Certified Defenses Against Near-Subspace Unrestricted Adversarial Attacks

Abstract

The security implications of adversarial examples for neural networks have motivated the adversarial learning community to obtain certifiable defenses against adversarial attacks. These certification schemes *obtain provable lower bounds, called robustness certificates, on the level of adversarial contamination per input sample up to which the classifier output is accurate. However, most existing certification schemes do not exploit sparse or low-rank properties of the input data distribution, leading to certificates that are too small, require expensive high probability computations, and are applicable to ℓ_p bounded contaminations. In this work, we assume that our data approximately lies in a union of low-dimensional linear subspaces, and develop a theory of adversarial robustness for subspace-sparse classifiers. The resultant geometric understanding of the behavior of our classifiers enables us to obtain norm-independent certification regions. In other words, we can provably defend against specific unrestricted adversarial attacks.*

1. Motivation and Contributions

Research in adversarial learning has shown over time that traditional neural network based classification models are extremely prone to adversarial perturbations, which lead to large degradations in the accuracy of classifiers. Accordingly, researchers have obtained *defenses* against such attacks, which can be classified into two broad types: empirical and certified. Empirical defenses [8, 13, 14, 17] modify the training algorithm, or employ preprocessing to obtain classifiers with improved performance against adversarially corrupted inputs. Such defenses are empirically observed to obtain moderate classification accuracy against attacks known in the literature, but do not provide a theoretical guarantee of performance under attack. Certified defenses [1, 2, 6, 10, 18] remedy this by providing provable lower bounds on their *certified accuracy* under any attack in a specific attack model, which specifies the way the adversary is allowed to modify the input.

In this work, we advocate for explicitly modelling the input data distribution while designing classifiers as well as their associated certification schemes. We work in a setting where our data distribution is supported near a union of low-dimensional linear subspaces. In this setting, we demonstrate that explicitly using the knowledge of the data distribution enables us to formulate classifiers that are inherently robust, without the need of expensive post-processing, like randomized smoothing [2]. We also demonstrate that the resultant certified regions have clear geometric representations, leading to clear and efficient certification algorithms, in contrast to the intricate schemes used by existing methods, like convex-relaxation-based approaches [16]. Specifically, we make two contributions in this work.

First, in the *on-subspace* case, our data domain $\mathcal{X}$ is exactly a union of low-dimensional linear subspaces, *i.e.*, $S_1, S_2, \ldots, S_k$, and the attack model is unbounded, but constrained to perturbations lying on the subspaces. We construct a robust classifier and obtain its associated certificates. We demonstrate that the certified regions have a simple description as the convex conic hull of selected points in the training data.

Second, in the *out-of-subspace* case, our data domain $\mathcal{X}$ is a union of low-dimensional linear subspaces perturbed by additive ℓ_2 bounded noise B_ϵ, and the attack model (which is still unbounded) now includes out-of-subspace perturbations. We generalize the classifier obtained earlier and obtain its associated certificates. We demonstrate that the certified region is now a general polyhedron, whose faces and extreme rays are described by selected training data-points.

We note here for clarity that our certified regions are not spherical in any ℓ_p norm, in contrast to most existing literature. Our certificates are *descriptions of conical regions* in constrast to *certification radii* obtained by existing literature [2, 18]. Hence, our methods provably defend against adversarial perturbations unrestricted in the ℓ_p norm. These perturbations produce semantically meaningful images, but are restricted to be near linear subspaces $S_1, \ldots, S_k$. Our methods complement related work on non-isotropic certificates [4, 5, 12] which do not model the underlying data distribution.

The remainder of this short paper is organized as follows. Sec. 2 introduces preliminaries and notation that we will use throughout. Sec. 3 obtains a classifier and the resultant certified regions for on-subspace attacks, followed by Sec. 4 which deals with out-of-subspace attacks. Finally, we show some qualitative examples in Fig. 3, and conclude in Sec. 5.

