MODELING AND ELIMINATING ADVERSARIAL EXAMPLES USING FUNCTION THEORY OF SEVERAL COMPLEX VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliability of a learning model is key to the successful deployment of machine learning in various industries. Creating a robust model, particularly one that is unaffected by adversarial attacks require a comprehensive understanding of the adversarial examples phenomenon. In this paper, we present a model and a solution for the existence and transfer of adversarial examples in analytic hypotheses. Grounding in function theory of several complex variables, we propose the class of complex-valued holomorphic hypotheses as a natural way for representing the submanifolds of decision boundary and samples simultaneously and specialize the definitions of the optimal Bayes and the maximum margin classifiers to this class of hypotheses. The approach is validated initially on both synthetic and real-world classification problems using polynomials. Backed by theoretical and experimental results, we believe the analysis to be applicable to other classes of analytic hypotheses such as neural networks.

1 INTRODUCTION

The state-of-the-art neural models are shown to suffer from the phenomenon of adversarial examples, where an artificial neural network (ANN) is fooled to return an undesirable output on particular inputs that an adversary carefully crafts. The phenomenon is peculiar because it seems to affect neural networks in every application globally and that the adversarial examples are invariant to changes in network architecture, transferring from one network to another. From the perspective of learning theory, the existence of these samples is paradoxical since the nonrobust networks show acceptable, even super-human, performance on the natural samples. In the literature, many attempts at resolving this paradox have been made, each revealing a different facet of the phenomenon.

Szegedy et al. (2013) explained the adversarial examples as small pockets in the domain of the hypothesis, where the hypothesis fails to be correct due to its highly nonlinear nature. In contrast, Goodfellow et al. (2014) proposed that the phenomenon is a side-effect of a linear hypothesis in high dimensions. Conversly, Ilyas et al. (2019) blamed useful nonrobust features that are effective in dealing with natural samples; but are a hindrance when the model is tested on adversarial examples. Barati et al. (2021) unified these different and opposing perspectives under the banner of pointwise convergence of the hypothesis to the optimal hypothesis. However, they came short of describing the transfer of the adversarial examples.

On a separate thread, Tanay & Griffin (2016) came up with a geometrical description of the phenomenon in which adversarial examples are a byproduct of tilting the decision boundary towards the natural samples. Shamir et al. (2019) observed that adversarial examples could be a natural consequence of the geometry of $\mathbb{R}^n$ with a Hamming metric. Following that, Shamir et al. (2021) put forward the dimpled manifold model, in which the decision boundary weaves through the submanifold of the samples, sitting very close to the natural samples.

In this paper, we attempt to unify all these proposals under what we call the diverging derivative model by analyzing the phenomenon through the lens of functions of several complex variables. The use of complex variables and functions enables us to consider the algebraic and the geometric properties of the phenomenon simultaneously, leading to a rigorous framework to study the phenomenon and a possible remedy for its effects. For this reason, we introduce a novel approach in

describing the submanifolds of samples and decision boundary using a complex-valued hypothesis and continue by showing that the holomorphicity of this hypothesis is necessary for its robustness. We then introduce the space of holomorphic hypotheses and show that, in the limit of infinite samples, all holomorphic hypotheses are forced to converge to the orthogonal projection of the optimal Bayes classifier into the space of holomorphic hypotheses, explaining the transfer of adversarial examples between analytic hypotheses. Finally, by generalizing the linear maximum margin classification rule to the case of holomorphic hypotheses, we pave the way for mitigating the effects of the adversarial examples phenomenon.

2 THE SPACE OF HOLOMORPHIC HYPOTHESES

We review holomorphic functions and discuss their various properties in appendix A. Here, we first motivate complex-valued classifiers as the primary model for analyzing the adversarial examples phenomenon in binary classification problems and continue by introducing the space of holomorphic hypotheses in learning theory terms.

Given the observations made about the phenomenon in the literature, we can see that there are two submanifolds that we need to consider in our analysis. The submanifold of natural samples $\mathcal{S}$ and the submanifold of the decision boundary $\mathcal{C}$. We assume that geometrical position of $\mathcal{C}$ and $\mathcal{S}$ are represented using the equations $u(x) = 0$ and $v(x) = 0$ in which $u : \mathcal{X} \to \mathbb{R}$ and $v : \mathcal{X} \to \mathbb{R}$ are smooth functions from the domain of samples to real numbers. Thus, $h(x) = u(x) + iv(x)$ is a smooth function from the domain of samples to the complex plane that encodes the information about the two submanifolds simultaneously. Having h, we can determine the label of a sample by looking at the sign of the real part of $h(x)$. Similarly, we can determine if a sample is natural by considering the imaginary part of $h(x)$. These definitions are consistent with the standard notions of the submanifolds of decision boundary and samples. For example, in the case of a 1D real-valued classifier, the submanifold of the samples is the real line, which is a curve in the complex plane, and the decision boundary is represented by the zeros of the hypothesis in which the real part is zero naturally. Hence, we can simultaneously learn the submanifolds of samples and the decision boundary by solving a regression and a classification problem.

Tanay & Griffin (2016) have shown that $\mathcal{S}$ and $\mathcal{C}$ are perpendicular in a robust linear classifier. As a result, if u and v were linear functions of x, their contours had to be perpendicular everywhere in Ω in order for h to be robust. From complex analysis, we know that this condition is known as conformality and that conformality and holomorphicity are equivalent. First, we generalize the result of Tanay & Griffin (2016) to the case of smooth complex-valued hypotheses.

Theorem 2.1 *Suppose that $\mathcal{C}$ and $\mathcal{S}$ are represented by contours of real and imaginary parts of a smooth function $h : \mathcal{X} \to \mathbb{C}$,*

$$\mathcal{C}(h) = \{x \in \mathcal{X} \mid \Re[h(x)] = 0\}, \quad \mathcal{S}(h) = \{x \in \mathcal{X} \mid \Im[h(x)] = 0\}. \tag{1}$$

Suppose that $c \in \mathcal{C}$ is a point on the decision boundary and that $\Re[h(x)]$ is a robust classifier. Then, h is holomorphic in a neighborhood of c.

We see that a robust complex-valued hypothesis has to be holomorphic in a vicinity of the decision boundary. As a consequence of the Cauchy-Riemann condition, constant functions are the only holomorphic functions that could be defined on a set $\mathcal{X} \subset \mathbb{R}$. Thus, for h to be robust, we have to assume that either the domain of h is complex, or that it is constant. While the transition from domains in $\mathbb{R}$ to domains in $\mathbb{C}$ is a necessity from the standpoint of theorem 2.1, it has the added benefit of enabling geometrical interpretations in scenarios where the previous proposals could not be applied.

Barati et al. (2021) showed that the phenomenon could be observed in 1D problems. Since it is impossible to extend the decision boundary outside the submanifold of samples in such cases, this observation immediately conflicts with the dimpled manifold and the boundary tilting models. In other words, since the decision boundary has a lower dimension than the domain of the hypothesis, it is forced to be 0D in a 1D classification problem. How would a set of points wrap the training samples? Or, from the boundary tilting perspective, How would one describe the angle between a set of points and a line? To use the geometrical models of the phenomenon in 1D problems, we need to assume some space for extending the decision boundary. Extending the domain of the hypothesis

from $\mathbb{R}$ to $\mathbb{C}$ would provide this necessary space. Consequently, we replace $x \in \mathbb{R}^d$ with $z \in \mathbb{C}^d$, and replace $\mathcal{X} \subset \mathbb{R}^d$ with $\Omega \subset \mathbb{C}^d$ to symbolize the transition from real variables to complex ones.

Theorem 2.1 shows that holomorphicity is enforced on a robust hypothesis in a limited sense. Precisely, the theorem asserts that h is analytic in a neighborhood of a point on the decision boundary. We emphasize that theorem 2.1 does not specify anything about the characteristics of h outside of a vicinity of the decision boundary. It is possible to construct smooth non-analytic functions that satisfy the conditions of theorem 2.1, e.g. a meromorphic function. Nevertheless, we only consider holomorphic functions in this paper since analysis of analytic functions is a well-known subject and it is the first step in analysis of other less well-behaved hypotheses. Moreover, when we assume that h is holomorphic, we can interpret h as a complex chart from Ω to $\mathbb{C}$ by definition. As a result, $\mathcal{C}$ and $\mathcal{S}$ would be submanifolds of this complex analytic manifold in the same sense that $\mathbb{R}$ is a submanifold of $\mathbb{C}$, namely, they are smooth and locally Euclidean. Compact Riemann surfaces are a prime example of the complex manifolds that we are considering in this paper.

In order to analyze a holomorphic function as a hypothesis in a learning algorithm, we first have to define the space of holomorphic functions as a space of learnable hypotheses. It is customary in analysis to represent the space of holomorphic functions on a domain Ω with $\mathcal{O}(\Omega)$. Then, the Bergman space $A^2(\Omega)$ is defined as follows,

$$A^2(\Omega) = \{f \in \mathcal{O}(\Omega) \mid \int_\Omega |f(z)|^2 dV(z)^{\frac{1}{2}} \equiv \|f\|_{A^2(\Omega)} < \infty\}, \tag{2}$$

in which $dV(z)$ is the volume differential of Ω. Since holomorphic functions has a unique power series representation, we can deduce that the VC-dimension of $A^2(\Omega)$ is upper bounded by the VC-dimension of the space of polynomial hypotheses on Ω.

