Invariance-Aware Randomized Smoothing Certificates

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

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

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

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

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

We opt for a randomized smoothing approach, as it allows us to focus on the interplay between invariances and robustness, rather than the specific means of implementing these invariances. 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 extends the certified region using information about the model's invariances. This post-processing-based certificate serves as a baseline for the provably tight gray-box certificates we derive in subsequent sections.

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

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 [28–32], LiDAR classification [33–36], drug discovery [37–39], particle physics [40–42] and protein folding [43–47].

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,
- a tight certificate for models invariant to translations in arbitrary dimensions,
- tight certificates for models invariant to rotations in 2D and elemental rotations in 3D.

We further demonstrate experimentally that the post-processing-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 [48] for a principled, high-level introduction into the realm of learning with invariances and equivariances.

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

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

Adversarial attacks on spatial data. Adversarial attacks on spatial data either modify [55–63], insert [64, 62, 65] or delete [66–68] 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 [55] more recent work has developed a rich assortment of domain-specific methods, for example to preserve object smoothness [61, 69], leverage critical points [66, 67] or craft physically realizable attacks (e.g. to attack models through LiDAR sensors) [70, 71, 58, 72]. Of particular note are attacks via isometries (e.g. rotations and translations) [73, 74], whose effectiveness motivates the use of invariant models. Note that we certify the robustness of invariant models to arbitrary point modifications attacks – not just isometry attacks (see also "orthogonal research directions" below). Like in other domains, empirical defenses have been proposed [56, 75, 76] and subsequently broken [77–79], motivating the development of black-box [80–82] and white-box [83] 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 [80] or as a special case of the respective certificate [82].

Orthogonal research directions. One related but orthogonal research direction is transformation-specific certification [24, 82–90]. 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 [24, 83–86] overapproximate 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 [87, 91] – randomize the transformation parameters to provide robustness guarantees for arbitrary models via statistical methods. While applying this paradigm to additive transformations (e.g. brightness changes) is relatively straightforward, non-composable transformations (e.g. rotations with interpolation) require a specialized toolset, e.g. statistical bounds on interpolation errors [87] or computation of transformation pre-images [91]. Later work generalized this principle to spatial data [82], vector-field deformations [89] and multiplicative parameters [90]. 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* (global max-pooling [83], message passing [92, 93] and batch normalization [94]). 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 on randomized smoothing

Randomized smoothing is a black-boxcertification technique that can be adapted to various data types, tasks and threat models [95–100, 88, 101–103]. Instead of directly certifying a classifier g, it constructs a smoothed classifier f that returns the most likely prediction under random perturbations of its input. It then certifies the robustness of this smoothed classifier. We present the tight Gaussian smoothing certificate derived by Cohen et al. [27] and its generalization to matrices [80, 82].

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

Let $y^* = f(\boldsymbol{X})$ be a smoothed prediction 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 F. One can lower-bound $p_{\boldsymbol{X}',y^*}$ by finding the *worst-case classifier* from a set of functions $\mathbb{H}$ with $g \in \mathbb{H}$:

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

For $\mathbb{H} = \left\{h : \mathbb{R}^{N \times D} \to \mathbb{Y} \mid \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}}\left[h(\boldsymbol{Z}) = y^*\right] \geq p_{\boldsymbol{X},y^*}\right\}$, the classifiers that are at least as likely as g to predict y^*, the exact solution is given by the Neyman-Pearson lemma [104] . 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.

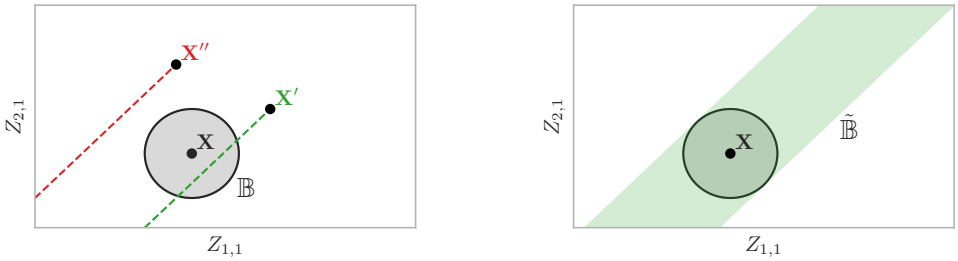
4 Problem setting

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

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

5 Post-processing-based gray-box certificates

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



(a) Pre-processing perspective

(b) Post-processing perspective

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

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

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

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

Importantly, note that Theorem 2 is equivalent to Theorem 1 and not a separate certificate. It is simply a different, implicit way of characterizing the region $\tilde{\mathbb{B}}$. Fig. 1 compares the two perspectives. Applying it to translation invariance shows that $h(\boldsymbol{X}') = y^*$ if $\left|\left|\boldsymbol{\Delta} - \boldsymbol{1}_N \overline{\boldsymbol{\Delta}}\right|\right|_2 < r$, where $\overline{\boldsymbol{X}}, \overline{\boldsymbol{\Delta}} \in \mathbb{R}^{1 \times D}$ are column-wise averages. For rotation invariance, Theorem 2 guarantees robustness if $\left|\left|\boldsymbol{X}'\boldsymbol{R}^T - \boldsymbol{X}\right|\right|_2 < r$, where $\boldsymbol{R}$ is an optimal rotation matrix defined by the singular value decomposition of $\boldsymbol{X}^T\boldsymbol{X}'$ [105, 106]. In Appendix C, we provide proves and discuss other invariances.

5.1 Application to randomized smoothing certificates

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

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

In other words: Isotropic Gaussian smoothing preserves invariance to any group of transformations composed of permutation and Euclidean isometries, i.e. rotation, reflection and translation. Theorem 3 follows from the fact that we smooth by adding zero-mean isotropic noise, meaning that applying an isometry (e.g. rotation) and then randomizing $\boldsymbol{X}$ is equivalent to first randomizing $\boldsymbol{X}$ and then applying the isometry – which has no effect on an isometry-invariant classifier. The formal proof can be found in Appendix C. We may thus instantiate Theorems 1 and 2 with $h = f$ and $r = \sigma\Phi^{-1}(p_{\boldsymbol{X},y^*})$.

6 Tight gray-box certificates

Now that we have derived the first gray-box certificates for invariant models, one may naturally wonder about their optimality. We answer this by deriving tight gray-box certificates, i.e. the best certificates that can be obtained for prediction y^* using only the invariances of base classifier g under

group $\mathbb{T}$ and its prediction probability $p_{\boldsymbol{X},y^*}$ under clean smoothing distribution $\mu_{\boldsymbol{X}}$. To do so, we find the worst-case invariant classifier, i.e. solve $\min_{h\in\mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}[h(\boldsymbol{Z})=y^*]$ with

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

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

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

6.1 Certification methodology

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

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 D.3, we prove that canonical maps let us discard the invariance constraints:

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

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

The unconstrained r.h.s. optimization problem in Theorem 4 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 (which, by Definition 3, is equivalent to a distribution over equivalence classes). We do so by simplifying the probability integrals in Theorem 4 via carefully chosen substitutions that allows us to express the distribution over equivalence classes as a marginal distribution over some alternative parameterization of $\mathbb{R}^{N\times D}$. We present a more formal discussion of our method in Appendix D.2.

6.2 Translation invariance

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

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

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

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

Substituting into 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 is identical to the post-processing-based certificate for translation invaraince from Section 5, i.e. the post-processing-based certificate is provably tight.

6.3 Rotation invariance in 2D

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

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

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

Proof in Appendix E. Our certification procedure, however, lets us derive the tight, strictly stronger certificates for rotation invariance in 2D (see Appendix D.5). In the following, let $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\mathrm{F}}$ be the Frobenius inner product and $\boldsymbol{R}(\theta) \in SO(2)$ the matrix that rotates counter-clockwise by angle θ.

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

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

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

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

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

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

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

Connection to prior work. In Appendix D.5 we show that the worst-case invariant classifier, i.e. the minimizer of our optimization problem, is a function that predicts y^* if and only if $\frac{\beta(\boldsymbol{X}',\boldsymbol{Z})}{\beta(\boldsymbol{X},\boldsymbol{Z})} \leq t$ with $\beta(\boldsymbol{X},\boldsymbol{Z}) = \int_0^{2\pi} \exp\left(\frac{1}{\sigma^2}\left\langle \boldsymbol{X}, \boldsymbol{Z}\boldsymbol{R}(\theta)^T \right\rangle_{\mathrm{F}}\right)\, d\theta$. Different from black-box randomized smoothing, where the worst-case classifier is a linear model [27], this classifier evaluates Gaussian kernels of $(\boldsymbol{X},\boldsymbol{Z})$ and $(\boldsymbol{X}',\boldsymbol{Z})$ and averages them over all possible rotations. This *group averaging* is a key technique for building invariant models [7, 107–111] Group-averaged kernels have been proposed in [16]. It is fascinating to see them naturally materialize from nothing but an invariance constraint.

6.4 Generalizations

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

7 Limitations and broader impact

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

Broader impact. With the growing prevalence of machine learning in safety-critical and sensitive domains like autonomous driving [112, 113] or healthcare [114, 115], trustworthy models promise to

become increasingly important. Certification is one pillar of trustworthiness, alongside concepts like fairness [116, 117] or differential privacy [118, 119]. Unique to our work is that we certify models with spatial invariances. These invariances naturally occur in the physical sciences, including those with a direct societal impact like pharmacology and biochemistry. Our work can be seen as a first step towards provable trustworthiness for tasks like drug discovery [37–39] or protein folding [43–47].

8 Experimental evaluation

We know that the post-processing-based certificate for translation invariance is tight and certifies robustness for an infinitely larger volume than black-box randomized smoothing (see Fig. 1). We therefore focus our experiments on the certificates for rotation invariance. Recall that the post-processed certificate guarantees robustness for the set $\left\{\boldsymbol{Z}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(D), ||\boldsymbol{Z} - \boldsymbol{X}||_2 \leq r\right\}$ with $r = \sigma\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right)$. In other words: It certifies robustness for perturbations with rotational components that can be eliminated to bring $\boldsymbol{X}'$ into distance r of $\boldsymbol{X}$. We want to understand whether the tight certificates offer any benefit beyond that, or if the strict inequality in Theorem 6 is due to some negligible ϵ. To this end, we first thoroughly examine the four-dimensional parameter space of the tight certificate for 2D rotation invariance, before applying our certificates to rotation invariant point cloud classifiers. All parameters and experimental details are specified in Appendix B. A reference implementation is made available with the supplementary material. We use 10000 samples per confidence bound and set $\alpha = 0.001$, i.e. all certificates hold with 99.9% probability.

8.1 Tight certificate parameter space

Recall that the tight certificate for rotation invariance depends on $||\boldsymbol{X}||_2$, $||\boldsymbol{\Delta}||_2$ and parameters ϵ_1 and ϵ_2, which capture the orientation of the perturbed point cloud and fulfill $\sqrt{\epsilon_1^2 + \epsilon_2^2} \leq ||\boldsymbol{X}||_2 \cdot ||\boldsymbol{\Delta}||_2$. To avoid clutter, we define $\tilde{\epsilon}_k := \epsilon_k \,/\, (||\boldsymbol{X}||_2 \cdot ||\boldsymbol{\Delta}||_2)$. As our metric for this section, we report $p_{\min}$, the smallest probability $p_{\boldsymbol{X},y^*}$ for which a prediction can still be certified [1]

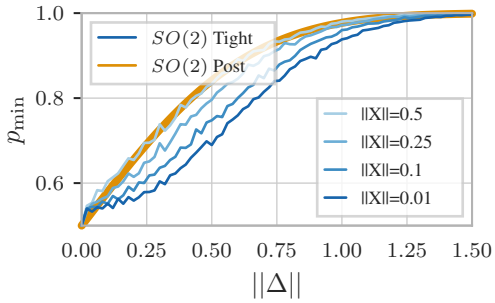


Figure 2: Comparison of tight and post-processing-based certificates applied to adversarial scaling for $\sigma = 0.5$ and varying $||\boldsymbol{\Delta}||$ and $||\boldsymbol{X}||$ (smaller $p_{\min}$ is better). As $||\boldsymbol{X}||$ increases, the difference between the certificates shrinks.

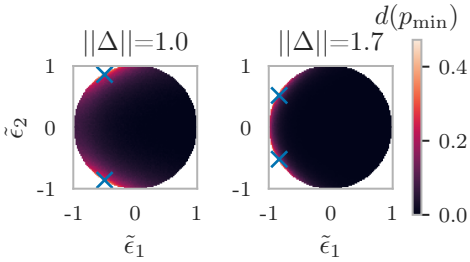


Figure 3: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for $\sigma = 0.5$, $||\boldsymbol{X}|| = 1.0$. Crosses correspond to adversarial rotations. Large $d(p_{\min})$ can be observed near adversarial rotations.

Adversarial scaling. First, we assume that $\boldsymbol{X}' = (1 + \beta)\boldsymbol{X}$, i.e. the input is adversarially scaled. In this case, we have $\tilde{\epsilon}_1 = 1$ and $\tilde{\epsilon}_2 = 0$. We then vary $||\boldsymbol{\Delta}||_2$ and $||\boldsymbol{X}||_2$ and evaluate our certificates. Note that such attacks have no rotational component, i.e. the post-processing-based certificate is identical to the black-box one. Fig. 2 shows that, even in the absence of rotations, the tight certificate can yield significantly stronger guarantees. For $\sigma = 0.5$ and $||\boldsymbol{X}||_2 = 0.01$, post-processing can only certify robustness for a prediction with $p_{\boldsymbol{X},y^*} = 0.8$ if $||\boldsymbol{\Delta}||_2 \leq 0.4$. The tight certificate can certify robustness up to $||\boldsymbol{\Delta}||_2 = 0.73$. However, the gap shrinks, as the norm of the clean data increases.

Effect of data norm. We would like to see if this is a pervasive pattern. To this end, we fix $||\boldsymbol{\Delta}||_2$, gradually increase $||\boldsymbol{X}||_2$, and evaluate the tight certificate for $\tilde{\epsilon}_1$, $\tilde{\epsilon}_2$ on a 100×100 rasterization of $[0, 1] \times [0, 1]$. We then measure the difference $d(p_{\min})$ to the black-box certificate (we will discuss

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

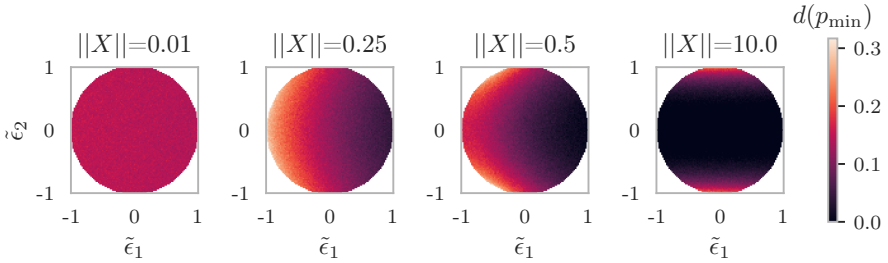


Figure 4: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for $\sigma = 0.5$, $||\boldsymbol{\Delta}|| = 0.5$ under varying $||\boldsymbol{X}||$, $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$. As $||\boldsymbol{X}||$ increases, the regions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

how these results relate to the post-processing-based one shortly). Fig. 4 shows that for small $||\boldsymbol{X}||_2$, the tight certificate outperforms the black-box certificate for arbitrary perturbations of norm $||\boldsymbol{\Delta}||_2$. But, as $||\boldsymbol{X}||_2$ increases, the regions where it outperforms the black-box one shrink.

Rotational components. Simple algebra (see Appendix H) shows that for any combination of $||\boldsymbol{X}||_2$ and $||\boldsymbol{\Delta}||_2$, adversarial rotations, i.e. $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{R}^T$ such that $||\boldsymbol{X}' - \boldsymbol{X}|| = ||\boldsymbol{\Delta}||$, correspond to two specific points with $\sqrt{\tilde{\epsilon}_1^2 + \tilde{\epsilon}_2^2} = 1$ in the certificate's parameter space. In Fig. 3 we fix $\sigma = 0.5$ and $||\boldsymbol{X}||_2 = 1$, vary $||\boldsymbol{\Delta}||$ and highlight these two points. We observe that the tight certificate only outperforms the black-box certificate for values of $\tilde{\epsilon_1}, \tilde{\epsilon_2}$ that are close to adversarial rotations. However, robustness to such attacks can also be readily certified via post-processing. Combined with our previous observations, we expect the tight and the post-processing-based certificate to perform similarly well, assuming that the smoothing standard deviation σ is small relative to $||\boldsymbol{X}||_2$.

8.2 Application to point cloud classification

To verify whether our observations hold in practice, we apply our certificates to point cloud classification. We consider two datasets: 3D point cloud maps of ModelNet40 [120], which consists of CAD models from 40 different categories, and 2D point cloud maps of MNIST [121]. We apply the same pre-processing steps as in [6]. As our base classifiers, we use rotation and translation (i.e. $SE(D)$) invariant versions of two well-established models: PointNet [6] and DGCNN [122]. To implement the invariances, we center the input data, perform principal component analysis, apply the model to all possible poses (see discussion in [123, 124]) and average the output logits (EnsPointNet and EnsDGCNN). In addition, we consider a more refined model [123] that combines canonical poses via a self-attention mechanism (AttnPointNet). Certification is performed on the default test sets.

Practical smoothing parameters. Fig. 5 shows the test set accuracies of randomly smoothed models under varying standard deviations σ. Values of σ that preserve an accuracy above 50% are small,

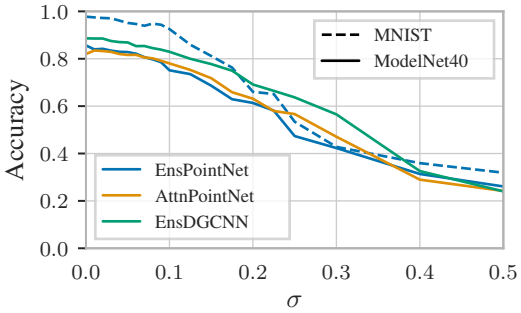


Figure 5: Test set accuracy of smoothed translation and rotation invariant point cloud classifiers on the ModelNet40 and MNIST pointcloud datasets, under varying standard deviation σ.

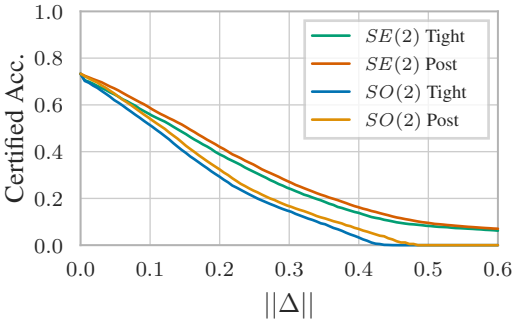
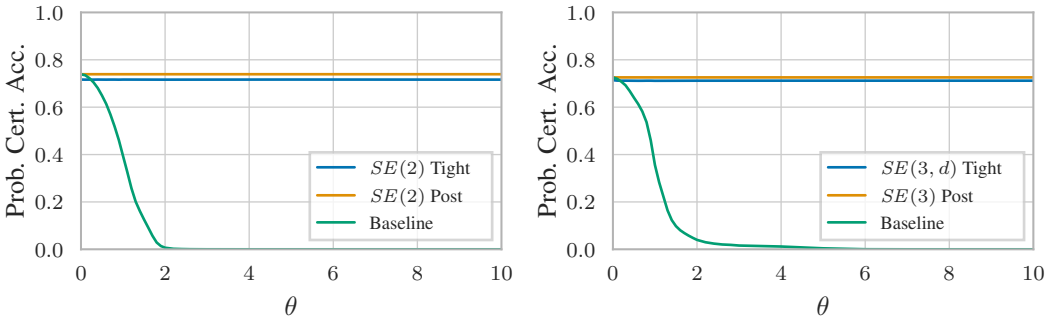


Figure 6: Comparison of certificates for adversarial scaling of MNIST with EnsPointNet and $\sigma = 0.15$. The tight and post-processing-based certificates yield similar certified accuracies.

relative to the average norm of the test sets (10.67 for MNIST, 19.17 for ModelNet40). Going by our previous results, we expect the tight and post-processing based certificates to perform similarly well.

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



(a) EnsPointNet on MNIST

(b) EnsDGCNN on ModelNet40

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

Rotational components. Finally, we study perturbations with rotational components. We fix $||\mathbf{\Delta}||$, randomly sample perturbations of the specified norm and then rotate $\boldsymbol{Z} = X + \mathbf{\Delta}$ by a specified angle $\theta \in [0, 10°]$ (in the case of ModelNet40, around one randomly chosen coordinate axis). For each element of the test set and each θ, we generate 10 such samples. We then compute the percentage of samples $\boldsymbol{X}'$ for which $f(\boldsymbol{X})$ is correct and $f(\boldsymbol{X}') = f(\boldsymbol{X})$ is provably guaranteed ('probabilistic certified accuracy"). Fig. 7 shows results for MNIST and ModelNet40 evaluated with $||\mathbf{\Delta}|| = \sigma = 0.1$ ($SE(3, d)$ refers to elemental rotations). The black-box baseline's probabilistic certified accuracy drops close to 0 for $\theta = 2°$. The gray-box certificates are almost constant in θ, i.e. effectively eliminate any induced rotation. However, the tight certificate did not offer any benefit beyond that. We did not observe a single sample for which only the tight certificate could guarantee robustness. In Appendix A.2.4 we consider arbitrary random rotations and observe that the post-processing-based certificate for arbitrary rotations is more effective than the tight one for elemental rotations.

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

9 Conclusion

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

Checklist

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[18] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International conference on computer aided verification*, pages 97–117. Springer, 2017.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[66] 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.

[67] 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.

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

[69] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, and Nenghai Yu. Shape-invariant 3d adversarial point clouds. *arXiv preprint arXiv:2203.04041*, 2022.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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*, 34:8857–8870, 2021.

[75] 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.

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

[77] 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.

[78] 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.

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

[80] 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.

[81] 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.

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

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

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

[89] 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.

[90] 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. doi: 10.24963/ijcai.2022/467.

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

[92] 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.

[93] 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.

[94] 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.

[95] 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.

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

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

[98] 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.

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

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

[101] 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.

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

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

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

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

[106] 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.

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

[108] 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.

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

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

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

[112] 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.

[113] 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.

[114] 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.

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

[116] 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.

[117] 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.

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

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

[120] 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.

[121] 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.

[122] 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.

[123] 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.

[124] 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.

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

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

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

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

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

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

[131] 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.

[132] 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.

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

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

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

A Additional experiments **20**
A.1 Tight certificate parameter space 21
A.2 Application to point cloud classification 26

B Full experimental setup **32**
B.1 Models, data and training parameters 32
B.2 Randomized smoothing parameters 33
B.3 Computational resources 33
B.4 Third-party assets 33

C Post-processing-based gray-box certificates **34**
C.1 Proof of Theorem 1 34
C.2 Proof of Theorem 2 34
C.3 Proof of Theorem 3 34
C.4 Application of Theorem 2 36

D Tight gray-box certificates **38**
D.1 Neyman-Pearson lemmmapa 38
D.2 Certification methodology 38
D.3 Proof of Theorem 4 40
D.4 Tight certificate for invariance under translation 42
D.5 Tight certificate for invariance under rotation in 2D 46
D.6 Tight certificate for invariance under elemental rotation in 3D 55
D.7 Tight certificate for simultaneous invariance under rotation and translation in 2D . . 59
D.8 Tight certificate for simultaneous invariance under elemental rotation and translation in 3D 64
D.9 Monte Carlo certification 67

E Proof of Theorem 6 **69**

F Multi-class certificates **71**

G Inverse certificates **73**

H Parameter space of the tight certificate for 2D rotation invariance **74**
H.1 Feasible parameter range 74
H.2 Adversarial rotations 74

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 post-processing-based counterparts. However, in practical scenarios, where the smoothing standard deviation is small relative to the norm of the clean data, the post-processing approach offers a very good approximation.

A.1 Tight certificate parameter space

For this section, recall that the tight certificate depends on $||\boldsymbol{X}||_2$, $||\boldsymbol{\Delta}||_2$ and parameters $\epsilon_1 = \langle \boldsymbol{X}, \boldsymbol{\Delta} \rangle_{\mathrm{F}}$, $\epsilon_2 = \left\langle \boldsymbol{X}\boldsymbol{R}(-\pi/2)^T, \boldsymbol{\Delta} \right\rangle_{\mathrm{F}}$, as well as smoothing standard deviation σ. As before, 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 H).

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 G). Smaller $p_{\min}$ 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 post-processing-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 feature any any rotational components. However, as $||\boldsymbol{X}||$ approaches σ, this difference shrinks, i.e. both approaches offer guarantees of similar strength.

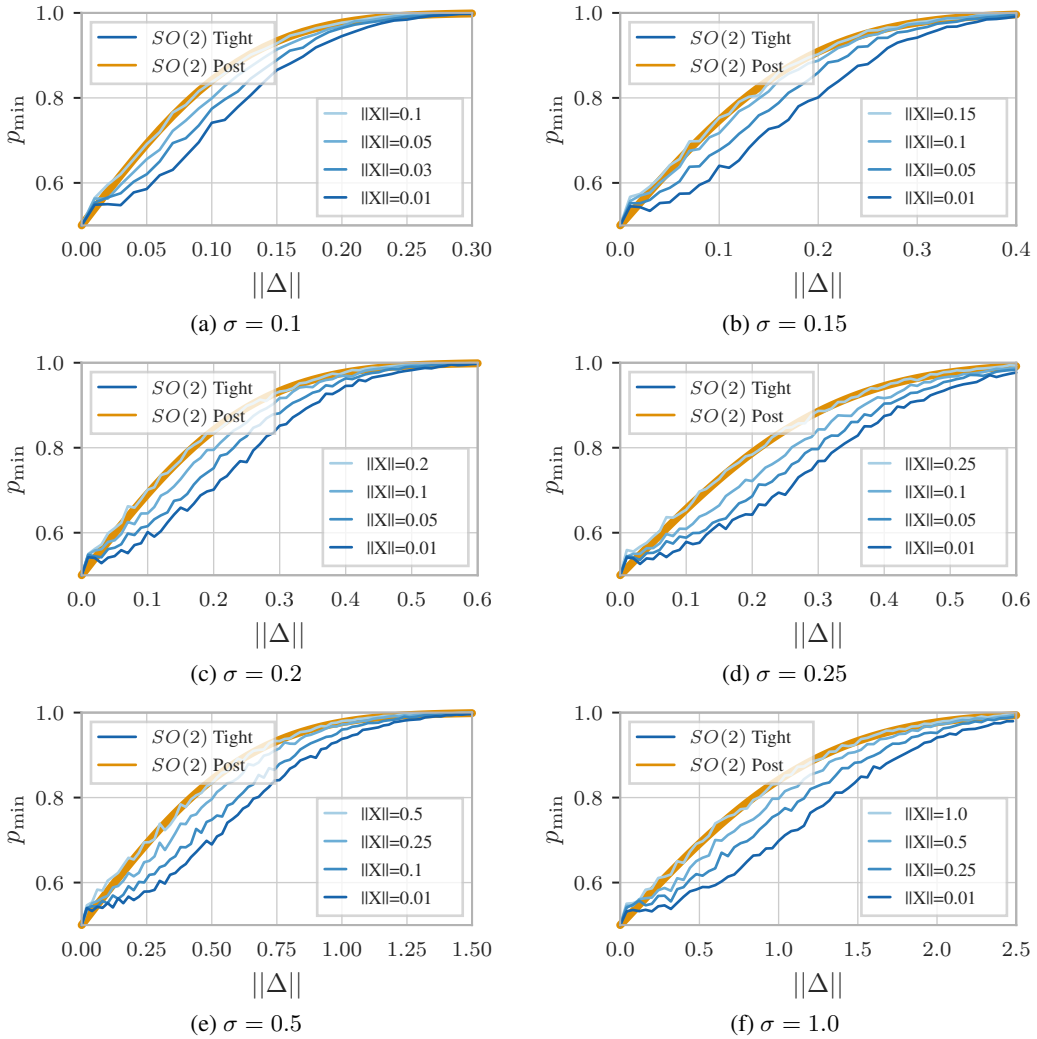


Figure 8: Comparison of tight and post-processing-based certificates for rotation invariance in 2D, applied to adversarial scaling for varying $||\boldsymbol{\Delta}||$, $||\boldsymbol{X}||$ and smoothing standard deviation σ (smaller $p_{\min}$ is better). As $||\boldsymbol{X}||$ increases, the difference 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 D.5, 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 H) 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 $||\boldsymbol{\Delta}||_2 \leq \sigma$, the tight certificate yields lower (i.e. better) $p_{\min}$ for arbitrary perturbations, not just those correponding 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 the case of 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 post-processing-based gray-box certificate for rotation invariance accounts for – or using the pre-processing perspective: eliminates – rotational components, this suggests that the post-processing-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 an arbitrary 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.