2. Notation, Problem Formulation and Main Results

In this paper, we are concerned with robust classification for data lying near the union of low-dimensional linear subspaces. Throughout this work, we assume access to a training dataset of M *clean* data-points $(\mathbf{s}_1, y_1), (\mathbf{s}_2, y_2), \dots, (\mathbf{s}_M, y_M)$ such that for all i, the point $\mathbf{s}_i$ lies on a union of K low-dimensional linear subspaces $S_1 \cup S_2 \cup \dots \cup S_K$. The corresponding label, given by a labelling function $y_i = \textsc{Label}(\mathbf{s}_i)$ identifies the associated linear subspace, *i.e.*, $\mathbf{s}_i \in S_{y_i}$. We will use the notation $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_M]$ for the training-data matrix, $\mathbf{y} = (y_1, y_2, \dots, y_M)$ for the training-labels, $\mathcal{X}$ for the data-domain, and $\mathcal{Y} = \{1, 2, \dots, K\}$ for the label domain. In this section, we will describe the problem formulation in the case where $\mathcal{X}$ is *on-subspace*, *i.e.*, $\mathcal{X} = S_1 \cup S_2 \cup \dots \cup S_K$. In Sec. 4, we will move to the case where points in the data-domain can lie slightly outside their linear subspaces.

Data-Space Restriction. The astute reader would notice that since the label y is determined exactly given the subspace that a data point $\mathbf{x} \in \mathcal{X}$ belongs to, the setup is not well defined when $\mathbf{x}$ belongs to the intersection of multiple subspaces. To remedy this, we remove all pairwise subspace intersections from $\mathcal{X}$, to obtain the restricted space $\bar{\mathcal{X}} = \cup_{k \in [K]} \bar{S}_k$, where the restricted subspaces are defined as $\bar{S}_k = S_k \setminus (\cup_{k' \neq k} S_k \cap S_{k'})$.

Problem Formulation. Given $\{(\mathbf{s}_i, y_i) \in \bar{\mathcal{X}} \times \mathcal{Y}\}_{i=1}^M$ the problem that we aim to solve in this paper, is to predict the label y for an arbitrary data-point $\mathbf{x} \in \bar{\mathcal{X}}$, in a fashion robust to arbitrary additive perturbations to $\mathbf{x}$. Before proceeding, we ponder over the cases when this goal is achievable.

Robustness. Can we expect a classifier $f\colon \bar{\mathcal{X}} \to \mathcal{Y}$ to be robust to an arbitrary additive perturbation $\mathbf{v}$, *i.e.*, $f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x})$? *No*, firstly because the data-domain $\bar{\mathcal{X}}$ consists of low-dimensional linear subspaces, and it might be possible that $\mathbf{x}' = \mathbf{x} + \mathbf{v} \notin \bar{\mathcal{X}}$ for an arbitrary $\mathbf{v}$, implying that f is undefined at the perturbed point $\mathbf{x}'$. Secondly, for a valid adversarial attack, the *true* label must also remain unchanged after a perturbation, *i.e.*, $\textsc{Label}(\mathbf{x} + \mathbf{v}) = \textsc{Label}(\mathbf{x})$. This is akin to saying that a perturbation should not be so strong as to change an image of a cat to that of a dog, *i.e.*, the perturbed image must remain *on-subspace*. In other words, we obtain the *attack model* at $\mathbf{x}$ as $V(\mathbf{x}) = \bar{S}_{\textsc{Label}(\mathbf{x})} - \mathbf{x}$.

The above attack model ensures that for all $\mathbf{x} \in \bar{\mathcal{X}}$, we have (a) $\mathbf{x}+\mathbf{v} \in \bar{\mathcal{X}}$ and (b) $\textsc{Label}(\mathbf{x}+\mathbf{v}) = \textsc{Label}(\mathbf{x})$ for all $\mathbf{v} \in V(\mathbf{x})$. Notice that this attack model is *unbounded*, as there is no norm bound on the attack vector $\mathbf{v}$. Additionally, $V(\mathbf{x})$ is also maximal, in the sense that any larger set of additive perturbations would violate either (a) or (b).

Given the above robustness setup, we can refine our problem formulation. Given training data $\mathbf{S}, \mathbf{y}$, our goal is to obtain a classifier $f\colon \bar{\mathcal{X}} \to \mathcal{Y}$ that is (I) robust to additive perturbations in V, *i.e.*, $f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x})$ $\forall \mathbf{v} \in V(\mathbf{x})$ and (II) accurate, *i.e.*, $f(\mathbf{x}) = \textsc{Label}(\mathbf{x})$, for any $\mathbf{x} \in \bar{\mathcal{X}}$.

