNet). Throughout this section, we report both probabilistic upper and lower bounds for the Monte Carlo evaluation of the tight certificates (recall Section 6.3).



(a) EnsPointNet on MNIST

(b) EnsDGCNN on ModelNet40

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

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 orbit-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 orbit-based certificate yield similar results. We further observe that enforcing translation invariance increases the certified accuracy and extends the range of $||\Delta||$ for which robustness can be certified.

Rotational components. Finally, we study perturbations with rotational components. We fix $||\boldsymbol{\Delta}||$, randomly sample perturbations of the specified norm and then rotate $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{\Delta}$ by a specified angle $\theta \in [0, 10°]$ (in the case of ModelNet40, around one randomly chosen 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 $||\boldsymbol{\Delta}|| = \sigma = 0.1$. 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 meaningful benefit beyond that. Using the lower bound for Monte Carlo evaluation, there was not 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 orbit-based 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 orbit-based procedure for certification that can be applied to arbitrary models with invariance to permutations and Euclidean isometries. We have proven that the orbit-based certificate for translation invariance is tight and derived strictly stronger certificates for rotation invariance. Our experiments are to be interpreted in two ways: Firstly, the fact that it is possible to derive tight invariance-aware certificates and that there exist scenarios in which they offer stronger guarantees for arbitrary perturbations should be an exciting inspiration for future work. Secondly, the fact that the orbit-based certificates are easily interpretable and offer good approximations of our tight certificates should invite their application to real-world tasks with inherent invariances.

Acknowledgements. The authors would like to thank Johannes Gasteiger for valuable discussions on invariant deep learning and Lukas Gosch for constructive criticism of the manuscript. This research was supported by the German Research Foundation, grant GU 1409/4-1.

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

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

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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*, 2018.

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

[21] 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 *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018.

[22] 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.

[23] 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.

[24] 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. 2018.

[25] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*, pages 656–672. IEEE, 2019.

[26] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

[27] 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.

[28] 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.

[29] 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*, 2017.

[30] 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.

[31] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2019.

[32] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34, 2021.

[33] Johannes Gasteiger, Chandan Yeshwanth, and Stephan Günnemann. Directional message passing on molecular graphs via synthetic coordinates. *Advances in Neural Information Processing Systems*, 34, 2021.

[34] Nicholas Gao and Stephan Günnemann. Ab-initio potential energy surfaces by pairing GNNs with neural wave functions. In *International Conference on Learning Representations*, 2022.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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*, 2020.

[45] 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.

[46] 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.

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

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

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

[50] 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.

[51] 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*, 2020.

[52] 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.

[53] Vasu Singla, Songwei Ge, Basri Ronen, and David Jacobs. Shift invariance can reduce adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[54] Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Higher-order certification for randomized smoothing. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

[56] Jan Schuchardt, Aleksandar Bojchevski, Johannes Klicpera, and Stephan Günnemann. Collective robustness certificates: Exploiting interdependence in graph neural networks. In *International Conference on Learning Representations*, 2021.

[57] Yan Scholten, Jan Schuchardt, Simon Geisler, Aleksandar Bojchevski, and Stephan Günnemann. Randomized message-interception smoothing: Gray-box certificates for graph neural networks. *Advances in Neural Information Processing Systems*, 35, 2022.

[58] 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*, 2018.

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

[60] 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.

[61] 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.

[62] 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.

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

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

[69] 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.

[70] 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.

[71] 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.

[72] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, and Nenghai Yu. Shape-invariant 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15335–15344, June 2022.

[73] Yulong Cao, Chaowei Xiao, Dawei Yang, Jing Fang, Ruigang Yang, Mingyan Liu, and Bo Li. Adversarial objects against lidar-based autonomous driving systems. *arXiv preprint arXiv:1907.05418*, 2019.

[74] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13716–13725, 2020.

[75] Mazen Abdelfattah, Kaiwen Yuan, Z Jane Wang, and Rabab Ward. Towards universal physical attacks on cascaded camera-lidar 3d object detection models. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3592–3596. IEEE, 2021.

[76] Yue Zhao, Yuwei Wu, Caihua Chen, and Andrew Lim. On isometry robustness of deep 3d point cloud models under adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2020.

[77] Wen Shen, Qihan Ren, Dongrui Liu, and Quanshi Zhang. Interpreting representation quality of dnns for 3d point cloud processing. *Advances in Neural Information Processing Systems*, 34, 2021.

[78] 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.

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

[80] 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.

[81] 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.

[82] Jiachen Sun, Karl Koenig, Yulong Cao, Qi Alfred Chen, and Zhuoqing Morley Mao. On adversarial robustness of 3d point cloud classification under adaptive attacks. In *32nd British Machine Vision Conference*, 2021.

[83] 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.

[84] 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.

[85] Wenda Chu, Linyi Li, and Bo Li. TPC: Transformation-specific smoothing for point cloud models. In *International Conference on Machine Learning*, 2022.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] Marc Fischer, Maximilian Baader, and Martin Vechev. Scalable certified segmentation via randomized smoothing. In *International Conference on Machine Learning*, 2021.

[92] Motasem Alfarra, Adel Bibi, Naeemullah Khan, Philip HS Torr, and Bernard Ghanem. DeformRS: Certifying input deformations with randomized smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 6, pages 6001–6009, 2022.

[93] Nikita Muravev and Aleksandr Petiushko. Certified robustness via randomized smoothing over multiplicative parameters of input transformations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3366–3372, 7 2022.

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

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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*, 2020.

[100] Alexander J Levine and Soheil Feizi. Improved, deterministic smoothing for L_1 certified robustness. In *International Conference on Machine Learning*, 2021.

[101] 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.

[102] 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, 2020.

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

[104] 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.

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

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

[107] Jerzy Neyman and Egon Sharpe Pearson. 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.

[108] Alfred Haar. Der Massbegriff in der Theorie der kontinuierlichen Gruppen. *Annals of mathematics*, pages 147–169, 1933.

[109] Henri Cartan. Sur le mesure de Haar. *Comptes Rendus de l'Académie des Sciences de Paris*, 211:759–762, 1940.

[110] Erik M Alfsen. A simplified constructive proof of the existence and uniqueness of Haar measure. *Mathematica Scandinavica*, 12(1):106–116, 1963.

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

[112] 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.

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

[114] 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.

[115] Ryan Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Relational pooling for graph representations. In *International Conference on Machine Learning*, 2019.

[116] Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022.

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

[118] 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.

[119] 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.

[120] 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.

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

[122] 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.

[123] 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.

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

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

[126] 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.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

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

[132] Nico Schlömer. quadpy. URL `https://github.com/sigma-py/quadpy`.

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

[134] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[135] Xiaoshui Huang, Guofeng Mei, Jian Zhang, and Rana Abbas. A comprehensive survey on point cloud registration. *arXiv preprint arXiv:2103.02690*, 2021.

[136] William Jake Johnson. Comparing variations of the neyman-pearson lemma. Master's thesis, Montana State University, 2016.

[137] Mark Aronovich Naimark. *Linear representations of the Lorentz group*. Elsevier, 1964.

[138] Francisco Eiras, Motasem Alfarra, Philip Torr, M. Pawan Kumar, Puneet K. Dokania, Bernard Ghanem, and Adel Bibi. ANCER: Anisotropic certification via sample-wise volume maximization. *Transactions of Machine Learning Research*, 2022.

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

[140] 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.

[141] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[142] Gerald J. Hahn and William Q. Meeker. *Statistical Intervals*. Wiley, August 1991.

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

[144] Kenneth Hung and William Fithian. Rank verification for exponential families. *The Annals of Statistics*, 47(2):758–782, 2019.

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 E for orbit-based certificates, Appendix F for tight certificates and Appendix G for the non-tightness of orbit-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 were submitted with the supplemental material during the review period. A reference implementation will be made available at https://www.cs.cit.tum.de/daml/invariance-smoothing. 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, see 3.a above.
   (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]

A Additional experiments **22**
A.1 Tight certificate parameter space . . . 23
A.2 Application to point cloud classification . . . 28

B Full experimental setup **33**
B.1 Models, data and training parameters . . . 33
B.2 Randomized smoothing parameters . . . 34
B.3 Computational resources . . . 34
B.4 Third-party assets . . . 35

C Group definitions **36**

D Proof of Theorem 1 **37**

E Orbit-based gray-box certificates **39**
E.1 Proof of Theorem 2 . . . 39
E.2 Explicit characterizations . . . 39

F Tight gray-box certificates **41**
F.1 Proof of Lemma 1 . . . 41
F.2 Neyman-Pearson Lemma . . . 42
F.3 Proof of Theorem 3 . . . 43
F.4 Group-specific certificates . . . 51
F.4.1 Translation invariance . . . 51
F.4.2 Rotation invariance in 2D . . . 53
F.4.3 Rotation invariance in 3D . . . 55
F.4.4 Roto-translation invariance in 2D and 3D . . . 57
F.5 Monte Carlo certification . . . 60

G Proof of Theorem 5 **62**

H Multi-class certificates **64**

I Inverse certificates **66**

J Parameter space of the tight certificate for rotation invariance in 2D **67**
J.1 Feasible parameter range . . . 67
J.2 Adversarial rotations . . . 67

A Additional experiments

In the following, we repeat the experiments from Section 8 for a wider range of certificate and model parameters. In Appendix A.1, we investigate the parameter space of the tight certificate for rotation invariance in 2D. In Appendix A.2, we again apply our certificates to rotation invariant point cloud classifiers, this time considering all three models (EnsPointNet, AttnPointNet and EnsDGCNN) and different smoothing distribution parameters. As before, we take 10000 samples per confidence bound and set $\alpha = 0.001$, i.e. all certificates hold with 99.9% probability.

All results support our main conclusion from Section 8: The tight certificates for rotation invariance can offer much stronger robustness guarantees than their orbit-based counterparts. However, in practical scenarios, where the smoothing standard deviation is small relative to the norm of the clean data, the orbit-based approach offers a very good approximation.

A.1 Tight certificate parameter space

For this section, recall that the tight certificate for 2D rotation invariance depends on $||\boldsymbol{X}||_2$, $||\boldsymbol{\Delta}||_2$ and parameters $\epsilon_1 = \langle \boldsymbol{X}, \boldsymbol{\Delta} \rangle_{\mathrm{F}}$, $\epsilon_2 = \left\langle \boldsymbol{X}\boldsymbol{R}\left(-\pi/2\right)^T, \boldsymbol{\Delta} \right\rangle_{\mathrm{F}}$, as well as smoothing standard deviation σ. We define $\tilde{\epsilon}_1 = \frac{\epsilon_1}{||\boldsymbol{X}||\cdot||\boldsymbol{\Delta}||}$ and $\tilde{\epsilon}_2 = \frac{\epsilon_2}{||\boldsymbol{X}||\cdot||\boldsymbol{\Delta}||}$. Note that $\sqrt{\tilde{\epsilon}_1^2 + \tilde{\epsilon}_2^2} \leq 1$ (see also Appendix J).

A.1.1 Adversarial scaling

We begin with adversarial scaling, which corresponds to $\tilde{\epsilon}_1 = 1, \tilde{\epsilon}_2 = 0$. We consider $\sigma \in \{0.1, 0.15, 0.2, 0.25, 0.5, 1.0\}$. For each σ, we vary $||\boldsymbol{X}||_2$ and $||\boldsymbol{\Delta}||$ and compute $p_{\min}$, the smallest value of $p_{\boldsymbol{X},y^*}$ for which a prediction can still be certified (see Appendix I). Lower $p_{\min}$ mean that the certificate can guarantee robustness for models that are less consistent in their predictions.

Fig. 8 shows that, if the clean data norm is small, e.g. $||\boldsymbol{X}|| = 0.01$, then the tight certificate yields much lower $p_{\min}$ than the orbit-based one, save for very small and very large values of $||\boldsymbol{\Delta}||$. That is, the tight certificate for rotation invariance is beneficial even for adversarial perturbations that do not have any rotational components. However, as $||\boldsymbol{X}||$ approaches σ, this difference shrinks, i.e. both approaches offer guarantees of similar strength.

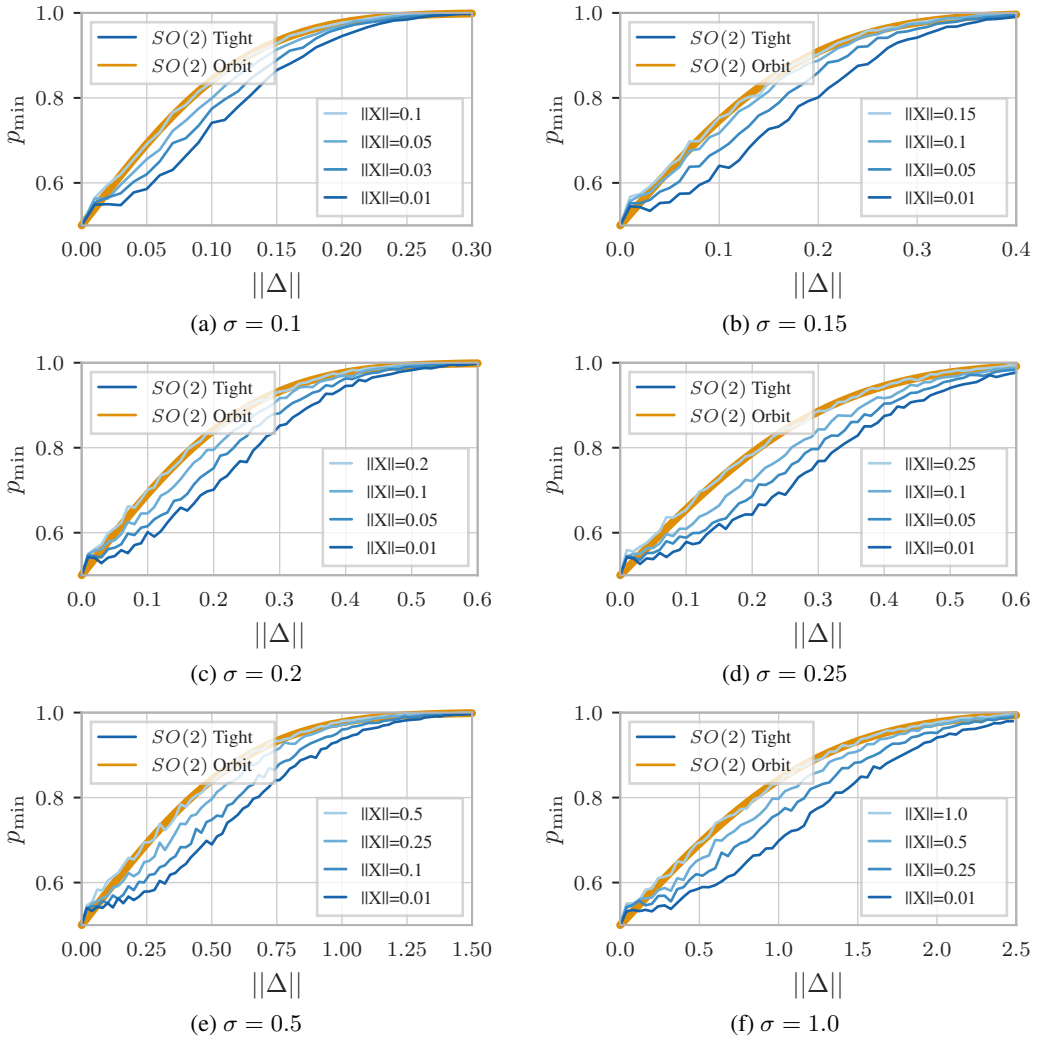


Figure 8: Comparison of tight and orbit-based certificates for rotation invariance in 2D, applied to adversarial scaling for varying $||\boldsymbol{\Delta}||$, $||\boldsymbol{X}||$ and smoothing standard deviation σ (smaller $p_{\min}$ is better). As $||\boldsymbol{X}||$ increases, the difference between the certificates shrinks.

A.1.2 Effect of clean data norm on certificates for arbitrary perturbations

Next, we study whether the same effect can be observed with arbitrary perturbations, i.e. arbitrary values of $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$. In Fig. 4 from Section 8.1, we fixed a specific combination of σ and $||\boldsymbol{\Delta}||$, varied $||\boldsymbol{X}||_2$ and evaluated the tight certificate for $\tilde{\epsilon}_1, \tilde{\epsilon}_2$ taken from a 100×100 rasterization of $[0, 1] \times [0, 1]$. We then measured the difference $d(p_{\min})$ to the black-box certificate.

Here, we consider different combinations of σ, $||\boldsymbol{\Delta}||$ and $||\boldsymbol{X}||$. We can reduce the dimensionality of the parameter space that needs to be explored by observing that both the tight gray-box and the black-box certificate do not depend on the absolute value of these parameters, but their value relative to σ. The black-box baseline has $p_{\min} = \Phi(||\boldsymbol{\Delta}||_2 \,/\, \sigma)$. As specified in Appendix F.4.2, the tight certificate depends on $||\boldsymbol{X}||_2^2 \,/\, \sigma^2$, $||\boldsymbol{\Delta}||_2^2 \,/\, \sigma^2$, $\epsilon_1 \,/\, \sigma^2 = \frac{1}{\sigma^2}||\boldsymbol{X}||_2 \cdot ||\boldsymbol{\Delta}||_2 \cdot \cos(\theta_1)$ and $\frac{1}{\sigma^2} \,/\, ||\boldsymbol{X}||_2 \cdot ||\boldsymbol{\Delta}||_2 \cdot \cos(\theta_1)$, where θ_1, θ_2 are angles between $2N$-dimensional vectors. Therefore, we can fix an arbitrary value of σ and then choose $||\boldsymbol{X}||_2$ and $||\boldsymbol{\Delta}||_2$ relative to it.

We set set $\sigma = 1$ and consider $||\boldsymbol{\Delta}||_2 \in \{\frac{1}{4}\sigma, \frac{1}{2}\sigma, \sigma, 2\sigma, 3\sigma\}$, i.e. perturbation norms that are much smaller, similar to or much larger than the smoothing standard deviation. For each $||\boldsymbol{\Delta}||_2$, we evaluate the tight certificate $||\boldsymbol{X}||_2 \in \{\frac{1}{100}\sigma, \frac{1}{2}\sigma, \sigma, 10\sigma\}$, i.e. clean data norms that are much smaller, similar to or much larger than the smoothing standard deviation.

Fig. 9 shows the resulting $d(p_{\min})$. In the case where $||\boldsymbol{\Delta}||_2 \leq 2||\boldsymbol{X}||_2$ we have additionally highlighted the two values of $(\tilde{\epsilon}_1, \tilde{\epsilon}_2)$ that correspond to adversarial rotations with blue crosses (see Appendix J) Note that, to improve readability, the colorbar is scaled differently for each choice of $||\boldsymbol{\Delta}||$, i.e. each row. Within each row, the same colorbar is used.

We can make four observations.

Firstly, in the case that $||\boldsymbol{\Delta}||_2 \leq \sigma$ and $||\boldsymbol{X}|| = 0.01$ the tight certificate yields lower (i.e. better) $p_{\min}$ for arbitrary perturbations, not just those corresponding to adversarial rotations.

Secondly, as the clean data norm $||\boldsymbol{X}||_2$ increases, the region where the tight certificate outperforms the black-box baseline shrinks. It concentrates around values of $(\tilde{\epsilon}_1, \tilde{\epsilon}_2)$ corresponding to adversarial rotations.

Thirdly, the difference in $p_{\min}$ is not as drastic when $||\boldsymbol{\Delta}||_2$ is small. This is to be expected. For instance, with $||\boldsymbol{\Delta}||_2 = \sigma \,/\, 4$, the black-box certificate has $p_{\min} = \Phi(\sigma \,/\, 4) \approx 0.6$. Here, the difference to the smallest possible value 0.5 is small, meaning there is little room for improvement. We already observed this in Fig. 8.

Finally, when $||\boldsymbol{\Delta}||_2$ is large relative to σ, and so large that $\boldsymbol{\Delta}$ cannot be the result of an adversarial rotation (e.g. $||\boldsymbol{\Delta}||_2 = 3, ||\boldsymbol{X}||_2 = 0.01$), then the tight certificate is almost identical to the black-box baseline. But, as $||\boldsymbol{X}||_2$ increases and an adversarial rotation becomes possible (e.g. $||\boldsymbol{\Delta}||_2 = 3$, $||\boldsymbol{X}||_2 = 10$) then the tight certificate does again yield much better $p_{\min}$ – but only for values of $(\tilde{\epsilon}_1, \tilde{\epsilon}_2)$ that are very close to adversarial rotations.

In summary, these observations support our claim that, if $||\boldsymbol{X}||$ is large relative to σ, then the tight certificate only improves upon the black-box baseline for perturbations that are similar to adversarial rotations. Combined with the fact that the orbit-based gray-box certificate for rotation invariance also accounts for rotational components, this suggests that the orbit-based certificate should be a good approximation in practice.

In Fig. 10, we repeat the same experiments with σ, $||\boldsymbol{X}||_2$ and $||\boldsymbol{\Delta}||_2$ scaled by a factor of $\frac{1}{2}$, to verify that it is indeed sufficient to only consider a single, arbitrary value of σ and the relative value of all other parameters. As expected, the results are identical to Fig. 9.

In Fig. 11 we perform the same experiment for fixed values of σ and $||\boldsymbol{X}||_2$ and two values of $||\boldsymbol{\Delta}||_2$ chosen such that adversarial rotations correspond to $\tilde{\epsilon} =\approx 0.5$ and $\tilde{\epsilon}_2 \approx \pm 0.5$, respectively (similar to Fig. 3 from Section 8.1). In other words: The adversarial rotations move in $30°$-increments in the parameter space. Again, we can observe that the regions with large $p_{\min}$ move with the adversarial rotations.

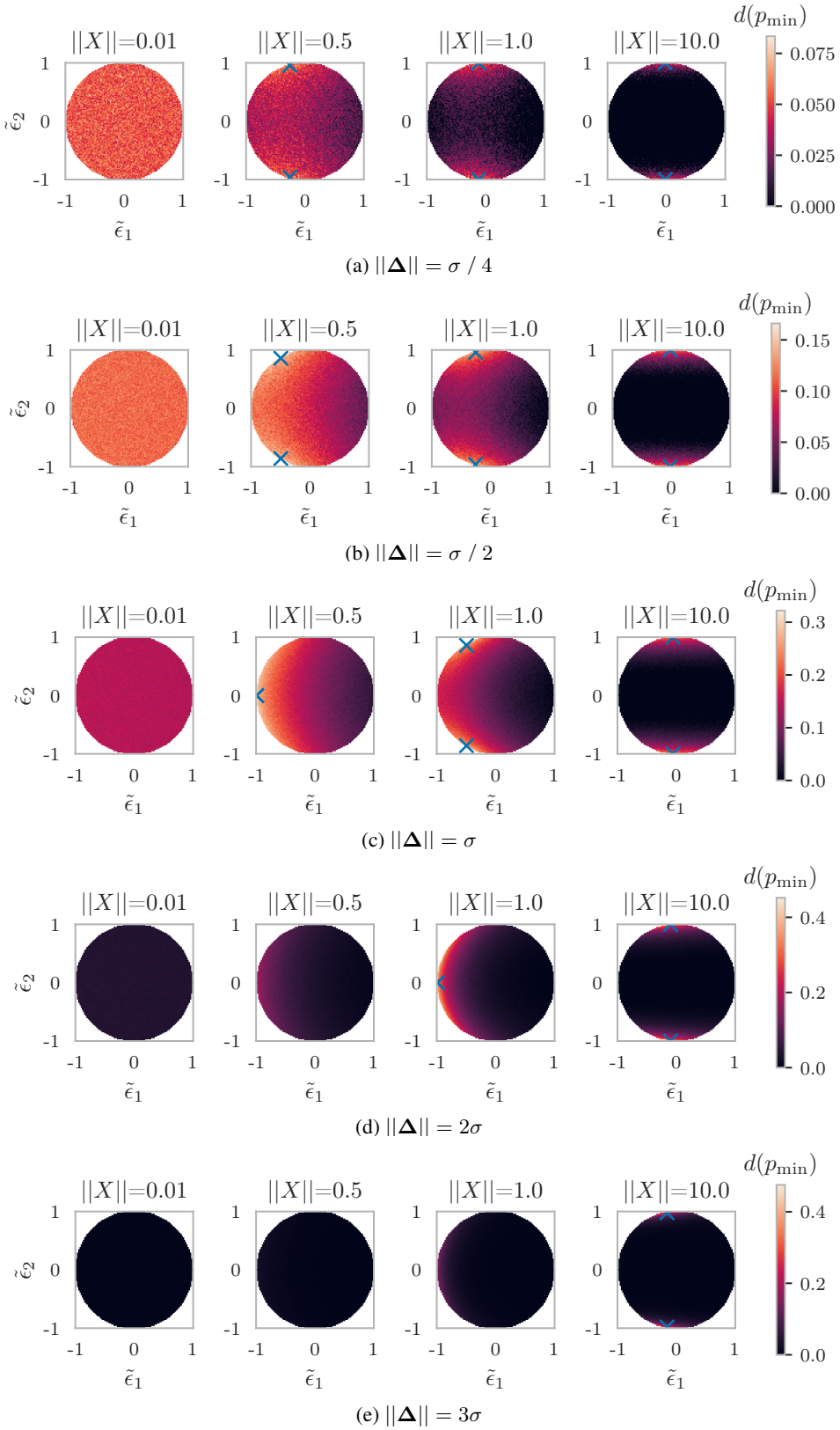
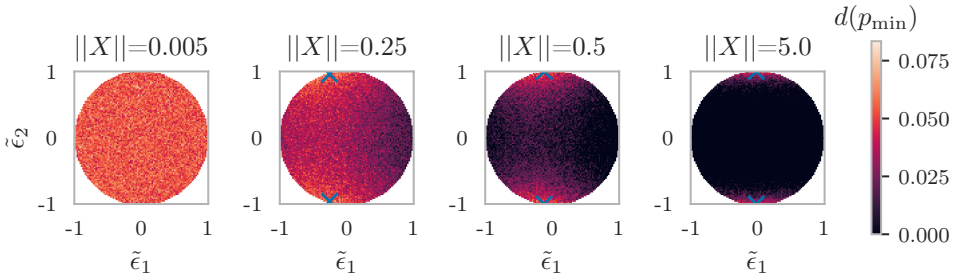
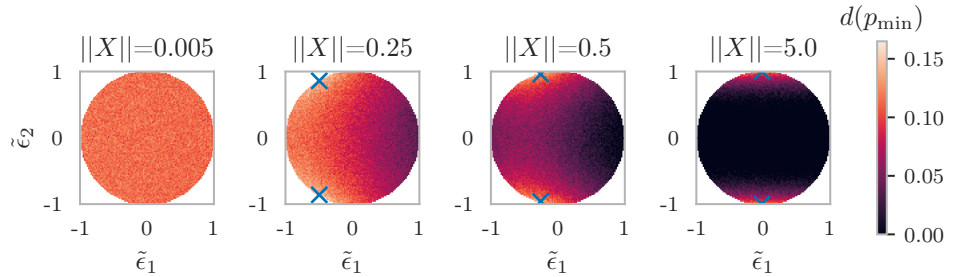


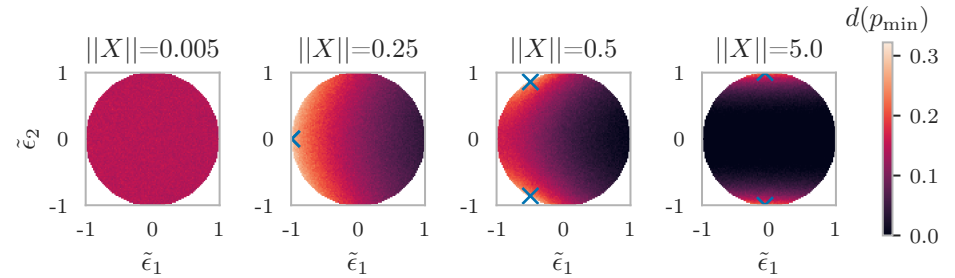
Figure 9: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for $\sigma = 1$ under varying $||\boldsymbol{X}||$, $||\mathbf{\Delta}||$, $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$. Blue crosses indicate adversarial rotations.