Theorem 2.2 *$A^2(\Omega)$ is nonuniform learnable.*

We emphasize that theorem 2.2 cannot be used to define a learnable space on all possible domains Ω. For example, if $\Omega = \mathbb{C}$, then the Bergman space does not exist. To give a complete account of the subject of domains of holomorphic functions, one needs to analyze it through the lens of domains of holomorphy. Nevertheless, as long as we choose Ω to be a convex and compact subset of $\mathbb{C}^d$, it is safe to ignore these subtleties. For example, there is no compact subset of $\mathbb{C}$ that cover $\mathbb{R}$ and thus theorem 2.2 could not be used to define a learnable hypotheses space that covers $\mathbb{R}$. On the other hand, $\mathbb{R} \cup \{\infty\}$ could be mapped to the unit circle using a Möbius transformation. Then, we can assume Ω to be a disk, which is compact and convex.

$A^2(\Omega)$ is a reproducing kernel Hilbert space with inner product $\langle f, g\rangle = \int_\Omega f(z)\overline{g(z)}dV(z)$. The Bergman kernel $K_\Omega(z, \zeta)$ is the unique function with the reproducing property

$$f(z) = \int_\Omega f(z)K_\Omega(z, \zeta)dV(\zeta), \quad \forall f \in A^2(\Omega). \tag{3}$$

Since $A^2(\Omega)$ is a subspace of $L^2(\Omega)$, an orthonormal basis $\{\varphi_j\}_{j=1}^\infty$ for $A^2(\Omega)$ exists. Due to holomorphicity, the hypotheses that we consider in this paper are the hypotheses that has a power series representation on their domain Ω,

$$h(z) = \sum_{|\alpha|\geq 0} w_\alpha \varphi_\alpha(z), \tag{4}$$

in which,

$$\begin{aligned} \gamma_\alpha &= \int_\Omega |z|^{2\alpha}\, dV(z), \\ \varphi_\alpha(z) &= \frac{z^\alpha}{\sqrt{\gamma_\alpha}}. \end{aligned} \tag{5}$$

In equation 4 and equation 5, we have used the multi-index notation,

$$\begin{aligned} \alpha =&(\alpha_1, \alpha_2, \cdots, \alpha_d) \quad \alpha_j \in \mathbb{N}_0, \\ |\alpha| =&\sum_{j=1}^{d} \alpha_j, \\ z^\alpha =&\prod_{j=1}^{d} z_j^{\alpha_j}. \end{aligned} \tag{6}$$

The next theorem is a well-known result in the function theory of several complex variables that provides a way for computing the Bergman kernel of Ω.

Theorem 2.3 *Let D be a compact subset of Ω. Then the series*

$$\sum_{j=1}^{\infty} \varphi_j(z)\overline{\varphi_j(\zeta)} \tag{7}$$

sums uniformly on $D \times D$ to the Bergman kernel $K_\Omega(z, \zeta)$.

A remarkable fact about theorem 2.3 is that it is true no matter what the choice of complete orthonormal basis $\{\varphi_j\}_{j=1}^{\infty}$ for $A^2(\Omega)$ is. In other words, if the features φ were represented using a Fourier series or some neural network instead of polynomials, equation 7 would still converge to the Bergman kernel of Ω as long as we keep the set of features complete and orthonormal on Ω.

The reproducing property of the Bergman kernel of Ω in conjunction with the fact that no hypothesis can get a better score than the optimal Bayes classifier provides the means to define the infinite sample limit of any learning rule on $A^2(\Omega)$ independently from the details of the implementation or training process.

Definition 2.1 (holomorphic optimal Bayes classifier) *The holomorphic optimal Bayes classifier is the orthogonal projection of the optimal Bayes classifier into $A^2(\Omega)$,*

$$o_{\mathcal{D}}(z) = \int_{\Omega} f_{\mathcal{D}}(\zeta) K_\Omega(z, \zeta)\, dV(\zeta), \tag{8}$$

in which $K_\Omega(z, \zeta)$ is the Bergman kernel of Ω and $f_{\mathcal{D}}$ is the optimal Bayes classifier.

As a consequence of definition 2.1, the adversarial examples of any two holomorphic hypotheses would be similar since they are converging to the same hypothesis as we introduce more samples to the training set.

Theorem 2.4 *Suppose that $h_1, h_2 \in A^2(\Omega)$ are two complex-valued hypotheses that are trained on two training sets $S_1, S_2 \sim \mathcal{D}$. Further, suppose that the size of the training samples are sufficiently large. Then, the adversarial examples of h_1 transfer to h_2 and vice versa.*

Theorem 2.4 provides a compelling explanation for the transfer of adversarial examples between different analytic hypotheses, such as neural networks or polynomials. Our proposal is an extension of the proposal of Goodfellow et al. (2014), in which the transfer of adversarial examples is attributed to convergence to the optimal linear classifier. Theorem 2.4 formally describes the transfer of adversarial examples in holomorphic hypotheses, including linear hypotheses.

3 MAXIMUM MARGIN CLASSIFICATION IN HOLOMORPHIC HYPOTHESES

In Tanay & Griffin (2016), it is argued that the adversarial examples phenomenon is a byproduct of a tilted decision boundary. It was also shown that linear maximum margin classifiers could be robust if regularized correctly. Here, we follow up this proposal by introducing the holomorphic maximum margin classifier, which generalizes the linear maximum margin classifiers to holomorphic hypotheses.

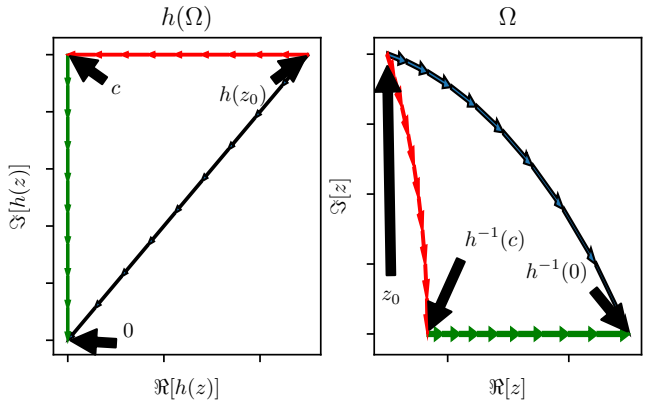

Figure 1: A depiction of the shortest path from a sample z_0 to $\mathcal{C}(h)$ and $\mathcal{Z}(h)$. the point $h^{-1}(c)$ represents the closest point on $\mathcal{C}$ to z_0.

From the perspective of a tilted boundary, the phenomenon occurs because the decision boundary somehow tilts towards the submanifold of samples. Since the hypotheses in $A^2(\Omega)$ are holomorphic, we can be assured that wherever $\mathcal{C}$ and $\mathcal{S}$ intersect, they form a right angle. Nevertheless, holomorphicity of h would not assert that the learning rule has maximized the margin of the decision boundary, which could result in the occurrence of adversarial regions. The connection between robustness and maximum margin classification has been studied before in the literature (Elsayed et al., 2018; Ding et al., 2020). A maximum margin classifier is a classifier that maximizes the distance between the training samples and the decision boundary. In general, the distance between two points on a manifold is defined as the length of the shortest curve that connects the two points without leaving the manifold. According to h, the shortest curve from a sample to $\mathcal{C}$ is the preimage of the straight path that connects the image of the sample z_0 to $\mathcal{C}$. Figure 1 illustrates the shortest path from z_0 to $\mathcal{C}$ in both the domain and the range of an arbitrary holomorphic hypothesis. The figure shows that the shortest path from a sample to the decision boundary is a curve and not a straight line in general.

Let $\mathcal{Z}(h) = \{z \in \Omega \,|\, h(z) = 0\}$ be the zero set of h. We can see that $\mathcal{Z}(h) = \mathcal{C}(h) \cap \mathcal{S}(h)$. Hence, the shortest path on the submanifold of the samples from $s \in \mathcal{S}(h)$ to $\mathcal{C}(h)$ is equal with the shortest path on the submanifold of samples from s to $\mathcal{Z}(h)$.

Theorem 3.1 *Consider a sample $z_0 \in \Omega \subset \mathbb{C}^d$ and a hypothesis $h \in A^2(\Omega)$. Suppose that $\gamma : [0, 1] \to \Omega$ is the shortest path from z_0 to $\mathcal{Z}(h)$. Then,*

$$\frac{d\gamma_j}{dt}(t) = \frac{dh}{\partial_j h(\gamma(t))} \quad j = \{1, \cdots, d\}. \tag{9}$$

Theorem 3.2 *Consider a hypothesis $h \in A^2(\Omega)$ and a sample $z_0 \in \Omega$. Suppose that γ_z and γ_c are the shortest paths from z_0 to $\mathcal{Z}(h)$ and $\mathcal{C}(h)$, respectively. Furthermore, assume that γ_s is the shortest path that connects the ends of γ_c and γ_z without leaving $\mathcal{C}(h)$. Then,*

$$\|\gamma_z\| - \|\gamma_s\| \leq \|\gamma_c\|, \tag{10}$$

With equality being true iff $\Im[h(z_0)] = 0$.

As a consequence of theorem 3.2, maximizing the margin of $h \in A^2(\Omega)$ is equivalent with minimizing $\|\gamma_s\|$ and maximizing $\|\gamma_z\|$ for every training sample. According to theorem 3.1, we can maximize $\|\gamma_z\|$ for all training samples by maximizing $|\frac{h(z)}{\partial_d h(z)}|^2$ everywhere in Ω. Moreover, by minimizing $|\Im[h(z_0)]|$ for all training samples, we can make the bound given by theorem 3.2 as tight as possible. In maximizing $|\frac{h(z)}{\partial_d h(z)}|^2$ everywhere in Ω, we cannot utilize $|h(z)|^2$ since we do not have access to $o_{\mathcal{D}}(z)$ and using $h(z)$ as a surrogate for $o_{\mathcal{D}}(z)$ would reinforce any wrong prediction of $h(z)$ as well. On the other hand, from definition 2.1 we know that $\partial_d o_{\mathcal{D}}(z)$ is as close to the zero function as possible since it is the orthogonal projection of a step function into $A^2(\Omega)$. Thus,

minimizing $|\partial_d h(z)|^2$ in Ω is a true heuristic for finding $o_{\mathcal{D}}(z)$. It is easy to see that minimizing $|\partial_d h(z)|^2$ in Ω is equivalent with minimizing the L^2 norm of $\partial_d h(z)$ for every dimension d. This result is in accordance with the success of regularizing the Jacobian of a hypothesis in training robust classifiers reported in literature (Ross & Doshi-Velez, 2018; Paknezhad et al., 2021).