Main Results. In our first main result, Theorem 3.1, we construct a robust classifier g and the associated *certified regions* $C(\mathbf{x})$ around each $\mathbf{x} \in \mathcal{X}$ where g is robust. In other words,

$$g(\mathbf{x} + \mathbf{v}) = g(\mathbf{x}) \text{ whenever } \mathbf{x} + \mathbf{v} \in C(\mathbf{x}). \qquad (1)$$

Notice that (1) implies that if g is accurate at $\mathbf{x}$, *i.e.*, $g(\mathbf{x}) = y$, then g is also accurate under all perturbations $\mathbf{v}$ such that $\mathbf{x} + \mathbf{v} \in C(\mathbf{x})$. This observation is standard in the provable robustness literature [2], and is the standard way certification schemes are used to provide a certified accuracy for a given dataset. In our second main result, Theorem 4.1, we extend our data and attack models to the *out-of-subspace* case, where the $\mathcal{X}$ is not constrained to lie perfectly on a union of subspaces, and the attacker can make out-of-subspace perturbations.

3. Adversarially Robust Classification in the On-Subspace Case

Given a data-point $\mathbf{x} \in \bar{\mathcal{X}}$, recall that our goal is to predict the label y such that $\mathbf{x} \in \bar{S}_y$ in a fashion that is robust to perturbations within the attack model $V(\mathbf{x})$. Assuming that the training data $\mathbf{S}$ is diverse enough so that $\bar{\mathcal{X}} \subseteq \text{range}(\mathbf{S})$, we will represent $\mathbf{x}$ as a linear combination of the columns of $\mathbf{S}$, *i.e.*, $\mathbf{x} = \mathbf{S}\mathbf{c}$. In this way, we hope to recover the correct subspace from the labels y_i associated with the indices i in the support of $\mathbf{c}$, defined as $\text{supp}(\mathbf{c}) = \{i\colon c_i \neq 0\}$.

The problem now reduces to how to obtain $\mathbf{c}$ such that $\mathbf{x} = \mathbf{S}\mathbf{c}$ and the true subspace S_y can be recovered from $\text{supp}(\mathbf{c})$. Following the sparse-subspace classification literature we expect that the representation of $\mathbf{x}$ by a *small number* of columns of $\mathbf{S}$ would select columns belonging to S_y. This can be relaxed to an optimization program,

$$\min_{\mathbf{c}} \|\mathbf{c}\|_1 \text{ sub. to } \mathbf{x} = \mathbf{S}\mathbf{c}. \qquad (2)$$

(2) is known as the primal form of the Basis-Pursuit problem, and has been studied under a variety of conditions on $\mathbf{S}$ in the sparse representation and subspace clustering literature [3, 7, 9, 11, 15, 19]. We pause to understand the implications of such conditions for our problem.

Given an optimal solution $\mathbf{c}^*(\mathbf{x})$ of Problem (2), how can we accurately predict the label y? One ideal situation could be that all columns in the support predict the same label, *i.e.*, y_i is identical for all $i \in \text{supp}(\mathbf{c}^*(\mathbf{x}))$. Indeed, this ideal case is well studied, and is ensured by necessary [9] and sufficient [11, 15, 19] conditions on the geometry of the subspaces $S_1, \dots, S_K$. Another situation could be that the *majority* of the columns in the support predict the correct label, i.e., $\texttt{Majority}(\{y_i : i \in \text{supp}(\mathbf{c}^*(\mathbf{x}))\})$ leads to accurate prediction. In what follows, we obtain robustness guarantees which work for *any* such aggregation function which can determine a single label from the support. Hence, our results can guarantee accurate, robust prediction even when classical conditions are not satisfied.