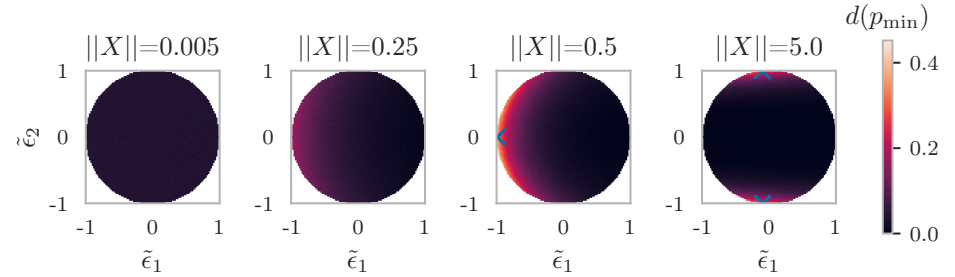
(a) $||\mathbf{\Delta}|| = \sigma \,/\, 4$



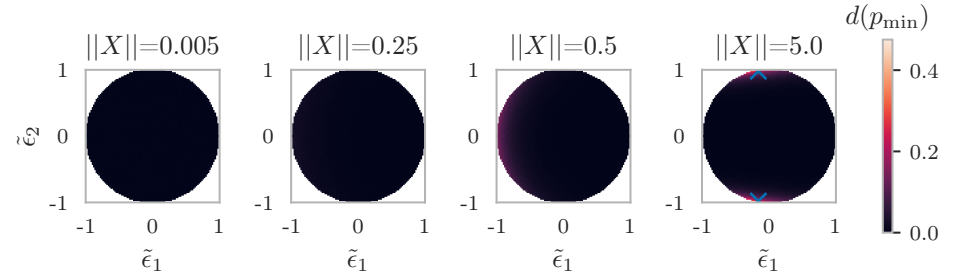
(b) $||\mathbf{\Delta}|| = \sigma \,/\, 2$



(c) $||\mathbf{\Delta}|| = \sigma$

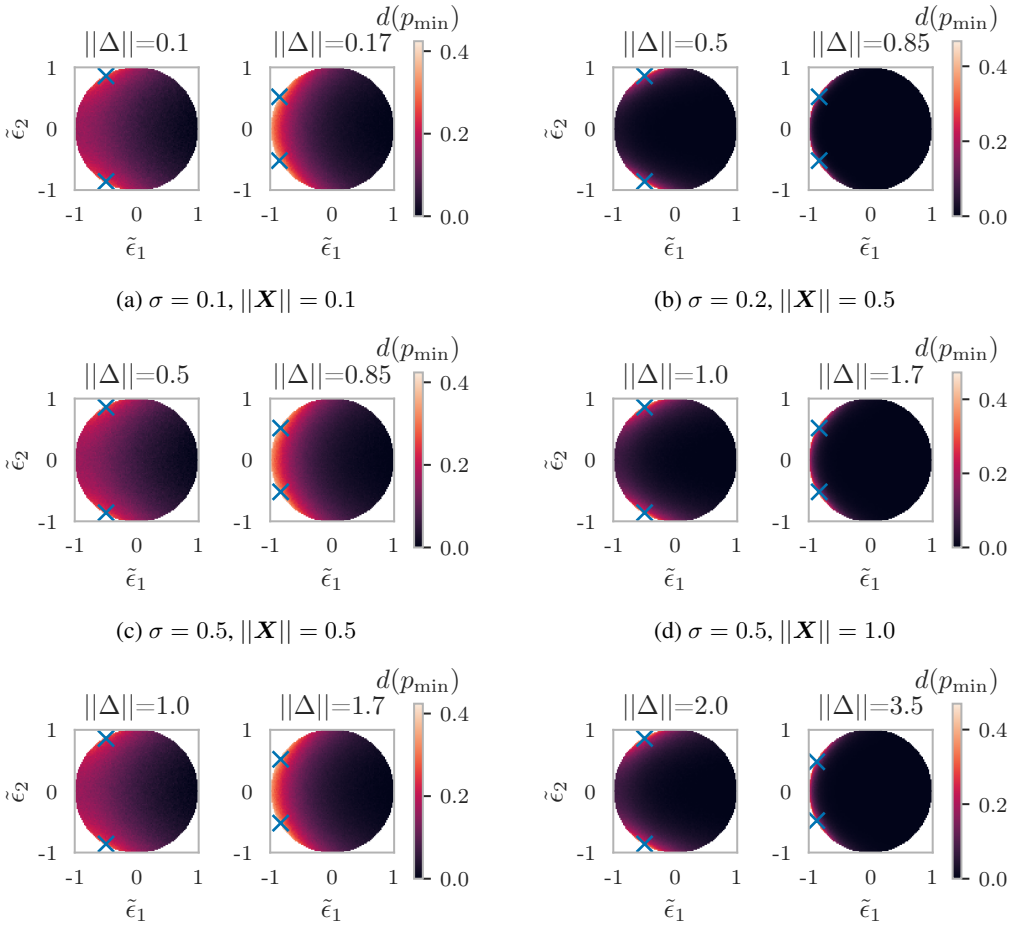


(d) $||\mathbf{\Delta}|| = 2\sigma$



(e) $||\mathbf{\Delta}|| = 3\sigma$

Figure 10: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for $\sigma = 1 \,/\, 2$ under varying $||\boldsymbol{X}||$, $||\mathbf{\Delta}||$, $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$. The results are identical to Fig. 9, because the certificate only depends on the value of the parameters relative to σ.



(a) $\sigma = 0.1$, $||\boldsymbol{X}|| = 0.1$

(b) $\sigma = 0.2$, $||\boldsymbol{X}|| = 0.5$

(c) $\sigma = 0.5$, $||\boldsymbol{X}|| = 0.5$

(d) $\sigma = 0.5$, $||\boldsymbol{X}|| = 1.0$

(e) $\sigma = 1.0$, $||\boldsymbol{X}|| = 1.0$

(f) $\sigma = 1.0$, $||\boldsymbol{X}|| = 2.0$

Figure 11: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for different combinations of σ, $||\boldsymbol{X}||_2$ and $||\boldsymbol{\Delta}||_2$. Parameter $||\boldsymbol{\Delta}||_2$ is chosen such that adversarial rotations (indicated with blue crosses) correspond to $\tilde{\epsilon} =\approx 0.5$ or $\tilde{\epsilon}_2 \approx \pm 0.5$. The regions with large $p_{\min}$ are concentrated around adversarial rotations.

A.2 Application to point cloud classification

Next, we repeat the experiments from Section 8.2 for diverse combinations of parameters and classifier architectures. We focus on the case $\sigma \leq 0.25$, where the randomly smoothed models achieve at least 50% accuracy (see Fig. 5), as even smaller accuracies are too low to be of any practical use.

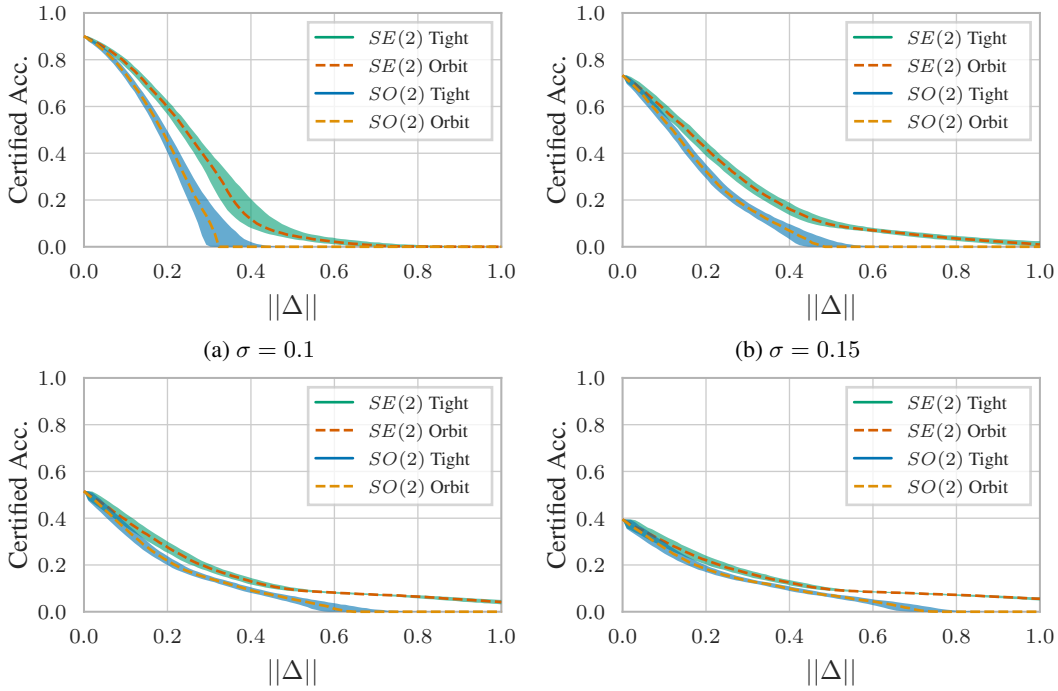
We first consider adversarial scaling on the point clouds constructed from MNIST (Appendix A.2.1). We then consider perturbations with random rotation components on MNIST (Appendix A.2.2) and ModelNet40 (Appendix A.2.3). For the tight certificates, we show both probabilistic upper and lower bounds obtained via Monte Carlo evaluation (recall Section 6.3). Note that only the lower bound is guaranteed to be a valid certificate with high probability.

A.2.1 Adversarial scaling on MNIST

Fig. 12 shows the certified accuracy (i.e. the number of correct and certifiably robust predictions) under adversarial scaling (i.e. $\tilde{\epsilon}_1 = 1, \tilde{\epsilon}_2 = 0$) for the randomly smoothed EnsPointNet with different σ on the MNIST test set. We evaluate the tight and orbit-based certificate for rotation invariance ($SO(2)$) and those for simultaneous rotation and translation invariance ($SE(2)$).

In all cases, the tight and orbit-based certificates yield similar certified accuracies. The certified accuracies of the tight certificate deviate slightly from the orbit-based ones, because we compute a probabilistic bound that holds with high probability $1 - \alpha$, rather than evaluating it exactly. The gap between the probalistic lower bound and the orbit-based certificate is particularly large for $\sigma = 0.1$, which can be explained by our observations from Appendix A.1: The smaller σ is, relative to $||\boldsymbol{X}||_2$ (which is defined by each datapoint of the test set and thus constant), the smaller the benefit of using it over the orbit-based baseline becomes.

Additionally enforcing translation invariance yields stronger robustness guarantees. For instance, with $\sigma = 0.2$, the $SE(2)$ certificates can still certify some predictions for $||\boldsymbol{\Delta}||_2 = 0.8$, whereas the $SO(2)$ certificates can not certify robustness beyond $||\boldsymbol{\Delta}||_2 \approx 0.62$.



(a) $\sigma = 0.1$ (b) $\sigma = 0.15$

(c) $\sigma = 0.2$ (d) $\sigma = 0.25$

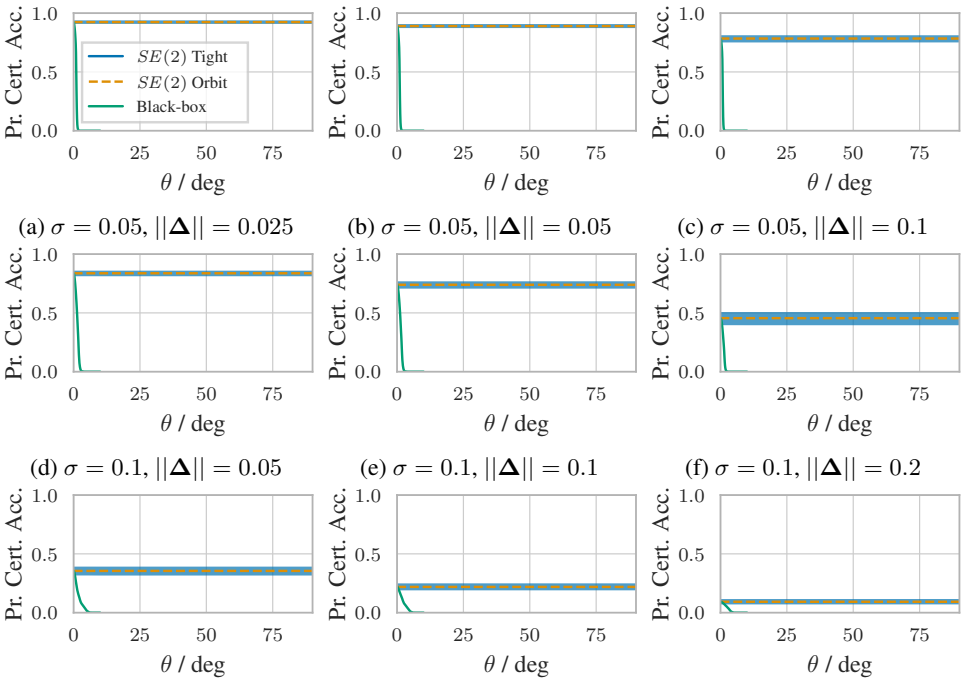
Figure 12: Comparison of certificates for adversarial scaling of MNIST with EnsPointNet and $\sigma \in \{0.1, 0.15, 0.2, 0.25\}$. Tight and orbit-based certificates yield similar certified accuracies.

A.2.2 Perturbations with random rotation components on MNIST

Like in Section 8.2, we construct perturbations with rotational components by fixing $||\mathbf{\Delta}||$, randomly sampling perturbations of the specified norm and then rotating $\boldsymbol{Z} = X + \mathbf{\Delta}$ by a specified angle θ. We obtain the original, unrotated perturbations via sampling from an isotropic matrix normal distribution and then scaling the sample to be of the desired norm. Per value of θ, we sample ten such perturbed inputs per element of the test set (i.e. 100000 per θ). We then compare the tight and the orbit-based certificate for simultaneous rotation and translation invariance (SE), as well as the black-box baseline. As our metric, we use probabilistic certified accuracy, i.e. the percentage of samples $\boldsymbol{X}'$ for which $f(\boldsymbol{X})$ is correct and $f(\boldsymbol{X}') = f(\boldsymbol{X})$ is provably guaranteed.

Fig. 13 shows the results for $\sigma \in \{0.05, 0.1, 0.2\}$, $||\mathbf{\Delta}||_2 \in \{\sigma \,/\, 2, \sigma, 2\sigma\}$ and $\theta \in [0°, 90°]$. While the black-box baseline fails to certify robustness even for small rotation angles, both gray-box certificates effectively eliminate the rotations, i.e. are constant in θ. However, the tight gray-box certificate does not offer any noteable benefit beyond that. The probabilistic lower bound never yields better probabilistic certified accuracy than the orbit-based certificate.

These results are consistent with our observartions about the tight certificate's parameter space: All values of σ that retain high accuracy are small, relative to the average norm of the test set (10.67).



(a) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.025$ (b) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.05$ (c) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.1$

(d) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.05$ (e) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.1$ (f) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.2$

(g) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.1$ (h) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.2$ (i) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.4$

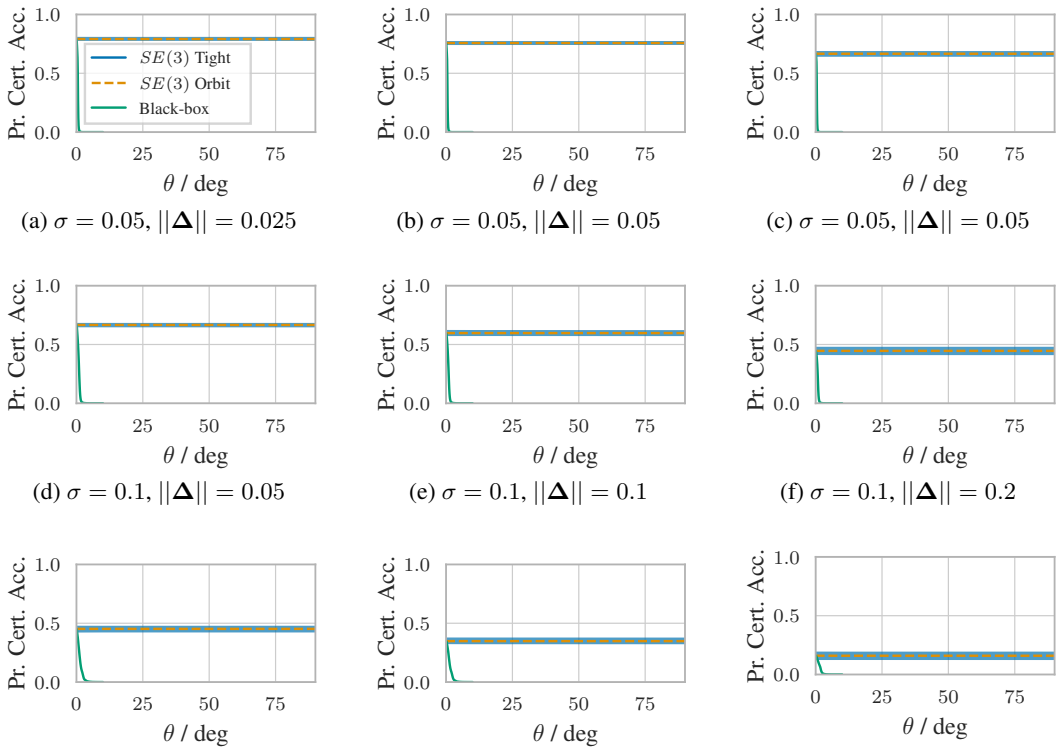
Figure 13: Comparison of tight gray-box, orbit-based gray-box and black-box certificates for MNIST with EnsPointNet and different σ, with respect to probabilistic certified accuracy. Perturbed inputs are generated by sampling perturbations with $||\mathbf{\Delta}||_2 \in \{\sigma \,/\, 2, \sigma, 2\sigma\}$ and rotating by angle θ. Group $SE(2)$ refers to simultaneous rotation and translation invariance. The orbit-based certificate is a good approximation of the tight certificate.

A.2.3 Perturbations with random elemental rotation components on ModelNet40

Next, we repeat the same experiment on ModelNet40. After sampling a perturbation $\boldsymbol{\Delta}$ of specified norm $||\boldsymbol{\Delta}||_2$, we randomly rotate $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$ by angle θ around a random axis. We obtain this axis by sampling from a three-dimensional isotropic normal distribution and normalizing the result.

We then compare the tight certificate for simultaneous invariance under rotation and translation to the orbit-based one and the black-box baseline.

Figs. 14 to 16 show the results for EnsPointNet, AttnPointNet and EnsDGCNN. In all cases, both the tight- and orbit-based certificate are constant in θ, but the tight certificate does not noticeably improve upon the orbit-based one w.r.t. probabilistic certified accuracy.

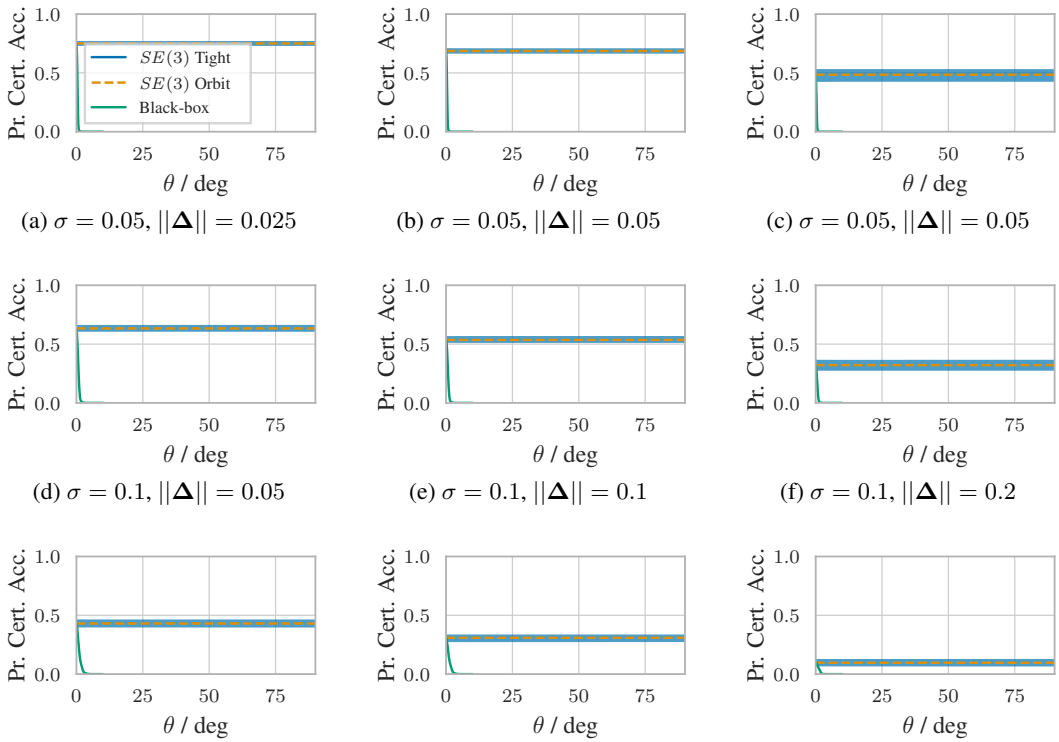


(a) $\sigma = 0.05$, $||\boldsymbol{\Delta}|| = 0.025$ (b) $\sigma = 0.05$, $||\boldsymbol{\Delta}|| = 0.05$ (c) $\sigma = 0.05$, $||\boldsymbol{\Delta}|| = 0.05$

(d) $\sigma = 0.1$, $||\boldsymbol{\Delta}|| = 0.05$ (e) $\sigma = 0.1$, $||\boldsymbol{\Delta}|| = 0.1$ (f) $\sigma = 0.1$, $||\boldsymbol{\Delta}|| = 0.2$

(g) $\sigma = 0.2$, $||\boldsymbol{\Delta}|| = 0.1$ (h) $\sigma = 0.2$, $||\boldsymbol{\Delta}|| = 0.2$ (i) $\sigma = 0.2$, $||\boldsymbol{\Delta}|| = 0.4$

Figure 14: Comparison of tight gray-box, orbit-based gray-box and black-box certificates for ModelNet40 with EnsPointNet and different σ, w.r.t. probabilistic certified accuracy. Perturbed inputs are generated by sampling perturbations with $||\boldsymbol{\Delta}||_2 \in \{\sigma \,/\, 2, \sigma, 2\sigma\}$ and rotating around a random axis by angle θ. Group $SE(3)$ refers to simultaneous rotation and translation invariance. The orbit-based certificate is a good approximation of the tight certificate.

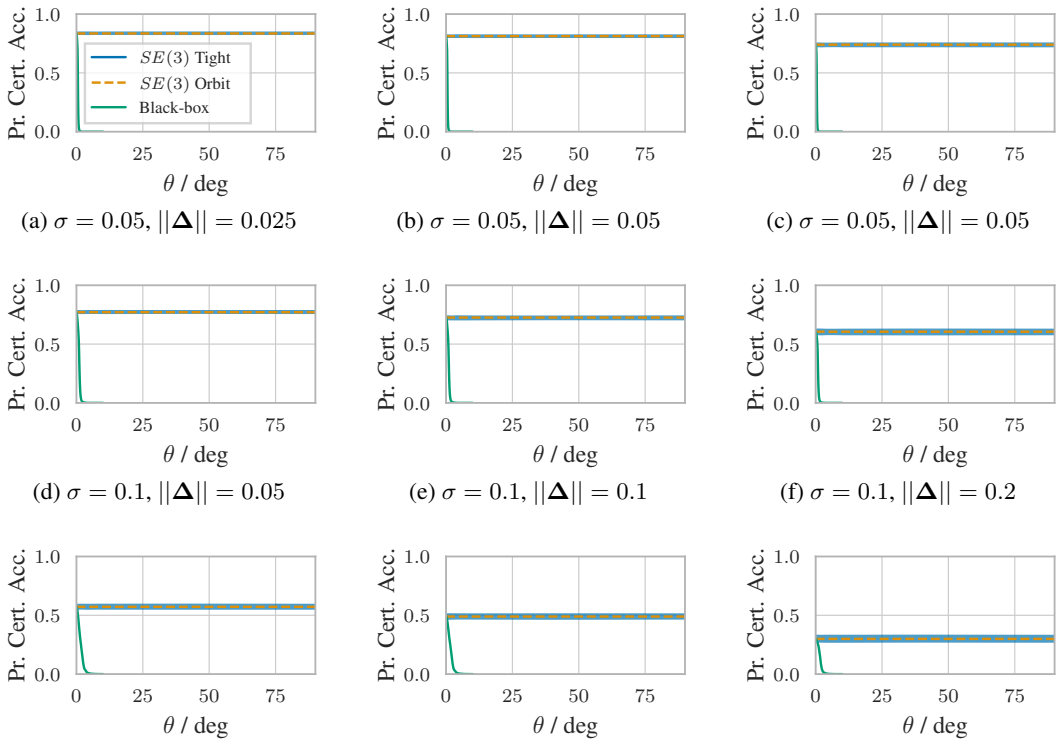


(a) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.025$ (b) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.05$ (c) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.05$

(d) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.05$ (e) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.1$ (f) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.2$

(g) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.1$ (h) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.2$ (i) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.4$

Figure 15: Comparison of tight gray-box, orbit-based gray-box and black-box certificates for ModelNet40 with AttnPointNet and different σ, w.r.t. probabilistic certified accuracy. Perturbed inputs are generated by sampling perturbations with $||\mathbf{\Delta}||_2 \in \{\sigma \,/\, 2, \sigma, 2\sigma\}$ and rotating around a random axis by angle θ. Group $SE(3)$ refers to simultaneous rotation and translation invariance. The orbit-based certificate is a good approximation of the tight certificate.



(a) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.025$ (b) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.05$ (c) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.05$

(d) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.05$ (e) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.1$ (f) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.2$

(g) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.1$ (h) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.2$ (i) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.4$

Figure 16: Comparison of tight gray-box, orbit-based gray-box and black-box certificates for ModelNet40 with EnsDGCNN and different σ, w.r.t. probabilistic certified accuracy. Perturbed inputs are generated by sampling perturbations with $||\mathbf{\Delta}||_2 \in \{\sigma \,/\, 2, \sigma, 2\sigma\}$ and rotating around a random axis by angle θ. Group $SE(3)$ refers to simultaneous rotation and translation invariance. The orbit-based certificate is a good approximation of the tight certificate.

B Full experimental setup

In the following, we provide a full specification of the used models, data sets and training parameters (Appendix B.1, randomized smoothing parameters (Appendix B.2), as well as the used computational resources (Appendix B.3) and third-party assets (Appendix B.4).

The code and all configuration files needed for reproducing the experimental results will be made available at https://www.cs.cit.tum.de/daml/invariance-smoothing.

B.1 Models, data and training parameters

B.1.1 Models

The experiments in Section 8.2 and Appendix A.2 are performed with three models: EnsPointNet, EnsDGCNN and AttnPointNet, which are rotation and translation invariant versions of PointNet [6], and the Dynamic Graph Convolutional Neural Network [128].

EnsPointNet is based on a standard PointNet architecture with an input T-Net but without a feature T-Net (we did not find the feature T-Net to improve the accuracy of the rotation invariant model). The input T-Net uses three convolutional layers with kernel size 1×1 (64, 128 and 1024 filters) and three linear layers (512, 256 and 9 neurons). The PointNet itself uses three convolutional layers with kernel size 1×1 (64, 128 and 1024 filters) and three linear layers (512, 256 and $|\mathbb{Y}|$ neurons). All layers, except the last one, use BatchNorm ($\epsilon = 1e-05$, $\text{momentum} = 0.1$). The second linear layer uses dropout ($p = 0.4$). To achieve rotation and translation invariance, the input data is centered and the two (for $2D$ data) or three (for $3D$ data) principal components are computed. Principal components are not unique. One has to account for order and sign ambiguities to ensure rotation invariance (see discussion in [129, 130]). If two or more eigenvalues are numerically close (relative tolerance $1e^{-5}$, absolute tolerance $1e^{-8}$), we iterate over all possible eigenvector signs and orders ($8 \cdot 6$), multiply the input data with the principal components, and pass it through the PointNet. If the eigenvalues are distinct, we sort them in ascending order and iterate over all possible signs (8). The 8 or 48 logit vectors are then averaged to obtain a prediction.

EnsDGCNN is based on a standard DGCNN architecture with an input spatial transform. The spatial transform uses uses three convolutional layers with kernel size 1×1 (64, 128 and 1024 filters) and three linear layers (512, 256 and 9 neurons). The DGCNN encodes the spatially transformed input with three DGC layers (64, 64, 64 and 128 filters and $k = 20$). The encoder output and residuals are concatenated and passed through a convolution with kernel size 1×1 (1024 filters), followed by max-pooling and three linear layers (512, 256, $|\mathbb{Y}|$ neurons). All layers use BatchNorm ($\epsilon = 1e-05$, $\text{momentum} = 0.1$). The first two linear layers use dropout ($p = 0.5$). We use the same ensembling approach as for EnsPointNet to achieve rotation and translation invariance.

AttnPointNet combines PointNet with the attention-based mechanism for combining canonical poses proposed in [129]. It uses the same encoder as EnsPointNet. After passing the different PCA-based canonical poses through the encoder, the hidden vectors are combined via a self-attention layer (1024 neurons each for query, key and value transform) and then passed through the same decoder as in EnsPointNet. To reduce memory usage, we only consider sign combinations that correspond to proper rotation matrices (see discussion in [130]).

B.1.2 Data

ModelNet40 [126] consists of 12311 CAD models from 40 categories, split into 9843 training samples and 2468 test samples. We subdivide the original training set into 80% train data and 20% validation data. The same split is used for all experiments. Each CAD model is then transformed into a 3D point cloud with 1024 points using the same pre-processing steps as in [6], i.e. randomly sampling from the mesh faces according to surface area and normalizing the resulting point cloud into the unit sphere.

MNIST [127] consists of 70000 handwritten digits, split into 60000 training samples and 10000 test samples. We subdivide the original training set into 80% train data and 20% validation data. The same split is used for all experiments. Each image is transformed into a 2D point with 1024 points using the same pre-processing steps as in [6], i.e. mapping all pixels with values greater than 128 to x-y coordinates, then uniformly sub-sampling or padding with a randomly chosen point and finally

normalizing into the unit sphere. We additionally combine classes 6 and 9, because these digits cannot be expected to be differentiated by a rotation-invariant model.

B.1.3 Training parameters

All models are trained on samples from the smoothing distribution (i.e. the model that is evaluated and certified with σ is trained on data augmented with Gaussian noise sampled from an isotropic normal distribution with standard deviation σ). Per element of a training batch, we take exactly one sample from the smoothing distribution. We do not use consistency regularization.

EnsPointNet is trained with cross entropy loss and the TNet regularization from [6] (weight 0.0001). Training is performed for 200 epochs with batch size 128 using Adam ($\beta_1 = 0.9, \beta_2 = 0.99$, $\epsilon = 1e-8$, weight_decay $= 1e^{-}4$). The learning rate starts at 0.001 and is multiplied with 0.7 every 20 epochs. The input data is randomly scaled by a factor from $[0.8, 1.25]$ and then transformed via principal component analysis, using a single set of eigenvectors with randomized sign and order.

EnsDGCNN is trained in the same manner as EnsPointNet, but with batch size 32, 400 epochs and learning rate decay every 40 epochs.

AttnPointNet is trained with batch size 64, 400 epochs and learning rate decay every 40 epochs. The other parameter values are identical to those for EnsPointNet. In order to train the self-attention layer, we do not transform the input with a single set of eigenvectors, but pass all possible eigenvectors corresponding to rotations through the network.

B.2 Randomized smoothing parameters

We use 1000 samples from the smoothing distribution to compute smoothed predictions. Abstentions, i.e. predictions for which we cannot guarantee that $p_{\boldsymbol{X},y^*} \geq 0.5$, are considered as incorrect.

For certification, we set the significance parameter α to 0.01. i.e. each certificate holds with probability 99.9%. We use 10000 samples per confidence bound, i.e. 10000 samples to bound $p_{\boldsymbol{X},y^*}$. For the tight gray-box certificates involving rotation invariance we additionally use 10000 samples to bound the threshold κ and 10000 samples to bound the optimal value itself (for more details, see Appendix F.5). We then use Holm-Bonferroni correction ([131]) to ensure that all three confidence bounds hold simultaneously with probability $1 - \alpha$.

Evaluating the tight certificates for $SO(3)$ and $SE(3)$ requires numerical integration over the two-dimensional parameter space $[-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, 2\pi]$ (see Appendix F.4.3). We use Clenshaw-Curtis quadrature with degree 20 in each dimension.

B.3 Computational resources

Training and Monte Carlo sampling of smoothed predictions was performed on a single NVIDIA A100 GPU ($40\,\mathrm{GB}$ VRAM) with an AMD EPYC 7543 CPU @ $2.8\,\mathrm{GHz}$. For EnsPointNet and AttnPointNet, $16\,\mathrm{GB}$ RAM were allocated. For EnsDGCNN, which requires additional memory for computation of k-nearest-neighbor graphs, $32\,\mathrm{GB}$ RAM were allocated. The average time for training EnsPointNet on MNIST was $42.8\,\mathrm{min}$. The average time for training EnsPointNet, AttnPoint and EnsDGCNN on ModelNet40 was $20.2\,\mathrm{min}$, $114\,\mathrm{min}$ and $158\,\mathrm{min}$, respectively. The average time for obtaining 11000 samples from EnsPointNet on MNIST was $1.1\,\mathrm{s}$. The average time for obtaining 11000 samples from EnsPointNet, AttnPoint and EnsDGCNN on ModelNet40 was $1.59\,\mathrm{s}$, $1.4\,\mathrm{s}$ and $24.91\,\mathrm{s}$, respectively.