Definition 3.1 (holomorphic maximum margin classifier) *Let h be a holomorphic hypothesis and let $S = \{(z_m, t_m) \in \Omega \times \{-1, 1\}\}$ be a set of training samples of size M. The holomorphic maximum margin classifier is the solution to the following program,*

$$\begin{aligned} \underset{h,\xi,\nu}{\arg\min} \quad & \lambda \int_{\Omega} \|\nabla h(z)\|^2 \, dV(z) + \frac{1}{M} \sum_{m=1}^{M} \xi_m + \nu_m \\ \text{subject to} \quad & 1 - \xi_m \leq t_m \Re[h(z_m)] \quad m = 1, \cdots, M, \\ & -\nu_m \leq \Im[h(z_m)] \leq \nu_m \quad m = 1, \cdots, M, \\ & 0 \leq \xi_m \quad m = 1, \cdots, M, \\ & 0 \leq \nu_m \quad m = 1, \cdots, M \end{aligned} \tag{11}$$

We can show that the linear maximum margin classifier is a special case of Definition 3.1. Assume that $h = w^H z + b$ is a linear classifier, then,

$$\|\nabla h(z)\| = \|w\|. \tag{12}$$

Since $\|w\|$ is independent of z, we can take it out of the integral. Thus, the remaining integral will be computing the volume of Ω which is a positive constant that does not depend on any of the optimization variables. Consequently, we can merge the integral with λ and we are left with a linear maximum margin classification program.

From the form of problem **??**, it is evident that the standard and holomorphic maximum margin classification are very similar and that we can make use of the kernel trick in the case of holomorphic maximum margin classification. Moreover, the notion of support vectors is the same in the standard maximum margin formulation and the holomorphic formulation, differing only in the addition of support vectors for the regression part of the holomorphic formulation.

4 SYNCHRONIZATION

According to Ilyas et al. (2019), the existence of adversarial examples is rooted in the existence of useful but nonrobust features. However, the perspective does not consider scenarios where nonrobust features are not only useful, but also essential to classify samples correctly. Furthermore, making use of nonrobust features does not automatically result in a nonrobust hypothesis. For example, consider the Fourier expansion of an optimal Bayes classifier, e.g. $\operatorname{sign}(x)$. The high-frequency Fourier bases are categorized as useful nonrobust features as stated by the definition of useful nonrobust features in such a scenario. However, we know that the Fourier expansion of the optimal Bayes classifier is robust and is equal with the optimal classifier and that the higher-frequency bases are essential for an accurate approximation of the classifier.

In this section we utilize definition 3.1 and introduce the concept of synchronized features as an improvement to the core idea of robust features of Ilyas et al. (2019).

Theorem 4.1 *Suppose that h is a holomorphic hypotheses with the following representation,*

$$h(z) = b + \sum_{j=1}^{n} w_j \varphi_j(z), \tag{13}$$

where $\{\varphi_j\}_{j=1}^n$ is a set of holomorphic features. Then, the dual problem of the holomorphic maximum hard margin program is as follows,

$$\begin{aligned} \underset{\lambda,\,\theta}{\arg\max} \quad & \sum_{m=1}^{M} \Re[\alpha_m] - \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} t_n t_m \overline{\alpha}_n \alpha_m \varphi_n^H \Sigma^{-1} \varphi_m, \\ \text{subject to} \quad & \sum_{m=1}^{M} t_m \alpha_m = 0, \\ & 0 \leq \lambda_m \quad m = 1, \cdots, M, \\ & 0 \leq \theta_m \quad m = 1, \cdots, M \end{aligned} \tag{14}$$

where $\alpha_m = \lambda_m + i\frac{\theta_m}{t_m}$ and

$$\Sigma = \int_\Omega J_\varphi(z) J_\varphi(z)^H \, dV(z), \tag{15}$$

where $J_\varphi(z)$ is the Jacobian matrix of $\varphi = [\varphi_j]$.

Σ is a positive definite square matrix that we call the synchronization matrix of $\{\varphi_j\}$. We can interpret Σ as a metric on the space of hypotheses spanned by φ and infer that the holomorphic maximum margin learning rule prioritizes hypotheses with smaller $\|w\|_\Sigma$. Computing the synchronization matrix paves the way to make use of nonrobust features robustly. Consequently, we see that dividing the useful features into robust and nonrobust categories is not helpful, and it is better to categorize feature sets instead of individual features. With robustness in mind, it is possible to isolate a particular category of feature sets identified with the property that their synchronization matrix is the identity matrix.

Definition 4.1 (synchronized feature set) *Let $\{\varphi_j\}$ be a set of features in which*

$$\int_\Omega (\nabla \varphi_j(z))^H (\nabla \varphi_k(z)) \, dV(z) = \begin{cases} 1 & \text{if } j \neq k \\ 0 & \text{if } j = k \end{cases}, \tag{16}$$

then $\{\varphi_j\}$ is a set of synchronized features and Ω is its synchronization domain.

For a set of synchronized features $\|w\|_\Sigma = \|w\|$. In other words, the objective of the holomorphic maximum margin learning rule coincides with minimizing the ℓ_2 norm of the weights. It should be clear from Definition 4.1 that the feature set of a linear hypothesis is synchronized up to a normalization constant. This fact explains the observation made in Tanay & Griffin (2016) about the effectiveness of ℓ_2 regularization of the weights in suppressing the adversarial examples phenomenon in linear classifiers. There is no reason to assume that the features of an artificial neural network are synchronized which explains the ineffectiveness of ℓ_2 regularization in training robust neural classifiers.

We can synchronize any set of features φ given the synchronization matrix Σ. Seeing that Σ is positive definite, its matrix square root exists, is unique, and it is a positive definite matrix. Thus, we can synchronize φ as follows,

$$\varphi^*(z) = \Sigma^{-\frac{1}{2}} \varphi(z). \tag{17}$$

Then, the synchronization matrix of φ^* would be the identity matrix,

$$\begin{aligned} \Sigma^* &= \int_\Omega J_{\varphi^*}(z) J_{\varphi^*}(z)^H \, dV(z), \\ &= \Sigma^{-\frac{1}{2}} \int_\Omega J_\varphi(z) J_\varphi(z)^H \, dV(z) \, \Sigma^{-\frac{H}{2}}, \\ &= \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{H}{2}} = I. \end{aligned} \tag{18}$$

5 THE DIVERGING DERIVATIVE MODEL

In this section, we put the proposed framework into practice and use it to analyze the phenomenon. Figure 2 visualizes the trained hypotheses by applying the maximum margin learning rule to a set

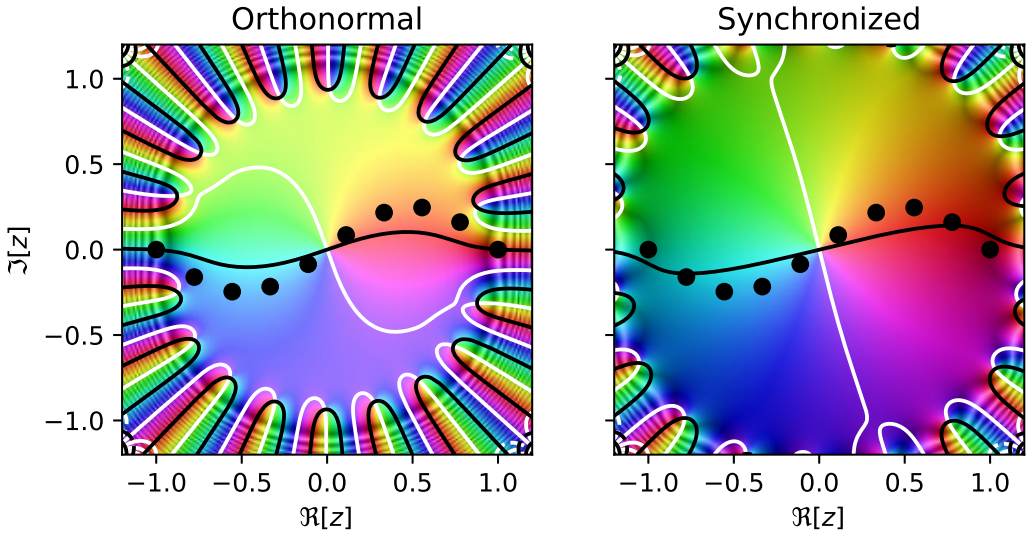

Figure 2: 1D orthogonal and synchronized hypotheses on the complex plane. The black and white contours represent the contours of imaginary and real parts of the hypotheses, respectively. The training samples are depicted using black dots. Hue and saturation of the colors represent the argument and the magnitude of $h(z)$, respectively.

of orthonormal and synchronized polynomial features on the unit disk $\mathbb{D}$. We omit the details of the implementation to make room for more content. The interested reader can find these details and more in appendix C. We can see in the figure that the submanifold of the decision boundary of the orthonormal hypothesis curves back and gets close to the submanifold of the samples, resulting in adversarial regions near the boundary of the disk. In contrast, the synchronized hypothesis does not show any adversarial regions in the unit disk.