Notice that there can be multiple solutions to Problem (2), with different supports. We found that this ambiguity causes difficulties in obtaining robustness certificates based on the support. To have a better control on the support of the solution, we turn to the dual problem for (2),

$$\max_{\mathbf{d}} \langle \mathbf{x}, \mathbf{d} \rangle \text{ sub. to } \mathbf{T}^\top \mathbf{d} \le \mathbf{1}, \tag{3}$$

where $\mathbf{T}$ is defined as the matrix containing the data-points and their negatives $\mathbf{T} = [\mathbf{S}, -\mathbf{S}]$. We will now show that the set of active constraints of the dual solution (3) is robust.

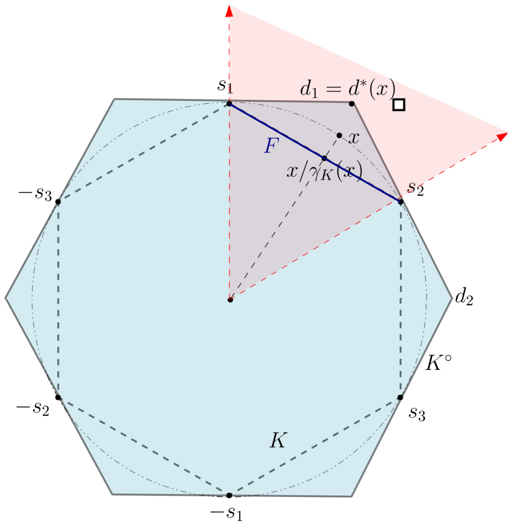

Figure 1. Geometry of the on-subspace dual problem (3). $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$ are the data-points. For all $\mathbf{x}'$ in the red shaded cone, Theorem 3.1 shows that the (minimal) set of active constraints are identical and equal to that at $\mathbf{d}^*(\mathbf{x})$. In other words, the dual classifier has the red cone as the unbounded certificate at $\mathbf{x}$.

Given a dual vector $\mathbf{d}$, we define the active constraint set B as $B(\mathbf{d}) = \{i : \langle \mathbf{t}_i^\top, \mathbf{d} \rangle = 1\}$. We denote the set of optimal solutions of the dual problem by $D^*(\mathbf{x})$. The *minimal* active constraint set at $\mathbf{x}$, denoted by $A(\mathbf{x})$, is now defined to be the columns of $\mathbf{T}$ that are active at all $\mathbf{d}^* \in D^*(\mathbf{x})$, *i.e.*, $A(\mathbf{x}) = \bigcap_{\mathbf{d}^* \in D^*(\mathbf{x})} B(\mathbf{d}^*)$. We pause to understand the set $A(\mathbf{x})$ by going back to Fig. 1. Whenever $\mathbf{x}'$ is in the interior of the red cone, the set $D^*(\mathbf{x}')$ has a single element, which is the vertex denoted in Fig. 1 by $\mathbf{d}_1$. In such a case, the intersection in $A(\mathbf{x})$ does not play any role, and set $A(\mathbf{x}')$ is simply the data-points $\mathbf{t}_1, \mathbf{t}_2$ contributing to the vertex $\mathbf{d}_1$. However, consider $\mathbf{x}' = \mathbf{s}_2$. Here, the set $D^*(\mathbf{s}_2)$ consists of the entire face of K° that contains $\mathbf{s}_2$. Specifically, $\mathbf{d}_1, \mathbf{d}_2 \in D^*(\mathbf{s}_2)$. For $\mathbf{d}_1$, we have the active constraints $B(\mathbf{d}_1) = \{\mathbf{t}_1, \mathbf{t}_2\}$, whereas for $\mathbf{d}_2$, we have the active constraints $B(\mathbf{d}_2) = \{\mathbf{t}_2, \mathbf{t}_3\}$. In this case, the intersection in $A(\mathbf{x})$ ensures that only the relevant data-point $\mathbf{t}_2$ remains in the intersection, thus ensuring that the *minimal* set of constraints $A(\mathbf{s}_2)$ is robust. With this understanding, we now have the following theorem showing that the minimal active set of constraints is robust.