Computation of certificates was performed on an Intel Xeon E5-2630 v4 CPU @ $2.2\,\mathrm{GHz}$, with $16\,\mathrm{GB}$ RAM allocated to each experiment. The average time for computing a tight certificate for 2D rotation invarance (i.e. computing threshold κ and bounding the optimal value, using 10000 samples each), was $0.05\,\mathrm{s}$. The average time for computing a tight certificate for 3D rotation invariance was $6.3\,\mathrm{s}$. The increase in computational cost is caused by the fact that we can only evaluate the worst-case classifier via numerical integration. The average time for computing the black-box randomized smoothing certificate and the tight certificate for translation invariance was in the sub-milisecond range and can thus not be accurately reported.

B.4 Third-party assets

Our work uses the publicly available MNIST [127] and ModelNet40 [126] datasets, as well as the quadpy quadrature library [132]. Our implementation additionally uses code from a publicly available implementation of PointNet and PointNet++ [133], as well as the author's reference implementation of DGCNN [128]. Both are available under MIT license.

C Group definitions

In the following, we define the different groups we consider in our work, each corresponding to a specific type of spatial invariance. Recall from Section 3.2 that a group is a set $\mathbb{T}$, associated with a binary operator $\cdot : \mathbb{T} \times \mathbb{T} \to \mathbb{T}$ that is closed and associative under the operation, has an inverse t^{-1} for each element $t \in \mathbb{T}$ and features an identity element e. Further recall that we associate each group with a group action $\circ : \mathbb{T} \times \mathbb{R}^{N \times D} \to \mathbb{R}^{N \times D}$ that describes how elements of the group modify elements of the input space.

C.1 Translation group $T(D)$

The translation group in D dimensions is the set of all D-dimensional continuous vectors, i.e. $T(D) = \mathbb{R}^D$. Two group elements $\boldsymbol{t}, \boldsymbol{t}' \in T(D)$ are combined via vector addition, i.e. $\boldsymbol{t} \cdot \boldsymbol{t}' = \boldsymbol{t} + \boldsymbol{t}'$. A group element $\boldsymbol{t} \in T(D)$ acts on an input $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ via row-wise addition, i.e. $\boldsymbol{t} \circ \boldsymbol{X} = \boldsymbol{X} + \mathbf{1}_N t^T$ with all-ones vector $\mathbf{1}_N \in \mathbb{R}^N$. The identity element e is the all-zeros vector $\mathbf{0}_D$.

C.2 Rotation group $SO(D)$

The rotation group in D dimensions, $SO(D)$ (short for *special orthogonal*), is the set of all D-dimensional rotation matrices, i.e. $SO(D) = \left\{ \boldsymbol{R} \in \mathbb{R}^{D \times D} \mid \boldsymbol{R}^T \boldsymbol{R} = \mathbf{I}_N \wedge \det(\boldsymbol{R}) = 1 \right\}$. Two group elements $\boldsymbol{R}, \boldsymbol{R}' \in SO(D)$ are combined via matrix multiplication, i.e. $\boldsymbol{R} \cdot \boldsymbol{R}' = \boldsymbol{R}\boldsymbol{R}'$. A group element $\boldsymbol{R} \in SO(D)$ acts on an input $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ via row-wise matrix-vector multiplication, i.e. $\boldsymbol{R} \circ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{R}^T$. The identity element e is the identity matrix $\mathbf{I}_D$.

C.3 Orthogonal group $O(D)$

The orthogonal group in D dimensions, $O(D)$, is the set of all D-dimensional orthogonal matrices, i.e. $O(D) = \left\{ \boldsymbol{A} \in \mathbb{R}^{D \times D} \mid \boldsymbol{R}^T \boldsymbol{R} = \mathbf{I}_N \right\}$. Note that $O(D) \supset SO(D)$. Like the rotation group, group elements are combined via matrix multiplication, i.e. $\boldsymbol{A} \cdot \boldsymbol{A}' = \boldsymbol{A}\boldsymbol{A}'$ and act on an input $\boldsymbol{X} \in \mathbb{A}^{N \times D}$ via row-wise matrix-vector multiplication, i.e. $\boldsymbol{A} \circ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{A}^T$. The identity element e is the identity matrix $\mathbf{I}_D$.

C.4 Roto-translation group $SE(D)$

The roto-translation group in D dimensions, $SE(D)$ (short for *special Euclidean*), is composed of all pairs of D-dimensional rotation matrices and translation vectors. That is, $SE(D) = SO(D) \times T(D)$. Two group elements $(\boldsymbol{R}, \boldsymbol{t}), (\boldsymbol{R}', \boldsymbol{t}') \in SE(D)$ are combined via a semidirect product, i.e. $(\boldsymbol{R}, \boldsymbol{t}) \cdot (\boldsymbol{R}', \boldsymbol{t}') = (\boldsymbol{R}\boldsymbol{R}', \boldsymbol{R}\boldsymbol{t}' + \boldsymbol{t})$. A group element $(\boldsymbol{R}, \boldsymbol{t}) \in SE(D)$ acts on an input $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ via row-wise matrix-vector multiplication, followed by a row-wise addition i.e. $(\boldsymbol{R}, \boldsymbol{t}) \circ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{R}^T + \mathbf{1}_N \boldsymbol{t}^T$. The identity element e is $(\mathbf{I}_D, \mathbf{0}_D)$.

C.5 Euclidean group $E(D)$

The Euclidean group in D dimensions, $E(D)$, corresponds to the set of all distance-preserving functions in Euclidean space. It is composed of all pairs of D-dimensional orthonormal matrices and translation vectors. That is, $E(D) = O(D) \times T(D)$. Note that $E(D) \supset SE(D)$. Like the roto-translation group, group elements combined via a semidirect product, i.e. $(\boldsymbol{A}, \boldsymbol{t}) \cdot (\boldsymbol{A}', \boldsymbol{t}') = (\boldsymbol{A}\boldsymbol{A}', \boldsymbol{A}\boldsymbol{t}' + \boldsymbol{t})$ and act on an input $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ via row-wise matrix-vector multiplication, followed by a row-wise addition i.e. $(\boldsymbol{A}, \boldsymbol{t}) \circ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{A}^T + \mathbf{1}_N \boldsymbol{t}^T$. The identity element e is $(\mathbf{I}_D, \mathbf{0}_D)$.

C.6 Permutation group $S(N)$

The permutation group in N dimensions, $S(N)$, is the set of all permutation matrices, i.e. $S(N) = \left\{ \boldsymbol{P} \in \{0,1\}^{N \times N} \mid \boldsymbol{P}^T \boldsymbol{P} = \mathbf{I}_N \right\}$. Two group elements $\boldsymbol{P}, \boldsymbol{P}' \in S(N)$ are combined via matrix multiplication, i.e. $\boldsymbol{P} \cdot \boldsymbol{P}' = \boldsymbol{P}\boldsymbol{P}'$. A group element $\boldsymbol{P} \in S(N)$ acts on an input $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ via matrix multiplication, i.e. $\boldsymbol{P} \circ \boldsymbol{X} = \boldsymbol{P}\boldsymbol{X}$. Note that the group action permutes the rows and not the columns. The identity element e is the identity matrix $\mathbf{I}_N$.

D Proof of Theorem 1

Theorem 1. *Let base classifier $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under group $\mathbb{T}$ with $\mathbb{T} \subseteq E(D)$ or $\mathbb{T} \subseteq S(N)$, where $E(D)$ is the Euclidean group and $S(N)$ is the permutation group. Then the isotropically smoothed classifier f, as defined in Section 3.1, is also invariant under $\mathbb{T}$.*

Recall that

$$f(\boldsymbol{X}) = \text{argmax}_{y\in\mathbb{Y}} p_{\boldsymbol{X},y}$$
$$\text{with } p_{\boldsymbol{X},y} = \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[g(\boldsymbol{Z}) = y\right]$$
$$\text{and } \mu_{\boldsymbol{X}}(\boldsymbol{Z}) = \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{Z}_{:,d} \mid \boldsymbol{X}_{:,d}, \sigma^2\mathbf{I}_N\right).$$

We shall prove Theorem 1 by showing that $\forall t \in \mathbb{T} : p_{\boldsymbol{X},y} = p_{(t\circ\boldsymbol{X}),y}$, i.e. the prediction probabilities are invariant. To do so, we need the following simple lemma:

Lemma 2. *Consider an arbitrary invertible matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, vectors $\boldsymbol{z}, \boldsymbol{m} \in \mathbb{R}^N$ and scalar $\sigma > 0$. Then*

$$\left|\det\left(\boldsymbol{A}^{-1}\right)\right| \mathcal{N}\left(\boldsymbol{A}^{-1}\boldsymbol{z} \mid \boldsymbol{m}, \sigma^2\mathbf{I}_N\right) = \mathcal{N}\left(\boldsymbol{z} \mid \boldsymbol{A}\boldsymbol{m}, \sigma^2\boldsymbol{A}\boldsymbol{A}^T\right).$$

Proof. It follows from the change of variables formula for densitiy functions that the l.h.s. term is the density function of random variable $\boldsymbol{z}' = \boldsymbol{A}\boldsymbol{z}$ with $\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{m}, \sigma^2\mathbf{1}_N\right)$. Due to the behavior of multivariate normal distributions under affine transformation, the r.h.s. term is also the density function of $\boldsymbol{z}'$. □

We begin by proving Theorem 1 for $\mathbb{T} \subseteq E(D)$, before considering $\mathbb{T} \subseteq S(N)$. In the following, let $\hat{g}(\boldsymbol{Z}) = \mathbb{1}\left[g(\boldsymbol{Z}) = y^*\right]$ indicate whether base classifier g classifies input $\boldsymbol{Z}$ as y^*.

Case 1: Assume that $\mathbb{T} \subseteq E(D)$ and consider an arbitrary $t \in \mathbb{T}$. By definition of the Euclidean group and the associated group action (see Appendix C.5) we must have $t = (\boldsymbol{A}, \boldsymbol{b})$ and $t \circ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{A}^T + \mathbf{1}_N\boldsymbol{b}^T$ for some orthogonal matrix $\boldsymbol{A} \in \mathbb{R}^{D\times D}$ and translation vector $\boldsymbol{b} \in \mathbb{R}^N$. By definition of the smoothing distribution, we have

$$\begin{aligned}
p_{(t\circ\boldsymbol{X}),y} &= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{Z})\mu_{(\boldsymbol{X}\boldsymbol{A}^T+\mathbf{1}_N\boldsymbol{b}^T)}(\boldsymbol{Z})\, d\boldsymbol{Z} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{Z}) \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Z}_n \mid \boldsymbol{A}\boldsymbol{X}_n + \boldsymbol{b}, \sigma^2\mathbf{I}_D\right)\, d\boldsymbol{Z} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{Z}) \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Z}_n - \boldsymbol{b} \mid \boldsymbol{A}\boldsymbol{X}_n, \sigma^2\mathbf{I}_D\right)\, d\boldsymbol{Z},
\end{aligned}$$

where in the last equality we have used that $\mathcal{N}(z \mid m + b) = \mathcal{N}(z - b \mid m)$. We can now perform two substitutions, $\boldsymbol{V} = \boldsymbol{Z} - \mathbf{1}_N\boldsymbol{b}^T$ and $\boldsymbol{W} = \boldsymbol{V}(\boldsymbol{A}^{-1})^T$, to transform this term into an expectation w.r.t. the original smoothing distribution $\mu_{\boldsymbol{X}}$:

$$\begin{aligned}
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V} + \mathbf{1}_N\boldsymbol{b}^T) \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{V}_n \mid \boldsymbol{A}\boldsymbol{X}_n, \sigma^2\mathbf{I}_D\right)\, d\boldsymbol{V} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V} + \mathbf{1}_N\boldsymbol{b}^T) \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{V}_n \mid \boldsymbol{A}\boldsymbol{X}_n, \sigma^2\boldsymbol{A}\boldsymbol{A}^T\right)\, d\boldsymbol{V} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V} + \mathbf{1}_N\boldsymbol{b}^T) \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{A}^{-1}\boldsymbol{V}_n \mid \boldsymbol{X}_n, \sigma^2\mathbf{I}_N\right)\, d\boldsymbol{V} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{W}\boldsymbol{A}^T + \mathbf{1}_N\boldsymbol{b}^T) \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{W}_n \mid \boldsymbol{X}_n, \sigma^2\mathbf{I}_N\right)\, d\boldsymbol{W}.
\end{aligned}$$

In the second step we have used the fact that $\boldsymbol{A}$ is orthogonal, i.e.$\boldsymbol{A}\boldsymbol{A}^T = \mathbf{I}_N$. In the third step, we have applied Lemma 2. Note that, because $|\det(\boldsymbol{A})| = 1$, we did not have to change the volume element. Finally, we can use the fact that g is invariant w.r.t. group $\mathbb{T}$ to prove that

$$
\begin{aligned}
p_{(t\circ\boldsymbol{X}),y} &= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{A}\boldsymbol{W} + \mathbf{1}_N \boldsymbol{b}^T) \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{W}_n \mid \boldsymbol{X}_n, \sigma^2 \mathbf{I}_N\right)\, d\boldsymbol{W} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{W}) \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{W}_n \mid \boldsymbol{X}_n, \sigma^2 \mathbf{I}_N\right)\, d\boldsymbol{W} \\
&= p_{\boldsymbol{X},y}.
\end{aligned}
$$

Case 2: Assume that $\mathbb{T} \subseteq S(N)$ and consider an arbitrary $t \in \mathbb{T}$. By definition of the permutation group and the associated group action (see Appendix C.6) we must have $t = \boldsymbol{P}$ and $t \circ \boldsymbol{X} = \boldsymbol{P}\boldsymbol{X}$ for some orthogonal matrix $\boldsymbol{P} \in \{0,1\}^{N\times N}$. The proof is virtually identical to that for the Euclidean group, except we now use the substitution $\boldsymbol{V} = \boldsymbol{P}\boldsymbol{Z}$:

$$
\begin{aligned}
p_{(t\circ\boldsymbol{X}),y} &= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{Z}) \mu_{(\boldsymbol{P}\boldsymbol{X})}(\boldsymbol{Z})\, d\boldsymbol{Z} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{Z}) \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{Z}_{:,d} \mid \boldsymbol{P}\boldsymbol{X}_{:,d}, \sigma^2 \mathbf{I}_N\right)\, d\boldsymbol{Z} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{Z}) \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{Z}_{:,d} \mid \boldsymbol{P}\boldsymbol{X}_{:,d}, \sigma^2 \boldsymbol{P}\boldsymbol{P}^T\right)\, d\boldsymbol{Z} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{Z}) \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{P}^{-1}\boldsymbol{Z}_{:,d} \mid \boldsymbol{X}_{:,d}, \sigma^2 \mathbf{I}_N\right)\, d\boldsymbol{Z} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{P}\boldsymbol{V}) \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{V}_{:,d} \mid \boldsymbol{X}_{:,d}, \sigma^2 \mathbf{I}_N\right)\, d\boldsymbol{V} \\
&= \int_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V}) \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{V}_{:,d} \mid \boldsymbol{X}_{:,d}, \sigma^2 \mathbf{I}_N\right)\, d\boldsymbol{V} \\
&= p_{\boldsymbol{X},y}.
\end{aligned}
$$

E Orbit-based gray-box certificates

In the following, we first prove the soundness of our orbit-based approach to gray-box robustness certification. We then present more explicit characterizations of the orbit-based certificates, which let us determine whether a specific perturbed input is part of the augmented certified region.

E.1 Proof of Theorem 2

Theorem 2. *Let $f \in \mathbb{R}^{N \times D} \to \mathbb{Y}$ be invariant under a group $\mathbb{T}$. Let $y^* = f(\boldsymbol{X})$ be a prediction that is certifiably robust to a set of perturbed inputs $\mathbb{B} \subseteq \mathbb{R}^{N \times D}$, i.e. $\forall \boldsymbol{Z} \in \mathbb{B} : f(\boldsymbol{Z}) = y^*$. Let $\tilde{\mathbb{B}} = \cup_{\boldsymbol{Z} \in \mathbb{B}} [\boldsymbol{Z}]_{\mathbb{T}}$. Then $\forall \boldsymbol{X}' \in \tilde{\mathbb{B}} : f(\boldsymbol{X}') = y^*$.*

Proof. Consider an arbitrary $\boldsymbol{X}' \in \tilde{\mathbb{B}}$. Due to the definition of $\tilde{\mathbb{B}}$, there must be a $\boldsymbol{Z} \in \mathbb{B}$ and a $t \in \mathbb{T}$ such that $\boldsymbol{X}' = t \circ \boldsymbol{Z}$. Since $t \in \mathbb{T}$, we know that $f(t \circ \boldsymbol{Z}) = f(\boldsymbol{Z})$. Since $\boldsymbol{Z} \in \mathbb{B}$, we know that $f(\boldsymbol{Z}) = y^*$. By the transitive property, we have $f(\boldsymbol{X}') = y^*$. □

E.2 Explicit characterizations

Recall that, for randomly smoothed models, a prediction $y^* = f(\boldsymbol{X})$ is certifiably robust within a Frobenius norm ball $\mathbb{B} = \{\boldsymbol{Z} \mid ||\boldsymbol{Z} - \boldsymbol{X}||_2 < r\}$ with $r = \sigma \Phi^{-1}(p_{\boldsymbol{X}, y^*})$. For this certificate, we can determine whether a specific perturbed input $\boldsymbol{X}'$ is in augmented certified region $\tilde{\mathbb{B}} = \cup_{\boldsymbol{Z} \in \mathbb{B}} [\boldsymbol{Z}]_{\mathbb{T}}$ by finding a transformation that minimizes the Frobenius distance between $\boldsymbol{X}'$ and $\boldsymbol{X}$:

Corollary 1. *Let $f \in \mathbb{R}^{N \times D} \to \mathbb{Y}$ be invariant under group $\mathbb{T}$. Let $y^* = f(\boldsymbol{X})$ 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. $\forall \boldsymbol{Z} \in \mathbb{B} : f(\boldsymbol{Z}) = y^*$. If $\min_{t \in \mathbb{T}} ||(t \circ \boldsymbol{X}') - \boldsymbol{X}||_2 < r$, then $f(\boldsymbol{X}') = y^*$.*

Proof. Let $t^* = \operatorname{argmin}_{t \in \mathbb{T}} ||(t \circ \boldsymbol{X}') - \boldsymbol{X}||_2$, define $\boldsymbol{Z} = t^* \circ \boldsymbol{X}'$ and assume that $||\boldsymbol{Z} - \boldsymbol{X}||_2 < r$. By definition of $\mathbb{B}$, we have $\boldsymbol{Z} \in \mathbb{B}$. Because $\mathbb{T}$ is a group, there must be an inverse element $(t^*)^{-1} \in \mathbb{T}$ with $(t^*)^{-1} \circ \boldsymbol{Z} = \boldsymbol{X}'$. Thus, by definition of orbits (see Definition 1), we have $\boldsymbol{X}' \in [\boldsymbol{Z}]_{\mathbb{T}}$. It follows from Theorem 2 that $f(\boldsymbol{X}') = y^*$. □

In the following, we discuss how to solve the optimization problem $\min_{\tau \in \mathbb{T}} ||\tau(\boldsymbol{X}') - \boldsymbol{X}||_2 < r$ for invariance under different groups $\mathbb{T}$. Before proceeding, remember that solving this optimization problem is not necessary for certifying robustness, i.e. specifying a set of inputs $\tilde{\mathbb{B}}$ such that $\forall \boldsymbol{X}' \in \tilde{\mathbb{B}} : f(\boldsymbol{X}') = y^*$ (see Theorem 2). It is only a way of performing membership inference, i.e. determining whether $\boldsymbol{X}' \in \tilde{\mathbb{B}}$ for a specific $\boldsymbol{X}' \in \mathbb{R}^{N \times D}$ – just like computing the Frobenius distance between $\boldsymbol{X}'$ and $\boldsymbol{X}$ can be used to determine whether $\boldsymbol{X}'$ is part of the original certified region $\mathbb{B}$.

E.2.1 Translation invariance

By definition of the translation group $T(D)$ and the associated group action (see Appendix C.1), we have $t \circ \boldsymbol{X}' = \boldsymbol{X}' + \boldsymbol{1}_N \boldsymbol{b}^T$ for some translation vector $\boldsymbol{b} \in \mathbb{R}^D$. Thus,

$$\min_{\tau \in \mathbb{T}} ||(\tau \circ \boldsymbol{X}') - \boldsymbol{X}||_2 = \min_{\boldsymbol{b} \in \mathbb{R}^D} ||(\boldsymbol{X}' + \boldsymbol{1}_N \boldsymbol{b}^T) - \boldsymbol{X}||_2 = ||\boldsymbol{\Delta} - \boldsymbol{1}_N \overline{\boldsymbol{\Delta}}||_2,$$

where $\boldsymbol{\Delta} = \boldsymbol{X}' - \boldsymbol{X}$ and $\overline{\boldsymbol{\Delta}} \in \mathbb{R}^{1 \times D}$ are the column-wise averages. The second equality can be shown by computing the gradients w.r.t. $\boldsymbol{b}$ and setting them to zero.

E.2.2 Rotation invariance

By definition of the rotation group $SO(D)$ and the associated group action (see Appendix C.2), we have $t \circ \boldsymbol{X}' = \boldsymbol{X}' \boldsymbol{R}^T$ for some rotation matrix $\boldsymbol{R}$. Thus,

$$\min_{\tau \in \mathbb{T}} ||(\tau \circ \boldsymbol{X}') - \boldsymbol{X}||_2 = \min_{\boldsymbol{R} \in SO(D)} ||\boldsymbol{X}' \boldsymbol{R}^T - \boldsymbol{X}||_2.$$

This is a special case of the orthogonal Procrustes problem, which can be solved via singular value decomposition [111]. The optimal rotation matrix $\boldsymbol{R}^*$ is given by $\boldsymbol{R}^* = \boldsymbol{V} \hat{S} \boldsymbol{U}^T$, where $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V} = (\boldsymbol{X}')^T \boldsymbol{X}$ and $\hat{S} = \operatorname{diag}\left(1, \ldots, 1, \operatorname{sign}\left(\det\left(\boldsymbol{V}\boldsymbol{U}^T\right)\right)\right)$.

E.2.3 Simultaneous rotation and reflection invariance

By definition of the orthogonal group $O(D)$ and the associated group action (see Appendix C.3), we have $t \circ \boldsymbol{X}' = \boldsymbol{X}'\boldsymbol{A}^T$ for some orthogonal matrix $\boldsymbol{A}$. Thus,

$$\min_{\tau \in \mathbb{T}} ||(\tau \circ \boldsymbol{A}') - \boldsymbol{X}||_2 = \min_{\boldsymbol{A} \in O(D)} ||\boldsymbol{X}'\boldsymbol{A}^T - \boldsymbol{X}||_2.$$

This is the orthogonal Procrustes problem, which can again be solved via singular value decomposition [111]. The optimal orthogonal matrix $\boldsymbol{A}^*$ is given by $\boldsymbol{A}^* = \boldsymbol{V}\boldsymbol{U}^T$, where $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V} = (\boldsymbol{X}')^T\boldsymbol{X}$. Here, accounting for the sign of the determinant is not necessary.

E.2.4 Simultaneous rotation and translation invariance

By definition of the special Euclidean group $SE(D)$ and the associated group action (see Appendix C.4), we have $t \circ \boldsymbol{X}' = \boldsymbol{X}'\boldsymbol{R}^T + \boldsymbol{1}_N\boldsymbol{b}^T$ for some rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{b} \in \mathbb{R}^D$. Thus,

$$\min_{\tau \in \mathbb{T}} ||(\tau \circ \boldsymbol{X}') - \boldsymbol{X}||_2 = \min_{\boldsymbol{R} \in O(D), \boldsymbol{c} \in \mathbb{R}^D} ||\boldsymbol{X}'\boldsymbol{R}^T + \boldsymbol{1}_N\boldsymbol{c}^T - \boldsymbol{X}||_2.$$

It can be shown that this is equivalent to centering $\boldsymbol{X}'$ and $\boldsymbol{X}$ and then solving the orthogonal Procrustes problem [112], i.e.

$$\min_{\tau \in \mathbb{T}} ||\tau\left(\boldsymbol{X}'\right) - \boldsymbol{X}||_2 = \min_{\boldsymbol{R} \in SO(D)} ||\left(\boldsymbol{X}' - \boldsymbol{1}_N\overline{\boldsymbol{X}'}\right)\boldsymbol{R}^T - \left(\boldsymbol{X} - \boldsymbol{1}_N\overline{\boldsymbol{X}}\right)||_2.$$

As discussed in Appendix E.2.2, this problem can be solved via singular value decomposition.

E.2.5 Permutation invariance

By definition of the permutation group $S(N)$ and the associated group action (see Appendix C.6), we have $t \circ \boldsymbol{X}' = \boldsymbol{P}\boldsymbol{X}'$ for some permutation matrix $\boldsymbol{P}$. Thus,

$$\min_{\tau \in \mathbb{T}} ||(\tau \circ \boldsymbol{X}') - \boldsymbol{X}||_2 = \min_{\boldsymbol{P} \in S(N)} ||\boldsymbol{P}\boldsymbol{X}' - \boldsymbol{X}||_2 = \sqrt{\min_{\boldsymbol{P} \in S(N)} ||\boldsymbol{P}\boldsymbol{X}' - \boldsymbol{X}||_2^2}.$$

The inner optimization problem is equivalent to finding an optimal matching in a bipartite graph with $2 \cdot N$ nodes whose cost matrix is given by $C_{n,m} = ||\boldsymbol{X}'_n - \boldsymbol{X}_m||_2^2$. This problem can be solved in polynomial time, for example via the Hungarian algorithm [134].

E.2.6 Simultaneous permutation, rotation and translation invariance

By definition of the permutation group $S(N)$ and the special Euclidean group $SE(N)$, we have

$$\min_{\tau \in \mathbb{T}} ||(\tau \circ \boldsymbol{X}') - \boldsymbol{X}||_2 = \min_{\boldsymbol{P} \in S(N)\boldsymbol{R} \in O(D), \boldsymbol{c} \in \mathbb{R}^D} ||\boldsymbol{P}\left(\boldsymbol{X}'\boldsymbol{R}^T + \boldsymbol{1}_N\boldsymbol{c}^T\right) - \boldsymbol{X}||_2.$$

Different from the previously discussed problems, the above is a challenging optimization problem known as point cloud registration, which does not have an efficiently computable solution. Approximate solutions to this problem are being actively studied (for a comprehensive survey, see [135]). Note that, if some upper bound $\hat{m}$ with $\min_{\tau \in \mathbb{T}} ||(\tau \circ \boldsymbol{X}') - \boldsymbol{X}||_2 < \hat{m}$ fulfills $\hat{m} < r$, then $\min_{\tau \in \mathbb{T}} ||(\tau \circ \boldsymbol{X}') - \boldsymbol{X}||_2 < r$. Thus, any approximate solution to the point cloud registration problem can be used for certification. If $\hat{m} < r$, we provably know that $\boldsymbol{X}' \in \tilde{\mathbb{B}}$ and thus $f(\boldsymbol{X}') = y^*$. However, there may be some $\boldsymbol{X}' \in \tilde{\mathbb{B}}$ with $\hat{m} > r$, which would be incorrectly declared as potential adversarial examples. Thus, an explicit characterization based on approximate solutions to the point cloud registration problem is sound, but not optimal.

F Tight gray-box certificates

In the following, we first prove Lemma 1, which lets us reduce the invariance-constrained optimization problem from Section 6 to a problem that is only constrained by the classifier's clean prediction probability. Then, we restate the Neyman-Pearson lemma from statistical testing, which provides an exact solution for the worst-case classifier, given its clean prediction probability. After that, we use the two lemmeta in proving Theorem 3, which summarizes tight certificates for translation, rotation and roto-translation invariance. Next, we discuss the certificates for each type of invariance in more detail and derive the results presented in Sections 6.2 to 6.5 from Theorem 3. We conclude by showing how to use Monte Carlo sampling to obtain narrow probabilistic bounds on the certificates involving rotation invariance (Appendix F.5).

F.1 Proof of Lemma 1

Recall from Section 6 that, in order to obtain tight gray-box certificates, we need to find the worst-case invariant classifier. That is, we need to solve $\min_{h \in \mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*]$, where $\mathbb{H}_\mathbb{T}$ is the set of all classifiers that are at least as likely as base classifier g to predict class $p_{\boldsymbol{X}, y^*}$ and have the same invariances. We want to prove that the invariance constraint can be eliminated via a canonical map $\gamma(\boldsymbol{Z})$, which maps each input $\boldsymbol{Z} \in \mathbb{R}^{N \times D}$ to a distinct representative of its orbit $[\boldsymbol{Z}]_\mathbb{T}$.

Lemma 1. *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 map for invariance under $\mathbb{T}$, then*

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

For the proof, recall

Definition 1. *The orbit of an input $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ w.r.t. a group $\mathbb{T}$ is $[\boldsymbol{X}]_\mathbb{T} = \{t \circ \boldsymbol{X} \mid t \in \mathbb{T}\}$.*

and

Definition 3. *A canonical map 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*

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

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

and

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

We begin our proof by deriving two lemmeta.

The first lemma states that, if an input $\boldsymbol{Z}$ is not the representative of its own orbit, then it cannot be the representative of any orbit.

Lemma 3. *Let γ be a canonical map for invariance under $\mathbb{T}$. Then, for all $\boldsymbol{Z} \in \mathbb{R}^{N \times D}$*

$$\boldsymbol{Z} \neq \gamma(\boldsymbol{Z}) \implies \nexists \boldsymbol{Z}' \in \mathbb{R}^{N \times D} : \gamma(\boldsymbol{Z}') = \boldsymbol{Z}.$$

Proof. Proof by contraposition. Assume there was some $\boldsymbol{Z}'$ with $\gamma(\boldsymbol{Z}') = \boldsymbol{Z}$. Eq. (3) from Definition 1 would imply that $\boldsymbol{Z} \in [\boldsymbol{Z}']_\mathbb{T}$. Eq. (4) from Definition 1 would then imply that $\gamma(\boldsymbol{Z}') = \gamma(\boldsymbol{Z})$. By the transitive property, we would have $\boldsymbol{Z} = \gamma(\boldsymbol{Z}') = \gamma(\boldsymbol{Z})$. □

The second lemma allows us to replace all equality constraints $\forall \boldsymbol{Z}' \in [\boldsymbol{Z}]_\mathbb{T} : h(\boldsymbol{Z}) = h(\boldsymbol{Z}')$ involving input $\boldsymbol{Z}$ and element of its equivalence class with a single equality constraint $h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z}))$.