Figure 2 shows that the region of the convergence of the power series corresponding to the orthonormal hypothesis does not cover the entirety of the unit disk. On the other hand, the convergence region of the synchronized hypothesis is a little bigger than the unit disk. This observation is the essence of our proposal; the diverging derivative model for the adversarial examples phenomenon. To be more precise, according to theorem A.1 in appendix A, a holomorphic hypothesis converges to the optimal hypothesis in both value and all of its derivatives. When we do not use a set of synchronized features, ℓ_2 regularization of the weights does not guide the learning rule to the hypothesis with the smallest derivative, the maximum margin hypothesis, but rather to the hypothesis with the smallest value. As a result, the derivative of the hypothesis is free to get as large as possible to fulfill both the training loss and the regularization objectives. In turn, this unbounded increase in the magnitude of the derivative shrinks the convergence region of the hypothesis, resulting in adversarial regions around the submanifold of the samples.

Compared to other proposals in the literature, the diverging derivative model has some unique advantages. First, it bridges the geometrical and functional approaches to describing the phenomenon. On top of that, the model connects these two approaches through the analyticity of the hypotheses space and enables the use of a host of powerful algebraic tools. Moreover, by representing the hypothesis by a power series, the diverging derivative model simplifies the analysis of different hypothesis by making the analysis of the hypothesis independent from the details of implementation of h, such as architecture. We provide more simulations and experiments in appendix C.

6 Conclusion and Future Research

This paper proposed a model to explain the existence and transfer of adversarial examples in holomorphic hypotheses. According to our proposal, the adversarial examples occur due to the divergence of the derivative of the trained hypothesis. These examples transfer to other analytic hy-

potheses because the projection of the optimal model into the space of holomorphic functions on a domain is uniquely decided by the Bergman kernel of the domain, and it is independent of the details of implementation or training. Moreover, we generalized the linear maximum margin classifier to holomorphic functions and introduced synchronized features as a solution to the phenomenon. We ground the model on rigorous analysis of the phenomenon and provide empirical evidence supporting the proposal.

Given the nature of the proposal, one may imagine that a similar approach could be applied to real analytic functions as well, and that there is no need to consider holomorphic functions and complex variables. Opposite to complex analytic functions, real analytic functions are not closed under uniform convergence. In other words, there is no guarantee that the limit of a sequence of analytic functions that converge uniformly is analytic. This is a major hurdle in considering real analytic functions as a learnable hypotheses space and complex variables and functions are unique in this regard.

The Achilles heel of our proposal is that computing the objective of problem 11 in high dimensions is NP-hard in the general case. In particular, implementing the holomorphic maximum margin classification for ANNs would require stochastic optimization techniques to the best of our knowledge (see appendix D). Nonetheless, we can see that it would be implementable for polynomials if we can come up with an efficient way to select useful polynomial features. Another promising approach to this problem is to use the dual formulation of problem 11, in which we would need the synchronized kernel of the domain or an approximation of it. On top of these, since the submanifold of samples is explicitly determined in a complex-valued hypothesis, our proposal could find use in implementing the solutions mentioned in Shamir et al. (2021) by providing a way to project the test sample into the submanifold of natural samples.

Another shortcoming of our approach is that it is based on the analysis of a binary classification problem. Even though our experiments suggest that the analysis generalizes to the multiclass classification problem, extending the proposal's reach in its current form to other machine learning applications does not seem feasible. We believe that the most helpful analysis would be when a sufficient robustness condition is derived from the training loss function. Our approach in defining the shortest path from a sample to the decision boundary can be generalized in this regard.

REFERENCES

Ramin Barati, Reza Safabakhsh, and Mohammad Rahmati. Towards explaining adversarial examples phenomenon in artificial neural networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7036–7042, 2021. doi: 10.1109/ICPR48806.2021.9412367.

Augustin Chevallier, Sylvain Pion, and Frédéric Cazals. Improved polytope volume calculations based on hamiltonian monte carlo with boundary reflections and sweet arithmetics. 2020.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkeryxBtPB.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Gamaleldin Fathy Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. 2018. URL https://arxiv.org/pdf/1803.05598.pdf.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Steven George Krantz. *Function theory of several complex variables*, volume 340. American Mathematical Soc., 2001.

T. Needham. *Visual Complex Analysis*. Comparative Pathobiology - Studies in the Postmodern Theory of Education. Clarendon Press, 1998. ISBN 9780198534464. URL https://books.google.com/books?id=ogz5FjmiqlQC.

Mahsa Paknezhad, Cuong Phuc Ngo, Amadeus Aristo Winarto, Alistair Cheong, Beh Chuen Yang, Wu Jiayang, and Lee Hwee Kuan. Explaining adversarial vulnerability with a data sparsity hypothesis, 2021.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance. *CoRR*, abs/1901.10861, 2019. URL http://arxiv.org/abs/1901.10861.

Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Thomas Tanay and Lewis Griffin. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

A FUNCTIONS OF SEVERAL COMPLEX VARIABLES

In this section, we give a brief introduction to the theory of functions of several complex variables. For a review of the subject see Needham (1998) and Krantz (2001). In order to establish notation, we first review the case of one complex variable. Let $\Omega \subset \mathbb{C} \sim \mathbb{R}^2$ be a domain in complex plane and let $f(x, y) = u(x, y) + iv(x, y)$ represent a complex-valued function on Ω. We write,

$$z = x + iy, \qquad \overline{z} = x - iy, \tag{19}$$

$$dz = dx + idy, \qquad d\overline{z} = dx - idy. \tag{20}$$

Then the differential of f at a point a is given by

$$\begin{aligned} df(a) &= \frac{\partial f}{\partial x}(a)dx + \frac{\partial f}{\partial y}(a)dy \\ &= \frac{1}{2}(\frac{\partial f}{\partial x} + \frac{1}{i}\frac{\partial f}{\partial y})(a)dz + \frac{1}{2}(\frac{\partial f}{\partial x} - \frac{1}{i}\frac{\partial f}{\partial y})(a)d\overline{z}. \end{aligned} \tag{21}$$

the following notation is introduced:

$$\frac{\partial f}{\partial z} = \frac{1}{2}(\frac{\partial f}{\partial x} + \frac{1}{i}\frac{\partial f}{\partial y}), \quad \frac{\partial f}{\partial \overline{z}} = \frac{1}{2}(\frac{\partial f}{\partial x} - \frac{1}{i}\frac{\partial f}{\partial y}) \tag{22}$$

to simplify equation 21 to

$$df(a) = \frac{\partial f}{\partial z}(a)dz + \frac{\partial f}{\partial \overline{z}}(a)d\overline{z}. \tag{23}$$

For a differentiable map, one has complex differentiability at a when the Cauchy-Riemann condition (C-R condition) holds at a:

$$\frac{\partial f}{\partial \overline{z}}(a) = 0. \tag{24}$$

If a function f is complex differentiable on every point of Ω, then it is called holomorphic.

Holomorphic functions are the main object of interest in our discussion. The interest in these functions comes from the fact that conformal maps and analytic functions are holomorphic. A conformal map is a complex-valued function in which the contours of the real and imaginary parts of that function are perpendicular. An analytic function is a function that is equal with its Taylor series. The

equivalence of complex differentiability, conformality and analyticity is a profound, nontrivial result in complex analysis. In other words, complex differentiability, analyticity and conformality of a complex-valued function are one and the same with holomorphicity of that function.

In case of $\Omega \subset \mathbb{C}^n$, a function $f : \Omega \to \mathbb{C}$ is complex differentiable if it is holomorphic in every dimension separately and we write

$$\overline{\partial}_j f = \frac{\partial f}{\partial \overline{z}_j} = \frac{1}{2}\left(\frac{\partial f}{\partial x_j} - \frac{1}{i}\frac{\partial f}{\partial y_j}\right) = 0 \quad j = 1, \cdots, n \tag{25}$$

or $\overline{\partial} f = 0$ for short and refer to this system of equations as the $\overline{\partial}$ equation.

We say that a sequence $\{f_j\}_{j=1}^{\infty}$ converges compactly on Ω if $\{f_j\}$ converges uniformly on each compact subset of Ω. In our discussion, a significant fact about $\mathcal{O}(\Omega)$ is that it is closed under compact convergence.

Theorem A.1 *Suppose $\{f_j\}_{j=1}^{\infty} \subset \mathcal{O}(\Omega)$ converges compactly in Ω to the function $f : \Omega \to \mathbb{C}$. Then $f \in \mathcal{O}(\Omega)$ and for each $\alpha \in \mathbb{N}^n$,*

$$\lim_{j \to \infty} \partial^\alpha f_j = \partial^\alpha f$$

compactly in Ω.

In theorem A.1 we have used the multi-index notation $\partial^\alpha = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_n^{\alpha_n}$.

B Proofs of the theorems

B.1 Theorem 2.1

First assume that samples are 1D. Without loss of generality we can assume that $h(c) = 0$. Since h is smooth, the real and imaginary part of $h(x) = u(x) + iv(x)$ could be approximated using a linear complex function g in a neighborhood of c,

$$g(x) = (x - c)\big(u'(c) + iv'(c)\big). \tag{26}$$

Since it is assumed that u is a linear robust classifier, we know from (Tanay & Griffin, 2016) that $\mathcal{C}(h)$ and $\mathcal{S}(h)$ are perpendicular on c. Thus, $v'(c)$ is a 90 degree rotation of $u'(c)$. Since multiplication by i could be interpreted as another 90 degree rotation, we can conclude that $u'(c) + iv'(c) = 0$. Hence, if u and v are functions of a real variable, they are forced to be constant around c,

$$\frac{du}{dx}(c) + i\frac{dv}{dx} = 0 \Rightarrow \frac{du}{dx}(c) = \frac{dv}{dx}(c) = 0. \tag{27}$$

In the case that x is a complex variable, $u'(c) + iv'(c) = 0$ would result in the Cauchy-Riemann condition,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \tag{28}$$

$$\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{29}$$

As a result, we can conclude that h is holomorphic in a neighborhood of a point $c \in \mathcal{C}(h)$.