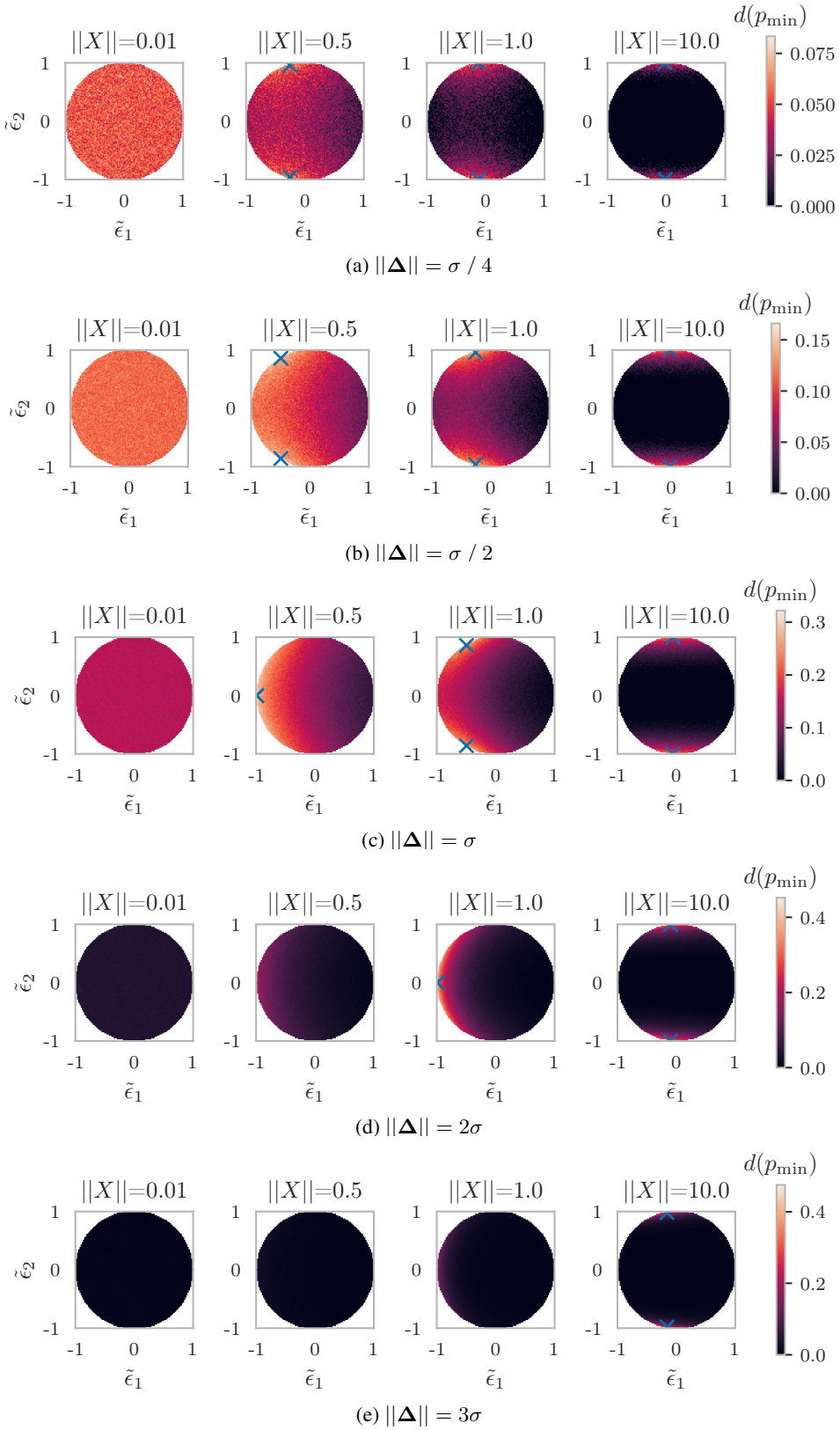
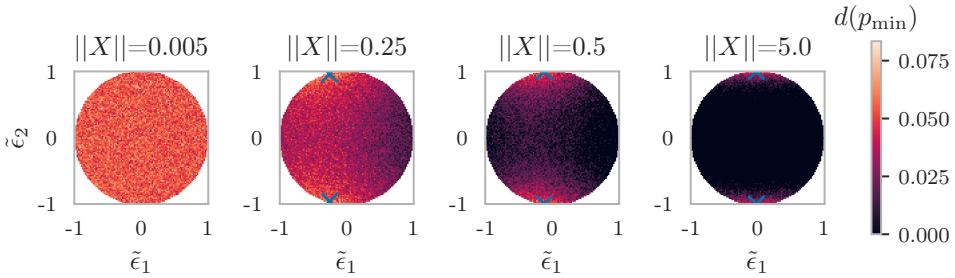
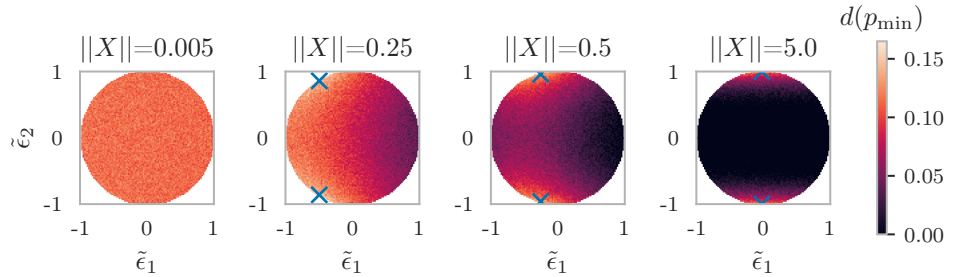


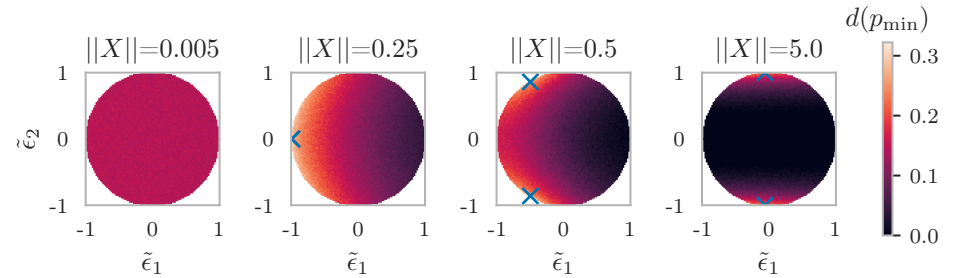
Figure 9: Difference in $p_{\min}$ between the tight certificate for 2D rotation invariance and the black-box baseline for $\sigma = 1$ under varying $||\boldsymbol{X}||$, $||\mathbf{\Delta}||$, $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$. Blue crosses indicate adversarial rotations.



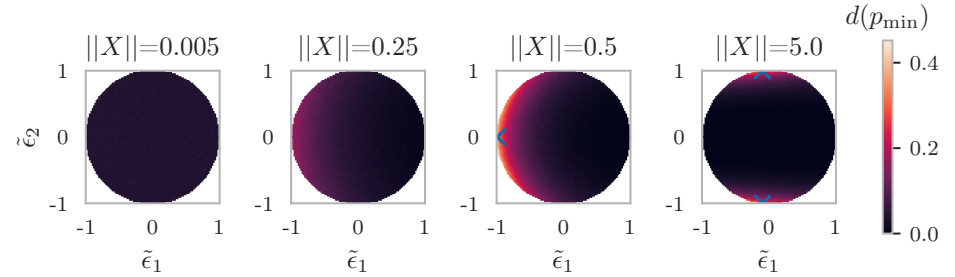
(a) $||\mathbf{\Delta}|| = \sigma \,/\, 4$



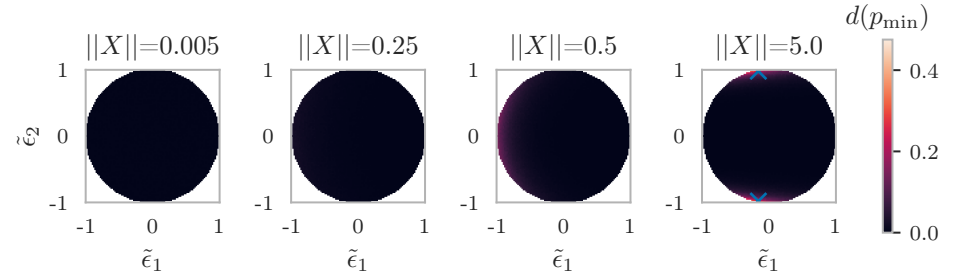
(b) $||\mathbf{\Delta}|| = \sigma \,/\, 2$



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

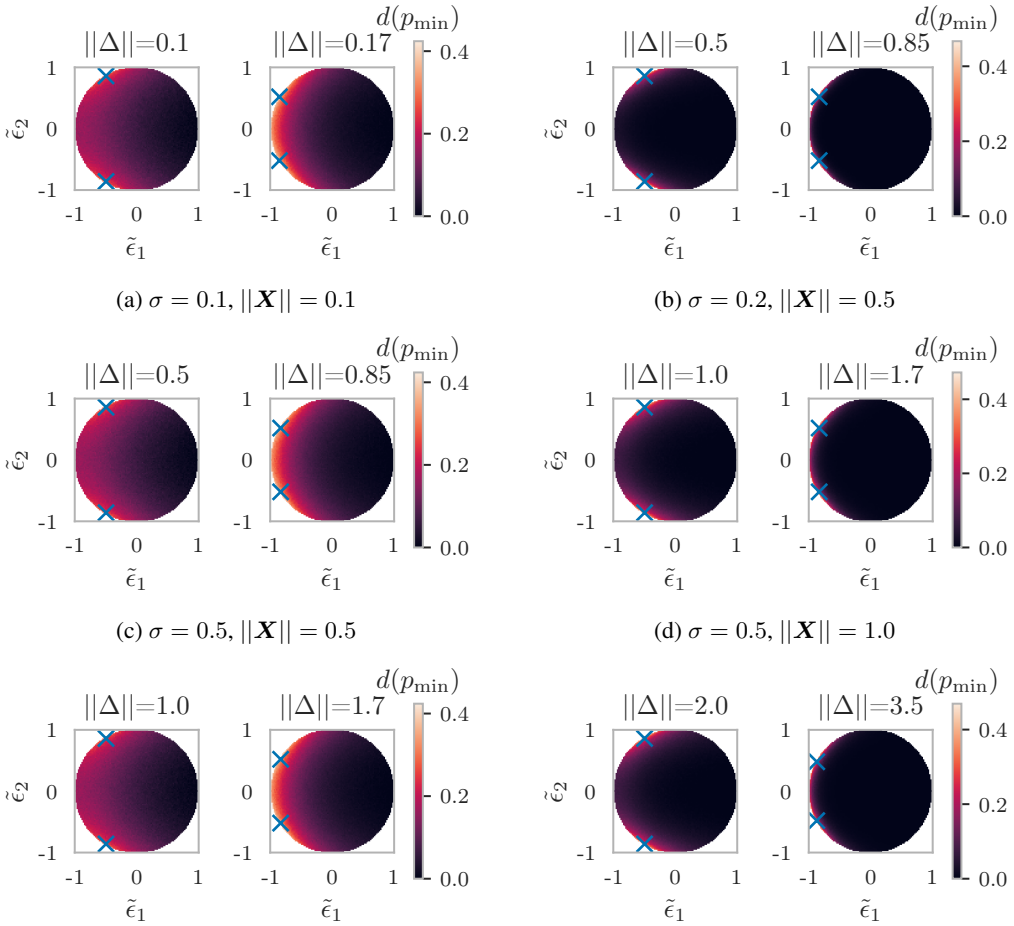


(d) $||\mathbf{\Delta}|| = 2\sigma$



(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.

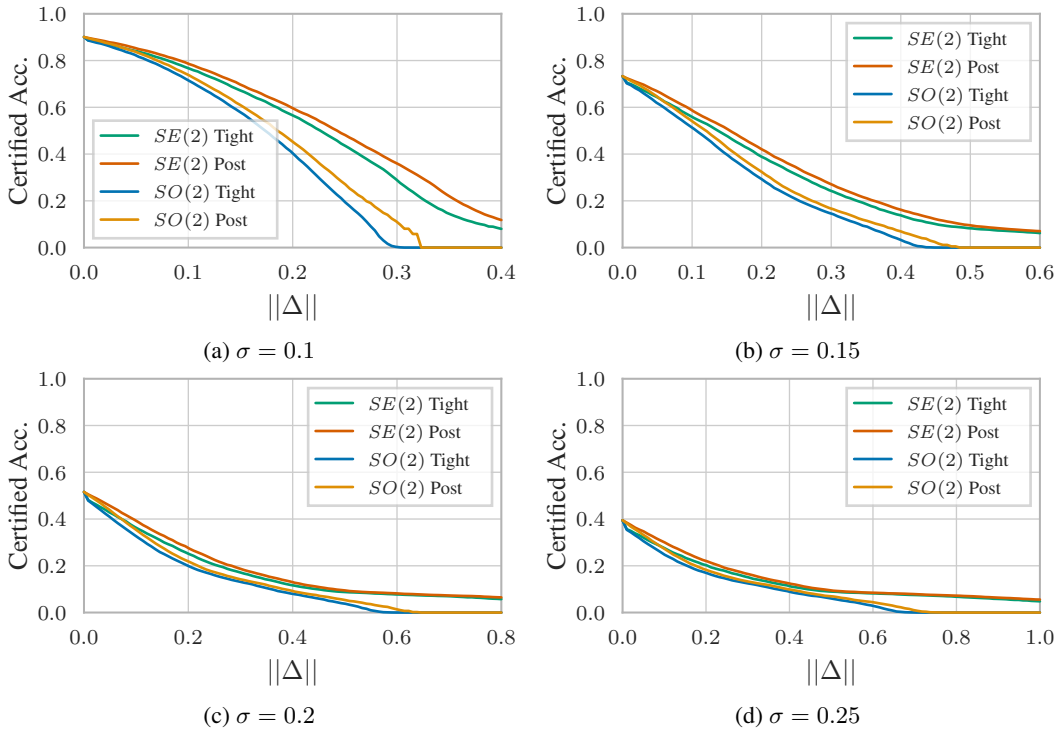
We first consider adversarial scaling on the point clouds constructed from MNIST (Appendix A.2.1). We then consider perturbations with random rotation components on MNIST (Appendix A.2.2) and random elemental rotation components on ModelNet40 (Appendix A.2.3). In Appendix A.2.4 we additionally consider arbitrary random rotations.

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 post-processing-based certificate for rotation invariance ($SO(2)$) and those for simultaneous rotation and translation invariance ($SE(2)$.

In all cases, the tight and post-processing-based certificates yield similar certified accuracies. The certified accuracies of the tight certificate are marginally lower, because we compute a probabilistic bound that holds with high probability $1 - \alpha$, rather than evaluating it exactly. This effect is amplified in the case $\sigma = 0.1$, 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 post-processing-based baseline becomes.

Furthermore, 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) $\sigma = 0.25$

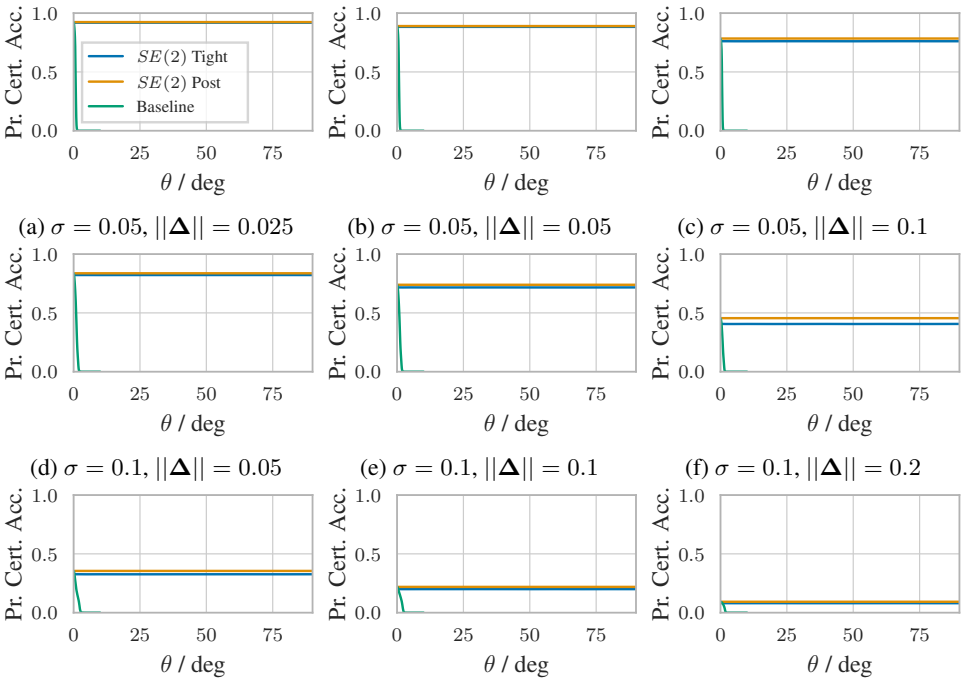
Figure 12: Comparison of certificates for adversarial scaling of MNIST with EnsPointNet and $\sigma \in \{0.1, 0.15, 0.2, 0.25\}$. Tight and post-processing-based certificates yield similar certified accuracies.

A.2.2 Perturbations with random rotation components on MNIST

Like in Section 8.2, we construct perturbations with rotational components by fixing $||\boldsymbol{\Delta}||$, randomly sampling perturbations of the specified norm and then rotating $\boldsymbol{Z} = X + \boldsymbol{\Delta}$ 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 post-processing-based certificate for simultaneous rotation and translation invariance (SE), as well as the black-box baseline w.r.t. 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 $\sigma \in \{0.05, 0.1, 0.2\}$, $||\boldsymbol{\Delta}||_2 \in \{\sigma \,/\, 2, \sigma, 2\sigma\}$ 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 benefit beyond that. It never yields higher probabilistic certified accuracy than the post-processing-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) $\sigma = 0.05$, $||\boldsymbol{\Delta}|| = 0.025$ (b) $\sigma = 0.05$, $||\boldsymbol{\Delta}|| = 0.05$ (c) $\sigma = 0.05$, $||\boldsymbol{\Delta}|| = 0.1$

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

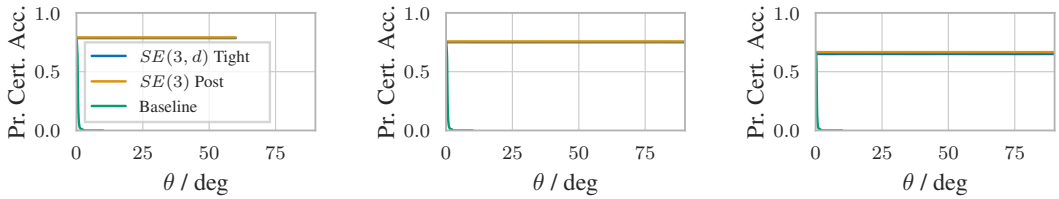
(g) $\sigma = 0.2$, $||\boldsymbol{\Delta}|| = 0.1$ (h) $\sigma = 0.2$, $||\boldsymbol{\Delta}|| = 0.2$ (i) $\sigma = 0.2$, $||\boldsymbol{\Delta}|| = 0.4$

Figure 13: Comparison of tight gray-box, post-processing-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 $||\boldsymbol{\Delta}||_2 \in \{\sigma \,/\, 2, \sigma, 2\sigma\}$ and rotating by angle θ. Group $SE(2)$ refers to simultaneous rotation and translation invariance. The tight gray-box certificate does not yield a higher probabilistic certified accuracy than the post-processing-based one.

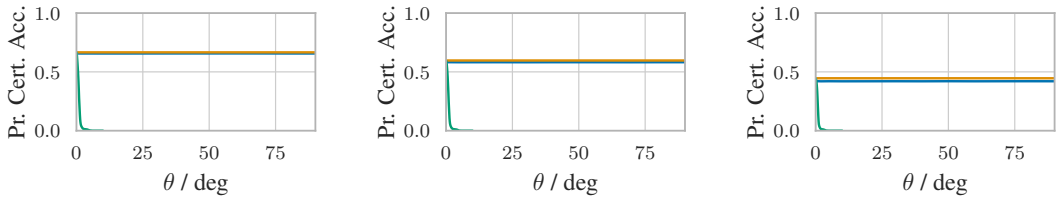
A.2.3 Perturbations with random elemental rotation components on ModelNet40

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 one of three coordinate axes (chosen randomly). We then compare the tight certificate for simultaneous invariance under elemental rotation and translation, as well as the post-processing-based certificate for simultaneous invariance under rotation and translation. We evaluate the tight certificate for all three axes and pick the best one.

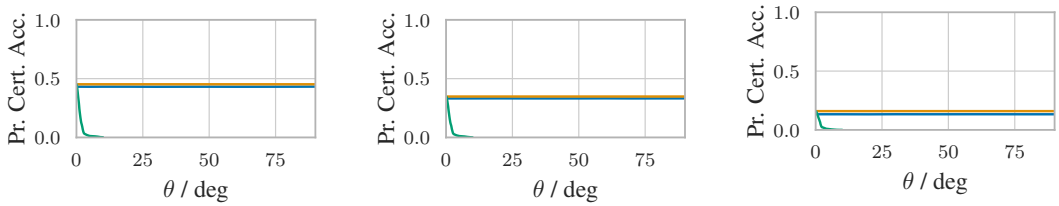
Figs. 14 to 16 show the results for EnsPointNet, AttnPointNet and EnsDGCNN. In all cases, both the tight- and post-processing-based certificate are constant in θ, but the tight certificate does not improve upon the post-processing-based one w.r.t. probabilistic certified accuracy.



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

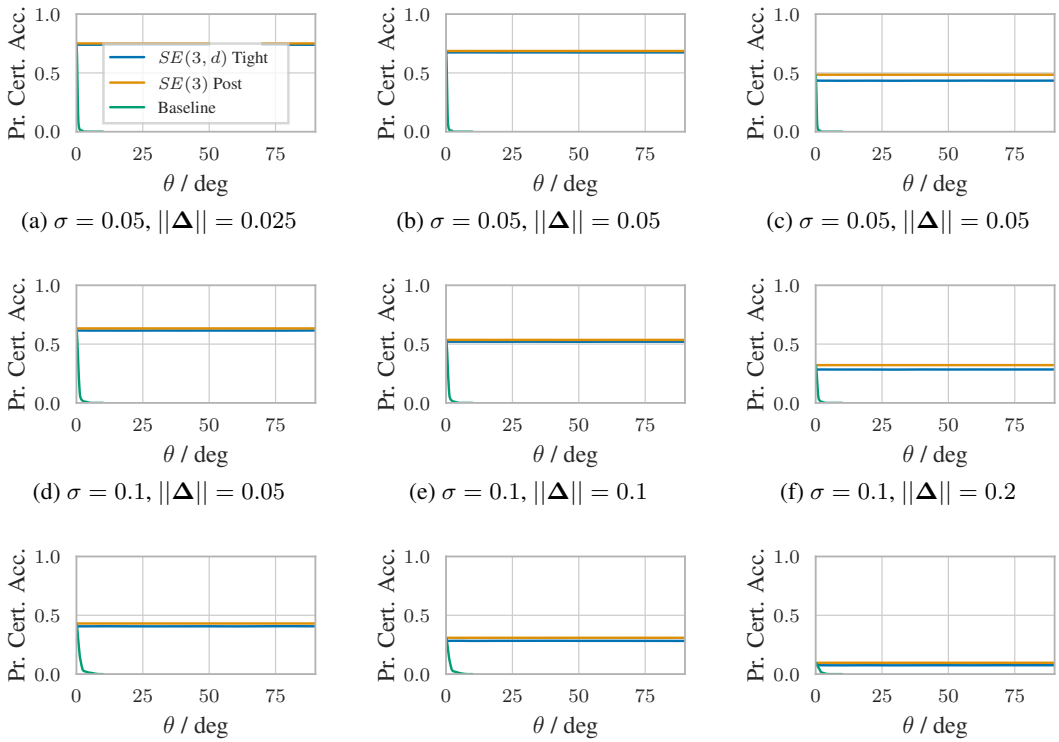


(d) $\sigma = 0.1$, $||\boldsymbol{\Delta}|| = 0.05$ (e) $\sigma = 0.1$, $||\boldsymbol{\Delta}|| = 0.1$ (f) $\sigma = 0.1$, $||\boldsymbol{\Delta}|| = 0.2$



(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, post-processing-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 coordinate axis by angle θ. Group $S(3)$ refers to simultaneous rotation and translation invariance and $SE(3, d)$ refers to invariance under elemental rotation and translation. The tight gray-box certificate does not yield a higher probabilistic certified accuracy than the post-processing-based one.