Theorem 3.1. *The minimal active constraint set $A(\mathbf{x})$ is robust, i.e., $A(\mathbf{x}') = A(\mathbf{x})$ for all $\mathbf{x}' \in C(\mathbf{x})$ where the cone $C(\mathbf{x})$ is defined as $C(\mathbf{x}) = \left\{ \sum_{\mathbf{t}_i \in A(\mathbf{x})} \alpha_i \mathbf{t}_i : \alpha_i > 0 \right\}$*

The above result demonstrates that the minimal active constraint set $A(\mathbf{x})$ is robust in a certified region $C(\mathbf{x})$ around any point $\mathbf{x}$. This shows that the following *dual* classifier is certified to be robust around any point $\mathbf{x}$:

$$g(\mathbf{x}) = \text{AGGREGATE}(\{y_i : \mathbf{t}_i \in A(\mathbf{x})\}), \tag{4}$$

where AGGREGATE is any function that takes a set of labels and outputs a single label, following some aggregation scheme, e.g, *predict the majority label.*

Implications. Having obtained a certifiably robust classifier g, we pause to understand some implications of the theory developed so far. We observe that the certified regions in Theorem 3.1 are unbounded, *i.e.*, there exist directions in the attack model V where the attacker can make *unbounded* additive perturbations, but still they would be unable to change the label predicted by g. This is in stark contrast to the ℓ_p bounded certified regions that can be obtained by most existing work on certification schemes. This is a demonstration of the power of modelling low-dimensional structure while constructing robust classifiers.

4. Adversarially Robust Classification in the Out-of-Subspace Case

In this section, we generalize the data model $\mathcal{X}$ to tolerate ℓ_2 bounded perturbations and similarly the attack model to out-of-subspace attacks. We construct the dual classifier as earlier and then provide certificates on its robustness.

Our data domain $\mathcal{X}^\epsilon$ will be the union of K low-dimensional linear subspaces perturbed by bounded ℓ_2 noise, as $\mathbf{x} = \mathbf{s} + \mathbf{n}$ where $\mathbf{s} \in S_1 \cup S_2 \cup \ldots \cup S_K$, $\mathbf{n} \in B_\epsilon = \{\|\mathbf{n}\|_2 \le \epsilon\}$. The label $y \in \mathcal{Y}$ would be given by the subspace that $\mathbf{s}$ belongs to, such that $\mathbf{s} \in S_y$.

Data-Space Restriction. Similar to Sec. 3, we to ensure that any data-point $\mathbf{x}$ can be unambiguously labelled, we will define the restricted space by removing *fat* intersections as $\bar{\mathcal{X}}^\epsilon = \cup_{k \in [K]} \bar{S}_k^\epsilon$, where $\bar{S}_i^\epsilon = (S_i + B_\epsilon) \setminus (\cup_{j \neq i}((S_i + B_\epsilon) \cap (S_j + B_\epsilon)))$.

Robustness. We now seek robustness to additive, *out-of-subspace* perturbations $\mathbf{v}$ lying in the unbounded attack model $V^\epsilon(\mathbf{x}) = \bar{S}^\epsilon_{\text{LABEL}(\mathbf{x})} - \mathbf{x}$.

Recall that we assume access to a clean training dataset $\mathbf{S}$, and our goal is to obtain an accurate classifier $f\colon \bar{\mathcal{X}}^\epsilon \to \mathcal{Y}$ whose predictions are robust to the attack model V^ϵ. Given an out-of-subspace data-point $\mathbf{x} \in \bar{\mathcal{X}}^\epsilon$, we propose to construct our robust classifier by obtaining $\mathbf{c}$ such that the on-subspace counterpart $\mathbf{s}$ can be represented as a linear combination of a small number of columns of $\mathbf{S}$, *i.e.*, $\mathbf{x} = \mathbf{S}\mathbf{c} + \mathbf{n}$. This can be obtained as, $\min_{\mathbf{c}} \|\mathbf{c}\|_1$ sub. to $\|\mathbf{x} - \mathbf{S}\mathbf{c}\|_2 \leq \epsilon$, whose dual is

$$\mathbf{d}_\lambda^*(\mathbf{x}) = \left(\arg\min_{\mathbf{d}} \|\lambda \mathbf{x} - \mathbf{d}\|_2 \text{ sub. to } \mathbf{T}^\top \mathbf{d} \leq \mathbf{1} \right). \quad (5)$$