In the case where samples are d dimensional, similar argument follows. First approximate h using linear functions around c. Then, fix all the coordinates of c except one, e.g. x_j. Similar to above, we reach the conclusion that h needs to satisfy Cauchy-Riemann condition in x_j. Since the choice of x_j was arbitrary, we conclude that h needs to satisfy the C-R condition in every dimension. Hence h is holomorphic in a vicinity of $c \in \mathcal{C}(h)$.

B.2 Theorem 2.2

Since $h \in A^2(\Omega)$ is holomorphic, it could be represented using a power series,

$$h(z) = \sum_{0 \leq |\alpha|} w_\alpha \varphi_\alpha(z), \tag{30}$$

in which $\{\varphi\}$ is a set of complete and orthonormal polynomials on Ω. Consequently, we define the truncated hypotheses space $\mathcal{H}_n^\Omega$,

$$\mathcal{H}_n^\Omega = \{h(z) \mid \sum_{0\leq|\alpha|\leq n} w_\alpha \varphi_\alpha(z)\}. \tag{31}$$

Thus, the VC-dimension of the real (or imaginary) part of $h \in \mathcal{H}_n^\Omega$ cannot get larger than the VC-dimension of the space of polynomial hypotheses of degree n. Hence, it is finite. Furthermore,

$$A^2(\Omega) = \bigcup_{n=0}^{\infty} \mathcal{H}_n^\Omega. \tag{32}$$

Thus, $A^2(\Omega)$ is nonuniform learnable.

B.3 THEOREM 2.4

Since $h_1, h_2 \in A^2(\Omega)$, they could be represented in a power series form,

$$h_j(z) = \sum_{0\leq|\alpha|} w_{j,\alpha}\varphi_\alpha(z). \tag{33}$$

Since h_j is trained on S_j, we can conclude that loss of h_j would be less than some ϵ_j with probability $1-\delta_j$ and that (ϵ_j, δ_j) are determined by the sample complexity of $A^2(\Omega)$ and the size of S_j.

We know that $o_\mathcal{D}$ has the lowest possible loss ϵ^* and that $o_\mathcal{D}$ can be represented in a similar form to h_j. Thus,

$$h_j(z) = o_\mathcal{D}(z) + \sum_{0\leq|\alpha|} a_{j,\alpha}\varphi_\alpha(z). \tag{34}$$

Hence,

$$\sup_{z\in\Omega} |\sum_{0\leq|\alpha|} a_{j,\alpha}\varphi_\alpha(z)| \leq \epsilon_j + \epsilon^*, \tag{35}$$

with probability $1-\delta_j$. Consequently,

$$h_1(z) - h_2(z) = \sum_{0\leq|\alpha|} a_{1,\alpha}\varphi_\alpha(z) - \sum_{0\leq|\alpha|} a_{2,\alpha}\varphi_\alpha(z) \tag{36}$$

Thus,

$$\sup_{z\in\Omega} |h_1(z) - h_2(z)| \leq 2\epsilon^* + \epsilon_i + \epsilon_j, \tag{37}$$

with probability $(1-\delta_1)(1-\delta_2)$.

We conclude that as we increase the size of the training samples, $\sup_z |h_1(z) - h_2(z)|$ gets smaller. As a result, for large enough training sets, $h_1(z)$ and $h_2(z)$ would be close in value.

B.4 THEOREM 3.1

Since $h \in A^2(\Omega)$ is holomorphic, it is equal with its Taylor series expansion around a point $z = \gamma(t)$,

$$h(z+\eta) = \sum_{0\leq|\alpha|} \frac{\partial^\alpha h(z)}{\alpha!}\eta^\alpha. \tag{38}$$

in which η is an arbitrary direction. Without loss of generality we can assume that all elements of η except η_j are zeros and that $\eta_j = \frac{d\gamma_j}{dt}(t)$. Then,

$$h(z+\eta) = h(z) + \eta_j \partial_j h(z). \tag{39}$$

Consequently,

$$\frac{d\gamma_j}{dt}(t) = \frac{dh}{\partial_j h(\gamma(t))}. \tag{40}$$

B.5 THEOREM 3.2

The length of a C^1 curve $\gamma : [0, 1] \to \Omega$ in $A^2(\Omega)$ is calculated as follows,

$$\|\gamma\| = \int_0^1 \Big(\sum_{j,k} g_{i,j}(\gamma(t))\gamma_i'(t)\overline{\gamma_j'(t)} \Big)^{\frac{1}{2}} dt, \tag{41}$$

in which $\{g_{i,j}\}$ is a Hermitian metric called the Bergman metric.

As depicted in the figure 1, we only need to use the fact that $h(z)$ is a holomorphic chart of a complex analytic manifold. Then, the theorem would simply result from the triangle inequality.

$$\|\gamma_z\| \leq \|\gamma_c\| + \|\gamma_s\| \Rightarrow \|\gamma_z\| - \|\gamma_s\| \leq \|\gamma_c\|. \tag{42}$$

B.6 THEOREM 4.1

The analogue hard margin problem of the holomorphic maximum margin classification rule is as follows,

$$\begin{aligned} \underset{w,\,b}{\arg\min} \quad & \frac{1}{2} w^H \Sigma w \\ \text{subject to} \quad & 1 \leq t_m \Re[h(z_m)] \quad m = 1, \cdots, M, \\ & \Im[h(z_m)] = 0 \quad m = 1, \cdots, M \end{aligned} \tag{43}$$

in which Σ is the synchronization matrix. Thus, the Lagrangian is defined as follows,

$$\mathcal{L} = \frac{1}{2} w^H \Sigma w + \sum_{m=1}^{M} \lambda_m (1 - t_m \Re[h(z_m)]) + \theta_m \Im[h(z_m)]. \tag{44}$$

We know,

$$\Re[z] = \frac{1}{2}(z + \overline{z}) \quad \Im[z] = \frac{1}{2i}(z - \overline{z}). \tag{45}$$

First, we solve for b,

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{m=1}^{M} t_m (\lambda_m + i\frac{\theta_m}{t_m}), \tag{46}$$

$$= \sum_{m=1}^{M} t_m \alpha_m. \tag{47}$$

Thus,

$$\sum_{m=1}^{M} t_m \alpha_m = 0. \tag{48}$$

Next, we solve for w,

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{1}{2}\Sigma\overline{w} - \sum_{m=1}^{M} t_m \lambda_m \frac{\partial}{\partial w} \Re[h(z_m)] + \theta_m \frac{\partial}{\partial w} \Im[h(z_m)], \tag{49}$$

in which we have used matrix calculus notation. In other words, $\frac{\partial \mathcal{L}}{\partial w}$ is a vector.

Thus,

$$w = \sum_{m=1}^{M} t_m \alpha_m \Sigma^{-1} \varphi. \tag{50}$$

We get the dual function by replacing w and b in the Lagrangian,

$$g(\lambda, \theta) = \sum_{m=1}^{M} \Re[\alpha_m] - \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} t_n t_m \overline{\alpha}_n \alpha_m \varphi_n^H \Sigma^{-1} \varphi_m. \tag{51}$$

Consequently, the dual problem of the holomorphic hard maximum margin is as follows,

$$\begin{aligned}
\underset{\lambda,\theta}{\arg\max} \quad & \sum_{m=1}^{M} \Re[\alpha_m] - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} t_n t_m \overline{\alpha}_n \alpha_m \varphi_n^H \Sigma^{-1} \varphi_m, \\
\text{subject to} \quad & \sum_{m=1}^{M} t_m \alpha_m = 0, \\
& 0 \leq \lambda_m \quad m = 1, \cdots, M, \\
& 0 \leq \theta_m \quad m = 1, \cdots, M
\end{aligned} \tag{52}$$

Since the primal problem is convex and Slater's condition is trivial for the equality constraints, strong duality holds.

C Detailed experiments

C.1 Synchronized polynomials

We will be needing the kernels of the orthonormal and synchronized polynomial features of the unit disk $\mathbb{D}$. Finding the kernel of the orthonormal features is straightforward. It is known that,

$$\begin{aligned}
\gamma_j &= \int_{\mathbb{D}} |z|^{2j}\, dV(z), \\
&= \frac{\pi}{j+1}, \\
\varphi_j(z) &= \frac{z^j}{\sqrt{\gamma_j}} \quad j \in \{0, \cdots, \infty\},
\end{aligned} \tag{53}$$

is an orthonormal basis for $A^2(\mathbb{D})$. Thus, the kernel of the orthonormal features is calculated as follows,

$$\begin{aligned}
K(z,\zeta) &= \sum_{j=1}^{\infty} \frac{j+1}{\pi} (z\overline{\zeta})^j, \\
&= \frac{1}{\pi(1-z\overline{\zeta})^2} - \frac{1}{\pi}.
\end{aligned} \tag{54}$$

In other words, the kernel of the orthonormal polynomials on Ω is the Bergman kernel of Ω minus the term that is contributed by the zero degree polynomial. Similarly,

$$\begin{aligned}
\gamma_j^* &= \int_{\mathbb{D}} j^2 |z|^{2(j-1)}\, dV(z), \\
&= j\pi, \\
\varphi_j^*(z) &= \frac{z^j}{\sqrt{\gamma_j^*}} \quad j \in \{1, \cdots, \infty\},
\end{aligned} \tag{55}$$

are the synchronized polynomials on $\mathbb{D}$. Thus, the kernel of the synchronized features is as follows,

$$\begin{aligned}
K^*(z,\zeta) &= \sum_{j=1}^{\infty} \frac{(z\overline{\zeta})^j}{j\pi}, \\
&= \frac{1}{\pi}\mathrm{Li}_1(z\overline{\zeta}), \\
&= -\frac{1}{\pi}\ln(1-z\overline{\zeta}),
\end{aligned} \tag{56}$$

in which Li_1 is the polylogarithm function of order 1.