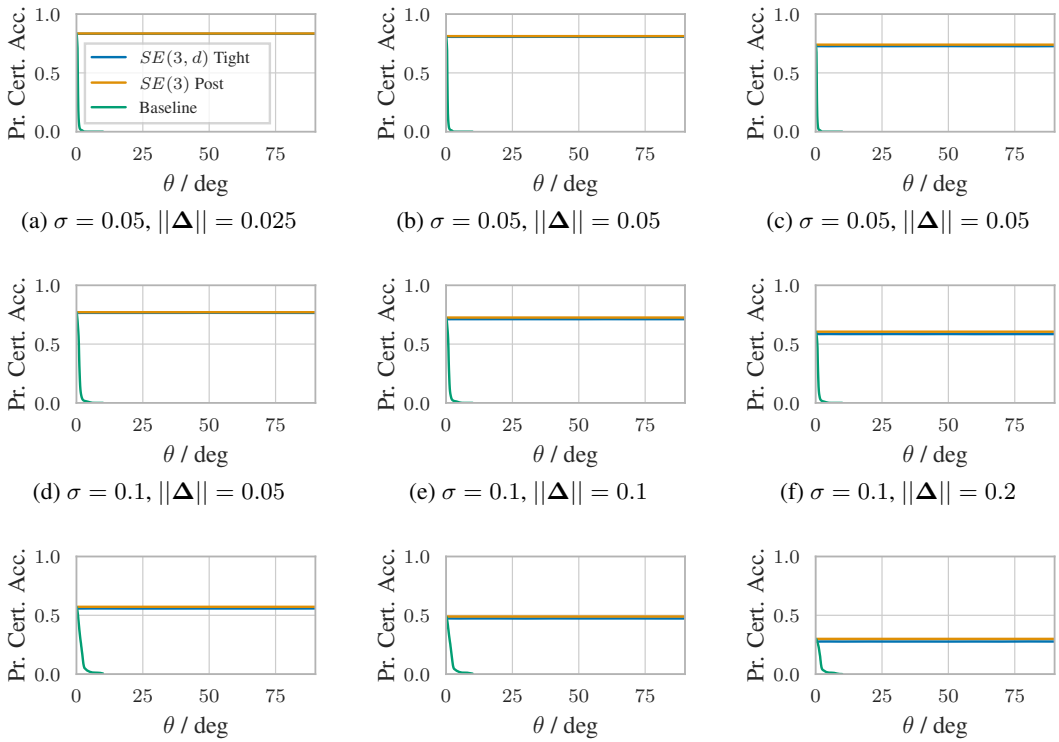


(a) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.025$ (b) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.05$ (c) $\sigma = 0.05$, $||\mathbf{\Delta}|| = 0.05$

(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, post-processing-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 coordinate axis by angle θ. Group $S(3)$ refers to simultaneous rotation and translation invariance and $SE(3, d)$ refers to invariance under elemental rotation and translation. The tight gray-box certificate does not yield a higher probabilistic certified accuracy than the post-processing-based one.



(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) $\sigma = 0.1$, $||\mathbf{\Delta}|| = 0.2$

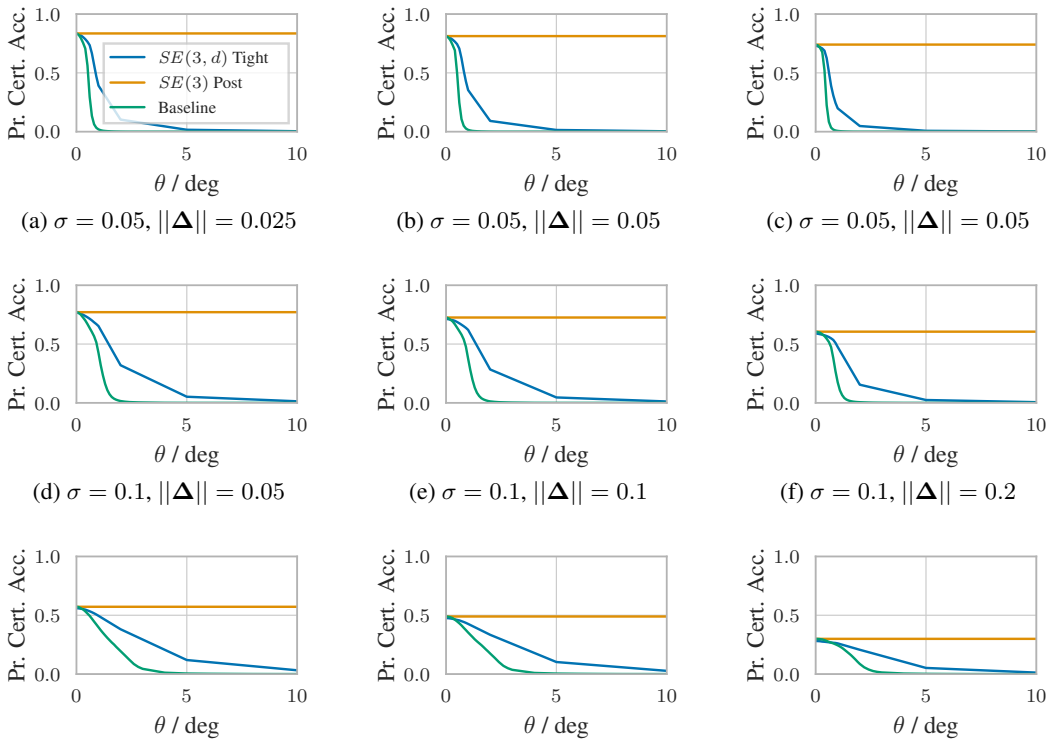
(g) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.1$ (h) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.2$ (i) $\sigma = 0.2$, $||\mathbf{\Delta}|| = 0.4$

Figure 16: Comparison of tight gray-box, post-processing-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 coordinate axis by angle θ. Group $S(3)$ refers to simultaneous rotation and translation invariance and $SE(3, d)$ refers to invariance under elemental rotation and translation. The tight gray-box certificate does not yield a higher probabilistic certified accuracy than the post-processing-based one.

A.2.4 Perturbations with arbitrary random rotation components on ModelNet40

Finally, we want to assess the ability of the tight certificate for invariance under *elemental* rotation and translation ($SE(3, d)$) to certify robustness under perturbations that have *arbitrary* rotation components. To this end, we sample perturbations $\mathbf{\Delta}$ of specified norm $||\mathbf{\Delta}||_2$ and randomly rotate $\boldsymbol{X}' = \boldsymbol{X} + \mathbf{\Delta}$ by angle θ around a randomly sampled axis. We obtain this axis by randomly sampling from a three-dimensional isotropic normal distribution and normalizing the result.

Fig. 17 shows the result of applying this evaluation scheme to EnsDGCNN on ModelNet40. The tight certificate outperforms the black-box one, as it can at least eliminate some of the rotational component induced by the sampling procedure. However, unlike the post-processing-based certificate for invariance under arbitrary rotations and translation ($SE(3)$, it is not constant in θ and can not certify robustness for $\theta \geq 10°$.



(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 17: Comparison of tight gray-box, post-processing-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, arbitrary axis by θ. Group $S(3)$ refers to simultaneous rotation and translation invariance and $SE(3, d)$ refers to invariance under elemental rotation and translation. The post-processing-based gray-box certificate outperforms the tight one, which can not account for arbitrary rotations.

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 are included with the supplementary material.

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 [122].

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 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 [123, 124]). 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 [123]. 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 [124]).

B.1.2 Data

ModelNet40 [120] 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 [121] 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 st eps 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 $p_{\boldsymbol{X},y^*}$ and – for the tight gray-box certificates involving rotation invariance – 10000 samples to bound the threshold t and 10000 samples to bound the optimal value itself (for more details, see Appendix D.9). For the tight gray-box certificates involving rotation invariance, we use Holm-Bonferroni correction ([125]) to ensure that all three confidence bounds hold simultaneously with probability $1 - \alpha$.

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 t 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 $0.06\,\mathrm{s}$. 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 [121] and ModelNet40 [120] datasets. Our implementation additionally uses code from a publicly available implementation of PointNet and PointNet++ [126], as well as the author's reference implementation of DGCNN [122]. Both are available under MIT license.

C Post-processing-based gray-box certificates

In the following, we first prove the soundness of our post-processing-based approach to gray-box robustness certification (Appendix C.1). We then prove the soundness of the equivalent pre-processing perspective for Frobenius norm certificates (Appendix C.2) and that isotropic Gaussian smoothing preserves invariance under groups composed of permutation and isometries of the Euclidean space (Appendix C.3). We conclude by discussing how the optimization problem underlying the pre-processing perspective can be solved for different invariances (Appendix C.4).

C.1 Proof of Theorem 1

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

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 $\tau \in \mathbb{T}$ such that $\boldsymbol{X}' = \tau(\boldsymbol{Z})$. Since $\tau \in \mathbb{T}$, we know that $h(\tau(\boldsymbol{Z})) = h(\boldsymbol{Z})$. Since $\boldsymbol{Z} \in \mathbb{B}$, we know that $h(\boldsymbol{Z}) = y^*$. By the transitive property, we have $h(\boldsymbol{X}') = y^*$. □

C.2 Proof of Theorem 2

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

Proof. If $\min_{\tau \in \mathbb{T}} ||\tau(\boldsymbol{X}') - \boldsymbol{X}||_2 < r$, then the minimizer $\tau^* = \operatorname{argmin}_{\tau \in \mathbb{T}} ||\tau(\boldsymbol{X}') - \boldsymbol{X}||_2$ fulfills $\tau^*(\boldsymbol{X}') \in \mathbb{B}$ and thus $h(\tau^*(\boldsymbol{X}')) = y^*$. Since $\tau^* \in \mathbb{T}$, we have $h(\tau^*(\boldsymbol{X}')) = h(\boldsymbol{X}')$. By the transitive property, we have $h(\boldsymbol{X}') = y^*$. □

C.3 Proof of Theorem 3

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

Recall that

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

We shall prove Theorem 3 by showing that $\forall \tau \in \mathbb{T} : p_{\boldsymbol{X},y} = p_{\tau(\boldsymbol{X}),y}$, i.e. the prediction probabilities are invariant. To do so, we need the following simple lemma:

Lemma 1. *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 $\boldsymbol{z}'$. □

In the following, let $\hat{g}(\boldsymbol{Z}) = \mathbb{1}[g(\boldsymbol{Z}) = y^*]$ indicate whether base classifier g classifies input $\boldsymbol{Z}$ as y^*. Now, consider an arbitrary $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and transformation $\tau \in \mathbb{T}$. Since

$\mathbb{T} \subseteq \{\boldsymbol{X} \mapsto \boldsymbol{P}\eta(\boldsymbol{X}) \mid \boldsymbol{P} \in S(N), \eta \in E(D)\}$ and $E(D)$ is composed of translations, rotations and reflections, we must have $\tau(\boldsymbol{X}) = \boldsymbol{P}\left(\boldsymbol{X}\boldsymbol{R}^T + \mathbf{1}_N \boldsymbol{c}^T\right)$ with $\boldsymbol{P} \in S(N)$, $\boldsymbol{R}^T \in O(D)$ and $\boldsymbol{c} \in \mathbb{R}^D$, where $O(D) \subseteq \mathbb{R}^{D\times D}$ is the orthogonal group, i.e. the set of all orthonormal matrices. Note that $\boldsymbol{P}$ is also orthonormal and that $|\det(\boldsymbol{P})| = |\det(\boldsymbol{R})| = 1$. We thus have

$$
\begin{aligned}
p_{\tau(\boldsymbol{X}),y} &= \int\limits_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{Z})\mu_{\tau(\boldsymbol{X})}(\boldsymbol{Z})d\boldsymbol{Z} \\
&= \int\limits_{\mathbb{R}^{N\times D}} \hat{g}(\tau(\boldsymbol{V}))\mu_{\tau(\boldsymbol{X})}(\tau(\boldsymbol{V}))d\boldsymbol{V} \\
&= \int\limits_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V})\mu_{\tau(\boldsymbol{X})}(\tau(\boldsymbol{V}))d\boldsymbol{V},
\end{aligned}
$$

where the second equality is due to a substitution $\boldsymbol{V} = \tau^{-1}(\boldsymbol{Z})$ and the third equality is due to $\tau \in \mathbb{T}$ and g being invariant under $\mathbb{T}$. We may now insert the definition of our isotropic matrix normal distribution μ and then use the fact that $\forall \boldsymbol{d} : \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{m} + \boldsymbol{d}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{z} - \boldsymbol{d} \mid \boldsymbol{m}, \boldsymbol{\Sigma})$ to eliminate the translation

$$
\begin{aligned}
&\int\limits_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V})\mu_{\tau(\boldsymbol{X})}(\tau(\boldsymbol{V}))d\boldsymbol{V} \\
&= \int\limits_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V}) \prod_{d=1}^{D} \mathcal{N}\left(\tau(\boldsymbol{Z})_{:,d} \mid \tau(\boldsymbol{X})_{:,d}, \sigma^2\mathbf{I}_N\right) d\boldsymbol{V} \\
&= \int\limits_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V}) \prod_{d=1}^{D} \mathcal{N}\left(\left(\boldsymbol{P}\left(\boldsymbol{X}\boldsymbol{R}^T + \mathbf{1}_N\boldsymbol{c}^T\right)\right)_{:,d} \mid \left(\boldsymbol{P}\left(\boldsymbol{X}\boldsymbol{R}^T + \mathbf{1}_N\boldsymbol{c}^T\right)\right)_{:,d}, \sigma^2\mathbf{I}_N\right) d\boldsymbol{V} \\
&= \int\limits_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V}) \prod_{d=1}^{D} \mathcal{N}\left(\left(\boldsymbol{P}\left(\boldsymbol{X}\boldsymbol{R}^T\right)\right)_{:,d} \mid \left(\boldsymbol{P}\left(\boldsymbol{X}\boldsymbol{R}^T\right)\right)_{:,d}, \sigma^2\mathbf{I}_N\right) d\boldsymbol{V} \\
&= \int\limits_{\mathbb{R}^{N\times D}} \hat{g}(\boldsymbol{V}) \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{P}\left(\boldsymbol{X}\boldsymbol{R}^T\right)_{:,d} \mid \boldsymbol{P}\left(\boldsymbol{X}\boldsymbol{R}^T\right)_{:,d}, \sigma^2\mathbf{I}_N\right) d\boldsymbol{V}.
\end{aligned}
$$

Next, we can use the fact that $\boldsymbol{P}$ is orthonormal, i.e. $\boldsymbol{P}\boldsymbol{P}^T = \mathbf{I}_N$ and $|\det(\boldsymbol{P}^{-1})| = 1$, to eliminate the permutation via Lemma 1:

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

Finally, we can express our isotropic matrix normal density as the product of N D-dimensional isotropic normal densities, rather than D N-dimensional isotropic normal densities and use the fact

that $\boldsymbol{R}$ is orthonormal, i.e. $\boldsymbol{R}\boldsymbol{R}^T = \mathbf{I}_N$ and $|\det(\boldsymbol{R}^{-1})| = 1$, to eliminate the rotation via Lemma 1:

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

By the transitive property, we have $p_{\tau(\boldsymbol{X}),y^*} = p_{\boldsymbol{X},y^*}$, which concludes our proof.

C.4 Application of Theorem 2

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

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 that, let us reiterate that using Theorem 2 and solving these optimization problems 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 1). 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 norm of $\boldsymbol{X}'$ can be used to determine whether it is part of the original certified region $\mathbb{B}$.

C.4.1 Translation invariance

In this case, we have $\mathbb{T} = \left\{\boldsymbol{X} \mapsto \boldsymbol{X} + \mathbf{1}_N\boldsymbol{c}^T \mid \boldsymbol{c} \in \mathbb{R}^D\right\}$, meaning

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

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. $\boldsymbol{c}$ and setting them to zero.

C.4.2 Rotation invariance

In this case, we have $\mathbb{T} = \left\{\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(D)\right\}$, meaning

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

This is a special case of the orthogonal Procrustes problem, which can be solved via singular value decomposition [105]. 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 $\hat{S} = \operatorname{diag}\left(1, \ldots, 1, \operatorname{sign}\left(\det\left(\boldsymbol{V}\boldsymbol{U}^T\right)\right)\right)$.

C.4.3 Simultaneous rotation and reflection invariance

In this case, we have $\mathbb{T} = \left\{\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T \mid \boldsymbol{R} \in O(D)\right\}$, where $O(D)$ is the orthogonal group, meaning

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

This is the orthogonal Procrustes problem, which can again be solved via singular value decomposition [105]. The optimal rotation matrix $\boldsymbol{R}^*$ is given by $\boldsymbol{R}^* = \boldsymbol{V}\boldsymbol{U}^T$, where $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V} = (\boldsymbol{X}')^T\boldsymbol{X}$. Here, accounting for the sign of the determinant is not necessary.

C.4.4 Permutation invariance

In this case, we have $\mathbb{T} = \{\boldsymbol{X} \mapsto \boldsymbol{P}\boldsymbol{X} \mid \boldsymbol{P} \in S(N)\}$, meaning

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

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.

C.4.5 Simultaneous rotation and translation invariance

In this case, we have $\mathbb{T} = \left\{\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T + \mathbf{1}_N \boldsymbol{c}^T \mid \boldsymbol{R} \in SO(D), \boldsymbol{c} \in \mathbb{R}^D\right\}$, meaning

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

It can be shown that this is equivalent to centering $\boldsymbol{X}'$ and $\boldsymbol{X}$ and then solving the orthogonal Procrustes problem [106], i.e.

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

As discussed in Appendix C.4.2, this problem can be solved via singular value decomposition.

C.4.6 Simultaneous permutation, rotation and translation invariance

In this case, we have $\mathbb{T} = \left\{\boldsymbol{X} \mapsto \boldsymbol{P}\left(\boldsymbol{X}\boldsymbol{R}^T + \mathbf{1}_N \boldsymbol{c}^T\right) \mid \boldsymbol{P} \in S(N), \boldsymbol{R} \in SO(D), \boldsymbol{c} \in \mathbb{R}^D\right\}$, meaning

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

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 [127]). Note that, if some upper bound $\hat{m}$ with $\min_{\tau \in \mathbb{T}} ||\tau(\boldsymbol{X}') - \boldsymbol{X}||_2 < \hat{m}$ fulfills $\hat{m} < r$, then $\min_{\tau \in \mathbb{T}} ||\tau(\boldsymbol{X}') - \boldsymbol{X}||_2 < r$. Thus, any approximate solution to the point cloud registration problem can be used for membership inference. If $\hat{m} < r$, we provably know that $\boldsymbol{X}' \in \tilde{\mathbb{B}}$ and thus $h(\boldsymbol{X}') = y^*$. 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, the membership inference via approximate solutions to the point cloud registration problem is sound, but not optimal.

D Tight gray-box certificates

In the following, we first restate the Neyman-Pearson lemma from statistical testing, which we shall use to exactly solve the optimization problem underlying our tight gray-box certificates (Appendix D.1). We then summarize our certification methodology (Appendix D.2) and prove the main theorem underlying this methodology (Appendix D.3). This methodology is then applied to translation invariance in arbitrary dimensions (Appendix D.4, rotation invariance in 2D (Appendix D.5 and invariance to elemental rotations in 3D (Appendix D.6). The latter two are further combined with translation invariance (Appendices D.7 and D.8). We conclude by discussing how to use Monte Carlo sampling to obtain narrow probabilistic bounds on the certificates involving rotation invariance (Appendix D.9).

D.1 Neyman-Pearson lemmmapa

For our purposes, the Neyman-Pearson lemma [104] can be formulated as follows:

Lemma 2 (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 $t \in \mathbb{R}_+$, the set $\left\{z \mid \frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} = t\right\}$ 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*

$$\left(\min_{h:\mathbb{A}\to\mathbb{Y}} \Pr_{z\sim\mu_{\boldsymbol{X}'}}[h(z) = y] \text{ s.t. } \Pr_{z\sim\mu_{\boldsymbol{X}}}[h(z) = y] \geq p\right) = \mathbb{E}_{z\sim\mu_{\boldsymbol{X}'}}[h^*(z)]$$
$$\text{with } h^*(z) = \mathbb{1}\left[\frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} \leq t\right] \text{ and } t \in \mathbb{R}_+ \text{ s.t. } \mathbb{E}_{z\sim\mu_{\boldsymbol{X}}}[h^*(z)] = p.$$

Here, indicator function h^* corresponds to a classifier that predicts class y if and only if the likelihood ratio is below a specific certain threshold. For an application of this variant of the Neyman-Pearson lemma to black-box robustness certification, as well as a discussion of its relation most powerful hypothesis tests, see [27]. For various other formulations of the lemma, see [128].

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 F.

Lemma 3 (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 $t \in \mathbb{R}_+$, the set $\left\{z \mid \frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} = t\right\}$ 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*

$$\left(\max_{h:\mathbb{A}\to\mathbb{Y}} \Pr_{z\sim\mu_{\boldsymbol{X}'}}[h(z) = y] \text{ s.t. } \Pr_{z\sim\mu_{\boldsymbol{X}}}[h(z) = y] \leq p\right) = \mathbb{E}_{z\sim\mu_{\boldsymbol{X}'}}[h^*(z)]$$
$$\text{with } h^*(z) = \mathbb{1}\left[\frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} \geq t\right] \text{ and } t \in \mathbb{R}_+ \text{ s.t. } \mathbb{E}_{z\sim\mu_{\boldsymbol{X}}}[h^*(z)] = p.$$

D.2 Certification methodology

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 $\mathbb{H}$ 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. Further recall that the invariance constraint can be eliminated by applying a canonical map, which maps each element of $\mathbb{R}^{N\times D}$ to a distinct representative of its equivalence class.

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

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

We shall prove this main theorem in Appendix D.3.

After applying Theorem 4 with an appropriate canonical map γ, we first notice that the optimization problem only depends on $\mathbb{1}[h(\boldsymbol{Z}) = y^*]$, i.e. whether an input is classified as y^* or some other class.

To avoid notational clutter, we may equivalently optimize over binary classifiers $h : \mathbb{R}^{N\times D} \to \{0,1\}$, with $h(\boldsymbol{Z}) = 1$ corresponding to the case that $\boldsymbol{Z}$ is classified as y^*:

$$\min_{h:\mathbb{R}^{N\times D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\gamma(\boldsymbol{Z}))\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right] \geq p_{\boldsymbol{X},y^*} \tag{9}$$

We can then transform the expectation integrals via a carefully chosen invertible substitution $\xi : \mathbb{R}^{N\times D} \to \mathbb{A}$ that maps all elements from $\mathbb{R}^{N\times D}$ into some alternative, K-dimensional parameter space $\mathbb{A} \subseteq \mathbb{R}^K$. We choose this substitution such that applying the canonical map γ is equivalent to setting $L < K$ parameters of the alternative parameter space to zero. That is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right] &= \int_{\mathbb{A}} h\left(\gamma\left(\xi^{-1}\left(\boldsymbol{v}\right)\right)\right)\mu_{\boldsymbol{X}}\left(\xi^{-1}\left(\boldsymbol{v}\right)\right)\left|\det\left(D_{\xi^{-1}}\right)\right| d\boldsymbol{v} \\ &= \int_{\mathbb{A}} h\left(\xi^{-1}\left(\begin{bmatrix}\boldsymbol{0}_L \\ \boldsymbol{v}_{(L+1):}\end{bmatrix}\right)\right)\mu_{\boldsymbol{X}}\left(\xi^{-1}\left(\boldsymbol{v}\right)\right)\left|\det\left(D_{\xi^{-1}}\right)\right| d\boldsymbol{v}.\end{aligned}$$

Since the inverse ξ^{-1} is matrix-valued, we define $D_{\xi^{-1}}$ to be the Jacobian of $\mathrm{vec}\left(\xi^{-1}\right)$.

Because the term involving h does not depend on the first L dimensions of $\boldsymbol{v}$, we may marginalize them out. Let $\mathbb{A}_{:L} = (\boldsymbol{v}_{:L} \mid \boldsymbol{v} \in \mathbb{A},)$ be the first L dimensions of the parameter space and let $\mathbb{A}_{(L+1):} = \left(\boldsymbol{v}_{(L+1):} \mid \boldsymbol{v} \in \mathbb{A},\right)$ be the remaining $K - L$ dimensions. Then,[2]

$$\begin{aligned}&\int_{\mathbb{A}} h\left(\xi^{-1}\left(\begin{bmatrix}\boldsymbol{0}_L \\ \boldsymbol{v}_{(L+1):}\end{bmatrix}\right)\right)\mu_{\boldsymbol{X}}\left(\xi^{-1}\left(\boldsymbol{v}\right)\right)\left|\det\left(D_{\xi^{-1}}\right)\right| \, d\boldsymbol{v} \\ =&\int_{\mathbb{A}_{(L+1):}} h\left(\xi^{-1}\left(\begin{bmatrix}\boldsymbol{0}_L \\ \boldsymbol{v}_{(L+1):}\end{bmatrix}\right)\right)\int_{\mathbb{A}_{:L}}\mu_{\boldsymbol{X}}\left(\xi^{-1}\left(\boldsymbol{v}\right)\right)\left|\det\left(D_{\xi^{-1}}\right)\right| \, d\boldsymbol{v}_{:L}\, d\boldsymbol{v}_{(L+1):} \\ :=&\int_{\mathbb{A}_{(L+1):}} h\left(\xi^{-1}\left(\begin{bmatrix}\boldsymbol{0}_L \\ \boldsymbol{v}_{(L+1):}\end{bmatrix}\right)\right)\nu_{\boldsymbol{X}}\left(\boldsymbol{v}_{(L+1):}\right) d\boldsymbol{v}_{(L+1):},\end{aligned}$$

where $\nu_{\boldsymbol{X}}$ is the marginal distribution over $\mathbb{A}_{(L+1):}$, the parameter space occupied by the representatives of the different equivalence classes. Inserting into the problem from Eq. (9) yields

$$\min_{h:\mathbb{R}^{N\times D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{w}\sim\nu_{\boldsymbol{X}'}}\left[h\left(\xi^{-1}\left(\begin{bmatrix}\boldsymbol{0}_L \\ \boldsymbol{w}\end{bmatrix}\right)\right)\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{w}\sim\nu_{\boldsymbol{X}}}\left[h\left(\xi^{-1}\left(\begin{bmatrix}\boldsymbol{0}_L \\ \boldsymbol{w}\end{bmatrix}\right)\right)\right] \geq p_{\boldsymbol{X},y^*},$$

where we have written $\boldsymbol{w} \in \mathbb{R}^{K-L}$ instead of $\boldsymbol{v}_{(L+1):}$ to clarify that the optimization problem no longer depends on the entire K-dimensional parameter space.

At this point, it is convenient to treat the function $h(\boldsymbol{Z})$ as a family of variables $(h_{\boldsymbol{Z}})$ indexed by $\mathbb{R}^{N\times D}$. Because ξ is a bijection and $\boldsymbol{w} \mapsto \begin{bmatrix}\boldsymbol{0}_L^T & \boldsymbol{w}^T\end{bmatrix}^T$ is an injective function, each distinct vector $\boldsymbol{w} \in \mathbb{A}_{(L+1):}$ indexes a distinct variable $h_{\boldsymbol{Z}}$. We may thus equivalently optimize over a family of variables $(h'_{\boldsymbol{w}}) \in \mathbb{Y}$ indexed by $\mathbb{A}_{(L+1):}$:

$$\min_{h':\mathbb{A}_{(L+1):}\to\{0,1\}} \mathbb{E}_{\boldsymbol{w}\sim\nu_{\boldsymbol{X}'}}\left[h'\left(\boldsymbol{w}\right)\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{w}\sim\nu_{\boldsymbol{X}}}\left[h'\left(\boldsymbol{w}\right)\right] \geq p_{\boldsymbol{X},y^*}. \tag{10}$$

Finally, we can exactly solve the lower-dimensional problem from Eq. (10) via Lemma 2:

$$\begin{aligned}\left(\min_{h':\mathbb{A}_{(L+1):}\to\{0,1\}} \mathbb{E}_{\boldsymbol{w}\sim\nu_{\boldsymbol{X}'}}\left[h'(z)\right] \text{ s.t. } \mathbb{E}\boldsymbol{w}\sim\nu_{\boldsymbol{X}}\left[h'(z)\right] \geq p_{\boldsymbol{X},y^*}\right) =& \mathbb{E}_{z\sim\mu_{\boldsymbol{X}'}}\left[h^*(z)\right] \\ \text{with } h^*(z) = \mathbb{1}\left[\frac{\nu_{\boldsymbol{X}'}(z)}{\nu_{\boldsymbol{X}}(z)} \leq t\right] &\text{ and } t \in \mathbb{R}_+ \text{ s.t. } \mathbb{E}_{z\sim\nu_{\boldsymbol{X}}}\left[h^*(z)\right] = p.\end{aligned}$$

We discuss which canonical map and substitution to use, how to derive the marginal distributions $\nu_{\boldsymbol{X}}$ and $\nu_{\boldsymbol{X}'}$ and how to apply the Neyman-Pearson lemma in the respective sections of the different invariances (see Appendices D.4 to D.8).