Lemma 4. *Let γ be a canonical map for invariance under $\mathbb{T}$. Then*

$$\forall \boldsymbol{Z}, \forall \boldsymbol{Z}' \in [\boldsymbol{Z}]_\mathbb{T} : h(\boldsymbol{Z}) = h(\boldsymbol{Z}') \iff \forall \boldsymbol{Z} : h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z})).$$

Proof.

$\Rightarrow$ Consider an arbitrary $\boldsymbol{Z} \in \mathbb{R}^{N \times D}$. Due to Eq. (3) from Definition 1, we know that $\gamma(\boldsymbol{Z}) \in [\boldsymbol{Z}]_{\mathbb{T}}$. Therefore, $\forall \boldsymbol{Z}' \in [\boldsymbol{Z}]_{\mathbb{T}} : h(\boldsymbol{Z}) = h(\boldsymbol{Z}')$ implies $h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z}))$.

$\Leftarrow$ Assume that $\forall \boldsymbol{Z} : h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z}))$. Consider an arbitrary pair of inputs $\boldsymbol{Z} \in \mathbb{R}^{N \times D}, \boldsymbol{Z}' \in [\boldsymbol{Z}]_{\mathbb{T}}$. Due to our assumption, we know that $h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z}))$. Due to Eq. (4) from Definition 3, we know that $\gamma(\boldsymbol{Z}) = \gamma(\boldsymbol{Z}')$ and thus $h(\gamma(\boldsymbol{Z})) = h(\gamma(\boldsymbol{Z}'))$. Again, due to our assumption, we know that $h(\gamma(\boldsymbol{Z}')) = h(\boldsymbol{Z}')$. By the transitive property, we have $h(\boldsymbol{Z}) = h(\boldsymbol{Z}')$. $\square$

We can now apply the two lemmata to prove Lemma 1. Lemma 4 allows us to restate $\mathbb{H}_{\mathbb{T}}$ as follows:

$$\mathbb{H}_{\mathbb{T}} = \left\{ h : \mathbb{R}^{N \times D} \to \mathbb{Y} \mid \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}} [h(\boldsymbol{Z}) = y] \geq p_{\boldsymbol{X}, y^*} \wedge \forall \boldsymbol{Z} : h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z})) \right\}.$$

Thus, the optimization problem $\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*]$ can be written as

$$\begin{aligned} &\min_{h : \mathbb{R}^{N \times D} \to \mathbb{Y}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] \\ &\quad \text{s.t.} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}} [h(\boldsymbol{Z}) = y^*] \geq p_{\boldsymbol{X}, y^*} \wedge \forall \boldsymbol{Z} : h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z})). \end{aligned}$$

Since we already know from the second constraint that any feasible solution must fulfill $h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z}))$, we may as well substitute $h(\gamma(\boldsymbol{Z}))$ for $h(\boldsymbol{Z})$ before solving the problem:

$$\begin{aligned} &\min_{h : \mathbb{R}^{N \times D} \to \mathbb{Y}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\gamma(\boldsymbol{Z})) = y^*] \\ &\quad \text{s.t.} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}} [h(\gamma(\boldsymbol{Z})) = y^*] \geq p_{\boldsymbol{X}, y^*} \wedge \forall \boldsymbol{Z} : h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z})). \end{aligned} \tag{13}$$

For each $\boldsymbol{Z} \in \mathbb{R}^{N \times D}$, we can now distinguish two cases: If $\boldsymbol{Z} = \gamma(\boldsymbol{Z})$, then the constraint $h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z}))$ is trivially fulfilled and can thus be dropped. If $\boldsymbol{Z} \neq \gamma(\boldsymbol{Z})$, then Lemma 3 shows that $h(\boldsymbol{Z})$ does not appear in the objective function or first constraint of Eq. (13), because there is no other $\boldsymbol{Z}'$ with $\gamma(\boldsymbol{Z}') = \boldsymbol{Z}$. We can thus ignore the second constraint, solve the optimization problem

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

and then let $h(\boldsymbol{Z}) \leftarrow h(\gamma(\boldsymbol{Z}))$ to obtain an optimal, feasible solution to Eq. (13).

F.2 Neyman-Pearson Lemma

For our purposes, the Neyman-Pearson lemma [107] can be formulated as follows:

Lemma 5 (Neyman-Pearson lower bound)**.** *Let $\mu_{\boldsymbol{X}'}$, $\mu_{\boldsymbol{X}}$, be two continuous distributions over a measurable set $\mathbb{A}$ such that, for all $\kappa \in \mathbb{R}_+$, the set $\left\{ z \mid \frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} = \kappa \right\}$ has measure zero. Consider an arbitrary label set $\mathbb{Y}$, a specific class label $y \in \mathbb{Y}$ and scalar $p \in [0, 1]$. Then*

$$\left(\min_{h : \mathbb{A} \to \mathbb{Y}} \Pr_{z \sim \mu_{\boldsymbol{X}'}} [h(z) = y] \text{ s.t. } \Pr_{z \sim \mu_{\boldsymbol{X}}} [h(z) = y] \geq p \right) = \mathop{\mathbf{E}}_{z \sim \mu_{\boldsymbol{X}'}} [h^*(z)]$$

$$\text{with } h^*(z) = \mathbb{1} \left[\frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} \leq \kappa \right] \text{ and } \kappa \in \mathbb{R}_+ \text{ such that } \mathop{\mathbf{E}}_{z \sim \mu_{\boldsymbol{X}}} [h^*(z)] = p.$$

Here, indicator function h^* corresponds to a classifier that predicts class y if and only if the likelihood ratio is below a specific threshold κ. For an application of this variant of the Neyman-Pearson lemma to black-box robustness certification, as well as a discussion of its relation to most powerful hypothesis tests, see [26]. For various other formulations of the lemma, see [136].

One can use the same approach to obtain an upper bound on the probability of predicting a specific class. This will be relevant for our discussion of multi-class certificates in Appendix H.

Lemma 6 (Neyman-Pearson upper bound)**.** *Let $\mu_{\boldsymbol{X}'}$, $\mu_{\boldsymbol{X}}$, be two continuous distributions over a measurable set $\mathbb{A}$ such that, for all $\kappa \in \mathbb{R}_+$, the set $\left\{ z \mid \frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} = \kappa \right\}$ has measure zero. Consider an arbitrary label set $\mathbb{Y}$, a specific class label $y \in \mathbb{Y}$ and scalar $p \in [0, 1]$. Then*

$$\left(\max_{h : \mathbb{A} \to \mathbb{Y}} \Pr_{z \sim \mu_{\boldsymbol{X}'}} [h(z) = y] \text{ s.t. } \Pr_{z \sim \mu_{\boldsymbol{X}}} [h(z) = y] \leq p \right) = \mathop{\mathbf{E}}_{z \sim \mu_{\boldsymbol{X}'}} [h^*(z)]$$

$$\text{with } h^*(z) = \mathbb{1} \left[\frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} \geq \kappa \right] \text{ and } \kappa \in \mathbb{R}_+ \text{ such that } \mathop{\mathbf{E}}_{z \sim \mu_{\boldsymbol{X}}} [h^*(z)] = p.$$

F.3 Proof of Theorem 3

Theorem 3. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under $\mathbb{T}$ with $\mathbb{T}$ chosen from $\{T(D), SO(D), SE(D)\}$. For $SO(D)$ and $SE(D)$, let $D \in \{2,3\}$. Let $\mathbb{H}_\mathbb{T}$ be defined as in Eq. (2) and η be a right Haar measure on $\mathbb{T}$. Define the indicator function $h^* : \mathbb{R}^{N\times D} \to \{0,1\}$ with*

$$h^*(\boldsymbol{Z}) = \mathbb{1}\left[\frac{\beta_{\boldsymbol{X}'}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}(\boldsymbol{Z})} \leq \kappa\right], \text{ where } \beta_{\boldsymbol{X}}(\boldsymbol{Z}) = \int_{t\in\mathbb{T}} \exp\left(\langle t\circ \boldsymbol{Z}, \boldsymbol{X}\rangle_{\mathrm{F}} \,/\, \sigma^2\right)\, d\eta(t) \tag{5}$$

$$\text{and } \kappa \in \mathbb{R} \text{ such that } \mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[h^*(\boldsymbol{Z})] = p_{\boldsymbol{X},y^*}. \tag{6}$$

Then

$$\min_{h\in\mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}[h(\boldsymbol{Z}) = y^*] = \mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}[h^*(\boldsymbol{Z})]. \tag{7}$$

We begin our proof by applying Lemma 1, which lets us restate the l.h.s. optimization problem from Eq. (7) as

$$\min_{h:\mathbb{R}^{N\times D}\to\mathbb{Y}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}[h(\gamma(\boldsymbol{Z})) = y^*] \text{ s.t. } \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[h(\gamma(\boldsymbol{Z})) = y^*] \geq p_{\boldsymbol{X},y^*} \tag{14}$$

with canonical map $\gamma : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times D}$. Note that Eq. (14) has the same form as the optimization problem solved by the Neyman Pearson lemma (Lemma 5), save for the canonical representation in the probability terms.

Our goal is to 1.) specify a canonical map γ for invariance w.r.t. group $\mathbb{T}$ 2.) bring the two probability terms from Eq. (14) into a form that does not depend γ, so that we can solve Eq. (14) exactly via the Neyman-Pearson lemma. To this end, we make the following proposition, which we shall later verify for the different considered groups $\mathbb{T}$:

Proposition 1. *Let $\mathbb{T} \in \{T(D), SO(D), SE(D)\}$. For $SO(D)$ and $SE(D)$, let $D \in \{2,3\}$. Let $e \in \mathbb{T}$ be the group's identity element and $\circ : \mathbb{T}\times\mathbb{R}^{N\times D} \to \mathbb{R}^{N\times D}$ be the group action (see Appendix C). There exist a parameter space $\Omega \subseteq \mathbb{R}^K$, a function $\tau : \Omega \to \mathbb{T}$ with $\tau(\mathbf{0}) = e$, as well as a parameter space $\Psi \subseteq \mathbb{R}^{N\cdot D-K}$ and a function $\xi : \psi \to \mathbb{R}^{N\times D}$ with $K \in \{1,\ldots,N\cdot D-1\}$ such that*

$$\lambda(\boldsymbol{\omega},\boldsymbol{\psi}) = \tau(\boldsymbol{\omega})\circ\xi(\boldsymbol{\psi}). \tag{15}$$

is a differentiable, surjective function from $\Omega\times\Psi$ to $\mathbb{R}^{N\times D}$, injective almost everywhere and fulfills

$$\forall t\in\mathbb{T}, \boldsymbol{\omega}\in\Omega, \boldsymbol{\psi}\in\Psi, \exists\boldsymbol{\omega}'\in\Omega: \quad t\circ\lambda(\boldsymbol{\omega},\boldsymbol{\psi}) = \lambda(\boldsymbol{\omega}',\boldsymbol{\psi}). \tag{16}$$

That is, the input space $\mathbb{R}^{N\times D}$ can be parameterized by a matrix $\xi(\boldsymbol{\psi}) \in \mathbb{R}^{N\times D}$ that is transformed by a group element $\tau(\boldsymbol{\omega}) \in \mathbb{T}$. As formalized in Eq. (16), this alternative parameterization lets us neatly disentangle the effect of a group action, which is a translation and/or rotation of the coordinate system, from the (group-invariant) geometry of $\boldsymbol{Z}$, i.e. $\xi(\boldsymbol{\psi})$.

Canonical map. Based on Proposition 1, we can define the following canonical map $\gamma : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times D}$ for $\boldsymbol{Z} = \lambda(\boldsymbol{\omega},\boldsymbol{\psi})$:

$$\gamma(\boldsymbol{Z}) = (\tau(\boldsymbol{\omega}))^{-1}\circ\lambda(\boldsymbol{\omega},\boldsymbol{\psi}). \tag{17}$$

Note that $(\tau(\boldsymbol{\omega}))^{-1} \in \mathbb{T}$ is the inverse of group element $\tau(\boldsymbol{\omega}) \in \mathbb{T}$ and not the inverse of the function τ. Further note that by Eq. (15), it holds for arbitrary $\boldsymbol{Z} = \lambda(\boldsymbol{\omega},\boldsymbol{\psi})$ that

$$\gamma(\boldsymbol{Z}) = (\tau(\boldsymbol{\omega}))^{-1}\circ\tau(\boldsymbol{\omega})\circ\xi(\boldsymbol{\psi}) = e\circ\xi(\boldsymbol{\psi}) = \tau(\mathbf{0})\circ\xi(\boldsymbol{\psi}) = \lambda(\mathbf{0},\boldsymbol{\psi}). \tag{18}$$

We can verify that γ is indeed a valid canonical map by testing the two criteria from Definition 3. Because $\gamma(\boldsymbol{Z})$ applies a group action, we have $\gamma(\boldsymbol{Z}) \in [\boldsymbol{Z}]_\mathbb{T}$, where $[\boldsymbol{Z}]_\mathbb{T}$ is the orbit of $\boldsymbol{Z}$ w.r.t. $\mathbb{T}$. Due to Eq. (16), we have $\forall\boldsymbol{Z}, \forall\boldsymbol{Z}\in[\boldsymbol{Z}]_\mathbb{T} : \gamma(\boldsymbol{Z}) = \gamma(\lambda(\boldsymbol{\omega},\boldsymbol{z})) = \lambda(\mathbf{0},\boldsymbol{\psi}) = \gamma(\lambda(\boldsymbol{\omega}',\boldsymbol{\psi})) = \gamma(\boldsymbol{Z}')$.

Substitution. Next, we use Proposition 1 and the canonical map from Eq. (17) to bring the probabilities from our optimization problem in Eq. (14) into a form that is compatible with the Neyman-Pearson lemma. To declutter the terms, we first introduce $\hat{h}(\boldsymbol{Z})$ as a shorthand for $\mathbb{1}[h(\boldsymbol{Z}) = y^*]$:

$$\Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[h(\gamma(\boldsymbol{Z})) = y^*] = \mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[\hat{h}(\gamma(\boldsymbol{Z}))\right].$$

We then perform the substitution $\boldsymbol{Z} = \lambda(\boldsymbol{\omega}, \boldsymbol{\psi})$ with $\boldsymbol{\omega} \in \Omega \subseteq \mathbb{R}^K$ and $\boldsymbol{\psi} \in \Psi \subseteq \mathbb{R}^{N \cdot D - K}$:

$$\begin{aligned} \mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}} \left[\hat{h}(\gamma(\boldsymbol{Z})) \right] &= \int_{\mathbb{R}^{N \times D}} \hat{h}(\gamma(\boldsymbol{Z})) \cdot \mu_{\boldsymbol{X}}(\boldsymbol{Z}) \, d\boldsymbol{Z} \\ &= \int_{\Psi} \int_{\Omega} \hat{h}(\gamma(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}))) \cdot \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})) \cdot |\det(\boldsymbol{J}_{\lambda}(\boldsymbol{\omega}, \boldsymbol{\psi}))| \; d\boldsymbol{\omega} d\boldsymbol{\psi}. \end{aligned} \tag{19}$$

where $\boldsymbol{J}_{\lambda}$ is the Jacobian of λ after vectorizing its domain and codomain. Next, we apply the canonical map defined in Eq. (17) to eliminate the group parameters $\boldsymbol{\omega}$ from the term involving classifier $\hat{h}$. This allows us to marginalize out $\boldsymbol{\omega}$:

$$\begin{aligned} &\int_{\Psi} \int_{\Omega} \hat{h}(\gamma(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}))) \cdot \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})) \cdot |\det(\boldsymbol{J}_{\lambda}(\boldsymbol{\omega}, \boldsymbol{\psi}))| \; d\boldsymbol{\omega} d\boldsymbol{\psi} \\ &= \int_{\Psi} \int_{\Omega} \hat{h}(\lambda(\mathbf{0}, \boldsymbol{\psi})) \cdot \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})) \cdot |\det(\boldsymbol{J}_{\lambda}(\boldsymbol{\omega}, \boldsymbol{\psi}))| \; d\boldsymbol{\omega} d\boldsymbol{\psi} \\ &= \int_{\Psi} \hat{h}(\lambda(\mathbf{0}, \boldsymbol{\psi})) \int_{\Omega} \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})) \cdot |\det(\boldsymbol{J}_{\lambda}(\boldsymbol{\omega}, \boldsymbol{\psi}))| \; d\boldsymbol{\omega} d\boldsymbol{\psi} \\ &:= \mathop{\mathbf{E}}_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}}} \left[\hat{h}(\lambda(\mathbf{0}, \boldsymbol{\psi})) \right] \\ &= \Pr_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}}} \left[h(\lambda(\mathbf{0}, \boldsymbol{\psi})) = y^* \right], \end{aligned} \tag{20}$$

where $\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) = \int_{\Omega} \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})) \cdot |\det(\boldsymbol{J}_{\lambda}(\boldsymbol{\omega}, \boldsymbol{\psi}))| \; d\boldsymbol{\omega}$ is the marginal density of parameters $\boldsymbol{\psi}$. Finally, we can insert our result into Eq. (14) to obtain the simplified optimization problem

$$\min_{h: \mathbb{R}^{N \times D} \to \mathbb{Y}} \Pr_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}'}} \left[h(\lambda(\mathbf{0}, \boldsymbol{\psi})) = y^* \right] \text{ s.t. } \Pr_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}}} \left[h(\lambda(\mathbf{0}, \boldsymbol{\psi})) = y^* \right] \geq p_{\boldsymbol{X}, y^*}. \tag{21}$$

Applying the Neyman-Pearson lemma. Eq. (21) is almost in the form required by the Neyman-Pearson lemma. But, unlike in Lemma 5, we optimize over a function defined on $\mathbb{R}^{N \times D}$ while only evaluating it on a subset of $\mathbb{R}^{N \times D}$, namely $\lambda(\mathbf{0}, \Psi)$. To resolve this mismatch, it is convenient to treat h as a family of variables $(h_{\boldsymbol{Z}})$ indexed by $\mathbb{R}^{N \times D}$. As specified in Proposition 1, λ is injective almost everywhere, meaning that each tuple $(\mathbf{0}, \boldsymbol{\psi})$ indexes a distinct variable $h_{\lambda(\mathbf{0}, \boldsymbol{\psi})}$, save for sets of measure zero, which do not influence the objective and constraint in Eq. (21). Therefore, we may equivalently optimize over a family of variables $(\tilde{h}_{\boldsymbol{\psi}})$ indexed by Ψ:

$$\min_{\tilde{h}: \Psi \to \mathbb{Y}} \Pr_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}'}} \left[\tilde{h}(\boldsymbol{\psi}) = y^* \right] \text{ s.t. } \Pr_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}}} \left[\tilde{h}(\boldsymbol{\psi}) = y^* \right] \geq p_{\boldsymbol{X}, y^*}. \tag{22}$$

According to Lemma 5, the minimizer is a function that classifies $\boldsymbol{\psi}$ as y^* iff $\tilde{h}^*(\boldsymbol{\psi}) = 1$ with

$$\tilde{h}^*(\boldsymbol{\psi}) = \mathbb{1} \left[\frac{\nu_{\boldsymbol{X}'}(\boldsymbol{\psi})}{\nu_{\boldsymbol{X}}(\boldsymbol{\psi})} \leq \kappa \right] \text{ with } \kappa \in \mathbb{R}_+ \text{ such that } \mathop{\mathbf{E}}_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}}} \left[\tilde{h}^*(\boldsymbol{\psi}) \right] = p_{\boldsymbol{X}, y^*}.$$

Consequently, the optimum of our original problem Eq. (21) is given by any function $h^*(\boldsymbol{Z})$ with $h^*(\lambda(\mathbf{0}, \boldsymbol{\psi})) = \tilde{h}(\boldsymbol{\psi})$, i.e. we can use an arbitrary classifier for all parts of the domain that do not appear in the probability terms of Eq. (21). We make the following proposition for our choice of worst-case classifier $h^*(\boldsymbol{Z})$:

Proposition 2. *Consider marginal density $\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) = \int_{\Omega} \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})) \cdot |\det(\boldsymbol{J}_{\lambda}(\boldsymbol{\omega}, \boldsymbol{\psi}))| \; d\boldsymbol{\omega}$. Let η be a right Haar measure on group $\mathbb{T} \in \{T(D), SO(D), SE(D)\}$. For $SO(D)$ and $SE(D)$, let $D \in \{2, 3\}$. Define the group-averaged kernel*

$$\beta_{\boldsymbol{X}}(\boldsymbol{Z}) = \int_{t \in \mathbb{T}} \exp\left(\langle t \circ \boldsymbol{Z}, \boldsymbol{X} \rangle_{\mathrm{F}} \,/\, \sigma^2 \right) \; d\eta(t) \tag{23}$$

Then,

$$\frac{\nu_{\boldsymbol{X}'}(\boldsymbol{\psi})}{\nu_{\boldsymbol{X}}(\boldsymbol{\psi})} \propto \frac{\beta_{\boldsymbol{X}'}(\lambda(\mathbf{0}, \boldsymbol{\psi}))}{\beta_{\boldsymbol{X}}(\lambda(\mathbf{0}, \boldsymbol{\psi}))},$$

where $\propto$ absorbs factors that are constant in $\boldsymbol{\psi}$.

Assuming Proposition 2 holds, we have by transitivity of the previous equalities that

$$\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \mathop{\mathbf{E}}_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}'}} [h^*(\lambda(\mathbf{0}, \boldsymbol{\psi}))] \tag{24}$$

with

$$h^*(\boldsymbol{Z}) = \mathbb{1}\left[\frac{\beta_{\boldsymbol{X}'}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}(\boldsymbol{Z})} \leq \kappa\right] \text{ and } \kappa \in \mathbb{R}_+ \text{ such that } \mathop{\mathbf{E}}_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}}} [h^*(\lambda(\mathbf{0}, \boldsymbol{\psi}))] = p_{\boldsymbol{X}, y^*}. \tag{25}$$

Resubstitution. Next, we need to transform the expectations w.r.t. marginal distribution $\nu_{\boldsymbol{X}}$ over $\Psi \subseteq \mathbb{R}^{N \cdot D - K}$ back into an expectation w.r.t. our original smoothing distribution $\mu_{\boldsymbol{X}}$ over $\mathbb{R}^{N \times D}$. Applying the steps from Eq. (19) to Eq. (20) in reverse order shows that

$$\mathop{\mathbf{E}}_{\boldsymbol{\psi} \sim \nu_{\boldsymbol{X}}} [h^*(\lambda(\mathbf{0}, \boldsymbol{\psi}))] = \mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}} [h^*(\gamma(\boldsymbol{Z}))]. \tag{26}$$

Exploiting invariance. Finally, we use the fact that our worst-case classifier h^* is invariant under group $\mathbb{T}$ to eliminate the canonical map γ from the expectation. Recall from Eq. (17) that, for any $\boldsymbol{Z} = \lambda(\boldsymbol{\omega}, \boldsymbol{\psi})$, we defined $\gamma(\boldsymbol{Z}) = (\tau(\boldsymbol{\omega}))^{-1} \circ \boldsymbol{Z}$, i.e. the canonical map lets a group element act on $\boldsymbol{Z}$. Since η is a right Haar measure, we have

$$\begin{aligned}
\beta_{\boldsymbol{X}}((\tau(\boldsymbol{\omega}))^{-1} \circ \boldsymbol{Z}) &= \int_{t \in \mathbb{T}} \exp\left(\left\langle t \circ (\tau(\boldsymbol{\omega}))^{-1} \circ \boldsymbol{Z}, \boldsymbol{X}\right\rangle_{\mathrm{F}} / \sigma^2\right) \, d\eta(t) \\
&= \int_{t \in \mathbb{T}} \exp\left(\left\langle (t \cdot (\tau(\boldsymbol{\omega}))^{-1}) \circ \boldsymbol{Z}, \boldsymbol{X}\right\rangle_{\mathrm{F}} / \sigma^2\right) \, d\eta(t) \\
&= \int_{t \in \mathbb{T}} \exp\left(\left\langle u \circ \boldsymbol{Z}, \boldsymbol{X}\right\rangle_{\mathrm{F}} / \sigma^2\right) \, d\eta(u) \\
&= \beta_{\boldsymbol{X}}(\boldsymbol{Z}),
\end{aligned}$$

where the second equality follows from the fact that $\circ$ is a group action The third equality holds because η is a right Haar measure, meaning we can make the substitution $u = t \cdot (\tau(\boldsymbol{\omega}))^{-1})$ without having to change the measure. Thus, we have

$$h^*(\gamma(\boldsymbol{Z})) = \mathbb{1}\left[\frac{\beta_{\boldsymbol{X}'}(\gamma(\boldsymbol{Z}))}{\beta_{\boldsymbol{X}}(\gamma(\boldsymbol{Z}))} \leq \kappa\right] = \mathbb{1}\left[\frac{\beta_{\boldsymbol{X}'}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}(\boldsymbol{Z})} \leq \kappa\right] = h^*(\boldsymbol{Z}).$$

Combined with Eq. (26) and Eqs. (24) and (25), this proves that the optimal value of the original variance-constrained problem is

$$\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h^*(\boldsymbol{Z})] \tag{27}$$

with

$$h^*(\boldsymbol{Z}) = \mathbb{1}\left[\frac{\beta_{\boldsymbol{X}'}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}(\boldsymbol{Z})} \leq \kappa\right] \text{ and } \kappa \in \mathbb{R}_+ \text{ such that } \mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}} [h^*(\boldsymbol{Z})] = p_{\boldsymbol{X}, y^*} = p_{\boldsymbol{X}, y^*}.$$

The last thing we need to do in order to conclude our proof is verify that Propositions 1 and 2 hold for each of the considered groups $\mathbb{T}$.

F.3.1 Verifying Propositions 1 and 2 for translation invariance

For $\mathbb{T} = T(D)$, we define $\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}) = \tau(\boldsymbol{\omega}) \circ \xi(\boldsymbol{\psi})$ with $\tau : \mathbb{R}^D \to T(D)$, $\xi : \mathbb{R}^{(N-1)D} \to \mathbb{R}^{N \times D}$ and

$$\begin{aligned}
\tau(\boldsymbol{\omega}) &= \boldsymbol{\omega} \\
\xi(\boldsymbol{\psi}) &= \begin{bmatrix} \mathbf{0}_D^T \\ \mathrm{vec}^{-1}(\boldsymbol{\psi}) \end{bmatrix},
\end{aligned}$$

where $\mathrm{vec}^{-1} : \mathbb{R}^{(N-1)D} \to \mathbb{R}^{(N-1) \times D}$ reshapes an input vector into a matrix. Due to the definition of group action $\circ$ from Appendix C.1, we have

$$\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}) = \begin{bmatrix} \mathbf{0}_D^T \\ \mathrm{vec}^{-1}(\boldsymbol{\psi}) \end{bmatrix} + \mathbf{1}_N \boldsymbol{\omega}^T,$$

with all-ones vector $\mathbf{1}_N \in \mathbb{R}^N$. In other words: We represent each element of $\mathbb{R}^{N\times D}$ as a matrix whose first row is zero, translated by a vector $\boldsymbol{\omega}$.

The function λ is evidently a differentiable bijection with inverse

$$\lambda^{-1}(\boldsymbol{Z}) = \left(\boldsymbol{Z}_1, \operatorname{vec}\left(\boldsymbol{Z}_{2:} - \mathbf{1}_D(\boldsymbol{Z}_1)^T\right)\right),$$

meaning it is surjective and injective. Furthermore, we have $\tau(\mathbf{0}_D) = \mathbf{0}_D$, with $\mathbf{0}_D$ being the identity element of $T(D)$. Lastly, we have for all $\boldsymbol{t} \in T(D)$, $\boldsymbol{\omega} \in \mathbb{R}^D$, $\boldsymbol{\psi} \in \mathbb{R}^{(N-1)D}$ that

$$\boldsymbol{t} \circ \lambda(\boldsymbol{\omega}, \boldsymbol{\psi}) = \lambda(\boldsymbol{t} + \boldsymbol{\omega}, \boldsymbol{\psi}).$$

Thus, all conditions from Proposition 1 are fulfilled.

To verify Proposition 2, we need to show that

$$\frac{\nu_{\boldsymbol{X}'}(\boldsymbol{\psi})}{\nu_{\boldsymbol{X}}(\boldsymbol{\psi})} \propto \frac{\beta_{\boldsymbol{X}'}(\lambda(\mathbf{0}, \boldsymbol{\psi}))}{\beta_{\boldsymbol{X}}(\lambda(\mathbf{0}, \boldsymbol{\psi}))},$$

with marginal density $\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) = \int_{\mathbb{R}^D} \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})) \cdot |\det(\boldsymbol{J}_\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}))| \; d\boldsymbol{\omega}$ and Haar integral $\beta_{\boldsymbol{X}}(\boldsymbol{Z}) = \int_{t\in\mathbb{R}^D} \exp\left(\langle t \circ \boldsymbol{Z}, \boldsymbol{X}\rangle_{\mathrm{F}} / \sigma^2\right) \; d\eta(t)$, where η is an arbitrary right Haar measure on translation group $T(D) = \mathbb{R}^D$. Firstly, we see that $|\det(\boldsymbol{J}_\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}))| = 1$, since we are only performing translations. Thus,

$$\begin{aligned}
\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) &= \int_{\mathbb{R}^D} \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})) \; d\boldsymbol{\omega} \\
&\propto \int_{\mathbb{R}^D} \prod_{n=1}^{N} \exp\left(-\frac{1}{2\sigma^2}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})_n - \boldsymbol{X}_n)^T \lambda(\boldsymbol{\omega}, \boldsymbol{\psi})_n - \boldsymbol{X}_n\right) \; d\boldsymbol{\omega} \\
&\propto \int_{\mathbb{R}^D} \prod_{n=1}^{N} \exp\left(\frac{1}{\sigma^2}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})_n)^T \lambda(\boldsymbol{\omega}, \boldsymbol{\psi})_n\right) \; d\boldsymbol{\omega} \\
&= \int_{\mathbb{R}^D} \exp\left(\langle \lambda(\boldsymbol{\omega}, \boldsymbol{\psi}), \boldsymbol{X}\rangle_{\mathrm{F}} / \sigma^2\right) \; d\boldsymbol{\omega} \\
&= \int_{\mathbb{R}^D} \exp\left(\langle \boldsymbol{\omega} \circ \lambda(\mathbf{0}_D, \boldsymbol{\psi}), \boldsymbol{X}\rangle_{\mathrm{F}} / \sigma^2\right) \; d\boldsymbol{\omega},
\end{aligned}$$

where $\propto$ absorbs factors that are constant in $\boldsymbol{\omega}$. In the above equalities we have first inserted the definition of our isotropic matrix normal smoothing distribution, then removed constant terms, then expressed the product of exponential functions more compactly using the Frobenius inner product and finally used that, by definition, $\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}) = \tau(\boldsymbol{\omega}) \circ \xi(\boldsymbol{\psi}) = \boldsymbol{\omega} \circ \lambda(\mathbf{0}_D, \boldsymbol{\psi})$. Finally, we note that the Lebesgue measure is translation-invariant, i.e. $\int_{\mathbb{R}^D} h(\boldsymbol{\omega} + \boldsymbol{c}) \; d\boldsymbol{\omega} = \int_{\mathbb{R}^D} h(\boldsymbol{\omega}) \; d\boldsymbol{\omega}$ for arbitrary functions h, meaning it is a Haar measure of the translation group. Since the translation group is a Lie group and thus locally compact, the Haar measure is unique up to a multiplicative constant [109, 110]. Thus, Proposition 2 holds.