We show that training a polynomial that is represented by a synchronized basis is robust. As our first experiment, we choose a training set S_n which is generated by the following rule,

$$\begin{aligned} x_{nj} &= 2\frac{j}{n} - 1, \\ z_{nj} &= x_{nj} + 0.25i \sin \pi x_{nj}, \\ S_{n+1} &= \{(z_{nj}, \operatorname{sign}(x_{nj}) \,|\, j = 0, \cdots, n\}. \end{aligned} \tag{57}$$

Next, we train two classifiers that make use of either an orthonormal set of polynomial features or a set of synchronized polynomials. We choose the following hypotheses spaces to apply our learning rule to,

$$\mathcal{H}_d = \{h(z) \,|\, h(z) = w_0 + \sum_{j=1}^{d} w_j \varphi_j(z) \quad w_k \in \mathbb{C}\}, \tag{58}$$

$$\mathcal{H}_d^* = \{h(z) \,|\, h(z) = w_0 + \sum_{j=1}^{d} w_j \varphi_j^*(z) \quad w_k \in \mathbb{C}\}, \tag{59}$$

where φ_j and φ_j^* are defined by equation 53 and equation 55.

Figure 2 visualizes the result of applying a learning rule similar to problem 11 to $\mathcal{H}_{20}$ and $\mathcal{H}_{20}^*$ with the sole difference that the integral in the objective is replaced by $\|w\|$. In the case of $\mathcal{H}_{20}$, the learning rule is similar to the standard maximum margin learning rule, and for $\mathcal{H}_{20}^*$ it is equivalent with the holomorphic maximum margin learning rule. We have also colored the figure using a domain coloring technique to help with the identification of $\mathcal{S}(h)$, $\mathcal{C}(h)$ and $\mathcal{Z}(h)$. The hue of the colors represent $\arg h(z)$ with cyan equaling π or $-\pi$, red 0, green-yellow $\frac{\pi}{2}$ and purple-blue $-\frac{\pi}{2}$ and the saturation of the colors represent $|h(z)|$. The zeros are identified with the property that all the possible hues surround them. Figure 2 shows that the decision boundary of the hypothesis with orthonormal features has curved back and became parallel with the submanifold of the samples near the boundary of the unit disk.

On the other hand, the holomorphic maximum margin learning rule resulted in a hypothesis that did not show a similar trait, and the decision boundary continued in a path perpendicular to the submanifold of the samples. We argue that Figure 2 describes the nonlinear version of the boundary tilting perspective that is conjectured by Tanay & Griffin (2016). The pattern of the colors also marks the domain of convergence of the hypothesis, and we can see that the synchronized features have resulted in a larger domain of convergence.

In the last experiment, the hypotheses were complex analytic. Next, we examine the scenario where the hypotheses are real analytic functions, in which the pointwise limit of a sequence of analytic functions is not guaranteed to be an analytic function. Similar to Barati et al. (2021), we choose the training samples to be,

$$S_{n+1} = \{(x, \operatorname{sign}(x)) \,|\, x = 2\frac{j}{n} - 1 \quad j = 0, \cdots, n\}. \tag{60}$$

Then, we train two maximum margin classifiers that use the orthonormal and the synchronized Chebyshev basis, respectively. It is known that,

$$\frac{dT_n(x)}{dx} = nU_{n-1}(x), \tag{61}$$

where T_n and U_{n-1} are Chebyshev basis of the first and second kind, respectively. $\{T_n\}$ form a sequence of orthonormal polynomials on $[-1, 1]$ with respect to the weight function $w(x) = \frac{1}{\sqrt{1-x^2}}$,

$$\int_{-1}^{1} T_n(x)\, T_m(x) w(x)\, dx = \begin{cases} 0 & \text{if } n \neq m, \\ \pi & \text{if } n = m = 0, \\ \frac{\pi}{2} & \text{if } n = m \neq 0. \end{cases} \tag{62}$$

$$\int_{-1}^{1} U_n(x)\, U_m(x) w(x)\, dx = \begin{cases} 2\pi \lceil \frac{\min(n,m)}{2} \rceil & \text{if } n, m \text{ odd}, \\ \pi + 2\pi \lfloor \frac{\min(n,m)}{2} \rfloor & \text{if } n, m \text{ even}, \\ 0 & \text{otherwise.} \end{cases} \tag{63}$$

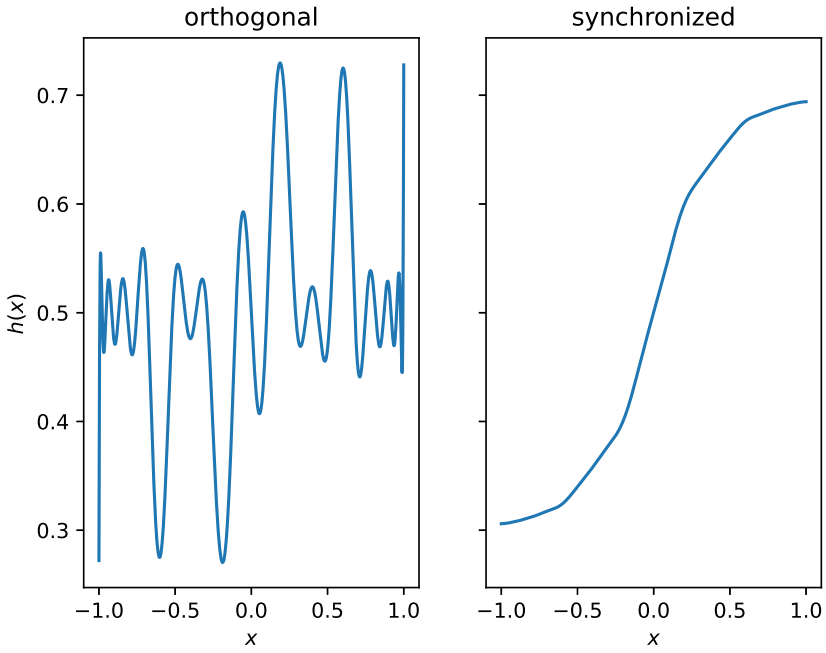

Figure 3: 1D orthonormal and synchronized Chebyshev hypotheses on the real line. We have used the basis functions up to degree 30 and S_6 in both cases.

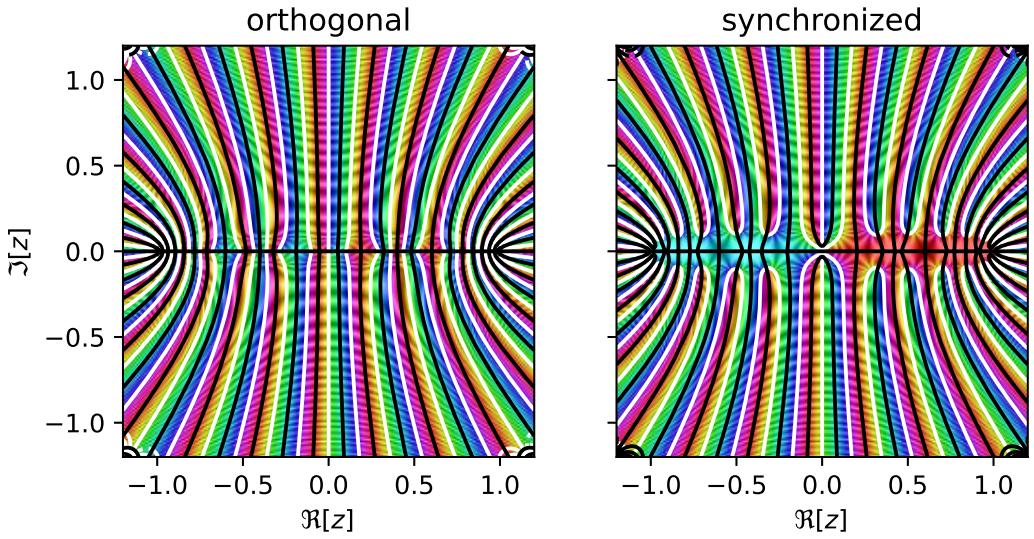

Figure 4: 1D orthonormal and synchronized hypotheses on the complex plane. The domain of convergence of the hypotheses does not cover the whole unit disk contrary to expectation. In case of orthonormal features, the domain of convergence does not even cover the whole of $[-1, 1]$ interval.

Similar to orthonormalization, the effect of weight functions on synchronization is to determine the importance of each point in the domain. We calculate the synchronization matrix of $\{T_n\}$ using equation 61 and equation 63.

We can see the results for training a Chebyshev polynomial of degree 30 on S_6 in Figures 3 and 4. Figure 3 shows the classifiers on the real line, and it demonstrates that the synchronized hypothesis is robust. Figure 4 shows the hypothesis on the complex plane. In this figure, we can see that $\mathcal{S}(h)$ and $\mathcal{C}(h)$ are perpendicular, irrespective of the robustness of the hypotheses. The phenomenon is enabled by parts of $\mathcal{S}(h)$ that have branched out from the real line into the complex plane. It is deducible from the figures that the synchronized hypothesis has maximized the distance between the training set and $\mathcal{Z}(h)$ and has a bigger domain of convergence.

Figure 4 suggest that adversarial regions are not always a result of a tilted boundary, but they are a result of copies of the decision boundary that are present in the complex plane and cross $\mathcal{S}(h)$ multiple times. These copies of the decision boundary are also available in the synchronized hypothesis, but they do not cross $\mathcal{S}(h)$ in a neighborhood of the synchronization domain.