[2] To simplify the notation, we assume the first L limits of integration to be independent of the remaining $K - L$, which is the case for the smoothing distribution and substitutions we use for our certificates.

D.3 Proof of Theorem 4

Before deriving the individual certificates, we need to prove our main theorem.

Theorem 4. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be invariant under group $\mathbb{T}$ and let $\mathbb{H}_\mathbb{T}$ be defined as in Eq.* (2)*. If $\gamma : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times D}$ is a canonical 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 2 (Equivalence classes)**.** *The equivalence class of an input $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ w.r.t. invariance under a group of transformations $\mathbb{T}$ is $[x]_\mathbb{T} = \{\tau(x) \mid \tau \in \mathbb{T}\}$.*

and

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*

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

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

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 of Theorem 4 by deriving two lemmeta.

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

Lemma 4. *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 2 would imply that $\boldsymbol{Z} \in [\boldsymbol{Z}']_\mathbb{T}$. Eq. (4) from Definition 2 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 next 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 5. *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 2, 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 arbitrary 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 $h(\boldsymbol{Z}) = h(\boldsymbol{Z}')$. □

Lemma 5 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:

$$\begin{aligned} \min_{h:\mathbb{R}^{N\times D}\to\mathbb{Y}} & \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}[h(\gamma(\boldsymbol{Z}))=y^*] \\ \text{s.t.} & \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[h(\gamma(\boldsymbol{Z}))=y^*] \geq p_{\boldsymbol{X},y^*} \wedge \forall \boldsymbol{Z}: h(\boldsymbol{Z}) = h(\gamma(\boldsymbol{Z})). \end{aligned} \tag{11}$$

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 4 shows that $h(\boldsymbol{Z})$ does not appear in the objective function or first constraint of Eq. (11), 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. (11).

D.4 Tight certificate for invariance under translation

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

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

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

Following our certification methodology, we derive this tight certificate for translation invariance, i.e. invariance under $\mathbb{T} = \left\{\boldsymbol{X} \mapsto \boldsymbol{X} + \boldsymbol{1}_N\boldsymbol{c}^T \mid \boldsymbol{c} \in \mathbb{R}^D\right\}$, by first specifying a canonical map (Appendix D.4.1) and then performing a suitable substitution to find the marginal distribution over a subset of an alternative parameter space occupied by the representatives of the different equivalence classes (Appendix D.4.2).

D.4.1 Canonical map

Lemma 6. *The function $\gamma(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{1}_N\boldsymbol{Z}_{1,:}^T$ is a canonical map for translation invariance.*

Proof. Based on the definition of canonical maps and translation invariance, we need to show that

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

and

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

$[\boldsymbol{Z}]_{\mathbb{T}} = \left\{\boldsymbol{Z} + \boldsymbol{1}_N\boldsymbol{c}^T \mid \boldsymbol{c} \in \mathbb{R}^{N\times D}\right\}$.

The function γ performs a translation with $\boldsymbol{c} = \boldsymbol{Z}_{1,:}$, so the first criterion is fulfilled.

Now, consider an arbitrary pair of matrices $\boldsymbol{Z}, \boldsymbol{Z}' \in \mathbb{R}^{N\times D}$ with $\boldsymbol{Z}' \in [\boldsymbol{Z}]_T$, i.e. $\boldsymbol{Z}' = \boldsymbol{Z} + \boldsymbol{1}_N\boldsymbol{c}^T$ for some $\boldsymbol{c}$. We know that $\boldsymbol{Z}'_{1,:} = \boldsymbol{Z}_{1,:} + \boldsymbol{c}$ and thus

$$\gamma(\boldsymbol{Z}') = \boldsymbol{Z}' - \boldsymbol{1}_N\boldsymbol{Z}'_{1,:} = \left(\boldsymbol{Z} + \boldsymbol{1}_N\boldsymbol{c}^T\right) - \boldsymbol{1}_N\left(\boldsymbol{Z}_{1,:} + \boldsymbol{c}\right)^T = \boldsymbol{Z} - \boldsymbol{1}_N\boldsymbol{Z}_{1,:}^T = \gamma(\boldsymbol{Z}),$$

meaning the second criterion is fulfilled as well. □

D.4.2 Proof for certificate

Our canonical map $\gamma(\boldsymbol{Z})$ subtracts the first row of $\boldsymbol{Z}$ from all rows. We perform a substitution into the alternative parameter space that specifies the first row $\boldsymbol{Z}_1$, as well $\boldsymbol{Z}_n - \boldsymbol{Z}_1$, i.e. the difference between all other rows to $\boldsymbol{Z}_1$. In this parameter space, applying the canonical map is equivalent to setting the first D parameters to 0.

Before deriving the marginal distribution over this lower-dimensional space and finding the worst-case classifier, we need the following lemma:

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 $\sum_{d=1}^D x_d = D\overline{\boldsymbol{x}}$. □

Further recall the following lemma from Appendix C.3:

Lemma 1. *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*

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

We begin our proof of Theorem 5 by reformulating the canonical map γ from Lemma 6 as $\gamma(\boldsymbol{Z}) = \boldsymbol{K}\boldsymbol{Z}$ with $\boldsymbol{K} = \begin{bmatrix} 0 & \mathbf{0}_{N-1}^T \\ -\mathbf{1}_{N-1} & \mathbf{I}_{N-1} \end{bmatrix}$. The multiplication $\boldsymbol{K}\boldsymbol{Z}$ sets the first row of $\boldsymbol{Z}$ to zero and subtracts the first row from all other rows, just like our originally defined γ.

Applying Theorem 4 with this canonical map shows that

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

We notice that the above optimization problem does not depend on the specific class predicted by h, but only on whether it is y^* or y. To avoid clutter, we formulate the equivalent problem

$$\min_{h:\mathbb{R}^{N\times D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{K}\boldsymbol{Z})\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\boldsymbol{K}\boldsymbol{Z})\right] \geq p_{\boldsymbol{X},y^*}, \tag{12}$$

where $h(\boldsymbol{Z}) = 1$ indicates that $\boldsymbol{Z}$ is classified as y^*.

Recall that $\mu_{\boldsymbol{X}}$ is an isotropic matrix normal distribution with variance σ^2. We thus have:

$$\mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\boldsymbol{K}\boldsymbol{Z})\right] = \int_{\mathbb{R}^{N\times D}} h(\boldsymbol{K}\boldsymbol{Z}) \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{V}_{:,d} \mid \boldsymbol{X}_{:,d}, \sigma^2\mathbf{I}\right)\, d\boldsymbol{Z} \tag{13}$$

Substitution. We can now use $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{Z}$ with $\boldsymbol{A} = \begin{bmatrix} 1 & \mathbf{0}_{N-1}^T \\ -\mathbf{1}_{N-1} & \mathbf{I}_{N-1} \end{bmatrix}$, $\boldsymbol{A}^{-1} = \begin{bmatrix} 1 & \mathbf{0}_{N-1} \\ \mathbf{1}_{N-1} & \mathbf{I}_{N} \end{bmatrix}$, where $\mathbf{1}_N \in \{1\}^{N-1}$ is a vector that has value 1 everywhere and $\mathbf{I}_{N-1} \in \{0,1\}^{(N-1)\times(N-1)}$ is the identity matrix.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\boldsymbol{K}\boldsymbol{Z})\right] &= \int_{\mathbb{R}^{N\times D}} h(\boldsymbol{K}\boldsymbol{A}^{-1}\boldsymbol{V}) |\det\left(\boldsymbol{A}^{-1}\right)|^D \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{A}^{-1}\boldsymbol{V}_{:,D} \mid \boldsymbol{X}_{:,d}, \sigma^2\mathbf{I}\right)\, d\boldsymbol{V} \\
&= \int_{\mathbb{R}^{N\times D}} h(\boldsymbol{K}\boldsymbol{A}^{-1}\boldsymbol{V}) \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{V}_{:,d} \mid \boldsymbol{A}\boldsymbol{X}_{:,d}, \sigma^2\boldsymbol{A}\boldsymbol{A}^T\right)\, d\boldsymbol{V}
\end{aligned}$$

The second equality follows from applying Lemma 1 D times.

Marginalization. Note that $\boldsymbol{K}\boldsymbol{A}^{-1} = \begin{bmatrix} 0 & \mathbf{0}_{N-1} \\ \mathbf{0}_{N-1} & \mathbf{I}_{N-1} \end{bmatrix}$. Since the first column is 0 everywhere, $h(\boldsymbol{K}\boldsymbol{A}^{-1}\boldsymbol{V})$ does not depend on the first row of $\boldsymbol{V}$, meaning we can marginalize out $\boldsymbol{V}_{1,1}, \ldots, \boldsymbol{V}_{1,D}$. For multivariate normal distributions, this is equivalent to dropping the first row of the mean and the first row and column of the covariance matrix. In the following, we use $\boldsymbol{A}_{2:} \in \mathbb{R}^{(N-1)\times D}$ to denote the remaining $N-1$ rows of $\boldsymbol{A}$.

$$= \int_{\mathbb{R}^{N\times D}} h\left(\begin{bmatrix} \mathbf{0}_D \\ \boldsymbol{V}_{2:} \end{bmatrix}\right) \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{V}_{2:,d} \mid \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}, \sigma^2\boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T\right)\, d\boldsymbol{V}_{2:}. \tag{14}$$

$$\tag{15}$$

Let $\nu_{\boldsymbol{X}}$ be this $((N-1)\times D)$-dimensional marginal distribution with density $\nu_{\boldsymbol{X}}(\boldsymbol{W}) = \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{W}_{:,d} \mid \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}, \sigma^2\boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T\right)$. We may then rewrite Eq. (12) as

$$\min_{h:\mathbb{R}^{N\times D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{W}\sim\nu_{\boldsymbol{X}'}}\left[h\left(\begin{bmatrix} \mathbf{0}_D^T \\ \boldsymbol{W} \end{bmatrix}\right)\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{W}\sim\nu_{\boldsymbol{X}}}\left[h\left(\begin{bmatrix} \mathbf{0}_D \\ \boldsymbol{W} \end{bmatrix}\right)\right] \geq p_{\boldsymbol{X},y^*}, \tag{16}$$

Importantly, note that Eq. (16) only depends on the function value of h evaluated for matrices whose first row is zero. To use the family of variables analogy discussed in Appendix D.2, Eq. (16) only depends on a subset of variables indexed by $\mathbb{R}^{(N-1)\times D}$. We may thus equivalently solve

$$\min_{h:\mathbb{R}^{(N-1)\times D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{W}\sim\nu_{\boldsymbol{X}'}}\left[h\left(\boldsymbol{W}\right)\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{W}\sim\nu_{\boldsymbol{X}}}\left[h\left(\boldsymbol{W}\right)\right] \geq p_{\boldsymbol{X},y^*}. \tag{17}$$

Solving the lower-dimensional problem. At this point, we could apply the Neyman-Pearson lemma to exactly solve Eq. (17). However, the exact solution has already been derived in prior work. Note that the density function of $\nu_{\boldsymbol{X}}$ is just the product of multiple multivariate normal density functions. It can be expressed as a single multivariate normal distribution with a block-diagonal covariance matrix. If $\boldsymbol{W} \sim \nu_{\boldsymbol{X}}$, then $\mathrm{vec}(\boldsymbol{W}) \sim \mathcal{N}\left(\boldsymbol{m}_{\boldsymbol{X}}, \boldsymbol{\Sigma}\right)$ with $\boldsymbol{m}_{\boldsymbol{X}} \in \mathbb{R}^{(N-1)D}$ and $\boldsymbol{\Sigma} \in \mathbb{R}_+^{(N-1)D\times(N-1)D}$, where

$$\boldsymbol{m}_{\boldsymbol{X}} = \mathrm{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}\right),$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{B} \end{bmatrix}$$

with $\boldsymbol{B} = \sigma^2 \boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T$. Since the function $\mathrm{vec}^{-1} : \mathbb{R}^{(N-1)D} \to \mathbb{R}^{(N-1)\times D}$ that de-vectorizes $\mathrm{vec}(\boldsymbol{W})$ is injective, we may restate our optimization problem from Eq. (17) as an optimization problem over a family of variables indexed by $\mathbb{R}^{(N-1)D}$:

$$\min_{h':\mathbb{R}^{(N-1)D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{m}_{\boldsymbol{X}'},\boldsymbol{\Sigma})}\left[h'\left(\boldsymbol{z}\right)\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{m}_{\boldsymbol{X}},\boldsymbol{\Sigma})}\left[h'\left(\boldsymbol{z}\right)\right] \geq p_{\boldsymbol{X},y^*}. \tag{18}$$

Eq. (18) is the problem of finding a tight certificate for a classifier smoothed with anisotropic Gaussian noise. Its optimal value has been derived in prior work (see [129] and Appendix A of [91]) 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

$$\sqrt{\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right)} = \frac{1}{\sigma}\left|\left|\boldsymbol{\Delta} - \boldsymbol{1}_N\overline{\boldsymbol{\Delta}}\right|\right|_2 \tag{19}$$

First, we use the fact that the inverse of a block-diagonal matrix is also a block-diagonal matrix and insert the definitions of $\boldsymbol{m}_{\boldsymbol{X}'}$, $\boldsymbol{m}_{\boldsymbol{X}}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{B}$ to show that:

$$\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{m}_{\boldsymbol{X}'} - \boldsymbol{m}_{\boldsymbol{X}}\right) \tag{20}$$

$$= \left(\mathrm{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'\right) - \mathrm{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}\right)\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathrm{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'\right) - \mathrm{vec}\left(\boldsymbol{A}_{2:}\boldsymbol{X}\right)\right) \tag{21}$$

$$= \sum_{d=1}^{D}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'_{:,d} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}\right)^T \boldsymbol{B}^{-1}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'_{:,d} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}\right) \tag{22}$$

$$= \sum_{d=1}^{D}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'_{:,d} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}\right)^T \frac{1}{\sigma^2}\left(\boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T\right)^{-1}\left(\boldsymbol{A}_{2:}\boldsymbol{X}'_{:,d} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}\right) \tag{23}$$

$$= \frac{1}{\sigma^2}\sum_{d=1}^{D}\left(\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,d}\right)^T\left(\boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T\right)^{-1}\left(\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,d}\right), \tag{24}$$

$$= \frac{1}{\sigma^2}\sum_{d=1}^{D}\left(\boldsymbol{\Delta}_{:,d}\right)^T\left(\boldsymbol{A}_{2:}\right)^T\left(\boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T\right)^{-1}\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,d}, \tag{25}$$

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

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

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

The matrix is sufficiently simple to be inverted by inspection:

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

Finally, matrix multiplication shows that

$$\left(\boldsymbol{A}_{2:}\right)^T \left(\boldsymbol{A}_{2:}\left(\boldsymbol{A}_{2:}\right)^T\right)^{-1} \boldsymbol{A}_{2:} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_{N,N} = \begin{bmatrix} 1-\frac{1}{N} & -\frac{1}{N} & \dots & -\frac{1}{N} \\ -\frac{1}{N} & 1-\frac{1}{N} & \dots & -\frac{1}{N} \\ \vdots & \vdots & \ddots & \frac{1}{N} \\ -\frac{1}{N} & -\frac{1}{N} & \dots & 1-\frac{1}{N} \end{bmatrix}. \tag{26}$$

Note that Eq. (26) transforms a vector by subtracting its average from all entries. Inserting into Eq. (25) shows that

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

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:

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

D.5 Tight certificate for invariance under rotation in 2D

The following theorem is our tight certificate for rotation invariance in 2D, i.e. invariance under $\mathbb{T} = \left\{ \boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(2) \right\}$. Different from the formulation in in Section 6.3, we also specify its exact dependence on the four parameters $||\boldsymbol{X}||_2$, $||\boldsymbol{\Delta}||$, ϵ_1 and ϵ_2:

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

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \Pr_{\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q}) \leq t] \tag{27}$$

$$\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(2)},\boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q}) \leq t] = p_{\boldsymbol{X},y^*} \tag{28}$$

$$\text{and } \rho(\boldsymbol{q}) = \mathcal{I}_0\left(\sqrt{q_1^2+q_2^2}\right) / \mathcal{I}_0\left(\sqrt{q_3^2+q_4^2}\right), \tag{29}$$

where $\mathcal{I}_0$ is the modified Bessel function of the first kind and order 0, and

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

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

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

Here, $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\mathrm{F}} = \sum_{n=1}^{N} \boldsymbol{A}_n^T \boldsymbol{B}_n^T = \mathrm{vec}(A)^T \mathrm{vec}(B)$ is the Frobenius inner product.

Following our certification methodology, we present a canonical map for invariance under rotation in 2D and then use it to derive our certificate via substitutions, marginalization and application of the Neyman-Pearson lemma. We additionally demonstrate that the worst-case invariant classifier is a Gaussian Haar integration kernel method (see Appendix D.5.3).

D.5.1 Canonical map

For our derivation, recall the definition of canonical maps.

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*

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

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

For rotation invariance, it is more convenient to work with polar coordinates. In the following, let $\lambda : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^{N\times 2}$ with $\lambda(\boldsymbol{r}, \boldsymbol{\theta}) = [\boldsymbol{r} \odot \cos(\boldsymbol{\theta}) \quad \boldsymbol{r} \odot \sin(\boldsymbol{\theta})]$ be the function that maps from polar coordinates into Cartesian coordiantes. Here, $\odot$ is the elementwise product and $\cos$ and $\sin$ are applied element-wise.

First, we define a rotation operator for matrices:

Definition 4. *The matrix rotation operator is the function $\mathrm{rot} : \mathbb{R}^{N\times 2} \times \mathbb{R} \to \mathbb{R}^{N\times 2}$ with*

$$\mathrm{rot}(\boldsymbol{Z}, a) = \boldsymbol{Z}\begin{bmatrix} \cos(a) & -\sin(a) \\ \sin(a) & \cos(a) \end{bmatrix}^T$$

that rotates all rows of matrix $\boldsymbol{Z}$ by angle a.

We use the matrix rotation operator to define our canonical map, which rotates all matrices $\boldsymbol{Z}$ such that their first non-zero row is aligned with the first coordinate axis, i.e. takes on value 0 in the second column.

Lemma 8. *The function* $\gamma(\boldsymbol{Z}) = \mathrm{rot}\left(\boldsymbol{Z}, -\alpha(\boldsymbol{Z})\right)$ *with*

$$\alpha(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } ||\boldsymbol{Z}||_2 = 0 \\ \arctan2\left(\boldsymbol{Z}_{n^*,2}, \boldsymbol{Z}_{n^*,1}\right) & \text{else} \end{cases}$$

where $n^* = \min\{n \mid ||\boldsymbol{Z}_n||_2 > 0\}$ *is the first non-zero row of matrix* $\boldsymbol{Z}$*, is a canonical map for rotation-invariance in* $2D$*.*

Proof. We can distinguish two cases.

Case 1. All entries of $\boldsymbol{Z}$ are 0, i.e. $\boldsymbol{Z} = \boldsymbol{0}_{N,2}$. In this case $||\boldsymbol{Z}||_2 = 0$ and thus $\gamma(\boldsymbol{Z}) = \boldsymbol{Z} = \boldsymbol{0}_{N,2}$. The equivalence class of $\boldsymbol{Z}$ is $[\boldsymbol{Z}]_{\mathbb{T}} = \left\{\boldsymbol{0}_{N,2}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(2)\right\} = \{\boldsymbol{0}_{N,2}\}$. Both criteria Eqs. (3) and (4) for canonical maps are thus fulfilled.

Case 2. The matrix $\boldsymbol{Z}$ has non-zero entries. Let $\boldsymbol{r}, \boldsymbol{\theta}$ be a map of $\boldsymbol{Z}$ in polar coordinates, i.e. $\boldsymbol{Z} = \lambda(\boldsymbol{r}, \boldsymbol{\theta})$. Let n^* be the index of the first non-zero row of $\boldsymbol{Z}$. Recall that rotation in cartesian coordinates corresponds to translation of the polar coordinate angles and thus $\gamma(\boldsymbol{Z}) = \lambda(\boldsymbol{r}, \boldsymbol{\theta} - \boldsymbol{1}_N\theta_{n^*})$.

Criterion Eq. (3) is fulfilled, because by construction of γ we have $\gamma(\boldsymbol{Z}) = \boldsymbol{Z}\boldsymbol{R}^T$ for some $\boldsymbol{R} \in SO(2)$.

Now, consider an arbitrary $\boldsymbol{Z}' \in [\boldsymbol{Z}]_{\mathbb{T}}$. By definition of equivalence class $[\boldsymbol{Z}]_{\mathbb{T}}$, there must be some $a \in \mathbb{R}$ such that $\boldsymbol{Z}' = \lambda\left(\boldsymbol{r}, \boldsymbol{\theta} + a\boldsymbol{1}_N\right)$. Note that rotation does not affect the radii, i.e. n^* is also the index of the first non-zero row of $\boldsymbol{Z}'$. Thus

$$\gamma(\boldsymbol{Z}') = \lambda\left(\boldsymbol{r}, \boldsymbol{\theta} + a\boldsymbol{1}_N - \boldsymbol{1}_N(\theta_{n^*} + a)\right) = \lambda\left(\boldsymbol{r}, \boldsymbol{\theta} - \boldsymbol{1}_N(\theta_{n^*})\right) = \gamma(\boldsymbol{Z}),$$

meaning criterion Eq. (4) is fulfilled as well. □

D.5.2 Proof for certificate

Our canonical map $\gamma(\boldsymbol{Z})$ rotates all rows of an input $\boldsymbol{Z} \in \mathbb{R}^{N\times D}$ such that the first non-zero row is aligned with the first coordinate-axis. Similar to our proof for translation invariance, we will use substitutions to state our problem in an alternative parameter space that specifies the angle θ_1 of the first non-zero row and all angle differences to this row, i.e. $\theta_n - \theta_1$, as well as the radii $\boldsymbol{r} \in \mathbb{R}_+^N$. We then find the marginal distribution over the representatives in this alternative parameter space and apply the Neyman-Pearson lemma. Finally, we transform our solution back into Cartesian coordinates and use the behavior of multivariate Normal distributions under affine transformations to prove our result.

Because the calculations for deriving the marginal distribution are somewhat more arduous, we shall handle them separately in Lemma 9, as to interfere with the rest of the proof.

Lemma 9. *Consider two multivariate normal distributions* $\mathcal{N}(\boldsymbol{M}_{:,1}, \boldsymbol{\Sigma}), \mathcal{N}(\boldsymbol{M}_{:,2}, \boldsymbol{\Sigma})$ *with* $\boldsymbol{M} \in \mathbb{R}^{N\times 2}$ *that share the same symmetric covariance matrix* $\boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{N\times D}$*. Further consider arbitrary* $\boldsymbol{\phi} \in \mathbb{R}^N$ *and* $\boldsymbol{r} \in \mathbb{R}^N$*. Define* $\boldsymbol{\xi} : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ *with* $\boldsymbol{\xi}(\alpha, \boldsymbol{\phi}) = \boldsymbol{\phi} + \boldsymbol{1}_N\alpha$*. Then*

$$\int_{[0,2\pi]} \mathcal{N}\left(\boldsymbol{r} \odot \cos\left(\boldsymbol{\xi}(\alpha, \boldsymbol{\phi})\right) \mid \boldsymbol{M}_{:,1}, \boldsymbol{\Sigma}\right) \mathcal{N}\left(\boldsymbol{r} \odot \sin\left(\boldsymbol{\xi}(\alpha, \boldsymbol{\phi})\right) \mid \boldsymbol{M}_{:,2}, \boldsymbol{\Sigma}\right)\, d\alpha$$

$$\propto \exp\left(-\frac{1}{2}\mathrm{vec}\left(\boldsymbol{M}\right)^T \begin{bmatrix} \boldsymbol{\Sigma}^{-1} & \boldsymbol{0} \\ 0 & \boldsymbol{\Sigma}^{-1} \end{bmatrix} \mathrm{vec}\left(\boldsymbol{M}\right)\right) \mathcal{I}_0\left(||\kappa\left(\lambda(r, \boldsymbol{\phi}), \boldsymbol{M}, \boldsymbol{\Sigma}\right)||_2\right)$$

$$\text{with } \kappa\left(\boldsymbol{Z}, \boldsymbol{M}, \boldsymbol{\Sigma}\right) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} & \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2} & -\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{Z}_{:,1} \\ \boldsymbol{Z}_{:,2} \end{bmatrix}.$$

where $\cos$ *and* $\sin$ *are applied element-wise,* $\odot$ *is the element-wise product and* $\propto$ *absorbs factors that are constant in* $\boldsymbol{M}$*. Note that* $\boldsymbol{w} \in \mathbb{R}^2$ *and that* $\mathcal{I}_0$ *is the modified Bessel function of the first kind and order* 0*.*

Proof. We begin by identifying terms that that are constant in both $\boldsymbol{M}$ and α and can thus be eliminated, as well as terms that are constant in α and can thus be moved outside the integral. We first eliminate the normalization constants:

$$\mathcal{N}(\boldsymbol{r} \odot \cos\left(\boldsymbol{\xi}(\alpha, \boldsymbol{\phi})\right) \mid \boldsymbol{M}_{:,1}, \boldsymbol{\Sigma})\mathcal{N}(\boldsymbol{r} \odot \sin\left(\boldsymbol{\xi}(\alpha, \boldsymbol{\phi})\right) \mid \boldsymbol{M}_{:,2}, \boldsymbol{\Sigma})$$

$$\propto \exp\left(\sum_{n=1}^{5} c_n\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\right)$$

with

$$
\begin{aligned}
c_1\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) &= -\frac{1}{2}\boldsymbol{M}_{:,1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} + \boldsymbol{M}_{:,2}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2},\\
c_2\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) &= \left(\boldsymbol{r}\odot\cos\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1},\\
c_3\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) &= \left(\boldsymbol{r}\odot\sin\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2},\\
c_4\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\boldsymbol{\Sigma}_{n,m}^{-1}r_n r_m\cos\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)_n\cos\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)_m,\\
c_5\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\boldsymbol{\Sigma}_{n,m}^{-1}r_n r_m\sin\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)_n\sin\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)_m.
\end{aligned}
$$

Note that c_4 and $c5$ correspond to $-\frac{1}{2}\cos\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)^T\boldsymbol{\Sigma}^{-1}\cos\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)$ and $-\frac{1}{2}\sin\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)^T\boldsymbol{\Sigma}^{-1}\sin\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)$, respectively.

c_1 is constant in α and can thus be moved outside the integral.

c_2 and c_3 are not constant in α and must therefore remain inside the integral.