F.3.2 Verifying Propositions 1 and 2 for rotation invariance in 2D

In this section, let

$$\boldsymbol{R}(\omega) = \begin{bmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{bmatrix} \tag{28}$$

be the matrix that rotates counter-clockwise by angle ω. Note that $SO(2) = \{\boldsymbol{R}(\omega) \mid \omega \in [0, 2\pi]\}$. To verify our propositions for $\mathbb{T} = SO(2)$, we define $\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}) = \tau(\omega) \circ \xi(\boldsymbol{\psi})$ with $\tau : [0, 2\pi] \to SO(2)$ and $\xi : \mathbb{R}_+ \times \mathbb{R}^{2(N-1)} \to \mathbb{R}^{N\times 2}$ (note that the first argument is non-negative) with

$$\begin{aligned}
\tau(\omega) &= \boldsymbol{R}(\omega) \\
\xi(\boldsymbol{\psi}) &= \begin{bmatrix} \psi_1 & 0 \\ \operatorname{vec}^{-1}(\boldsymbol{\psi}_{2:}) \end{bmatrix},
\end{aligned}$$

where $\operatorname{vec}^{-1} : \mathbb{R}^{2(N-1)} \to \mathbb{R}^{(N-1)\times 2}$ reshapes an input vector into a matrix. Due to the definition of group action $\circ$ from Appendix C.2, we have

$$\lambda(\omega, \boldsymbol{\psi}) = \begin{bmatrix} \psi_1 & 0 \\ \operatorname{vec}^{-1}(\boldsymbol{\psi}_{2:}) \end{bmatrix} \boldsymbol{R}(\omega)^T$$

In other words: We represent each element of $\mathbb{R}^{N\times 2}$ as a matrix whose first row is aligned with the x-axis, rotated counter-clockwise by an angle ω.

The function λ is evidently surjective. Any $\boldsymbol{Z} \in \mathbb{R}^{N\times 2}$ can be represented via

$$\boldsymbol{Z} = \begin{bmatrix} ||\boldsymbol{Z}||_2 & 0 \\ \boldsymbol{Z}_{2:}\boldsymbol{R}(-\omega^*)^T \end{bmatrix} \boldsymbol{R}(\omega^*)^T$$

with $\omega^* = \operatorname{arctan2}(Z_{1,2}, Z_{1,1})$. It is also injective, save for the set $\{\boldsymbol{Z} \in \mathbb{R}^{N\times 2} \mid \boldsymbol{Z}_1 = \boldsymbol{0}\}$, which has measure zero. The first row $\boldsymbol{Z}_1$ uniquely defines $\psi_1 = ||\boldsymbol{Z}_1||_2$ and $\omega = \operatorname{arctan2}(Z_{1,2}, Z_{1,1})$, because polar coordinates for non-zero vectors are unique. The parameter vector $\boldsymbol{z}_{2:}$ must fulfill

$$\operatorname{vec}^{-1}(\boldsymbol{\psi}_{2:})\boldsymbol{R}(\omega) = \boldsymbol{Z}_{2:},$$

which has a unique solution because rotation matrix $\boldsymbol{R}(\omega)$ is invertible. In addition to surjectivity and injectivity almost everywhere, λ is differentiable and we have $\tau(0) = \mathbf{I}_N$, with identity matrix $\mathbf{I}_N$ being the identity element of $SO(2)$. Furthermore, we have for all $\boldsymbol{R}(\alpha)$, $\boldsymbol{\omega}$, $\boldsymbol{\psi}$ that

$$\boldsymbol{R}(\alpha) \circ \lambda(\omega, \boldsymbol{\psi}) = \lambda(\alpha + \omega, \boldsymbol{\psi}).$$

Thus, all conditions from Proposition 1 are fulfilled.

To verify Proposition 2, we need to show that

$$\frac{\nu_{\boldsymbol{X}'}(\boldsymbol{\psi})}{\nu_{\boldsymbol{X}}(\boldsymbol{\psi})} \propto \frac{\beta_{\boldsymbol{X}'}(\lambda(\boldsymbol{0}, \boldsymbol{\psi}))}{\beta_{\boldsymbol{X}}(\lambda(\boldsymbol{0}, \boldsymbol{\psi}))},$$

with marginal density $\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) = \int_{[0,2\pi]} \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})) \cdot |\det(\boldsymbol{J}_\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}))| \; d\boldsymbol{\omega}$ and Haar integral $\beta_{\boldsymbol{X}}(\boldsymbol{Z}) = \int_{\boldsymbol{R}\in SO(2)} \exp\left(\langle t \circ \boldsymbol{Z}, \boldsymbol{X}\rangle_{\mathrm{F}} \,/\, \sigma^2\right) \; d\eta(t)$. We begin by calculating the Jacobian

$$\begin{aligned}
\boldsymbol{J}_\lambda(\omega, \boldsymbol{\psi}) &= \begin{bmatrix} \frac{\partial \operatorname{vec}(\lambda)}{\partial \omega} & \frac{\partial \operatorname{vec}(\lambda)}{\partial \psi_1} & \frac{\partial \operatorname{vec}(\lambda)}{\partial \psi_2} & \cdots & \frac{\partial \operatorname{vec}(\lambda)}{\partial \psi_{2N-1}} \end{bmatrix} \\
&= \begin{bmatrix} -\psi_1 \sin(\omega) & \cos(\omega) & \boldsymbol{0} \\ \psi_1 \cos(\omega) & \sin(\omega) & \boldsymbol{0} \\ \boldsymbol{a} & \boldsymbol{0} & \boldsymbol{B} \end{bmatrix},
\end{aligned}$$

with some vector $\boldsymbol{a} \in \mathbb{R}^{2(N-1)}$ and block-diagonal matrix $\boldsymbol{B} \in \mathbb{R}^{2(N-1)\times 2(N-1)}$ with

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{R}(\omega) & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{R}(\omega) \end{bmatrix}.$$

Due to the block structure, we have

$$|\det(\boldsymbol{J}_\lambda(\omega, \boldsymbol{\psi}))| = \left|\det\left(\begin{bmatrix} -\psi_1 \sin(\omega) & \cos(\omega) \\ \psi_1 \cos(\omega) & \sin(\omega) \end{bmatrix}\right)\right| \prod_{n=2}^{N} |\det(\boldsymbol{R}(\omega))| = |\psi_1|.$$

Thus, our marginal density is

$$\begin{aligned}
\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) &= \int_{[0,2\pi]} |\psi_1| \cdot \mu_{\boldsymbol{X}}(\lambda(\omega, \boldsymbol{\psi})) \; d\boldsymbol{\omega} \\
&\propto \int_{[0,2\pi]} \prod_{n=1}^{N} \exp\left(-\frac{1}{2\sigma^2}(\lambda(\omega, \boldsymbol{\psi})_n - \boldsymbol{X}_n)^T \lambda(\omega, \boldsymbol{\psi})_n - \boldsymbol{X}_n\right) \; d\omega \\
&\propto \int_{[0,2\pi]} \prod_{n=1}^{N} \exp\left(\frac{1}{\sigma^2}(\lambda(\omega, \boldsymbol{\psi})_n)^T \lambda(\omega, \boldsymbol{\psi})_n\right) \; d\omega \\
&= \int_{[0,2\pi]} \exp\left(\langle \lambda(\omega, \boldsymbol{\psi}), \boldsymbol{X}\rangle_{\mathrm{F}} \,/\, \sigma^2\right) \; d\omega \\
&= \int_{[0,2\pi]} \exp\left(\langle \boldsymbol{R}(\omega) \circ \lambda(\boldsymbol{0}_D, \boldsymbol{\psi}), \boldsymbol{X}\rangle_{\mathrm{F}} \,/\, \sigma^2\right) \; d\omega,
\end{aligned}$$

where $\propto$ absorbs factors that are constant in ω. Because the composition of two rotations corresponds to a translation of rotation angles, the translation-invariant Lebesgue measure is an invariant measure for group $SO(2)$ (in this angle-based parameterization): $\int_{[0,2\pi]} h(\boldsymbol{R}(\omega) \cdot \boldsymbol{R}(\omega')) \; d\omega = \int_{[0,2\pi]} h(\boldsymbol{R}(\omega + \omega')) \; d\omega = \int_{[0,2\pi]} h(\boldsymbol{R}(\omega)) \; d\omega$ for arbitrary functions h. Since $SO(2)$ is a Lie group and thus locally compact, the Haar measure is unique up to a multiplicative constant [109, 110]. Thus, Proposition 2 holds.

F.3.3 Verifying Propositions 1 and 2 for rotation invariance in 3D

In this section, let

$$\boldsymbol{R}(\boldsymbol{\omega}) = \begin{bmatrix} \cos(\omega_1) & -\sin(\omega_1) & 0 \\ \sin(\omega_1) & \cos(\omega_1) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos(\omega_2) & 0 & \sin(\omega_2) \\ 0 & 1 & 0 \\ -\sin(\omega_2) & 0 & \cos(\omega_2) \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\omega_3) & -\sin(\omega_3) \\ 0 & \sin(\omega_3) & \cos(\omega_3) \end{bmatrix}$$

be the matrix that performs an *intrinsic* rotation around the z-axis by angle ω_1, followed by a rotation around the new y-axis by angle ω_2 and then a rotation around the new x-axis by angle ω_3.

Note that $SO(3) = \{\boldsymbol{R}(\boldsymbol{\omega}) \mid \boldsymbol{\omega} \in \Omega\}$ with $\Omega = [0, 2\pi] \times [-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, 2\pi]$. To verify our propositions for $\mathbb{T} = SO(3)$, we define $\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}) = \tau(\omega) \circ \xi(\boldsymbol{\psi})$ with $\tau : \Omega \to SO(3)$ and $\xi : \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}^{3(N-2)} \to \mathbb{R}^{N\times 3}$ (note that the first and third argument are non-negative) with

$$\begin{aligned} \tau(\boldsymbol{\omega}) &= \boldsymbol{R}(\boldsymbol{\omega}) \\ \xi(\boldsymbol{\psi}) &= \begin{bmatrix} \psi_1 & 0 & 0 \\ \psi_2 & \psi_3 & 0 \\ \multicolumn{3}{c}{\mathrm{vec}^{-1}(\boldsymbol{\psi}_{4:})} \end{bmatrix}, \end{aligned}$$

where $\mathrm{vec}^{-1} : \mathbb{R}^{3(N-1)} \to \mathbb{R}^{(N-1)\times 3}$ reshapes an input vector into a matrix. Due to the definition of group action $\circ$ from Appendix C.2, we have

$$\lambda(\omega, \boldsymbol{\psi}) = \begin{bmatrix} \psi_1 & 0 & 0 \\ \psi_2 & \psi_3 & 0 \\ \multicolumn{3}{c}{\mathrm{vec}^{-1}(\boldsymbol{\psi}_{4:})} \end{bmatrix} \boldsymbol{R}(\boldsymbol{\omega})^T$$

In other words: We represent each element of $\mathbb{R}^{N\times 3}$ as a matrix whose first row is aligned with the x-axis, and whose second row is in the first or second quadrant of the x-y-plane (because ψ_3 is non-negative) which is then intrinsically rotated by z-y-x angles $\omega_1, \omega_2, \omega_3$.

The function $\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})$ is injective and surjective, save for the set $\{\boldsymbol{Z} \in \mathbb{R}^{N\times 3} \mid \boldsymbol{Z}_1 = \boldsymbol{0} \vee \exists c \in \mathbb{R} : \boldsymbol{Z}_2 = c \cdot \boldsymbol{Z}_1\}$, which has measure zero. That is, for any $\boldsymbol{Z}$ outside this set, $\boldsymbol{Z} = \lambda(\boldsymbol{\omega}, \boldsymbol{\psi})$ has a unique solution. Firstly, $(\psi_1, \omega_1, \omega_2)$ are spherical coordinates of $\boldsymbol{Z}_1$, which are unique for $\boldsymbol{Z}_1 \neq \boldsymbol{0}$. Secondly, the unique angles ω_1 and ω_2 constrain the x-axis after the intrinsic rotation to be co-linear with $\boldsymbol{Z}_1$. Thus, and because we assume that $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are not co-linear, there must be unique angle ω_3 for rotation around the x-axis that ensures that $\boldsymbol{Z}_2$ is in the first or second quadrant of the new x-y-plane. Finally, with the rotation angles $\boldsymbol{\omega}$ and parameter ψ_1 fixed, the remaining parameter values are determined by

$$\begin{bmatrix} \psi_2 & \psi_3 & 0 \\ \multicolumn{3}{c}{\mathrm{vec}^{-1}(\boldsymbol{\psi}_{4:})} \end{bmatrix} \boldsymbol{R}(\boldsymbol{\omega})^T = \boldsymbol{Z}_{2:} \iff \begin{bmatrix} \psi_2 & \psi_3 & 0 \\ \multicolumn{3}{c}{\mathrm{vec}^{-1}(\boldsymbol{\psi}_{4:})} \end{bmatrix} = \boldsymbol{Z}_{2:}(\boldsymbol{R}(\boldsymbol{\omega})^{-1})^T.$$

The function is also surjective on the entirety of $\mathbb{R}^{N\times 3}$. If $\boldsymbol{Z}_1 = \boldsymbol{0}$, we have $\psi_1 = 0$ and ω_1, ω_2 can be chosen arbitrarily. If $\exists c : \boldsymbol{Z}_2 = c \cdot \boldsymbol{Z}_1$, then ω_3 can be chosen arbitrarily. The remaining parameters can be chosen using the procedure described above. Finally, the function is differentiable, we have $\tau(\boldsymbol{0}) = \mathbf{I}_N$ and τ is a surjective function into $SO(3)$, meaning

$$\tau(\boldsymbol{\omega}') \circ \lambda(\boldsymbol{\omega}, \boldsymbol{\psi}) = (\boldsymbol{R}(\boldsymbol{\omega}') \cdot \boldsymbol{R}(\boldsymbol{\omega})) \circ \xi(\boldsymbol{v}) = \lambda(\boldsymbol{\omega}'', \boldsymbol{\psi})$$

for some $\omega'' \in \Omega$. Thus, all criteria from Proposition 1 are fulfilled.

Like in previous sections, we verify Proposition 2 by first calculating the Jacobian of λ. In the following, we use the shorthands $s_i = \sin(\omega_i)$ and $c_i = \cos(\omega_i)$. We have

$$\begin{aligned} \boldsymbol{J}_\lambda(\omega, \boldsymbol{\psi}) &= \begin{bmatrix} \frac{\partial \mathrm{vec}(\lambda)}{\partial \psi_1} & \frac{\partial \mathrm{vec}(\lambda)}{\partial \omega_1} & \frac{\partial \mathrm{vec}(\lambda)}{\partial \omega_2} & \frac{\partial \mathrm{vec}(\lambda)}{\partial \omega_3} & \frac{\partial \mathrm{vec}(\lambda)}{\partial \psi_2} & \cdots & \frac{\partial \mathrm{vec}(\lambda)}{\partial \psi_{(3(N-1))}} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{D} & \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{E} & \boldsymbol{F} & \boldsymbol{C} \end{bmatrix}, \end{aligned}$$

with

$$\boldsymbol{A} = \begin{bmatrix} c_1 c_2 & -\psi_1 c_2 s_1 & -\psi_1 c_1 s_2 \\ c_2 s_1 & \psi_1 c_1 c_2 & -\psi_1 s_1 s_2 \\ -s_2 & 0 & -\psi_1 c_2 \end{bmatrix},$$

$$\boldsymbol{B} = \begin{bmatrix} \psi_3 \left(c_1 c_\gamma s_2 + s_1 s_\gamma\right) & c_1 c_2 & c_1 s_2 s_\gamma - c_\gamma s_1 \\ \psi_3 \left(-c_1 s_\gamma + c_\gamma s_1 s_2\right) & c_2 s_1 & c_1 c_\gamma + s_1 s_2 s_\gamma \\ \psi_3 \left(c_2 c_\gamma\right) & -s_2 & c_2 s_\gamma, \end{bmatrix}$$

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{R}(\boldsymbol{\omega}) & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{R}(\boldsymbol{\omega}) \end{bmatrix}.$$

Due to the block structure and because $\omega_2 \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and $\boldsymbol{\psi}_3 \in \mathbb{R}_+$, we have

$$\begin{aligned} |\det\left(\boldsymbol{J}_\lambda(\omega, \boldsymbol{\psi})\right)| &= \det\left(\boldsymbol{A}\right) \cdot \det\left(\boldsymbol{B}\right) \cdot \det\left(\boldsymbol{C}\right) \\ &= |\psi_1^2 \cos(\omega_2)| \cdot |\psi_3| \cdot |1| \\ &= \psi_1^2 \cdot \psi_3 \cdot \cos(\omega_2). \end{aligned}$$

Thus, our marginal density $\nu_{\boldsymbol{X}}(\boldsymbol{\psi})$ is

$$\begin{aligned} \nu_{\boldsymbol{X}}(\boldsymbol{\psi}) &= \int_\Omega \psi_1^2 \cdot \psi_3 \cdot \cos(\omega_2) \cdot \mu_{\boldsymbol{X}}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi}))\, d\boldsymbol{\omega} \\ &\propto \int_\Omega \cos(\omega_2) \cdot \prod_{n=1}^N \exp\left(-\frac{1}{2\sigma^2}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})_n - \boldsymbol{X}_n)^T \lambda(\boldsymbol{\omega}, \boldsymbol{\psi})_n - \boldsymbol{X}_n\right)\, d\boldsymbol{\omega} \\ &\propto \int_\Omega \cos(\omega_2) \cdot \prod_{n=1}^N \exp\left(\frac{1}{\sigma^2}(\lambda(\boldsymbol{\omega}, \boldsymbol{\psi})_n)^T \lambda(\boldsymbol{\omega}, \boldsymbol{\psi})_n\right)\, d\boldsymbol{\omega} \\ &= \int_\Omega \cos(\omega_2) \cdot \exp\left(\langle \lambda(\boldsymbol{\omega}, \boldsymbol{\psi}), \boldsymbol{X}\rangle_{\mathrm{F}} \,/\, \sigma^2\right)\, d\boldsymbol{\omega} \\ &= \int_\Omega \cos(\omega_2) \cdot \exp\left(\langle \boldsymbol{R}(\boldsymbol{\omega}) \circ \lambda(\boldsymbol{0}_D, \boldsymbol{\psi}), \boldsymbol{X}\rangle_{\mathrm{F}} \,/\, \sigma^2\right)\, d\boldsymbol{\omega}. \end{aligned}$$

This is, up to a multiplicative constant, the unique Haar integral for this angle-based parameterization of $SO(D)$ (see, for instance [137, Chapter 1]).[3] Thus, Proposition 2 holds.

F.3.4 Verifying Propositions 1 and 2 for roto-translation invariance in 2D and 3D

Finally, we prove that the propositions hold for $\mathbb{T} = SE(D)$ with $D \in \{2, 3\}$, which amounts to combining the results from the previous sections. In the following, let $\tau_{\mathrm{rot}} : \Omega_{\mathrm{rot}} \to SO(D)$ with $\tau_{\mathrm{rot}}(\boldsymbol{\omega}) = \boldsymbol{R}(\boldsymbol{\omega})$ be the the parameterization of $SO(D)$ defined in Appendix F.3.2 or Appendix F.3.3. Further let $\xi_{\mathrm{rot}} : \Psi_{\mathrm{rot}} \to \mathbb{R}^{(N-1)\times D}$ be the same function as in Appendix F.3.2 or Appendix F.3.3, but for matrices with $N-1$ instead of N rows.

We begin by defining $\tau : \Phi_{\mathrm{rot}} \times \mathbb{R}^D :\to SE(D)$ and $\xi : \Psi_{\mathrm{rot}} \to \mathbb{R}^{N\times D}$ as follows:

$$\begin{aligned} \tau(\boldsymbol{\omega}, \boldsymbol{b}) &= (\tau_{\mathrm{rot}}(\boldsymbol{\omega}), \boldsymbol{b}) = (\boldsymbol{R}(\boldsymbol{\omega}), \boldsymbol{b}) \\ \xi(\boldsymbol{\psi}) &= \begin{bmatrix} \boldsymbol{0}_D^T \\ \xi(\boldsymbol{\psi}) \end{bmatrix}, \end{aligned}$$

Due to the definition of group action $\circ$ from Appendix C.4, we have

$$\lambda((\boldsymbol{\omega}, \boldsymbol{b}), \boldsymbol{\psi}) = \tau(\boldsymbol{\omega}, \boldsymbol{b}) \circ \xi(\boldsymbol{\psi}) = \begin{bmatrix} \boldsymbol{0}_D^T \\ \xi(\boldsymbol{\psi}) \end{bmatrix} \boldsymbol{R}(\boldsymbol{\omega})^T + \boldsymbol{1}_N \boldsymbol{b}^T.$$

In other words: We represent each element of $\mathbb{R}^{N\times D}$ as a matrix whose first row is zero and whose second row is aligned with the x-axis which is then rotated and finally translated. For $D = 3$ we

[3] Note that they have a factor $\sin$ instead of $\cos$, because they parameterize $SO(D)$ via z-x-z Euler angles. The proof for intrinsic z-y-x rotation is analogous.

additionally constrain the third row to be in the first or second quadrant of the x-y plane (for more details, see Appendix F.3.3).

The function $\lambda((\boldsymbol{\omega}, \boldsymbol{b}), \boldsymbol{\psi})$ is surjective and injective, save for

- $\{\boldsymbol{Z} \in \mathbb{R}^{N\times 2} \mid \boldsymbol{Z}_2 - \boldsymbol{Z}_1 = 0\}$ (if $D = 2$),
- $\{\boldsymbol{Z} \in \mathbb{R}^{N\times 3} \mid \boldsymbol{Z}_2 - \boldsymbol{Z}_1 = 0 \vee \exists c \in \mathbb{R} : \boldsymbol{Z}_2 - \boldsymbol{Z}_1 = c(\boldsymbol{Z}_3 - \boldsymbol{Z}_1)\}$ (if $D = 3$),

which have measure zero. That is, given an $\boldsymbol{Z} \in \mathbb{R}^{N\times D}$ the solution to $\boldsymbol{Z} = \lambda((\boldsymbol{\omega}, \boldsymbol{b}), \boldsymbol{\psi})$ is unique. Evidently, we must have $\boldsymbol{b} = \boldsymbol{Z}_1$. With this parameter fixed, the remaining parameters can be found by solving the equation $\boldsymbol{Z}_{2:} - \boldsymbol{1}_{N-1}\boldsymbol{Z}_1^T = \tau_{\mathrm{rot}}(\boldsymbol{\omega}) \circ \xi_{\mathrm{rot}}(\boldsymbol{\psi})$, whose unique solution we discussed in Appendix F.3.2 and Appendix F.3.3. The function is also surjective on the entirety of $\mathbb{R}^{N\times D}$ for the reasons presented in Appendix F.3.2 and Appendix F.3.3. Further note that the function $\lambda((\boldsymbol{\omega}, \boldsymbol{b}), \boldsymbol{\psi})$ is differentiable and that $\tau(\boldsymbol{0}, \boldsymbol{0}) = (\mathbf{I}_D, \boldsymbol{0}_D)$, which is the identity element of $SE(D)$ (see Appendix C.4). Finally, we see that

$$
\begin{aligned}
(\boldsymbol{R}(\boldsymbol{\omega}'), \boldsymbol{b}') \circ \lambda((\boldsymbol{\omega}, \boldsymbol{b}), \boldsymbol{\psi}) &= ((\boldsymbol{R}(\boldsymbol{\omega}'), \boldsymbol{b}') \cdot (\boldsymbol{R}(\boldsymbol{\omega}), \boldsymbol{b}) \circ \xi(\boldsymbol{\psi}) \\
&= (\boldsymbol{R}(\boldsymbol{\omega}')\boldsymbol{R}(\boldsymbol{\omega}), \boldsymbol{R}(\boldsymbol{\omega})\boldsymbol{b} + \boldsymbol{b}') \circ \xi(\boldsymbol{\psi}) \\
&= \lambda((\boldsymbol{\omega}'', \boldsymbol{b}''), \boldsymbol{\psi})
\end{aligned}
$$

for some $\boldsymbol{\omega}'' \in \Omega, \boldsymbol{b}'' \in \mathbb{R}^D$. Thus, all criteria from Proposition 1 are fulfilled.

Next, we verify Proposition 2 by again showing that the marginal density $\nu_{\boldsymbol{X}}(\boldsymbol{\psi})$ is a Haar integral. The function λ is a composition of two functions: The function λ_{rot} from the previous sections and a translation. Thus, the Jacobian determinant is the product of the two corresponding Jacobians. As discussed in Appendix F.3.1, the Jacobian determinant for translation is 1. Let $\boldsymbol{J}_{\lambda_{\mathrm{rot}}}(\boldsymbol{\omega}, \boldsymbol{\psi})$ be the Jacobian of λ_{rot}. Then

$$
\begin{aligned}
\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) &= \int_{\Omega_{\mathrm{rot}}} |\det\left(\boldsymbol{J}_{\lambda_{\mathrm{rot}}}(\boldsymbol{\omega}, \boldsymbol{\psi})\right)| \cdot \int_{\mathbb{R}^D} \mu_{\boldsymbol{X}}(\lambda((\boldsymbol{\omega}, \boldsymbol{b}), \boldsymbol{\psi}))\, d\boldsymbol{b}\, d\boldsymbol{\omega} \\
&\propto \int_{\Omega_{\mathrm{rot}}} |\det\left(\boldsymbol{J}_{\lambda_{\mathrm{rot}}}(\boldsymbol{\omega}, \boldsymbol{\psi})\right)| \cdot \int_{\mathbb{R}^D} \exp\left(\langle \lambda((\boldsymbol{\omega}, \boldsymbol{b}), \boldsymbol{\psi}), \boldsymbol{X}\rangle_{\mathrm{F}} / \sigma^2\right)\, d\boldsymbol{b} d\,\boldsymbol{\omega} \\
&= \int_{\Omega_{\mathrm{rot}}} |\det\left(\boldsymbol{J}_{\lambda_{\mathrm{rot}}}(\boldsymbol{\omega}, \boldsymbol{\psi})\right)| \cdot \int_{\mathbb{R}^D} \exp\left(\langle (\boldsymbol{R}(\boldsymbol{\omega}), \boldsymbol{b}) \circ \lambda((\boldsymbol{0}_D, \boldsymbol{0}_D), \boldsymbol{\psi}), \boldsymbol{X}\rangle_{\mathrm{F}} / \sigma^2\right)\, d\boldsymbol{b}\, d\boldsymbol{\omega},
\end{aligned}
$$

where $\propto$ absorbs factors that are constant in $\boldsymbol{\omega}$ and $\boldsymbol{b}$. This is a Haar integral, because the inner and outer integral are Haar integrals for $T(D)$ and $SO(D)$, respectively. Let us verify this by considering an arbitrary $(\boldsymbol{R}(\boldsymbol{\omega}'), \boldsymbol{b}') \in SE(D)$. To avoid clutter, define the shorthand $f(\boldsymbol{Z}) = \exp\left(\langle \boldsymbol{Z}, \boldsymbol{X}\rangle_{\mathrm{F}} / \sigma^2\right)$. We have

$$
\begin{aligned}
&\int_{\Omega_{\mathrm{rot}}} |\det\left(\boldsymbol{J}_{\lambda_{\mathrm{rot}}}(\boldsymbol{\omega}, \boldsymbol{\psi})\right)| \cdot \int_{\mathbb{R}^D} f\left(((\boldsymbol{R}(\boldsymbol{\omega}), \boldsymbol{b}) \cdot (\boldsymbol{R}(\boldsymbol{\omega}'), \boldsymbol{b}')) \circ \lambda((\boldsymbol{0}_D, zeros_D), \boldsymbol{\psi})\right)\, d\boldsymbol{b}\, d\boldsymbol{\omega} \\
&= \int_{\Omega_{\mathrm{rot}}} |\det\left(\boldsymbol{J}_{\lambda_{\mathrm{rot}}}(\boldsymbol{\omega}, \boldsymbol{\psi})\right)| \cdot \int_{\mathbb{R}^D} f\left(\lambda(\boldsymbol{0}_D, \boldsymbol{\psi})(\boldsymbol{R}(\omega)\boldsymbol{R}(\boldsymbol{\omega}')) + \boldsymbol{1}_N\left(\boldsymbol{R}(\omega)\boldsymbol{b}' + \boldsymbol{b}\right)^T\right)\, d\boldsymbol{b}\, d\boldsymbol{\omega} \\
&= \int_{\Omega_{\mathrm{rot}}} |\det\left(\boldsymbol{J}_{\lambda_{\mathrm{rot}}}(\boldsymbol{\omega}, \boldsymbol{\psi})\right)| \cdot \int_{\mathbb{R}^D} f\left(\lambda(\boldsymbol{0}_D, \boldsymbol{\psi})(\boldsymbol{R}(\omega)\boldsymbol{R}(\boldsymbol{\omega}')) + \boldsymbol{1}_N\boldsymbol{b}^T\right)\, d\boldsymbol{b}\, d\boldsymbol{\omega} \\
&= \int_{\Omega_{\mathrm{rot}}} |\det\left(\boldsymbol{J}_{\lambda_{\mathrm{rot}}}(\boldsymbol{\omega}, \boldsymbol{\psi})\right)| \cdot \int_{\mathbb{R}^D} f\left(\lambda(\boldsymbol{0}_D, \boldsymbol{\psi})\boldsymbol{R}(\omega) + \boldsymbol{1}_N\boldsymbol{b}^T\right)\, d\boldsymbol{b}\, d\boldsymbol{\omega} \\
&= \int_{\Omega_{\mathrm{rot}}} |\det\left(\boldsymbol{J}_{\lambda_{\mathrm{rot}}}(\boldsymbol{\omega}, \boldsymbol{\psi})\right)| \cdot \int_{\mathbb{R}^D} f\left((\boldsymbol{R}(\boldsymbol{\omega}), \boldsymbol{b}) \circ \lambda((\boldsymbol{0}_D, \boldsymbol{0}_D), \boldsymbol{\psi})\right)\, d\boldsymbol{b}\, d\boldsymbol{\omega}.
\end{aligned}
$$

Here, we have first applied the definition of the group action and group operator from Appendix C.4 and then used the fact that we are integrating over Haar measures for $T(D)$ and then $SO(D)$. Since $SE(D)$ is a Lie group and thus locally compact, the Haar measure is unique up to a multiplicative constant [109, 110]. This confirms that Proposition 2 holds.