As done earlier, we define the set of active constraints as $A_\lambda(\mathbf{x}) = \{\mathbf{t}_i \colon \langle \mathbf{t}_i, \mathbf{d}_\lambda^*(\mathbf{x})\rangle = 1\}$. Note that $A_\lambda(\mathbf{x})$ is a much simpler definition than $A(\mathbf{x})$ as the set of optimal dual solutions now has a single element. Geometrically, $A_\lambda(\mathbf{x})$ identifies the face of K° which contains the projection of $\lambda\mathbf{x}$, if $A_\lambda(\mathbf{x})$ is non-empty. Otherwise, if $A_\lambda(\mathbf{x})$ is empty, then $\lambda\mathbf{x}$ lies inside the polyhedron K°.

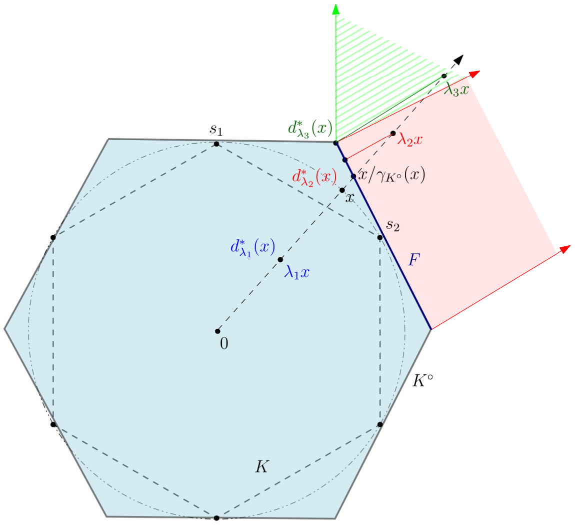

Figure 2. Geometry of the out-of-subspace dual problem (5). At $\lambda = \lambda_1$, the point $\lambda_1\mathbf{x}$ lies in the interior of K°. Hence, $S(\mathbf{x})$ is empty and $\operatorname{supp}(\mathbf{c}^*(\mathbf{x}))$ is also empty. As λ increases, a non-empty support is obtained for the first time at $\lambda = 1/\gamma_{K^\circ}(\mathbf{x})$. For all $\lambda_2\mathbf{x}$ in the red shaded polyhedron, the projection $\mathbf{d}_{\lambda_2}^*(\mathbf{x}) = \operatorname{Proj}_{K^\circ}(\lambda_2\mathbf{x})$ lies on the face F. As λ increases further we reach the green polyhedron. Further increases in λ do not change the dual solution, which will always remain at the vertex $\mathbf{d}_{\lambda_3}^*(\mathbf{x})$. Thus, depending on λ, the dual classifier enjoys the red or green region as the unbounded certificate for the data-point $\mathbf{x}$.

From Fig. 2, we can see visually that whenever $\mathbf{x}, \mathbf{x}'$ both lie in the same shaded polyhedron (red or green), their projections would lie on the same face of K°. This is shown formally in Theorem 4.1.

Theorem 4.1. *The set of active constraints A_λ is robust, i.e., $A_\lambda(\mathbf{x}') = A_\lambda(\mathbf{x})$ for all $\lambda\mathbf{x}' \in C(\mathbf{x})$, where the polyhedron $C(\mathbf{x})$ is defined as $C(\mathbf{x}) = F(\mathbf{x}) + V(\mathbf{x})$, with $F \subseteq K^\circ$ being a facet of the polyhedron K° that $\mathbf{x}$ orthogonally projects to, and V being the cone generated by the constraints active at (i.e., normal to) F.*

The above theorem shows that the following *dual* classifier is certified to be robust,

$$g_\lambda(\mathbf{x}) = \text{AGGREGATE}(\{y_i \colon \mathbf{t}_i \in A_\lambda(\mathbf{x})\}). \quad (6)$$

Theorem 4.1 then gives us a region $C(\mathbf{x})$ around $\mathbf{x}$ where the output of g_λ does not change.