$c_4 + c_5$ can be shown to be constant w.r.t. both $\boldsymbol{M}$ and α and can thus be eliminated. By the trigonometric identity $\cos(a)\cos(b) + \sin(a)\sin(b) = \cos(a-b)$, we have

$$
c_4\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) + c_5\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\boldsymbol{\Sigma}_{n,m}^{-1}r_n r_m\cos\left(\xi(\alpha,\boldsymbol{\phi})_n - \xi(\alpha,\boldsymbol{\phi})_m\right).
$$

Recall that $\boldsymbol{\xi}(\alpha,\boldsymbol{\phi}) = \boldsymbol{\phi} + \mathbf{1}_N\alpha$, i.e. "$+\alpha$" appears in every row. Subtraction thus eliminates α for all combinations of n and m.

Based on these observations, we may conclude that

$$
\int_{[0,2\pi]} \mathcal{N}\left(\boldsymbol{r}\odot\cos\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right) \mid \boldsymbol{M}_{:,1}, \boldsymbol{\Sigma}\right)\mathcal{N}\left(\boldsymbol{r}\odot\sin\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right) \mid \boldsymbol{M}_{:,2}, \boldsymbol{\Sigma}\right)\, d\alpha \tag{30}
$$

$$
\propto \exp\left(c_1\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\right)\int_{[0,2\pi]} \exp\left(c_2\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) + c_3\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\right)\, d\alpha. \tag{31}
$$

Next, we use the trigonometric identites $\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$ and $\sin(a+b) = \sin(a)\cos(b) + \cos(a)\sin(b)$ to move α outside the inner products of c_2 and c_3:

$$
\begin{aligned}
&c_2\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\\
&= \left(\boldsymbol{r}\odot\cos\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1}\\
&= \left(\boldsymbol{r}\odot\cos\left(\mathbf{1}_N\alpha + \boldsymbol{\phi}\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1}\\
&= \left(\boldsymbol{r}\odot\cos\left(\mathbf{1}_N\alpha\right)\odot\cos\left(\boldsymbol{\phi}\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} - \left(\boldsymbol{r}\odot\sin\left(\mathbf{1}_N\alpha\right)\odot\sin\left(\boldsymbol{\phi}\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1}\\
&= \cos(\alpha)\left(\boldsymbol{r}\odot\cos\left(\boldsymbol{\phi}\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} - \sin(\alpha)\left(\boldsymbol{r}\odot\sin\left(\boldsymbol{\phi}\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1}\\
&= \cos(\alpha)\lambda\left(\boldsymbol{r},\boldsymbol{\phi}\right)_{:,1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} - \sin(\alpha)\lambda\left(\boldsymbol{r},\boldsymbol{\phi}\right)_{:,2}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1}
\end{aligned} \tag{32}
$$

and

$$
\begin{aligned}
&c_3\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\\
&= \left(\boldsymbol{r}\odot\sin\left(\boldsymbol{\xi}(\alpha,\boldsymbol{\phi})\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2}\\
&= \left(\boldsymbol{r}\odot\sin\left(\mathbf{1}_N\alpha + \boldsymbol{\phi}\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2}\\
&= \sin(\alpha)\left(\boldsymbol{r}\odot\cos\left(\boldsymbol{\phi}\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2} + \cos(\alpha)\left(\boldsymbol{r}\odot\sin\left(\boldsymbol{\phi}\right)\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2}\\
&= \sin(\alpha)\lambda\left(\boldsymbol{r},\boldsymbol{\phi}\right)_{:,1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2} + \cos(\alpha)\lambda\left(\boldsymbol{r},\boldsymbol{\phi}\right)_{:,2}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2}.
\end{aligned} \tag{33}
$$

Stacking the $\boldsymbol{\Sigma}^{-1}\boldsymbol{M}$ terms, we can express $c_2 + c_3$ as follows:

$$\begin{aligned}
&c_2\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) + c_3\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) \\
=&cos(\alpha)\begin{bmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1}\\ \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2}\end{bmatrix}^T\begin{bmatrix}\lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,1}\\ \lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,2}\end{bmatrix} + sin(\alpha)\begin{bmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2}\\ -\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1}\end{bmatrix}^T\begin{bmatrix}\lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,1}\\ \lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,2}\end{bmatrix} \\
:=&\cos(\alpha)\eta_1 + \sin(\alpha)\eta_2.
\end{aligned}$$

We can now substitute our results into Eq. (31) and simplify the integral using the fact that $\cos(\alpha)\eta_1 + \sin(\alpha)\eta_2 = \cos(\alpha+\omega)\sqrt{\eta_1^2+\eta_2^2)}$ for some $\omega \in [0, 2\pi]$ (see also Eq. 29 of [130]).

$$\begin{aligned}
&\exp\left(c_1\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\right) \int\limits_{[0,2\pi]} \exp\left(c_2\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right) + c_3\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\right)\ d \\
=&\exp\left(c_1\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\right) \int\limits_{[0,2\pi]} \exp\left(\cos(\alpha)\eta_1 + \sin(\alpha)\eta_2\right)\ d\alpha \\
=&\exp\left(c_1\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\right) \int\limits_{[0,2\pi]} \exp\left(\cos(\alpha+\omega)\sqrt{\eta_1^2+\eta_2^2)}\right)\ d\alpha \\
=&\exp\left(c_1\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\right) \int\limits_{[0,2\pi]} \exp\left(\cos(\alpha)\sqrt{\eta_1^2+\eta_2^2)}\right)\ d\alpha \\
=&\exp\left(c_1\left(\alpha, \boldsymbol{M}, \boldsymbol{r}, \boldsymbol{\phi}\right)\right)\frac{\pi}{2}\mathcal{I}_0\left(\sqrt{\eta_1^2+\eta_2^2)}\right),
\end{aligned}$$

where the second to last equality is due to the fact that we integrate over an entire period of $\cos$ and the last equality uses the integral map of the modified Bessel function of the first kind with order 0 (see Eq. 10.32.1 of [131]):

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

Restating $\sqrt{\eta_1^2+\eta_2^2)}$ as

$$\left|\left|\begin{bmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} & \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2}\\ \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2} & -\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1}\end{bmatrix}^T\begin{bmatrix}\lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,1}\\ \lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,2}\end{bmatrix}\right|\right|_2$$

concludes our proof. □

Applying the canonical map. Now, we can turn to deriving the certificate itself. Applying Theorem 4 with our canonical map from Lemma 8 shows that

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

Like in our previous proof, we notice that the above optimization problem does not depend on the specific class predicted by h, but only on whether it is y^* or y. We thus formulate the equivalent problem

$$\min_{h:\mathbb{R}^{N\times D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\gamma(\boldsymbol{Z}))\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right] \geq p_{\boldsymbol{X},y^*}, \tag{34}$$

where $h(\boldsymbol{Z}) = 1$ indicates that $\boldsymbol{Z}$ is classified as y^*.

Substitution. We begin by going from Cartesian to polar coordinates: As defined in the previous section, let $\lambda : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^{N\times 2}$ with $\lambda\left(\boldsymbol{r}, \boldsymbol{\theta}\right) = [\boldsymbol{r}\odot\cos(\boldsymbol{\theta}) \quad \boldsymbol{r}\odot\sin(\boldsymbol{\theta})]$ be the function that maps from polar coordinates into cartesian coordiantes.

$$\mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right] \tag{35}$$

$$= \int_{\mathbb{R}^{N\times D}} h(\gamma(\boldsymbol{Z}))\mu_{\boldsymbol{X}}(\boldsymbol{Z})\ d\boldsymbol{Z} \tag{36}$$

$$= \int_{\mathbb{R}^N}\left(\prod_{n=1}^{N} r_n\right)\int_{[0,2\pi]^N} h(\gamma(\lambda(\boldsymbol{r},\boldsymbol{\theta})))\mu_{\boldsymbol{X}}\left(\lambda(\boldsymbol{r},\boldsymbol{\theta})\right))\ d\boldsymbol{\theta}\ d\boldsymbol{r} \tag{37}$$

The r_n in the last equality are due to the fact that λ transforms N rows from polar to Cartesian coordinates, with each transformation having a Jacobian determinant r.

Now, we would like to actually apply our canonical map γ, which performs a translation of the angles $\boldsymbol{\theta}$. Recall from Lemma 8 that γ requires a case distinction to find the first non-zero row of $\lambda(\boldsymbol{r}, \boldsymbol{\theta})$. However, due to the factor $\prod_{n=1}^{N} r_n$, the entire integrand becomes zero whenever any row of $\lambda(\boldsymbol{r}, \boldsymbol{\theta})$ is zero. We can therefore always assume that we are in the case with $r_1 > 0$, i.e. rotate by θ_1, without influencing the value of the integral. Thus

$$\int_{\mathbb{R}^N} \left(\prod_{n=1}^{N} r_n\right) \int_{[0,2\pi]^N} h(\gamma(\lambda(\boldsymbol{r}, \boldsymbol{\theta}))) \mu_{\boldsymbol{X}} \left(\lambda(\boldsymbol{r}, \boldsymbol{\theta})\right) d\boldsymbol{\theta}\, d\boldsymbol{r} \tag{38}$$

$$= \int_{\mathbb{R}^N} \left(\prod_{n=1}^{N} r_n\right) \int_{[0,2\pi]^N} h\left(\lambda(\boldsymbol{r}, \boldsymbol{\theta} - \mathbf{1}_N \theta_1)\right) \mu_{\boldsymbol{X}} \left(\lambda(\boldsymbol{r}, \boldsymbol{\theta})\right) d\boldsymbol{\theta}\, d\boldsymbol{r}. \tag{39}$$

Next, we apply a similar substitution to the one used in our translation certificate: Let $\boldsymbol{\psi} = \boldsymbol{A}\boldsymbol{\theta}$ with $\boldsymbol{A} = \begin{bmatrix} 1 & \mathbf{0}_{N-1}^T \\ -\mathbf{1}_{N-1} & \mathbf{I}_{N-1} \end{bmatrix}$ and $\boldsymbol{A}^{-1} = \begin{bmatrix} 1 & \mathbf{0}_{N-1} \\ \mathbf{1}_{N-1} & \mathbf{I}_N \end{bmatrix}$. Note that $\boldsymbol{A}^{-1}\left(\boldsymbol{\psi} - \mathbf{1}_N \psi_1\right) = \begin{bmatrix} 0 \\ \boldsymbol{\psi}_{2:} \end{bmatrix}$, where $\boldsymbol{\psi}_{2:}$ are the last $N-1$ entries of $\boldsymbol{\psi}$. In other words: Applying canonical map γ is equivalent to setting ψ_1 to 0. Further note that $\det\left(\boldsymbol{A}^{-1}\right) = 1$ and that we have to adjust the integration boundaries, because we are now integrating over angle differences. Thus

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}} \left[h(\gamma(\boldsymbol{Z}))\right] \\ &= \int_{\mathbb{R}^N} \left(\prod_{n=1}^{N} r_n\right) \int_{[0,2\pi]} \int_{[-\psi_1, 2\pi - \psi_1]^{N-1}} h\left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\psi}_{2:} \end{bmatrix}\right)\right) \mu_{\boldsymbol{X}} \left(\lambda(\boldsymbol{r}, \boldsymbol{A}^{-1}\boldsymbol{\psi})\right) d\boldsymbol{\psi}_{2:} d\psi_1\, d\boldsymbol{r}. \end{aligned}$$

Marginalization. The function λ that maps from polar into Cartesian coordinates is 2π-periodic w.r.t. its second argument and, by definition, we have $\boldsymbol{A}^{-1}\boldsymbol{\psi} = \begin{bmatrix} 0 \\ \boldsymbol{\psi}_{2:} \end{bmatrix} + \mathbf{1}_{N-1}\psi_1$. This means that for any fixed $\boldsymbol{\psi}_1$, integrating $\boldsymbol{\psi}_{2:}$ over $[-\psi_1, 2\pi - \psi_1]^{N-1}$ is equivalent to integrating over $[0, 2\pi]$. Further note that the term involving h does not depend on $\boldsymbol{\psi}$, meaning it can be moved out.

$$= \int_{\mathbb{R}^N} \left(\prod_{n=1}^{N} r_n\right) \int_{[0,2\pi]} \int_{[0,2\pi]^{N-1}} h\left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\psi}_{2:} \end{bmatrix}\right)\right) \mu_{\boldsymbol{X}} \left(\lambda(\boldsymbol{r}, \boldsymbol{A}^{-1}\boldsymbol{\psi})\right) d\boldsymbol{\psi}_{2:} d\psi_1\, d\boldsymbol{r} \tag{40}$$

$$= \int_{\mathbb{R}^N} \int_{[0,2\pi]^{N-1}} h\left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\psi}_{2:} \end{bmatrix}\right)\right) \left(\prod_{n=1}^{N} r_n\right) \int_{[0,2\pi]} \mu_{\boldsymbol{X}} \left(\lambda(\boldsymbol{r}, \boldsymbol{A}^{-1}\boldsymbol{\psi})\right) d\psi_1 d\boldsymbol{\psi}_{2:}\, d\boldsymbol{r} \tag{41}$$

This demonstrates that $\mathbb{E}_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}} \left[h(\gamma(\boldsymbol{Z}))\right] = \mathbb{E}_{\boldsymbol{r}, \boldsymbol{\psi}_{2:} \sim \nu_{\boldsymbol{X}}} \left[h\left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\psi}_{2:} \end{bmatrix}\right)\right)\right]$ with $\nu_{\boldsymbol{X}}$ being a distribution over $\mathbb{R}^N \times [0, 2\pi]^{N-1}$ with density function

$$\nu_{\boldsymbol{X}}(\boldsymbol{r}, \boldsymbol{\psi}_{2:}) = \left(\prod_{n=1}^{N} r_n\right) \int_{[0,2\pi]} \mu_{\boldsymbol{X}} \left(\lambda(\boldsymbol{r}, \boldsymbol{A}^{-1}\boldsymbol{\psi})\right) d\psi_1 \tag{42}$$

$$\left(\prod_{n=1}^{N} r_n\right) \int_{[0,2\pi]} \mu_{\boldsymbol{X}} \left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\psi}_{2:} \end{bmatrix} + \mathbf{1}_N \psi_1\right)\right) d\psi_1 \tag{43}$$

Inserting into our original optimization problem Eq. (34) and replacing $\boldsymbol{\psi}_{2:}$ with $\boldsymbol{\omega} \in [0, 2\pi]^{N-1}$ to clarify that the problem no longer depends on the entire vector of angles $\boldsymbol{\psi}$ results in

$$\begin{aligned} \min_{h: \mathbb{R}^{N \times D} \to \{0,1\}} \; & \mathbb{E}_{\boldsymbol{r}, \boldsymbol{\omega} \sim \nu_{\boldsymbol{X}'}} \left[h\left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\omega} \end{bmatrix}\right)\right)\right] \\ \text{s.t. } & \mathbb{E}_{\boldsymbol{r}, \boldsymbol{\omega} \sim \nu_{\boldsymbol{X}}} \left[h\left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\omega} \end{bmatrix}\right)\right)\right] \geq p_{\boldsymbol{X}, y^*}. \end{aligned} \tag{44}$$

Note that λ is an injective function from $\mathbb{R}^N \times [0, 2\pi]^N$ into $\mathbb{R}^{N\times D}$, meaning that each pair $(\boldsymbol{r}, \omega)$ in the support of $\nu_{\boldsymbol{X}}$ and $\nu_{\boldsymbol{X}'}$ corresponds to a distinct function value $h(\boldsymbol{Z})$. We may thus equivalently find the worst-case classifier in $\mathbb{R}^N \times [0, 2\pi]^{N-1}$:

$$\min_{h:\mathbb{R}^N\times[0,2\pi]^{N-1}\to\{0,1\}} \mathbb{E}_{\boldsymbol{r},\boldsymbol{\omega}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{r},\boldsymbol{\omega})\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{r},\boldsymbol{\omega}\sim\mu_{\boldsymbol{X}}}\left[h(\boldsymbol{r},\boldsymbol{\omega})\right] \geq p_{\boldsymbol{X},y^*}. \tag{45}$$

Solving the lower-dimensional problem. Applying the Neyman-Pearson lemma (Lemma 2) to the above problem shows that its optimal value is

$$\mathbb{E}_{\boldsymbol{r},\boldsymbol{\omega}\sim\mu_{\boldsymbol{X}'}}\left[h^*(\boldsymbol{r},\boldsymbol{\omega})\right] \tag{46}$$

$$\text{with } h^*(\boldsymbol{r},\boldsymbol{\omega}) = \mathbb{1}\left[\frac{\nu_{\boldsymbol{X}'}(\boldsymbol{r},\boldsymbol{\omega})}{\nu_{\boldsymbol{X}}(\boldsymbol{r},\boldsymbol{\omega})} \leq t\right] \text{ and } t \in \mathbb{R} \text{ s.t. } \mathbb{E}_{\boldsymbol{r},\boldsymbol{\omega}\sim\mu_{\boldsymbol{X}}}\left[h^*(\boldsymbol{r},\boldsymbol{\omega})\right] = p_{\boldsymbol{X},y^*}. \tag{47}$$

Transformation into Cartesian coordinates. To conclude our proof, we transform the expectations in Eqs. (46) and (47) over this lower-dimensional parameter space back to our Cartesian space $\mathbb{R}^{N\times D}$. Before doing so, we take a minor precaution to make the integration bounds easier to work with. We define a function $\hat{h} : \mathbb{R}^N \times \mathbb{R}^{N-1} \to \{0, 1\}$ that allows h to be evaluated for arbitrary polar coordiante angles, not just those in $[0, 2\pi]$: For this, we let $\hat{h}(\boldsymbol{r}, \boldsymbol{\omega}) = h^*(\boldsymbol{r}, (\chi(\omega))$, where $\chi(\omega) = \omega + 2k\pi$ s.t. $\omega + 2k \in [0, 2\pi]$. Clearly $\mathbb{E}_{\boldsymbol{r},\boldsymbol{\omega}\sim\mu_{\boldsymbol{X}'}}\left[\hat{h}(\boldsymbol{r},\boldsymbol{\omega})\right] = \mathbb{E}_{\boldsymbol{r},\boldsymbol{\omega}\sim\mu_{\boldsymbol{X}'}}\left[h^*(\boldsymbol{r},\boldsymbol{\omega})\right]$. Therefore

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{r},\boldsymbol{\omega}\sim\mu_{\boldsymbol{X}}}\left[h^*(\boldsymbol{r},\boldsymbol{\omega})\right] \\
&= \int_{\mathbb{R}^N\times[0,2\pi]^{N-1}} \hat{h}(\boldsymbol{r},\boldsymbol{\omega})\nu_{\boldsymbol{X}}(\boldsymbol{r},\boldsymbol{\omega})\, d\boldsymbol{\omega}\, d\boldsymbol{r} \\
&= \int_{\mathbb{R}^N\times[0,2\pi]^{N-1}} \hat{h}(\boldsymbol{r},\boldsymbol{\omega}) \left(\prod_{n=1}^N r_n\right) \int_{[0,2\pi]} \mu_{\boldsymbol{X}}\left(\lambda\left(\boldsymbol{r}, \begin{bmatrix}0\\ \boldsymbol{\omega}\end{bmatrix} + \mathbf{1}_N\beta\right)\right)\, d\beta\, d\boldsymbol{\omega}\, d\boldsymbol{r} \\
&= \int_{\mathbb{R}^N\times[0,2\pi]^{N}} \hat{h}(\boldsymbol{r},\boldsymbol{\omega}) \left(\prod_{n=1}^N r_n\right) \mu_{\boldsymbol{X}}\left(\lambda\left(\boldsymbol{r}, \begin{bmatrix}0\\ \boldsymbol{\omega}\end{bmatrix} + \mathbf{1}_N\beta\right)\right)\, d\boldsymbol{\omega} d\beta\, d\boldsymbol{r}
\end{aligned}$$

The second equality uses the definition of $\nu_{\boldsymbol{X}}$ from Eq. (43) and the last equality combines the integration bounds of the two integrals.

We now apply the substitution $\boldsymbol{\theta} = \boldsymbol{A}^{-1}\begin{bmatrix}\beta\\ \boldsymbol{\omega}\end{bmatrix}$ with $\boldsymbol{A}^{-1} = \begin{bmatrix}1 & \mathbf{0}_{N-1}\\ \mathbf{1}_{N-1} & \mathbf{I}_N\end{bmatrix}$ to transform back into standard polar coordinates. We then simplify the integration bounds by using the fact that, for any fixed θ_1 all terms of the integrand are 2π-periodic in $\boldsymbol{\theta}_{2:}$ – similar to how we simplified the integration bounds when transforming into our alternative parameter space in Eqs. (40) and (41):

$$= \int_{\mathbb{R}^N\times[0,2\pi]\times[\theta_1,2\pi+\theta_1]^{N-1}} \hat{h}(\boldsymbol{r},\boldsymbol{\theta}_{2:} - \mathbf{1}_{N-1}\theta_1) \left(\prod_{n=1}^N r_n\right) \mu_{\boldsymbol{X}}\left(\lambda\left(\boldsymbol{r},\boldsymbol{\theta}\right)\right)\, d\theta_{2:}\, d\theta_1\, d\boldsymbol{r} \tag{48}$$

$$= \int_{\mathbb{R}^N\times[0,2\pi]^{N}} \hat{h}(\boldsymbol{r},\boldsymbol{\theta}_{2:} - \mathbf{1}_{N-1}\theta_1) \left(\prod_{n=1}^N r_n\right) \mu_{\boldsymbol{X}}\left(\lambda\left(\boldsymbol{r},\boldsymbol{\theta}\right)\right)\, d\boldsymbol{\theta}\, d\boldsymbol{r}. \tag{49}$$

Finally, recall from Eq. (47) that $\hat{h}$ depends on the likelihood ratios of the marginal distributions $\nu_{\boldsymbol{X}'}$ and $\nu_{\boldsymbol{X}}$. Lemma 9 provides a closed-form expression for these likelihood ratios, up to constant factors in $\boldsymbol{X}'$ and $\boldsymbol{X}$. This closed-form expression lets us express the likelihood ratio as a function of

$\lambda(\boldsymbol{r}, \boldsymbol{\theta})$.

$$
\begin{aligned}
&\hat{h}(\boldsymbol{r}, \boldsymbol{\theta}_{2:} - \mathbf{1}_{N-1}\theta_1) = 1 \\
\iff &\frac{\left(\prod_{n=1}^{N} r_n\right) \int\limits_{[0,2\pi]} \mu_{\boldsymbol{X}'} \left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\theta}_{2:} - \mathbf{1}_{N-1}\theta_1 \end{bmatrix} + \mathbf{1}_N \beta\right)\right) \, d\beta}{\left(\prod_{n=1}^{N} r_n\right) \int\limits_{[0,2\pi]} \mu_{\boldsymbol{X}} \left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\theta}_{2:} - \mathbf{1}_{N-1}\theta_1 \end{bmatrix} + \mathbf{1}_N \beta\right)\right) \, d\beta} \leq t \\
\iff &\frac{\int\limits_{[0,2\pi]} \mu_{\boldsymbol{X}'} \left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\theta}_{2:} - \mathbf{1}_{N-1}\theta_1 \end{bmatrix} + \mathbf{1}_N \theta_1 + \mathbf{1}_N \beta\right)\right) \, d\beta}{\int\limits_{[0,2\pi]} \mu_{\boldsymbol{X}} \left(\lambda\left(\boldsymbol{r}, \begin{bmatrix} 0 \\ \boldsymbol{\theta}_{2:} - \mathbf{1}_{N-1}\theta_1 \end{bmatrix} + \mathbf{1}_N \theta_1 + \mathbf{1}_N \beta\right)\right) \, d\beta} \leq t \\
\iff &\frac{\int\limits_{[0,2\pi]} \mu_{\boldsymbol{X}'} \left(\lambda\left(\boldsymbol{r}, \boldsymbol{\theta} + \mathbf{1}_N \beta\right)\right) \, d\beta}{\int\limits_{[0,2\pi]} \mu_{\boldsymbol{X}} \left(\lambda\left(\boldsymbol{r}, \boldsymbol{\theta} + \mathbf{1}_N \beta\right)\right) \, d\beta} \leq t \\
\iff &c \cdot \frac{\mathcal{I}_0\left(||\kappa\left(\lambda(\boldsymbol{r}, \boldsymbol{\theta}), \boldsymbol{X}', \sigma^2 \mathbf{I}_N\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\lambda(\boldsymbol{r}, \boldsymbol{\theta}), \boldsymbol{X}, \sigma^2 \mathbf{I}_N\right)||_2\right)} \leq t
\end{aligned}
$$

with $\kappa : \mathbb{R}^{N \times D} \times \mathbb{R}^{N \times D} \times \mathbb{R}^{N \times N} \to \mathbb{R}^2$ defined as in Lemma 9, $\mathcal{I}_0$ being the modified Bessel function of the first kind and order 0 and c being constant in $\boldsymbol{r}$ and $\boldsymbol{\theta}$. In the second equality, we have cancelled out the the r_n and used the fact that λ is 2π-periodic in its second argument to add $\mathbf{1}_N \theta_1$ without changing the value of the integral.

Inserting back into Eq. (49) and transforming into Cartesian coordinates shows that

$$
\mathbb{E}_{\boldsymbol{r}, \boldsymbol{\omega} \sim \mu_{\boldsymbol{X}}}\left[h^*(\boldsymbol{r}, \boldsymbol{\omega})\right] = \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}}\left[c \cdot \frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}', \sigma^2 \mathbf{I}_N\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2 \mathbf{I}_N\right)||_2\right)} \leq t\right].
$$

Inserting into the optimal solution of our lower-dimensional optimization problem (Eqs. (46) and (47)) shows that

$$
\min_{h \in \mathbb{H}_{\mathrm{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right] = \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}}\left[c \cdot \frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}', \sigma^2 \mathbf{I}_N\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2 \mathbf{I}_N\right)||_2\right)} \leq t\right] \tag{50}
$$

$$
\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}}\left[c \cdot \frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}', \sigma^2 \mathbf{I}_N\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2 \mathbf{I}_N\right)||_2\right)} \leq t\right]. \tag{51}
$$

Note that we can absorb the constant c into the threshold t.

$$
\min_{h \in \mathbb{H}_{\mathrm{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right] = \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}}\left[\frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}', \sigma^2 \mathbf{I}_N\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2 \mathbf{I}_N\right)||_2\right)} \leq t'\right] \tag{52}
$$

$$
\text{with } t' \in \mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}}}\left[\frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}', \sigma^2 \mathbf{I}_N\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2 \mathbf{I}_N\right)||_2\right)} \leq t'\right]. \tag{53}
$$

If t' fulfills the conditions in Eq. (53), then $t = \frac{t'}{c}$ fulfills the conditions in Eq. (51).