F.4 Group-specific certificates

In this section, we prove the results for specific invariances from Sections 6.2 to 6.5, i.e. translation invariance (Appendix F.4.1), rotation invariance in 2D (Appendix F.4.2 and 3D (Appendix F.4.3, as well as roto-translation invariance in 2D and 3D (Appendix F.4.4).

F.4.1 Translation invariance

Theorem 4. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under $\mathbb{T} = T(D)$ and $\mathbb{H}_\mathbb{T}$ be defined as in Eq.* (2)*. Then*

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

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

During our proof of Theorem 3 in Appendices F.3 and F.3.1, we have shown that

$$\min_{h\in\mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h^*(\boldsymbol{Z}) = y^*] = \mathop{\mathbf{E}}_{\boldsymbol{\psi}\sim\nu_{\boldsymbol{X}'}} \left[\tilde{h}^*(\boldsymbol{\psi})\right] \tag{29}$$

with

$$\tilde{h}^*(\boldsymbol{\psi}) = \mathbb{1}\left[\frac{\nu_{\boldsymbol{X}'}(\boldsymbol{\psi})}{\nu_{\boldsymbol{X}}(\boldsymbol{\psi})} \leq \kappa\right] \text{ and } \kappa \in \mathbb{R}_+ \text{ such that } \mathop{\mathbf{E}}_{\boldsymbol{\psi}\sim\nu_{\boldsymbol{X}}} \left[\tilde{h}^*(\boldsymbol{\psi})\right] = p_{\boldsymbol{X},y^*} \tag{30}$$

with $\boldsymbol{\psi} \in \mathbb{R}^{(N-1)\cdot D}$ and marginal distribution

$$\begin{aligned}
\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) &= \int_{\mathbb{R}^D} \mu_{\boldsymbol{X}}\left(\begin{bmatrix}\mathbf{0}_D^T\\ \mathrm{vec}^{-1}(\boldsymbol{\psi})\end{bmatrix} + \mathbf{1}_N\boldsymbol{\omega}^T\right) d\boldsymbol{\omega}\\
&= \int_{\mathbb{R}^D} \prod_{d=1}^{D} \mathcal{N}\left(\begin{bmatrix}0\\ \mathrm{vec}^{-1}(\boldsymbol{\psi})_{:,d}\end{bmatrix} + \mathbf{1}_N\boldsymbol{\omega}_d \mid \boldsymbol{X}_{:,d}, \sigma^2\mathbf{I}_N\right) d\boldsymbol{\omega},
\end{aligned}$$

where we have inserted the definition of our isotropic matrix normal smoothing distribution for the second equality. Using the matrix $\boldsymbol{A} = \begin{bmatrix}1 & \mathbf{0}_{N-1}^T\\ -\mathbf{1}_{N-1} & \mathbf{I}_{N-1}\end{bmatrix} \in \mathbb{R}^{N\times N}$ with inverse $\boldsymbol{A}^{-1} = \begin{bmatrix}1 & \mathbf{0}_{N-1}^T\\ \mathbf{1}_{N-1} & \mathbf{I}_{N-1}\end{bmatrix}$ and all-zeros vector $\mathbf{0}_{N-1} \in \mathbb{R}^{N-1}$, all-ones vector $\mathbf{1}_{N-1} \in \mathbb{R}^{N-1}$ and identity matrix $\mathbf{I}_{N-1}$, the marginal density can equivalently be written as

$$\begin{aligned}
\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) &= \int_{\mathbb{R}^D} \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{A}^{-1}\begin{bmatrix}\omega_d\\ \mathrm{vec}^{-1}(\boldsymbol{\psi})_{:,d}\end{bmatrix} \mid \boldsymbol{X}_{:,d}, \sigma^2\mathbf{I}_N\right) d\boldsymbol{\omega},\\
&= \int_{\mathbb{R}^D} \prod_{d=1}^{D} \frac{1}{|\det(\boldsymbol{A})|}\mathcal{N}\left(\begin{bmatrix}\omega_d\\ \mathrm{vec}^{-1}(\boldsymbol{\psi})_{:,d}\end{bmatrix} \mid \boldsymbol{A}\boldsymbol{X}_{:,d}, \sigma^2\boldsymbol{A}\boldsymbol{A}^T\right) d\boldsymbol{\omega},\\
&= \int_{\mathbb{R}^D} \prod_{d=1}^{D} \mathcal{N}\left(\begin{bmatrix}\omega_d\\ \mathrm{vec}^{-1}(\boldsymbol{\psi})_{:,d}\end{bmatrix} \mid \boldsymbol{A}\boldsymbol{X}_{:,d}, \sigma^2\boldsymbol{A}\boldsymbol{A}^T\right) d\boldsymbol{\omega},
\end{aligned}$$

where the second equality follows from the change of variable formula for densities (see also Lemma 2) and the third equality is due to $\det(\boldsymbol{A}) = 1$. Evidently, we are marginalizing out the first dimension of each of the D densities. For normal distributions, this is equivalent to dropping the first row of the mean as well as the first row and column of the covariance matrix, i.e.

$$\begin{aligned}
\nu_{\boldsymbol{X}}(\boldsymbol{\psi}) &= \prod_{d=1}^{D} \mathcal{N}\left(\mathrm{vec}^{-1}(\boldsymbol{\psi})_{:,d} \mid \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}, \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)\\
&= \mathcal{N}\left(\boldsymbol{\psi} \mid \boldsymbol{m}_{\boldsymbol{X}}, \boldsymbol{\Sigma}\right)
\end{aligned}$$

with

$$\boldsymbol{m}_{\boldsymbol{X}} = \operatorname{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}\right),$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{B} \end{bmatrix}$$

and $\boldsymbol{B} = \sigma^2 \boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T$.

Inserting back into Eqs. (29) and (30), we see that the optimal value of our variance-constrained optimization problem $\min_{h\in\mathbb{H}_{\mathrm{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h^*(\boldsymbol{Z}) = y^*\right]$ equals the optimal value of the black-box certification problem for anisotropic normal smoothing distribution $\mathcal{N}\left(\boldsymbol{m}_{\boldsymbol{X}}, \boldsymbol{\Sigma}\right)$. This optimal value has been derived in prior work (see [138] and Appendix A of [94]) and is

$$\Phi\left(\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right) - \sqrt{\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right)}\right)$$

We can conclude our proof by showing that

$$\sqrt{\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right)} = \frac{1}{\sigma}\left|\left|\boldsymbol{\Delta} - \boldsymbol{1}_N\overline{\boldsymbol{\Delta}}\right|\right|_2 \tag{31}$$

To this end, we need the following result:

Lemma 7. *Consider an arbitrary vector $\boldsymbol{x} \in \mathbb{R}^D$ and let $\overline{\boldsymbol{x}} = \frac{1}{D}\sum_{d=1}^D x_d$ be its average. Then*

$$\boldsymbol{x}^T\left(\boldsymbol{x} - \boldsymbol{1}_D\overline{\boldsymbol{x}}\right) = \left(\boldsymbol{x} - \boldsymbol{1}_D\overline{\boldsymbol{x}}\right)^T\left(\boldsymbol{x} - \boldsymbol{1}_D\overline{\boldsymbol{x}}\right).$$

Proof. We substract the left-hand side from the right-hand side

$$\begin{aligned} & \boldsymbol{x}^T\left(\boldsymbol{x} - \boldsymbol{1}_D\overline{\boldsymbol{x}}\right) = \left(\boldsymbol{x} - \boldsymbol{1}_D\overline{\boldsymbol{x}}\right)^T\left(\boldsymbol{x} - \boldsymbol{1}_D\overline{\boldsymbol{x}}\right) \\ \iff & 0 = \left(-\boldsymbol{1}_D\overline{\boldsymbol{x}}\right)^T\left(\boldsymbol{x} - \boldsymbol{1}_D\overline{\boldsymbol{x}}\right) \\ \iff & 0 = \overline{\boldsymbol{x}}\sum_{d=1}^D \overline{\boldsymbol{x}} - \overline{\boldsymbol{x}}\sum_{d=1}^D x_d \\ \iff & 0 = \overline{\boldsymbol{x}}D\overline{\boldsymbol{x}} - \overline{\boldsymbol{x}}D\overline{\boldsymbol{x}}, \end{aligned}$$

where the last equality follows from the fact that $\sum_{d=1}^D x_d = D\overline{\boldsymbol{x}}$. □

Now, we can proceed by using the fact that the inverse of a block-diagonal matrix is also a block-diagonal matrix and inserting the definitions of $\boldsymbol{m}_{\boldsymbol{X}'}$, $\boldsymbol{m}_{\boldsymbol{X}}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{B}$ to show that:

$$\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right) \tag{32}$$
$$= \left(\operatorname{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'\right) - \operatorname{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}\right)\right)^T \boldsymbol{\Sigma}^{-1}\left(\operatorname{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'\right) - \operatorname{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}\right)\right) \tag{33}$$
$$= \sum_{d=1}^D \left(\boldsymbol{A}_{2:}\boldsymbol{X}'_{:,d} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}\right)^T \boldsymbol{B}^{-1}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'_{:,d} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}\right) \tag{34}$$
$$= \sum_{d=1}^D \left(\boldsymbol{A}_{2:}\boldsymbol{X}'_{:,d} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}\right)^T \frac{1}{\sigma^2}\left(\boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T\right)^{-1}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'_{:,d} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}\right) \tag{35}$$
$$= \frac{1}{\sigma^2}\sum_{d=1}^D \left(\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,d}\right)^T\left(\boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T\right)^{-1}\left(\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,d}\right), \tag{36}$$
$$= \frac{1}{\sigma^2}\sum_{d=1}^D \left(\boldsymbol{\Delta}_{:,d}\right)^T\left(\boldsymbol{A}_{2:}\right)^T\left(\boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T\right)^{-1}\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,d}, \tag{37}$$

where for the second to last last equality we have used that $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$.

Our next step is to compute $(\boldsymbol{A}_{2:})^T \left(\boldsymbol{A}_{2:} (\boldsymbol{A}_{2:})^T\right)^{-1} \boldsymbol{A}_{2:}$. Recall that $\boldsymbol{A}_{2:} \in \mathbb{R}^{(N-1)\times D}$ and $\boldsymbol{A}_{2:} = [-\mathbf{1}_{N-1} \quad \mathbf{I}_{N-1}]$. Matrix multiplication shows that

$$\boldsymbol{A}_{2:} (\boldsymbol{A}_{2:})^T = [-\mathbf{1}_{N-1} \quad \mathbf{I}_{N-1}] [-\mathbf{1}_{N-1} \quad \mathbf{I}_{N-1}]^T = \mathbf{1}_{N-1,N-1} + \mathbf{I}_{N-1} = \begin{bmatrix} 2 & 1 & \dots & 1 \\ 1 & 2 & \dots & 1 \\ \vdots & \vdots & \ddots & 1 \\ 1 & 1 & \dots & 2 \end{bmatrix}.$$

The matrix is sufficiently simple to be inverted by inspection:

$$\left(\boldsymbol{A}_{2:} (\boldsymbol{A}_{2:})^T\right)^{-1} = \mathbf{I}_{N-1} - \frac{1}{N}\mathbf{1}_{N-1,N-1} = \begin{bmatrix} 1-\frac{1}{N} & -\frac{1}{N} & \dots & -\frac{1}{N} \\ -\frac{1}{N} & 1-\frac{1}{N} & \dots & -\frac{1}{N} \\ \vdots & \vdots & \ddots & \frac{1}{N} \\ -\frac{1}{N} & -\frac{1}{N} & \dots & 1-\frac{1}{N} \end{bmatrix}.$$

Finally, matrix multiplication shows that

$$(\boldsymbol{A}_{2:})^T \left(\boldsymbol{A}_{2:} (\boldsymbol{A}_{2:})^T\right)^{-1} \boldsymbol{A}_{2:} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_{N,N} = \begin{bmatrix} 1-\frac{1}{N} & -\frac{1}{N} & \dots & -\frac{1}{N} \\ -\frac{1}{N} & 1-\frac{1}{N} & \dots & -\frac{1}{N} \\ \vdots & \vdots & \ddots & \frac{1}{N} \\ -\frac{1}{N} & -\frac{1}{N} & \dots & 1-\frac{1}{N} \end{bmatrix}. \tag{38}$$

Note that the matrix in Eq. (38) transforms a vector by subtracting its average from all entries. Inserting into Eq. (37) shows that

$$\begin{aligned} &(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}) \\ =&\frac{1}{\sigma^2} \sum_{d=1}^{D} (\boldsymbol{\Delta}_{:,d})^T \left(\boldsymbol{\Delta}_{:,d} - \mathbf{1}_D \overline{\boldsymbol{\Delta}_{:,d}}\right) \\ =&\frac{1}{\sigma^2} \sum_{d=1}^{D} \left(\boldsymbol{\Delta}_{:,d} - \mathbf{1}_D \overline{\boldsymbol{\Delta}_{:,d}}\right)^T \left(\boldsymbol{\Delta}_{:,d} - \mathbf{1}_D \overline{\boldsymbol{\Delta}_{:,d}}\right) \\ =&\frac{1}{\sigma^2} \sum_{n=1}^{N} \sum_{d=1}^{D} \left(\boldsymbol{\Delta} - \mathbf{1}_N \overline{\boldsymbol{\Delta}}\right)_{n,d}^2, \end{aligned}$$

where $\overline{\boldsymbol{\Delta}_{:,d}} \in \mathbb{R}$ is the average of column $\boldsymbol{\Delta}_{:,d}$, $\overline{\boldsymbol{\Delta}} \in \mathbb{R}^{1\times D}$ are the column-wise averages of matrix $\boldsymbol{\Delta}$, the second equality is due to Lemma 7 and the third equality uses the definition of inner products.

Taking the square root and using the definition of the Frobenius norm yields our desired result:

$$\sqrt{(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}})} = \sqrt{\frac{1}{\sigma^2} \sum_{n=1}^{N} \sum_{d=1}^{D} \left(\boldsymbol{\Delta} - \mathbf{1}_N \overline{\boldsymbol{\Delta}}\right)_{n,d}^2} = \frac{1}{\sigma} \left|\left|\boldsymbol{\Delta} - \mathbf{1}_N \overline{\boldsymbol{\Delta}}\right|\right|_2.$$

F.4.2 Rotation invariance in 2D

Theorem 6. *Let $g : \mathbb{R}^{N\times 2} \to \mathbb{Y}$ be invariant under $\mathbb{T} = SO(2)$ and $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2). *Define the indicator function $\rho : \mathbb{R}^4 \to \{0,1\}$ with*

$$\rho(\boldsymbol{q}) = \mathbb{1}\left[\mathcal{I}_0\left(\sqrt{q_1^2 + q_2^2}\right) / \mathcal{I}_0\left(\sqrt{q_3^2 + q_4^2}\right) \leq \kappa\right],$$
$$\textit{with } \kappa \in \mathbb{R} \textit{ such that } \mathop{\mathbf{E}}_{\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(2)},\boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q})] = p_{\boldsymbol{X},y^*}.$$

Then

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

where

$$\boldsymbol{m}^{(1)} = \frac{1}{\sigma^2}\begin{bmatrix} 2\epsilon_1 + ||\boldsymbol{X}||_2^2 + ||\boldsymbol{\Delta}||_2^2 \\ 0 \\ \epsilon_1 + ||\boldsymbol{X}||_2^2 \\ \epsilon_2 \end{bmatrix}, \quad \boldsymbol{m}^{(2)} = \frac{1}{\sigma^2}\begin{bmatrix} \epsilon_1 + ||\boldsymbol{X}||_2^2 \\ -\epsilon_2 \\ ||\boldsymbol{X}||_2^2 \\ 0 \end{bmatrix},$$

$$\boldsymbol{\Sigma} = \frac{1}{\sigma^2}\begin{bmatrix} 2\epsilon_1 + ||\boldsymbol{X}||_2^2 + ||\boldsymbol{\Delta}||_2^2 & 0 & \epsilon_1 + ||\boldsymbol{X}||_2^2 & \epsilon_2 \\ 0 & 2\epsilon_1 + ||\boldsymbol{X}||_2^2 + ||\boldsymbol{\Delta}||_2^2 & -\epsilon_2 & \epsilon_1 + ||\boldsymbol{X}||_2^2 \\ \epsilon_1 + ||\boldsymbol{X}||_2^2 & -\epsilon_2 & ||\boldsymbol{X}||_2^2 & 0 \\ \epsilon_2 & \epsilon_1 + ||\boldsymbol{X}||_2^2 & 0 & ||\boldsymbol{X}||_2^2. \end{bmatrix},$$

with clean data norm $||\boldsymbol{X}||_2$*, perturbation norm* $||\boldsymbol{\Delta}||_2$ *and parameters* $\epsilon_1 = \langle \boldsymbol{X}, \boldsymbol{\Delta}\rangle_{\mathrm{F}}$*,* $\epsilon_2 = \langle \boldsymbol{X}\boldsymbol{R}\left(-\pi/2\right)^T, \boldsymbol{\Delta}\rangle_{\mathrm{F}}$*.*

We know from Theorem 3 and our derivations in Appendix F.3.2 that

$$\min_{h\in\mathbb{H}_{\mathrm{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right] = \mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h^*(\boldsymbol{Z})\right], \tag{39}$$

with

$$h^*(\boldsymbol{Z}) = \mathbb{1}\left[\frac{\beta_{\boldsymbol{X}'}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}(\boldsymbol{Z})} \leq \kappa\right] \text{ and } \kappa \in \mathbb{R}_+ \text{ such that } \mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h^*(\boldsymbol{Z})\right] = p_{\boldsymbol{X},y^*}, \tag{40}$$

$$\beta_{\boldsymbol{X}}(\boldsymbol{Z}) = \int_{[0,2\pi]} \exp\left(\left\langle \boldsymbol{Z}\boldsymbol{R}(\omega)^T, \boldsymbol{X}\right\rangle_{\mathrm{F}} / \sigma^2\right)\, d\omega,$$

$$\boldsymbol{R}(\omega) = \begin{bmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{bmatrix}.$$

To prove Theorem 6, we shall first calculate the integral $\beta_{\boldsymbol{X}}(\boldsymbol{Z})$ characterizing our worst-case classifier and then use the fact that it applies an affine transformation to our random input data sampled from $\mu_{\boldsymbol{X}}$ and $\mu_{\boldsymbol{X}'}$.

We begin by factoring out $\cos(\omega)$ and $\sin(\omega)$ from the exponent:

$$\begin{aligned}
\left\langle \boldsymbol{Z}\boldsymbol{R}(\omega)^T, \boldsymbol{X}\right\rangle_{\mathrm{F}} &= \sum_{n=1}^{N} (\boldsymbol{R}(\omega)\boldsymbol{Z}_n)(\boldsymbol{X}_n)^T \\
&= \sum_{n=1}^{N} \cos(\omega)\boldsymbol{Z}_{n,1}\boldsymbol{X}_{n,1} - \sin(\omega)\boldsymbol{Z}_{n,2}\boldsymbol{X}_{n,1} + \sin(\omega)\boldsymbol{Z}_{n,1}\boldsymbol{X}_{n,2} + \cos(\omega)\boldsymbol{Z}_{n,2}\boldsymbol{X}_{n,2} \\
&= \cos(\omega)\begin{bmatrix}\boldsymbol{X}_{:,1}\\ \boldsymbol{X}_{:,2}\end{bmatrix}^T \begin{bmatrix}\boldsymbol{Z}_{:,1}\\ \boldsymbol{Z}_{:,2}\end{bmatrix} + \sin(\omega)\begin{bmatrix}\boldsymbol{X}_{:,2}\\ -\boldsymbol{X}_{:,1}\end{bmatrix}^T \begin{bmatrix}\boldsymbol{Z}_{:,1}\\ \boldsymbol{Z}_{:,2}\end{bmatrix} \\
&= \cos(\omega)\cdot \mathrm{vec}(\boldsymbol{X})^T\mathrm{vec}(\boldsymbol{Z}) + \sin(\omega)\cdot \mathrm{vec}\left(\boldsymbol{X}\boldsymbol{R}\left(-\pi/2\right)^T\right)^T \mathrm{vec}(\boldsymbol{Z}).
\end{aligned}$$

We can then show that

$$\begin{aligned}
&\beta_{\boldsymbol{X}}(\boldsymbol{Z}) \\
&= \int_{[0,2\pi]} \exp\left(\frac{1}{\sigma^2}\left(\cos(\omega)\cdot \mathrm{vec}(\boldsymbol{X})^T\mathrm{vec}(\boldsymbol{Z}) + \sin(\omega)\cdot \mathrm{vec}\left(\boldsymbol{X}\boldsymbol{R}\left(-\pi/2\right)^T\right)^T \mathrm{vec}(\boldsymbol{Z})\right)\right)\, d\omega \\
&= \int_{[0,2\pi]} \exp\left(\cos(\omega)\cdot\sqrt{\left(\mathrm{vec}(\boldsymbol{X})^T\mathrm{vec}(\boldsymbol{Z})/\sigma^2\right)^2 + \left(\mathrm{vec}\left(\boldsymbol{X}\boldsymbol{R}\left(-\pi/2\right)^T\right)^T \mathrm{vec}(\boldsymbol{Z})/\sigma^2\right)^2}\right)\, d\omega \\
&= 2\pi\cdot\mathcal{I}_0\left(\sqrt{\left(\mathrm{vec}(\boldsymbol{X})^T\mathrm{vec}(\boldsymbol{Z})/\sigma^2\right)^2 + \left(\mathrm{vec}\left(\boldsymbol{X}\boldsymbol{R}\left(-\pi/2\right)^T\right)^T \mathrm{vec}(\boldsymbol{Z})/\sigma^2\right)^2}\right),
\end{aligned}$$

where the second equality follows from the fact that $\cos(\omega)\eta_1 + \sin(\omega)\eta_2 = \cos(\omega+\alpha)\sqrt{\eta_1^2+\eta_2^2}$ for some $\alpha \in [0, 2\pi]$ (see also Eq. 29 of [139]) and the third equality is due to the integral representation

of the modified Bessel function of the first kind with order 0 (see Eq. 10.32.1 of [140]):

$$\mathcal{I}_0(z) = \frac{1}{\pi} \int\limits_{[0,\pi]} \exp\left(\cos(\alpha) z\right) \, d\alpha.$$

We can now insert this expression for the worst-case classifier into the expectations w.r.t. clean and perturbed smoothing distributions $\mu_{\boldsymbol{X}}, \mu'_{\boldsymbol{X}}$ from Eqs. (39) and (40) . Note that, by definition, we have $\mathrm{vec}(\boldsymbol{Z}) \sim \mathcal{N}(\mathrm{vec}(\boldsymbol{X}), \sigma^2 \cdot \mathbf{I}_{2N})$ for $\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}$. Thus:

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h^*(\boldsymbol{Z})] &= \mathop{\mathbf{E}}_{\boldsymbol{q} \sim \mathcal{N}(\boldsymbol{m}^{(1)}, \boldsymbol{\Sigma})} \left[\mathbb{1} \left[\mathcal{I}_0 \left(\sqrt{q_1^2 + q_2^2} \right) / \, \mathcal{I}_0 \left(\sqrt{q_3^2 + q_4^2} \right) \leq \kappa \right] \right], \\
\mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}} [h^*(\boldsymbol{Z})] &= \mathop{\mathbf{E}}_{\boldsymbol{q} \sim \mathcal{N}(\boldsymbol{m}^{(2)}, \boldsymbol{\Sigma})} \left[\mathbb{1} \left[\mathcal{I}_0 \left(\sqrt{q_1^2 + q_2^2} \right) / \, \mathcal{I}_0 \left(\sqrt{q_3^2 + q_4^2} \right) \leq \kappa \right] \right],
\end{aligned}$$

with means $\boldsymbol{m}^{(1)} = \boldsymbol{W} \mathrm{vec}(\boldsymbol{X}')$, $\boldsymbol{m}^{(2)} = \boldsymbol{W} \mathrm{vec}(\boldsymbol{X})$ and covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{W} \boldsymbol{W}^T$, where

$$\boldsymbol{W} = \frac{1}{\sigma^2} \cdot \begin{bmatrix} \mathrm{vec}(\boldsymbol{X}')^T \\ \mathrm{vec}\left(\boldsymbol{X}' \boldsymbol{R}\left(-\pi \,/\, 2\right)^T \right) \\ \mathrm{vec}(\boldsymbol{X})^T \\ \mathrm{vec}\left(\boldsymbol{X} \boldsymbol{R}\left(-\pi \,/\, 2\right)^T \right) \end{bmatrix}.$$

Calculating the matrix-vector and matrix-matrix products yields the values from Theorem 6, thus proving that the certificate is valid.

F.4.3 Rotation invariance in 3D

As discussed in Section 6.4, we can apply Theorem 3 to $\mathbb{T} = SO(3)$, but do not have a closed-form expression for the Haar integral $\beta_{\boldsymbol{X}}(\boldsymbol{Z})$ characterizing our worst-case classifier. We can however evaluate it using numerical integration. In the following, we first show how we can reduce the number of integration variables to facilitate numerical integration. We then show that, similar to the 2D case, the Monte Carlo certification procedure presented in Section 6.3 only requires sampling from a 16-dimensional normal distribution, which allows us to obtain tight bounds via a large number of samples at little computational cost.

Reducing the number of integration variables. Recall from our proof of Theorem 3 in Appendices F.3 and F.3.3 that

$$\beta_{\boldsymbol{X}}(\boldsymbol{Z}) = \int_{\Omega} \cos(\omega_2) \cdot \exp\left(\left\langle \boldsymbol{Z} \boldsymbol{R}(\boldsymbol{\omega})^T, \boldsymbol{X} \right\rangle_{\mathrm{F}} / \, \sigma^2 \right) \, d\boldsymbol{\omega}, \tag{41}$$

with $\Omega = [0, 2\pi] \times [-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, 2\pi]$ and rotation matrix

$$\begin{aligned}
\boldsymbol{R}(\boldsymbol{\omega}) &= \begin{bmatrix} \cos(\omega_1) & -\sin(\omega_1) & 0 \\ \sin(\omega_1) & \cos(\omega_1) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos(\omega_2) & 0 & \sin(\omega_2) \\ 0 & 1 & 0 \\ -\sin(\omega_2) & 0 & \cos(\omega_2) \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\omega_3) & -\sin(\omega_3) \\ 0 & \sin(\omega_3) & \cos(\omega_3) \end{bmatrix} \\
&= \begin{bmatrix} \cos(\omega_1) & -\sin(\omega_1) & 0 \\ \sin(\omega_1) & \cos(\omega_1) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos(\omega_2) & \sin(\omega_2)\sin(\omega_3) & \cos(\omega_3)\sin(\omega_2) \\ 0 & \cos(\omega_3) & -\sin(\omega_3) \\ -\sin(\omega_2) & \cos(\omega_2)\sin(\omega_3) & \cos(\omega_2)\cos(\omega_3) \end{bmatrix} \\
&:= \begin{bmatrix} \cos(\omega_1) & -\sin(\omega_1) & 0 \\ \sin(\omega_1) & \cos(\omega_1) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \tilde{\boldsymbol{R}}(\boldsymbol{\omega}_{2:}),
\end{aligned}$$

with $\tilde{\boldsymbol{R}}(\boldsymbol{\omega}_{2:})$ being the matrix that rotates around the y- and z-axis by angles ω_2 and ω_3, respectively. While we cannot evaluate this term analytically, we can calculate the inner integral over ω_1 in order

to reduce it to a double integral. We begin by factoring out $\cos(\omega_1)$ and $\sin(\omega_1)$ from the exponent:

$$
\begin{aligned}
&\left\langle \boldsymbol{Z}\boldsymbol{R}(\boldsymbol{\omega})^T, \boldsymbol{X} \right\rangle_{\mathrm{F}} / \sigma^2 \\
=&\sum_{n=1}^{N} (\boldsymbol{X}_n)^T \boldsymbol{R}(\boldsymbol{\omega}) \boldsymbol{Z}_n / \sigma^2 \\
=&\cos(\omega_1) \left(\sum_{n=1}^{N} \boldsymbol{X}_{n,1} \left(\tilde{\boldsymbol{R}}(\boldsymbol{\omega}_{2:}) \boldsymbol{Z}_n \right)_1 + \boldsymbol{X}_{n,2} \left(\tilde{\boldsymbol{R}}(\boldsymbol{\omega}_{2:}) \boldsymbol{Z}_n \right)_2 \right) / \sigma^2 \\
&+ \sin(\omega_1) \left(\sum_{n=1}^{N} \boldsymbol{X}_{n,2} \left(\tilde{\boldsymbol{R}}(\boldsymbol{\omega}_{2:}) \boldsymbol{Z}_n \right)_1 - \boldsymbol{X}_{n,1} \left(\tilde{\boldsymbol{R}}(\boldsymbol{\omega}_{2:}) \boldsymbol{Z}_n \right)_2 \right) / \sigma^2 \\
&+ \sum_{n=1}^{N} \boldsymbol{X}_{n,3} \left(\tilde{\boldsymbol{R}}(\boldsymbol{\omega}_{2:}) \boldsymbol{Z}_n \right)_3 / \sigma^2. \\
=&\cos(\omega_1) \cdot \chi_1(\boldsymbol{\omega}_{2:}, \boldsymbol{X}^T \boldsymbol{Z}) + \sin(\omega_1) \cdot \chi_2(\boldsymbol{\omega}_{2:}, \boldsymbol{X}^T \boldsymbol{Z}) + \chi_3(\boldsymbol{\omega}_{2:}, \boldsymbol{X}^T \boldsymbol{Z}),
\end{aligned}
$$

where χ_1, χ_2, χ_3 are shorthands we introduce to avoid clutter in the following derivations. Due to the entries of $\tilde{\boldsymbol{R}}(\boldsymbol{\omega}_2 :)$, the shorthand $\chi_1 : \mathbb{R}^2 \times \mathbb{R}^{3\times 3} \to \mathbb{R}$ is defined as

$$
\begin{aligned}
&\chi_1(\boldsymbol{\omega}_{2:}, \boldsymbol{X}^T \boldsymbol{Z})) \\
:=&\frac{1}{\sigma^2} \sum_{n=1}^{N} \Big(\boldsymbol{X}_{n,1} \left(\cos(\omega_2) \boldsymbol{Z}_{n,1} + \sin(\omega_2)\sin(\omega_3) \boldsymbol{Z}_{n,2} + \cos(\omega_3)\sin(\omega_2) \boldsymbol{Z}_{n,3} \right) \\
&+ \boldsymbol{X}_{n,2} \left(\cos(\omega_3) \boldsymbol{Z}_{n,2} - \sin(\omega_3) \boldsymbol{Z}_{n,3} \right) \Big) \\
=&\cos(\omega_2) (\boldsymbol{X}_{:,1})^T \boldsymbol{Z}_{:,1} + \sin(\omega_2)\sin(\omega_3) (\boldsymbol{X}_{:,1})^T \boldsymbol{Z}_{:,2} + \cos(\omega_3)\sin(\omega_2) (\boldsymbol{X}_{:,1})^T \boldsymbol{Z}_{:,3} \\
&+ \cos(\omega_3) (\boldsymbol{X}_{:,2})^T \boldsymbol{Z}_{:,2} - \sin(\omega_3) (\boldsymbol{X}_{:,2})^T \boldsymbol{Z}_{:,3} \\
=&\cos(\omega_2) (\boldsymbol{X}^T \boldsymbol{Z})_{1,1} + \sin(\omega_2)\sin(\omega_3) (\boldsymbol{X}^T \boldsymbol{Z})_{1,2} + \cos(\omega_3)\sin(\omega_2) (\boldsymbol{X}^T \boldsymbol{Z})_{1,3} \\
&+ \cos(\omega_3) (\boldsymbol{X}^T \boldsymbol{Z})_{2,2} - \sin(\omega_3) (\boldsymbol{X}^T \boldsymbol{Z})_{2,3}
\end{aligned}
$$