Since powers of z are holomorphic and form an orthogonal basis on the unit disk, we expect that the trained hypothesis be holomorphic on the unit disk $\mathbb{D}$. However, it is evident from Figure 4 that the domain of convergence does not cover the whole disk in either hypotheses. A similar effect is observable in function approximation when we approximate analytic functions with singularities using polynomials. This observation suggests that the holomorphic optimal Bayes classifier is singular somewhere on the complex plane. The holomorphic optimal Bayes classifier $o_{\mathcal{D}}(z)$ is calculated as follows,

$$\begin{aligned} o_{\mathcal{D}}(z) &= \int_{\mathbb{D}} f_{\mathcal{D}}(\zeta) K(z,\zeta)\, dV(\zeta) \\ &= \int_{-\pi}^{\pi} \int_{0}^{1} \frac{\operatorname{sign}(r\cos\theta)}{\pi(1-zre^{-i\theta})^2}\, r dr d\theta \\ &= \frac{i}{\pi}(\ln(-z-i) - \ln(-z+i)) + \cdots . \end{aligned} \tag{64}$$

where

$$f_{\mathcal{D}}(z) = \operatorname{sign}(\Re z), \tag{65}$$

is the optimal Bayes classifier and

$$K(z,\zeta) = \frac{1}{\pi(1-z\bar{\zeta})^2}, \tag{66}$$

is the Bergman kernel of $\mathbb{D}$.

The existence of adversarial regions could be analyzed from two different perspectives; the holomorphic maximum margin classifier's primal and the dual formulations. Nevertheless, the duality of the two formulations suggests that both of these interpretations are the same and that it would be incorrect to prefer one over the other.

The primal formulation suggests that the phenomenon occurs because the derivative of the trained hypotheses is diverging. From equation 64 it is evident that $o_{\mathcal{D}}(z)$ has two logarithmic branch points on i and $-i$ and that it is a transcendental function. Thus, it cannot be expressed in terms of a finite sequence of algebraic operations. Consequently, the space of polynomial hypotheses of finite degree is agnostic to $o_{\mathcal{D}}(z)$. As a result, the learning rule has to compensate for this shortcoming somehow. The singularities in the complex plane force the learning rule to prefer higher degree polynomials to lower degree ones, even when the decision boundary could be easily represented using a linear hypothesis. Since a nonrobust learning rule only seeks the uniform convergence of h, it is fooled into choosing a hypothesis that does not uniformly converge in its derivative, producing adversarial regions near the training points.

The dual of the holomorphic maximum margin optimization problem suggests another perspective on the phenomenon. An orthonormal set of polynomial features would result in a kernel that has a pole for its singularity as demonstrated by equation 54 whereas the singularities of equation 64 are logarithmic branch points. The discrepancy between the singularities results in the wrong hypothesis to be learned by the maximum margin learning rule. In contrast, the kernel of a synchronized set of

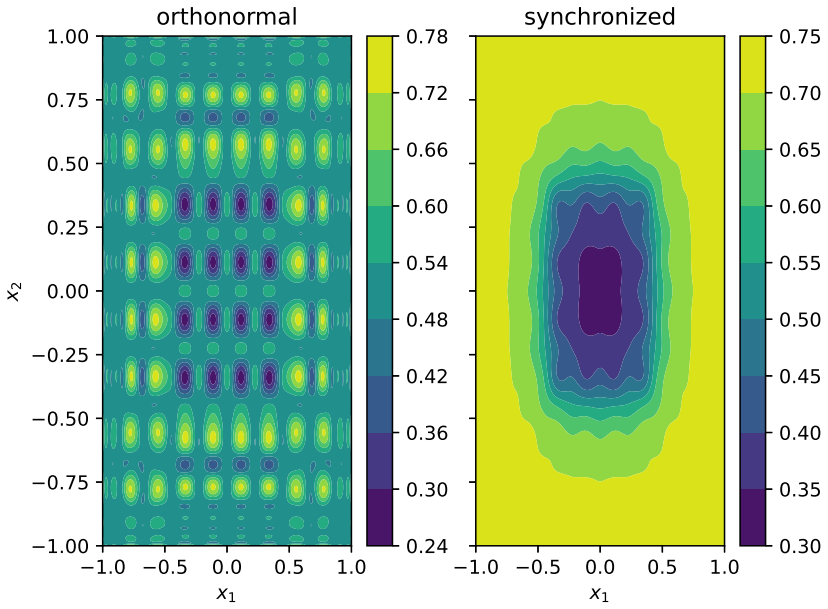

Figure 5: 2D orthonormal and synchronized hypotheses. We have used basis functions up to degree 30 in both dimensions and 100 equispaced samples to train the classifiers.

polynomials exhibits the correct type of singularity as demonstrated by equation 56, which would then be able to represent the optimal hypothesis correctly.

We conducted an experiment on a 2D synthetic problem using orthonormal and synchronized 2D Chebyshev polynomials as well and reported the result in Figure 5. The optimal Bayes classifier is

$$f_{\mathcal{D}}(x) = \operatorname{sign}(\|x\|_\infty - 0.5) \tag{67}$$

in this experiment and the synchorization matrix is computed using equation 61, equation 62, equation 63. The results agree with the dimpled manifold model in the sense that the decision boundary appear to be weaving through the submanifold of the samples. Nonetheless, adversarial regions are positioned between the training points. This observation is not predicted by the dimpled manifold model.

C.2 Synchronization in real-world problems

For our next experiment, we train a few maximum margin classifiers on the UCI ML handwritten digits dataset (Dua & Graff, 2017) that use orthonormal and synchronized Chebyshev polynomials as features. The samples are 8-by-8 black and white pictures of handwritten digits. The reason for choosing this dataset over MNIST is that it is relatively low dimensional compared to the MNIST family while being complicated enough to be a good representation of a real-world dataset. Nevertheless, the combinatorial explosion of possible polynomial features is still severe. To keep the training procedure computationally feasible, we limit the degree of the polynomials in a manner inspired by Markov random fields. More precisely, we first create an 8-by-8 lattice graph and connect each node to itself. Then, we enumerate all unique walks on this graph up to a certain length. The nodes decide the uniqueness of a walk. For example,

$$(3,2) \to (3,2) \to (3,3)$$

and

$$(3,2) \to (3,3) \to (3,2)$$

are considered the same. Finally, we assign each dimension to its corresponding node in the graph and choose the degree by counting the occurrence of each node in a walk. For example,

$$(3,2) \to (3,2) \to (3,3),$$

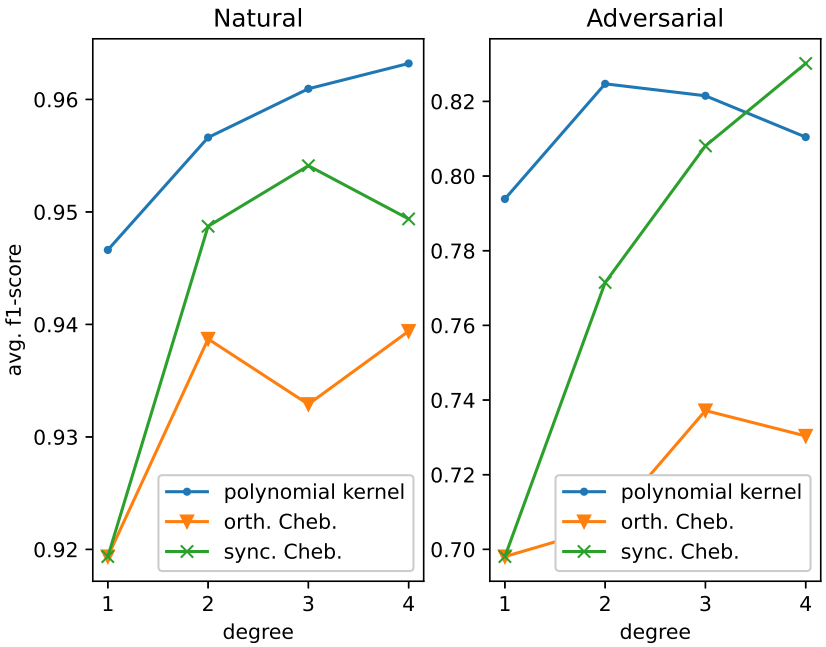

Figure 6: The weighted average f1-score of the models in natural and adversarial settings on the digits dataset.

corresponds to $T_2(x_{32})T_1(x_{33})$. We calculate the synchronization matrix with respect to the normalized weight function $w(x) = \frac{1}{\pi\sqrt{1-x^2}}$, so that Σ_{jk} does not become so large that elements of $\Sigma^{-\frac{1}{2}}$ approach zero.

After splitting the dataset into train and test sets using 50% of samples for each, we train maximum margin classifiers with polynomial kernel up to degree 4 on the train set and attack these classifiers using a l_2 normalized, one-step attack on the test set. Next, we train new classifiers for Chebyshev polynomials with the same degree and compute the weighted average of f1-score of the classifiers on the natural test samples and the adversarial test samples of the polynomial kernel classifier. We have reported the results of the experiment in Figure 6.

Figure 6 reports the accuracy of maximum margin classifiers that were trained on the UCI ML handwritten digits dataset (Dua & Graff, 2017) and used different polynomials as features on natural and adversarial test samples. One of the models uses the polynomial kernel, and the other two use the orthonormal and synchronized Chebyshev polynomials as features. We have used the classifier with the polynomial kernel as the baseline since it employs all the polynomial features up to a certain degree. In contrast, due to the combinatorial explosion of possible polynomial features in high dimensions, the Chebyshev polynomial features were selected based on a scheme founded on the position of the pixels in the image. The adversarial examples are generated by attacking the baseline using a one-step, ℓ_2 normalized, gradient-based attack.