Likelihood ratio distribution. Finally, recall from Lemma 9 that

$$
\kappa\left(\boldsymbol{Z}, \boldsymbol{M}, \boldsymbol{\Sigma}\right) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} & \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2} & -\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{Z}_{:,1} \\ \boldsymbol{Z}_{:,2} \end{bmatrix}.
$$

Inside the probability terms appearing in Eqs. (52) and (53), we have

$$
\kappa\left(\boldsymbol{Z}, \boldsymbol{X}', \sigma^2 \mathbf{I}_N\right) = \frac{1}{\sigma^2} \begin{bmatrix} \boldsymbol{X}'_{:,1} & \boldsymbol{X}'_{:,2} \\ \boldsymbol{X}'_{:,2} & -\boldsymbol{X}'_{:,1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{Z}_{:,1} \\ \boldsymbol{Z}_{:,2} \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} \mathrm{vec}(\boldsymbol{X}')^T \\ \mathrm{vec}\left(\boldsymbol{X}'\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T\right)^T \end{bmatrix} \mathrm{vec}(\boldsymbol{Z}),
$$

and

$$\kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2 \mathbf{I}_N\right) = \frac{1}{\sigma^2}\begin{bmatrix}\boldsymbol{X}_{:,1} & \boldsymbol{X}_{:,2} \\ \boldsymbol{X}_{:,2} & -\boldsymbol{X}_{:,1}\end{bmatrix}^T \begin{bmatrix}\boldsymbol{Z}_{:,1} \\ \boldsymbol{Z}_{:,2}\end{bmatrix} = \frac{1}{\sigma^2}\begin{bmatrix}\operatorname{vec}(\boldsymbol{X})^T \\ \operatorname{vec}\left(\boldsymbol{X}\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T\right)^T\end{bmatrix}\operatorname{vec}(\boldsymbol{Z}),$$

where $\boldsymbol{R}\left(-\frac{\pi}{2}\right)$ is the rotation matrix that rotates clockwise by angle $\frac{\pi}{2}$.

In Eq. (52), the random variable $\boldsymbol{Z}$ follows distribution $\mu_{\boldsymbol{X}'}$. Thus, $\operatorname{vec}(\boldsymbol{Z}) \sim \mathcal{N}\left(\operatorname{vec}(\boldsymbol{X}'), \sigma^2 \mathbf{I}_{2N}\right)$. Therefore, the random variable $q \in \mathbb{R}^4 = \begin{bmatrix}\kappa\left(\boldsymbol{Z}, \boldsymbol{X}', \sigma^2\mathbf{I}_N\right) \\ \kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2\mathbf{I}_N\right)\end{bmatrix}$, which is an affine transformation of $\operatorname{vec}(\boldsymbol{Z})$ follows distribution $\mathcal{N}\left(\boldsymbol{m}^{(1)}, \boldsymbol{S}\right)$ with $\boldsymbol{m}^{(1)} = \boldsymbol{C}\boldsymbol{X}'$ and $\boldsymbol{S} = \sigma^2\boldsymbol{C}\boldsymbol{C}^T$, where

$$\boldsymbol{C} = \frac{1}{\sigma^2}\begin{bmatrix}\operatorname{vec}(\boldsymbol{X}')^T \\ \operatorname{vec}\left(\boldsymbol{X}'\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T\right)^T \\ \operatorname{vec}(\boldsymbol{X})^T \\ \operatorname{vec}\left(\boldsymbol{X}\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T\right)^T\end{bmatrix}.$$

Similarly, in Eq. (53), the random variable $\boldsymbol{Z}$ follows distribution $\mu_{\boldsymbol{X}}$. Therefore, the random variable $q \in \mathbb{R}^4 = \begin{bmatrix}\kappa\left(\boldsymbol{Z}, \boldsymbol{X}', \sigma^2\mathbf{I}_N\right) \\ \kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2\mathbf{I}_N\right)\end{bmatrix}$, which is an affine transformation of $\operatorname{vec}(\boldsymbol{Z})$ follows distribution $\mathcal{N}\left(\boldsymbol{m}^{(2)}, \boldsymbol{S}\right)$ with $\boldsymbol{m}^{(2)} = \boldsymbol{C}\boldsymbol{X}$ and $\boldsymbol{S} = \sigma^2\boldsymbol{C}\boldsymbol{C}^T$.

Computing the matrix-vector and matrix-matrix products and using the definition that $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$ shows that

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

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

with $\epsilon_1 = \langle \boldsymbol{X}, \boldsymbol{\Delta}\rangle_{\mathrm{F}}$, $\epsilon_2 = \left\langle \boldsymbol{X}\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T, \boldsymbol{\Delta}\right\rangle_{\mathrm{F}}$. This concludes our proof.

D.5.3 Worst-case classifier

In Section 6.3, we claimed that the worst-case classifier, i.e. the minimizer of $\min_{h\in\mathbb{H}_{\mathrm{T}}} \Pr_{\boldsymbol{Z}\sim\mu'_{\boldsymbol{X}}}\left[h(\boldsymbol{Z}) = y^*\right]$ was a Gaussian Haar integration kernel method [16], i.e. a function that predicts y^* if and only if $\frac{\beta(\boldsymbol{X}',\boldsymbol{Z})}{\beta(\boldsymbol{X},\boldsymbol{Z})} \leq t$ with $\beta(\boldsymbol{X}, \boldsymbol{Z}) = \int_0^{2\pi} \exp\left(\frac{1}{\sigma^2}\left\langle \boldsymbol{X}, \boldsymbol{Z}\boldsymbol{R}(\alpha)^T\right\rangle_{\mathrm{F}}\right)\, d\alpha.$, where $\boldsymbol{R}(\alpha) \in SO(2)$ is the matrix that rotates counter-clockwise by angle α.

Here, we verify this claim. In fact, we have already implicitly done so in the derivation of our certificate. Recall from Eq. (52) that

$$\min_{h\in\mathbb{H}_{\mathrm{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right] = \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[\frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}', \sigma^2\mathbf{I}_N\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2\mathbf{I}_N\right)||_2\right)} \leq t\right]$$

for some $t \in \mathbb{R}$. Let $\boldsymbol{r}, \boldsymbol{\theta} \in \mathbb{R}^N$ be the radii and angles of $\boldsymbol{Z}$, i.e. $\boldsymbol{Z} = \lambda(\boldsymbol{r}, \boldsymbol{\theta})$. During our proof of Lemma 9, we have shown that

$$\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2\mathbf{I}_N\right)||_2\right) \propto \int_{[0,2\pi]} \exp\left(c_2\left(\alpha, \boldsymbol{X}, \boldsymbol{r}, \boldsymbol{\theta}\right) + c_3\left(\alpha, \boldsymbol{X}, \boldsymbol{r}, \boldsymbol{\theta}\right)\right)\, d\alpha \tag{54}$$

$$\text{with } \cos(\alpha)\lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,1}^T \frac{1}{\sigma^2}\boldsymbol{X}_{:,1} - \sin(\alpha)\lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,2}^T \frac{1}{\sigma^2}\boldsymbol{X}_{:,1} \tag{55}$$

$$\text{and } \sin(\alpha)\lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,1}^T \frac{1}{\sigma^2}\boldsymbol{X}_{:,2} + \cos(\alpha)\lambda\left(\boldsymbol{r}, \boldsymbol{\phi}\right)_{:,2}^T \frac{1}{\sigma^2}\boldsymbol{X}_{:,2}, \tag{56}$$

see Eqs. (31) to (33) with $\boldsymbol{X}$ substituted for $\boldsymbol{M}$ and $\frac{1}{\sigma^2}\mathbf{I}_N$ substituted for Σ^{-1}. We have omitted terms that are constant and can thus be absorbed into classification threshold t.

By substituting $\boldsymbol{Z}$ for $\lambda(\boldsymbol{r}, \boldsymbol{\theta})$ and explicitly writing out the inner products in Eqs. (55) and (56) as summations over the N rows, Eq. (54) can be equivalently expressed as

$$\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \boldsymbol{X}, \sigma^2\mathbf{I}_N\right)||_2\right) \propto \int\limits_{[0,2\pi]} \exp\left(\frac{1}{\sigma^2}\sum_{n=1}^{N} \boldsymbol{X}_n^T \begin{bmatrix}\cos(\alpha) & -\sin(\alpha)\\ \sin(\alpha) & \cos(\alpha)\end{bmatrix} \boldsymbol{Z}_n\right) d\alpha. \tag{57}$$

Since $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_{\mathrm{F}} = \sum_{n=1}^{N}\left(\boldsymbol{A}_n^T\boldsymbol{B}_n\right)$, the above is in fact identical to

$$\int_0^{2\pi} \exp\left(\frac{1}{\sigma^2}\left\langle \boldsymbol{X}, \boldsymbol{Z}\boldsymbol{R}(\alpha)^T\right\rangle_{\mathrm{F}}\right) d\alpha.$$

D.6 Tight certificate for invariance under elemental rotation in 3D

Next, we generalize our certificate for rotation invariance in 2D to invariance under $\mathbb{T} = \left\{\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(3, d^*) \in \mathbb{R}^3\right\}$, with $SO(3, d) = \{\boldsymbol{R} \in SO(3) \mid \boldsymbol{R}_{d^*} = \boldsymbol{e}_d\}$, where $\boldsymbol{e}_d$ is an indicator vector and $d^* \in \{1, 2, 3\}$ is the rotation axis.

In the following, let $\boldsymbol{A}_{:,-d^*} \in \mathbb{R}^{N\times 2}$ be the matrix composed of the two columns that are not $\boldsymbol{A}_{:,d^*}$.

Theorem 8. *Let $g : \mathbb{R}^{N\times 3} \to \mathbb{Y}$ be invariant to elemental rotations around axis d^* and let $\mathbb{H}_\mathbb{T}$ be defined as in Eq. (2). Then*

$$\min_{h\in\mathbb{H}_\mathbb{T}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \Pr_{o\sim\mathcal{N}(u^{(1)},s),\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)} [\exp(o)\rho(\boldsymbol{q}) \leq t] \tag{58}$$

$$\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{o\sim\mathcal{N}(u^{(2)},s),\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(2)},\boldsymbol{\Sigma}\right)} [\exp(o)\rho(\boldsymbol{q}) \leq t] = p_{\boldsymbol{X},y^*} \tag{59}$$

$$\text{and } \rho(\boldsymbol{q}) = \mathcal{I}_0\left(\sqrt{q_1^2 + q_2^2}\right) / \mathcal{I}_0\left(\sqrt{q_3^2 + q_4^2}\right), \tag{60}$$

where $\mathcal{I}_0$ is the modified Bessel function of the first kind and order 0, and

$$u^{(1)} = \frac{1}{\sigma^2}||\boldsymbol{\Delta}_{:,d^*}||_2^2, \quad u^{(2)} = -\frac{1}{\sigma^2}||\boldsymbol{\Delta}_{:,d^*}||_2^2, \quad s = \frac{1}{\sigma}||\boldsymbol{\Delta}_{:,d^*}||_2.$$

$$\boldsymbol{m}^{(1)} = \frac{1}{\sigma^2}\begin{bmatrix} 2\epsilon_1 + \epsilon_3 + \epsilon_4 \\ 0 \\ \epsilon_1 + \epsilon_3 \\ \epsilon_2 \end{bmatrix}, \quad \boldsymbol{m}^{(2)} = \frac{1}{\sigma^2}\begin{bmatrix} \epsilon_1 + \epsilon_3 \\ -\epsilon_2 \\ \epsilon_3 \\ 0 \end{bmatrix},$$

$$\boldsymbol{\Sigma} = \frac{1}{\sigma^2}\begin{bmatrix} 2\epsilon_1 + \epsilon_3 + \epsilon_4 & 0 & \epsilon_1 + \epsilon_3 & \epsilon_2 \\ 0 & 2\epsilon_1 + \epsilon_3 + \epsilon_4 & -\epsilon_2 & \epsilon_1 + \epsilon_3 \\ \epsilon_1 + \epsilon_3 & -\epsilon_2 & \epsilon_3 & 0 \\ \epsilon_2 & \epsilon_1 + \epsilon_3 & 0 & \epsilon_3 \end{bmatrix}.$$

with $\epsilon_1 = \left\langle \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{\Delta}} \right\rangle_{\mathrm{F}}$, $\epsilon_2 = \left\langle \tilde{\boldsymbol{X}}\boldsymbol{R}\left(-\frac{\pi}{2}\right)^T, \tilde{\boldsymbol{\Delta}} \right\rangle_{\mathrm{F}}$ $\epsilon_3 = ||\boldsymbol{X}_{:,-d^*}||_2^2$ and $\epsilon_4 = ||\boldsymbol{X}_{:,-d^*}||_2^2$.

The certificate is very similar to that from Appendix D.5, but additionally depends on the amount of perturbation $||\boldsymbol{\Delta}_{:,d^*}||_2$ along perturbation axis d^*. It further differs from the previous certificate in that the probabilities in Eqs. (58) and (59) involve a fifth normally distributed random variable $o \in \mathbb{R}$.

D.6.1 Canonical map

The canonical map differs from that for rotations in 2D in that we have to make case distinctions between $d^* = 1$, $d^* = 2$ and $d^* = 3$. We shall formulate these case distinctions in a way that makes the case $d^* = 3$ as similar as possible to the canonical map for rotation invariance in 2D, which will help simplify the proof of our certificate.

For our derivation, recall the definition of canonical maps.

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*

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

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

For rotation invariance, it is more convenient to work with cylindrical coordinates. In the following, let $l : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N \times \{1, 2, 3\} \to \mathbb{R}^{N\times 3}$ be the function that maps from cylindrical coordinates into Cartesian coordinates with

$$l(\boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{z}, d) = \begin{cases} \begin{bmatrix} \boldsymbol{z} & \lambda(\boldsymbol{r}, \boldsymbol{\theta})_{:,1} & \lambda(\boldsymbol{r}, \boldsymbol{\theta})_{:,2} \end{bmatrix} & \text{if } d = 1 \\ \begin{bmatrix} \lambda(\boldsymbol{r}, \boldsymbol{\theta})_{:,2} & \boldsymbol{z} & \lambda(\boldsymbol{r}, \boldsymbol{\theta})_{:,1} \end{bmatrix} & \text{if } d = 2 \\ \begin{bmatrix} \lambda(\boldsymbol{r}, \boldsymbol{\theta})_{:,1} & \lambda(\boldsymbol{r}, \boldsymbol{\theta})_{:,2} & \boldsymbol{z} \end{bmatrix} & \text{if } d = 3 \end{cases} \tag{61}$$

and $\lambda(\boldsymbol{r}, \boldsymbol{\theta}) = [\boldsymbol{r} \odot \cos(\boldsymbol{\theta}) \quad \boldsymbol{r} \odot \sin(\boldsymbol{\theta})]$ defined like in Appendix D.5.

Next, we define an elemental rotation operator for matrices:

Definition 5. *The elemental rotation operator is the function* $\mathrm{erot} : \mathbb{R}^{N\times 3} \times \mathbb{R} \times \{1,2,3\} \to \mathbb{R}^{N\times 2}$ *with*

$$\mathrm{erot}(\boldsymbol{Z}, a, d) = \begin{cases} \boldsymbol{Z}\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(a) & -\sin(a) \\ 0 & \sin(a) & \cos(a) \end{bmatrix}^T & \textit{if } d = 1 \\ \boldsymbol{Z}\begin{bmatrix} \cos(a) & 0 & \sin(a) \\ 0 & 1 & 0 \\ -\sin(a) & 0 & \cos(a) \end{bmatrix}^T & \textit{if } d = 2 \\ \boldsymbol{Z}\begin{bmatrix} \cos(a) & -\sin(a) & 0 \\ \sin(a) & \cos(a) & 0 \\ 0 & 0 & 1 \end{bmatrix}^T & \textit{if } d = 3 \end{cases}$$

that rotates all rows of matrix $\boldsymbol{Z}$ *by angle* a *around axis* d*.*

We use the elemental matrix rotation operator to define our canonical map, which rotates an input matrix $\boldsymbol{Z}$ around axis d^* such that the first row n^* with non-zero radius r_{n^*} takes on angle 0:

Lemma 10. *The function* $\gamma(\boldsymbol{Z}) = \mathrm{erot}\left(\boldsymbol{Z}, -\alpha(\boldsymbol{Z}), d^*\right)$ *with*

$$\alpha(\boldsymbol{Z}) = \begin{cases} 0 & \textit{if } ||\Lambda(\boldsymbol{Z}, d^*)||_2 = 0 \\ \arctan2\left(\Lambda(\boldsymbol{Z})_{n^*,2}, \Lambda(\boldsymbol{Z})_{n^*,1}\right) & \textit{else} \end{cases}$$

where $n^* = \min\{n \mid ||\Lambda(\boldsymbol{Z}_n)||_2 > 0\}$ *is the first non-zero row of matrix* $\Lambda(\boldsymbol{Z}), d^*$ *with*

$$\Lambda(\boldsymbol{Z}, d) = \begin{cases} \begin{bmatrix} \boldsymbol{Z}_{:,2} & \boldsymbol{Z}_{:,3} & \textit{if } d = 1 \\ \boldsymbol{Z}_{:,3} & \boldsymbol{Z}_{:,1} & \textit{if } d = 2 \\ \boldsymbol{Z}_{:,1} & \boldsymbol{Z}_{:,2} & \textit{if } d = 3 \end{bmatrix} \end{cases}$$

is a canonical map for elemental rotations around axis d^**.*

Proof. We can distinguish two cases.

Case 1. All entries of $\Lambda(\boldsymbol{Z}, d^*)$ are 0, i.e. $\Lambda(\boldsymbol{Z}, d^*) = \boldsymbol{0}_{N,2}$. In this case $\alpha(\boldsymbol{Z}) = 0$ and thus $\gamma(\boldsymbol{Z}) = \boldsymbol{Z}$. The equivalence class of $\boldsymbol{Z}$ is $[\boldsymbol{Z}]_{\mathbb{T}} = \{\boldsymbol{Z}\}$, because rotation around d^* will not make $\Lambda(\boldsymbol{Z}, d^*)$ non-zero. Both criteria Eqs. (3) and (4) for canonical maps are thus fulfilled.

Case 2. The matrix $\Lambda(\boldsymbol{Z}, d^*)$ has non-zero entries. Let $(\boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{z})$ be a map of $\boldsymbol{Z}$ in cylindrical coordinates, i.e. $\boldsymbol{Z} = l(\boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{z}, d^*)$. Let n^* be the index of the first non-zero row of $\Lambda(\boldsymbol{Z}, d^*)$. Recall that rotation around axis d^* in Cartesian coordinates corresponds to translation of the cylindrical coordinate angles and thus $\gamma(\boldsymbol{Z}) = l(\boldsymbol{r}, \boldsymbol{\theta} - \boldsymbol{1}_N\theta_{n^*}, \boldsymbol{z}, d^*)$.

Criterion Eq. (3) is fulfilled, because by construction of γ we have $\gamma(\boldsymbol{Z}) = \boldsymbol{Z}\boldsymbol{R}^T$ for some $\boldsymbol{R} \in SO(3, d^*)$.

Now, consider an arbitrary $\boldsymbol{Z}' \in [\boldsymbol{Z}]_{\mathbb{T}}$. By definition of $[\boldsymbol{Z}]_{\mathbb{T}}$, there must be some $a \in \mathbb{R}$ such that $\boldsymbol{Z}' = l\left(\boldsymbol{r}, \boldsymbol{\theta} + a\boldsymbol{1}_N, \boldsymbol{z}, d^*\right)$. Note that rotation does not affect the radii, i.e. n^* is also the index of the first non-zero row of $\boldsymbol{Z}'$. Thus

$$\gamma(\boldsymbol{Z}') = l\left(\boldsymbol{r}, \boldsymbol{\theta} + a\boldsymbol{1}_N - \boldsymbol{1}_N(\theta_{n^*} + a), \boldsymbol{z}, d^*\right) = \lambda\left(\boldsymbol{r}, \boldsymbol{\theta} - \boldsymbol{1}_N(\theta_{n^*}), \boldsymbol{z}, d^*\right) = \gamma(\boldsymbol{Z}),$$

meaning criterion Eq. (4) is fulfilled as well. □

D.6.2 Proof for certificate

In the following, we provide the proof of Theorem 8 for the case $d^* = 3$. The other cases follow analogously. Mapping into cylindrical coordinates is equivalent to transforming the two columns that are not d^* into polar coordinates. Similarly, applying the canonical map for elemental rotations is equivalent to applying the canonical map for rotation in 2D (see Appendix D.5) to the two columns that are not d^* into polar coordinates. We use this fact to reuse much of our proof from Appendix D.5.

Applying Theorem 4 with our canonical map from Lemma 10 shows that

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

Like in our previous proof, we notice that the above optimization problem does not depend on the specific class predicted by h, but only on whether it is y^* or y. We thus formulate the equivalent problem

$$\min_{h:\mathbb{R}^{N\times D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\gamma(\boldsymbol{Z}))\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right] \geq p_{\boldsymbol{X},y^*}, \tag{62}$$

where $h(\boldsymbol{Z}) = 1$ indicates that $\boldsymbol{Z}$ is classified as y^*.

Substitution. We begin by going from Cartesian to cylindrical coordinates: As defined in Eq. (61), let $l : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N \times \{1,2,3\} \to \mathbb{R}^{N\times 3}$ be the function that maps from cylindrical coordinates into Cartesian coordiantes.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right]\\
=&\int_{\mathbb{R}^{N\times D}} h(\gamma(\boldsymbol{Z}))\mu_{\boldsymbol{X}}(\boldsymbol{Z})\, d\boldsymbol{Z}\\
=&\int_{\mathbb{R}^N}\int_{\mathbb{R}^N}\left(\prod_{n=1}^N r_n\right)\int_{[0,2\pi]^N} h(\gamma(l(\boldsymbol{r},\boldsymbol{\theta},\boldsymbol{Z}_{:,d^*},d^*)))\mu_{\boldsymbol{X}}\left(l(\boldsymbol{r},\boldsymbol{\theta},\boldsymbol{Z}_{:,d^*},d^*)\right)\, d\boldsymbol{\theta}\, d\boldsymbol{r}\, d\boldsymbol{Z}_{:,d^*}
\end{aligned}$$

The r_n in the last equality are due to the fact that l transforms N rows from cylindrical to Cartesian coordinates, with each transformation having a Jacobian determinant r.

Similarity to 2D rotation invariance. Now, we use the fact that we are in case d^* to show that the above integral is almost identical to the one from our proof for rotation invariance in 2D.

Firstly, we know from the definition of $l(\cdot)$ in Eq. (61) that $l(\boldsymbol{r},\boldsymbol{\theta},\boldsymbol{Z}_{:,d^*},d^*) = l(\boldsymbol{r},\boldsymbol{\theta},\boldsymbol{Z}_{:,3},3) = [\lambda(\boldsymbol{r},\boldsymbol{\theta}) \quad \boldsymbol{Z}_{:,3}]$, where $\lambda(\boldsymbol{r},\boldsymbol{\theta}) = [\boldsymbol{r}\odot\cos(\boldsymbol{\theta}) \quad \boldsymbol{r}\odot\sin(\boldsymbol{\theta})]$ is the function that maps from polar into Cartesian coordinates.

Secondly, let γ' be the canonical map for rotation invariance in 2D, as defined in Lemma 8. We know from the definition of our canonical map for elemental rotations in Lemma 10 that $\gamma\left([\lambda(\boldsymbol{r},\boldsymbol{\theta}) \quad \boldsymbol{Z}_{:,3}]\right) = \left[\gamma'\left(\lambda(\boldsymbol{r},\boldsymbol{\theta})\right) \quad \boldsymbol{Z}_{:,3}\right]$.

Thirdly, recall that $\mu_{\boldsymbol{X}}$ is an isotropic matrix normal distribution, i.e. all columns are independent. Let $\mu_{\boldsymbol{X}_{:,:2}}$ be the distribution of the first two dimensions and $\mu_{\boldsymbol{X}_{:,3}}$ be the distribution of the third dimension, i.e.

$$\mu_{\boldsymbol{X}}(\boldsymbol{Z}) = \mu_{\boldsymbol{X}_{:,:2}}\left(\boldsymbol{Z}_{:,:2}\right)\mu_{\boldsymbol{X}_{:,3}}(\boldsymbol{Z}_{:,3}).$$

Combining all three steps shows that

$$\begin{aligned}
&\int_{\mathbb{R}^N}\int_{\mathbb{R}^N}\left(\prod_{n=1}^N r_n\right)\int_{[0,2\pi]^N} h(\gamma(l(\boldsymbol{r},\boldsymbol{\theta},\boldsymbol{Z}_{:,d^*},d^*)))\mu_{\boldsymbol{X}}\left(l(\boldsymbol{r},\boldsymbol{\theta},\boldsymbol{Z}_{:,d^*},d^*)\right)\, d\boldsymbol{\theta}\, d\boldsymbol{r}\, d\boldsymbol{Z}_{:,d^*}\\
=&\int_{\mathbb{R}^N}\mu_{\boldsymbol{X}_{:,3}}\left(\boldsymbol{Z}_{:,3}\right)\Bigg(\\
&\int_{\mathbb{R}^N}\left(\prod_{n=1}^N r_n\right)\int_{[0,2\pi]^N} h\left([\gamma'\left(\lambda(\boldsymbol{r},\boldsymbol{\theta})\right) \quad \boldsymbol{Z}_{:,3}]\right)\mu_{\boldsymbol{X}_{:,:2}}\left(\gamma'\left(\lambda(\boldsymbol{r},\boldsymbol{\theta})\right)\right)\, d\boldsymbol{\theta}\, d\boldsymbol{r}\Bigg)\, d\boldsymbol{Z}_{:,3}
\end{aligned}$$

This expectation is identical to the one we derived in Eq. (37) of Appendix D.5, save for an independent variable $\boldsymbol{Z}_{:,3}$ that contributes factor $\mu_{\boldsymbol{X}_{:,3}}\left(\boldsymbol{Z}_{:,3}\right)$ to the likelihood. We may thus perform exactly the same steps as in the derivation of our certificate for rotation invariance in 2D, starting at Eq. (37), i.e. further substitutions, marginalization, application of the Neyman Pearson lemma and transformation back into Cartesian coordinates, to show that

$$\min_{h\in\mathbb{H}_{\mathrm{T}}}\Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right] = \Pr_{\boldsymbol{Z}_{:,3}\sim\mu_{\boldsymbol{X}'_{:,3}},\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)}\left[\frac{\mu_{\boldsymbol{X}'_{:,3}}(\boldsymbol{Z}_{:,3})}{\mu_{\boldsymbol{X}_{:,3}}(\boldsymbol{Z}_{:,3})}\rho(\boldsymbol{q}) \leq t\right] \tag{63}$$

$$\text{with } t\in\mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{Z}_{:,3}\sim\mu_{\boldsymbol{X}_{:,3}},\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(2)},\boldsymbol{\Sigma}\right)}\left[\frac{\mu_{\boldsymbol{X}'_{:,3}}(\boldsymbol{Z}_{:,3})}{\mu_{\boldsymbol{X}_{:,3}}(\boldsymbol{Z}_{:,3})}\rho(\boldsymbol{q}) \leq t\right] = p_{\boldsymbol{X},y^*} \tag{64}$$

$$\text{and } \rho(\boldsymbol{q}) = \mathcal{I}_0\left(\sqrt{q_1^2+q_2^2}\right)/\mathcal{I}_0\left(\sqrt{q_3^2+q_4^2}\right), \tag{65}$$

Likelihood ratio distribution. The last step of our derivation is to find the distribution of the likelihood ratios in Eqs. (63) and (64). Since $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$, we have

$$\frac{\mu_{\boldsymbol{X}'_{:,3}}(\boldsymbol{Z}_{:,3})}{\mu_{\boldsymbol{X}_{:,3}}(\boldsymbol{Z}_{:,3})} = \frac{\mathcal{N}\left(\boldsymbol{Z}_{:,3} \mid \boldsymbol{X}', \sigma^2 \mathbf{I}_N\right)}{\mathcal{N}\left(\boldsymbol{Z}_{:,3} \mid \boldsymbol{X}, \sigma^2 \mathbf{I}_N\right)} = \exp\left(\frac{1}{2\sigma^2}\left(2\boldsymbol{\Delta}_{:,3}^T \boldsymbol{Z}_{:,3} - ||\boldsymbol{\Delta}_{:,3}||_2^2 - 2\boldsymbol{\Delta}_{:,3}^T \boldsymbol{X}_{:,3}\right)\right),$$

see also Appendix A of [27].