Similarly, the shorthand $\chi_2 : \mathbb{R}^2 \times \mathbb{R}^{3\times 3} \to \mathbb{R}$ is defined as

$$
\begin{aligned}
&\chi_2(\boldsymbol{\omega}_{2:,}, \boldsymbol{X}^T \boldsymbol{Z}) \\
:=&\cos(\omega_2) (\boldsymbol{X}^T \boldsymbol{Z})_{2,1} + \sin(\omega_2)\sin(\omega_3) (\boldsymbol{X}^T \boldsymbol{Z})_{2,2} + \cos(\omega_3)\sin(\omega_2) (\boldsymbol{X}^T \boldsymbol{Z})_{2,3} \\
&- \cos(\omega_3) (\boldsymbol{X}^T \boldsymbol{Z})_{1,2} + \sin(\omega_3) (\boldsymbol{X}^T \boldsymbol{Z})_{1,3}.
\end{aligned}
$$

The shorthand $\chi_3 : \mathbb{R}^2 \times \mathbb{R}^{3\times 3} \to \mathbb{R}$ is defined as

$$
\begin{aligned}
&\chi_3(\boldsymbol{\omega}_{2:,}, \boldsymbol{X}^T \boldsymbol{Z}) \\
:=&-\sin(\omega_2) (\boldsymbol{X}^T \boldsymbol{Z})_{3,1} + \cos(\omega_2)\sin(\omega_3) (\boldsymbol{X}^T \boldsymbol{Z})_{3,2} + \cos(\omega_2)\cos(\omega_3) (\boldsymbol{X}^T \boldsymbol{Z})_{3,3}.
\end{aligned}
$$

Just like in Appendix F.4.2, we can use the integral formula for the modified Bessel function of the first kind and order zero to eliminate the integral w.r.t. ω_1:

$$
\begin{aligned}
&\beta_{\boldsymbol{X}}(\boldsymbol{Z}) \\
=&\int_{\Omega} \cos(\omega_2) \cdot \exp\left(\chi_3(\boldsymbol{\omega}_{2:}, \boldsymbol{X}^T \boldsymbol{Z}) \right) \cdot \exp\left(\cos(\omega_1) \cdot \chi_1(\boldsymbol{\omega}_{2:}, \boldsymbol{X}, \boldsymbol{Z}) + \sin(\omega_1) \cdot \chi_2(\boldsymbol{\omega}_{2:}, \boldsymbol{X}^T \boldsymbol{Z}) \right) \, d\boldsymbol{\omega} \\
=&2\pi \cdot \int_{[-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, 2\pi]} \cos(\omega_2) \cdot \exp\left(\chi_3(\boldsymbol{\omega}_{2:}, \boldsymbol{X}^T \boldsymbol{Z}) \right) \cdot \mathcal{I}_0 \left(\sqrt{\chi_1(\boldsymbol{\omega}_{2:}, \boldsymbol{X}, \boldsymbol{Z})^2 + \chi_2(\boldsymbol{\omega}_{2:}, \boldsymbol{X}, \boldsymbol{Z})^2} \right) \, d\boldsymbol{\omega}_{2:} \\
:=&\hat{\beta}(\boldsymbol{X}^T \boldsymbol{Z}).
\end{aligned}
$$

We introduce $\hat{\beta}(\boldsymbol{X}^T \boldsymbol{Z})$ to avoid clutter in the next equations and to highlight that this function characterizing our worst-case classifier only depends on $\boldsymbol{X}^T \boldsymbol{Z} \in \mathbb{R}^{3\times 3}$.

Efficient Monte Carlo certification. Just like in Appendix F.4.2, we can now use the fact that our worst-case classifier only depends on a small number of variables that are the result of linearly

transforming our randomized inputs $\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}$, namely the 9 entires of $\boldsymbol{X}^T\boldsymbol{Z}$. Let $\zeta : \mathbb{R}^8 \to \mathbb{R}^{3\times 3}$ with

$$\zeta(\boldsymbol{q}) = \begin{bmatrix} q_1 & q_3 & q_6 \\ 0 & q_4 & q_7 \\ q_2 & q_5 & q_8 \end{bmatrix}$$

be a function that zero-pads its 8-dimensional input vector before devectorizing into shape 3×3. Since $\mathrm{vec}(\boldsymbol{Z}) \sim \mathcal{N}\left(\mathrm{vec}(\boldsymbol{X}), \sigma^2 \mathbf{I}_{3N}\right)$ if $\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}$, we have by definition of $\hat{\beta}$ that

$$\mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[\frac{\beta_{\boldsymbol{X}'}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}(\boldsymbol{Z})} \leq \kappa\right] = \mathop{\mathbf{E}}_{\boldsymbol{q}\sim\mathcal{N}(\boldsymbol{m}_{\boldsymbol{X}},\boldsymbol{\Sigma})}\left[\frac{\hat{\beta}(\zeta(\boldsymbol{q}_{1:8}))}{\hat{\beta}(\zeta(\boldsymbol{q}_{9:16}))} \leq \kappa\right],$$

with $\boldsymbol{m}_{\boldsymbol{X}} = \boldsymbol{W}\mathrm{vec}(\boldsymbol{X})$, $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{W}\boldsymbol{W}^T$ and

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{X}'_{:,1} & \boldsymbol{X}'_{:,3} & \mathbf{0} & \mathbf{0} & \boldsymbol{X}_{:,1} & \boldsymbol{X}_{:,3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{X}' & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{X}' & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{X} \end{bmatrix}^T.$$

Thus, the expectation that provides the optimal value of our variance-constrained optimization problem can be probabilistically bounded via Monte Carlo sampling from a 16-dimensional normal distribution, regardless of the data dimensionality.

F.4.4 Roto-translation invariance in 2D and 3D

For the roto-granslation group $SE(D)$, we prove that evaluating the tight certificate is equivalent to evaluating the tight certificate for rotation group $SO(D)$ after centering the clean and perturbed data $\boldsymbol{X}, \boldsymbol{X}'$ by substracting their averages.

Theorem 7. *Let $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \mathbf{1}_N\overline{\boldsymbol{X}}$ with column-wise averages $\overline{\boldsymbol{X}}, \overline{\boldsymbol{X}'} \in \mathbb{R}^{1\times D}$. Further let η be a right Haar measure on $SO(3)$ and $\mu_{\boldsymbol{X}}$ our isotropic matrix normal smoothing distribution. Let*

$$\beta_{\boldsymbol{X}}^{SE(D)} = \int_{SE(D)}\int_{\mathbb{R}^D} \exp\left(\left\langle \boldsymbol{Z}\boldsymbol{R}^T + \mathbf{1}_N\boldsymbol{b}^T, \boldsymbol{X}\right\rangle_{\mathrm{F}} / \sigma^2\right)\, d\boldsymbol{b}\, d\eta(\boldsymbol{R})$$

be the Haar integral characterizing the worst-case classifier for roto-translation invariance (see also Appendix F.3.4) and

$$\beta_{\boldsymbol{X}}^{SO(D)} = \int_{SO(D)}\int_{\mathbb{R}^D} \exp\left(\left\langle \boldsymbol{Z}\boldsymbol{R}^T, \boldsymbol{X}\right\rangle_{\mathrm{F}} / \sigma^2\right)\, d\eta(\boldsymbol{R})$$

be the Haar integral characterizing the worst-case classifier for rotation invariance (see also Appendices F.3.2 and F.3.3). Then, for all $\boldsymbol{V} \in \mathbb{R}^{N\times D}$ and $\kappa \in \mathbb{R}_+$,

$$\mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{V}}}\left[\mathbb{1}\left[\frac{\beta_{\boldsymbol{X}'}^{SE(D)}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}^{SE(D)}(\boldsymbol{Z})} \leq \kappa\right]\right] = \mathop{\mathbf{E}}_{\boldsymbol{Z}\sim\mu_{\tilde{\boldsymbol{V}}}}\left[\mathbb{1}\left[\frac{\beta_{\tilde{\boldsymbol{X}}'}^{SO(D)}(\boldsymbol{Z})}{\beta_{\tilde{\boldsymbol{X}}}^{SO(D)}(\boldsymbol{Z})} \leq \kappa\right]\right].$$

Note that these (with $\boldsymbol{V} = \boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{X}'$) are exactly the expectations determining the optimal value of our variance-constrained optimization problem (see Theorem 3).

To prove this theorem, we first simplify the integrand of $\beta_{\boldsymbol{X}}^{SE(D)}(\boldsymbol{Z})$ by moving the rotation matrix into the second argument of the Frobenius inner product (note that, since $\boldsymbol{R}$ is a rotation matrix, we have $\boldsymbol{R}^{-1} = \boldsymbol{R}^T$):

$$\begin{aligned}
\left\langle \boldsymbol{Z}\boldsymbol{R}^T + \mathbf{1}_N\boldsymbol{b}^T, \boldsymbol{X}\right\rangle_{\mathrm{F}} &= \left\langle \boldsymbol{Z}\boldsymbol{R}^T + \mathbf{1}_N\boldsymbol{b}^T\boldsymbol{R}\boldsymbol{R}^T, \boldsymbol{X}\right\rangle_{\mathrm{F}} \\
&= \left\langle \left(\boldsymbol{Z} + \mathbf{1}_N\boldsymbol{b}^T\boldsymbol{R}\right)\boldsymbol{R}^T, \boldsymbol{X}\right\rangle_{\mathrm{F}} \\
&= \left\langle \boldsymbol{Z} + \mathbf{1}_N\boldsymbol{b}^T\boldsymbol{R}, \boldsymbol{X}\boldsymbol{R}\right\rangle_{\mathrm{F}} \\
&= \left\langle \boldsymbol{Z} + \mathbf{1}_N(\boldsymbol{R}^T\boldsymbol{b})^T, \boldsymbol{X}\boldsymbol{R}\right\rangle_{\mathrm{F}}
\end{aligned}$$

Next, we bring our integral into the same form as in our derivation of the certificate for translation invariance (see Appendix F.4.1), so that we can easily integrate out the translation vector $\boldsymbol{b}$. First, we eliminate the rotation matrix $\boldsymbol{R}^T$ via the substitution $\boldsymbol{c} = \boldsymbol{R}^T\boldsymbol{b}$:

$$\begin{aligned}\beta_{\boldsymbol{X}}^{SE(D)}(\boldsymbol{Z}) &= \int_{SO(D)} \int_{\mathbb{R}^D} \exp\left(\left\langle \boldsymbol{Z} + \mathbf{1}_N(\boldsymbol{R}^T\boldsymbol{b})^T, \boldsymbol{X}\boldsymbol{R}\right\rangle_{\mathrm{F}} / \sigma^2\right)\, d\boldsymbol{b}\, d\eta(\boldsymbol{R}) \\ &= \int_{SO(D)} \int_{\mathbb{R}^D} \exp\left(\left\langle \boldsymbol{Z} + \mathbf{1}_N\boldsymbol{c}^T, \boldsymbol{X}\boldsymbol{R}\right\rangle_{\mathrm{F}} / \sigma^2\right)\, d\boldsymbol{c}\, d\eta(\boldsymbol{R})\end{aligned}$$

For the next step, let $\boldsymbol{A} = \begin{bmatrix} 1 & \mathbf{0}_{N-1}^T \\ -\mathbf{1}_{N-1} & \mathbf{I}_{N-1}\end{bmatrix} \in \mathbb{R}^{N\times N}$ with inverse $\boldsymbol{A}^{-1} = \begin{bmatrix} 1 & \mathbf{0}_{N-1}^T \\ \mathbf{1}_{N-1} & \mathbf{I}_{N-1}\end{bmatrix}$ and all-zeros vector $\mathbf{0}_{N-1} \in \mathbb{R}^{N-1}$, all-ones vector $\mathbf{1}_{N-1} \in \mathbb{R}^{N-1}$ and identity matrix $\mathbf{I}_{N-1}$. Multiplying a vector or matrix with $\boldsymbol{A}$ is equivalent to subtracting the first row from all other rows. We introduce the redundant factor $\boldsymbol{A}^{-1}\boldsymbol{A}$ and make the substitution $\boldsymbol{u} = \boldsymbol{c} + \boldsymbol{Z}_1$ to show that

$$\begin{aligned}\beta_{\boldsymbol{X}}^{SE(D)}(\boldsymbol{Z}) &= \int_{SO(D)} \int_{\mathbb{R}^D} \exp\left(\left\langle \boldsymbol{A}^{-1}\boldsymbol{A}\left(\boldsymbol{Z} + \mathbf{1}_N\boldsymbol{c}^T\right), \boldsymbol{X}\boldsymbol{R}\right\rangle_{\mathrm{F}} / \sigma^2\right)\, d\boldsymbol{c}\, d\eta(\boldsymbol{R}) \\ &= \int_{SO(D)} \int_{\mathbb{R}^D} \exp\left(\left\langle \boldsymbol{A}^{-1}\begin{bmatrix} \boldsymbol{Z}_1^T + \boldsymbol{c}^T \\ (\boldsymbol{Z}_2 - \boldsymbol{Z}_1)^T \\ \cdots \\ (\boldsymbol{Z}_N - \boldsymbol{Z}_1)^T\end{bmatrix}, \boldsymbol{X}\boldsymbol{R}\right\rangle_{\mathrm{F}} / \sigma^2\right)\, d\boldsymbol{c}\, d\eta(\boldsymbol{R}) \\ &= \int_{SO(D)} \int_{\mathbb{R}^D} \exp\left(\left\langle \boldsymbol{A}^{-1}\begin{bmatrix} \boldsymbol{u}^T \\ (\boldsymbol{Z}_2 - \boldsymbol{Z}_1)^T \\ \cdots \\ (\boldsymbol{Z}_N - \boldsymbol{Z}_1)^T\end{bmatrix}, \boldsymbol{X}\boldsymbol{R}\right\rangle_{\mathrm{F}} / \sigma^2\right)\, d\boldsymbol{u}\, d\eta(\boldsymbol{R}) \\ &\propto \int_{SO(D)} \int_{\mathbb{R}^D} \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{A}^{-1}\begin{bmatrix} \boldsymbol{u}_d \\ \boldsymbol{Z}_{2,d} - \boldsymbol{Z}_{1,d} \\ \cdots \\ \boldsymbol{Z}_{N,d} - \boldsymbol{Z}_{1,d}\end{bmatrix} \mid \boldsymbol{X}\boldsymbol{R}_{:,d}, \sigma^2\mathbf{I}_N\right)\, d\boldsymbol{u}\, d\eta(\boldsymbol{R}).\end{aligned}$$

Next, we use the behavior of multivariate normal densities under affine transformation (see also Lemma 2), as well as the fact that marginalizing out one dimension of a multivariate normal density is equivalent to dropping the correspond row of the mean and corresponding row and column of the adjacency matrix, to show that

$$\begin{aligned}\beta_{\boldsymbol{X}}^{SE(D)}(\boldsymbol{Z}) &\propto \int_{SO(D)} \int_{\mathbb{R}^D} \prod_{d=1}^{D} \mathcal{N}\left(\begin{bmatrix} \boldsymbol{u}_d \\ \boldsymbol{Z}_{2,d} - \boldsymbol{Z}_{1,d} \\ \cdots \\ \boldsymbol{Z}_{N,d} - \boldsymbol{Z}_{1,d}\end{bmatrix} \mid \boldsymbol{A}\boldsymbol{X}\boldsymbol{R}_{:,d}, \sigma^2\boldsymbol{A}\boldsymbol{A}^T\right)\, d\boldsymbol{u}\, d\eta(\boldsymbol{R}) \\ &= \int_{SO(D)} \prod_{d=1}^{D} \mathcal{N}\left(\begin{bmatrix} \boldsymbol{Z}_{2,d} - \boldsymbol{Z}_{1,d} \\ \cdots \\ \boldsymbol{Z}_{N,d} - \boldsymbol{Z}_{1,d}\end{bmatrix} \mid \boldsymbol{A}_{2:}\boldsymbol{X}\boldsymbol{R}_{:,d}, \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)\, d\eta(\boldsymbol{R}) \\ &= \int_{SO(D)} \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,d} \mid \boldsymbol{A}_{2:}\boldsymbol{X}\boldsymbol{R}_{:,d}, \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)\, d\eta(\boldsymbol{R}) \\ &\propto \int_{SO(D)} \prod_{d=1}^{D} \exp\left(\frac{1}{\sigma^2}\boldsymbol{Z}_{:,d}^T(\boldsymbol{A}_{2:})^T\left(\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\boldsymbol{R}_{:,d}\right)\, d\eta(\boldsymbol{R})\end{aligned}$$

Finally, reusing the fact that $(\boldsymbol{A}_{2:})^T\left(\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)^{-1}\boldsymbol{A}_{2:}\boldsymbol{X} = \tilde{\boldsymbol{X}}$ with $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \mathbf{1}_N\overline{\boldsymbol{X}}$ from Appendix F.4.1 proves that

$$\begin{aligned}\beta_{\boldsymbol{X}}^{SE(D)}(\boldsymbol{Z}) &\propto \int_{SO(D)} \prod_{d=1}^{D} \exp\left(\frac{1}{\sigma^2}\boldsymbol{Z}_{:,d}^T\tilde{\boldsymbol{X}}\boldsymbol{R}_{:,d}\right)\, d\eta(\boldsymbol{R}) \\ &= \int_{SO(D)} \exp\left(\left\langle \boldsymbol{Z}\boldsymbol{R}^T, \tilde{\boldsymbol{X}}\right\rangle / \sigma^2\right)\, d\eta(\boldsymbol{R}).\end{aligned}$$

The last term is the Haar integral $\beta_{\tilde{\boldsymbol{X}}}^{SO(D)}(\boldsymbol{Z})$ characterizing the worst-case classifier for $SO(D)$, after centering $\boldsymbol{X}$. This means that we have

$$\frac{\beta_{\boldsymbol{X}'}^{SE(D)}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}^{SE(D)}(\boldsymbol{Z})} = \frac{\beta_{\tilde{\boldsymbol{X}}'}^{SO(D)}(\boldsymbol{Z})}{\beta_{\tilde{\boldsymbol{X}}}^{SO(D)}(\boldsymbol{Z})}.$$

Note that, when using $\propto$ in the previous steps, we have only removed constant factors that appear in the enumerator and denominator and thus cancel out.

The last thing we need to do is verify that not only the classifiers, but also their expectations under $\boldsymbol{V}$ and $\tilde{\boldsymbol{V}}$ are identical (see Theorem 7). We use the fact that the worst-case classifier only depends on a low-dimensional linear projection of our random input variable $\boldsymbol{Z}$ (similar to our derivations for Monte Carlo evaluation in Appendices F.4.2 and F.4.3):

$$\begin{aligned}
\beta_{\boldsymbol{X}}^{SE(D)}(\boldsymbol{Z}) &\propto \int_{SO(D)} \prod_{d=1}^{D} \exp\left(\frac{1}{\sigma^2} \boldsymbol{Z}_{:,d}^T \tilde{\boldsymbol{X}} \boldsymbol{R}_{:,d}\right) \, d\eta(\boldsymbol{R}) \\
&= \int_{SO(D)} \prod_{d=1}^{D} \exp\left(\frac{1}{\sigma^2} (\tilde{\boldsymbol{X}}^T \boldsymbol{Z}_{:,d})^T \boldsymbol{R}_{:,d}\right) \, d\eta(\boldsymbol{R}) \\
&= \int_{SO(D)} \exp\left(\left\langle \tilde{\boldsymbol{X}}^T \boldsymbol{Z}, \boldsymbol{R} \right\rangle / \, \sigma^2\right) \, d\eta(\boldsymbol{R})
\end{aligned}$$

By definition of our smoothing distribution, we have $\mathrm{vec}(\boldsymbol{Z}) \sim \mathcal{N}(\mathrm{vec}(\boldsymbol{V}) \mid \mathrm{vec}(\boldsymbol{V}), \sigma^2 \mathbf{I}_{D \cdot N})$ for $\boldsymbol{Z} \sim \mu_{\boldsymbol{V}}$. Because our worst-case classifier for $SE(D)$ performs a linear transformation of $\mathrm{vec}(\boldsymbol{Z})$, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{V}}} \left[\mathbb{1} \left[\frac{\beta_{\boldsymbol{X}'}^{SE(D)}(\boldsymbol{Z})}{\beta_{\boldsymbol{X}}^{SE(D)}(\boldsymbol{Z})} \leq \kappa \right] \right] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{V}}} \left[\mathbb{1} \left[\frac{\int_{SO(D)} \exp\left(\left\langle \tilde{\boldsymbol{X}}'^T \boldsymbol{Z}, \boldsymbol{R} \right\rangle / \, \sigma^2\right) \, d\eta(\boldsymbol{R})}{\int_{SO(D)} \exp\left(\left\langle \tilde{\boldsymbol{X}}^T \boldsymbol{Z}, \boldsymbol{R} \right\rangle / \, \sigma^2\right) \, d\eta(\boldsymbol{R})} \leq \kappa \right] \right] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{q} \sim \mathcal{N}(\boldsymbol{m}_{\boldsymbol{V}}, \boldsymbol{\Sigma})} \left[\mathbb{1} \left[\frac{\int_{SO(D)} \exp\left(\left\langle \mathrm{vec}^{-1}(\boldsymbol{q}_{1:9}), \boldsymbol{R} \right\rangle / \, \sigma^2\right) \, d\eta(\boldsymbol{R})}{\int_{SO(D)} \exp\left(\left\langle \mathrm{vec}^{-1}(\boldsymbol{q}_{10:18}), \boldsymbol{R} \right\rangle / \, \sigma^2\right) \, d\eta(\boldsymbol{R})} \leq \kappa \right] \right],
\end{aligned}$$

with $\boldsymbol{m}_{\boldsymbol{V}} = \boldsymbol{W} \mathrm{vec}(\boldsymbol{V})$ and $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T$, where

$$\boldsymbol{W} = \begin{bmatrix} \tilde{\boldsymbol{X}}' & \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{X}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{X}}' & \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{X}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{X}}' & \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{X}} \end{bmatrix}^T.$$

Similarly, we have for $SO(3)$ (after centering $\boldsymbol{V}$), that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{Z} \sim \mu_{\tilde{\boldsymbol{V}}}} \left[\mathbb{1} \left[\frac{\beta_{\tilde{\boldsymbol{X}}'}^{SO(D)}(\boldsymbol{Z})}{\beta_{\tilde{\boldsymbol{X}}}^{SO(D)}(\boldsymbol{Z})} \leq \kappa \right] \right] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{q} \sim \mathcal{N}(\boldsymbol{m}_{\tilde{\boldsymbol{V}}}, \boldsymbol{\Sigma})} \left[\mathbb{1} \left[\frac{\int_{SO(D)} \exp\left(\left\langle \mathrm{vec}^{-1}(\boldsymbol{q}_{1:9}), \boldsymbol{R} \right\rangle / \, \sigma^2\right) \, d\eta(\boldsymbol{R})}{\int_{SO(D)} \exp\left(\left\langle \mathrm{vec}^{-1}(\boldsymbol{q}_{10:18}), \boldsymbol{R} \right\rangle / \, \sigma^2\right) \, d\eta(\boldsymbol{R})} \leq \kappa \right] \right],
\end{aligned}$$

with $\boldsymbol{m}_{\tilde{\boldsymbol{V}}} = \boldsymbol{W} \mathrm{vec}(\tilde{\boldsymbol{V}})$ and $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T$. Finally, due to the fact (see Lemma 7) that calculating an inner product between a centered and an uncentered vector (here: columns of $\tilde{\boldsymbol{X}}$ and $\boldsymbol{V}$) is equivalent to centering both vectors before calculating the inner product (here: columns of $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{V}}$), we have $\boldsymbol{m}_{\boldsymbol{V}} = \boldsymbol{m}_{\tilde{\boldsymbol{V}}}$. Thus, both expectations are equal, which concludes our proof.

F.5 Monte Carlo certification

All of the discussed tight certificates, save for the one for translation invariance, are of the form

$$\begin{aligned}\min_{h\in\mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z})=y^*\right] &= \Pr_{V\sim\upsilon^{(1)}}\left[\rho(V)\leq\kappa\right]\\ \text{with } \kappa\in\mathbb{R} \text{ such that } &\Pr_{V\sim\upsilon^{(2)}}\left[\rho(V)\leq\kappa\right]=p_{\boldsymbol{X},y^*}\\ \text{and } p_{\boldsymbol{X},y^*} &:= \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[g(\boldsymbol{Z})=y^*\right],\end{aligned}$$

where $g:\mathbb{R}^{N\times D}\to\mathbb{Y}$ is our base classifier, $\mu_{\boldsymbol{X}},\mu_{\boldsymbol{X}'}$ are the clean and perturbed smoothing distribution, $\upsilon^{(1)}$ and $\upsilon^{(2)}$ are distributions over some set $\mathbb{S}$, and $\rho:\mathbb{S}\to\mathbb{R}$ is some arbitrary scalar-valued function.

As discussed in Section 6.3, we propose to compute a narrow probabilistic lower bound on $\min_{h\in\mathbb{H}_\mathbb{T}}\Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z})=y^*\right]$ by combining three confidence bounds.

Before computing these bounds, we have to inspect the monotonicity of our certificate. Evidently, $\Pr_{V\sim\upsilon^{(1)}}\left[\rho(V)\leq\kappa\right]$ is monotonically decreasing in κ. Furthermore, κ is monotonically decreasing in $p_{\boldsymbol{X},y^*}$: If $p_{\boldsymbol{X},y^*}$ decreases, then a smaller κ is sufficient for fulfilling the constraint $\Pr_{V\sim\upsilon^{(2)}}\left[\rho(V)\leq\kappa\right]=p_{\boldsymbol{X},y^*}$. Thus, we can compute a lower bound on our certificate by

1. lower-bounding $p_{\boldsymbol{X},y^*}$, i.e. $\underline{p_{\boldsymbol{X},y^*}}\leq p_{\boldsymbol{X},y^*}$,
2. lower-bounding the $\kappa\in\mathbb{R}$ that fulfills $\Pr_{V\sim\upsilon^{(2)}}\left[\rho(V)\leq\kappa\right]=\underline{p_{\boldsymbol{X},y^*}}$, i.e. $\underline{\kappa}\leq\kappa$
3. and then lower-bounding $\Pr_{V\sim\upsilon^{(1)}}\left[\rho(V)\leq\underline{\kappa}\right]$.

The random variable $\mathbb{1}\left[g(\boldsymbol{Z})=y^*\right]$ is a Bernoulli random variable. Just like other randomized smoothing methods (e.g. [26]), we can lower-bound $p_{\boldsymbol{X},y^*}$ by evaluating g on N_1 samples from $\mu_{\boldsymbol{X}}$ and computing a binomial proportion confidence bound, such as the the Clopper-Pearson confidence bound [141].

Lower-bounding κ requires computing a lower bound on $F^{-1}(\underline{p_{\boldsymbol{X},y^*}})$, where F^{-1} is the quantile function of $\rho(V)$ with $V\sim\upsilon^{(2)}$. A non-parametric lower confidence bound on this quantile can be constructed by evaluating $\rho(V)$ on N_2 samples from $\upsilon^{(2)}$ and returning the largest order statistic $R^{(n)}$ such that $\left[\rho(V)\leq R^{(n)}\right]\leq\underline{p_{\boldsymbol{X},y^*}}$ holds with high probability (see Section 5.2.1. of [142] and Algorithm 1 below).

The random variable $\left[\rho(V)\leq\underline{\kappa}\right]$ is another Bernoulli random variable and can thus be lower-bounded by evaluating $\rho(V)$ on N_3 samples from $\mu_{\boldsymbol{X}'}$ and computing a binomial proportion confidence bound.

We want all three confidence bounds to simultaneously hold with high probability $1-\alpha$. We ensure this by using Holm-Bonferroni [131] correction to account for the multiple comparisons problem.[4] In our case, this corresponds to computing the first bound with significance α, the second one with significance $\alpha/2$ and the last one with significance $\alpha/3$.

Algorithm 1 summarizes our certification procedure. LOWERCOUNFBOUND refers to the Clopper-Pearson lower confidence bound. BINPVALUE$(n, N_2, \geq, \underline{p_{\boldsymbol{X},y^*}})$ refers to the p-value of a Binomial test with the null-hypothesis that the success probability is greater or equal $\underline{p_{\boldsymbol{X},y^*}}$. Note that only the first confidence bound requires evaluating the base classifier g. For the other two confidence bounds a large number of samples can be evaluated at little computational cost.

[4] Holm-Bonferroni correction has already been used in the context of randomized smoothing, see [91].