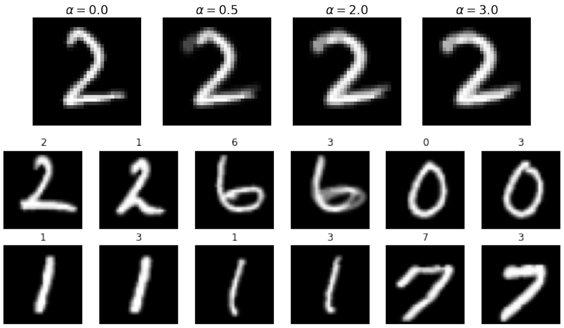

Figure 3. Top Row: Example showing our unrestricted certificates. We show $\mathbf{x} + \alpha\mathbf{v}$ for some $\mathbf{v} \in C(\mathbf{x})$, where $\|\mathbf{x}\|_2 = \|\mathbf{v}\|_2 = 1$. We are able to certify $g(\mathbf{x} + \alpha\mathbf{v}) = 2$ for a relative perturbation norm α much larger than existing methods. Bottom Rows: We train a standard neural network for classifying MNIST, and obtain large ℓ_2 norm adversarial perturbations $\mathbf{v}_\epsilon$ via Projected Gradient Descent, where we also project $\mathbf{v}_\epsilon$ to our certified regions. This ensures that $\mathbf{v}_\epsilon$ does not change the predicted class under our model, i.e., $g_\lambda(\mathbf{x}) = g_\lambda(\mathbf{x} + \mathbf{v}_\epsilon)$. Each pair of images shows $(\mathbf{x}, \operatorname{Proj}_{C(\mathbf{x})}(\mathbf{x} + \mathbf{v}_\epsilon))$, where $\mathbf{v}_\epsilon$ is the adversarial example with norm ϵ, Proj is the projection operator, and $C(\mathbf{x})$ is our certified region at $\mathbf{x}$. The title shows the class predicted by the NN, demonstrating that $\mathbf{v}_\epsilon$ makes the NN misclassify $\mathbf{x}$, while our method is certified to be correct. This demonstrates our resilience to unrestricted, but semantically meaningful, adversarial perturbations.

5. Conclusion

In this paper, we studied the question of adversarial robustness for a classification task where the training data lies on a union of low-dimensional linear subspaces. In this setting, we constructed robust classifiers for the unbounded, on-subspace and out-of-subspace attack models, and obtained their associated robustness guarantees.

References

[1] Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified defenses for adversarial patches. In *8th International Conference on Learning Representations (ICLR 2020)(virtual)*. In-

ternational Conference on Learning Representations, 2020. 1

[2] J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *ICML 2019*. 1, 2

[3] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003. 2

[4] 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. 1

[5] Ecenaz Erdemir, Jeffrey Bickford, Luca Melis, and Sergul Aydore. Adversarial robustness with non-uniform perturbations. *Advances in Neural Information Processing Systems*, 34:19147–19159, 2021. 1

[6] 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. 1

[7] Simon Foucart and Holger Rauhut. An invitation to compressive sensing. In *A mathematical introduction to compressive sensing*, pages 1–39. Springer, 2013. 2

[8] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. 1

[9] M. D Kaba, C. You, D. P. Robinson, E. Mallada, and R. Vidal. A nullspace property for subspace-preserving recovery. In *ICML 2021*. 2

[10] Alexander Levine and Soheil Feizi. (de) randomized smoothing for certifiable defense against patch attacks. *Advances in Neural Information Processing Systems*, 33:6465–6475, 2020. 1

[11] Chun-Guang Li, Chong You, and René Vidal. On geometric analysis of affine sparse subspace clustering. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1520–1533, 2018. 2

[12] Chen Liu, Ryota Tomioka, and Volkan Cevher. On certifying non-uniform bounds against adversarial attacks. In *International Conference on Machine Learning*, pages 4072–4081. PMLR, 2019. 1

[13] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016. 1

[14] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019. 1

[15] M. Soltanolkotabi and E. J Candes. A geometric analysis of subspace clustering with outliers. *Ann. Stats 2012*. 2

[16] E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML 2018*. 1

[17] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2019. 1

[18] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020. 1

[19] Chong You and René Vidal. Geometric conditions for subspace-sparse recovery. In *International Conference on Machine Learning*, pages 1585–1593. PMLR, 2015. 2