The argument of the exponential function is an affine transformation of random variable $\boldsymbol{Z}_{:,3}$, $o = \frac{1}{2\sigma^2}\left(2\boldsymbol{\Delta}_{:,3}^T \boldsymbol{Z}_{:,3} - ||\boldsymbol{\Delta}_{:,3}||_2^2 - 2\boldsymbol{\Delta}_{:,3}^T \boldsymbol{X}_{:,3}\right)$.

In Eq. (63), we have $\boldsymbol{Z}_{:,3} \sim \mu_{\boldsymbol{X}'_{:,3}} := \mathcal{N}\left(\boldsymbol{X}_{:,3} + \boldsymbol{\Delta}_{:,3}, \sigma^2 \mathbf{I}_N\right)$ and thus $o \sim \mathcal{N}(u^{(1)}, s)$ with mean $u^{(1)} = \frac{1}{2\sigma^2}\left(2||\boldsymbol{\Delta}_{:,3}||_2^2 - ||\boldsymbol{\Delta}_{:,3}||_2^2\right) = \frac{1}{2\sigma^2}\left(||\boldsymbol{\Delta}_{:,3}||_2^2\right)$ and variance $s^2 = \frac{1}{\sigma^2}\boldsymbol{\Delta}_{:,3}^T \sigma^2 \mathbf{I}_N \boldsymbol{\Delta}_{:,3} \frac{1}{\sigma^2} = \frac{1}{\sigma^2}||\boldsymbol{\Delta}_{:,3}||_2^2$.

In Eq. (64) we have $\boldsymbol{Z}_{:,3} \sim \mu_{\boldsymbol{X}_{:,3}} := \mathcal{N}\left(\boldsymbol{X}_{:,3}, \sigma^2 \mathbf{I}_N\right)$ and thus $o \sim \mathcal{N}(u^{(2)}, s)$ with mean $u^{(2)} = \frac{1}{2\sigma^2}\left(-||\boldsymbol{\Delta}_{:,3}||_2^2\right) = -\frac{1}{2\sigma^2}\left(||\boldsymbol{\Delta}_{:,3}||_2^2\right)$ and variance $s^2 = \frac{1}{\sigma^2}||\boldsymbol{\Delta}_{:,3}||_2^2$.

Therefore, we have

$$\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}}[h(\boldsymbol{Z}) = y^*] = \Pr_{o \sim \mathcal{N}(u^{(1)}, s), \boldsymbol{q} \sim \mathcal{N}\left(\boldsymbol{m}^{(1)}, \boldsymbol{\Sigma}\right)}[\exp(o)\rho(\boldsymbol{q}) \leq t] \tag{66}$$

$$\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{o \sim \mathcal{N}(u^{(2)}, s), \boldsymbol{q} \sim \mathcal{N}\left(\boldsymbol{m}^{(2)}, \boldsymbol{\Sigma}\right)}[\exp(o)\rho(\boldsymbol{q}) \leq t] = p_{\boldsymbol{X}, y^*} \tag{67}$$

$$\text{and } \rho(\boldsymbol{q}) = \mathcal{I}_0\left(\sqrt{q_1^2 + q_2^2}\right) / \mathcal{I}_0\left(\sqrt{q_3^2 + q_4^2}\right), \tag{68}$$

which concludes our proof.

D.7 Tight certificate for simultaneous invariance under rotation and translation in 2D

Next, we consider the case that base classifier g is rotation and translation invariant, i.e. invariant under group $\mathbb{T} = \left\{ \boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T + \mathbf{1}_N \boldsymbol{c}^T \mid \boldsymbol{R} \in SO(2), \boldsymbol{c} \in \mathbb{R}^D \right\}$.

In the following, let $\tilde{\boldsymbol{A}} \in \mathbb{R}^{N \times D}$ be matrix $A \in \mathbb{R}^{N \times D}$ after centering, i.e. $\tilde{\boldsymbol{A}} = \boldsymbol{A} - \mathbf{1}_N \overline{\boldsymbol{A}}$, where $\overline{\boldsymbol{A}} \in \mathbb{R}^{1 \times D}$ are the column-wise averages. Further let $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\mathrm{F}} = \sum_{n=1}^{N} \boldsymbol{A}_n^T \boldsymbol{B}_n^T = \mathrm{vec}(A)^T \mathrm{vec}(B)$ be the Frobenius inner product.

Theorem 9. *Let $g : \mathbb{R}^{N \times 2} \to \mathbb{Y}$ be rotation and translation invariant and $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. Then*

$$\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \Pr_{\boldsymbol{q} \sim \mathcal{N}\left(\boldsymbol{m}^{(1)}, \boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q}) \leq t] \tag{69}$$

$$\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{q} \sim \mathcal{N}\left(\boldsymbol{m}^{(2)}, \boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q}) \leq t] = p_{\boldsymbol{X}, y^*} \tag{70}$$

$$\text{and } \rho(\boldsymbol{q}) = \mathcal{I}_0 \left(\sqrt{q_1^2 + q_2^2} \right) / \mathcal{I}_0 \left(\sqrt{q_3^2 + q_4^2} \right), \tag{71}$$

where $\mathcal{I}_0$ is the modified Bessel function of the first kind and order 0*, and*

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

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

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

The above certificate is identical to the tight certificate for rotation invariance in 2D (see Appendix D.5), save for the fact that the clean input $\boldsymbol{X}$ and the perturbation $\boldsymbol{\Delta}$ have been centered. Since $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$, this is equivalent to centering $\boldsymbol{X}$ and $\boldsymbol{X}'$ and then evaluating the certificate from Appendix D.5.

D.7.1 Canonical map

For our derivation, recall the definition of canonical maps:

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*

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

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

In the following, let γ_{rot} be the canonical map for invariance under $\mathbb{T}_{\mathrm{rot}} := \left\{ \boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(2) \right\}$, as defined in Appendix D.5. Further let γ_{trans} be the canonical map for invariance under $\mathbb{T}_{\mathrm{trans}} := \left\{ \boldsymbol{X} \mapsto \mathbf{1}_N \boldsymbol{c}^T \mid \boldsymbol{c} \in \mathbb{R}^2 \right\}$, as defined in Appendix D.4.

Composing these two canonical maps yields a canonical map for simultaneous rotation and translation invariance.

Lemma 11. *The function $\gamma(\boldsymbol{Z}) = \gamma_{\mathrm{rot}}\left(\gamma_{\mathrm{trans}}(\boldsymbol{Z})\right)$ is a canonical map for simultaneous rotation and translation invariance.*

Proof. Consider an arbitrary $\boldsymbol{Z} \in \mathbb{R}^{N \times 2}$. As defined in Appendices D.4 and D.5, γ performs some translation, followed by some rotation. That is, $\gamma(\boldsymbol{Z}) = \gamma_{\mathrm{rot}}\left(\boldsymbol{Z} + \mathbf{1}_N \boldsymbol{c}^T\right)$ for some $\boldsymbol{c} \in \mathbb{R}^2$ and $\gamma_{\mathrm{rot}}(\boldsymbol{Z} + \mathbf{1}_N \boldsymbol{c}^T) = \left(\boldsymbol{Z} + \mathbf{1}_N \boldsymbol{c}^T\right) \boldsymbol{R}^T = \boldsymbol{Z}\boldsymbol{R}^T + \mathbf{1}_N \left(\boldsymbol{R}\boldsymbol{c}\right)^T$ for some $\boldsymbol{R} \in SO(2)$. Evidently $\gamma(\boldsymbol{Z}) \in [\boldsymbol{Z}]_{\mathbb{T}}$, meaning criterion Eq. (3) is fulfilled.

Now, consider an arbitrary pair $\boldsymbol{Z} \in \mathbb{R}^{N\times 2}, \boldsymbol{Z}' \in [\boldsymbol{Z}]_{\mathbb{T}}$. Since $\boldsymbol{Z} \in [\boldsymbol{Z}]_{\mathbb{T}}$, we must have $\boldsymbol{Z}' = \boldsymbol{Z}\boldsymbol{R}^T + \mathbf{1}_N \boldsymbol{c}^T$ for some $\boldsymbol{R} \in SO(2)$ and $\boldsymbol{c} \in \mathbb{R}^2$. Recall that $\gamma_{\text{trans}}(\boldsymbol{Z})$ maps input $\boldsymbol{Z}$ to a representative by subtracting the first row from all other rows, i.e.

$$\gamma_{\text{trans}}(\boldsymbol{Z}) = \begin{bmatrix} 0 & \mathbf{0}_{N-1}^T \\ -\mathbf{1}_{N-1} & \mathbf{I}_{N-1} \end{bmatrix} := \boldsymbol{K}\boldsymbol{Z}.$$

Further note that γ_{trans} is a canonical map for translation invariance and therefore fulfills criterion Eq. (4). Therefore, we have $\gamma_{\text{trans}}(\boldsymbol{Z}') = \boldsymbol{K}\left(\boldsymbol{Z}\boldsymbol{R}^T + \mathbf{1}_N \boldsymbol{c}^T\right) = \boldsymbol{K}\left(\boldsymbol{Z}\boldsymbol{R}^T\right) = (\boldsymbol{K}\boldsymbol{Z})\boldsymbol{R}^T$, where second equality is due to criterion Eq. (4) and the last equality is due to associativity. Finally, note that since γ_{rot} is a canonical map for translation invariance, we must have $\gamma_{\text{rot}}(\gamma_{\text{trans}}(\boldsymbol{Z}')) = \gamma_{\text{rot}}((\boldsymbol{K}\boldsymbol{Z})\boldsymbol{R}^T) = \gamma_{\text{rot}}(\boldsymbol{K}\boldsymbol{Z}) = \gamma_{\text{rot}}(\gamma_{\text{trans}}(\boldsymbol{Z}))$. This shows that γ also fulfills criterion Eq. (4), i.e. it is a canonical map for simultaneous translation and rotation invariance. □

D.7.2 Proof of certificate

To derive our certificate, we shall combine the substitutions we used for the translation invariance certificate (see Appendix D.4) with those we used for rotation invariance (see Appendix D.5). We shall see that this leads to an optimization problem that is almost identical to that for 2D rotation invariance, save for the fact that we have anisotropic normal distributions. Like in Appendix D.4, the mean and covariance of these anisotropic distributions naturally lead to a centering of the certificate parameters.

Before proceeding, we need the following lemma:

Lemma 12. *Consider arbitrary vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$ and let $\overline{\boldsymbol{x}} = \frac{1}{D}\sum_{d=1}^D x_d$, $\overline{\boldsymbol{y}} = \frac{1}{D}\sum_{d=1}^D y_d$ be their averages. Then*

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

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

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

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

Applying Theorem 4 with our canonical map $\gamma(\boldsymbol{Z}) = \gamma_{\text{rot}}(\gamma_{\text{trans}}(\boldsymbol{Z}))$ leads to optimization problem

$$\min_{h:\mathbb{R}^{N\times D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\gamma(\boldsymbol{Z}))\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right] \geq p_{\boldsymbol{X},y^*}, \tag{72}$$

where $h(\boldsymbol{Z}) = 1$ indicates that $\boldsymbol{Z}$ is classified as y^*.

Substitution. First, we perform exactly the same substitution. as in our proof for translation invariance to eliminate γ_{trans} (see Eq. (13) to Eq. (14) in Appendix D.4). That is, $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{Z}$ with $\boldsymbol{A} = \begin{bmatrix} 1 & \mathbf{0}_{N-1}^T \\ -\mathbf{1}_{N-1} & \mathbf{I}_{N-1} \end{bmatrix}$.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right] \\
&= \int_{\mathbb{R}^{N\times 2}} h(\gamma_{\text{rot}}(\gamma_{\text{trans}}(\boldsymbol{Z}))) \prod_{d=1}^2 \mathcal{N}\left(\boldsymbol{V}_{:,d} \mid \boldsymbol{X}_{:,d}, \sigma^2\mathbf{I}\right)\, d\boldsymbol{Z} \\
&= \int_{\mathbb{R}^{(N-1)\times 2}} h\left(\gamma_{\text{rot}}\left(\begin{bmatrix}\mathbf{0}_2^T \\ \boldsymbol{V}_{2:}\end{bmatrix}\right)\right) \prod_{d=1}^2 \mathcal{N}\left(\boldsymbol{V}_{:,d} \mid \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}, \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)\, d\boldsymbol{V}_{2:} \\
&:= \int_{\mathbb{R}^{(N-1)\times 2}} h\left(\gamma_{\text{rot}}\left(\begin{bmatrix}\mathbf{0}_2^T \\ \boldsymbol{W}\end{bmatrix}\right)\right) \eta_{\boldsymbol{X}}(\boldsymbol{W})\, d\boldsymbol{W},
\end{aligned}$$

where we have defined $\eta_{\boldsymbol{X}}(\boldsymbol{W}) = \prod_{d=1}^{2} \mathcal{N}\left(\boldsymbol{W}_{:,d} \mid \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}, \sigma^2 \boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)$ to be the marginal distribution over $\mathbb{R}^{N\times 2}$.

Next, we eliminate γ_{rot} by transforming into polar coordinates. For this, let $\lambda(\boldsymbol{r}, \boldsymbol{\theta}) = [\boldsymbol{r} \odot \cos(\boldsymbol{\theta}) \quad \boldsymbol{r} \odot \sin(\boldsymbol{\theta})]$ be the function that maps from polar into Cartesian coordinates.

$$= \int_{\mathbb{R}^{(N-1}} \prod_{n=1}^{N-1} r_n \int_{[0,2\pi]} h\left(\gamma_{\text{rot}}\left(\begin{bmatrix} \mathbf{0}_2^T \\ \lambda(\boldsymbol{r}, \boldsymbol{\theta}) \end{bmatrix}\right)\right) \eta_{\boldsymbol{X}}(\lambda(\boldsymbol{r}, \boldsymbol{\theta}))\, d\boldsymbol{r}\boldsymbol{\theta}$$

Recall from Appendix D.5 that γ_{rot} rotates its argument such that its first non-zero row is aligned with the first coordinate axis. Evidently, the first row of its argument is zero. Due to the factor $\prod_{n=1}^{N-1} r_n$, the entire integrand is zero whenver any of the other rows are zero (recall our discussion of Eq. (38) in Appendix D.5). We may thus assume that γ_{rot} always rotates by θ_1:

$$= \int_{\mathbb{R}^{(N-1}} \prod_{n=1}^{N-1} r_n \int_{[0,2\pi]} h\left(\begin{bmatrix} \mathbf{0}_2^T \\ \lambda(\boldsymbol{r}, \boldsymbol{\theta} - \mathbf{1}_{N-1}\theta_1) \end{bmatrix}\right) \eta_{\boldsymbol{X}}(\lambda(\boldsymbol{r}, \boldsymbol{\theta}))\, d\boldsymbol{r}\boldsymbol{\theta} \tag{73}$$

Since $\boldsymbol{W} \mapsto \begin{bmatrix} \mathbf{0}_2^T \\ \boldsymbol{W} \end{bmatrix}$ is an injective function from $\mathbb{R}^{(N-1)\times 2}$ into $\mathbb{R}^{N\times 2}$, we may formulate an equivalent optimization problem over a family of variables (h') indexed by $\mathbb{R}^{(N-1)\times 2}$, which leads to expectation

$$\int_{\mathbb{R}^{(N-1}} \prod_{n=1}^{N-1} r_n \int_{[0,2\pi]} h'\left(\lambda(\boldsymbol{r}, \boldsymbol{\theta} - \mathbf{1}_{N-1}\theta_1)\right) \eta_{\boldsymbol{X}}(\lambda(\boldsymbol{r}, \boldsymbol{\theta}))\, d\boldsymbol{r}\boldsymbol{\theta}$$

Similarity to 2D rotation invariance. We now observe that this expectation is almost identical to Eq. (39) from Appendix D.5, the expectation we obtained after eliminating the canonical map for 2D rotation invariance. The only difference is that we have a distribution $\eta_{\boldsymbol{X}}$ that is a product of two anisotropic multivariate normal distributions, instead of $\mu_{\boldsymbol{X}}$, which was a product of two isotropic multivariate normal distributions. However, our proof did not make any assumptions about the parameters of the distribution. We may thus perform the same steps, starting at Eq. (39), i.e. further substitution, marginalization, application of the Neyman Pearson lemma and transformation back into Cartesian coordinates, to show that

$$\min_{h\in\mathbb{H}_{\mathrm{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right] = \Pr_{\boldsymbol{W}\sim\eta_{\boldsymbol{X}'}}\left[\frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{W}, \boldsymbol{A}_{2:}\boldsymbol{X}', \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{W}, \boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)||_2\right)} \leq t\right] \tag{74}$$

$$\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{W}\sim\eta_{\boldsymbol{X}}}\left[\frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{W}, \boldsymbol{A}_{2:}\boldsymbol{X}', \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{W}, \boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)||_2\right)} \leq t\right], \tag{75}$$

with κ defined as in Lemma 9.

Likelihood ratio distribution. Finally, recall from Lemma 9 that

$$\kappa\left(\boldsymbol{W}, \boldsymbol{M}, \boldsymbol{\Sigma}\right) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} & \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,2} & -\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{:,1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{W}_{:,1} \\ \boldsymbol{W}_{:,2} \end{bmatrix}.$$

We shall now use our knowledge about the distribution of $\boldsymbol{W}$, as well as some simple algebra to show that Eqs. (74) and (75) are equivalent to evaluating the tight certificate for 2D rotation invariance, after centering $\boldsymbol{X}'$ and $\boldsymbol{X}$. Let $\boldsymbol{B} = \boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T$. First, we can absorb the parameter $\boldsymbol{B}$ into the second argument of κ:

$$\begin{aligned}\kappa\left(\boldsymbol{W}, \boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\boldsymbol{B}\right) =& \frac{1}{\sigma^2}\begin{bmatrix} \left(\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,1} & \left(\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,2} \\ \left(\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,2} & -\left(\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{W}_{:,1} \\ \boldsymbol{W}_{:,2} \end{bmatrix} \\ =& \kappa\left(\boldsymbol{W}, \boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\mathbf{I}_N\right)\end{aligned}$$

Furthermore, we know from Eqs. (74) and (75) that $\boldsymbol{W} \sim \eta_{\boldsymbol{X}}$ or $\boldsymbol{W} \sim \eta'_{\boldsymbol{X}}$ with $\eta_{\boldsymbol{X}}(\boldsymbol{W}) = \prod_{d=1}^{2} \mathcal{N}\left(\boldsymbol{W} \mid \boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)$. Let us begin with the case that $\boldsymbol{W} \sim \eta_{\boldsymbol{X}}$. The random

variable $\boldsymbol{W}$ is an affine transformation of random variable $\boldsymbol{E} \in \mathbb{R}^{N\times 2}$ that is distributed according to a zero-mean isotropic matrix normal distribution $\mu_{\mathbf{0}}(\boldsymbol{E}) = \prod_{d=1}^{2} \mathcal{N}\left(\boldsymbol{Z} \mid \mathbf{0}, \sigma^2 \mathbf{I}_N\right)$:

$$\begin{aligned}\boldsymbol{W} =& \begin{bmatrix} \boldsymbol{A}_{2:} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_{2:} \end{bmatrix} (\boldsymbol{E} + \boldsymbol{X}) \\ :=& \boldsymbol{S}\left(\boldsymbol{E} + \boldsymbol{X}\right).\end{aligned}$$

We may thus restate the probability from Eq. (75) as

$$\Pr_{\boldsymbol{E}\sim\mu_{\mathbf{0}}}\left[\frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{S}\left(\boldsymbol{E}+\boldsymbol{X}\right), \boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}', \sigma^2\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{S}\left(\boldsymbol{E}+\boldsymbol{X}\right), \boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\right)||_2\right)} \leq t\right], \tag{76}$$

Next, we can absorb $\boldsymbol{S}$ into the second argument of κ:

$$\begin{aligned}&\kappa\left(\boldsymbol{S}(\boldsymbol{E}+\boldsymbol{X}), \boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\mathbf{I}_N\right) \\ =&\frac{1}{\sigma^2}\begin{bmatrix} \left(\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,1} & \left(\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,2} \\ \left(\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,2} & -\left(\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,1} \end{bmatrix}^T \begin{bmatrix} \left(\boldsymbol{A}_{2:}\left(\boldsymbol{E}+\boldsymbol{X}\right)\right)_{:,1} \\ \left(\boldsymbol{A}_{2:}\left(\boldsymbol{E}+\boldsymbol{X}\right)\right)_{:,2} \end{bmatrix} \\ =&\frac{1}{\sigma^2}\begin{bmatrix} \left((\boldsymbol{A}_{2:})^T B^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,1} & \left((\boldsymbol{A}_{2:})^T\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,2} \\ \left((\boldsymbol{A}_{2:})^T\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,2} & -\left((\boldsymbol{A}_{2:})^T\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}\right)_{:,1} \end{bmatrix}^T \begin{bmatrix} \left(\boldsymbol{E}+\boldsymbol{X}\right)_{:,1} \\ \left(\boldsymbol{E}+\boldsymbol{X}\right)_{:,2} \end{bmatrix} \\ =&\kappa\left(\boldsymbol{E}+\boldsymbol{X}, (\boldsymbol{A}_{2:})^T\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\mathbf{I}_N\right).\end{aligned}$$

We know from Eq. (26) of our derivation of the translation invariance certificate that

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

i.e. $\left(\boldsymbol{A}_{2:}\right)^T\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}$ is the matrix that subtracts the mean from every vector it is multiplied with. Thus $\kappa\left(\boldsymbol{E}+\boldsymbol{X}, (\boldsymbol{A}_{2:})^T\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\mathbf{I}_N\right) = \kappa\left(\boldsymbol{E}+\boldsymbol{X}, \tilde{\boldsymbol{X}}, \sigma^2\mathbf{I}_N\right)$, where $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \mathbf{1}_N\overline{\boldsymbol{X}}$.

Next, we can apply Lemma 12, which shows that computing the inner product between a centered and an uncentered vector is identical to centering both vectors and then computing the inner product:

$$\kappa\left(\boldsymbol{E}+\boldsymbol{X}, \tilde{\boldsymbol{X}}, \sigma^2\mathbf{I}_N\right) = \frac{1}{\sigma^2}\begin{bmatrix} \tilde{\boldsymbol{X}}_{:,1} & \tilde{\boldsymbol{X}}_{:,2} \\ \tilde{\boldsymbol{X}}_{:,2} & \tilde{\boldsymbol{X}}_{:,1} \end{bmatrix}^T \begin{bmatrix} \left(\boldsymbol{E}+\boldsymbol{X}\right)_{:,1} \\ \left(\boldsymbol{E}+\boldsymbol{X}\right)_{:,2} \end{bmatrix} \tag{77}$$

$$= \frac{1}{\sigma^2}\begin{bmatrix} \tilde{\boldsymbol{X}}_{:,1} & \tilde{\boldsymbol{X}}_{:,2} \\ \tilde{\boldsymbol{X}}_{:,2} & \tilde{\boldsymbol{X}}_{:,1} \end{bmatrix}^T \begin{bmatrix} \left(\boldsymbol{E}+\tilde{\boldsymbol{X}}\right)_{:,1} \\ \left(\boldsymbol{E}+\tilde{\boldsymbol{X}}\right)_{:,2} \end{bmatrix} \tag{78}$$

$$= \kappa\left(\boldsymbol{E}+\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}, \sigma^2\mathbf{I}_N\right). \tag{79}$$

Finally, inserting this result into Eq. (76) and using linearity of expectation shows that

$$\begin{aligned}&\Pr_{\boldsymbol{E}\sim\mu_{\mathbf{0}}}\left[\frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{S}\left(\boldsymbol{E}+\boldsymbol{X}\right), \boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}', \sigma^2\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{S}\left(\boldsymbol{E}+\boldsymbol{X}\right), \boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\boldsymbol{X}, \sigma^2\right)||_2\right)} \leq t\right] \\ =& \Pr_{\boldsymbol{E}\sim\mu_{\mathbf{0}}}\left[\frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{E}+\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}', \sigma^2\mathbf{I}_N\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{E}+\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}, \sigma^2\mathbf{I}_N\right))||_2\right)} \leq t\right] \\ =& \Pr_{\boldsymbol{Z}\sim\mu_{\tilde{\boldsymbol{X}}}}\left[\frac{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \tilde{\boldsymbol{X}}', \sigma^2\mathbf{I}_N\right)||_2\right)}{\mathcal{I}_0\left(||\kappa\left(\boldsymbol{Z}, \tilde{\boldsymbol{X}}, \sigma^2\mathbf{I}_N\right))||_2\right)} \leq t\right].\end{aligned}$$

The case $\boldsymbol{W} \sim \eta_{\boldsymbol{X}'}$ follows analogously. We therefore have

$$\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \Pr_{\boldsymbol{Z} \sim \mu_{\tilde{\boldsymbol{X}}'}} \left[\frac{\mathcal{I}_0 \left(||\kappa \left(\boldsymbol{Z}, \tilde{\boldsymbol{X}}', \sigma^2 \mathbf{I}_N \right))||_2 \right)}{\mathcal{I}_0 \left(||\kappa \left(\boldsymbol{Z}, \tilde{\boldsymbol{X}}, \sigma^2 \mathbf{I}_N \right))||_2 \right)} \leq t \right] \tag{80}$$

$$\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{Z} \sim \mu_{\tilde{\boldsymbol{X}}'}} \left[\frac{\mathcal{I}_0 \left(||\kappa \left(\boldsymbol{Z}, \tilde{\boldsymbol{X}}', \sigma^2 \mathbf{I}_N \right))||_2 \right)}{\mathcal{I}_0 \left(||\kappa \left(\boldsymbol{Z}, \tilde{\boldsymbol{X}}, \sigma^2 \mathbf{I}_N \right))||_2 \right)} \leq t \right], \tag{81}$$

This solution is identical to the one we derived for rotation invariance (see Eqs. (52) and (53) in Appendix D.5), except that all occurrences of $\boldsymbol{X}$ have been replaced with the centered $\tilde{\boldsymbol{X}}$ and all occurrences of $\boldsymbol{X}'$ have been replaced with the centered $\tilde{\boldsymbol{X}}'$. Thus, performing the subsequent derivation steps from Appendix D.5 yields the desired certificate.