Algorithm 1 Monte Carlo certification procedure

function PROBCERTIFY(y^*, g, $\mu_{\boldsymbol{X}}$, $\upsilon^{(1)}$, $\upsilon^{(2)}$, N_1, N_2, N_3, α)

$\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(N_1)} \leftarrow$ SAMPLE($\mu_{\boldsymbol{X}}$, N_1)

$\text{count}_1 \leftarrow \sum_{n=1}^{N_1} \mathbb{1}\left[g(\boldsymbol{Z}^{(n)}) = y^*\right]$

$\underline{p_{\boldsymbol{X},y^*}} \leftarrow$ LOWERCONFBOUND(count_1, N_1, $1-\alpha$) ▷ First bound

$n^* \leftarrow \max\left\{n \mid \text{BINPVALUE}(n, N_2, \geq, \underline{p_{\boldsymbol{X},y^*}}) < \frac{\alpha}{2}\right\}$

$V^{(1)}, \ldots, V^{(N_2)} \leftarrow$ SAMPLE($\upsilon^{(2)}$, N_2)

$R^{(1)}, \ldots, R^{(N_2)} \leftarrow$ SORTASCENDING($\rho(V^{(1)}), \ldots, \rho(V^{(N_2)})$)

$\underline{\kappa} \leftarrow R^{(n^*)}$ ▷ Second bound

$V^{(1)}, \ldots, V^{(N_3)} \leftarrow$ SAMPLE($\upsilon^{(1)}$, N_3)

$\text{count}_2 \leftarrow \sum_{n=1}^{N_3} \mathbb{1}\left[\rho(V) \leq \underline{\kappa}\right]$

return LOWERCONFBOUND(count_2, N_3, $1-\frac{\alpha}{3}$) ▷ Third bound

G Proof of Theorem 5

Next, we prove that the post-processing-based certificate for rotation-invariance ($\mathbb{T} = SO(D)$) is not tight, which we formalized in Section 6.3 as follows:

Theorem 5. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under $\mathbb{T} = SO(D)$ and $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. Assume that perturbed input $\boldsymbol{X}'$ is not obtained via rotation of $\boldsymbol{X}$, i.e. $\nexists \boldsymbol{R} \in SO(D) : \boldsymbol{X}' = \boldsymbol{X}\boldsymbol{R}^T$. Further assume that $p_{\boldsymbol{X},y^*} \in (0,1)$. Then, for all $\boldsymbol{R} \in SO(D)$:*

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] > \Phi\left(\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right) - \frac{1}{\sigma}\left|\left|\boldsymbol{X}'\boldsymbol{R}^T - \boldsymbol{X}\right|\right|_2\right). \tag{8}$$

The right-hand side term is the optimal value of the black-box optimization problem evaluated for perturbed input $\boldsymbol{X}'\boldsymbol{R}^T$, i.e.

$$\min_{h\in\mathbb{H}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'\boldsymbol{R}^T}} [h(\boldsymbol{Z}) = y^*],$$

with $\mathbb{H} = \left\{h : \mathbb{R}^{N\times D} \to \mathbb{Y} \mid \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}} [h(\boldsymbol{Z}) = y^*] \geq p_{\boldsymbol{X},y^*}\right\}$. The left-hand side term is the optimal value of the gray-box optimization problem evaluated for perturbed input $\boldsymbol{X}'$. Note that $\mathbb{H}_{\mathbb{T}}$ is a set of rotation invariant classifiers. In Appendix D we have proven that rotation of the smoothing distribution's mean does not have an effect on the prediction probabilities of such classifiers, i.e.

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'\boldsymbol{R}^T}} [h(\boldsymbol{Z}) = y^*].$$

Thus, we can prove Theorem 5 by proving the following, more general statement:

Lemma 8. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under $\mathbb{T} = SO(D)$ and $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. Further assume that $\boldsymbol{X}' \neq \boldsymbol{X}$ and that $p_{\boldsymbol{X},y^*} \in (0,1)$. Then:*

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] > \min_{h\in\mathbb{H}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] \tag{42}$$

We can do so by showing that no optimal solution to the r.h.s. black-box problem from Eq. (42) is a feasible solution to the l.h.s. gray-box problem. More formally, let $\mathbb{H}^* \subseteq \mathbb{H}$ be the set of all classifiers minimizing the r.h.s. black box problem. We have $\mathbb{H}^* \subseteq \mathbb{H}$ and $\mathbb{H}_{\mathbb{T}} \subseteq \mathbb{H}$. By showing that $\mathbb{H}^* \cap \mathbb{H}_{\mathbb{T}} = \emptyset$, we prove that $\mathbb{H}_{\mathbb{T}} \subseteq \mathbb{H} \setminus \mathbb{H}^*$, i.e. all rotation invariant classifiers from $\mathbb{H}_{\mathbb{T}}$ yield strictly larger optimal values.

Recall that *an* optimal solution to the r.h.s. black-box problem in Eq. (42) is given by the Neyman-Pearson lemma [107]. Later work on hypothesis testing has shown that the Neyman-Pearson lemma is not only a sufficient, but a necessary condition for optimality[143]: Every most powerful test must fulfill the likelihood ratio inequalities, save for sets of zero measure. For our formulation of the Neyman-Pearson lemma (see Lemma 5), this means that any classifier h that is an optimal solution to the r.h.s. black-box problem from Eq. (42) must fulfill

$$h(\boldsymbol{Z}) \in \begin{cases} \{y^*\} & \text{if } \frac{\mu_{\boldsymbol{X}'}(\boldsymbol{Z})}{\mu_{\boldsymbol{X}}(\boldsymbol{Z})} < \kappa \\ \mathbb{Y} \setminus \{y^*\} & \text{if } \frac{\mu_{\boldsymbol{X}'}(\boldsymbol{Z})}{\mu_{\boldsymbol{X}}(\boldsymbol{Z})} > \kappa \end{cases}$$

for some $\kappa \in \mathbb{R}$, save for sets of zero measure. Cohen et al. [26] have shown that if $\boldsymbol{X}' \neq \boldsymbol{X}$ and $p \in (0,1)$ and $\mu_{\boldsymbol{X}}, \mu_{\boldsymbol{X}'}$ are isotropic Gaussian distributions, then these likelihood ratio inequalities correspond to a linear decision boundary, i.e. any optimal solution most fulfill

$$h(\boldsymbol{Z}) \in \begin{cases} \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z}\rangle_{\mathrm{F}} + b < 0 \\ \mathbb{Y} \setminus \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z}\rangle_{\mathrm{F}} + b > 0 \end{cases} \tag{43}$$

for some $b \in \mathbb{R}$ and $\boldsymbol{W} \in \mathbb{R}^{N\times D}$ with $\boldsymbol{w} \neq \boldsymbol{0}_{N,D}$, save for sets of zero measure. Here, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_{\mathrm{F}} = \mathrm{vec}(\boldsymbol{W})^T\mathrm{vec}(\boldsymbol{Z})$ is the Frobenius inner product.

Therefore, we can prove Lemma 8 by showing that classifiers complying with Eq. (43) (save for sets of zero measure) are not rotation invariant, i.e. they are not feasible solutions to the gray-box optimization problem.

We first do this for the case that D is even and then for the case that D is odd.

Lemma 9. *Assume that $D = 2k$ for some $k \in \mathbb{N}$. Let $h : \mathbb{R}^{N \times D} \to \mathbb{Y}$ be a classifier that fulfills*

$$h(\boldsymbol{Z}) \in \begin{cases} \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b < 0 \\ \mathbb{Y} \setminus \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b > 0 \end{cases} \tag{44}$$

for some $b \in \mathbb{R}$ and $\boldsymbol{w} \in \mathbb{R}^{N \times D}$ with $\boldsymbol{w} \neq \mathbf{0}_{N \cdot D}$, save for sets of zero measure. Then, there is an input $\hat{\boldsymbol{Z}}$ and a rotation matrix $\boldsymbol{R} \in SO(D)$ such that $h(\hat{\boldsymbol{Z}}) \neq h(\hat{\boldsymbol{Z}}\boldsymbol{R}^T)$.

Proof. **Case 1.** Assume that $b \geq 0$.

Consider the set $\mathbb{S} = \{\boldsymbol{Z} \mid \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} < b\}$. By construction, we have $\forall \boldsymbol{Z} \in \mathbb{S} : \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b < 0$ and $\forall \boldsymbol{Z} \in \mathbb{S} : \langle \boldsymbol{W}, -\boldsymbol{Z} \rangle_{\mathrm{F}} + b > 0$.

There must be at least one $\hat{\boldsymbol{Z}} \in \mathbb{S}$ with $h(\hat{\boldsymbol{Z}}) = y^*$ and $h(\hat{\boldsymbol{Z}}) \neq y^*$. Otherwise, all points $\boldsymbol{Z} \in \mathbb{S}$ would need to fulfill $h(\boldsymbol{Z}) \neq y^*$ or $h(-\boldsymbol{Z}) = y^*$, in which case we would have sets of non-zero measure violating the likelihood ratio inequalities from Eq. (44).

Note that $-\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}(-\mathbf{I}_N)^T$ and $-\mathbf{I}_N \in SO(D)$, because $-\mathbf{I}_N$ is orthonormal and $\det(-\mathbf{I}_N) = 1$, due to D being even. Thus, we have found an input $\hat{\boldsymbol{Z}}$ and a rotation matrix $\boldsymbol{R} \in SO(D)$ such that $h(\boldsymbol{Z}) \neq h(\boldsymbol{Z}\boldsymbol{R}^T)$.

Case 2. Assume that $b < 0$. This case follows analogously by constructing the set $\mathbb{S} = \{\boldsymbol{Z} \mid \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} > b\}$ and considering a point $\hat{\boldsymbol{Z}} \in \mathbb{S}$ with $h(\hat{\boldsymbol{Z}}) \neq y^*$ and $h(-\hat{\boldsymbol{Z}}) = y^*$ □

Now, we consider the case that D is odd. Here, $-\mathbf{I}_N \notin SO(D)$, because $\det(-\mathbf{I}_N) = -1$. Therefore, we will need to use slightly more complicated constructions.

Lemma 10. *Assume that $D = 2k - 1$ for some $k \in \mathbb{N}$. Let $h : \mathbb{R}^{N \times D} \to \mathbb{Y}$ be a classifier that fulfills*

$$h(\boldsymbol{Z}) \in \begin{cases} \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b < 0 \\ \mathbb{Y} \setminus \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b > 0 \end{cases} \tag{45}$$

for some $b \in \mathbb{R}$ and $\boldsymbol{w} \in \mathbb{R}^{N \times D}$ with $\boldsymbol{w} \neq \mathbf{0}_{N \cdot D}$, save for sets of zero measure. Then, there is an input $\hat{\boldsymbol{Z}}$ and a rotation matrix $\boldsymbol{R} \in SO(D)$ such that $h(\hat{\boldsymbol{Z}}) \neq h(\hat{\boldsymbol{Z}}\boldsymbol{R}^T)$.

Proof. In the following, let $\boldsymbol{A} = \begin{bmatrix} 1 & \mathbf{0}_D^T \\ \mathbf{0}_{D-1} & -\mathbf{I}_{D-1} \end{bmatrix}$. We have $\boldsymbol{A} \in SO(D)$, because $\boldsymbol{A}$ is orthonormal and $\det(\boldsymbol{A}) = 1$, due to D being odd.

Case 1. Assume that $b \geq 0$.

We know that $\boldsymbol{W} \neq \mathbf{0}_{N,D}$. Without loss of generality, assume that $\boldsymbol{W}_{:,2:} \neq \mathbf{0}_{N,(D-1)}$, i.e. at least one of the the last $D - 1$ columns is non-zero.

Consider the set $\mathbb{S} = \left\{\boldsymbol{Z} \mid \boldsymbol{W}_{:,1}^T \boldsymbol{Z}_{:,1} \in [0, 1] \wedge \langle \boldsymbol{W}_{:,2:}, \boldsymbol{Z}_{:,2:} \rangle_{\mathrm{F}} < b - 1\right\}$. By construction, we have $\forall \boldsymbol{Z} \in \mathbb{S} : \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b < 0$ and $\forall \boldsymbol{Z} \in \mathbb{S} : \left\langle \boldsymbol{W}, \boldsymbol{Z}\boldsymbol{A}^T \right\rangle_{\mathrm{F}} + b > 0$.

There must be at least one $\hat{\boldsymbol{Z}} \in \mathbb{S}$ with $h(\hat{\boldsymbol{Z}}) = y^*$ and $h(\hat{\boldsymbol{Z}}\boldsymbol{A}^T) \neq y^*$. Otherwise, all points $\boldsymbol{Z} \in \mathbb{S}$ would need to fulfill $h(\boldsymbol{Z}) \neq y^*$ or $h(\boldsymbol{Z}\boldsymbol{A}^T) = y^*$, in which case we would have sets of non-zero measure violating Eq. (44).

Case 2. Assume that $b < 0$. This case follows analogously by constructing the set $\mathbb{S} = \left\{\boldsymbol{Z} \mid \boldsymbol{W}_{:,1}^T \boldsymbol{Z}_{:,1} \in [-1, 0] \wedge \langle \boldsymbol{W}_{:,2:}, \boldsymbol{Z}_{:,2:} \rangle_{\mathrm{F}} > b + 1\right\}$ and considering a point $\hat{\boldsymbol{Z}} \in \mathbb{S}$ with $h(\hat{\boldsymbol{Z}}) \neq y^*$ and $h(\boldsymbol{A}\hat{\boldsymbol{Z}}) = y^*$. □

Lemmas 9 and 10 combined prove Lemma 8, which in turn proves Theorem 5.

H Multi-class certificates

In our discussions in Sections 3.1, 5 and 6, we have only considered binary certificates[5]. That is, prediction y^* of smoothed classifier is certifiably robust to a perturbed input $\boldsymbol{X}'$ if the probability of base classifier g predicting y^* under perturbed smoothing distribution $\mu_{\boldsymbol{X}'}$ is greater than the probability of all other classes combined. That is, $p_{\boldsymbol{X}',y^*} > \frac{1}{2}$. We have then, for different invariances, computed lower bounds $\underline{p_{\boldsymbol{X}',y^*}} \leq p_{\boldsymbol{X}',y^*}$. If $\underline{p_{\boldsymbol{X}',y^*}} > \frac{1}{2}$, then $p_{\boldsymbol{X}',y^*} > \frac{1}{2}$ and the prediction is certifiably robust.

Alternatively, one can certify robustness by showing that y^* remains more likely than the second most likely class under the perturbed smoothing distribution [26], i.e. $p_{\boldsymbol{X}',y^*} > \max_{y \neq y^*} p_{\boldsymbol{X}',y}$. To this end, one can compute a lower bound $\underline{p_{\boldsymbol{X}',y^*}} \leq p_{\boldsymbol{X}',y^*}$ and an upper-bound $\overline{\max_{y \neq y^*} p_{\boldsymbol{X}',y}} \geq \max_{y \neq y^*} p_{\boldsymbol{X}',y}$. If $\underline{p_{\boldsymbol{X}',y^*}} > \overline{\max_{y \neq y^*} p_{\boldsymbol{X}',y}}$, then robustness is guaranteed. Due to the monotonicity of randomized smoothing certificates w.r.t. to the prediction probabilities, this can be achieved by showing that $\underline{p_{\boldsymbol{X}',y^*}} > \overline{p_{\boldsymbol{X}',y'}}$, where $y' = \operatorname{argmax}_{y \neq y^*} p_{\boldsymbol{X},y}$ is the second most likely class under the clean smoothing distribution. This second most likely class y' can be found via a binomial test, see [26, 144].

The lower bound $\underline{p_{\boldsymbol{X}',y^*}}$ can be obtained using the same formulae discussed throughout the main text (e.g. $\underline{p_{\boldsymbol{X}',y^*}} = \Phi\left(\Phi^{-1}(p_{\boldsymbol{X},y'}) - \frac{||\Delta||_2}{\sigma}\right)$ for black-box randomized smoothing). The upper bound $\overline{p_{\boldsymbol{X}',y'}}$ can be found via the Neyman-Pearson upper bound discussed in Appendix F.2:

Lemma 6 (Neyman-Pearson upper bound)**.** *Let $\mu_{\boldsymbol{X}'}$, $\mu_{\boldsymbol{X}}$, be two continuous distributions over a measurable set $\mathbb{A}$ such that, for all $\kappa \in \mathbb{R}_+$, the set $\left\{z \mid \frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} = \kappa\right\}$ has measure zero. Consider an arbitrary label set $\mathbb{Y}$, a specific class label $y \in \mathbb{Y}$ and scalar $p \in [0,1]$. Then*

$$\left(\max_{h:\mathbb{A}\to\mathbb{Y}} \Pr_{z\sim\mu_{\boldsymbol{X}'}}[h(z)=y] \text{ s.t. } \Pr_{z\sim\mu_{\boldsymbol{X}}}[h(z)=y] \leq p\right) = \mathop{\mathbf{E}}_{z\sim\mu_{\boldsymbol{X}'}}[h^*(z)]$$

$$\text{with } h^*(z) = \mathbb{1}\left[\frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} \geq \kappa\right] \text{ and } \kappa \in \mathbb{R}_+ \text{ such that } \mathop{\mathbf{E}}_{z\sim\mu_{\boldsymbol{X}}}[h^*(z)] = p.$$

Note that it is identical to the lower bound, save for replacing "$\leq$" with "$\geq$" and vice-versa.

The resulting black-box upper bound is $\overline{p_{\boldsymbol{X}',y'}} = \Phi\left(\Phi^{-1}(p_{\boldsymbol{X},y'}) + \frac{||\Delta||_2}{\sigma}\right)$. It is identical to the black-box lower bound, save for replacing "$-$" with "$+$". Substituting both bounds into $\underline{p_{\boldsymbol{X}',y^*}} > \overline{\max_{y \neq y^*} p_{\boldsymbol{X}',y}}$ and applying the inverse-normal CDF Φ^{-1} to both sides shows that the classifier is certifiably robust if $||\boldsymbol{\Delta}|| \leq \frac{\sigma}{2}\left(\Phi^{-1}(p_{\boldsymbol{X}',y^*}) - \Phi^{-1}(p_{\boldsymbol{X}',y'})\right)$ [26].

Evidently, the orbit-based versions of the multi-class certificates are identical to their binary counterparts, except that the radius of the underlying black-box certificate changes.

For the tight gray-box approach, upper bounds can be derived by going through the same derivations as in Appendix F, but using the Neyman-Pearson upper bounds from Lemma 6, i.e. replacing "$\leq$" with "$\geq$" and vice-versa. (see Lemma 6). In the case of translation invariance, this results in upper bound $\Phi\left(\Phi^{-1}(p_{\boldsymbol{X},y'}) + \frac{||\Delta - \mathbf{1}_N\overline{\Delta}||_2}{\sigma}\right)$). In the cases involving rotation invariance, this results in bounds of the form

$$\begin{aligned} &\Pr_{V\sim\upsilon^{(1)}}[\rho(V) \geq \kappa] \\ \text{with } \kappa \in \mathbb{R} \text{ such that } &\Pr_{V\sim\upsilon^{(2)}}[\rho(V) \geq \kappa] = p_{\boldsymbol{X},y'} \\ \text{and } p_{\boldsymbol{X},y^*} &:= \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[g(\boldsymbol{Z}) = y'], \end{aligned}$$

i.e. the same certificates, but with "$\geq \kappa$" instead of "$\leq \kappa$".

Our Monte Carlo certification procedure proposed in Appendix F.5 can also easily be adapted to computing an upper bound by replacing "$\leq$" with "$\geq$" and vice-versa.:

[5] Not to be confused with certificates that are limited to binary classifiers.

Algorithm 2 Monte Carlo certification procedure (Upper bound)

function PROBCERTIFYUPPER(y^*, g, $\mu_{\boldsymbol{X}}$, $\upsilon^{(1)}$, $\upsilon^{(2)}$, N_1, N_2, N_3, α)
  $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(N_1)} \leftarrow$ SAMPLE($\mu_{\boldsymbol{X}}$, N_1)
  $\text{count}_1 \leftarrow \sum_{n=1}^{N_1} \mathbb{1}\left[g(\boldsymbol{Z}^{(n)}) = y^*\right]$
  $\overline{p_{\boldsymbol{X},y'}} \leftarrow$ UPPERCONFBOUND(count_1, N_1, $1-\alpha$) ▷ First bound
  $n^* \leftarrow \max\left\{n \mid \text{BINPVALUE}(n, N_2, \leq, 1-\overline{p_{\boldsymbol{X},y'}}) < \frac{\alpha}{2}\right\}$
  $V^{(1)}, \ldots, V^{(N_2)} \leftarrow$ SAMPLE($\upsilon^{(2)}$, N_2)
  $R^{(1)}, \ldots, R^{(N_2)} \leftarrow$ SORTASCENDING($\rho(V^{(1)}), \ldots, \rho(V^{(N_2)})$)
  $\underline{\kappa} \leftarrow R^{(n^*)}$ ▷ Second bound
  $V^{(1)}, \ldots, V^{(N_3)} \leftarrow$ SAMPLE($\upsilon^{(1)}$, N_3)
  $\text{count}_2 \leftarrow \sum_{n=1}^{N_3} \mathbb{1}\left[\rho(V) \geq \underline{\kappa}\right]$
  return UPPERCONFBOUND(count_2, N_3, $1-\frac{\alpha}{3}$) ▷ Third bound

Note that we still need to lower-bound threshold κ, because this increases the probability $\Pr_{V\sim\upsilon^{(1)}}\left[\rho(V) \geq t\right]$ and thus leads to more pessimistic, sound multi-class certificates.

Finally note that, because the multi-class certificates are just a combination of two binary certificates, our evaluation of binary invariance-aware certificates in in Section 8 is a good indicator of the performance of invariance-aware multi-class certificates.

I Inverse certificates

In Section 8.1, we evaluate inverse certificates, i.e. compute the smallest prediction probability $p_{\min}$ for which robustness can still be certified, given the remaining certificate parameters such as $||\boldsymbol{\Delta}||_2$ or $||\boldsymbol{X}||_2$.

In the case of the black-box randomized smoothing baseline, $p_{\min}$ can be calculated by solving $\Phi\left(\Phi^{(-1)}(p) - \frac{||\boldsymbol{\Delta}||_2}{\sigma}\right) = \frac{1}{2}$ for p, which results in $p_{\min} = \Phi\left(\frac{||\boldsymbol{\Delta}||_2}{\sigma}\right)$.

Similarly, the gray-box certificates for translation invariance has $p_{\min} = \Phi\left(\frac{||\boldsymbol{\Delta} - \mathbf{1}_N \overline{\boldsymbol{\Delta}}||_2}{\sigma}\right)$, where $\overline{\boldsymbol{\Delta}} \in \mathbb{R}^{1 \times D}$ are the column-wise averages.

The tight gray-box certificates involving rotation invariance do not have closed-form analytic expressions. Instead, they are of the form

$$\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*]] = \Pr_{V \sim \upsilon^{(1)}} [\rho(V) \leq \kappa]$$
$$\text{with } \kappa \in \mathbb{R} \text{ such that } \Pr_{V \sim \upsilon^{(2)}} [\rho(V) \leq \kappa] = p_{\boldsymbol{X}, y^*}.$$

Here, an inverse certificate can be computed by first finding a threshold κ such that $\Pr_{V \sim \upsilon^{(1)}} [\rho(V) \leq \kappa] = \frac{1}{2}$ and then evaluating $\Pr_{V \sim \upsilon^{(2)}} [\rho(V) \leq \kappa]$. That is,

$$p_{\min} = \Pr_{V \sim \upsilon^{(2)}} [\rho(V) \leq \kappa]$$
$$\text{with } \kappa \in \mathbb{R} \text{ such that } \Pr_{V \sim \upsilon^{(1)}} [\rho(V) \leq \kappa] = \frac{1}{2}.$$

To ensure a fair comparison with our baselines, we modify the Monte Carlo certification procedure proposed in Appendix F.5 to compute a probabilistic upper bound on $p_{\min}$ that holds with high probability $1 - \alpha$.

Algorithm 3 Monte Carlo inverse certification procedure

function INVERSEPROBCERTIFY($\upsilon^{(1)}$, $\upsilon^{(2)}$, N_1, N_2, α)
  $n^* \leftarrow \min\{n \mid \text{BINPVALUE}(n, N_1, \leq, 0.5) < \alpha\}$
  $V^{(1)}, \ldots, V^{(N_2)} \leftarrow$ SAMPLE($\upsilon^{(1)}$, N_1)
  $R^{(1)}, \ldots, R^{(N_2)} \leftarrow$ SORTASCENDING($\rho(V^{(1)}), \ldots, \rho(V^{(N_2)})$)
  $\overline{\kappa} \leftarrow R^{(n^*)}$ ▷ First bound
  $V^{(1)}, \ldots, V^{(N_2)} \leftarrow$ SAMPLE($\upsilon^{(2)}$, N_3)
  $\text{count}_2 \leftarrow \sum_{n=1}^{N_2} \mathbb{1}[\rho(V) \leq \overline{\kappa}]$
  return UPPERCONFBOUND(count_2, N_2, $1 - \frac{\alpha}{2}$) ▷ Second bound

Note that, different from Appendix F.5 we compute a probabilistic upper bound on threshold κ by finding the smallest order statistic $R^{(n)}$ such that $\left[\rho(V) \leq R^{(n)}\right] \geq \frac{1}{2}$ holds with high probability.

J Parameter space of the tight certificate for rotation invariance in 2D

Our tight certificates for 2D rotation invariance depend on $||\boldsymbol{X}||_2$, $||\boldsymbol{\Delta}||_2$, as well as parameters $\epsilon_1 = \langle \boldsymbol{X}, \boldsymbol{\Delta} \rangle_{\mathrm{F}}$, $\epsilon_2 = \left\langle \boldsymbol{X}\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T, \boldsymbol{\Delta} \right\rangle_{\mathrm{F}}$, which capture the orientation of the perturbed pointcloud, relative to the clean point cloud.

In this section, we determine the feasible range of ϵ_1 and ϵ_2 and calculate the two pairs of values corresponding to adversarial rotations, which is relevant for our experiments in Section 8.1.

J.1 Feasible parameter range

Let $\hat{\boldsymbol{X}} = \frac{\boldsymbol{X}}{||\boldsymbol{X}||_2}$. Then, parameters ϵ_1 and ϵ_2 can be equivalently stated as follows:

$$\epsilon_1 = \langle \boldsymbol{X}, \boldsymbol{\Delta} \rangle_{\mathrm{F}} = \mathrm{vec}(\boldsymbol{X})^T \mathrm{vec}(\boldsymbol{\Delta}) = \mathrm{vec}(\hat{\boldsymbol{X}})^T \mathrm{vec}(||\boldsymbol{X}||_2 \boldsymbol{\Delta}),$$
$$\epsilon_2 = \left\langle \boldsymbol{X}\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T, \boldsymbol{\Delta} \right\rangle_{\mathrm{F}} = \mathrm{vec}\left(\hat{\boldsymbol{X}}\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T\right)^T \mathrm{vec}(||\boldsymbol{X}||_2 \boldsymbol{\Delta}).$$

Note that $\mathrm{vec}(\hat{\boldsymbol{X}})$ and $\mathrm{vec}\left(\hat{\boldsymbol{X}}\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T\right)$ have norm 1 and are orthogonal to each other. Thus, they are the first two elements of an orthonormal basis of $\mathbb{R}^{2N}$. Parameters ϵ_1 and ϵ_2 are projections of $\mathrm{vec}(||\boldsymbol{X}||_2 \boldsymbol{\Delta})$ onto the first two elements of this basis. Basis changes preserve the norm of vectors. Thus, the values ϵ_1, ϵ_2 must fulfill $\sqrt{\epsilon_1^2 + \epsilon_2^2} \leq ||\boldsymbol{X}||_2 \cdot ||\boldsymbol{\Delta}||_2$.

J.2 Adversarial rotations

Known rotation angle. Consider an arbitrary $\boldsymbol{X} \in \mathbb{R}^{N \times 2}$. Further consider an adversarially perturbed $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{R}(\theta)^T$ that is the result of rotating all rows of $\boldsymbol{X}$ by angle θ. Let $\boldsymbol{\Delta} = \boldsymbol{X}' - \boldsymbol{X}$. Recall that, for any vector $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^D$, $\boldsymbol{a}^T\boldsymbol{b}^T = ||\boldsymbol{a}||_2 ||\boldsymbol{b}||_2 \cos(\angle \boldsymbol{a}\boldsymbol{b})$. Therefore

$$\epsilon_1 = \langle \boldsymbol{X}, \boldsymbol{\Delta} \rangle_{\mathrm{F}} = \sum_{n=1}^{N} \boldsymbol{X}_n^T(\boldsymbol{X}'_n - \boldsymbol{X}_n) = \sum_{n=1}^{N} ||\boldsymbol{X}_n^T||_2^2 \cos(\theta) - ||\boldsymbol{X}_n^T||_2^2 = ||\boldsymbol{X}||_2^2(\cos(\theta) - 1),$$

where the last equality follows from the definition of the Frobenius norm. Similarly, we have

$$\epsilon_2 = ||\boldsymbol{X}||_2^2(\cos(\theta + \pi/2) - \cos(\pi/2)) = -||\boldsymbol{X}||_2^2 \sin(\theta).$$

Unknown rotation angle. Now assume that we know $||\boldsymbol{X}||_2$ and $||\boldsymbol{\Delta}||_2$, but do not know the rotation angle θ. We have

$$\begin{aligned}
||\boldsymbol{\Delta}||_2^2 =& \left\langle \boldsymbol{X}\boldsymbol{R}(\theta)^T - \boldsymbol{X}, \boldsymbol{X}\boldsymbol{R}(\theta)^T - \boldsymbol{X} \right\rangle_{\mathrm{F}} \\
=& ||\boldsymbol{X}\boldsymbol{R}(\theta)^T||_2^2 + ||\boldsymbol{X}||_2^2 - 2\left\langle \boldsymbol{X}, \boldsymbol{X}\boldsymbol{R}(\theta)^T \right\rangle_{\mathrm{F}} \\
=& ||\boldsymbol{X}||_2^2 + ||\boldsymbol{X}||_2^2 - 2||\boldsymbol{X}||_2^2 \cos(\theta) \\
=& 2||\boldsymbol{X}||_2^2(1 - \cos(\theta)).
\end{aligned}$$

Evidently, such an adversarial rotation is only possible if $||\boldsymbol{\Delta}|| \leq 2||\boldsymbol{X}||$, which corresponds to a rotation by angle π. Solving for θ yields

$$\theta = \pm \arccos\left(1 - \frac{||\boldsymbol{\Delta}||_2^2}{2||\boldsymbol{X}||_2^2}\right)$$

Inserting into our formulae for ϵ_1 and ϵ_2 and using that $\sin(\arccos(a)) = \sqrt{1 - a^2}$ shows that

$$\epsilon_1 = -\frac{1}{2}||\boldsymbol{\Delta}||_2^2,$$
$$\epsilon_2 = \pm\frac{1}{2}\sqrt{||\boldsymbol{\Delta}||_2^2\,(4||\boldsymbol{X}||_2^2 - ||\boldsymbol{\Delta}||_2^2)}.$$