We can see that as we increase the degree of the polynomial, the model's performance improves on the natural samples as expected. In the adversarial setting, the performance of the models increases as we move from a linear hypothesis to a polynomial of degree 2. This observation follows the linearity hypothesis of Goodfellow et al. (2014), and we conclude that a linear hypothesis is not strong enough to represent the robust optimal hypothesis. However, as we increase the degree of the polynomial, the performance of non-synchronized hypotheses is dominated by the effect of the diverging derivative and worsens, a trait that is not shared with the synchronized hypothesis. We present these results as evidence for useful nonrobust features that are essential in constructing a robust hypothesis if regularized correctly.

Table 1: The accuracy of the neural and the Bergman hypothesis on the adversarial examples in the direct and transferred scenarios. The results suggest that the two hypothesis represent the same decision boundary.

		source	
		MLP	BKM
target	MLP	0.07	0.30
	BKM	0.09	0.27

C.3 TRANSFER OF ADVERSARIAL EXAMPLES

This section is dedicated to examining the transfer of adversarial examples between holomorphic hypothesis. In an experiment we compare the adversarial examples of a kernel machine with the Bergman kernel and an artificial neural network on the MNIST dataset. For the neural network hypothesis, we use a standard multi-layer perceptron (MLP) with $\tanh$ activation function and the space of kernel machine hypotheses (BKM) is as follows,

$$\mathcal{H}_M = \{h(z) \,|\, h(z) = w_0 + \sum_{m=1}^{M} w_m K(z, \zeta_m)\}, \tag{68}$$

in which K is the Bergman kernel of the d-dimensional poly-disk $\mathbb{D}^d$,

$$K(z, \zeta) = \frac{1}{\pi^d} \prod_{j=1}^{d} \frac{1}{\pi(1 - z\bar{\zeta}_j)^2}. \tag{69}$$

We scale the pixel values of samples of MNIST to $[-1, 1]$ range so that the samples fall into the poly-disk. We train both hypotheses using gradient descent and cross-entropy loss with ℓ_2 regularization of the parameters of the hypotheses. The size of the MLP was $784 \times 512 \times 512 \times 10$, and the kernel machine had 100 kernels. Both hypotheses use softmax as the final activation function. The parameters of the MLP are real numbers, whereas the parameters of the kernel machine are complex. In our experiments, the parameters of the Bergman kernels ζ_j had to be pure imaginary; otherwise, the learning rule would fail. As of this writing, we have not been able to find the reason behind this requirement, but we guess it has something to do with the singularities of the Bergman kernels.

After training for ten epochs, the accuracy of the neural and Bergman hypotheses on the test set is 0.88 and 0.91, respectively. Then, we attack each hypothesis with a white-box, one-step, l_∞-normalized gradient-based attack. The results of the experiment are reported in Table 1. The result shows that the adversarial examples of the two hypotheses transfer, and the performance of the hypotheses are almost identical.

To better demonstrate the extent of similarity between the two hypotheses, we graphed each hypothesis in the neighborhood of the adversarial path in the complex plane in Figure 7. To be more precise, if s_n and s_a represent the natural and adversarial samples, let

$$\eta = s_n - s_a \tag{70}$$

to represent the adversarial direction for sample s_n and k and l to represent the label that the hypothesis assign to the natural and adversarial examples, respectively. Figure 7 visualizes the following function for each hypothesis h,

$$g(z) = h_k(s_n + z\eta) - h_l(s_n + z\eta) \quad z \in \mathbb{C}. \tag{71}$$

We emphasize that the hypothesis h is not activated with softmax in $g(z)$. Since we compare the real parts of h_k and h_l to decide the label of the sample, $\mathcal{C}(g)$ is the decision boundary between classes k and l. Considering that our analysis does not cover the case of multiclass classification, we will leave the interpretation of $\mathcal{S}(g)$ to future works.

From the figure, we can see that the position of the submanifold of the decision boundary is similar in both hypotheses. The only difference is that the multi-layer perceptron is a periodic function in the complex plane due to its activation function. Nevertheless, this seems to be a technical issue, and both hypotheses are identical with regard to the geometrical position of their decision boundaries as far as we are concerned. This observation supports our proposal that all analytic hypotheses converge to the holomorphic optimal Bayes classifier regardless of their implementations.

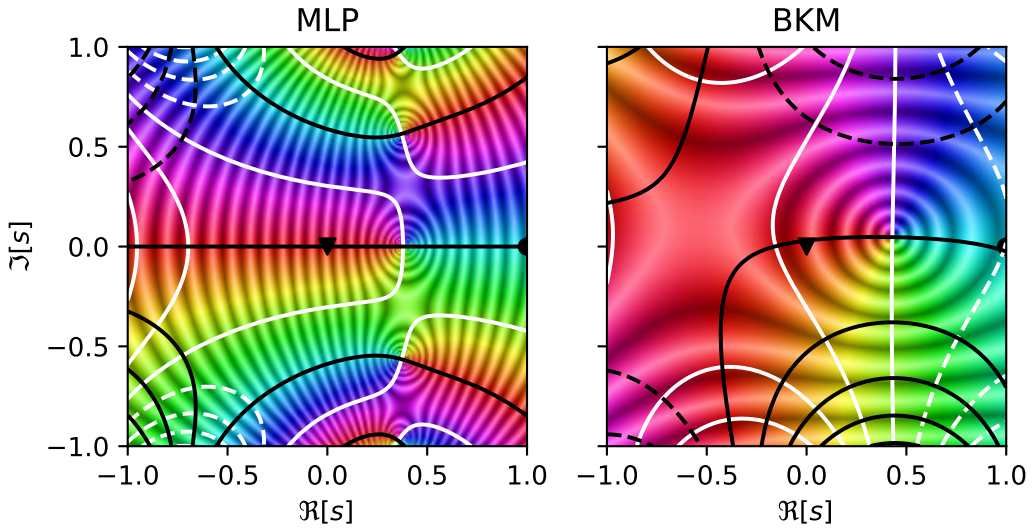

Figure 7: Visualization of the slice of a Bergman kernel machine (BKM) and a multi-layer perceptron (MLP) in the adversarial direction in the complex plane for an arbitrary test sample. The triangle and the circle represent the position of the natural and adversarial samples respectively.

D HARDNESS OF ROBUST LEARNING

We start the discussion from the perspective of the primal problem of holomorphic maximum margin classification. The exact computation of the objective of the primal problem is NP-hard in general. For example, consider a neural network with a single hidden layer and ReLU activation and suppose that the samples are points in $\mathcal{X} = [0, 1]^d$. To compute the synchronization matrix of the neurons, we have to compute

$$\begin{aligned}\Sigma_{jk} &= \int_{\mathcal{X}} (\nabla \sigma_j(x))^T (\nabla \sigma_k(x))\, dV(x), \\ &= w_i^T w_j \int_{\mathcal{X}} H_j(x) H_k(x)\, dV(x),\end{aligned} \tag{72}$$

in which $H_j(x)$ is the Heaviside step function indicating the half-space where neuron j is activated. Hence, we can see that computing Σ_{jk} is equivalent to computing the volume of a somewhat arbitrary polytope. It is known that this problem is $\mathcal{P}\#$-complete, and thus it is NP-hard (Chevallier et al., 2020). Consequently, training a maximum margin neural classifier is a stochastic optimization problem and is outside the scope of this paper.

Next, we will inspect the solution from the perspective of the dual of holomorphic maximum margin classification problem. In Section 3, we showed that it would be possible to train robust classifiers if we have the kernel corresponding to the synchronized features on a domain. We also calculated the kernel for the unit disk $\mathbb{D}$. Here, we will try to do the same for the poly-disk $\mathbb{D}^d$. First, we need a basis for $A^2(\mathbb{D}^d)$. It is known that $\{z^\alpha\}$ is an orthogonal basis for $A^2(\mathbb{D}^d)$ in which we have used the multi-index notation,

$$\begin{aligned}\alpha &= (\alpha_1, \alpha_2, \cdots, \alpha_d) \quad \alpha_j \in \mathbb{N}_0, \\ z^\alpha &= \prod_{j=1}^{d} z_j^{\alpha_j}.\end{aligned} \tag{73}$$

Similar to equation 55, the synchronized polynomial features are calculated as follows,

$$\begin{aligned}\gamma_\alpha^* &= \int_{\mathbb{D}^d} \sum_{j=1}^{d} |\frac{\partial z^\alpha}{\partial z_j}|^2 \, dV(z), \\ &= \frac{\pi^d \sum_{j=1}^{d} \alpha_j(\alpha_j+1)}{\prod_{k=1}^{d}(\alpha_k+1)}, \\ \varphi_\alpha^*(z) &= \frac{z^\alpha}{\sqrt{\gamma_\alpha^*}}. \end{aligned} \tag{74}$$

Thus, the kernel of the synchronized polynomial features of $\mathbb{D}^d$ is computed by,

$$K^*(z,\zeta) = \sum_\alpha \frac{\prod_{k=1}^{d}(\alpha_k+1)}{\pi^d \sum_{j=1}^{d} \alpha_j(\alpha_j+1)} (z\bar{\zeta})^\alpha. \tag{75}$$

There is no simple expression for $K^*(z,\zeta)$ to the best of our knowledge, and we have to leave this solution to future works as well.

It seems that there is no easy way to solve the holomorphic maximum margin optimization problem in high dimensions. Nevertheless, it is still possible to simplify the hypotheses space so that robust training becomes computationally feasible. We can achieve efficiency by limiting α, i.e., assuming independence between dimensions and only using certain degrees of polynomials as features.