D.8 Tight certificate for simultaneous invariance under elemental rotation and translation in 3D

Finally, we consider the case that base classifier g is invariant under group $\mathbb{T} = \left\{\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T + \boldsymbol{1}_N \boldsymbol{c}^T \mid \boldsymbol{R} \in SO(3, d^*), \boldsymbol{c} \in \mathbb{R}^3\right\}$, with $SO(3, d) = \{\boldsymbol{R} \in SO(3) \mid \boldsymbol{R}_{d^*} = \boldsymbol{e}_d\}$, where $\boldsymbol{e}_d$ is an indicator vector and d^* is the rotation axis.

In the following, let $\tilde{\boldsymbol{A}} \in \mathbb{R}^{N \times D}$ be matrix $A \in \mathbb{R}^{N \times D}$ after centering, i.e. $\tilde{\boldsymbol{A}} = \boldsymbol{A} - \boldsymbol{1}_N \overline{\boldsymbol{A}}$, where $\overline{\boldsymbol{A}} \in \mathbb{R}^{1 \times D}$ are the column-wise averages. Further let $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\mathrm{F}} = \sum_{n=1}^{N} \boldsymbol{A}_n^T \boldsymbol{B}_n^T = \mathrm{vec}(A)^T \mathrm{vec}(B)$ be the Frobenius inner product.

Theorem 10. *Let $g : \mathbb{R}^{N \times 3} \to \mathbb{Y}$ be translation invariant and invariant to elemental rotations around axis d^* and let $\mathbb{H}_{\mathbb{T}}$ be defined as in Eq.* (2)*. Then*

$$\min_{h \in \mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z} \sim \mu_{\boldsymbol{X}'}} [h(\boldsymbol{Z}) = y^*] = \Pr_{o \sim \mathcal{N}(u^{(1)}, s), \boldsymbol{q} \sim \mathcal{N}\left(\boldsymbol{m}^{(1)}, \boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q}) \leq t] \tag{82}$$

$$\textit{with } t \in \mathbb{R} \textit{ s.t. } \Pr_{o \sim \mathcal{N}(u^{(2)}, s), \boldsymbol{q} \sim \mathcal{N}\left(\boldsymbol{m}^{(2)}, \boldsymbol{\Sigma}\right)} [\rho(\boldsymbol{q}) \leq t] = p_{\boldsymbol{X}, y^*} \tag{83}$$

$$\textit{and } \rho(\boldsymbol{q}) = \mathcal{I}_0\left(\sqrt{q_1^2 + q_2^2}\right) / \mathcal{I}_0\left(\sqrt{q_3^2 + q_4^2}\right), \tag{84}$$

where $\mathcal{I}_0$ is the modified Bessel function of the first kind and order 0*, and*

$$u^{(1)} = \frac{1}{\sigma^2} ||\tilde{\boldsymbol{\Delta}}_{:,d^*}||_2^2, \quad u^{(2)} = -\frac{1}{\sigma^2} ||\tilde{\boldsymbol{\Delta}}_{:,d^*}||_2^2, \quad s = \frac{1}{\sigma} ||\tilde{\boldsymbol{\Delta}}_{:,d^*}||_2.$$

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

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

The certificate is identical to centering $\boldsymbol{X}$ and $\boldsymbol{X}'$ and evaluating the tight certificate for invariance under elemental rotation derived in Appendix D.6.

D.8.1 Canonical map

In the following, let γ_{erot} be the canonical map for invariance under $\mathbb{T}_{\mathrm{erot}} := \left\{\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(3, d^*)\right\}$, as defined in Appendix D.6. Further let γ_{trans} be the canonical map for invariance under $\mathbb{T}_{\mathrm{trans}} := \left\{\boldsymbol{X} \mapsto \boldsymbol{1}_N \boldsymbol{c}^T \mid \boldsymbol{c} \in \mathbb{R}^2\right\}$, as defined in Appendix D.4.

Composing these two canonical maps yields a canonical map for simultaneous rotation and translation invariance.

Lemma 13. *The function $\gamma(\boldsymbol{Z}) = \gamma_{\mathrm{erot}}\left(\gamma_{\mathrm{trans}}(\boldsymbol{Z})\right)$ is a canonical map for simultaneous rotation and translation invariance.*

Proof. Identical to the proof of Lemma 10, replacing every occurence of "$SO(2)$" with $SO(3, d^*)$. □

D.8.2 Proof for certificate

Like in Appendix D.6, we only consider the case $d^* = 3$, the other cases follow analogously. As in our proof for simultaneous rotation and translation invariance in Appendix D.7, we first eliminate γ_{trans} via the substitution we used for our translation invariance certificate. We then make the same observation as in our proof for elemental rotations: The optimization problem is identical to that for 2D rotation invariance, save for an independent factor in the density function. We may thus reuse much of the proof and only have to find the distribution of an independent likelihood ratio.

Applying Theorem 4 with our canonical map $\gamma(\boldsymbol{Z}) = \gamma_{\text{rot}}(\gamma_{\text{trans}}(\boldsymbol{Z}))$ leads to optimization problem

$$\min_{h:\mathbb{R}^{N\times D}\to\{0,1\}} \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\gamma(\boldsymbol{Z}))\right] \text{ s.t. } \mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right] \geq p_{\boldsymbol{X},y^*}, \tag{85}$$

where $h(\boldsymbol{Z}) = 1$ indicates that $\boldsymbol{Z}$ is classified as y^*.

Substitution. Performing the same steps as in Appendix D.7 to eliminate γ_{trans} shows that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right]\\ &= \int_{\mathbb{R}^{(N-1)\times 3}} h\left(\gamma_{\text{rot}}\left(\begin{bmatrix}\mathbf{0}_2^T\\ \boldsymbol{W}\end{bmatrix}\right)\right)\eta_{\boldsymbol{X}}(\boldsymbol{W})\,d\boldsymbol{W},\end{aligned}$$

where we have defined $\eta_{\boldsymbol{X}}(\boldsymbol{W}) = \prod_{d=1}^{3}\mathcal{N}\left(\boldsymbol{W}_{:,d} \mid \boldsymbol{A}_{2:}\boldsymbol{X}_{:,d}, \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)$ to be the marginal distribution over $\mathbb{R}^{N\times 3}$, with $\boldsymbol{A} = \begin{bmatrix}1 & \mathbf{0}_{N-1}^T\\ -\mathbf{1}_{N-1} & \mathbf{I}_{N-1}\end{bmatrix}$.

Note that $\eta_{\boldsymbol{X}}$ is the product of three independent multivariate normal distributions, i.e. all columns are independent. In the following, let $\eta_{\boldsymbol{X}_{:,:2}}$ be the distribution of the first two dimensions and $\eta_{\boldsymbol{X}_{:,3}}$ be the distribution of the third dimension, i.e.

$$\eta_{\boldsymbol{X}}(\boldsymbol{W}) = \eta_{\boldsymbol{X}_{:,:2}}\left(\boldsymbol{W}_{:,:2}\right)\eta_{\boldsymbol{W}_{:,3}}(\boldsymbol{Z}_{:,3}).$$

Similarity to invariance under elemental rotation. Next, we transform into cylindrical coordinates. Because we are in the case d^*, this is equivalent to transforming $\boldsymbol{Z}_{:,:2}$ into polar coordinates (recall our discussion in Appendix D.6). For this, let $\lambda(\boldsymbol{r},\boldsymbol{\theta}) = [\boldsymbol{r}\odot\cos(\boldsymbol{\theta}) \quad \boldsymbol{r}\odot\sin(\boldsymbol{\theta})]$ be the function that maps from polar into Cartesian coordinates.

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\gamma(\boldsymbol{Z}))\right]\\ &= \int_{\mathbb{R}^{(N-1)}} \eta_{\boldsymbol{X}_{:,3}(\boldsymbol{W}_{:,3})}\Bigg(\\ &\int_{\mathbb{R}^{(N-1)}}\left(\prod_{n=1}^{N-1} r_n\right)\int_{[0,2\pi]} h\left(\gamma_{\text{rot}}\left(\begin{bmatrix}\mathbf{0}_2^T & 0\\ \lambda(\boldsymbol{r},\boldsymbol{\theta}) & \boldsymbol{Z}_{:,3}\end{bmatrix}\right)\right)\eta_{\boldsymbol{X}_{:,:2}}(\lambda(\boldsymbol{r},\boldsymbol{\theta}))\,d\boldsymbol{r}\,d\boldsymbol{\theta}\Bigg)\,d\boldsymbol{W}_{:,3}\end{aligned}$$

This expectation is identical to Eq. (73) from Appendix D.8, save for the independent factor $\eta_{\boldsymbol{X}_{:,3}(\boldsymbol{W}_{:,3})}$ in the densitiy function. We may thus perform the same derivations, i.e. further substitutions, marginalization, application of the Neyman Pearson lemma and transformation back into Cartesian coordinates, to show that

$$\min_{h\in\mathbb{H}_{\text{T}}}\Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right] = \Pr_{\boldsymbol{W}_{:,3}\sim\eta_{\boldsymbol{X}'_{:,3}},\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)}\left[\frac{\eta_{\boldsymbol{X}'_{:,3}}(\boldsymbol{W}_{:,3})}{\eta_{\boldsymbol{X}_{:,3}}(\boldsymbol{W}_{:,3})}\rho(\boldsymbol{q}) \leq t\right] \tag{86}$$

$$\text{with } t\in\mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{W}_{:,3}\sim\eta_{\boldsymbol{X}_{:,3}},\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(2)},\boldsymbol{\Sigma}\right)}\left[\frac{\eta_{\boldsymbol{X}'_{:,3}}(\boldsymbol{W}_{:,3})}{\eta_{\boldsymbol{X}_{:,3}}(\boldsymbol{W}_{:,3})}\rho(\boldsymbol{q}) \leq t\right] = p_{\boldsymbol{X},y^*} \tag{87}$$

$$\text{and } \rho(\boldsymbol{q}) = \mathcal{I}_0\left(\sqrt{q_1^2+q_2^2}\right)/\mathcal{I}_0\left(\sqrt{q_3^2+q_4^2}\right), \tag{88}$$

Likelihood ratio distribution. The last step of our derivation is to find the distribution of the likelihood ratios in Eqs. (86) and (87). Recall that $\eta_{:,3}(\boldsymbol{W}_{:,3}) = \mathcal{N}\left(\boldsymbol{W}_{:,3} \mid \boldsymbol{A}_{2:}\boldsymbol{X}_{:,3}, \sigma^2\boldsymbol{A}_{2:}(\boldsymbol{A}_{2:})^T\right)$. Evidently, it is an affine transformation $\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3}$ of the random variable $\boldsymbol{Z}_{:,3}\in\mathbb{R}^N$ which follows isotropic normal distribution $\mu_{\boldsymbol{X}_{:,3}}(\boldsymbol{Z}_{:,3}) = \mathcal{N}\left(\boldsymbol{Z}_{:,3} \mid \boldsymbol{X}_{:,3}, \sigma^2\mathbf{I}_n\right)$. We may thus restate Eqs. (86) and (87) as

$$\min_{h\in\mathbb{H}_{\text{T}}}\Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}\left[h(\boldsymbol{Z}) = y^*\right] = \Pr_{\boldsymbol{Z}_{:,3}\sim\mu_{\boldsymbol{X}'_{:,3}},\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)}\left[\frac{\mu_{\boldsymbol{X}'_{:,3}}(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3})}{\mu_{\boldsymbol{X}_{:,3}}(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3})}\rho(\boldsymbol{q}) \leq t\right] \tag{89}$$

$$\text{with } t\in\mathbb{R} \text{ s.t. } \Pr_{\boldsymbol{Z}_{:,3}\sim\mu_{\boldsymbol{X}_{:,3}},\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(2)},\boldsymbol{\Sigma}\right)}\left[\frac{\mu_{\boldsymbol{X}'_{:,3}}(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3})}{\mu_{\boldsymbol{X}_{:,3}}(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3})}\rho(\boldsymbol{q}) \leq t\right] = p_{\boldsymbol{X},y^*} \tag{90}$$

Next, we can compute the likelihood ratio. Define $\boldsymbol{B} = (\boldsymbol{A}_{2:}\boldsymbol{A}_{2:})^T$ and recall that $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$. Then:

$$
\begin{aligned}
&\frac{\mu_{\boldsymbol{X}'_{:,3}}(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3})}{\mu_{\boldsymbol{X}_{:,3}}(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3})} \\
=&\frac{\mathcal{N}\left(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3} \mid \boldsymbol{A}_{2:}\boldsymbol{X}'_{:,3}, \sigma^2\boldsymbol{B}\right)}{\mathcal{N}\left(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3} \mid \boldsymbol{A}_{2:}\boldsymbol{X}_{:,3}, \sigma^2\boldsymbol{B}\right)} \\
=&\frac{\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,3} - \boldsymbol{A}_{2:}\boldsymbol{X}'_{:,3}\right)^T \boldsymbol{B}^{-1}\left(\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,3} - \boldsymbol{A}_{2:}\boldsymbol{X}'_{:,3}\right)\right)}{\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,3} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,3}\right)^T \boldsymbol{B}^{-1}\left(\boldsymbol{A}_{2:}\boldsymbol{\Delta}_{:,3} - \boldsymbol{A}_{2:}\boldsymbol{X}_{:,3}\right)\right)} \\
=&\exp\left(\frac{1}{2\sigma^2}\left(\boldsymbol{\Delta}_{:,3}^T(\boldsymbol{A}_{2:})^T\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}\left(2\boldsymbol{Z}_{:,3} - \boldsymbol{\Delta}_{:,3} - 2\boldsymbol{X}_{:,3}\right)\right)\right)
\end{aligned}
$$

Now, recall from Eq. (26) of our derivation of the translation invariance certificate that

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

i.e. $(\boldsymbol{A}_{2:})^T\boldsymbol{B}^{-1}\boldsymbol{A}_{2:}$ is the matrix that subtracts the mean from every vector it is multiplied with. Further recall Lemma 12, i.e. the inner product between a centered and an uncentered vector is identical to centering both vectors and then computing the inner product. Thus

$$
\frac{\mu_{\boldsymbol{X}'_{:,3}}(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3})}{\mu_{\boldsymbol{X}_{:,3}}(\boldsymbol{A}_{2:}\boldsymbol{Z}_{:,3})} = \exp\left(\frac{1}{2\sigma^2}\left(2\tilde{\boldsymbol{\Delta}}_{:,3}^T\boldsymbol{Z}_{:,3} - ||\tilde{\boldsymbol{\Delta}}_{:,3}||_2^2 - 2\tilde{\boldsymbol{\Delta}}_{:,3}\tilde{\boldsymbol{X}}_{:,3}\right)\right).
$$

The argument of the exponential function is an affine transformation of random variable $\boldsymbol{Z}_{:,3}$, $o = \frac{1}{2\sigma^2}\left(2\tilde{\boldsymbol{\Delta}}_{:,3}^T\boldsymbol{Z}_{:,3} - ||\tilde{\boldsymbol{\Delta}}_{:,3}||_2^2 - 2\tilde{\boldsymbol{\Delta}}_{:,3}\tilde{\boldsymbol{X}}_{:,3}\right)$.

In Eq. (89), we have $\boldsymbol{Z}_{:,3} \sim \mu_{\boldsymbol{X}'_{:,3}} := \mathcal{N}\left(\boldsymbol{X}_{:,3} + \boldsymbol{\Delta}_{:,3}, \sigma^2\mathbf{I}_N\right)$ and thus $o \sim \mathcal{N}(u^{(1)}, s)$ with mean $u^{(1)} = \frac{1}{2\sigma^2}\left(2||\tilde{\boldsymbol{\Delta}}_{:,3}||_2^2 - ||\tilde{\boldsymbol{\Delta}}_{:,3}||_2^2\right) = \frac{1}{2\sigma^2}\left(||\tilde{\boldsymbol{\Delta}}_{:,3}||_2^2\right)$ and variance $s^2 = \frac{1}{\sigma^2}\tilde{\boldsymbol{\Delta}}_{:,3}^T\sigma^2\mathbf{I}_N\tilde{\boldsymbol{\Delta}}_{:,3}\frac{1}{\sigma^2} = \frac{1}{\sigma^2}||\tilde{\boldsymbol{\Delta}}_{:,3}||_2^2$. Note that we have again used Lemma 12 in computing the inner product between $\tilde{\boldsymbol{\Delta}}_{:,3}$ and the mean of our random distribution.

In Eq. (90) we have $\boldsymbol{Z}_{:,3} \sim \mu_{\boldsymbol{X}_{:,3}} := \mathcal{N}\left(\boldsymbol{X}_{:,3}, \sigma^2\mathbf{I}_N\right)$ and thus $o \sim \mathcal{N}(u^{(2)}, s)$ with mean $u^{(2)} = \frac{1}{2\sigma^2}\left(-||\tilde{\boldsymbol{\Delta}}_{:,3}||_2^2\right) = -\frac{1}{2\sigma^2}\left(||\tilde{\boldsymbol{\Delta}}_{:,3}||_2^2\right)$ and variance $s^2 = \frac{1}{\sigma^2}||\tilde{\boldsymbol{\Delta}}_{:,3}||_2^2$. Note that we have again used Lemma 12 in computing the inner product between $\tilde{\boldsymbol{\Delta}}_{:,3}$ and the mean of our random distribution.

Therefore, we have

$$
\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}}[h(\boldsymbol{Z}) = y^*] = \Pr_{o\sim\mathcal{N}(u^{(1)},s),\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(1)},\boldsymbol{\Sigma}\right)}[\exp(o)\rho(\boldsymbol{q}) \leq t] \tag{91}
$$

$$
\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{o\sim\mathcal{N}(u^{(2)},s),\boldsymbol{q}\sim\mathcal{N}\left(\boldsymbol{m}^{(2)},\boldsymbol{\Sigma}\right)}[\exp(o)\rho(\boldsymbol{q}) \leq t] = p_{\boldsymbol{X},y^*} \tag{92}
$$

$$
\text{and } \rho(\boldsymbol{q}) = \mathcal{I}_0\left(\sqrt{q_1^2+q_2^2}\right)/\mathcal{I}_0\left(\sqrt{q_3^2+q_4^2}\right), \tag{93}
$$

which concludes the proof for our final tight certificate.

D.9 Monte Carlo certification

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

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

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, $\Pr_{V\sim\upsilon^{(1)}}\left[\rho(V)\leq t\right]$ is monotonically decreasing in t. Furthermore, t is monotonically decreasing with $p_{\boldsymbol{X},y^*}$: If $p_{\boldsymbol{X},y^*}$ decreases, then a smaller t is sufficient for fulfilling the constraint $\Pr_{V\sim\upsilon^{(2)}}\left[\rho(V)\leq t\right]$. 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 $t\in\mathbb{R}$ that fulfills $\Pr_{V\sim\upsilon^{(2)}}\left[\rho(V)\leq t\right]=\underline{p_{\boldsymbol{X},y^*}}$, i.e. $\underline{t}\leq t$
3. and then lower-bounding $\Pr_{V\sim\upsilon^{(1)}}\left[\rho(V)\leq \underline{t}\right]$.

The random variable $\mathbb{1}\left[g(\boldsymbol{Z})=y^*\right]$ is a Bernoulli random variable. Just like other randomized smoothing methods (e.g. [27]), 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 [132].

Lower-bounding t 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 [133] and Algorithm 1 below).

The random variable $\left[\rho(V)\leq\underline{t}\right]$ 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 [125] correction to account for the multiple comparisons problem.[3] 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 the Binomial test with 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.

[3] Holm-Bonferroni correction has already been used in the context of randomized smoothing, see [88].

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($\rho(V^{(1)}), \ldots, \rho(V^{(N_2)})$)
  $\underline{t} \leftarrow R^{(n^*)}$ ▷ Second bound
  $V^{(1)}, \ldots, V^{(N_3)} \leftarrow$ SAMPLE($\upsilon^{(1)}$, N_3)
  $\text{count}_2 \leftarrow \sum_{n=1}^{N_3} \mathbb{1}\left[\rho(V) \leq \underline{t}\right]$
  return LOWERCONFBOUND(count_2, N_3, $1-\frac{\alpha}{3}$) ▷ Third bound

E Proof of Theorem 6

Next, we prove that the post-processing-based certificate for rotation-invariance, i.e. invariance under $\mathbb{T} = \left\{\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^T \mid \boldsymbol{R} \in SO(D)\right\}$ is not tight, which we formalized in Section 6.3 as follows:

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

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} \left[h(\boldsymbol{Z}) = y^*\right] > \Phi\left(\Phi^{-1}\left(p_{\boldsymbol{X},y^*}\right) - \frac{\left|\left|\boldsymbol{X}'\boldsymbol{R}^T - \boldsymbol{X}\right|\right|_2}{\sigma}\right). \tag{94}$$

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.

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

with $\mathbb{H} = \left\{h : \mathbb{R}^{N\times D} \to \mathbb{Y} \mid \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}\left[h(\boldsymbol{Z}) = y^*\right] \geq p_{\boldsymbol{X},y^*}\right\}$. The 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 C.3 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.

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

Thus, we can prove Theorem 6 by proving the following, more general statement:

Lemma 14. *Let $g : \mathbb{R}^{N\times D} \to \mathbb{Y}$ be rotation invariant 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:*

$$\min_{h\in\mathbb{H}_{\mathbb{T}}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} \left[h(\boldsymbol{Z}) = y^*\right] > \min_{h\in\mathbb{H}} \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}'}} \left[h(\boldsymbol{Z}) = y^*\right] \tag{95}$$

We can do so by showing that no optimal solution to the r.h.s. black-box problem from Eq. (95) 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. (95) is given by the Neyman-Pearson lemma [104]. Later work on hypothesis testing has shown that the Neyman-Pearson lemma is not only a sufficient, but a necessary condition for optimality[134]: 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 2), this means that any classifier h that is an optimal solution to the r.h.s. black-box problem from Eq. (95) must fulfill

$$h(\boldsymbol{Z}) \in \begin{cases} \{y^*\} & \text{if } \frac{\mu_{\boldsymbol{X}'}(\boldsymbol{Z})}{\mu_{\boldsymbol{X}}(\boldsymbol{Z})} < t \\ \mathbb{Y} \setminus \{y^*\} & \text{if } \frac{\mu_{\boldsymbol{X}'}(\boldsymbol{Z})}{\mu_{\boldsymbol{X}}(\boldsymbol{Z})} > t \end{cases}$$

for some $t \in \mathbb{R}$, save for sets of zero measure. Cohen et al. [27] 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

$$h(\boldsymbol{Z}) \in \begin{cases} \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z}\rangle_{\mathrm{F}} + b < 0 \\ \mathbb{Y} \setminus \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z}\rangle_{\mathrm{F}} + b > 0 \end{cases} \tag{96}$$

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 14 by showing that classifiers complying with Eq. (96) (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 15. *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*

$$h(\boldsymbol{Z}) \in \begin{cases} \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b < 0 \\ \mathbb{Y} \setminus \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b > 0 \end{cases} \tag{97}$$

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. (97).

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 16. *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*

$$h(\boldsymbol{Z}) \in \begin{cases} \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b < 0 \\ \mathbb{Y} \setminus \{y^*\} & \text{if } \langle \boldsymbol{W}, \boldsymbol{Z} \rangle_{\mathrm{F}} + b > 0 \end{cases} \tag{98}$$

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. (97).

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 15 and 16 combined prove Lemma 14, which in turn proves Theorem 6.

F Multi-class certificates

In our discussions in Sections 3, 5 and 6, we have only considered binary certificates[4]. 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 [27], 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 [27, 135].

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 D.1:

Lemma 3 (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 $t \in \mathbb{R}_+$, the set $\left\{z \mid \frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} = t\right\}$ 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*

$$\left(\max_{h:\mathbb{A}\to\mathbb{Y}} \Pr_{z\sim\mu_{\boldsymbol{X}'}}[h(z)=y] \text{ s.t. } \Pr_{z\sim\mu_{\boldsymbol{X}}}[h(z)=y] \leq p\right) = \mathbb{E}_{z\sim\mu_{\boldsymbol{X}'}}[h^*(z)]$$
$$\text{with } h^*(z) = \mathbb{1}\left[\frac{\mu_{\boldsymbol{X}'}(z)}{\mu_{\boldsymbol{X}}(z)} \geq t\right] \text{ and } t \in \mathbb{R}_+ \text{ s.t. } \mathbb{E}_{z\sim\mu_{\boldsymbol{X}}}[h^*(z)] = p.$$

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 $||\boldsymbol{\Delta}|| \leq \frac{\sigma}{2}\left(\Phi^{-1}(p_{\boldsymbol{X}',y^*}) - \Phi^{-1}(p_{\boldsymbol{X}',y'})\right)$ [27].

Evidently, the post-processing-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 D, but using the Neyman-Pearson upper bounds from Lemma 3, i.e. replacing "$\leq$" with "$\geq$" and vice-versa. (see Lemma 3). 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

$$\Pr_{V\sim\upsilon^{(1)}}[\rho(V) \geq t]$$
$$\text{with } t \in \mathbb{R} \text{ s.t. } \Pr_{V\sim\upsilon^{(2)}}[\rho(V) \geq t] = p_{\boldsymbol{X},y'}$$
$$\text{and } p_{\boldsymbol{X},y^*} := \Pr_{\boldsymbol{Z}\sim\mu_{\boldsymbol{X}}}[g(\boldsymbol{Z}) = y'],$$

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

Our Monte Carlo certification procedure proposed in Appendix D.9 can also easily be adapted to computing an upper bound by replacing "$\leq$" with "$\geq$" and vice-versa.:

[4] 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($\rho(V^{(1)}), \ldots, \rho(V^{(N_2)})$)
  $\underline{t} \leftarrow R^{(n^*)}$ ▷ Second bound
  $V^{(1)}, \ldots, V^{(N_3)} \leftarrow$ SAMPLE($\upsilon^{(1)}$, N_3)
  $\text{count}_2 \leftarrow \sum_{n=1}^{N_3} \mathbb{1}\left[\rho(V) \geq \underline{t}\right]$
  return UPPERCONFBOUND(count_2, N_3, $1 - \frac{\alpha}{3}$) ▷ Third bound

Note that we still need to lower-bound threshold t, 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.

G 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

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

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

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

To ensure a fair comparison with our baselines, we modify the Monte Carlo certification procedure proposed in Appendix D.9 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($\rho(V^{(1)}), \ldots, \rho(V^{(N_2)})$)
  $\bar{t} \leftarrow R^{(n^*)}$ ▷ First bound
  $V^{(1)}, \ldots, V^{(N_2)} \leftarrow$ SAMPLE($\upsilon^{(2)}$, N_3)
  $\text{count}_2 \leftarrow \sum_{n=1}^{N_2} \mathbb{1}\left[\rho(V) \leq \bar{t}\right]$
  return UPPERCONFBOUND(count_2, N_2, $1 - \frac{\alpha}{2}$) ▷ Second bound

Note that, different from Appendix D.9 we compute a probabilistic upper bound on threshold t 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.

H Parameter space of the tight certificate for 2D rotation invariance

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.

H.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:

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

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 $\sqrt{\epsilon_1^2 + \epsilon_2^2} \leq ||\boldsymbol{X}||_2 \cdot ||\boldsymbol{\Delta}||_2$.

H.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

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